ОРГАНИЗАЦИОННЫЙ КОМИТЕТ КОНФЕРЕНЦИИ



Председатель
Богуш В.А., ректор учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», д-р физ.-мат. наук, профессор

ЧЛЕНЫ ОРГАНИЗАЦИОННОГО КОМИТЕТА



Араш Косари, доцент кафедры электротехники и информационных технологий «Иранская исследовательская организация по науке и технологиям», канд. техн. наук, доцент, Исламская Республика Иран



Батура М.П., научный руководитель НИЛ 6.3 учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», д-р техн. наук, профессор, академик Международной академии наук высшей школы, заслуженный работник образования Республики Беларусь



Давыдов М.В., первый проректор учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. техн. наук, доцент



Дик С.К., заведующий кафедрой инженерной и компьютерной графики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. физ.-мат. наук, доцент



Исмаилова А.А., PhD, ассоциированный профессор, и.о. профессора ГОП «Информационные системы», НАО «Казахский агротехнический исследовательский университет имени С.Сейфулина», Республика Казахстан



Казак Т.В., заведующий кафедрой инженерной психологии и эргономики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», д-р психол. наук, профессор



Касанин С.Н., заместитель генерального директора по научной работе Государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси»



Кругликов С.В., главный научный сотрудник Государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси»



Кудж С.А., ректор федерального государственного бюджетного образовательного учреждения высшего образования «МИРЭА – Российский технологический университет», д-р техн. наук



Лихачевский Д.В., заместитель начальника государственного учреждения «Секретариат Наблюдательного совета Парка высоких технологий», канд. техн. наук, доцент



Мезилов Г.А., ректор «Инженерно-технологического университета Туркменистана имени Огуз-хана», д-р техн. наук



Пархименко В.А., заведующий кафедрой экономики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. экон. наук, доцент



Петренко В.И., заведующий кафедрой организации и технологии защиты информации ФГАО ВО «Северо-Кавказский федеральный университет», канд. техн. наук, доцент



Пискун Г.А., заместитель декана факультета компьютерного проектирования учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. техн. наук, доцент



Рябова А.Н., начальник Государственного учреждения «Секретариат Наблюдательного совета Парка высоких технологий»



Стемпицкий Р.А., проректор по научной работе учреждения образования «Белорусский государственный университет информатики радиоэлектроники», канд. техн. наук, доцент



Тебуева Т.Б., заведующая кафедрой прикладной математики и компьютерной безопасности ФГАОУ ВО «Северо-Кавказский федеральный университет (г. Ставрополь), д-р физ.-мат. наук, доцент, Российская Федерация



Тузиков А.В., заведующий лабораторией математической кибернетики Государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси», д-р физ.-мат. наук, профессор



Тусупов Д.А., заведующий кафедрой информационных систем Евразийского национального университета им. Л.Н. Гумилева, д-р физ.-мат. наук, профессор, Республика Казахстан



Ульянов Н.И., декан факультета компьютерных систем и сетей учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. техн. наук, доцент



Целых А.А., заместитель директора Института компьютерных технологий и информационной безопасности Южного федерального университета, канд. техн. наук, доцент, Российская Федерация



Шнейдеров Е.Н., проректор учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», канд. техн. наук, доцент

ОРГАНИЗАТОРЫ КОНФЕРЕНЦИИ

Министерство образования Республики Беларусь

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»

Национальная академия наук Беларуси

Государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси»

МИРЭА – Российский технологический университет,
Российская Федерация

ФГАОУ ВО «Северо-Кавказский федеральный университет»,
Российская Федерация

Евразийский национальный университет имени Л. Н. Гумилева,
Республика Казахстан

СПОНСОР КОНФЕРЕНЦИИ

Министерство образования
Республики Беларусь

Резидент Парка высоких технологий КБК ТехЛаб
Команда опытных разработчиков и инженеров, которые создают современные технологические решения для бизнеса. Помогают компаниям трансформировать идеи в эффективные цифровые продукты.

Резидент Парка высоких технологий ISsoft
Компания ISsoft – один из крупнейших белорусских разработчиков IT решений для рынков США и Западной Европы. Основана в Минске в 2004 г. как дочерняя компания корпорации Coherent Solutions, Inc. (США).
Резидент Парка Высоких Технологий с 2007 года.

ОГЛАВЛЕНИЕ

К.С. Николаев Методика построения модели определения параметров технологического режима. Использование метода векторного опорных векторов 12

П.Ю. Бранцевич, И.В. Григорьева, П.П. Королевич, Е.Н. Базылев Методы обработки цифровых сигналов электроэнцефалограмм при исследовании их информативно-значимых параметров.....19

Н.А. Кириенко Преобразование иерархических транзисторных описаний комбинационных логических схем в системе логического проектирования.....31

М.П. Батура, И.В. Марахина, В.А. Пархименко К вопросу о системной оценке эффективности деятельности организаций в социальных сетях 37

В.В. Верняховская, О.М. Раптунович, В.А. Усова Эффективное управление большими данными: сравнительный анализ платформ и технологий для обработки информации.....44

Н.В. Щербина, Е.Н. Бычек Использование искусственного интеллекта в промышленной безопасности.....52

Г.Б. Никифоров, В.И. Пименов, Д.А. Андреев О соответствии автоматической процедуры разметки методом лучшего-худшего масштабирования решению задач анализа тональности финансовых новостей.....58

Г.К. Стамкулова, С.М. Салиев Проектирование центра мониторинга информационной безопасности (SOC) в инфраструктуре университета с применением Big Data и Advanced Analitycs.....64

Г.К. Стамкулова, Э. Накыпов Разработка программного решения для цифровизации экологического мониторинга и управления отходами 73

М.С. Борисенков, С.Н. Шульженко Формирование признакового пространства для предиктивного анализа состояния дисковых подсистем на основе атрибутов S.M.A.R.T. .80

Д.А. Чернов, С.Н. Шульженко Сравнительный анализ применения графовых моделей при оценке качества программных систем на языке GO и других языках программирования.....85

А.В. Казаков Разработка голосового помощника для приемной комиссии университета на основе BIG DATA и RAG-архитектуры.....90

Ю.С. Шевнина, А.В. Прокофьев, О.А. Ашарафьянова Прогнозирование состояния образовательного процесса с применением нечетких когнитивных карт и алгоритмов нелинейного обучения.....93

D.M. Rahel Predicting target variable values and constructing a ROC curve 100

D.M. Rahel Consumer behavior models in the mobile communications market.....103

Э.Е. Дадашев, М.Д. Самакаева Выравнивание динамических рядов в аналитике больших данных в цифровой экономике 107

А.М. Найда, Э.Д. Кочешкова Интеграция классических методов стратегического анализа и моделей Big Data в системе принятия управленческих решений 114

Н.К. Крючков, А.В. Воробей, И.В. Андриалович Программное обеспечение д ля автоматизированной оценки эффективности инвестиционных проектов 121

И.П. Кобяк Большие скорости и нанопараметры частиц в подпространствах атома водорода.....131

Н.Ф. Николаев, А.В. Макаров Гибридные методы машинного обучения в прогнозировании качества продукции для потоковых данных в концепции Индустрия 4.0 138

З. Джафаров, О.В. Бойправ, Э. Имамалиев Применение метода машинного обучения KNN для краткосрочного прогнозирования интенсивности трафика в сетях 5G (на примере видеоконференций) 144

Л.В. Примакович, С.Ф. Миксюк Адаптивное прогнозирование страхового запаса: от модели к программной реализации в эпоху Big Data 149

Г.А. Хришкевич, В.И. Пименов, Д.А. Андреев Утраты монументальной живописи как показатели Big Data: объективно-ориентированное представление дефектов на основе многоканальной сегментации 157

B.A. AI-Nami The Role of Advanced Analytics and Big Data in Automating and Personalizing Graphic Design for Dynamic Web Interfaces in IT Solutions 164

Т.В. Казак, А.А. Ковальчук, А.Н. Василькова Выявление психоэмоциональных паттернов на основе анализа больших данных в цифровых платформах психологической поддержки 172

Н.И. Липницкая, А.Ю. Герасимович, Е.П. Ременчик Прогнозирование демографической пирамиды Республики Беларусь с использованием методов расширенной аналитики (Advanced Analytics) в Python 183

М.Т. Мырадов Построение высокоточной интеллектуальной системы обработки естественного туркменского языка на основе комбинированного трансформенного подхода 191

А.А. Гулис, М.А. Калугина Влияние методов обучения с подкреплением на стратегию поведения автономных агентов 195

Y. Liashchevich, Cai Yuxin Classic and modern encryption algorithms for Big Data processing 203

Д.В. Лихачевский, С.А. Мигалевич, В.О. Агапова Разработка программного комплекса для инвентаризации и мониторинга состояния основных средств организации 208

Д.А. Лагутик, А.Ю. Савицкий, А.Е. Козлов Графовые нейронные сети как направление развития в системах безопасности силовых ведомств 213

А.В. Бардашевич, А.Ю. Савицкий, В.А. Федоренко Применение векторных баз данных для организации поиска технической и учебной документации в области телекоммуникаций 220

П.А. Красев, М.А. Калугина Влияние методов дообучения на адаптацию языковых моделей для анализа процессов разработки 225

М.М. Ходжамаммедов Обнаружение аномалий на основе Big Data для выявления мошенничества в цифровых системах 231

Ван Там Лэ, С.М. Боровиков, А.В. Будник, С.К. Дик Практика оценки ожидаемой надежности прикладных компьютерных программ для информационных систем 234

Е.И. Полоско Комплексный статистический анализ больших данных: интеграция методов снижения размерности, кластеризации и регрессии 243

Н.И. Липницкая, М.А. Миткевич, В.В. Смертьев Модели машинного обучения для прогнозирования уровня тревожности по поведенческим и социально-демографическим признакам 251

Е.И. Полоско, П.Ю. Сороко, А.Д. Тригубович Оценка влияния посещаемости занятий на академическую успеваемость студентов с использованием Python-аналитики 258

И.В. Андриалович, И.П. Галяк, Е.А. Касперец Разработка легковесного алгоритма детекции «живости» лица (Liveness Detection) для систем веб-идентификации 265

К.А. Абраев Симбиоз невидимых сетей: технологии Lorawan и Big Data в архитектуре цифрового будущего 271

М.М. Ходжамаммедов Применение технологий Big Data для мониторинга и выявления мошеннических действий в реальном времени 274

В.В. Венгеренко Применение VLM для анализа визуальных сцен 276

А.В. Шкрабов, Н.А. Резников, С.К. Дик, В.М. Бондарик, И.И. Ревинская Интеграция систем генерации на основе поиска и семантического кэширования для повышения достоверности предметно-ориентированных LLM 286

Н.А. Резников, А.В. Шкрабов, С.К. Дик, Г.Д. Ситник, И.И. Ревинская Методика повышения устойчивости сверточной нейронной сети с использованием аугментации данных на примере модели для сегментации изображений глазных яблок 295

С.С. Бекарев, М.А. Калугина Неавторегрессионная модульная система синтеза речи с явным моделированием длительности на базе CTC-выравнивания 302

G.O. Orazdurdyyeva, M.B. Bekiyeva Adaptive multi-scale heterogeneous graph learning for real-time anomaly detection in industrial cyber-physical systems 308

Л.В. Примакович, И.П. Логинова Стратегии адаптивного распараллеливания в алгоритме сортировке шелла для задач предобработки данных в Big Data системах 314

С.Н. Резаи Вахид Гибридный подход к прогнозированию заполненности помещений для оптимизации энергопотребления высших учебных заведений 322

С.Н. Касанин, С.В. Кругликов О перспективах правового регулирования технологий искусственного интеллекта в Республике Беларусь 326

К.А. Хаджинова, А.Г. Савчиц, Р.В. Земляник, Д.А. Хаджинов Разработка алгоритма детекции светофоров и распознавание их сигналов для автономного транспорта 331

Ю.А. Засинец, В.А. Хуцишвили, М.В. Давыдов Интеллектуальная система поддержки принятия решений в диагностике дисфункций ВНЧС 337

V.A. Kovalev, E.V. Snezhko, D.S. Karpenko, A.G. Varvashevich Evaluating vulnerabilities and protections of ai methods in computer-assisted image-based diagnosis 341

А.В. Мосендз, Г.А. Хацкевич Влияние интеграции геймификации персонализации на удержание пользователей в цифровых образовательных приложениях 349

УДК 004.852

МЕТОДИКА ПОСТРОЕНИЯ МОДЕЛИ ОПРЕДЕЛЕНИЯ ПАРАМЕТРОВ ТЕХНОЛОГИЧЕСКОГО РЕЖИМА С ИСПОЛЬЗОВАНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ



К.С. Николаев
Старший преподаватель института
СПИНТех, НИУ МИЭТ
knpreacher@gmail.com

К.С. Николаев
Окончил Московский институт электронной техники. Область научных интересов связана с разработкой интеллектуальных информационных систем.

Аннотация. Предложена методика построения модели определения параметров технологического режима на основе метода опорных векторов (SVR). Методика включает очистку данных, исключение выбросов, стандартизацию признаков и подбор гиперпараметров с использованием кросс-валидации. Для каждого выходного параметра формируется отдельная модель. Результаты подтверждают эффективность подхода при ограниченном объёме данных. Разработанные модели позволяют сократить число экспериментов и решить задачу «холодного старта» при внедрении новых технологических процессов.

Ключевые слова: метод опорных векторов, SVR, машинное обучение, математическое моделирование, технологический режим, подбор гиперпараметров, кросс-валидация, обработка данных, исключение выбросов, холодный старт.

Введение. Технологический режим в контексте лазерной обработки представляет собой набор (вектор) данных из числовых и категориальных значений. Эти значения делятся на группы [1], а именно:

- параметры материала – физические и химические свойства исходного материала обрабатываемого образца (например, теплоемкость);
- исходные параметры образца – физические параметры непосредственно обрабатываемого образца, например, толщина;
- исходные параметры установки – параметры и диапазоны, настроенные разработчиком установки;
- варьируемые параметры установки – значения, которые выставляет технолог для проведения операции (например, частота повторения импульсов, контурная скорость);
- результирующие параметры – параметры, полученные в ходе технологической обработки (например, в случае лазерной резки – ширина и глубина реза).

На рисунке 1 представлен пример организации данных технологических режимов в информационной системе.

Измеряемые					Задаваемые					
Геометрия изделия		Качество изделия		Производительность	Лазер			Кинематическая система машины	Пневмо и газовая система	
Ширина реза *мкм*	Глубина реза *мкм*	Зона загрязнения поверхности *мкм*	Зона термического влияния *мкм*	Время обработки изделия *с*	Частота повторения импульсов *кГц*	Длительность импульса *пс*	Средняя мощность *Вт*	Контурная скорость *мм/мин*	Тип защитного газа	Давление защитного газа *бар*
Асферическая (16 мм)										
200		510	100	60	25	0,2	8,2	500,000000	Воздух	6
190		410	100	60	25	0,4	8,2	500,000000	Воздух	6
200		215	111	60	25	0,78	8,2	500,000000	Воздух	6
200	600	189	100	60	25	4,6	8,2	500,000000	Воздух	6

Рисунок 1. Пример организации данных технологических режимов в информационной системе

Проблемная ситуация обусловлена противоречием между высокой степенью нелинейности и многофакторности физико-химических процессов, протекающих в зоне взаимодействия лазерного излучения с веществом, и ограниченными возможностями эмпирических методов настройки оборудования, используемых инженерно-техническим персоналом [2].

Существующая практика формирования технологического режима базируется преимущественно на накопленном экспертном опыте операторов и итеративном методе «проб и ошибок».

Данный подход характеризуется системными ограничениями, среди которых: многомерность пространства параметров (полный перебор комбинаций в ручном режиме является ресурсоемким и экономически нецелесообразным), стохастическая природа процессов (наличие случайных возмущений) и отсутствие универсальных детерминированных моделей (ввиду сложности теплофизических явлений, таких как фазовые переходы, гидродинамика расплава, плазмообразование [3]).

В рамках решения задачи определения параметров технологического режима следует рассмотреть применение искусственных нейронных сетей (ИНС) и методов машинного обучения.

Нейронные сети обладают универсальным свойством приближения: при достаточном числе нейронов они способны аппроксимировать произвольную непрерывную функцию f: $R^n \to R$ с заданной точностью.

Однако практическая реализация данного свойства требует выполнения ключевого условия – наличия репрезентативной обучающей выборки большого объема.

Для задач моделирования технологических режимов характерны следующие особенности:

- малый объем выборки – получение экспериментальных данных связано с проведением реальных испытаний, затратами материалов, времени и производственных ресурсов;
- высокая стоимость эксперимента – каждая точка данных соответствует реальному технологическому прогону, что делает невозможным формирование больших датасетов, типичных для задач компьютерного зрения или обработки текста;
- гетерогенность и неполнота данных – технологические параметры могут иметь пропуски, различную шкалу измерений, а также быть подвержены случайным возмущениям, что требует дополнительной обработки и увеличивает неопределенность модели.

Соответственно, несмотря на высокую гибкость и нелинейность аппроксимации, нейронные сети не являются оптимальным инструментом для решения задачи определения параметров технологического режима при ограниченном объеме экспериментальных данных. В условиях малых выборок более обоснованным представляется использование методов, основанных на принципе структурной минимизации риска, в частности метода опорных векторов (SVM).

Сам по себе SVM обычно используется в задачах классификации [4].

В контексте решаемой задачи, следует использовать расширение SVM – метод опорных векторов для регрессии (SVR).

Цель метода – построить функцию $f(x)=\langle w,\phi(x)\rangle+b$, которая отклоняется от истинных значений не более чем на заданную величину и при этом обладает минимальной сложностью.

Разработка методики построения модели определения параметров технологического режима. Разработка универсальной модели определения параметров технологического режима представляет собой сложную научно-техническую задачу вследствие существенной разнородности физических процессов, лежащих в основе различных технологических операций (резка, сварка, абляция и др.).

Каждый из указанных процессов характеризуется собственной физической природой, отличающимися механизмами тепломассопереноса, фазовыми превращениями, динамикой расплава, а также спецификой взаимодействия инструмента или энергетического источника с обрабатываемым материалом.

Так, при резке доминируют процессы локального плавления и удаления материала; при сварке – формирование сварочной ванны и кристаллизация; при абляции – интенсивное испарение и плазмообразование.

Соответственно, выходные показатели качества (глубина проплавления, ширина шва, шероховатость, зона термического влияния и др.) могут отличаться, в зависимости от выполняемой операции.

Методика построения модели определения параметров технологического режима для конкретной операции определяется следующими этапами:

1. Определение входных и выходных числовых параметров режима, согласно специфике операции;
2. Предобработка данных (обработка пропусков и выбросов, нормализация);
3. Проверка на статистическую зависимость входных параметров между собой;
4. Обучение модели SRV для каждого выхода отдельно (т.е. требуется обучить столько моделей, сколько выходных выходных параметров) и подбор гиперпараметров C;
5. Оценка качества (RMSE, MAE, R^2).

Рассмотрим применение описанной методики на примере импульсной лазерной резки.

Задача состоит в том, чтобы по требованиям заказчика (параметры образца и критерии качества) определить значения варьируемых параметров станка.

В случае резки имеем (таблица 1):

Таблица 1. Перечень входных и выходных параметров импульсной лазерной резки

Входные параметры	**Выходные параметры**
Ширина реза (мкм) Глубина реза (мкм) Зона загрязнения поверхности (мкм) Зона термического влияния (мкм) Время обработки (с) Толщина материала (мкм)	Частота повторения импульсов (кГц) Длительность импульса (пс) Средняя мощность (Вт)

Имеем исходный набор размера (1542; 9). Фрагмент набора данных представлен на рисунке 2.

cut_width	cut_depth	surface_co	termo_affe	product_pr	material_th	repeat_freq	pulse_dura	avg_power
0.0	20.0	0.0	0.0	0.0	0.0	10.0	0.01	0.3
0.0	20.0	0.0	0.0	0.0	0.0	25.0	0.01	1.5
0.0	400.0	0.0	0.0	0.0	0.4	25.0	0.01	10.0
0.0	400.0	0.0	0.0	0.0	0.4	25.0	0.01	8.0
0.0	400.0	0.0	0.0	0.0	0.4	25.0	0.01	5.0
0.0	400.0	0.0	0.0	0.0	0.4	25.0	0.01	7.0
50.0	0.0	0.0	0.0	0.0	0.0	50.0	13.0	19.0
50.0	420.0	15.0	115.0	2.0	1.0	50.0	20.0	18.0
40.0	300.0	70.0	0.0	0.0	0.3	50.0	15.0	15.5
50.0	150.0	77.0	0.0	0.0333	0.3	10.0	40.0	13.0
80.0	300.0	115.0	0.0	0.0	0.4	25.0	0.01	4.0
22.0	400.0	77.0	1205.0	0.0	0.4	25.0	0.01	5.0
83.0	300.0	95.0	520.0	0.0	0.4	25.0	0.01	8.0
75.0	300.0	110.0	316.0	0.0	0.8	25.0	0.01	5.0
90.0	300.0	130.0	220.0	0.0	0.8	25.0	0.01	5.0
80.0	300.0	115.0	240.0	0.0	0.8	25.0	0.01	4.0
42.0	450.0	60.0	300.0	0.005	0.8	50.0	15.0	18.0

Рисунок 2. Фрагмент исходных данных

На следующем шаге уберем неполные и пустые записи. В качестве пропуска значения параметра в исходном наборе данных используется число 0. Но, согласно физическому смыслу, некоторые параметры действительно могут иметь нулевое значение. Будем считать, что запись неполная, если нулевых параметров в ней 7 и более.

Далее, следует очистить данные от выбросов, применив метод межквартильного размаха (IQR) [5]. Вычисляется первый и третий квартили Q1 (25%) и Q3 (75%) как медианы нижней и верхней половин данных. Сам размах следует считать по формуле IQR=Q3-Q1. Получив значение размаха, определим границы допустимых значений: zmin=Q1-k ·IQR, zmax=Q3+k ·IQR, где k – коэффициент чувствительности (k=1.5). Соответственно, будем считать значение выбросом, если оно не принадлежит отрезку [zmin, zmax]. Этот метод применяется для всех столбов исходных данных поочередно, и в случае нахождения выбросов удаляется полная запись.

После применения указанных операций размер выборки составил (1361; 9). Для нормализации данных используется стандартизация признаков (z-нормализация) – xi=xi-ii.

В рамках анализа экспериментальных данных технологических режимов целесообразно предварительно оценить статистическую зависимость между входными параметрами, поскольку высокая коррелированность признаков может приводить к избыточности информации [6], ухудшению интерпретируемости модели и нестабильности оценки гиперпараметров.

Построим матрицы корреляций по Пирсону и Спирмену, чтобы удостовериться в статистической независимости параметров (рисунок 3, 4). По получившимся матрицам можно полагать, что параметры статистически независимы.

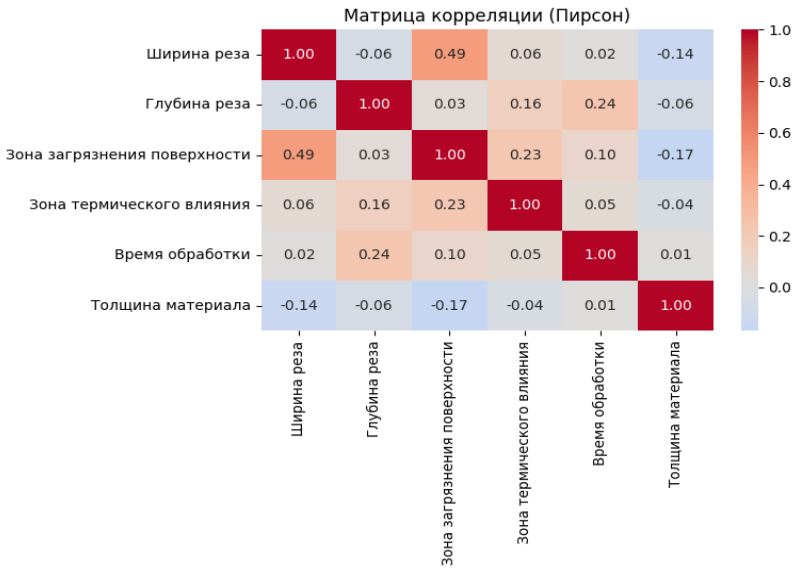


Рисунок 3. Матрица корреляций по Пирсону

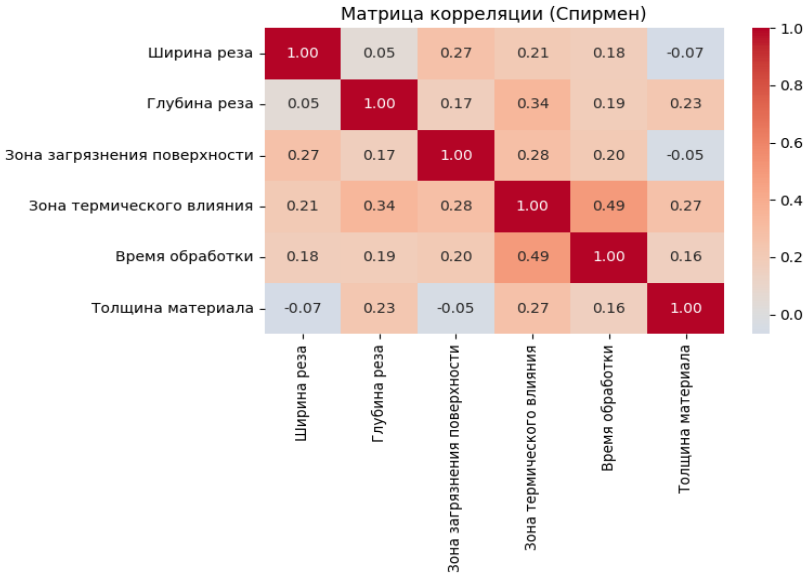


Рисунок 4. Матрица корреляций по Спирмену

Для обучения модели SVR используется метод Grid Search («сетка»; это метод подбора гиперпараметров, основанный на полном переборе заранее заданных значений параметров модели) [7]. В качестве ядра будет использоваться радиальная базисная функция (RBF), что позволяет учитывать нелинейные зависимости, измеряя «близость» объектов в пространстве признаков. Значения сетки параметров представлены в таблице 2.

Таблица 2. Значения сетки параметров для обучения SVR

Параметр обучения	**Значения**
C	[1, 10, 100, 300]
	[0.01, 0.05, 0.1, 0.2]
	["scale", 0.01, 0.05, 0.1, 0.2]

Значение «scale» для параметра означает, что параметр ядра RBF вычисляется автоматически на основе дисперсии входных данных, т.е. $\gamma = 1n \cdot \mathrm{Var}(x)$, где n – число признаков; $\mathrm{Var}(x)$ – дисперсия данных обучающей выборки.

Стоит отметить, что вследствие использования нормализации по стандартным отклонениям $\mathrm{Var}(x) \approx 1$.

Параметры SVR, полученные после обучения, представлены в таблице 3.

Таблица 3. Результаты обучения SVR

Параметр режима	C			**CV RMSE**	**RMSE**	**MAE**	**R^2**
Частота повторения импульсов	10	0.1	0.2	0.6522	0.6176	0.3406	0.5044
Длительность импульса	10	0.01	0.1	0.5329	7.2359	1.1615	-0.0257
Средняя мощность	100	0.01	scale	0.6001	0.3014	0.1278	0.9457

Проведем анализ полученных результатов:

- Частота повторения импульсов – ошибка на тесте сопоставима с кросс-валидацией, а значит, нет оснований предполагать переобучение; модель объясняет около 50% дисперсии.
- Длительность импульса – $R^2 < 0$, что означает, что модель непригодна для определения значения параметра. Причина такого поведения в исходном наборе данных: 84% значений этого параметра не заполнены.
- Средняя мощность – нет оснований предполагать переобучение; модель объясняет более 94% дисперсии.

Графическая иллюстрация полученных результатов представлена на рисунке 5.

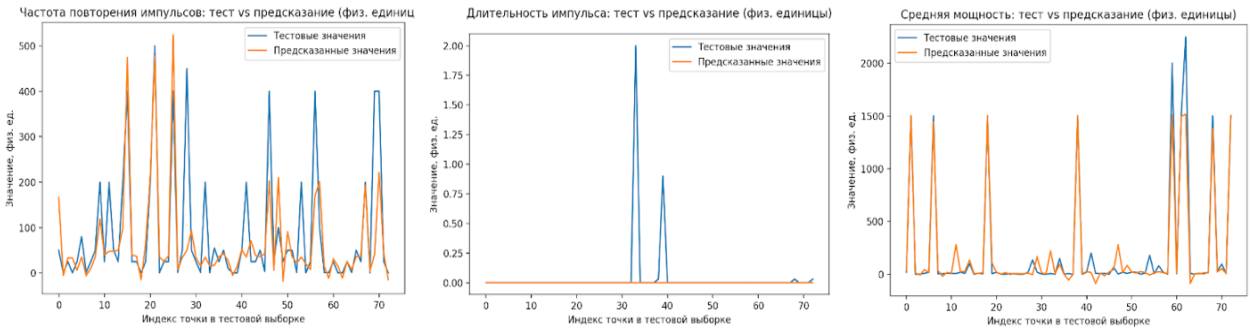


Рисунок 5. Тестовые и предсказанные значения на графиках

Разработанные по описанной методике модели могут представлять практическую ценность для инженеров-технологов.

Даже в случае получения не идеально точных прогнозных значений модель позволяет сформировать обоснованную стартовую оценку технологического режима.

Это способствует сокращению количества реальных экспериментов, снижению затрат времени и материалов, а также решению задачи «холодного старта» при внедрении нового материала, оборудования или технологического процесса.

Использование модели в качестве инструмента предварительного расчёта позволяет сузить область поиска оптимального режима и повысить эффективность последующих экспериментов.

Заключение. В рамках настоящего исследования была разработана методика построения модели определения параметров технологического режима на основе метода опорных векторов (SVR).

Полученные результаты показали, что для большинства исследуемых выходных характеристик модель SVR обеспечивает высокую или удовлетворительную точность прогнозирования.

Результаты исследования подтверждают целесообразность применения метода опорных векторов для моделирования технологических режимов в условиях ограниченного объема данных.

Список литературы

[1] Magdum V. B., Kittur J. K. A review on laser beam machining process modeling and optimization //IOP Conference Series: Materials Science and Engineering. – IOP Publishing, 2021. – Т. 1104. – №. 1. – С. 012015.

[2] Cabrera M. L. et al. An overview of strategies for identifying manufacturing process window through design of experiments and machine learning techniques while considering the uncertainty associated with //The International Journal of Advanced Manufacturing Technology. – 2024. – Т. 134. – №. 11. – С. 4981-5019.

[3] Агаповичев А. В., Сотов А. В., Смелов В. Г. Математическое моделирование процесса селективного лазерного сплавления порошка титанового сплава ВТ6 //VESTNIK of Samara University. Aerospace and Mechanical Engineering. – 2020. – Т. 19. – №. 2. – С. 53-62.

[4] Yu H., Kim S. SVM Tutorial-Classification, Regression and Ranking //Handbook of Natural computing. – 2012. – Т. 1. – С. 479-506.

[5] Дорофеев В. С., Волосатова Т. М. Ансамблирование методов обнаружения выбросов при подготовке обучающей выборки данных //Моделирование, оптимизация и информационные технологии. – 2022. – Т. 10. – №. 3. – С. 1.

[6] Dhal P., Azad C. A comprehensive survey on feature selection in the various fields of machine learning //Applied intelligence. – 2022. – Т. 52. – №. 4. – С. 4543-4581.

[7] Lameski P. et al. SVM parameter tuning with grid search and its impact on reduction of model over-fitting //Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing: 15th International Conference, RSFDGrC 2015, Tianjin, China, November 20-23, 2015, Proceedings. – Cham : Springer International Publishing, 2015. – С. 464-474.

Авторский вклад

Николаев Константин Сергеевич – постановка задачи исследования, разработка методики построения модели определения параметров технологического режима, оценка полученных результатов.

METHODOLOGY FOR BUILDING A MODEL FOR DETERMINING TECHNOLOGICAL MODE PARAMETERS USING THE SUPPORT VECTOR MACHINE

K.S. Nikolaev

Senior Lecturer, SPINTECH Institute,
National Research University MIET

Abstract. A methodology for constructing a model for determining technological mode parameters based on the support vector method (SVR) is proposed. The methodology includes data cleaning, outlier exclusion, feature standardization, and hyperparameter selection using cross-validation. A separate model is formed for each output parameter. The results confirm the effectiveness of the approach with a limited amount of data. The developed models make it possible to reduce the number of experiments and solve the problem of a “cold start” when introducing new technological processes.

Keywords: support vector method, SVR, machine learning, mathematical modeling, technological mode, hyperparameter selection, cross-validation, data processing, outlier exclusion, cold start.

УДК 616.89-08:004.3:004.42:004.9

МЕТОДЫ ОБРАБОТКИ ЦИФРОВЫХ СИГНАЛОВ ЭЛЕКТРОЭНЦЕФАЛОГРАММ ПРИ ИССЛЕДОВАНИИ ИХ ИНФОРМАТИВНО-ЗНАЧИМЫХ ПАРАМЕТРОВ



П.Ю. Бранцевич
Профессор кафедры программного обеспечения информационных технологий БГУИР, доктор технических наук
branc@bsuir.by



И.В. Григорьева
Ученый секретарь ГУ «Республиканский научно-практический центр психического здоровья», кандидат медицинсских наук
Inessgrig@mail.ru



П.П. Королевич
Врач психиатр-нарколог, врач функциональной диагностики, научный сотрудник отдела психических и поведенческих расстройств ГУ «Республиканский научно-практический центр психического здоровья».
p.korolevich@gmail.com



Е.Н. Базылев
Старший преподаватель кафедры программного обеспечения информационных технологий БГУИР, исследователь
e.bazylev@bsuir.by

П.Ю. Бранцевич

С 1985 года работает в области разработки архитектуры, методов, алгоритмов и программного обеспечения компьютерных систем и комплексов для решения задач испытаний, вибрационного контроля, мониторинга, диагностики и автоматики защиты сложных механизмов и агрегатов с вращательным движением. С 1995 по 2019 годы научный руководитель НИЛ «Систем вибродиагностики» БГУИР. Научный руководитель заданий ГНТП, ГПНИ, хозяйственных договоров и контрактов.

И.В. Григорьева

Занимается научной, лечебной и преподавательской деятельностью. Сфера научных и практических интересов: психотерапевтическая кинезиология, арт-терапия, аддиктология, психоонкология, нейролингвистическое программирование, эриксоновский гипноз, трансактный анализ, психокатарзис, танатотерапия.

П.П. Королевич

Окончил Белорусский государственный медицинский университет. Высшая квалификационная категория. Окончил ординатуру по специальности «психиатрия-наркология», окончил аспирантуру. Специалист в области электросудорожной терапии, электроэнцефалографии. Принимает участие в исследованиях особенностей биоэлектрической активности головного мозга пациентов с психическими и поведенческими расстройствами

Е.Н. Базылев

Окончил Белорусский государственный университет информатики и радиоэлектроники. Окончил аспирантуру с квалификацией исследователь. Принимает участие в выполнении заданий ГПНИ.. Разрабатывает аппаратные и программные средства встроенных систем.

Аннотация. Цифровые сигналы электроэнцефалограмм (ЭЭГ) отражают определённое состояние нейронной сети головного мозга человека. Рассмотрены методы обработки сигналов ЭЭГ, определения их информативно-значимых параметров и характеристик. Показана возможность сравнения ЭЭГ с использованием полосовых спектров. Предложен метод обработки сигналов ЭЭГ в узких частотных полосах, которые являются модулированными гармониками.

Ключевые слова: сигнал, электроэнцефалограмма, параметр, характеристика, цифровая обработка.

Введение. Важность обсуждения вопроса использования искусственного интеллекта в настоящее время превысила спрос на понимание работы естественного интеллекта. Таким образом отсутствие полного понимания функционирования естественного интеллекта затрудняет использование его потенциала и уменьшает социальную значимость. Представляется, что элементом естественного интеллекта является человек.

Человек – это органическая система принятия решений, управляемая самообучающейся нейронной сетью головного мозга, получающая с помощью органов чувств информацию о состоянии её носителя (самом человеке), об окружающей действительности, происходящих событиях и формирующая управляющие сигналы для органов человека, которые воспроизводят ответные реакции и действия, обладающая способностью создавать новые знания [1].

У человека пять основных органов чувств, позволяющих воспринимать мир: глаза (зрение), уши (слух), нос (обоняние), язык (вкус) и кожа (осязание), которые вместе с вестибулярным аппаратом (равновесие) получают соответствующие данные и передают их в нейронную сеть головного мозга для обработки и десять систем органов: опорно-двигательная, кровеносная, пищеварительная, дыхательная, выделительная, покровная, эндокринная, нервная, лимфатическая, половая, которые по указанию нейронной сети головного мозга воспроизводят ответные действия и реакции, что определяет его уникальность.

Нейронная сеть головного мозга человека включает поступательный механизм самообучения в течении процесса жизнедеятельности. Поэтому в настоящее время невозможно надеть шапочку, нажать кнопку и занести в нейронную сеть человека знания. Человек получает знания из той среды, в которой находится и не из какой-то другой. Однако вполне возможно, что через какое-то время будут созданы средства для занесения знаний в нейронную сеть головного мозга человека, извлечения знаний из нейронной сети головного мозга человека, и средства переноса состояния нейронной сети головного мозга человека в другую органическую или неорганическую среду.

Что касается отдельно взятого человека, как носителя интеллекта, то он получает первоначальную настройку своей нейронной сети, своих базовых способностей от родителей, но его дальнейшее развитие определяется средой и процессом самообучения. Эмоции – это тоже деятельность нейронной сети головного мозга человека. А интуиция всё-таки основывается на полученных знаниях. Интуитивные решения принимаются относительно той сферы, в которой человек находится и развивается. И эти интуитивные решения есть ни что иное, как создание новых знаний, что является основополагающим свойством естественного интеллекта.

Естественный интеллект можно рассматривать и «в большом» формате, как развитие разумной жизни человечества, когда каждое поколение получает знания от предыдущего поколения, создаёт новые знания и передаёт накопленные знания следующему поколению.

Нейронную сеть мозга образуют нейроны через сложные физические соединения, синапсы, формируя разветвленные цепи передачи информации. Они обмениваются электрическими импульсами, укрепляя связи при частом использовании и связи отмирают, если связь неактивна. Развитие идет от генетической основы к опыту и обучению [2, 3].

Нейроны состоят из тела, дендритов, которые принимают сигналы, и аксона, через которые производится передача сигнала. Передача информации происходит в местах контакта между нейронами, где сигнал переходит от аксона одного нейрона к дендриту другого. Когда

аксон одного нейрона приближается к дендриту другого, между ними возникает точка связи – синапс [2, 3].

Передача информации происходит не напрямую, а через выброс специальных веществ − нейромедиаторов. Они преодолевают синаптическую щель и активируют соседнюю клетку. Если два нейрона активируются одновременно или часто взаимодействуют, связь между ними укрепляется. Это называется синаптической пластичностью. Таким образом формируются устойчивые цепочки, отвечающие за память и навыки.

В мозге взрослого человека образуется около 100 триллионов таких связей. Новые пути возникают при обучении, путешествиях или занятиях творчеством. Для формирования сложной структуры отростков нейрону требуется около 40 – 45 дней, а для появления новой связи в результате практики − регулярные повторения.

Нейроны, которые активируются вместе, соединяются вместе. Повторение действий укрепляет связи, выделяя больше нейромедиаторов и увеличивая синапсы. Активные нейроны развивают новые дендриты (отростки), увеличивая количество связей (до 30000 на клетку). Неактивные связи отмирают, оптимизируя работу мозга.

Развитие нейронной сети включает пролиферацию (деление), миграцию (перемещение) и дифференцировку (созревание) стволовых клеток в нейроны с их последующим встраиванием в сеть. Процессы формирования связей продолжаются всю жизнь, завися от обучения, опыта и эмоционального состояния [2−4].

Работа головного мозга, вследствие прохождения электрических импульсов по аксонам нейронов, сопровождается изменением электромагнитного поля на поверхности головы, которое можно зафиксировать специальными первичными преобразователями и представить в виде изменяющихся параметров тока или напряжения, что и происходит, когда снимают электроэнцефалограмму (ЭЭГ) [4−8].

ЭЭГ – это запись колебаний электрического потенциала с поверхности головы, которые генерирует огромное количество нервных клеток головного мозга при своей работе. Эти потенциалы очень слабые, поэтому требуется специальная аппаратура для их регистрации. Слабые изменения электрических сигналов на поверхности головы человека улавливаются специальными электродами, которые устанавливаются на коже головы. Затем эти сигналы усиливаются, оцифровываются и передаются в память компьютера, откуда затем сохраняются в виде файлов на носителях информации.

Стандартной системой размещения электродов на поверхности головы, рекомендованной Международной федерацией электроэнцефалографии и клинической нейрофизиологии, является система «10−20%» [7, 8].

Данные файлов электроэнцефалограммы являются цифровыми сигналами, которые можно обрабатывать и исследовать с помощью методов цифровой обработки сигналов (ЦОС) [9−13]. Целью исследования ЭЭГ является определение информативно значимых признаков, по изменению которых можно делать выводы о нарушении работы нейронной сети головного мозга человека, либо об эффективности применяемых методик лечения. Учитывая трудность формального разграничения нормы и патологии, для ЭЭГ можно выделить, по крайней мере, три регистра: 1) нормальная ЭЭГ; 2) пограничные (между нормой и патологией) варианты; 3) патологические изменения (нарушения) биоэлектрической активности головного мозга. Нормальными считаются варианты ЭЭГ, соответствующие возрасту и функциональному состоянию обследуемого по частотным, амплитудным и пространственно-временным характеристикам. К патологическим видам относят варианты ЭЭГ, параметры которых значительно отличаются от нормальных. Чем «дальше» отстоят выявляемые изменения от нормы, тем тяжелее лежащие в их основе нарушения. В клинической нейрофизиологии наиболее часто употребляемой шкалой для качественной оценки тяжести регистрируемых нарушений является следующий ряд: легкие − умеренные − выраженные − грубые − очень грубые нарушения. Определяемые при визуально-логическом и математическом анализе незначительные изменения биоэлектрической активности головного мозга на ЭЭГ и

отсутствие явной патологической активности свойственны пограничным вариантам. Пограничные изменения на ЭЭГ могут рассматриваться как условно патологические: для их интерпретации необходимым условием является наличие или отсутствие клинических проявлений [4, 8].

Таким образом, важнейшей задачей при анализе ЭЭГ становится выявление и определение информативно-значимых признаков, которые наиболее достоверно отражают фактическое состояние обследуемого человека и позволяют дать объективную оценку действенности применяемых методов воздействия или лечения.

Цифровая обработка сигналов электроэнцефалограмм. Электроэнцефалограмма является набором цифровых электрических сигналом, отражающих спонтанную электрическую активность мозга в течение определенного периода времени, полученных с нескольких электродов на поверхности головы.

В настоящее время для оценки сигналов ЭЭГ анализируется мощность сигнала в частотных полосах, называемых: альфа (8-13 Гц); бета (13-35 Гц); гамма (35-70 Гц); дельта (0,3-4 Гц); тэта (4-8 Гц) [6, 8, 10]. Однако эти параметры являются весьма обобщающими и не в полной мере отражают процессы, происходящие в нейронной сети головного мозга. Поэтому для исследования параметров и характеристик ЭЭГ применяют и другие методы цифровой обработки сигналов, такие как спектральный анализ, вейвлет-анализ, преобразование Гильберта-Хуанга, частотный полосовой анализ, выделение огибающих [13−19]. Экспериментальные результаты применения спектрального анализа показывает, что при изменении частотного разрешения изменяются параметры гармонических составляющих, представляющих сигнал ЭЭГ на разных временных интервалах, при этом заметно определённое сохранение интенсивности сигнала в определённых частотных полосах [14, 15, 17, 18].

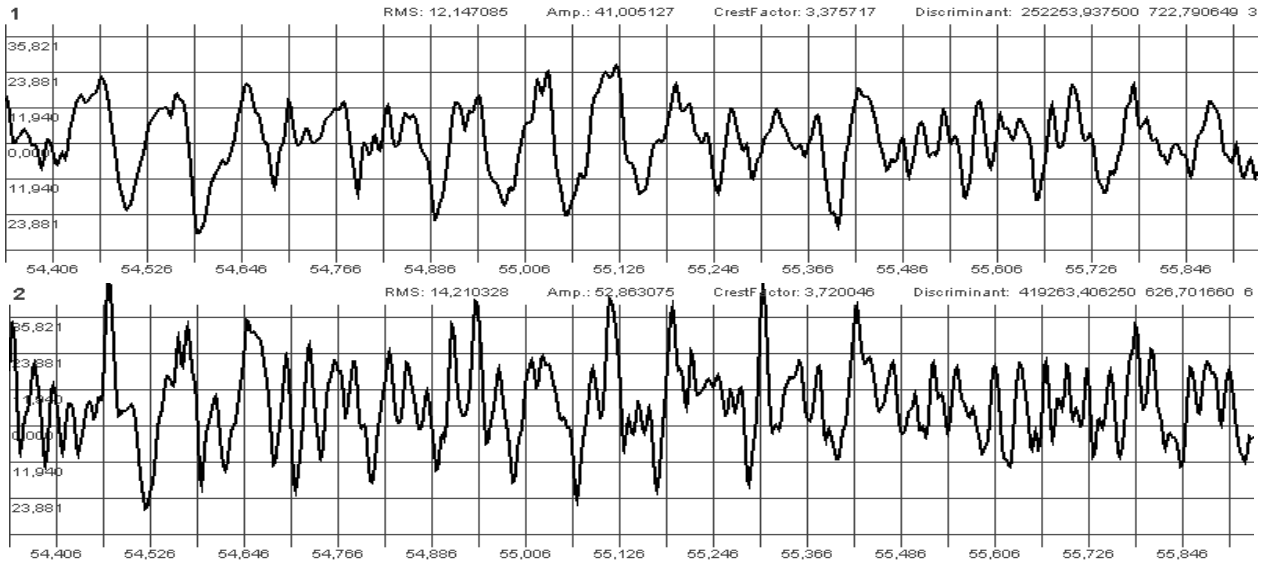


Рисунок 1. Сигналы ЭЭГ в точках контроля O1-A1 и O2-A2

Поэтому процедура выбора длительности временного интервала для определения параметров ЭЭГ требует дальнейшего исследования. Проведен ряд экспериментальных исследований параметров и характеристик ЭЭГ с целью выявления информативно-значимых параметров, по которым можно было бы проводить их кластеризацию. В качестве примера на рисунке 1 показаны временные реализации сигналов ЭЭГ, полученных при частоте дискретизации 500 Гц. Амплитудные спектры этих сигналов приведены на рисунке 2. Временной интервал анализа выбран равным примерно 16 секунд, с числом дискретных точек 8192. Наибольшая интенсивность ЭЭГ наблюдается в альфа-диапазоне.

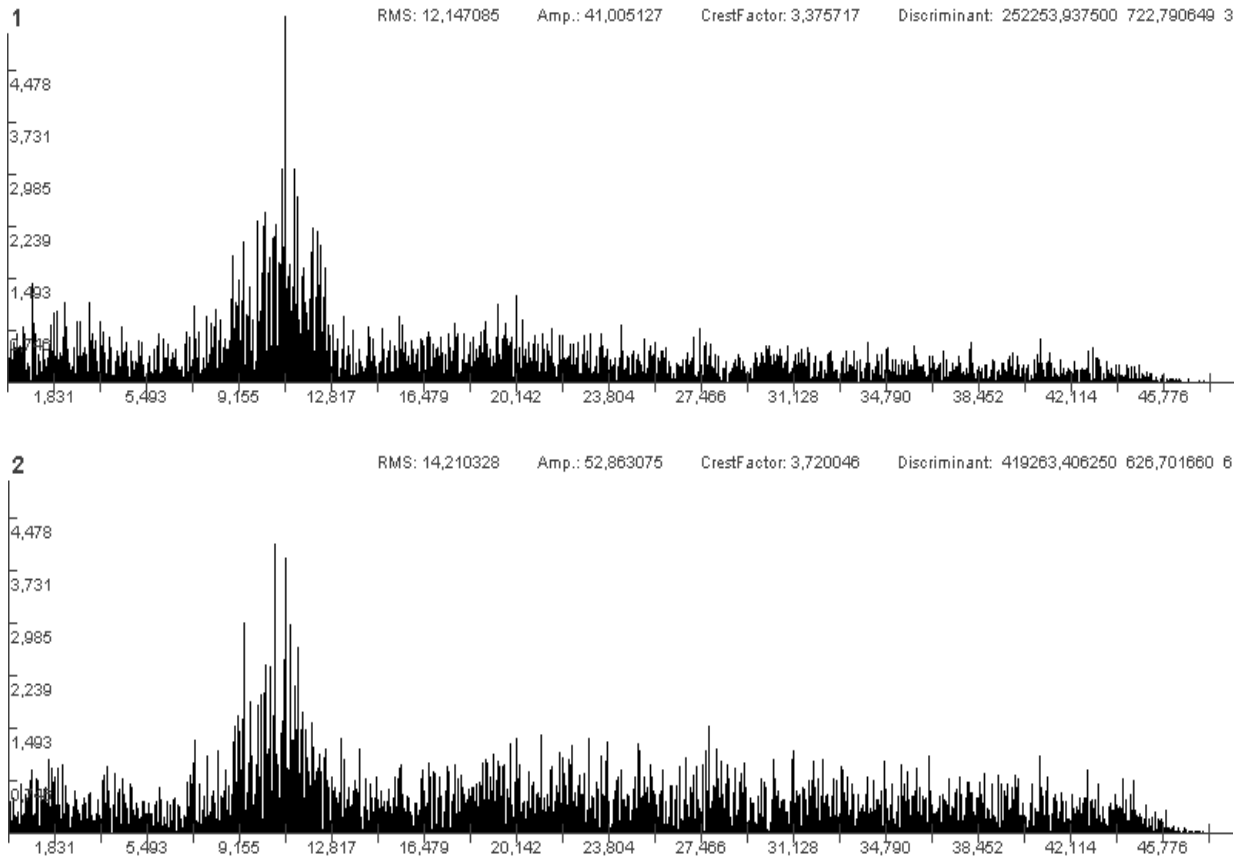


Рисунок 2. Амплитудные спектры сигналов ЭЭГ в точках контроля O1-A1 и O2-A2
T_a=16,384 с; Δf = 0,061 Гц

Для сравнения на рисунке 3 изображен амплитудный спектр сигнала O1-A1 при интервале анализа 32,768 секунды и числе дискретных точек 16384. Для большей наглядности и более удобного восприятия можно использовать поло-совой амплитудный спектр с узкими частотными полосами [20]. Полосовые спектры рассматриваемых сигналов показаны на рисунке 4.

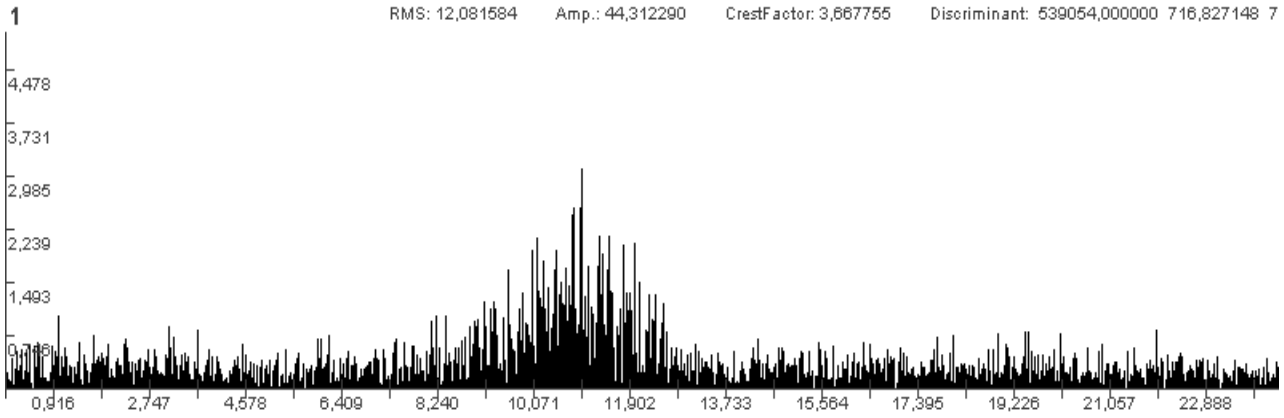


Рисунок 3. Амплитудные спектры сигналов ЭЭГ в точках контроля O1-A1 и O2-A2
T_a=32,768 с; Δf = 0,0305 Гц

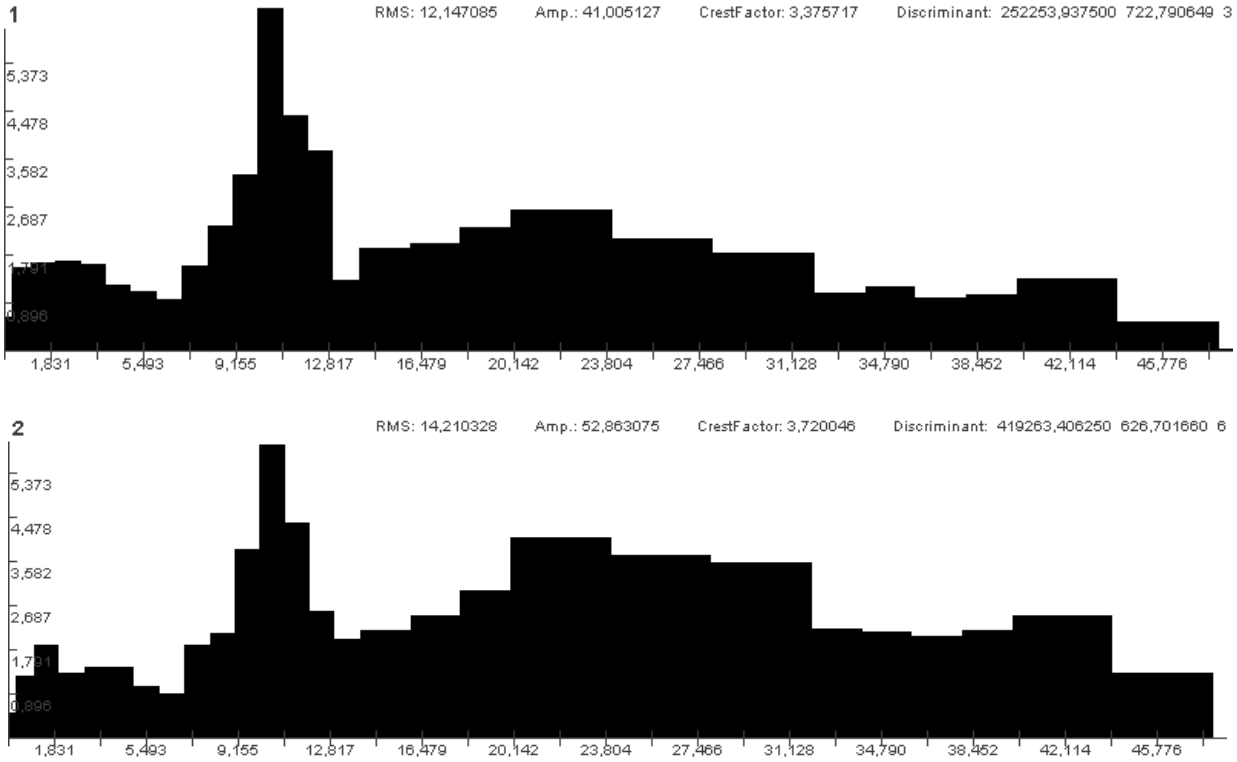


Рисунок 4. Полосовые спектры сигналов ЭЭГ в точках контроля O1-A1 и O2-A2
Полосы 0-0,3 Гц; 0,3-1 Гц; 1-2 Гц; 2-3 Гц; 3-4 Гц; 4-5 Гц; 5-6 Гц; 7-8 Гц; 8-9 Гц; 9-10 Гц; 10-11 Гц; 11-12 Гц; 12-13 Гц; 13-14 Гц; 14-16 Гц; 16-18 Гц; 18-20 Гц; 20-24 Гц; 24-28 Гц; 28-32 Гц; 32-34 Гц; 34-36 Гц; 36-38 Гц; 38-40 Гц; 40-44 Гц; 44-48 Гц; 48-70 Гц

Такое представление позволяет визуально оценить особенности интенсивностей сигналов ЭЭГ в каждой из частотных полос − альфа, бета, гамма, дельта, тэта. Для сравнения на рисунке 5 показаны временные реализации сигналов ЭЭГ, полученных во втором эксперименте, а на рисунке 6 полосовые спектры этих сигналов. Для этих ЭЭГ характерно подавленное значение сигналов в альфа-диапазоне и значительно более интенсивные частотные составляющие бета-диапазона.

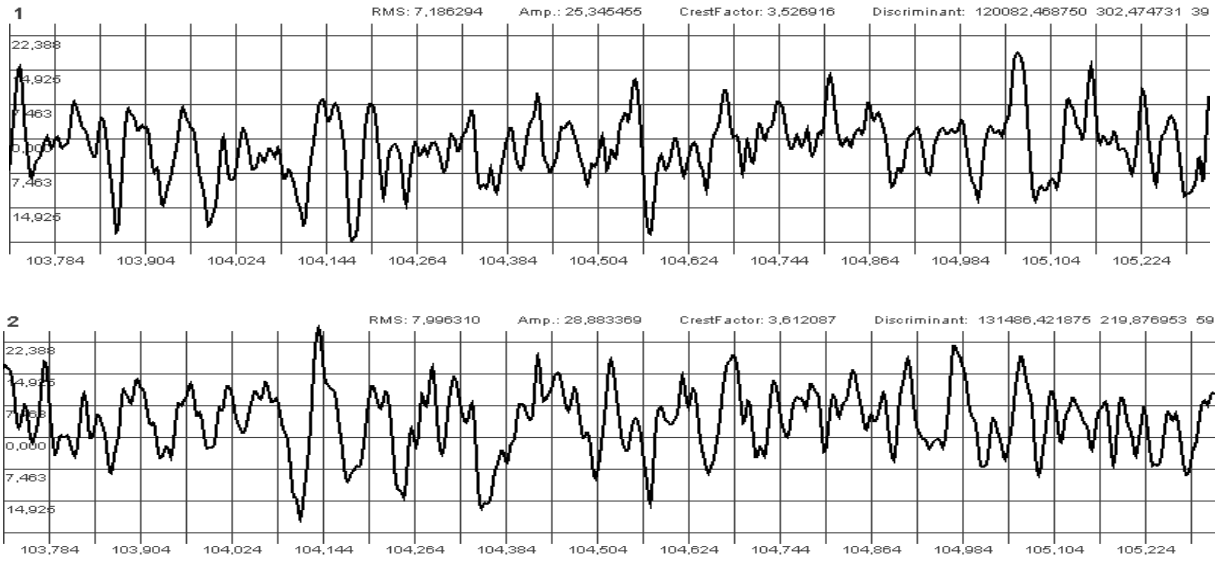


Рисунок 5. Сигналы ЭЭГ (эксперимент 2) в точках контроля O1-A1 и O2-A2

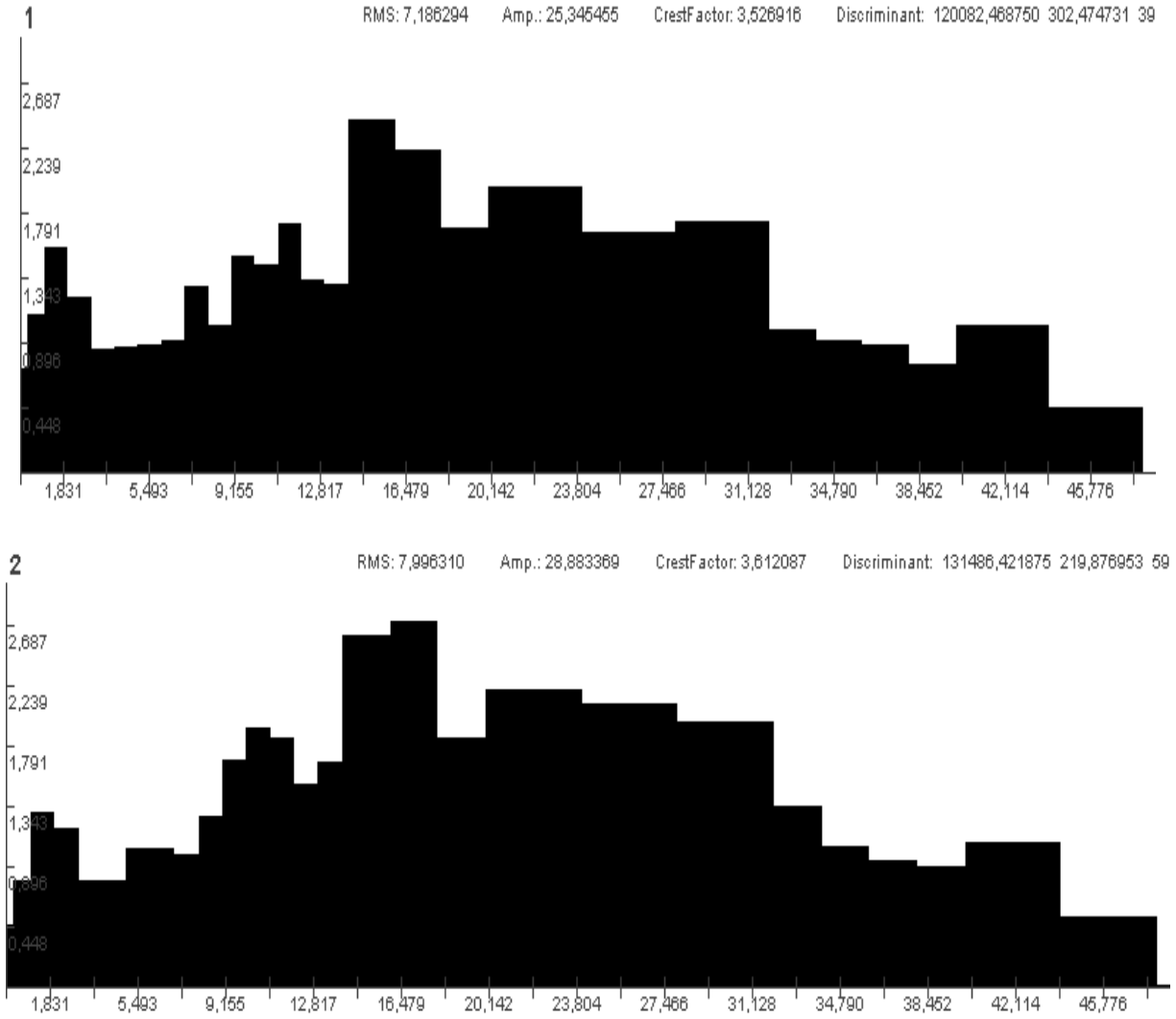


Рисунок 6. Полосовые спектры сигналов ЭЭГ (эксперимент 2) в точках контроля O1-A1 и O2-A2

Полосы 0-0,3 Гц; 0,3-1 Гц; 1-2 Гц; 2-3 Гц; 3-4 Гц; 4-5 Гц; 5-6 Гц; 7-8 Гц; 8-9 Гц; 9-10 Гц; 10-11 Гц; 11-12 Гц; 12-13 Гц; 13-14 Гц; 14-16 Гц; 16-18 Гц; 18-20 Гц; 20-24 Гц; 24-28 Гц; 28-32 Гц; 32-34 Гц; 34-36 Гц; 36-38 Гц; 38-40 Гц; 40-44 Гц; 44-48 Гц; 48-70 Гц

Более расширенное информационное представление о сигнале ЭЭГ можно получить построив временные тренды параметров этого сигнала (рисунок 7).

Они позволяют оценить влияние внешних факторов (например, открытие и закрытие глаз, световое воздействие) на изменения, возникающие в работе нейронной сети головного мозга при их воздействии.

Ряд проведенных экспериментов показал [17−20], что в узких частотных полосах сигнал ЭЭГ представляет модулированную гармонику.

А сам сигнал ЭЭГ можно представить суммой таких модулированных гармоник (рисунки 6, 9) [19−20].

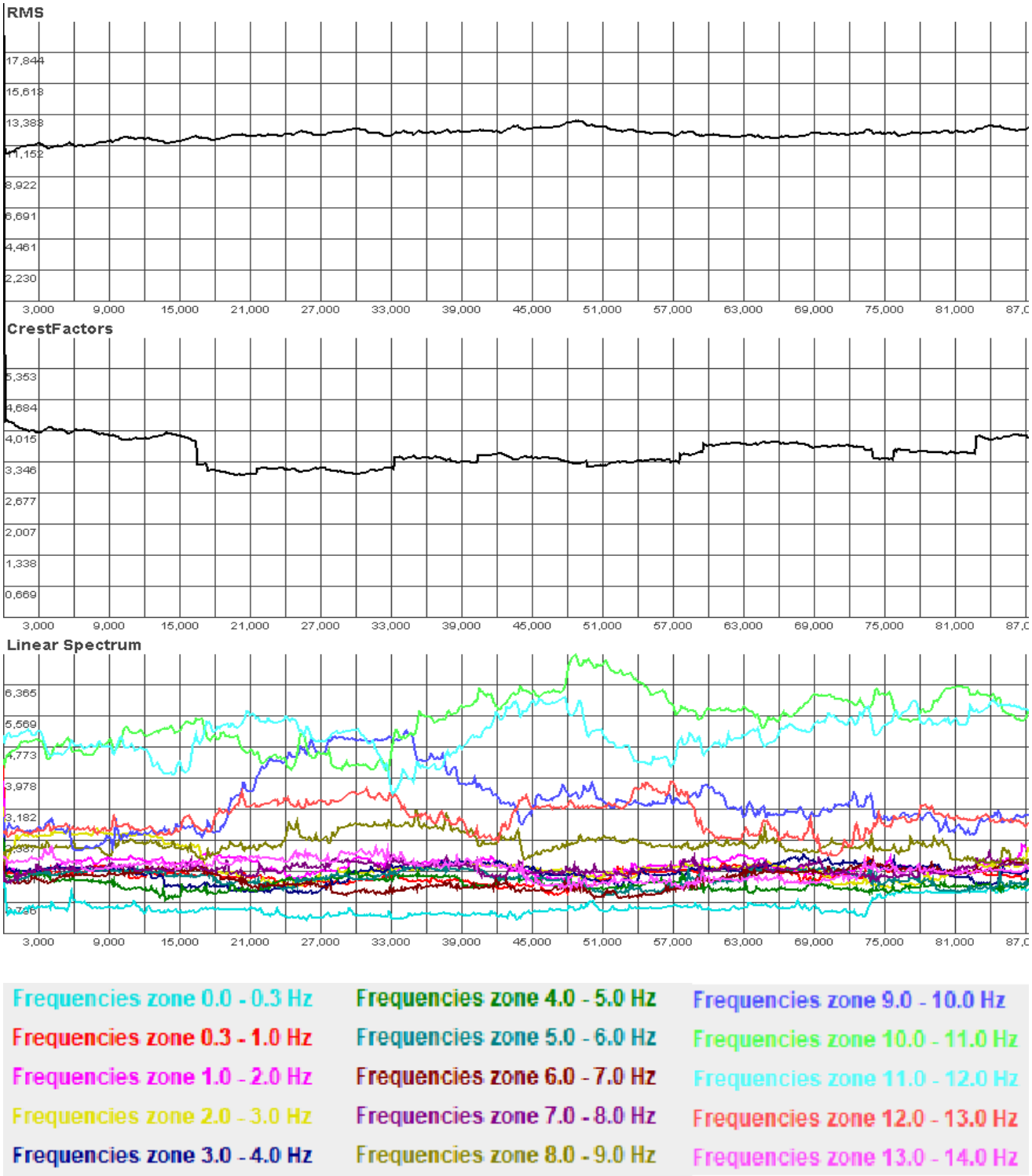


Рисунок 7. Временные тренды среднего квадратического значения (СКЗ) сигнала ЭЭГ O1-A1, пик-фактора этого сигнала, СКЗ этого сигнала в узких частотных полосах

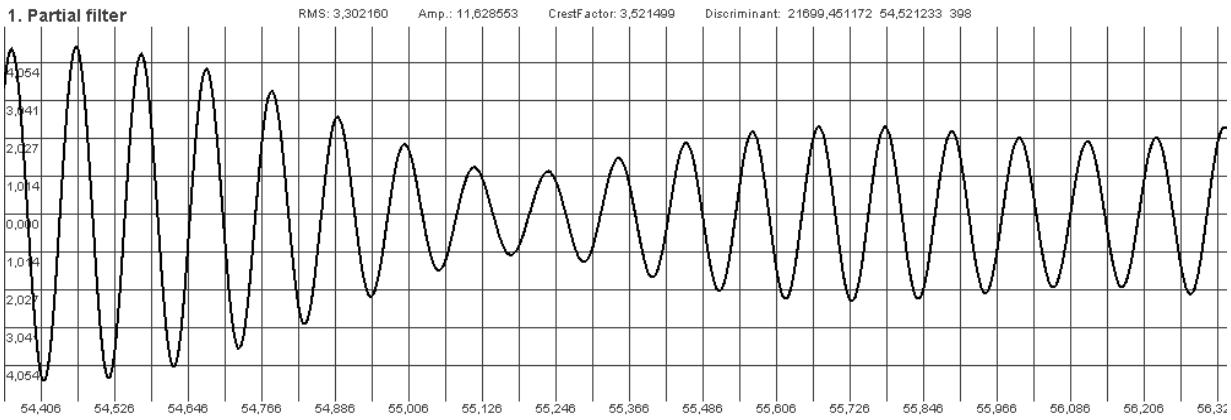


Рисунок 8. Сигнал ЭЭГ в частотной полосе 9-10 Гц для точки контроля O1-A1

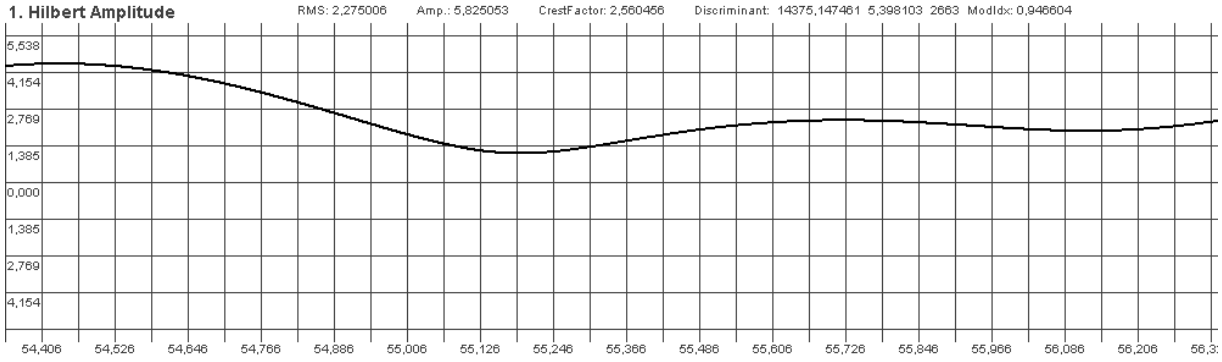


Рисунок 9. Огибающая сигнала ЭЭГ в частотной полосе 9-10 Гц для точки контроля O1-A1

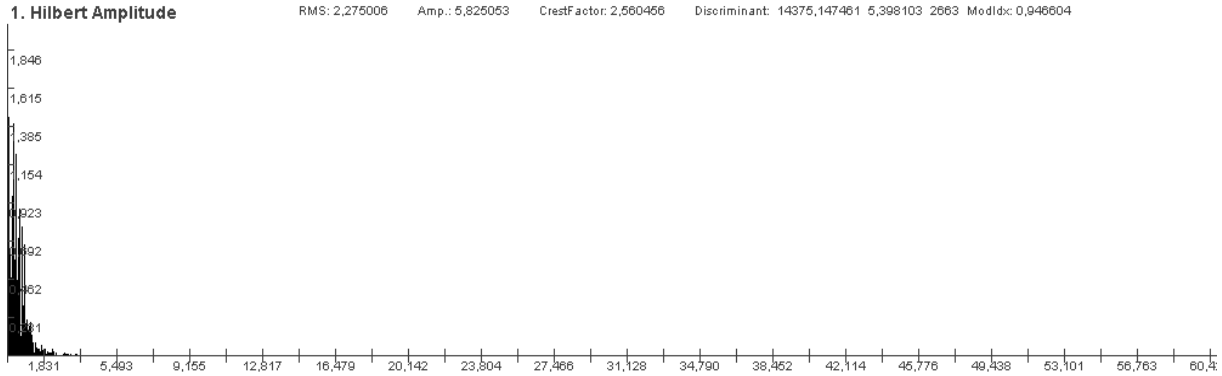


Рисунок 10. Амплитудный спектр огибающей сигнала ЭЭГ в частотной полосе 9-10 Гц для точки контроля O1-A1

Для подобных сигналов свойства модулирующего воздействия оцениваются огибающей (рисунки 9, 12) и её амплитудным спектром (рисунки 10, 13).

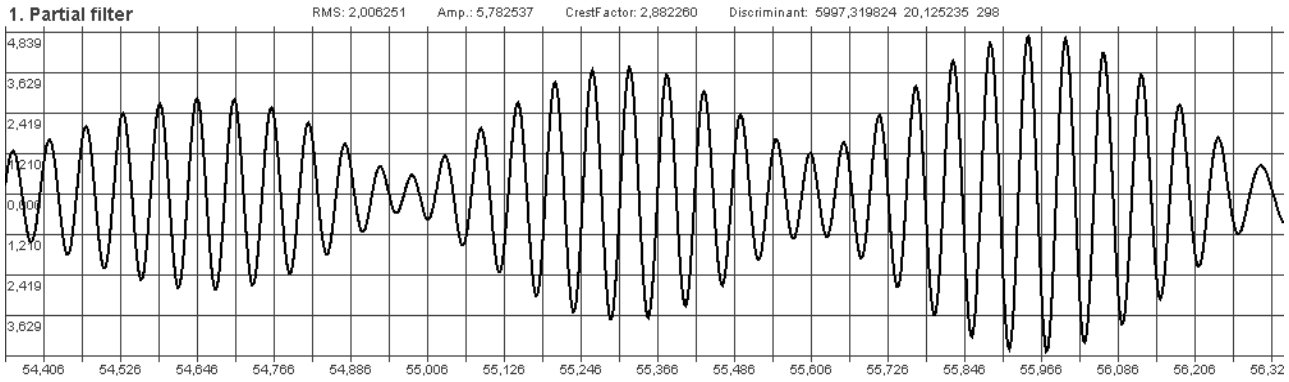


Рисунок 11. Сигнал ЭЭГ в частотной полосе 16-18 Гц для точки контроля O1-A1

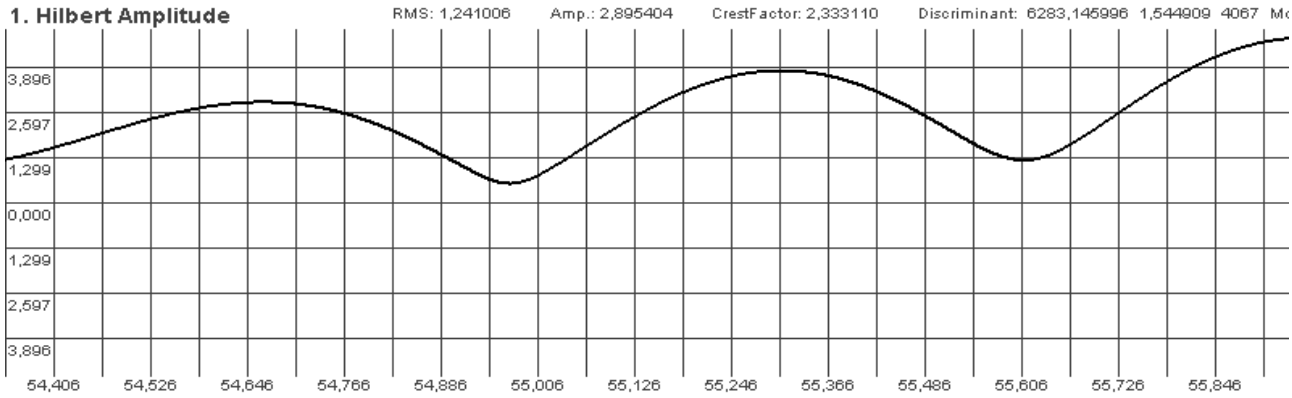


Рисунок 12. Огибающая сигнала ЭЭГ в частотной полосе 16-18 Гц для точки контроля O1-A1

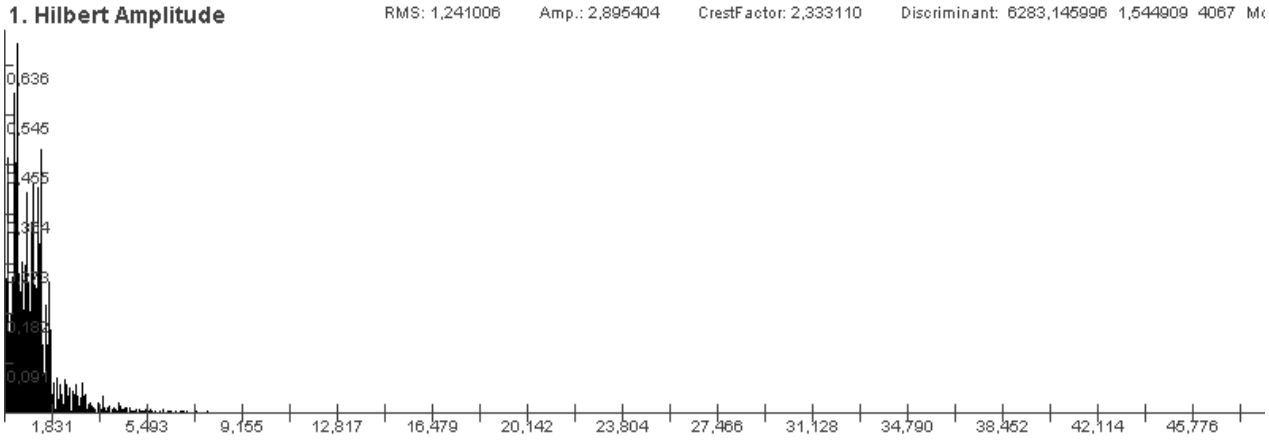


Рисунок 13. Амплитудный спектр огибающей сигнала ЭЭГ в частотной полосе 16-18 Гц для точки контроля O1-A1

Можно предположить, что параметры составляющих ЭЭГ, представляющих собой модулированные гармоники, являются информативно-значимыми параметрами при сравнении ЭЭГ или оценке их изменений.

Заключение. Сигналы электроэнцефалографии являются сложно структурированными, весьма разнообразными и отражают изменение электромагнитного поля в конкретной точке на поверхности головы.

Можно предположить, что эффект этого изменения есть следствие прохождения потенциалов действия по аксонам нейронов [21−23].

Величина влияния конкретного потенциала действия на электромагнитное поле в определенной точке получения сигнала ЭЭГ зависит от его удаленности от этой точки и скорости продвижения по аксону и, следовательно, может увеличиваться и уменьшаться за какой-то временной интервал.

Такое влияющее изменение электрического потенциала можно описать гармоникой определенной частоты с некоторой функцией модуляцией ее амплитуды.

Весь сложный сигнал ЭЭГ при таком подходе декомпозируется в виде суммы модулированных гармоник, возможно не только по амплитуде, но и по фазе.

Экспериментальные исследования показывают, что в узких частотных диапазонах сигнал ЭЭГ действительно представляет собой модулированную гармонику.

Таким образом, исследование параметров и характеристик таких компонент сигналов ЭЭГ ориентировано на получение различных новых данных о работе нейронной сети головного мозга человека, об эффективности применяемых методов лечения и функционального воздействия.

Список литературы

[1] Бранцевич, П. Ю. Автоматизация принятия решений при анализе состояния сложных объектов и ситуаций / П. Ю. Бранцевич // IV Международный форум IT-Академграда «Искусственный интеллект в Беларуси»: доклады, Минск, 13–14 октября 2025 г. – Минск : ОИПИ НАН Беларуси, 2025. – С. 63–73.

[2] Картер, Р. Как работает мозг / Р. Картер. Пер. с англ. П. Петрова. – Москва : Издательство АСТ: CORPUS, 2016. – 224 с.

[3] Блум, Ф. Мозг, разум и поведение / Ф. Блум, А. Лейзерсон, Л. Хофстедтер. Пер. с англ. – М.: Мир,. 1988. –248 с,

[4] Павлова, Л.П. Доминанты деятельного мозга человека. Системный психофизиологический подход к анализу ЭЭГ/ Л. П. Павлова. СПб.: Информ-навигатор, 2017. – 430 с.

[5] Электроэнцефалография: руководство / М.В. Александров, Л.Б. Иванов, С.А. Лытаев [и др.]. Под ред. М.В. Александрова. 3-е изд., перераб. и доп. – Санкт-Петербург: СпецЛит, 2020. − 224 с.

[6] Зенков, Л. Р. линическая электроэнцефалография (с элементами эпилептологии) / Л. Р. Зенков. − 2-е изд., испр. и доп. − М.: МЕДпресс-информ, 2002. − 368 с.

[7] Докукина, Т. В. Визуальная и компьютерная ЭЭГ в клинической практике / Т. В. Докукина, Н. Н. Мисюк. − Минск: Кнігазбор, 2011. − 187 с.

[8] Татум, У.О. Клиническая интерпретация электроэнцефалографии / У. О. Татум, А. М. Хусейн, С. Р. Бенбадис, П. В. Каплан. Пер. с англ. − М.: Издательский дом БИНОМ. 2020. − 264 с

[9] Лайонс, Р. Цифровая обработка сигналов / Р. Лайонс. – М.: ООО «Бином-Пресс», 2006. – 656 с.

[10] Кулаичев, А. П. Компьютерная электрофизиология и функциональная диагностика. Москва: ИНФРА-М; 2018. – 469 с.

[11] Бранцевич, П. Ю. Цифровая обработка вибрационных сигналов / П. Ю. Бранцевич. – Минск: Бестпринт, 2022. – 297 с.

[12] Бранцевич, П. Ю. Компьютерные системы и комплексы обработки вибрационных сигналов / П. Ю. Бранцевич. – Минск: Бестпринт, 2023. – 282 с.

[13] Бранцевич, П. Ю. Цифровая обработка сигналов для встроенных систем : учеб.-метод. пособие / П. Ю. Бранцевич, Д. В. Деменковец, В. А. Леванцевич. – Минск : БГУИР, 2025. – 283 с.

[14] Бранцевич, П.Ю. Цифровая обработка электроэнцефалограмм / П.Ю. Бранцевич // Science and innovation. – Ташкент: LLC «Science and innovation», 2023. – Special Issue 3. – с. 930–934.

[15] Бранцевич П.Ю. Сравнительный анализ электроэнцефалограмм. BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics: сб. науч. ст. IX Междунар. науч.-практ. конф. В 2 ч. Ч. 1 (Республика Беларусь, Минск, 17–18 мая 2023 года). Редкол.: В. А. Богуш [и др.]. – Минск: БГУИР, 2023. – с. 132–144.

[16] Яковлева, Т. В. Анализ временных рядов ЭЭГ пациентов с алкогольной зависимостью с помощью нейронных сетей / Т. В. Яковлева, А. В. Крысько // Российский журнал биомеханики. – 2024. – Т. 28, № 1. – С. 125-142.

[17] Бранцевич, П. Ю. Цифровая обработка сигналов и сравнительный анализ электроэнцефалограмм / П. Ю. Бранцевич // BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics : сб. науч. ст. X Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 13 марта 2024 года). В 2 ч. Ч. 1 / редкол. : В. А. Богуш [и др.]. – Минск: БГУИР, 2024. – С. 101–112.

[18] Бранцевич, П. Ю. Информативные признаки и характеристики электроэнцефалограмм / П. Ю. Бранцевич, Т. В. Докукина, Н. Н. Мисюк, П. П. Королевич // Медэлектроника – 2024. Средства медицинской электроники и новые медицинские технологии : сб. науч. ст. XIV Междунар. науч.- техн. конф. (Республика Беларусь, Минск, 5-6 декабря 2024 года). – Минск : БГУИР, 2024. – С. 63–69.

[19] Бранцевич, П. Ю. Декомпозиция цифровых сигналов электроэнцефалограмм при их сравнительном анализе / П. Ю. Бранцевич // BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics : сб. науч. ст. XI Междунар. науч.-практ. конф. (Республика Беларусь, Минск, 23-24 апреля 2025 года) / редкол. : В. А. Богуш [и др.]. – Минск: БГУИР, 2025. – С. 85–91.

[20] Бранцевич, П. Ю. Определение информативно-значимых параметров электроэнцефалограмм при цифровой обработке методом разложения на модулированные гармоники / П. Ю. Бранцевич // Неразрушающий контроль и диагностика. − 2025. − № 2. – С. 12–22.

[21] Смирнов, В.М. Физиология сенсорных систем и высшая нервная деятельность / В.М. Смирнов, С.М. Будылина. − М. : Медицина, 2003. − 304 с.

[22] Гайтон, А.К. Медицинская физиология / А.К. Гайтон, Дж.Э. Холл / Пер. с англ.; Под ред. В.И. Кобрина. − М. : Логосфера, 2008. − 1296 с.

[23] Николлс, Д. От нейрона к мозгу / Д. Николлс, Р. Мартин, Б. Валлас, П. Фукс. Пер. с англ. − М. : Едиториал УРСС, 2003. − 672 с.

Авторский вклад

Бранцевич Петр Юльянович – цифровая обработка сигналов ЭЭГ, формирование статьи.
Григорьева Инесса Викторовна – постановка задачи исследования.
Королевич Павел Павлович – регистрация и первичная обработка сигналов ЭЭГ.
Базылев Евгений Николаевич – разработка программного средства для исследований ЭЭГ.

METHODS FOR PROCESSING DIGITAL ELECTROENCEPHALOGRAM SIGNALS IN THE STUDY OF THEIR INFORMATIVELY SIGNIFICANT PARAMETERS

P.J. Brancevich
Professor of the Department of Software and Information Technology at BSUIR, Doctor of Science in Technical Sciences

I.V. Grigorieva
Scientific Secretary of the Republican Scientific and Practical Center for Mental Health, Candidate of Medical Sciences

P.P. Korolevich
Psychiatrist-Narcologist, Functional Diagnostics Doctor, Researcher of Department of Mental and Behavioral Disorders of Republican Scientific and Practical Center for Mental Health

E.N. Bazyleu
Senior Lecturer, Department of Software and Information Technology, BSUIR, Research

Abstract. Digital electroencephalogram (EEG) signals reflect the specific state of the human brain's neural network. Methods for processing EEG signals and determining their information-relevant parameters and characteristics are discussed. The possibility of comparing EEG signals using bandpass spectra is demonstrated. A method for processing EEG signals in narrow frequency bands, which are modulated harmonics, is proposed.

Keywords: signal, electroencephalogram, parameter, characteristic, digital processing

УДК 681.325

ПРЕОБРАЗОВАНИЕ ИЕРАРХИЧЕСКИХ ТРАНЗИСТОРНЫХ ОПИСАНИЙ КОМБИНАЦИОННЫХ ЛОГИЧЕСКИХ СХЕМ В СИСТЕМЕ ЛОГИЧЕСКОГО ПРОЕКТИРОВАНИЯ



Н.А. Кириенко
Доцент кафедры
экономической информатики БГУИР,
кандидат технических наук, доцент
kirienkonaminsk@mail.ru

Н.А. Кириенко

Окончила Минский радиотехнический институт. Область научных интересов связана с автоматизацией процессов логического проектирования дискретных устройств, преобразованием и оптимизацией функциональных описаний логических схем.

Аннотация. Преобразование представлений исходных описаний проектируемых цифровых устройств имеет большое значение в процессе оптимизации и синтеза. Рассматриваются виды процедур оптимизации и преобразований в экспериментальной системе логического проектирования. Представлены метод и алгоритм преобразования многоуровневых транзисторных описаний логических схем в одноуровневые. Сделан вывод о применении преобразования в процессе перепроектирования схем.

Ключевые слова: логическая схема, функциональное описание комбинационной схемы, иерархическое транзисторное описание, процедура устранения иерархии, VHDL, КМОП элемент, система логического проектирования.

Введение. Синтез комбинационных логических схем в заданной библиотеке (базисе) проектирования всегда предваряется оптимизацией представлений тех систем булевых функций, которые требуется реализовать в схеме. В процессе оптимизации функциональных описаний будущей схемы выполняются процедуры минимизации, факторизации, декомпозиции, разбиения на подсистемы систем булевых функций, и многие другие преобразования. Набор процедур оптимизации присутствует в экспериментальной системе логического проектирования FLC-2 [1], разработанной и эксплуатирующейся в ОИПИ НАН РБ. Главное окно системы можно видеть на рисунке 1.

В качестве средств описания функционирования логического устройства в системе FLC-2 используются специализированные языки VHDL [2], SF [1]. Если VHDL является популярным языком систем проектирования, то язык SF (структурно-функционального описания) был разработан в лаборатории логического проектирования ОИПИ НАН РБ для поддержки ряда систем логического проектирования, в том числе FLC-2. Система предлагает различные процедуры преобразования исходных функциональных описаний с целью их оптимизации, моделирования, верификации, синтеза топологии, конвертации VHDL – SF и SF – VHDL.

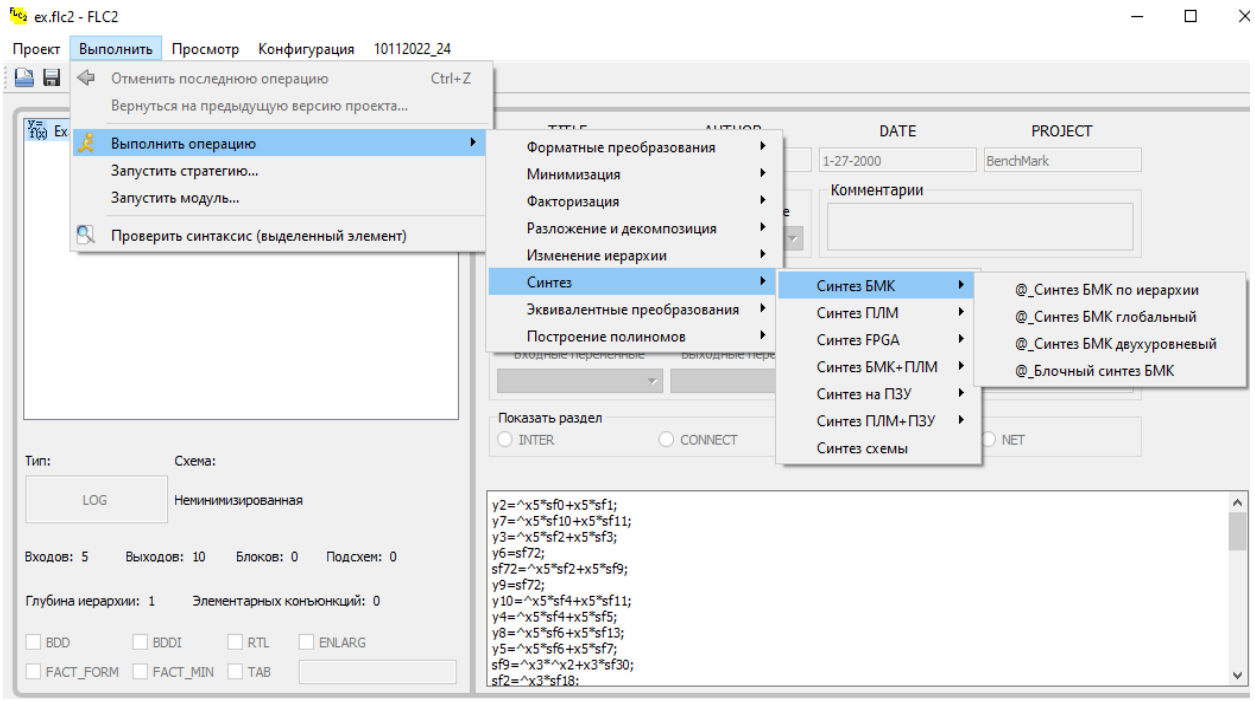


Рисунок 1. Главное окно системы FLC-2

В настоящей работе рассматривается постановка задачи и алгоритм устранения иерархии транзисторных описаний - преобразования многоуровневых описаний логических схем в одноуровневые транзисторные описания.

Постановка задачи. Среди процедур системы логического проектирования особое место занимает группа процедур изменения иерархии функциональных описаний, представленных на рисунке 2.

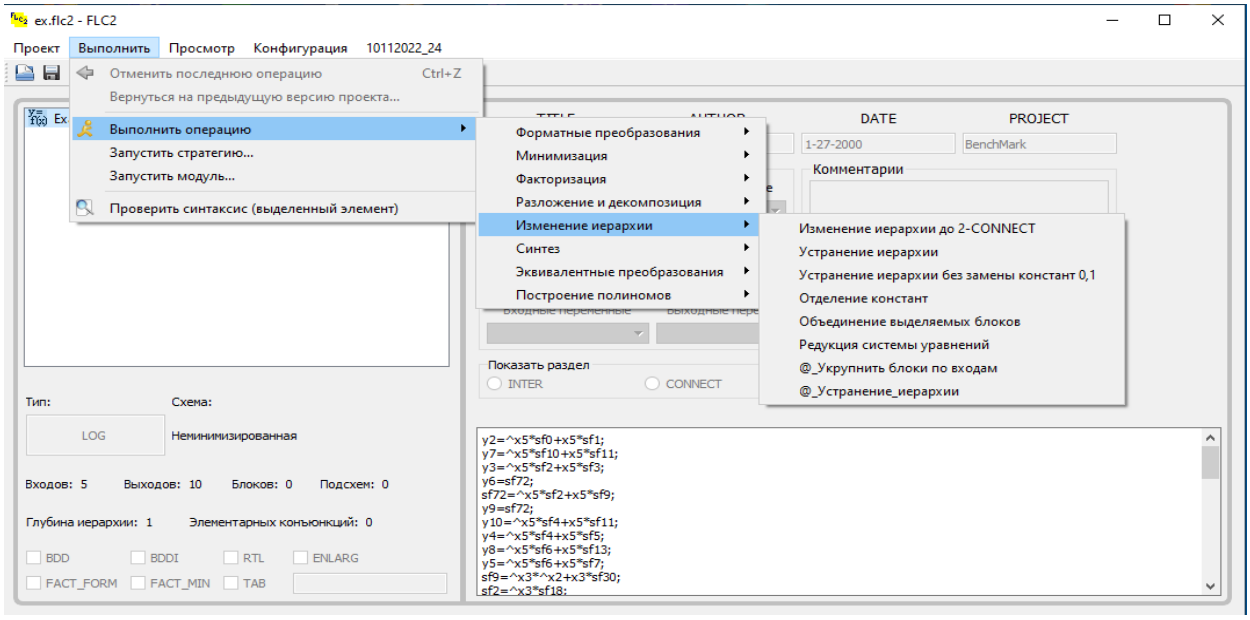


Рисунок 2. Набор процедур изменения иерархии описаний системы FLC-2

Различные процедуры системы логического проектирования требуют соответствующих форм представления функциональных описаний схем. SF-описание состоит из блоков и

обладает иерархической структурой, т. е. каждый блок описания может быть самостоятельным иерархическим описанием. Листьевым SF-описанием может быть система логических уравнений, пара матриц (булева и троичная), или транзисторное описание. Рассмотрим далее метод и алгоритм преобразования многоуровневого транзисторного описания в одноуровневое.

Представление функциональных описаний логических схем. Исходные описания, поступающие на вход САПР СБИС, имеют, как правило, иерархическую структуру. Проектировщик представляет задание на проектирование в виде схемы взаимосвязанных логических блоков, каждый из которых также может быть представлен совокупностью взаимосвязанных логических блоков, образуя многоуровневое структурное описание. В описании обязательно присутствуют листьевые блоки в виде стандартных элементов заданной библиотеки проектирования, логика функционирования которых может описываться либо системой логических уравнений, либо системой дизъюнктивных нормальных форм в матричном виде, либо структурной схемой из транзисторов. В последнем случае мы имеем дело с транзисторными описаниями объектов проектирования.

Многоуровневое транзисторное описание логической схемы может быть получено в результате преобразования структурного описания устройства, реализованного в базисе заданной библиотеки проектирования, в SPICE-формат (формат описания устройства для системы моделирования SPICE [3]). Структурное описание может быть результатом синтеза устройства по заданному исходному описанию поведения в одной из промышленных систем проектирования, например, LeonardoSpectrum фирмы Mentor Graphics [1]. Рассмотрим небольшой пример комбинационной схемы CIRCUIT1, функционирование которой задано уравнением (1).

$$f1=(x3 \wedge x1) \vee x2 \vee x3 \qquad (1)$$

В листинге 1 представлено описание поведения логической схемы CIRCUIT1 на языке VHDL.

Листинг 1. Исходное описание поведения логической схемы CIRCUIT1 на языке VHDL.

```
library ieee;
use ieee.std_logic_1164.all;
entity CIRCUIT1 is
port(x1, x2, x3 : in std_logic;
f1:out std_logic
    );
end CIRCUIT1;
architecture CIRCUIT1_arch of CIRCUIT1 is
begin
f1<=((NOT x3 AND NOT x1) OR x2 OR x3);
end CIRCUIT1_arch;
```

Описание синтезированной с помощью системы LeonardoSpectrum схемы представлено в листинге 2. Схема синтезирована в базисе КМОП-элементов библиотеки проектирования s3lib [4]. В описании представлены определения входных (x1, x2, x3), выходных (f1) переменных, основных элементов схемы (ix5, ix66, ix68). Для каждого элемента определен его тип, входные и выходные сигналы. Типы элементов библиотеки s3lib представлены в таблице 1.

Таблица 1. Логические функции КМОП-элементов библиотеки s3lib.

N	y=A
NA3	y=ABC

Листинг 2. Синтезированная схема CIRCUIT1 в базисе КМОП элементов.

```
library IEEE;
use IEEE.STD_LOGIC_1164.all;
```

```
entity CIRCUIT1 is
   port (
      x1 : IN std_logic ;
      x2 : IN std_logic ;
      x3 : IN std_logic ;
      f1 : OUT std_logic) ;
end CIRCUIT1 ;
architecture CIRCUIT1_arch of CIRCUIT1is
   signal nx65, nx67: std_logic ;
begin
   ix5 : NA3 port map ( Y=>f1, A=>x1, B=>nx65, C=>nx67);
   ix66 : N port map ( Y=>nx65, A=>x2);
   ix68 : N port map ( Y=>nx67, A=>x3);
end CIRCUIT1_arch ;
```

Файл описания схемы CIRCUIT1 в формате SPICE для моделирования в системе схемотехнического моделирования AccuSimII (фирмы Mentor Graphics) [2] представлен в листинге 3. Описание каждого элемента схемы преобразовано в соответствии с требованиями AccuSimII. SPICE-формат двухуровневого описания схемы CIRCUIT1 удобен в дальнейшем для явного выделения подсхем.

Листинг 3. SPICE-формат двухуровневого описания схемы CIRCUIT1.

```
.include s3lib.cir
.include models_035.mod
RVCC VCC VCC3 1m
.subckt CIRCUIT1 f1 x1 x2 x3
X_ix5 f1 x1 nx65 nx67  NA3
X_ix66 nx65 x2  N
X_ix68 nx67 x3  N
.ends CIRCUIT1
X_CIRCUIT1 f1 x1 x2 x3 CIRCUIT1
.OPTIONS TNOM=27
.TEMP 27
.param VHI=3.3 VLO=0 TDELAY=0 TRISE=1n TFALL=1n TSAMPLE=100n
Vx1 x1 0 pattern VHI VLO TDELAY TRISE TFALL TSAMPLE 1001100101 R
Vx2 x2 0 pattern VHI VLO TDELAY TRISE TFALL TSAMPLE 0011110001 R
Vx3 x3 0 pattern VHI VLO TDELAY TRISE TFALL TSAMPLE 1011011000 R
.TRAN 1E-09 {TSAMPLE*10} 0
```

Алгоритм преобразования многоуровневого транзисторного описания в одноуровневое. Описание схемы в формате SPICE содержит три элемента библиотеки проектирования двух типов: NA3, N. Задача устранения иерархии (или компиляции) транзисторных описаний состоит в том, чтобы вместо каждой строки описания элемента встроить строки определения этого элемента на транзисторном уровне. Для выполнения этой задачи необходимо иметь описание библиотеки элементов на транзисторном уровне. Пример описания элемента типа N на транзисторном уровне представлен в листинге 4.

Листинг 4. Пример описания элемента типа N на транзисторном уровне

```
* Component: N
subckt N Y A
M2 Y A VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11 ps=1.36e-
05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
M1 Y A 0 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06
pd=9.6e-06 nrs=0.92 nrd=0.92
.ends N
```

Описание каждой подсхемы состоит из раздела внешних имен, тела описания и раздела конца описания. Раздел внешних имен имеет следующий вид:

```
.SUBCKT  <имя подсхемы>  <имена внешних полюсов>.
```

Имена внешних полюсов схемы есть перечень формальных параметров, которые при компиляции должны быть заменены фактическими из строки обращения к подсхеме. Тело описания может содержать описания как подсхем (вложенные подсхемы), так и транзисторов. Строка описания вложенной подсхемы имеет следующий вид (смотри листинг 3):

```
<имя подсхемы> <имена полюсов> <имя типа подсхемы>.
```

Имя подсхемы состоит из схемного имени подсхемы, перед которым записывается префикс «X_». В разделе <имена полюсов> должны быть перечислены имена реальных (фактических) переменных, соответствующих формальным переменным подсхемы. Фактические переменные должны быть указаны в том порядке, который был использован в списке имен внешних полюсов (формальных переменных) в заголовке описания соответствующего типа подсхемы (разделе .SUBCKT). Строка описания транзистора имеет следующий вид:

```
<имя транзистора><номера узлов><физические характеристики>
```

Имя транзистора состоит из заглавной латинской буквы M и номера. Далее следуют четыре идентификатора, означающих имена узлов стока, затвора, истока, подложки транзистора, соответственно. Далее следуют тип модели транзистора (MODP или MODN) и его физические характеристики. Раздел конца описания:

```
.ENDS <имя типа подсхемы>
```

Задача компиляции решается путем поэтапного раскрытия подсхем в исходном описании и замене листьевых подсхем их транзисторными описаниями. В описании более высокого уровня подсхема описывается обращением к ней - заголовком, в котором выписаны фактические параметры (имена переменных, подаваемых на входы и снимаемых с выходов схемы). В заголовке описания подсхемы перечислены формальные параметры. Соответствие имен формальных и фактических параметров задается порядком их перечисления. В процессе компиляции схемы обращение к схеме должно быть заменено описанием ее внутренней структуры, причем формальные параметры должны быть заменены на фактические во всем описании. Имена транзисторов и внутренние имена переменных подсхемы должны быть подвергнуты переименованию, так как в противном случае, при раскрытии одинаковых типов схем получится многоразовое использование одинаковых имен.

В качестве новых имен транзисторов используется тот же символ “M” со сквозной последовательной нумерацией. Внутренние имена в транзисторных описаниях элементов заменяются на числовой порядковый номер. Имена 0 и VCC3 резервируются за узлами земли и питания. Решение задачи компиляции заключается в поуровневом просмотре многоуровневого описания. По мере раскрытия очередной подсхемы, описания ее транзисторов с замененными именами узлов переписываются в результирующее описание, а о каждой подсхеме заполняются таблицы нераскрытых на данном уровне подсхем. Тип подсхемы запоминается в таблице имен типов, а список фактических параметров - таблице фактических параметров.

При выполнении компиляции на i-м уровне, составляются таблицы нераскрытых подсхем для (i+1)-го уровня. Если таблицы для (i+1)-го уровня пусты, то компиляция схемы закончена. Если таблицы для (i+1) уровня не пусты, то происходит обработка подсхем из этих таблиц. Пример скомпилированной подсхемы CIRCUIT1 представлен в листинге 5.

Листинг 5. Пример скомпилированной подсхемы CIRCUIT1.

```
.subckt CIRCUIT1 f1 x1 x2 x3
M12 N4 x1 0 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06 pd=9.6e-06 nrs=0.92 nrd=0.92
M11 N3 nx65 N4 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06 pd=9.6e-06 nrs=0.92 nrd=0.92
M8 f1 nx67 N3 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06 pd=9.6e-06 nrs=0.92 nrd=0.92
```

```
    M9 f1 nx67 VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11
ps=1.36e-05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
    M2 f1 nx65 VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11
ps=1.36e-05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
    M1 f1 x1 VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11 ps=1.36e-
05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
    M13 nx65 x2 VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11
ps=1.36e-05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
    M14 nx65 x2 0 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06
pd=9.6e-06 nrs=0.92 nrd=0.92
    M15 nx67 x3 VCC3 VCC3 MODP L=3.5e-07 W=4.5e-06 as=1.04e-11 ad=1.04e-11
ps=1.36e-05 pd=1.36e-05 nrs=0.511111 nrd=0.511111
    M16 nx67 x3 0 0 MODN L=3.5e-07 W=2.5e-06 as=5.75e-12 ad=5.75e-12 ps=9.6e-06
pd=9.6e-06 nrs=0.92 nrd=0.92
    .ends CIRCUIT1
```

Заключение. Рассмотренный алгоритм положен в основу процедуры преобразования иерархических транзисторных описаний комбинационных логических схем, функционирующей в экспериментальной системе логического проектирования FLC-2.

Реализованная процедура может быть полезной для выполнения перепроектирования схемы в базисе других КМОП элементов. Также процедура полезна для тестирования результатов проектирования, которое позволяет установить, соответствует ли описание проекта на уровне транзисторов спецификации проектируемого устройства.

Список литературы

[1] Бибило П.Н., Романов В.И. Система логической оптимизации функционально-структурных описаний цифровых устройств на основе продукционно-фреймовой модели представления знаний // Проблемы разработки перспективных микро- и наноэлектронных систем. – 2020. Сб. трудов / под общ. ред. акад. РАН А.Л. Стемпковского. М.: ИППМ РАН. 2020. N 4. С. 9–16.

[2] Бибило, П. Н. Системы проектирования интегральных схем на основе языка VHDL. StateCAD, ModelSim, LeonardoSpectrum / П. Н. Бибило. – М.: СОЛОН-Пресс, 2005. – 384 с.

[3] SPICE Simulation Models – Mode of access: https://www.ni.com/en/shop/electronic-test-instrumentation/application-software-for-electronic-test-and-instrumentation-category/what-is-multisim/spice-simulation-fundamentals/spice-simulation-models.html– Date of access: 04.01.2026.

[4] Бибило, П.Н. Система логического проектирования функциональных блоков заказных КМОП СБИС с пониженным энергопотреблением/ П.Н. Бибило, Н.А. Авдеев, С.Н. Кардаш, Н.А. Кириенко, Ю.Ю. Ланкевич, И.П. Логинова, Д.И. Черемисинов, Л.Д. Черемисинова / Микроэлектроника, 2018г., №1 – С. 72–88.

TRANSFORMATION OF HIERARCHICAL TRANSISTOR DESCRIPTIONS OF COMBINATIONAL LOGIC CIRCUITS IN THE SYSTEM FOR LOGIC DESIGN

N.A. Kirienko
Associate Professor,
Department of economic informatics
PhD of Technical sciences, Associate Professor

Abstract. Transformation of representations of initial descriptions of designed digital devices is of great importance in the process of optimization and synthesis. The types of optimization and transformations procedures in the experimental system of logical design are considered. A method and algorithm for transforming multi-level transistor descriptions of logic circuits into single-level ones are presented. A conclusion is made about the application of the transformation in the process of circuit redesign.

Key words: logic circuit, functional description of the combinational circuit, hierarchical transistor description, hierarchy elimination procedure, VHDL, CMOS element, system of logical design.

УДК 339.138:004.738.5

К ВОПРОСУ О СИСТЕМНОЙ ОЦЕНКЕ ЭФФЕКТИВНОСТИ ДЕЯТЕЛЬНОСТИ ОРГАНИЗАЦИЙ В СОЦИАЛЬНЫХ СЕТЯХ



М.П. Батура
Заведующий НИЛ 6.3 БГУИР, доктор технических наук, профессор
bmpbel@bsuir.by



И.В. Марахина
Доцент кафедры экономики БГУИР, кандидат экономических наук, доцент
inamarahina@gmail.com



В.А. Пархименко
Заведующий кафедрой экономики БГУИР, кандидат экономических наук, доцент
uladzimir.parkhimenka@gmail.com

М.П. Батура

Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: системный анализ, управление и обработка информации в технических и организационных системах.

И.В. Марахина

Окончила Белорусский государственный университет информатики и радиоэлектроники. Сферы интересов: инновационное развитие, теория организации, маркетинг, поисковое продвижение и веб-аналитика.

В.А. Пархименко

Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: маркетинг, экономическая теория, межотраслевой анализ, цифровая трансформация.

Аннотация. Интенсификация использования социальных сетей в качестве канала маркетинговых коммуникаций и соответствующий рост затрат на SMM актуализируют проблему оценки экономической отдачи от данных вложений. В статье представлен критический анализ существующих подходов к измерению результативности деятельности организаций в социальных сетях

Ключевые слова: социальная сеть, SMM, оценка эффективности, метрики эффективности, коммуникационная эффективность, коммерческая эффективность

Введение. В последнее десятилетие значимость соцсетей значительно выросла, и они превратились в ключевой канал маркетинговых коммуникаций.

Сегодня присутствие в соцсетях – это обязательный компонент стратегии развития для большинства организаций.

Компании вкладывают значительные средства в создание контента, таргетированную рекламу и развитие сообществ. Однако, несмотря на понимание необходимости присутствия в соцсетях, перед менеджментом остро встает вопрос окупаемости затрат.

Ключевые аспекты проблемы оценки эффективности работы организаций в социальных сетях.

Следует учитывать два взаимосвязанных направления: оценку коммуникационной (качество взаимодействия с аудиторией) и коммерческой (влияние на продажи) эффективности.

1. Коммуникационная эффективность касается результативности взаимодействия организации с целевой аудиторией в социальных сетях, отражающей степень достижения

качественных (нефинансовых) целей коммуникации. Количественно она может быть выражена через отношение достигнутых коммуникационных эффектов (роста лояльности, уровня узнаваемости бренда, позитивного имиджа, вовлеченности аудитории) к затратам на реализацию SMM.

2. Коммерческая эффективность характеризует вклад SMM в достижение финансовых и бизнес-целей организации.

3. Данный показатель отражает отношение экономических результатов, полученных благодаря присутствию и активности в социальных сетях (выручка, количество целевых действий – покупок, регистраций, лидов, звонков, прирост доли рынка), к объему инвестированных в SMM ресурсов (затрат).

4. Идеальная работа в соцсетях подразумевает, что высокая коммуникационная эффективность является драйвером для коммерческой. На практике эта связь не всегда линейна и очевидна. Проблема заключается в том, что высокие коммуникационные показатели (лайки, репосты, комментарии и т.д.) не всегда конвертируются в коммерческий результат, а попытки напрямую «привязать» продажи к каждому посту разрушают доверие аудитории.

Как показывают обзоры литературы, количество комплексных исследований, рассматривающих проблему оценки эффективности SMM в единстве коммуникационных и коммерческих аспектов, остается относительно невелико. В тоже время можно выделить следующие общие направления научных и практических исследованиях по этому вопросу.

1. Переход **от отдельных метрик в сторону комплексных моделей**, увязанных с бизнес-целями. В рамках исследования такие подходы были систематизированы и представлены в форме таблицы 1.

2. Смещением фокуса от простого количественного анализа **к качественному, контекстному измерению, интегрированному с бизнес-показателями**.

При этом, с одной стороны, возрастает роль алгоритмических и платных инструментов: оценка неразрывно связана с метриками эффективности рекламы (CPC, CPM, CTR, ROAS и т.д.) и анализом каналов в рамках моделей многоканальной атрибуции, таких как PESO (Paid, Earned, Shared, Owned) [7]. С другой стороны, происходит критическое переосмысление самих принципов измерения, в основе которого лежит отказ от «метрик тщеславия» [1].

Происходит переход от показателей активности к показателям результативности, требующем не только критического осмысления данных, но и их интеграции в единую систему, позволяющую оценивать все затраты с точки зрения вклада в конечный бизнес-результат.

Таблица 1. Подходы, используемые для оценки эффективности деятельности организаций в социальных сетях

Подход	**Сущностные особенности**	**Ключевые метрики / принципы**	**Достоинства**	**Недостатки**
1	2	3	4	5
Подход на основе воронки / иерархии целей или эффектов	Связывает активность с этапами пути клиента (воронки): от осведомленности до действия. Базируется на моделях AIDA, иерархии эффектов.	Определяются стадией воронки, целью и могут включать такие метрики, как охват, вовлеченность, трафик, лиды, конверсии, продажи.	Интуитивно понятен, позволяет выстраивать логику от коммуникации к коммерции.	Не учитывает множественность касаний и сложные пути клиента; метрики верхних уровней могут быть «метриками тщеславия».

Продолжение таблицы1

1	2	3	4	5
Программные логические модели (как развитие иерархического подхода)[1]	Оценка через цепочку: входы → действия → выходы → результаты → воздействие (итоговое изменение). Четкое разделение того, что делает компания и что происходит с рынком (аудиторией).	Для каждого элемента цепочки подбирают показатели: вход (затраты), действия (посты, активность в социальных сетях и т.д.), выход (охват, настроение, просмотры и т.д.), краткосрочный результат (подписчики, лайки, комментарии, переходы по ссылкам, удовлетворенность и т.д.), среднесрочный результат (намерение действовать, регистрации и т.д.), долгосрочный результат (удовлетворенность клиентов, лояльность и т.д.), итоговое изменение (рост продаж, рост инвестиций и т.д.) [1]	Позволяет избежать «ошибки подстановки», учитывает долгосрочные эффекты, интегрирует обучение	Требует предварительного построения модели и сбора данных на всех этапах; сложность для разовых кампаний.

Продолжение таблицы 1

1	2	3	4	5
Сбалансированная система показателей	Многомерная оценка по нескольким перспективам: финансовая, клиентская, внутренних процессов, обучения и роста [2].	Финансовая перспектива: показатели экономической отдачи от SMM-активности; клиентская перспектива: показатели восприятия бренда и взаимодействия с аудиторией; перспектива внутренних процессов: показатели операционной эффективности управления социальными сетями; перспектива обучения и роста: показатели развития компетенций и инновационного потенциала. Детализация данных перспектив с выделением 65 конкретных показателей представлена в работе [3], где на основе экспертных оценок также определены весовые коэффициенты для каждой группы.	Комплексный взгляд, связь стратегии и операционных метрик, баланс кратко- и долгосрочных целей.	Сложность внедрения, требует каскадирования целей; причинно-следственные связи не всегда очевидны [4].

Продолжение таблицы 1

1	2	3	4	5
Многокритериальные модели	Оценка предполагает формирование системы сгруппированных показателей, каждый из которых отражает определенную функциональную область или цель присутствия в социальных сетях.	В.Л. Музыкант [5, с. 338] предлагает оценивать эффективность SMM по ряду индикаторов, объединенных по типам эффективности: медийные, финансовые, анализ поведения аудитории. Зарубежные исследователи предлагают модель «4 Social C», включающую четыре критериальные группы: коммерция (Commerce), контент (Content), сообщество (Community), клиентский сервис (Customer Care) [6].	Учитывает разные функции SMM, гибкость в настройке под специфику компании.	Может быть избыточной для малого бизнеса; требует определения весов каждой группы и расчета большого числа показателей.

Источник: разработано с использованием [1–6]

3. Переход от фрагментарного ретроспективного анализа **к целостному циклу измерения, оценки и обучения** (MEL – Measurement, Evaluation, Learning).

В отличие от традиционных подходов, лишь констатирующих достигнутое или указывающих на неэффективность коммуникаций, модель MEL встраивает данные в цикл непрерывного улучшения стратегий.

Она поддерживает перспективное планирование и непрерывное совершенствование, превращая данные из инструмента отчетности в основу для стратегического развития [1].

Модель MEL включает следующие элементы [1]:

1.Четыре фундаментальных компонента:

- Теория изменений – определяет желаемые результаты и выстраивает логику «обратного проектирования».
- Программные логические модели – визуализируют теорию изменений, определяя входные данные, действия, результаты и итоги, необходимые для достижения целей.
- Ключевые концепции и принципы – три типа оценки (формирующая, процессная, итоговая), SMART-цели, KPI.
- Таксономия – категоризированный перечень методов и показателей с указанием методологии (количественный/качественный), метода сбора и этапа оценки (деятельность, результат, воздействие).

2.Два практико-ориентированных компонента:

- Конкретные методы и показатели для проектов и кампаний (с использованием программного обеспечения).
- Подходы к обучению персонала и применению полученных знаний для непрерывного совершенствования.

3.Организационные ценности: подотчетность и прозрачность; принятие решений на основе фактических данных; поддержка MEL на всех уровнях компании.

4.Организационное слушание, включая процессы сбора и анализа обратной связи от аудитории, анализ данных для выявления изменений и их причин.

Представленный обзор современных подходов к оценке позволяет взглянуть на оценку эффективности под разными углами.

Однако само наличие столь различных подходов указывает на то, что ни один из них не является универсальным и не снимает полностью проблем, возникающих при попытке измерить реальную отдачу от SMM.

На практике организации по-прежнему сталкиваются с рядом фундаментальных вызовов, как объективно присущих цифровой среде, так и связанных с ограничениями существующих инструментов и метрик.

Рассмотрим ключевые из них подробнее.

- «Метрики тщеславия». Организации часто попадают в ловушку погони за такими показателями, как количество подписчиков, общий охват и суммарное число лайков, которые напрямую не связаны с эффективностью работы.
- Стандартные отчеты, предоставляемые самими социальными платформами, перегружены такими показателями. Например, высокий уровень вовлеченности может быть следствием провокационного поста или негативной новости о компании. Анализ без учета контекста публикаций приводит к неверным стратегическим выводам.
- Эмоциональный тон и качество обратной связи. Современные алгоритмы анализа тональности несовершенны.
- Проблема атрибуции в многоканальной среде. Покупке часто предшествует сложный путь взаимодействий с пользователем: реклама в одной сети, обсуждение в другой, переход из поисковика, изучение сайта. Модель «последнего клика» не учитывает этот путь, искажая реальный вклад социальных сетей. Для офлайн-бизнеса проблема усугубляется – отследить связь между постом в социальных сетях и визитом в магазин без специальных инструментов (промокоды, опросы) практически невозможно.
- Временная задержка эффекта. Финансовые вложения в создание полезного контента (статьи, обзоры) могут не приносить мгновенных продаж, но формируют долгосрочный авторитет. Клиент, прочитавший экспертную статью, при возникновении потребности с большей вероятностью выберет именно авторитетную для него компанию (и это может произойти спустя месяцы и даже годы).

Стандартные отчеты зачастую не фиксируют такую отложенную конверсию.

Из сказанного следуют, что скорее всего не может существовать единого универсального набора показателей, одинаково пригодного для оценки эффективности любой организации в социальных сетях.

То, что является успехом для некоммерческой организации (укрепление доверия), может быть провалом для интернет-магазина (отсутствие прямых продаж).

Таким образом, преодоление сложившегося методологического противоречия видится в отказе от универсальных метрик в пользу разработки собственных (отраслевых) показателей, отражающих реальные бизнес-задачи конкретной компании или группы компаний (отрасли).

Заключение. Несмотря на многолетнюю практику использования социальных сетей в маркетинге, проблема достоверной оценки их эффективности до сих пор не имеет универсального решения.

Ключевая проблема современных организаций заключается в использовании стандартизированных метрик, которые не способны отразить ни реальную коммуникационную глубину, ни точную коммерческую отдачу.

Наиболее перспективный путь преодолеть этот разрыв и доказать окупаемость затрат является отказ от универсальных шаблонов в пользу разработки индивидуальной (или отраслевой) системы показателей эффективности.

Такая система должна строиться на основе конкретных бизнес-задач, данных сквозной аналитики и включать в себя как количественные, так и качественные критерии.

Список литературы

[1] Macnamara, J. Jim Macnamara's MEL Manual for Public Communication / J.Macnamara. – 2024. – 42 p.

[2] Kaplan, R. S. Strategy maps : converting intangible assets into tangible outcomes / R. S. Kaplan, D. P. Norton. – Boston : Harvard Business School Publishing, 2004.

[3] Jami pour, M. A Framework for Social Media Marketing Evaluation in SMEs: A Mixed-methods Approach / M. Jami pour, S. M. Jafari, F. Javidi // Journal of Modern Marketing Research. – 2023. – № 12(4). – PP. 1–28.

[4] Madsen, D. Balanced Scorecard: History, Implementation, and Impact / D. Madsen // Encyclopedia. – 2025. URL:https://www.researchgate.net/publication/389984126_Balanced_Scorecard_History_Implementation_and_Impact/references.

[5] Музыкант, В. Л. Основы интегрированных коммуникаций: теория и современные практики в 2 ч. Часть 2. SMM, рынок M&A / В. Л. Музыкант. – 2-е изд. – М : Юрайт, 2021 – 507 с.

[6] Zaugg, A. D. Social Media Controlling – die 4 Social C / A. D. Zaugg, U. Egle // HMD Praxis der Wirtschaftsinformatik. – 2014. – № 50(5). – S. 86–92.

[7] Macnamara, J. «PESO» media strategy shifts to «SOEP»: Opportunities and ethical dilemmas / J. Macnamara, M. Lwin, A. Adi, A. Zerfass // Public Relations Review. – 2016. – № 42(3). – PP. 377–385.

Авторский вклад

Батура Михаил Павлович – постановка проблемы и цели исследования, определение общей концепции и методологии работы, научное руководство исследованием, формулирование основных положений, ключевых выводов, разработка структуры статьи.

Марахина Инна Викторовна – разработка структуры статьи, проведение сравнительного анализа подходов к оценке эффективности SMM, формулирование основных положений и заключения.

Пархименко Владимир Анатольевич – сбор и критический анализ литературных источников, выявление основных проблем оценки эффективности, участие в обсуждении результатов.

ON THE ISSUE OF SYSTEMIC PERFORMANCE EVALUATION OF ORGANIZATIONS IN SOCIAL MEDIA

M.P. Batura
Head of the Research Laboratory, Research Laboratory 6.3, BSUIR, Doctor of Technical Sciences, Professor

I.V. Marakhina
Associate Professor at the Department of Economics, BSUIR, Candidate of Economic Sciences (PhD in Economics), Associate Professor

U.A. Parkhimenko
Head of Economics, BSUIR, Candidate of Economic Sciences (PhD in Economics), Associate Professor

Abstract. The intensified use of social media as a channel for marketing communications and the corresponding growth of SMM budgets bring to the fore the issue of evaluating the economic return on these investments. The article presents a critical analysis of existing approaches to measuring the effectiveness of organizational activities in social media.

Keywords: social media, SMM, performance evaluation, performance metrics, communication effectiveness, commercial effectiveness.

УДК 004.657:004.75-047.44

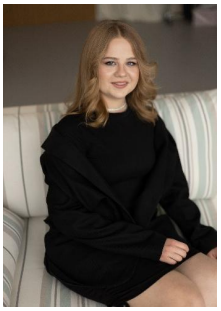
ЭФФЕКТИВНОЕ УПРАВЛЕНИЕ БОЛЬШИМИ ДАННЫМИ: СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПЛАТФОРМ И ТЕХНОЛОГИЙ ДЛЯ ОБРАБОТКИ ИНФОРМАЦИИ



В.В. Верняховская
Заместитель декана инженерно-экономического факультета БГУИР, магистр экономических наук
verniahovskaya@bsuir.by



О.М. Раптунович
Старший преподаватель кафедры экономической информатики БГУИР, магистр
oraptunovich@gmail.com



В.А. Усова
Студентка инженерно-экономического факультета, специальности Информационные системы и технологии БГУИР
uvikamail@gmail.com

В.В. Верняховская

Окончила Белорусский государственный университет информатики и радиоэлектроники. Магистр экономических наук. Работает заместителем декана инженерно-экономического факультета БГУИР. Направления исследований: трансфер технологий, инновационная деятельность, информационные технологии в маркетинге.

О.М. Раптунович

Окончила Частый Институт Управления и Предпринимательства. Область научных интересов связана с актуальными вопросами экономики, логистики, маркетинга и информационных технологий и способами их решений.

В.А. Усова

Обучается в Белорусском государственном университете информатики и радиоэлектроники. на третьем курсе инженерно-экономического факультета. Направления исследований: информационные технологии в экономике, искусственный интеллект, машинное обучение, анализ данных.

Аннотация. В статье рассматривается эффективное использование больших данных и анализируются платформы и технологии для анализа данных. Используются различные инструменты и методы для анализа, хранения и обработки данных. Внимание уделено платформам Hadoop, Apache Spark и другим, а также технологии машинного обучения, влияющие на производительность и масштабируемость обработки данных.

Ключевые слова: большие данные, Hadoop, Apache Spark, машинное обучение, MapReduce, NoSQL базы данных, анализ данных, Kafka, Flink.

Введение. Если посмотреть на то, как изменился мир за последние десять лет, можно заметить, что данные стали неотъемлемой частью нашей жизни. По мере того, как развивались технологии объемы информации, которая поступала из разных источников, стали расти в геометрической прогрессии. В какой-то момент стало понятно, что традиционные методы хранения и анализа не справляются с таким объемом. После этого и появилась концепция «большие данные» (Big Data). На наш взгляд, важно понимать, что «большие данные» – это не про огромные размеры, а про объемы информации, которые превышают возможности обычных, традиционных систем для их обработки и хранения. Такие данные отличаются не

только внушительными размерами, но и разнообразием форматов, скоростью поступления и непостоянностью [1].

После того как стали распространены Интернета вещей, социальные сети, мобильные устройства и цифровые платформы объем создаваемых данных продолжает увеличиваться с экспоненциальной скоростью. Только за последние несколько лет было создано около 90% всех мировых данных. Если рассматривать более продолжительный период, то с 2010 года общий объем информации в мире вырос более чем в 66 раз (рисунок 1). В 2024 году данные достигли 147 зеттабайт, что почти на 22,5% больше, чем в предыдущем году. К 2025 году данные выросли до 181 зеттабайт, а к 2026 – до 221 зеттабайт [2].

Таким образом, только за один год объем произведенных данных во всем мире увеличился почти на 40 зеттабайт, что составляет рост примерно на 22,09% по сравнению с 2025 годом. На наш взгляд такая динамика наглядно подтверждает, что с каждым годом объемы данных будут только увеличиваться, и их влияние на все сферы человеческой жизни также будет только возрастать. Хранение, анализ, эффективная и правильная обработка информации станут одними из главных задач для бизнеса, науки и государства [3].



Рисунок 1. Количество данных, создаваемых в мире

Чтобы работать с данными нужны специальные инструменты и платформы, которые смогут расти вместе с объемами информации и эффективно работать с ними. Для этого приходится использовать сложные методы, такие как аналитические алгоритмы и распределенные вычисления, которые смогли бы извлекать нужную и полезную информацию из необработанных данных. Важно, чтобы платформа умела работать с разными типами данных, не только с текстом и числами, но и видео, аудио и другими. На данный момент таких технологий не мало, и каждая хороша по-своему. Выбор зависит от того, с какими данными планируется работа, какая стоит задача у проекта.

Цель нашей работы – сравнить платформы и технологии для работы с большими данными, разобрать их ключевые характеристики, преимущества и недостатки и определить для каких задач каждая подходит лучше.

Что такое большие данные? Большими данными (Big Data) называют огромные массивы структурированной и неструктурированной информации, с которыми обычные, традиционные базы данных и методы обработки уже не справляются. Сюда входит и текст, и видео, и аудио, и графика – все то, что ежедневно накапливается в самых разных сферах: от финансов и медицины до соцсетей и интернета вещей.

Описать большие данные можно через модель «3V». Объем (Volume) – это сами огромные массивы информации, и с каждым днем их становится все больше и больше. Например, посты в социальных сетях, покупки в интернете, и другие данные, которые каждый день генерируют люди могут достигать колоссальных масштабов.

Скорость (Velocity) – это про то, как данные быстро и часто поступают и их нужно обрабатывать практически мгновенно. В качестве примера можно привести поведение пользователей в реальном времени или данные с устройств интернета вещей. Разнообразие (Veriety) – информация приходит в самых разных формах: текст, видео, аудио, числа, из-за этого ее становится очень сложно анализировать.

Сейчас к этим трем характеристикам все чаще добавляют и другие. Появились модели «4V» и «5V». Сначала модель «3V» придумали, чтобы просто описать основные, главные черты больших данных, но спустя время стало понятно, что есть и другие важные аспекты. Первый – это достоверность (Veracity), то есть надежность данных. Если исходные данные неточные или противоречивые, то и выводы по ним получатся ошибочные. Второй – это ценность (Value). Большие данные сами по себе ничего не дают, важно уметь извлекать из них полезную информацию и применять ее на практике.

В некоторых случаях добавляют еще визуализацию (Visualization) и вариативность (Variability), чтобы подчеркнуть, как важно правильно представлять данные и учитывать, что они меняются со временем.

Источники больших данных самые разные. Интернет вещей (IoT) – это датчики и устройства, которые постоянно передают информацию о своем состоянии. Соцсети генерируют огромные объемы текстов, фото, видео, по которым можно изучать поведение людей. Транзакции в электронной коммерции, медицинские исследования, научные эксперименты – все это тоже требует работы с большими объемами данных, для того чтобы получить точные и надежные результаты.

Но то, что данных много и они сложные, вовсе не значит, что их нельзя обработать. Современные технологии с этим справляются. Распределенные вычисления и новые подходы к хранению, такие как Hadoop и NoSQL-базы данных, позволяют хранить информацию, которая не помещается в обычные реляционные базы. А инструменты вроде Apache Spark помогают быстро анализировать огромные массивы. Все это открывает новые возможности – от бизнеса до научных исследований [4].

В бизнесе большие данные помогают улучшить аналитику. Прогнозирование спроса, предоставление клиентам того, что им действительно нужно, улучшение обслуживания – это все становится возможным благодаря анализу данных. В медицине большие данные помогают диагностировать и лечить заболевания, а также разрабатывать новые лекарства. В науке они используются для математических моделей, изучения окружающей среды и освоения космоса. Но, несмотря на все преимущества, в работе с большими данными есть проблемы. Безопасность и конфиденциальность остаются важными и актуальными вопросами.

Получается, большие данные – это не просто модное слово, тренд, а важный инструмент, который меняет подход к принятию решений. С их помощью повседневная жизнь становится удобнее и открываются новые возможности в разных сферах.

Платформы и технологии для обработки больших данных. Один из ключевых подходов к работе с большими данными – распределенные вычисления. Суть подхода проста: задачи делятся на части и распределяются между множества узлов – это и позволяет эффективно обрабатывать гигантские массивы информации. На такой основе создаются мощные системы для хранения и анализа данных. Один из инструментов – Apache Hadoop. Это целая платформа для распределенного хранения и обработки данных. В ее основе лежит технология MapReduce, которая выполняет задачи параллельно на разных узлах по схеме «главный – подчиненные». В качестве главного выступает сервер JobTracker, раздающий задания подчиненным узлам кластера и контролирующий их выполнение (рисунок 2) [5].

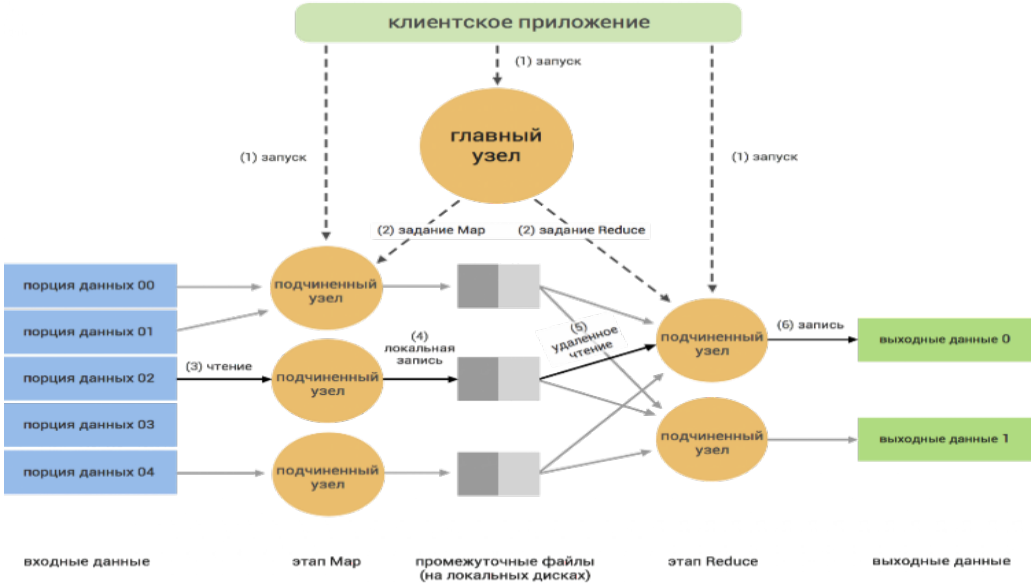


Рисунок 2. Архитектура Hadoop

Это открывает возможности для обработки петабайтов информации. Hadoop использует распределенную файловую систему HDFS (Hadeer Distributed File System), которая эффективно справляется с хранением больших объемов данных на различных машинах.

Но у MapReduce есть свои недостатки, например он не может быстро обрабатывать данные в реальном времени. Чтобы решить данную проблему разработали платформу Apache Spark, она более гибкая и быстрая для обработки больших данных.

В отличие от классического обработчика ядра Apache Hadoop с двухуровневой концепцией MapReduce на базе дискового хранилища, Spark использует специализированные примитивы для рекуррентной обработки в оперативной памяти. Благодаря этому многие вычислительные задачи реализуются в Смарк значительно быстрее. Сравнение Apache Hadeer и Spark представлено на рисунке 3.

Hadoop MapReduce	Apache Spark
Fast	100x faster than MapReduce
Batch Processing	Real-time Processing
Stores Data on Disk	Stores Data in Memory
Written in Java	Written in Scala

Рисунок 3. Сравнение Apache Hadoop и Spark

Стоит отметить, что Spark Streaming, в отличие от, например, Apache Storm, Flink или Samza, не обрабатывает потоки Big Data целиком. Для этого существует микропакетный подход. В нем поток данных разбивается на небольшие пакеты временных интервалов. Обычные реляционные базы данных, которым нужны жесткие схемы данных и которые не могут эффективно работать с огромными объемами информации не всегда подходят для хранения и обработки больших данных. В этой связи широко распространены NoSOL базы

данных, которые предназначены для обработки неструктурированных и полуструктурированных данных, таких как текст, изображения и видео [6].

Среди популярных NoSOL баз данных можно выделить следующие. Cassandra – масштабируемость и высокая доступность. Apache Cassandra – это распределенная NoSQL база данных, предназначенная для хранения больших объемов данных с высокой доступностью и масштабируемостью. Она была разработана для того, чтобы решать проблемы, возникающие при работе с огромными массивами данных в реальном времени. Cassandra выбирают компании, которым нужно, чтобы система продолжала работать, даже если несколько узлов базы данных вышли из строя. Основной особенностью Cassandra является архитектура peer-to-peer, в которой все узлы в кластере равны. Тут нет единой точки отказа, данные дублируются на нескольких узлах, поэтому они остаются доступными даже при сбоях. А если копии хранятся на серверах в разных географических точках, надежность становится еще выше. Cassandra используют там, где доступность критически важна, а именно в соцсетях, финансах, интернете вещей.

Еще плюсом выделяют гибкость в работе с данными. Cassandra использует модель колоночных семей, что позволяет удобно хранить и обрабатывать информацию, которая не укладывается в жесткие структуры или постоянно меняется. Система хорошо работает с огромными объемами данных разного формата и масштабируется по мере роста записей.

И главное – это то, что Cassandra выдерживает высокие нагрузки и отвечает на вопросы с минимальной задержкой, поэтому ее так любят компании, которым нужно обрабатывать большие данные в реальном времени [7].

MongoDB – одна из самых популярных NoSQL-баз. Она хранит данные в виде документов. В отличие от Cassandra, в ней используется формат похожий на JSON, поэтому MongoDB отлично подходит для работы с неструктурированными или не полностью структурированными данными, например, данные из соцсетей, веб-приложений или мобильных сервисов. Каждый документ может содержать разные типы данных: массивы, вложенные объекты – что угодно. Это дает разработчикам свободу: не нужно жестко задавать структуру заранее, ее можно менять по ходу дела, если меняются требования бизнеса или логика работы с данными.

MongoDB умеет ускорять запросы с помощью индексов, также поддерживает агрегацию, с помощью которой можно легко посчитать сумму, среднее или другие статистические показатели. Еще в ней есть встроенные инструменты для сложных запросов, поэтому для аналитики она подходит хорошо. Если данных становится слишком много, то можно добавить новые узлы в кластер, база масштабируется горизонтально без лишних сложностей. Чаще всего MongoDB используют в веб-разработке, мобильных приложениях, интернет-магазинах и задачах аналитики. Она прекрасно подходит для приложений, требующих быстрого доступа к данным, например, для ведения каталогов товаров или хранения пользовательских данных и предпочтений.

Apache HBase – это распределенная база данных, которая хорошо масштабируется. Она построена на основе Google Bigtable и предназначена для хранения и обработки больших объемов структурированной информации. HBase используется там, где важна высокая скорость записи и чтения, а также когда нужно работать с большими таблицами. Она хорошо подходит для хранения данных, которые удобно представляются в табличном виде, например, аналитические платформы, где хранятся большие наборы данных и требуется быстрая обработка.

Также как и Cassandra, HBase поддерживает репликацию, которая дает отказоустойчивость и надежный доступ к данным, но в отличие от MongoDB, где данные хранятся в виде документов, HBase использует табличную структуру и формат «ключ – значение», это позволяет эффективно работать с огромными объемами.

Одно из главных достоинств HBase – способность обрабатывать данные с очень высокой скоростью записи, поэтому ее часто выбирают для приложений, которые работают в реальном времени, например, системы мониторинга, веб-аналитики, обработки транзакций.

HBase не требует строгой системы, в ней можно хранить данные с разной структурой, система гибко масштабируется по мере роста объемов. Правда, в отличие от Cassandra и MongoDB, HBase заточена на работу с большими данными в кластерном режиме.

Cassandra, MongoDB и HBase – три мощные NoSQL-базы, у каждой свои сильные стороны и особенности, которые подходят для разных типов задач. Какую из них выбрать зависит от конкретного проекта, объемов данных и того, что нужно делать с приложением.

С ростом скорости генерации данных все важнее становится обработка в реальном времени. Платформы, которые это умеют, дают серьезное преимущество в бизнес-аналитике, мониторинге и управлении.

Apache Kalfka и Apache Flink представляют собой две платформы, которые часто применяются вместе для обработки данных в потоковом режиме. Kalfka является распространенной платформой для передачи потоковых данных. Ее характеристики – это высокая пропускная способность, масштабируемость и гарантии доставки сообщений. Платформа поддерживает хранение данных в заданный промежуток времени, что обеспечивает возможность их повторного анализа. Kalfka разделяет данные на темы, что позволяет улучшить отказоустойчивость и параллельную обработку.

Apache Flink ориентирован на обработку потоковых данных в реальном времени. Он берет данные из потоков и позволяет делать с ними сложные операции: аналитику, обработку событий, обучение моделей. Flink умеет управлять состоянием вычислений, работать с временными окнами и обрабатывать как бесконечные потоки, так и пакетные данные. Отметим, что события обрабатываются строго в том порядке, в котором поступили.

Вместе Kalfka и Flink работают просто: первый доставляет данные, второй обрабатывает. Такая пара хорошо подходит для мониторинга, аналитики в реальном времени и задач, связанных с машинным обучением. Вместе они дают высокую производительность, гибкость и возможность строить масштабируемые системы обработки потоков [8].

Технологии анализа данных и их применение. За последние 10 лет технологии анализа данных приобрели высокую значимость, потому что стали важной частью многих сфер (бизнес, наука, здравоохранение, государственное управление). Современные инструменты позволяют находить и получать полезную информацию из больших объемов данных, это открывает возможности для более обоснованных решений и оптимизации процессов.

Машинное обучение (Machine learning, ML) – это область искусственного интеллекта, в которой компьютер учится улучшать свои результаты на основе данных, без прописывания правил. Алгоритмы выявляют закономерности в примерах и потом применяют их к новым данным, которые раньше не встречались [9].

В зависимости от способа обучения выделяют три подхода. Обучение с учителем, при котором алгоритм обучается на заранее подготовленных данных, то есть на тех, где для каждого примера есть ответы. Обучение без учителя работает без заранее подготовленных данных, алгоритм сам ищет скрытые структуры, для этого группирует объекты (кластеризация), выявляет связи между ними. Этот подход применяется в рекомендательных системах. Обучение с подкреплением строится на взаимодействии с окружающей средой. Модель делает действие, получает положительную или отрицательную оценку и постепенно учится выбирать действия, которые приносили бы максимальную положительную оценку. Такой подход чаще используется в автономных системах.

Одним из популярных направлений машинного обучения в последнее время стало глубокое обучение. Он основан на многослойных нейронных сетях. Такие модели хорошо работают с большими объемами данных и сложными закономерностями.

Анализ данных шире чем машинное обучение. Он включает в себя и статистические методы и подходы на основе ИИ. Задачей анализа является выявление паттернов, трендов и

связи данных. Информация для анализа данных может поступать из разных источников, например, базы данных, файлы, социальные сети, веб-страницы. Перед анализом данные очищают: убирают пропуски, устраняют выбросы, нормализуют значения – от этого зависит качество следующих результатов. Далее применяют статистические методы (проверка гипотез, расчет средних и отклонений) или алгоритмы машинного обучения. Визуализируем результаты с помощью графиков, диаграмм, интерактивных панелей, которые помогают увидеть данные с новой стороны и делают результаты более наглядными. На основе полученных данных и прогнозов принимаются обоснованные решения.

Машинное обучение и анализ данных продолжают развиваться. Сочетание мощных алгоритмов с большими объемами информации позволяет строить более точные прогнозы, оптимизировать процессы и создавать новые продукты и услуги.

Проблемы и вызовы при работе с большими данными. Когда мы работаем с большими данными, появляется возможность для анализа и принятия решений, но еще с этим появляются проблемы, которые приходится решать, чтобы использование данных было эффективным.

Одна из ключевых проблем – это защита чувствительной информации, такой как личные данные пользователей, коммерческая тайна и другие сведения, неправомерное использование которых может иметь серьезные последствия. Чтобы минимизировать риски применяют шифрование, механизмы контроля доступа и другие методы защиты, но с ростом объемов собираемых данных обеспечивать их безопасность становится труднее.

Объем сам по себе не гарантирует ценности данных, их полезность определяется точностью, полнотой и актуальностью. На практике обычно встречаются ошибки в записях, пропущенные значения, дублирование, противоречивая или устаревшая информация. Некачественные данные дают ошибочные выводы и делают ненадежными модели, построенные на их основе, именно поэтому предобработка становится обязательным этапом. Но даже при использовании отработанных методов с ростом количества данных гарантировать отличное качество данных сложно.

Чем больше данных, тем больше нужно вычислительных мощностей, чтобы их обработать. Если сервера слабые или каналы передачи данных узкие, то сложные задачи выполнятся долго. В некоторых случаях время очень важно, например, когда нужно мониторить потоки данных в реальном времени, обновлять прогнозные модели без задержек. Справиться с этим помогают распределенные системы. Они разбивают работу на части и раскладывают их по разным узлам. Это немного улучшает производительность, но не решает проблему полностью. Данных становится все больше, задачи усложняются, и даже хорошо настроенная распределенная инфраструктура рано или поздно достигает своего предела.

Будущее технологий больших данных. Дальнейшее развитие технологий больших данных связано с ростом объемов информации, генерируемой людьми, устройствами и автоматизированными системами. Эффективное извлечение значимых сведений из этих массивов становится ключевым фактором для бизнеса, научных исследований и управления системами.

Одно из основных направлений – это интеграция больших данных с методами искусственного интеллекта и машинного обучения. Такое совмещение позволяет повысить точность анализа и прогнозирования, а также сократить время принятия решений. На сегодняшний день предсказательная аналитика уже применяется для оптимизации бизнес-процессов, прогнозирования спроса и диагностики заболеваний. В дальнейшем подобные решения будут внедряться в различные сферы жизни, например, в городское управление, где анализ данных в реальном времени с использованием ИИ может использоваться для регулирования транспортных потоков, контроля энергопотребления и мониторинга экологической обстановки.

Сейчас все больше внимания уделяют тому, как и где хранятся данные. Старые методы работы с большими массивами уже не справляются с потребностями новых типов данных,

потому что появляется много «тяжелых» форматов: видео, аудио, изображение. Приходится искать новые подходы, облака и распределенные системы вроде тех, что работают на облачных платформах, помогают гибко расширять мощности под такие нагрузки и делают хранение более надежным. В ближайшее время начнут серьезнее относиться к вопросам этики и безопасности данных. Чем больше информации собираем, тем выше риск, что она утечет или ее будут использовать не по назначению. Поэтому для компаний, которые работают с данными, на первый план выйдут не просто технические решения, а соблюдение законов, защита конфиденциальности пользователей и внятные стандарты безопасности. Технологии, такие как блокчейн, тоже могут сыграть ключевую роль в обеспечении прозрачности и безопасности обработки данных. В целом, все идет к тому, что технологии больших данных будут становиться только интереснее, так как каждый год появляются новые инструменты, которые позволяют лучше понимать мир вокруг и принимать более обоснованные решения. Такие технологии будут продолжать развиваться, будут менять не только способы ведения бизнеса, но и повседневную жизнь людей, делая ее удобнее и безопаснее.

Заключение. Большие данные прочно вошли в современную информационную среду.

Благодаря таким технологиям как Hadoop, NoSQL-базы и другие, организации теперь могут работать с объемами информации, которые раньше казались неподъемными. Но дело не только в том, чтобы хранить и обрабатывать это данные, подключение методов машинного обучения и искусственного интеллекта позволяет капать глубже и принимать решения не «на глазок», а на основе того, что реально видно из анализа. При этом как и в любой сложной системе, есть свои подводные камни. Безопасность, качество данных, производительность – все это остается вызовом, который не пропадает, а только усиливается со временем. Перспективы использования больших данные связаны с дальнейшем развитием технологий таких, как облачные вычисления, квантовые вычисления и автоматизацию процессов обработки данных.

Список литературы

[1] Тесленко И.Б., Губернаторов А.М., Дигилина О.Б., Крылов В.Е. BIG DATA=Большие данные: учебное пособие. Владимир: ВлГУ, 2021. – 123 с.

[2] Сколько данных создается каждый день? [Электронный ресурс]. – Режим доступа: https://inclient.ru/data-create-stats/. – Дата доступа: 01.03.2026

[3] Статистика больших данных за 2026 год [Электронный ресурс]. – Режим доступа: https://www.demandsage.com/big-data-statistics/. – Дата доступа: 01.03.2026

[4] Обработка потоковых данных [Электронный ресурс]. – Режим доступа: https://datareview.info/article/obrabotka-potokovyx-dannyx-storm-spark-i-samza/. – Дата доступа: 01.03.2026

[5] Обработка больших данных: первые шаги в понимании Hadoop MapReduce и Spark [Электронный ресурс]. – Режим доступа: https://devby.io/news/luxoft-big-data#google_vignette. – Дата доступа: 01.03.2026

[6] Модели данных в NoSQL [Электронный ресурс]. – Режим доступа: https://habr.com/ru/companies/otus/articles/760226/. – Дата доступа: 01.03.2026

[7] Модель данных Apache Cassandra [Электронный ресурс]. – Режим доступа: https://bigdataschool.ru/wiki/cassandra#. – Дата доступа: 01.03.2026

[8] Построение архитектуры данных реального времени с помощью Apache Kafka, Flink и Druid [Электронный ресурс]. – Режим доступа: https://clck.ru/3GSU89. – Дата доступа: 01.03.2026

[9] В. В. Воронина, А. В. Михеев, Н. Г. Ярушкина, К. В. Святов. Теория и практика машинного обучения: учебное пособие. Ульяновск : УлГТУ, 2017. – 290 с.

Авторский вклад

Верняховская Вероника Владимировна – сформулировала задачу исследования, разработала структуру статьи, провела анализ платформ обработки больших данных, участвовала в формировании теоретической и методологической базы.

Раптунович Ольга Михайловна – выполнила сравнительный анализ технологий обработки и хранения данных, описала архитектуру и принципы работы рассматриваемых систем, подготовила иллюстрации и участвовала в написании основной части статьи.

Усова Виктория Александровна – исследовала методы анализа данных и машинного обучения, встроила полученные результаты в общую концепцию работы, сформулировала выводы и заключение, выполнила редактирование и подготовку статьи к публикации.

EFFECTIVE MANAGEMENT OF BIG DATA: A COMPARATIVE ANALYSIS OF PLATFORMS AND TECHNOLOGIES FOR INFORMATION PROCESSING

V.V. Verniahovskaya
Master of economics, Deputy Dean of the faculty of engineering and Economics at BSUIR

O.M. Raptunovich
Senior Lecturer at the Department of Economic Informatics BSUIR, master

V.A. Usova
Student of BSUIR, Faculty of Engineering and Economics, specialty Information Systems and Technologies

Abstract. The article examines the effective use of big data and analyzes platforms and technologies for data analysis. Various tools and methods are used for data analysis, storage, and processing. Attention is paid to the platforms Hadoop, Apache Spark and others, as well as machine learning technologies that affect the performance and scalability of data processing.

Keywords: big data, Hadoop, Apache Spark, machine learning, MapReduce, NoSQL databases, data analysis, Kafka, Flink.

УДК 004.896

ИСПОЛЬЗОВАНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ПРОМЫШЛЕННОЙ БЕЗОПАСНОСТИ



Е. Н. Бычек
Студентка кафедры инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроники
bychekeliza@gmail.com



Н. В. Щербина
Старший преподаватель кафедры инженерной психологии и эргономики Белорусского государственного университета информатики и радиоэлектроник, магистр технических наук
shcherbina@bsuir.by

Е. Н. Бычек

Обучается на третьем курсе Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с информационными системами и технологиями в обеспечении промышленной безопасности».

Н. В. Щербина

Окончила Белорусский национальный технический университет. Область научных интересов связана с исследованием проблем саморегуляции функционального состояния человека, аспектов деятельности человека-оператора в системах контроля и управления, промышленной безопасностью, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация: Проанализирована роль технологий искусственного интеллекта (ИИ) в трансформации систем промышленной безопасности на основе актуальных отраслевых источников. Рассмотрены ключевые направления применения ИИ: от предиктивной аналитики и компьютерного зрения до интеллектуальных средств индивидуальной защиты. Особое внимание уделено сдвигу от реактивных к проактивным стратегиям предотвращения инцидентов. В работе также систематизированы основные этические и организационные вызовы, связанные с внедрением ИИ, включая вопросы доверия, прозрачности и приватности.

Ключевые слова: промышленная безопасность, искусственный интеллект, компьютерное зрение, интеллектуальные носимые устройства, предиктивная аналитика, этика, безопасность труда, мониторинг, отказ оборудования.

Введение. Концепция Индустрии 4.0, или Четвертой промышленной революции, центрирована вокруг данных, аналитики и взаимосвязанных систем. Ее ядро составляет способность не просто собирать информацию, но и использовать ее для прогнозирования и решения проблем до их возникновения, что имеет прямое отношение к сфере промышленной безопасности. Внедрение искусственного интеллекта знаменует переход от традиционных, зачастую реактивных методов обеспечения безопасности (чек-листы, периодические проверки) к проактивной, предсказывающей модели. Как отмечают эксперты, ИИ способен анализировать массивы данных с беспрецедентной скоростью, обнаруживая паттерны, невидимые человеческому глазу, что позволяет прогнозировать риски и предотвращать инциденты. Актуальность этой трансформации подчеркивается глобальной статистикой: по данным Международной организации труда, ежегодно происходит около 317 миллионов несчастных случаев на производстве [1].

Основная часть. Практическое применение искусственного интеллекта в промышленной безопасности реализуется через несколько взаимодополняющих технологических направлений:

1 *Предиктивная аналитика и мониторинг оборудования.* Одно из наиболее значимых применений ИИ – прогнозирование отказов оборудования, которое основывается на обработке и анализе больших объемов данных. Алгоритмы машинного обучения анализируют исторические и текущие данные с датчиков IoT для отслеживания ключевых показателей (вибрация, температура, давление и др.), выявляют незначительные отклонения от нормальных закономерностей, и как следствие, прогнозируют потенциальные отказы до их возникновения, что позволяет находить долгосрочные решения. Такой подход позволяет перейти от планово-предупредительного обслуживания к обслуживанию по фактическому состоянию, минимизируя незапланированные простои и снижая риски аварий, связанных с внезапным выходом техники из строя. На рисунке 1 показаны преимущества ИИ в предоставлении отказов оборудования. Данный подход формирует методологическую основу Индустрии 4.0, в рамках которой данные служат инструментом для упреждающего выявления и нивелирования потенциальных проблем [2].



Рисунок 1. Преимущества ИИ в предотвращении отказов оборудования

2 *Компьютерное зрение для анализа поведения и условий труда*. Системы на основе компьютерного зрения используют камеры и сенсоры для непрерывного автоматического мониторинга производственной среды в реальном времени. Они способны распознавать потенциальные опасности, такие как: отсутствие предписанного средства индивидуальной защиты (СИЗ), несанкционированное нахождение в опасной зоне, небезопасное обращение с оборудованием или появление таких рисков, как разливы жидкостей. Данные системы предоставляют инструмент для постоянного аудита соблюдения протоколов безопасности, выступая в роли «дополнительной пары глаз», которая не устает и не отвлекается.

На рисунке 2 показаны преимущества компьютерного зрения для анализа поведения и условий труда на промышленных объектах [3-5].

3 *Интеллектуальные носимые устройства и биометрический мониторинг*. Интеллектуальные носимые устройства и системы биометрического мониторинга стали неотъемлемой частью концепции Индустрии 4.0, превращая традиционные СИЗ в активные цифровые инструменты.

Современные СИЗ эволюционируют в сторону интеграции с ИИ.

Умные каски, жилеты и браслеты, оснащенные датчиками, отслеживают не только окружающую среду (наличие газов, экстремальные температуры и т.п.), но и физиологическое состояние работника (уровень усталости, тепловой стресс, частоту сердечных сокращений и т.п).

В случае выявления опасных отклонений устройство может подать предупреждение самому работнику и/или диспетчеру.

Кроме того, разрабатываются AI-экзоскелеты, которые снижают физическую нагрузку на оператора, предотвращая травмы опорно-двигательного аппарата при работе с тяжестями [1].



Рисунок 2. Преимущества компьютерного зрения

На рисунке 3 показаны преимущества интеллектуальных носимых устройств и биометрического мониторинга [6-8].

4 *Автоматизация опасных задач и инспекций.* ИИ и робототехника берут на себя выполнение задач в средах с недопустимо высоким риском для человека.

Это включает использование автономных дронов для инспекции труднодоступных или опасных объектов (например, высотных конструкций, нефтехимических резервуаров) и развертывание роботов для работ с токсичными веществами или в зонах возможных обрушений.

Таким образом, минимизируется непосредственная экспозиция персонала к угрозам.



Рисунок 3. Преимущества интеллектуальных носимых устройств и биометрического мониторинга

На рисунке 4 показаны преимущества использования автоматизации опасных задач и инспекций на производственных объектах [9-11].

Американское общество инженеров по безопасности (ASSP) в своей позиции обозначило четыре фундаментальных принципа ответственного использования ИИ в безопасности, которые служат ориентиром для организаций [4]:

1 *Доверие*: ИИ должен усиливать навыки и способности специалистов по безопасности, а не заменять человеческое суждение и принятие решений. Окончательный контроль и контекстуальная оценка рисков должны оставаться за человеком.

2 *Прозрачность*: Работники, менеджеры и руководители должны быть информированы о возможностях и ограничениях, используемых ИИ-технологий. Открытость в отношении того, как системы работают и какие данные обрабатывают, формирует культуру доверия и подотчетности.

3 *Справедливость*: Необходимо гарантировать, что ИИ-технологии не усугубляют существующие и не создают новые формы дискриминации в области охраны труда и промышленной безопасности. Алгоритмы должны быть свободны от предвзятостей, которые могут негативно сказаться на определенных группах работников.



Рисунок 4. Преимущества использования автоматизации опасных задач и инспекций на производственных объектах

4 *Конфиденциальность*: Организации обязаны внедрять надежные меры защиты для предотвращения несанкционированного доступа, неправомерного использования или эксплуатации конфиденциальной информации, собираемой ИИ-системами. Защита персональных данных работников является критически важной.

Несмотря на ряд преимуществ использования ИИ в промышленной безопасности, практическое внедрение ИИ-технологий в производственный процесс затруднено из-за ряда финансовых (высокие инвестиции), технических (сложность интеграции), социальных (сопротивление работников) проблем, проблем кибербезопасности, а также этических аспектов (справедливость и непредвзятость, прозрачность и объяснимость, ответственность, приватность и безопасность данных).

Заключение. Искусственный интеллект утвердился в качестве трансформационного инструмента в области промышленной безопасности, предлагая переход от исторически сложившейся реактивной модели к проактивной, основанной на данных. Технологии предиктивной аналитики, компьютерного зрения и интеллектуальных носимых устройств демонстрируют значительный потенциал для предотвращения аварий, снижения травматизма и оптимизации производственных процессов. Однако, как и любая мощная технология, ИИ требует ответственного и взвешенного подхода к внедрению. Ключом к успешной интеграции

является не слепая автоматизация, а создание симбиотической системы, где ИИ выступает в роли мощного аналитического инструмента, расширяющего возможности специалистов по безопасности, а не заменяющего их профессиональное суждение. Соблюдение принципов доверия, прозрачности, справедливости и конфиденциальности, сформулированных профессиональным сообществом, является обязательным условием для построения эффективных, этичных и принимаемых работниками систем безопасности будущего. Таким образом, будущее промышленной безопасности лежит не в противопоставлении человека и машины, а в их разумной интеграции, где технология служит для защиты самого ценного ресурса – человеческой жизни и здоровья. Внедрение ИИ-систем потребует психологического сопровождения в производственном секторе, которое может быть представлено следующими направлениями:

1) технологическая тревога: психологическая подготовкой работников, объясняющая, что ИИ – это инструмент, а не замена работника, а также профилактика цифрового отчуждения для предотвращения чувства контроля над процессом, когда работник перестает понимать, почему система приняла то или иное решение;

2) поддержка в условиях проактивной модели: помочь в адаптации к работе в условиях неопределенности и доверия ИИ-системам, а также помочь в оптимизации режимов труда работника во избежание потери концентрации внимания;

3) этический и психологический мониторинг: помочь психолога в реальном времени либо через обучение ИИ-систем принятию решения о критических показателях жизнедеятельности работника, а также этическая безопасность персональных данных;

4) цифровое доверие: недопущение гипердоверия и преодоления недоверия.

Список литературы

[1] How AI is transforming workplace safety for engineers [Электронный ресурс]. – Режим доступа : https://www.eit.edu.au/ai-is-transforming-workplace-safety-for-engineers/. – Дата доступа: 16.01.2026.

[2] Industry 4.0 is all about analytics and industrial data [Электронный ресурс]. – Режим доступа : https://www.wizata.com/knowledge-base/industry-4.0-is-all-about-analytics-and-industrial-data. – Дата доступа: 18.01.2026.

[3] The disrupt Labs [Электронный ресурс]. – Режим доступа : https://thedisruptlabs.com/what-is-computer-vision-ai/. – Дата доступа: 20.01.2026.

[4] Defining the Role of AI in Safety [Электронный ресурс]. – Режим доступа : https://www.assp.org/news-and-articles/defining-the-role-of-ai-in-safety. – Дата доступа: 18.01.2026.

[5] Zest Minds Making Technology Work for you [Электронный ресурс]. – Режим доступа : https://www.zestminds.com/blog/real-world-applications-computer-vision/#:~:text=When%20combined%20with%20IoT%20data,and%20clinician%20support%2C%20not%20replacement. – Дата доступа: 20.01.2026.

[6] Ajprotech [Электронный ресурс]. – Режим доступа : https://ajprotech.com/blog/internet-of-things/top-wearables-of-2026-trends-in-health-and-fitness.html#:~:text=AI%2DDriven%20Sensor%20Fusion%20and,security%20by%20minimizing%20cloud%20dependency. – Дата доступа: 20.01.2026.

[7] Cerexio [Электронный ресурс]. – Режим доступа : https://cerexio.com/blog/benefits-of-employing-wearable-technology-in-manufacturing-cerexio-singapore#:~:text=It%20is%20evident%20that%20wearable,production%20lines%20and%20reduces%20mistakes. – Дата доступа: 20.01.2026.

[8] Министерство по чрезвычайным ситуациям Республики Беларусь [Электронный ресурс]. – Режим доступа : https://niipb.mchs.gov.by/novosti/504370/. – Дата доступа: 20.01.2026.

[9] Yamato [Электронный ресурс]. – Режим доступа : https://www.yamatoscale.co.uk/automation-in-manufacturing-the-impact-on-safety-and-efficiency/#:~:text=Outside%20of%20safety%20concerns%2C%20there,results%20in%20a%20shorter%20timeframe. – Дата доступа: 21.01.2026.

[10] Field Eagle [Электронный ресурс]. – Режим доступа : https://www.fieldeagle.com/blog/top-technology-trends-watch-industrial-inspections-2026/#:~:text=Align%20safety%20and%20environmental%20goals,December%2029%2C%202025. – Дата доступа: 21.01.2026.

[11] Национальный правовой Интернет-портал Республики Беларусь [Электронный ресурс]. – Режим доступа : https://pravo.by/novosti/novosti-pravo-by/2022/december/72610/. – Дата доступа: 20.01.2026.

Авторский вклад

Авторы внесли равнозначный вклад.

USE OF ARTIFICIAL INTELLIGENCE IN INDUSTRIAL SAFETY

E. N. Bychek
Student of the Department of Engineering Psychology and Ergonomics, BSUIR

N. V. SHCHerbina
Senior Lecturer, Department of Engineering Psychology and Ergonomics, BSUIR, Master of Technical Sciences

Abstract. The article analyzes the role of artificial intelligence (AI) technologies in the transformation of industrial safety systems based on current industry sources. Key areas of AI application are considered: from predictive analytics and computer vision to intelligent personal protective equipment. Special attention is paid to the shift from reactive to proactive incident prevention strategies. The work also systematizes the main ethical and organizational challenges associated with the implementation of AI, including issues of trust, transparency, and privacy.

Keywords: industrial safety, artificial intelligence, computer vision, intelligent wearables, predictive analytics, ethics, occupational safety, monitoring, equipment failure.

УДК 004.8:004.912

О СООТВЕТСТВИИ АВТОМАТИЧЕСКОЙ ПРОЦЕДУРЫ РАЗМЕТКИ МЕТОДОМ ЛУЧШЕГО-ХУДШЕГО МАСШТАБИРОВАНИЯ РЕШЕНИЮ ЗАДАЧ АНАЛИЗА ТОНАЛЬНОСТИ ФИНАНСОВЫХ НОВОСТЕЙ







Г.Б. Никифоров
Ассистент отделения информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ
nikiforov.gb@pskgu.ru

В.И. Пименов
Заведующий кафедрой информационных технологий, СПбГУПТД, доктор технических наук, с.н.с.
v_pim@mail.ru

Д.А. Андреев
Заведующий отделением информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ, кандидат технических наук
d.andreev@pskgu.ru

Г.Б. Никифоров
Окончил Псковский государственный университет. Область научных интересов связана с разработкой методов и алгоритмов обработки естественного языка и машинного обучения в финансовой сфере.

В.И. Пименов
Окончил Балтийский государственный технический университет «Военмех» имени Д.Ф. Устинова. Область научных интересов связана с разработкой методов и алгоритмов интеллектуального анализа данных, когнитивных технологий, компьютерных систем обработки и анализа данных, компьютерной графики, распознавания образов, мультимедиа технологий и 3D-моделирования.

Д.А. Андреев
Окончил Псковский государственный политехнический институт. Область научных интересов связана с разработкой моделей, алгоритмов и показателей качества формализованного описания и анализа технологий производства продукции.

Аннотация. В работе рассматривается автоматическая процедура разметки методом лучшего-худшего масштабирования (BWS) для непрерывного анализа тональности финансовых новостных заголовков на русском языке. На наборе данных из 532 финансовых новостных заголовков выполнено сопоставление автоматической процедуры разметки относительно ручной. Показано, что автоматическая процедура разметки сохраняет межаннотаторскую согласованность (*к ≈ 0,30) и внутреннюю надёжность при расщеплении (SHR ≈ 0,87), обеспечивает сопоставимость итоговых оценок (R2 ≈ 0,49). Результаты подтверждают применимость предобученной большой языковой модели (LLM) без дообучения на наборе данных для масштабируемой количественной разметки финансовых текстов в условиях ограниченного объёма данных.*

Ключевые слова: искусственный интеллект, машинное обучение, предобученные большие языковые модели, автоматическая разметка, анализ тональности.

Введение. Количественный анализ тональности финансовых новостей является значимым компонентом моделей прогнозирования рыночных показателей и систем поддержки инвестиционных решений [1, 2].

В отличие от задач классификации с дискретными метками, в прикладных финансовых исследованиях востребованы непрерывные оценки тональности, отражающие интенсивность и направленность информационного воздействия.

Формирование таких оценок осложняется ограниченным объёмом размеченных наборов данных и выраженной субъективностью интерпретации финансовых новостей [3].

Одним из подходов к получению непрерывных оценок является метод лучшего-худшего масштабирования (Best-Worst Scaling, BWS), основанный на относительном сравнении текстов и обеспечивающий устойчивость итоговых оценок при сравнении объектов.

Ручная процедура разметки методом лучшего-худшего масштабирования характеризуется высокой трудоёмкостью и низкой масштабируемостью, что затрудняет её применение при потоковой обработке данных.

Умеренная согласованность между аннотаторами дополнительно увеличивает требования к объёму разметки для обеспечения устойчивости оценок.

Развитие предобученных больших языковых моделей создаёт предпосылки для автоматизации сравнительной разметки, включая автоматическую процедуру разметки методом лучшего-худшего масштабирования без дополнительного обучения на специализированных наборах данных [4].

При этом в условиях отсутствия объективного эталона разметки оценка качества автоматической разметки требует сопоставления с ручными оценками и анализа метрик качества разметки [5].

Это обуславливает необходимость эмпирической проверки сопоставимости автоматической и ручной процедур разметки методом лучшего-худшего масштабирования по характеристикам согласованности и практической применимости в задачах количественного анализа тональности.

Постановка задачи. В исследовании использован набор данных FiNeS [6], содержащий финансовые новостные заголовки и их ручную разметку.

Ручная разметка выполнена методом лучшего-худшего масштабирования, предполагающего попарные сравнительные задачи выбора наиболее и наименее позитивного текста.

Пусть задано множество финансовых новостных заголовков:

$$T=t1,t2,\ldots,tN, \quad (1)$$

где t_i – один финансовый новостной заголовок; $i = 1, 2,\ldots, N$ – индекс финансового новостного заголовка; $N = 532$ – количество финансовых новостных заголовков в наборе данных FiNeS.

Для каждого финансового новостного заголовка t_i определяется непрерывная оценка тональности $si \in [-1, 1]$.

В рамках исследования рассматриваются две её реализации: H – результат ручной процедуры разметки методом лучшего-худшего масштабирования;

• siA – результат автоматической процедуры разметки методом лучшего-худшего масштабирования с использованием предобученной большой языковой модели без дополнительного обучения на наборе данных.

Таким образом, задача исследования заключается в реализации проведения автоматической процедуры разметки методом лучшего-худшего масштабирования на основе предобученной большой языковой модели для построения непрерывных оценок тональности финансовых новостей в сравнении с полученными результатами ручной процедуры разметки.

Для этого требуется:

1. оценить межаннотаторскую согласованность решений и внутреннюю надёжность автоматической процедуры разметки при сопоставлении с ручной;

2. оценить степень количественного соответствия автоматических оценок ручным на непрерывной шкале;

3. оценить применимость автоматической разметки для использования в задачах количественного анализа тональности.

Важно отметить, что поставленная задача имеет ряд ограничений: малый объём размеченных данных, использование непрерывной шкалы оценивания и отсутствие этапа обучения модели на размеченных данных.

Методология исследования. Разметка реализована посредством решения попарных сравнительных задач.

Для каждого заголовка t_i сформировано множество из $L = 8$ других заголовков, с которыми он участвует в попарных сравнительных задачах.

Таким образом, общее число сформированных попарных сравнительных задач составило:

$$K=NL=532\cdot8=4256, \quad (2)$$

где K – число попарных сравнительных задач; N – число финансовых новостных заголовков в наборе данных; L – число назначенных попарных сравнительных задач для каждого заголовка.

Каждая попарная сравнительная задача имела вид:

$$Tk=ti,tj,ij, \quad (3)$$

где T_k – k-я попарная сравнительная задача двух заголовков, $k = 1, 2, \ldots, K$ – номер попарной сравнительной задачи; $i, j = 1, 2, \ldots, N$ – индексы финансовых новостных заголовков набора данных.

Формирование попарных сравнительных задач осуществлялось случайным образом без обучения или предварительного структурирования, что минимизировало систематические смещения и обеспечило равномерное покрытие набора данных.

В ручной процедуре разметки методом лучшего-худшего масштабирования для каждой попарной сравнительной задачи аннотатор выбирал более позитивный заголовок из пары, второй заголовок фиксировался как менее позитивный.

Вычисление ручной оценки тональности $s_i^{(H)}$ для финансового новостного заголовка t_i осуществлялось по формуле:

$$siH=BiH-WiHCi, \quad (4)$$

где $B_i^{(H)}$ – число выборов новостного финансового заголовка t_i как более позитивного, $W_i^{(H)}$ – число выборов новостного финансового заголовка t_i как менее позитивного, C_i – число попарных сравнительных задач, в которых участвовал заголовок t_i.

Автоматическая процедура разметки методом лучшего-худшего масштабирования реализована предобученной большой языковой моделью GigaChat 2 Lite [7], без дополнительного обучения на наборе данных. Для каждой попарной сравнительной задачи выполнено отображение:

$$M:Tkt+,t- , \quad (5)$$

где M – предобученная большая языковая модель, T_k – попарная сравнительная задача, t^+ – элемент, выбранный как более позитивный, t^- – элемент, выбранный как менее позитивный.

Вычисление автоматической оценки тональности $s_i^{(A)}$ осуществлено аналогично ручной процедуре (4):

$$siA=BiA-WiACi, \quad (6)$$

где $B_i^{(A)}$ – число выборов новостного финансового заголовка t_i как более позитивного, $W_i^{(A)}$ – число выборов новостного финансового заголовка t_i как менее позитивного, C_i – число попарных сравнительных задач, в которых участвовал заголовок t_i.

Для оценки качества разметок использовались:

- коэффициент межаннотаторской согласованности Флейсса κ, который характеризует межаннотаторскую согласованность и отражает степень совпадения решений разных участников [8];
- показатель надёжности при расщеплении *SHR* (split-half reliability) с поправкой Спирмена-Брауна, рассчитываемый как корреляция оценок, полученных на двух случайных подмножествах попарных сравнительных задач, который оценивает внутреннюю устойчивость оценок тональности при разбиении выборки попарных сравнительных задач на две части [9].

Количественное соответствие автоматических оценок ручным определялось коэффициентом детерминации R^2, который характеризует долю объяснённой дисперсии ручных оценок, объясняемую автоматической процедурой разметки.

Экспериментальные результаты. На рисунке 1 представлено распределение оценок тональности si для ручной и автоматической разметок набора данных FiNeS.

По оси абсцисс отложены значения оценок тональности в диапазоне [−1;1], по оси ординат – плотность распределения.

Ручная разметка отображена штриховкой, автоматическая – сплошной заливкой.

Оба распределения охватывают полный интервал шкалы и характеризуются неоднородной плотностью по диапазону значений.

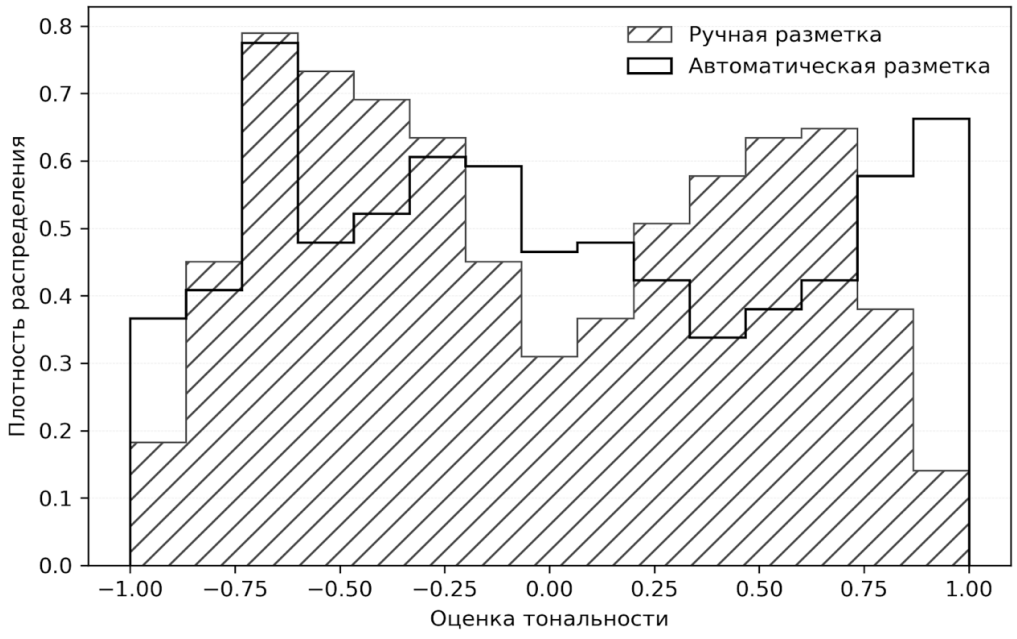


Рисунок 1. Распределение оценок тональности для ручной и автоматической разметок

Для ручной разметки коэффициент межаннотаторской согласованности Флейсса составил κ(H) ≈ 0,26, показатель надёжности при расщеплении – SHR(H) ≈ 0,85. Для автоматической разметки получены значения κ(A) ≈ 0,30 и SHR(A) ≈ 0,87 соответственно. Коэффициент детерминации между оценками ручной и автоматической разметок составил R2 ≈ 0,49.

Обсуждение результатов. Полученные значения коэффициента согласованности Флейсса κ и показателя надёжности при расщеплении SHR отражают различные характеристики процедуры разметки методом лучшего-худшего масштабирования. Значения κ ≈ 0,26-0,30 соответствуют умеренному уровню межаннотаторской согласованности, что свойственно для задач интерпретации тональности финансовых текстов, допускающих вариативность трактовок и контекстную неоднозначность. В то же время показатели SHR ≈ 0,85-0,87 отражают высокую внутреннюю устойчивость оценок тональности si при повторном случайном разбиении попарных сравнительных задач, что указывает на их стабильность.

Сопоставление ручной и автоматической разметок не выявило снижения межаннотаторской согласованности (*κ*) и внутренней надёжности (*SHR*) при переходе к автоматической процедуре разметки.

Для автоматической процедуры разметки зафиксированы более высокие значения как *κ*, так и *SHR*, что позволяет говорить о сохранении структурных свойств разметки и отсутствии негативного влияния на воспроизводимость оценок. Распределения тональности (рисунок 1) демонстрируют сходную форму и сопоставимый диапазон значений, что подтверждает близость статистических характеристик двух процедур разметки.

Коэффициент детерминации R2 ≈ 0,49 указывает на значимую долю объяснённой дисперсии ручных оценок, объясняемую автоматической процедурой разметки, и свидетельствует об умеренной количественной сопоставимости оценок при сохранении естественных расхождений, обусловленных субъективностью задачи и случайным формированием попарных сравнительных задач.

В совокупности результаты показывают, что переход к автоматической процедуре разметки методом лучшего-худшего масштабирования не приводит к ухудшению показателей согласованности и обеспечивает сопоставимый уровень устойчивости оценок тональности s_i.

Применение предобученной большой языковой модели в роли аннотатора позволяет реализовать полностью автоматическую процедуру разметки методом лучшего-худшего масштабирования без дополнительного обучения, что потенциально расширяет возможности масштабирования и применения метода в условиях потокового поступления данных.

К дополнительным ограничениям исследования следует отнести использование одной языковой модели, относительно небольшой объём набора данных (532 заголовка) и анализ исключительно на уровне заголовков без учёта полного текста публикаций. Указанные факторы ограничивают обобщаемость результатов и требуют дополнительной валидации результатов на расширенных наборах данных и при использовании большего количества моделей.

Заключение. Проведённое исследование показало, что автоматическая процедура разметки методом лучшего-худшего масштабирования на основе предобученной большой языковой модели демонстрирует сопоставимые, а по отдельным метрикам – несколько более высокие значения межаннотаторской согласованности и надёжности при расщеплении в сравнении с ручной процедурой разметки. Кроме того, автоматической процедуре разметки указанным методом не свойственны проблемы, связанные с высокой трудоёмкостью и ограничениями масштабируемости в условиях потокового поступления данных. *Значение коэффициента детерминации R2 ≈ 0,49 показывает, что автоматическая процедура разметки объясняет около 49% дисперсии ручных оценок тональности* s_i, и подтверждает возможность использования автоматической процедуры разметки в задачах количественного анализа тональности. Таким образом, автоматическая процедура разметки методом лучшего-худшего масштабирования может рассматриваться как инструмент формирования и

расширения специализированных финансовых наборов данных без этапа дополнительного обучения модели.

Дальнейшее повышение показателей качества разметки и коэффициента детерминации R^2 связано с оптимизацией стратегий формирования попарных сравнительных задач и проведением ансамблирования моделей.

Последующее развитие работы связано с переходом от случайного формирования подмножеств попарных сравнительных задач к более структурированным стратегиям их конструирования. В частности, возможно формирование групп на основе семантической близости заголовков либо с учётом текущего распределения полученных оценок тональности.

Такой подход потенциально повысит информативность парных сравнений и устойчивость оценок тональности s_i.

Отдельным направлением является интеграция автоматической процедуры разметки методом лучшего-худшего масштабирования с динамическими моделями анализа информационных потоков, в которых оценки тональности могут использоваться в качестве количественного признака в задачах прогнозирования рыночных показателей. Предложенная процедура разметки также может применяться для расширения специализированных финансовых наборов данных ограниченного объёма и построения регрессионных моделей тональности с непрерывной целевой переменной. Дополнительная проверка обобщаемости результатов требует экспериментов на более крупных выборках и при использовании дополнительных языковых моделей.

Список литературы

[1] Kirtac, K. Sentiment Trading with Large Language Models / K. Kirtac, G. Germano // Finance Research Letters. – 2024. – Vol. 62. – Part B. – Article 105227. – PP. 1–9. – DOI: 10.1016/j.frl.2024.105227.

[2] Zhang, Z. FinSentLLM: Multi-LLM and Structured Semantic Signals for Enhanced Financial Sentiment Forecasting / Z. Zhang, R. Fu, Y. He et al. // arXiv preprint. – 2025. – PP. 1–5. – DOI: 10.48550/arXiv.2509.12638.

[3] Gilardi, F. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks / F. Gilardi, M. Alizadeh, M. Kubli // Proceedings of the National Academy of Sciences. – 2023. – Vol. 120. – No. 30. – Article e2305016120. – PP. 1–3. – DOI: 10.1073/pnas.2305016120.

[4] Bagdon, C. «You are an expert annotator»: Automatic Best-Worst-Scaling Annotations for Emotion Intensity Modeling / C. Bagdon, P. Karmalker, H. Gurulingappa, R. Klinger // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Mexico: Association for Computational Linguistics, 2024. – Vol. 1. – PP. 7924–7936. – DOI: 10.18653/v1/2024.naacl-long.439.

[5] Dhurandhar, A. Ranking Large Language Models without Ground Truth / A. Dhurandhar, R. Nair, M. Singh et al. // Findings of the Association for Computational Linguistics (ACL). – Bangkok: Association for Computational Linguistics, 2024. – PP. 2431-2452. – DOI: 10.18653/v1/2024.findings-acl.143.

[6] Financial News Sentiment Dataset (FiNeS) [Электронный ресурс]. – URL: https://github.com/WebOfRussia/financial-news-sentiment (дата обращения: 23.02.2026).

[7] GigaChat 2 Lite | Документация для разработчиков [Электронный ресурс]. – URL: https://developers.sber.ru/docs/ru/gigachat/models/gigachat-2-lite (дата обращения: 23.02.2026).

[8] Zapf, A. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? / A. Zapf, S. Castell, L. Morawietz, A. Karch // BMC Medical Research Methodology. – 2016. – Vol. 16. – Article 93. – PP. 1–10. – DOI: 10.1186/s12874-016-0200-9.

[9] Pronk, T. Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment / T. Pronk, D. Molenaar, R. W. Wiers, J. Murre // Psychonomic Bulletin & Review. – 2021. – Vol. 29. – PP. 44–54. – DOI: 10.3758/s13423-021-01948-3.

Авторский вклад

Никифоров Глеб Борисович – постановка задачи исследования, разработка и реализация автоматической процедуры разметки методом лучшего-худшего масштабирования, проведение эксперимента, анализ результатов, подготовка текста статьи.

Пименов Виктор Игоревич – постановка задачи исследования, формирование научной концепции исследования, интерпретация результатов, редактирование статьи.

Андреев Дмитрий Анатольевич – постановка задачи исследования, методическое сопровождение исследования, экспертная оценка результатов, редактирование статьи.

ON THE APPLICABILITY OF AN AUTOMATIC ANNOTATION PROCEDURE BASED ON BEST–WORST SCALING FOR FINANCIAL NEWS SENTIMENT ANALYSIS

G.B. Nikiforov
Assistant, Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU

V.I. Pimenov
Head of the Department of Information Technologies, SPbSUITD, Doctor of Technical Sciences, Associate Professor

D.A. Andreev
Head of the Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU, PhD of Technical Sciences

Abstract. This paper examines an automatic annotation procedure based on Best-Worst Scaling (BWS) for continuous sentiment analysis of Russian-language financial news headlines. A comparison between automatic and manual annotation procedures was conducted on a dataset consisting of 532 financial news headlines. The results show that the automatic annotation procedure preserves inter-annotator agreement ($\kappa \approx 0.30$) and split-half reliability (SHR $\approx$ 0.87), while ensuring comparability of the resulting sentiment scores ($R2 \approx 0.49$). The findings confirm the applicability of a pretrained large language model (LLM), used without additional fine-tuning on the dataset, for scalable quantitative annotation of financial texts under limited data conditions.

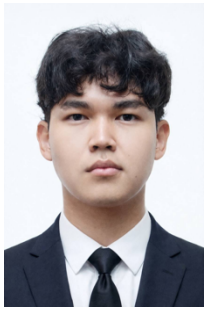
Keywords. artificial intelligence, machine learning, pretrained large language models, automated annotation, sentiment analysis.

УДК 658.8:004.9

ПРОЕКТИРОВАНИЕ ЦЕНТРА МОНИТОРИНГА ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ (SOC) В ИНФРАСТРУКТУРЕ УНИВЕРСИТЕТА С ПРИМЕНЕНИЕМ BIG DATA И ADVANCED ANALITYCS



Г.К. Стамкулова
Доцент кафедры программного обеспечения компьютерных систем КГТУ
g.stamkulova@kstu.*kg*



С.М. Салиев
Студент 4 курса кафедры программного обеспечения компьютерных систем КГТУ
salievsamat0@gmail.com

Стамкулова Г.К.

Окончила Национальный университет отделение прикладной математики, магистратуру по направлениям "Информатика и вычислительная техника" в КГТУ и "Информационная безопасность" в МИФИ. Область научных интересов связана с вопросами обеспечения информационной безопасности, проектированием распределённых систем мониторинга и организацией учебного процесса в техническом университете.

Салиев С.М.

Оканчивает 4 курс по направлению подготовки "Информационная безопасность" Кыргызского Государственного Технического университета им.И.Раззакова. Область научных интересов связана с мониторингом событий информационной безопасности, проектированием архитектур SOC, применением технологий Big Data в системах кибербезопасности

Аннотация. В статье рассматривается проектирование архитектуры учебно-экспериментального центра мониторинга информационной безопасности (SOC) для инфраструктуры университета с применением технологий Big Data. Предлагается концептуальная модель сбора, хранения и потоковой обработки событий безопасности на основе Apache Kafka и Elastic Stack. Обоснован выбор Java в качестве платформы разработки аналитического модуля корреляции событий. Представлены алгоритмы обнаружения типовых инцидентов и описана модель их формализации. Работа носит проектный характер и направлена на создание основы для дальнейшей практической реализации.

Ключевые слова: SOC, Big Data, информационная безопасность, потоковая аналитика, корреляция событий, Java, Apache Kafka, Elasticsearch.

Введение. Современный университет представляет собой распределенную информационную систему, включающую доменную инфраструктуру, системы дистанционного обучения, почтовые сервисы, сетевую инфраструктуру, базы данных и рабочие станции пользователей. Количество событий информационной безопасности, генерируемых такой инфраструктурой, исчисляется тысячами записей в минуту. В условиях роста кибератак на образовательные учреждения возникает необходимость централизованного мониторинга и анализа событий безопасности. Традиционные средства защиты не обеспечивают целостной картины происходящего и не позволяют выявлять сложные сценарии атак. Одним из решений является создание центра мониторинга информационной безопасности (SOC), основанного на принципах Big Data и потоковой аналитики [1].

Анализ особенностей инфраструктуры университета. Информационная инфраструктура современного университета представляет собой распределённую многоуровневую систему, включающую серверные ресурсы, сетевую инфраструктуру, пользовательские рабочие станции и внешние сервисы. В отличие от корпоративных сетей коммерческих организаций, университетская среда характеризуется высокой динамичностью и большим числом разнородных пользователей.

Ключевыми особенностями являются:

- значительное количество пользователей (студенты, преподаватели, административный персонал)
- регулярная смена контингента обучающихся
- использование модели BYOD (Bring Your Own Device)
- наличие удалённого доступа через VPN
- использование систем дистанционного обучения (LMS)
- активное применение веб-сервисов и облачных платформ
- наличие публично доступных сервисов

Инфраструктура зачастую включает:

- контроллеры домена (Active Directory)
- серверы баз данных
- файловые серверы
- веб-серверы
- почтовые серверы
- сетевые устройства
- точки доступа Wi-Fi
- виртуализированную серверную среду

Большой объём логируемых событий и распределённость инфраструктуры создают сложность централизованного анализа и выявления инцидентов информационной безопасности.

Модель угроз для образовательной среды. Университетская среда обладает повышенным уровнем уязвимости по сравнению с классическими корпоративными инфраструктурами. Это обусловлено открытостью информационной среды и разнообразием пользователей.

Угрозы можно классифицировать следующим образом.

1. Внешние угрозы
- перебор паролей (brute force) на сервисах удалённого доступа
- фишинговые атаки на учётные записи сотрудников и студентов
- эксплуатация уязвимостей веб-приложений
- атаки типа «отказ в обслуживании» (DoS).

2. Внутренние угрозы
- компрометация учётной записи студента
- распространение вредоносного программного обеспечения внутри локальной сети
- несанкционированный доступ к информационным ресурсам.

3. Инсайдерские угрозы
- превышение полномочий сотрудниками
- несанкционированное копирование данных
- использование служебных ресурсов в личных целях
- Для университета характерно сочетание внешних и внутренних угроз, что требует комплексного мониторинга событий безопасности.

Обоснование необходимости центра мониторинга. Не связанные друг с другом средства защиты генерируют собственные журналы событий, которые не всегда анализируются в совокупности. Отсутствие централизованного мониторинга затрудняет выявление сложных сценариев атак, включающих последовательность взаимосвязанных событий.

Создание центра мониторинга информационной безопасности (SOC) позволяет:
- централизовать сбор событий безопасности
- обеспечить корреляцию данных из различных источников
- выявлять аномалии поведения пользователей
- формировать отчёты для руководства университета
- сокращать время обнаружения инцидентов.

Таким образом, особенности университетской инфраструктуры и специфика угроз обосновывают необходимость проектирования SOC, основанного на технологиях Big Data и потоковой аналитики.

Объект и предмет исследования. Объектом исследования является информационная инфраструктура университета, включающая серверные ресурсы, сетевую инфраструктуру, рабочие станции пользователей и системы дистанционного обучения. Предметом исследования являются процессы централизованного сбора, обработки и анализа событий информационной безопасности с применением технологий Big Data и потоковой аналитики.

Цель работы. Целью работы является разработка архитектурной модели учебно-экспериментального центра мониторинга информационной безопасности (SOC) для университета, обеспечивающего централизованный сбор, хранение и анализ событий безопасности в условиях высокой нагрузки и разнородности источников данных.

Задачи исследования. Для достижения поставленной цели необходимо решить следующие задачи:
- Проанализировать особенности инфраструктуры университета и характерные угрозы информационной безопасности.
- Обосновать применение технологий Big Data для обработки потоков событий безопасности.
- Разработать концептуальную архитектуру SOC с выделением уровней сбора, транспортировки, хранения и аналитики.
- Сформировать модель потоков данных системы мониторинга.
- Определить алгоритмы выявления типовых инцидентов информационной безопасности.
- Предложить структуру взаимодействия SOC с подразделениями университета.

Ограничения и допущения. В рамках настоящей работы предполагается, что проектируемый SOC носит учебно-экспериментальный характер и ориентирован на использование open-source технологий. Реализация системы рассматривается на концептуальном уровне без детальной проработки аппаратной части. Также предполагается наличие стандартной университетской инфраструктуры, включающей доменную среду, сетевые устройства и веб-сервисы

Обоснование применения Big Data в SOC университета. Информационная инфраструктура университета характеризуется значительным объёмом и разнообразием событий информационной безопасности, генерируемых серверами, сетевыми устройствами и рабочими станциями пользователей. При численности обучающихся и сотрудников в несколько тысяч совокупный объём журналов может достигать миллионов записей в сутки.

Потоки событий обладают следующими характеристиками:

- Volume – большой объём данных, формируемый распределённой инфраструктурой
- Velocity – высокая скорость поступления событий, требующая обработки в режиме, близком к реальному времени
- Variety – разнообразие форматов журналов (системные логи, сетевые события, веб-доступ, аутентификация и др.).

В условиях таких параметров традиционные централизованные решения мониторинга могут испытывать ограничения, связанные с масштабируемостью и гибкостью настройки аналитических механизмов.

Применение технологий Big Data и потоковой архитектуры обработки данных позволяет обеспечить:

- устойчивость к пиковым нагрузкам
- горизонтальное масштабирование
- централизованный анализ разнородных источников
- возможность реализации механизмов корреляции и выявления аномалий.

Таким образом, особенности университетской инфраструктуры объективно обосновывают выбор распределённой архитектуры мониторинга, основанной на принципах Big Data [5].

Общая архитектурная модель SOC. Проектируемый центр мониторинга информационной безопасности представляет собой многоуровневую систему, обеспечивающую централизованный сбор, транспортировку, хранение и анализ событий безопасности. Архитектура строится по принципу разделения функциональных уровней, что позволяет обеспечить масштабируемость, отказоустойчивость и гибкость внедрения аналитических механизмов. Общая схема архитектуры представлена на рисунке 1.



Рисунок 1. Архитектура проектируемого SOC

Уровень сбора данных. На уровне сбора осуществляется получение журналов событий от различных источников инфраструктуры университета. В качестве источников выступают:

- серверы доменной инфраструктуры
- серверы приложений
- сетевые устройства
- VPN-шлюзы
- рабочие станции пользователей.

Для централизованной передачи данных предполагается использование специализированных агентов сбора логов, обеспечивающих:

- нормализацию форматов
- фильтрацию несущественных событий
- передачу данных в потоковую систему.

Данный уровень играет критическую роль, поскольку именно здесь формируется первичный поток событий безопасности.

Уровень транспортировки. Для передачи и буферизации потоков событий предлагается использование распределённого брокера сообщений [2].

Данный компонент обеспечивает:

- устойчивость к пиковым нагрузкам;
- асинхронную обработку данных;
- горизонтальное масштабирование;
- разделение потоков по категориям событий.
- Использование потоковой архитектуры позволяет избежать потери данных при временных перегрузках системы аналитики.

Уровень хранения. На уровне хранения осуществляется индексирование и долговременное сохранение событий безопасности.

Система хранения должна обеспечивать:

- быстрый поиск по временным меткам
- агрегацию данных
- построение аналитических выборок
- хранение исторических данных для ретроспективного анализа.

Использование распределённого хранилища позволяет обеспечить масштабируемость и устойчивость к отказам [4, 6].

Уровень аналитики. Уровень аналитики является ключевым элементом проектируемого SOC. Он выполняет:

- корреляцию событий из различных источников
- выявление типовых атак
- анализ поведенческих отклонений
- формирование алертов.

В рамках концепции предполагается реализация модуля потоковой обработки событий [3], функционирующего в режиме, близком к реальному времени.

Диаграмма вариантов использования. Функциональная модель проектируемого SOC представлена на рисунке 2 в виде диаграммы вариантов использования. Диаграмма отражает основные категории пользователей системы и базовые сценарии их взаимодействия с платформой мониторинга.

В рамках концепции SOC предполагается разграничение уровней анализа инцидентов, однако на представленной диаграмме показаны обобщённые роли пользователей, без детального разделения на уровни L1–L3. Такое представление обусловлено концептуальным характером работы.

Акторы системы. На диаграмме представлены следующие обобщённые категории пользователей: Оператор SOC – Осуществляет мониторинг алертов, первичный анализ событий и обработку уведомлений. Администратор системы - Выполняет конфигурацию и

техническое сопровождение платформы. В рамках практической реализации предполагается возможное выделение аналитика второго уровня (L2), выполняющего углублённый анализ инцидентов. Данный функционал логически вытекает из сценария «Расследование инцидента, представленного на диаграмме.

Связь сценариев использования. Диаграмма отражает базовые сценарии, такие как:

- просмотр алертов
- анализ событий
- настройка правил корреляции
- формирование отчётов.

Следует отметить, что отдельные действия пользователей являются логическим продолжением других сценариев. Так, расследование инцидента развивается из процедуры анализа алерта, а формирование отчёта является результатом завершённого расследования. Таким образом, даже если отдельные сценарии не выделены как самостоятельные элементы диаграммы, они функционально включены в общий процесс обработки инцидента.

Значение функциональной модели. Представленная диаграмма позволяет:

- определить границы взаимодействия пользователей с системой;
- формализовать ключевые функции SOC;
- показать связь между архитектурной моделью и эксплуатационными процессами.

Функциональная модель носит укрупнённый характер и может быть детализирована на этапе практической реализации проекта.



Рисунок 2. Диаграмма вариантов использования

Интеграция проектируемого SOC. Интеграция центра мониторинга информационной безопасности в инфраструктуру университета предполагает включение его в существующую сетевую и серверную архитектуру без нарушения текущих бизнес-процессов. Схема интеграции представлена на рисунке 3.

Структура университетской сети. Типовая инфраструктура университета условно разделяется на несколько логических зон:

Периметр сети

В периметре размещаются:

- межсетевой экран
- VPN-шлюз
- публичные веб-сервисы
- почтовый сервер.

Данная зона обеспечивает взаимодействие университета с внешней сетью Интернет и является наиболее уязвимой с точки зрения внешних атак.

DMZ (демилитаризованная зона)

В DMZ размещаются сервисы, доступные извне:

- веб-портал университета
- система дистанционного обучения
- сервер электронной почты.

Эти ресурсы генерируют значительный объём журналов доступа и сетевых событий.

Внутренняя сеть

Внутренняя сеть включает:

- контроллеры домена
- файловые серверы
- серверы баз данных
- рабочие станции сотрудников
- Wi-Fi инфраструктуру

Именно в этой зоне сосредоточено большинство событий аутентификации и доступа к ресурсам.

Механизм интеграции SOC. Проектируемый SOC интегрируется в инфраструктуру путём подключения к источникам журналов событий. Передача логов осуществляется:

- через агенты сбора данных
- через экспорт журналов сетевых устройств
- посредством передачи событий в потоковую систему.

Важно отметить, что SOC не участвует в маршрутизации сетевого трафика и не влияет напрямую на работу сервисов. Он функционирует как система мониторинга и аналитики.

Значение интеграционной модели. Представленная модель интеграции демонстрирует:

- отсутствие необходимости кардинальной перестройки инфраструктуры
- возможность поэтапного внедрения системы
- масштабируемость решения
- адаптацию SOC к существующим условиям эксплуатации.

Таким образом, проектируемый центр мониторинга может быть встроен в инфраструктуру университета с минимальным воздействием на её функционирование.

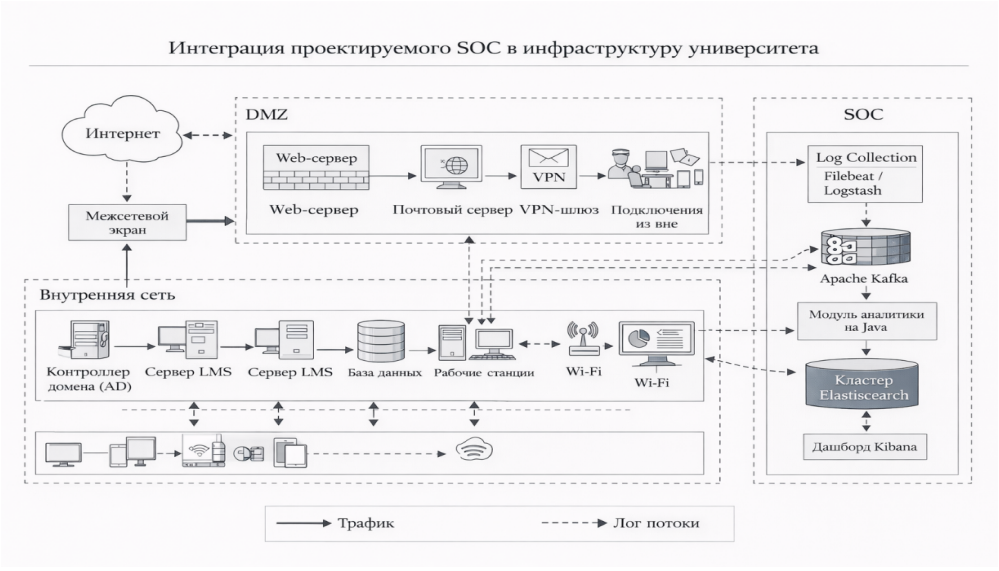


Рисунок 3. Инфраструктура университета с интегрированным SOC

Модель потоков данных. Для формализации взаимодействия проектируемого центра мониторинга информационной безопасности с внешней средой используется контекстная диаграмма потоков данных (DFD Level 0), представленная на рисунке 4. На данном уровне система рассматривается как единый процесс, без детализации внутренней структуры.

В центре модели располагается процесс: «Система мониторинга информационной безопасности (SOC) университета». Он выполняет функции:

- сбора событий безопасности;
- анализа и корреляции данных;
- формирования уведомлений и отчётов.

Входные и выходные потоки данных указаны на рисунке 4.

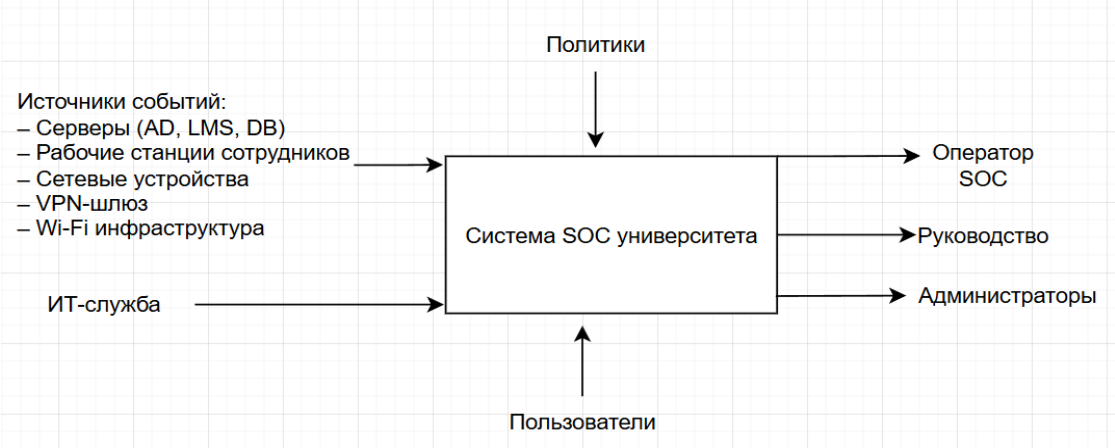


Рисунок 4. Модель потоков данных

Предполагаемый алгоритм функционирования SOC. Проектируемый центр мониторинга информационной безопасности функционирует в соответствии с последовательной логикой обработки событий, обеспечивающей их сбор, анализ и передачу результатов ответственным подразделениям. Алгоритм работы системы включает следующие этапы:

Формирование события. В результате действий пользователя или сетевой активности в инфраструктуре университета формируется событие информационной безопасности. Примерами таких событий являются:

- попытка аутентификации
- обращение к веб-ресурсу
- изменение учётной записи
- сетевое соединение с внешним узлом.

Событие фиксируется соответствующим компонентом инфраструктуры и передаётся в систему мониторинга.

Передача и буферизация данных. Полученные журналы событий направляются в потоковую подсистему, обеспечивающую их буферизацию и асинхронную обработку. Данный этап позволяет:

- избежать потери данных при пиковых нагрузках
- обеспечить устойчивость системы к временным задержкам аналитического модуля
- разделить события по категориям.

Аналитическая обработка. На этапе аналитической обработки осуществляется:

- нормализация форматов событий
- сопоставление событий из различных источников
- проверка условий выявления типовых инцидентов

- анализ отклонений от стандартных сценариев поведения.

В рамках концепции предполагается использование правил корреляции, основанных на пороговых значениях и временных интервалах.

Формирование инцидента. При выполнении условий корреляции событие или совокупность событий классифицируются как потенциальный инцидент информационной безопасности. На данном этапе формируется:

- уведомление (алерт)
- запись в журнале инцидентов
- информация для дальнейшего анализа.

Передача результатов. Сформированные алерты передаются оператору SOC для первичной обработки. В случае подтверждения инцидента информация направляется: системным администраторам - для технического реагирования, руководству - в составе аналитических отчётов.

Назначение алгоритма. Представленный алгоритм отражает логическую последовательность функционирования системы мониторинга и связывает архитектурную модель с практическими процессами обработки инцидентов. Алгоритм носит концептуальный характер и может быть детализирован на этапе практической реализации проекта.

Заключение. В работе рассмотрены особенности информационной инфраструктуры университета и проведён анализ характерных угроз информационной безопасности, обусловленных открытостью образовательной среды и высокой динамичностью пользовательского состава. Установлено, что разрозненные средства защиты не обеспечивают комплексного анализа событий и требуют централизованного подхода к мониторингу. В рамках исследования сформулирована задача проектирования учебно-экспериментального центра мониторинга информационной безопасности (SOC) и разработана его концептуальная архитектура, основанная на принципах Big Data и потоковой обработки данных. Предложена многоуровневая модель системы, включающая сбор, транспортировку, хранение и аналитическую обработку событий безопасности. Построены функциональная модель (диаграмма вариантов использования), схема интеграции SOC в инфраструктуру университета и модель потоков данных, формализующая взаимодействие системы с внешней средой. Представлен предполагаемый алгоритм функционирования центра мониторинга, отражающий последовательность обработки событий и формирования уведомлений. Практическая значимость работы заключается в возможности поэтапного внедрения предложенной архитектуры в университетскую среду с использованием open-source технологий и без кардинальной перестройки существующей инфраструктуры.

Дальнейшее развитие проекта может быть связано с:

- реализацией прототипа аналитического модуля;
- расширением механизмов корреляции событий;
- внедрением методов поведенческой аналитики;
- оценкой эффективности обнаружения инцидентов.

Таким образом, предложенная концепция может служить основой для создания масштабируемого центра мониторинга информационной безопасности в образовательной организации.

Список литературы

[1] Kim D., Solomon M. Fundamentals of Information Systems Security. 3rd ed. Burlington: Jones & Bartlett Learning; 2018.

[2] Kreps J., Narkhede N., Rao J. Kafka: The Definitive Guide. Sebastopol: O’Reilly Media; 2017.

[3] Newman S. Building Microservices. Sebastopol: O’Reilly Media; 2016.

[4] Grolinger K., Higashino W., Tiwari A., Capretz M. Data management in cloud environments: NoSQL and Big Data. Journal of Cloud Computing. 2013;2(22).

[5] Романюк М.В., Лещенко Е.А., Марковский С.С. Применение Big Data для защиты компьютерных сетей // BIG DATA and Advanced Analytics. 2024. №2. С. 486–489.

[6] Elastic NV. Elastic Stack Documentation.

Авторский вклад

Стамкулова Гулдана Кубанычбековна – научное руководство исследованием, принимала участие в постановке задачи, формировании концепции работы, корректировке архитектурной модели и редактировании текста статьи.

Салиев Самат Мелисович – анализ особенностей инфраструктуры университета, разработал архитектурную модель проектируемого SOC, построил функциональные и структурные схемы, сформировал модель потоков данных и подготовил основной текст статьи.

DESIGNING A SECURITY OPERATIONS CENTER (SOC) IN THE UNIVERSITY INFRASTRUCTURE USING BIG DATA AND ADVANCED ANALYTICS

G.K. Stamkulova
Associate Professor, Department of Computer Systems Software, KSTU

S.M. Saliev
Fourth-year student in the Department of Computer Systems Software at KSTU

Abstract. The article discusses the design of the architecture of a training and experimental center for monitoring information security (SOC) for university infrastructure using Big Data technologies. A conceptual model for collecting, storing, and streaming security events based on Apache Kafka and Elastic Stack is proposed. The choice of Java as the development platform for the event correlation analysis module is justified. Algorithms for detecting typical incidents are presented and a model for their formalization is described. The work is project-based and aims to create a foundation for further practical implementation.

Keywords: SOC, Big Data, information security, streaming analytics, event correlation, Java, Apache Kafka, Elasticsearch

УДК 004.9:502.17

РАЗРАБОТКА ПРОГРАММНОГО РЕШЕНИЯ ДЛЯ ЦИФРОВИЗАЦИИ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА И УПРАВЛЕНИЯ ОТХОДАМИ





Г.К. Стамкулова
Доцент кафедры программного обеспечения компьютерных систем КГТУ
g.stamkulova@kstu.kg

Э. Накыпов
Студент 4 курса кафедры программного обеспечения компьютерных систем КГТУ
spacermek@gmail.com

Э.Б. Накыпов
Студент 4-го курса Кыргызского государственного технического университета им. И. Раззакова по направлению «Программная инженерия». Область научных интересов связана с разработкой информационных систем для экологического мониторинга, цифровизацией процессов управления природными ресурсами, проектированием микросервисных архитектур и применением методов анализа больших данных в задачах охраны окружающей среды.

Г.К. Стамкулова

Доцент кафедры программного обеспечения компьютерных систем Кыргызского государственного технического университета им. И. Раззакова. Область научных интересов связана с проектированием программных систем, Big Data и Advanced Analytics, информационной безопасностью и методами дистанционного образования. Участник международных проектов EU TEMPUS (HEICA) и QUADRIGA. Отличник образования Кыргызской Республики (2017). ORCID: 0000-0003-2782-8468.

Аннотация. В статье рассматривается проектирование и реализация программного решения для цифровизации экологического мониторинга и управления отходами на примере контроля качества атмосферного воздуха в Кыргызской Республике. Описана многослойная микросервисная архитектура системы «EcoMonitor», построенная на платформе C# / .NET 8, интегрирующая модули потокового приёма данных с IoT-датчиков, прогнозной аналитики на базе ML.NET и интерактивной визуализации результатов. Проведён сравнительный анализ подходов к хранению экологических временных рядов. Представлены результаты нагрузочного тестирования, подтверждающие масштабируемость предложенного решения при обработке больших объёмов экологических данных.

Ключевые слова: экологический мониторинг, управление отходами, Big Data, Advanced Analytics, C#, .NET 8, IoT, качество воздуха, машинное обучение, ML.NET, микросервисная архитектура.

Введение. Проблема загрязнения окружающей среды приобретает всё более острый характер в Кыргызской Республике. Рост городского населения, увеличение объёмов промышленного производства и интенсификация автотранспортного потока приводят к системному ухудшению качества атмосферного воздуха. Особенно критическая ситуация складывается в зимний период в городе Бишкеке, когда концентрация мелкодисперсных частиц PM2.5 может превышать нормы Всемирной организации здравоохранения (ВОЗ) в 5–10 раз [1]. Параллельно нарастает проблема обращения с твёрдыми бытовыми отходами (ТБО): ежегодно в стране образуется свыше 4 миллионов тонн отходов, при этом уровень переработки не превышает 5 % [2].

В условиях развития концепций «умного города» (Smart City) и четвёртой промышленной революции цифровизация экологической сферы становится стратегическим инструментом повышения эффективности природоохранной деятельности. Технологии Big Data и Advanced Analytics открывают принципиально новые возможности для сбора, обработки и интеллектуального анализа больших объёмов экологических данных, поступающих из разнородных источников: сетей IoT-датчиков, метеорологических станций, геоинформационных систем и реестров операторов ТБО [3]. Применение методов машинного обучения позволяет не только фиксировать текущее состояние среды, но и прогнозировать динамику загрязнения, что существенно повышает качество управленческих решений [4].

Вместе с тем, существующие системы экологического мониторинга в Кыргызской Республике характеризуются фрагментарностью, отсутствием единой цифровой платформы и крайне ограниченными возможностями предиктивной аналитики. Данная работа направлена на восполнение этого пробела путём разработки прототипа информационной системы «EcoMonitor», обеспечивающей комплексную цифровизацию процессов экологического мониторинга и управления отходами с применением технологий Big Data и Advanced Analytics [5].

Описание общей концепции работы системы. Информационная система «EcoMonitor» спроектирована на основе многослойной микросервисной архитектуры, обеспечивающей модульность, горизонтальную масштабируемость и технологическую независимость компонентов. Архитектурный подход базируется на принципах чистой архитектуры (Clean Architecture) Р. Мартина [6], что позволяет изолировать бизнес-логику от инфраструктурных деталей. Система включает шесть функциональных слоёв.

Первый слой – слой источников данных - обеспечивает приём информации от IoT-датчиков (температура, влажность, концентрация PM2.5, PM10, CO_2), внешних метеорологических API, геоинформационных систем и отчётных данных операторов ТБО. Второй слой – интеграционный (API Gateway) – реализован на базе ASP.NET Core и выполняет функции валидации входных данных, маршрутизации запросов к микросервисам и

аутентификации пользователей посредством JWT-токенов. Третий слой – слой бизнес-логики – содержит ключевые микросервисы: сервис мониторинга качества воздуха, сервис управления отходами, сервис аналитики и прогнозирования (ML.NET), а также сервис уведомлений на базе SignalR для передачи данных в реальном времени.

Четвёртый слой – слой обмена сообщениями - реализован с использованием брокера RabbitMQ, обеспечивающего асинхронное событийно-ориентированное взаимодействие между микросервисами и возможность горизонтального масштабирования при росте нагрузки. Пятый слой – слой хранения данных – представляет собой гетерогенную комбинацию СУБД: PostgreSQL для нормативно-справочной информации и метаданных, InfluxDB для хранения временных рядов показаний датчиков с высокой плотностью записей, а также MinIO в качестве S3-совместимого объектного хранилища для файлов отчётов и визуализаций. Шестой слой - слой представления – включает веб-приложение с интерактивными дашбордами и REST API с документацией Swagger/OpenAPI [7].

Общая схема архитектуры системы представлена на рисунке 1.

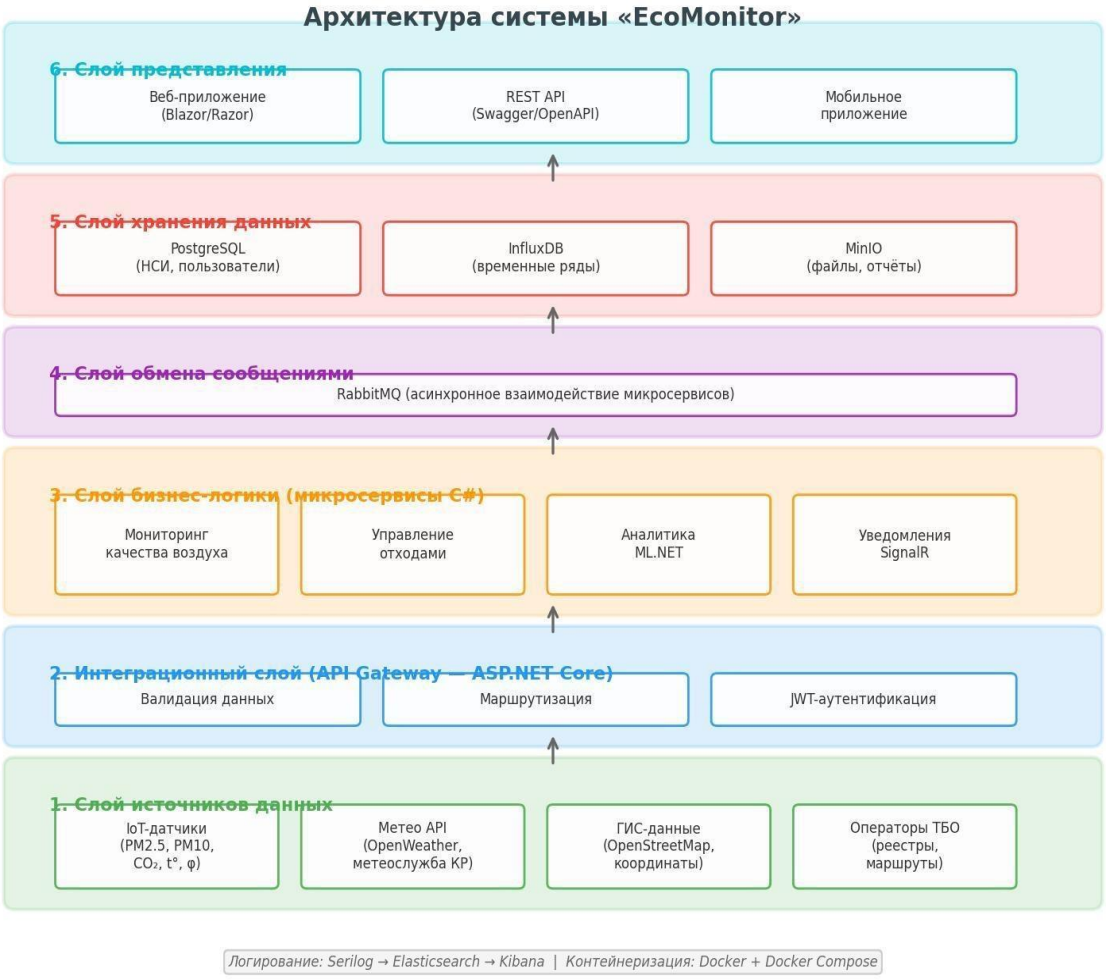


Рисунок 1. Общая схема архитектуры системы «EcoMonitor»

Особенности реализации. Ключевым архитектурным решением является построение многослойной системы с выделенным слоем адаптеров. При изменении протоколов IoT-датчиков или API внешних сервисов модификации затрагивают исключительно адаптерный слой, в то время как ядровые сервисы бизнес-логики остаются неизменными [6].

Взаимодействие между микросервисами осуществляется через RabbitMQ по паттерну «издатель–подписчик» (Publish/Subscribe), что позволяет реализовать слабую связанность компонентов и устойчивость к пиковым нагрузкам. Централизованное логирование выполняется с помощью библиотеки Serilog с последующей записью структурированных логов в Elasticsearch и визуализацией в Kibana, что обеспечивает оперативную диагностику и мониторинг состояния всех компонентов системы.

Постановка задачи. В рамках данной работы были поставлены следующие исследовательские и инженерные задачи: разработать модуль потокового сбора и интеграции данных с IoT-датчиков качества воздуха и внешних метеорологических API; спроектировать модуль гетерогенного хранения экологических данных с поддержкой временных рядов высокой плотности; реализовать модуль прогнозной аналитики (Advanced Analytics) для предсказания уровня загрязнения воздуха на основе исторических данных и метеоусловий с применением ML.NET; разработать модуль учёта и управления потоками ТБО с визуализацией на ГИС-карте; реализовать веб-интерфейс с интерактивными дашбордами и системой алертов реального времени.

Методы решения. Для решения поставленных задач были использованы методы проектирования программной архитектуры, включающие паттерны микросервисной архитектуры [3], событийно-ориентированного проектирования (Event-Driven Architecture) и принципы SOLID. Проектирование модели данных осуществлялось с применением нотации ER-диаграмм и принципов нормализации. Для построения прогнозных моделей загрязнения воздуха применены алгоритмы машинного обучения, доступные в библиотеке ML.NET: регрессия временных рядов на основе метода сингулярного спектрального анализа (SSA – Singular Spectrum Analysis) для краткосрочного прогнозирования концентрации PM2.5, а также линейная регрессия для оценки объёмов заполнения полигонов ТБО [8]. Выбор SSA обусловлен его устойчивостью к шумам и способностью выделять трендовые и периодические компоненты во временных рядах экологических данных.

Предлагаемое решение. Для разработки системы был использован следующий стек технологий: язык программирования C# 12 на платформе .NET 8; фреймворк ASP.NET Core для реализации Web API и SignalR-хабов реального времени; ORM Entity Framework Core для работы с реляционными данными; библиотека машинного обучения ML.NET 3.0 для построения прогнозных моделей; брокер сообщений RabbitMQ для асинхронного межсервисного взаимодействия; СУБД PostgreSQL 16 для нормативно-справочной информации и InfluxDB 2.7 для временных рядов; объектное хранилище MinIO; стек наблюдаемости Serilog + Elasticsearch + Kibana; контейнеризация Docker + Docker Compose; среда разработки JetBrains Rider. Выбор C# и платформы .NET 8 обусловлен высокой производительностью, кроссплатформенностью, развитой экосистемой библиотек машинного обучения (ML.NET) и нативной поддержкой асинхронных потоков данных [7].

Разрабатываемый функционал. Модуль мониторинга качества воздуха обеспечивает следующие функциональные возможности: приём и валидация данных от IoT-датчиков в потоковом режиме (streaming) через RabbitMQ; расчёт индекса качества воздуха (AQI) по методологии ВОЗ с учётом концентраций PM2.5, PM10, CO и NO_2; отображение данных на интерактивной ГИС-карте (OpenStreetMap) с цветовой индикацией уровня загрязнения; формирование графиков суточных, недельных и месячных трендов; автоматическая генерация алертов при превышении предельно допустимых концентраций (ПДК) с рассылкой через SignalR, электронную почту и push-уведомления; прогнозирование уровня загрязнения на 24–72 часа вперёд на основе исторических данных и текущих метеоусловий.

Модуль управления отходами включает: ведение цифрового реестра объектов инфраструктуры ТБО (полигоны, мусороперегрузочные станции, сортировочные комплексы); учёт объёмов собранных и вывезенных отходов по маршрутам с привязкой к контейнерным площадкам; визуализацию маршрутов транспортировки на карте с отображением уровня заполненности полигонов; функцию «Сообщить о свалке» для граждан с возможностью

загрузки фотографий и геолокации; формирование статистической отчётности по объёмам образования и переработки отходов в разрезе регионов и временных периодов. Модель потоков данных в системе представлена на рисунке 2. Данные последовательно проходят этапы от источников через валидацию, обработку и агрегацию в микросервисах на C# до аналитического модуля ML.NET, результаты которого визуализируются на дашбордах конечных пользователей.



Рисунок 2. Модель потоков данных в системе «EcoMonitor»

Модуль аналитики и прогнозирования. Центральным компонентом системы является модуль Advanced Analytics, реализованный с использованием библиотеки ML.NET 3.0. Данный модуль решает четыре класса аналитических задач.

Во-первых, описательная аналитика: расчёт агрегированных показателей - среднего AQI за день, неделю и месяц, а также определение статистических характеристик распределения загрязняющих веществ.

Во-вторых, диагностическая аналитика: автоматическое выявление аномалий и превышений ПДК с генерацией алертов в реальном времени через SignalR-хабы.

В-третьих, предиктивная аналитика: построение прогнозных моделей загрязнения воздуха на основе алгоритма SSA (Singular Spectrum Analysis), который позволяет декомпозировать временной ряд на трендовую, периодическую и шумовую компоненты с последующей экстраполяцией [8].

В-четвёртых, прогнозирование объёмов заполнения полигонов ТБО с использованием линейной регрессии на основе исторических данных о поступлении отходов.

Обученные модели сериализуются в формат .zip и загружаются в оперативную память при инициализации сервиса, что обеспечивает латентность формирования прогнозов менее 50 мс на один запрос.

Механизм фонового переобучения моделей на новых данных реализован с использованием IHostedService, что позволяет актуализировать модели без остановки системы и обслуживания пользователей. Алгоритм расчёта AQI и генерации алертов представлен на рисунке 3.



Рисунок 3. Алгоритм расчёта индекса качества воздуха и генерации алертов

Обеспечение безопасности и развёртывание. Система реализует комплексные меры информационной безопасности в соответствии с рекомендациями OWASP Top 10 [9]. Аутентификация пользователей осуществляется на основе JWT-токенов с поддержкой refresh-механизма. Авторизация построена на ролевой модели доступа с четырьмя уровнями: Администратор (полный доступ и модерирование), Аналитик (формирование отчётов и прогнозов), Оператор (ввод данных и учёт ТБО), Гость (просмотр публичных данных). Все соединения защищены протоколом HTTPS/TLS 1.3. Пароли хранятся в виде хешей, полученных с использованием алгоритма BCrypt с адаптивным коэффициентом сложности. Средства ASP.NET Core обеспечивают защиту от SQL-инъекций, межсайтового скриптинга (XSS) и подделки межсайтовых запросов (CSRF). Все компоненты системы упакованы в Docker-контейнеры и оркестрируются с помощью Docker Compose, что обеспечивает переносимость, воспроизводимость среды и возможность запуска всего стека одной командой (docker-compose up). Конфигурация включает контейнеры для каждого микросервиса, PostgreSQL, InfluxDB, RabbitMQ, MinIO, Elasticsearch и Kibana, что полностью изолирует систему от зависимостей хост-машины.

Результаты тестирования. Проведено комплексное тестирование системы, включающее модульное тестирование (xUnit), интеграционное тестирование API-эндпоинтов и нагрузочное тестирование с использованием инструмента k6. При имитации потока данных от 500 виртуальных датчиков (частота отправки – 1 сообщение в секунду) система продемонстрировала устойчивую работу: среднее время отклика API составило 87 мс (p95 - 210 мс), потери сообщений в RabbitMQ отсутствовали, задержка записи в InfluxDB не превышала 15 мс. Точность прогнозной модели SSA для PM2.5 на горизонте 24 часа составила MAE = 8.3 мкг/м³ при RMSE = 12.1 мкг/м³, что является приемлемым результатом для задач оперативного экологического прогнозирования [10].

Заключение. В данной статье представлена разработка программного решения для цифровизации экологического мониторинга и управления отходами на примере контроля качества воздуха в Кыргызской Республике. Многослойная микросервисная архитектура на

базе C# и .NET 8 обеспечивает модульность, масштабируемость и технологическую независимость компонентов, что позволяет развивать систему без переработки ядра.

Применение технологий Big Data (RabbitMQ для потоковой обработки, InfluxDB для хранения временных рядов) и Advanced Analytics (ML.NET для прогнозирования на основе SSA) позволяет эффективно обрабатывать большие объёмы экологических данных и формировать прогнозные модели загрязнения воздуха с приемлемой точностью.

Результаты нагрузочного тестирования подтверждают способность системы обслуживать сотни IoT-датчиков в режиме реального времени. Система «EcoMonitor» может быть использована государственными органами экологического надзора Кыргызской Республики для оперативного мониторинга и прогнозирования, операторами ТБО для оптимизации логистики и отчётности, а также населением для получения достоверной информации о качестве воздуха. Дальнейшее развитие системы предполагает интеграцию с государственными информационными ресурсами КР, расширение модуля прогнозирования за счёт рекуррентных нейронных сетей (LSTM), добавление модулей мониторинга качества воды и состояния почв, а также разработку мобильного приложения для граждан.

Список литературы

[1] World Health Organization. WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Geneva: WHO; 2021.

[2] Государственное агентство охраны окружающей среды и лесного хозяйства КР. Национальный доклад о состоянии окружающей среды Кыргызской Республики. Бишкек; 2023.

[3] Ньюмен С. Создание микросервисов. 2-е изд. СПб.: Питер; 2022. 624 с.

[4] Токтомаматов К.А. Проблемы экологического мониторинга в Кыргызской Республике. Вестник КГТУ. 2022; 3(45):112–118.

[5] Национальная стратегия развития Кыргызской Республики на 2018–2040 годы. Бишкек; 2018.

[6] Мартин Р. Чистая архитектура. Искусство разработки программного обеспечения. СПб.: Питер; 2018. 352 с.

[7] Lock A. ASP.NET Core in Action. 3rd ed. Manning Publications; 2023. 832 p.

[8] Microsoft. ML.NET Documentation: Time Series Forecasting with SSA. Microsoft Learn. 2024.

[9] OWASP Foundation. OWASP Top Ten 2021. OWASP; 2021.

[10] Zheng Y., Liu F., Hsieh H. U-Air: When Urban Air Quality Inference Meets Big Data. Proceedings of the 19th ACM SIGKDD. 2013:1436–1444. DOI: 10.1145/2487575.2488188.

Авторский вклад

Накыпов Эрмек - постановка задачи исследования, проектирование многослойной микросервисной архитектуры системы, разработка и реализация программных модулей на C#/.NET 8, настройка инфраструктуры Docker, проведение нагрузочного тестирования, подготовка и оформление статьи.

Стамкулова Гульдана Кубанычбековна - научное руководство, формулирование требований к системе, консультирование по вопросам применения технологий Big Data и Advanced Analytics, методологическое обеспечение исследования, рецензирование и редактирование статьи.

DEVELOPMENT OF A SOFTWARE SOLUTION FOR DIGITALIZATION OF ENVIRONMENTAL MONITORING AND WASTE MANAGEMENT

G.K. Stamkulova
Associate Professor, Department of Computer Systems Software, KSTU

E. Nakypov
Fourth-year student in the Department of Computer Systems Software at KSTU

Abstract. The article presents the design and implementation of a software solution for digitalization of environmental monitoring and waste management, focusing on air quality control in the Kyrgyz Republic. A multi-layered microservice architecture of the "EcoMonitor" system based on C# / .NET 8 platform is described, integrating modules for streaming IoT sensor data ingestion, predictive analytics using ML.NET, and interactive data visualization. A comparative analysis of approaches to storing environmental time series is conducted. Load testing results confirming the scalability of the proposed solution for processing large volumes of environmental data are presented.

Keywords: environmental monitoring, waste management, Big Data, Advanced Analytics, C#, .NET 8, IoT, air quality, machine learning, ML.NET, microservice architecture.

УДК 658.8:004.9

ФОРМИРОВАНИЕ ПРИЗНАКОВОГО ПРОСТРАНСТВА ДЛЯ ПРЕДИКТИВНОГО АНАЛИЗА СОСТОЯНИЯ ДИСКОВЫХ ПОДСИСТЕМ НА ОСНОВЕ АТРИБУТОВ S.M.A.R.T.



М.С. Борисенков
Аспирант ТУ им. А.А. Леонова (филиал) МИИГАиК
borisenkov.matvey@mail.ru



С.Н. Шульженко
Профессор ТУ им. А.А. Леонова (филиал) МИИГАиК, доктор технических наук
shulzh79@mail.ru

М.С. Борисенков

Окончил Технологический университет имени А.А. Леонова.. Область научных интересов связана с ракетокосмической областью и разработкой информационно-прогностических систем.

С.Н. Шульженко

Окончил Тульский государственный университет. Область научных интересов связана с разработкой и совершенствованием интеллектуальных информационных систем и их применением в производственных задачах

Аннотация. В статье предложена методика оптимизации признакового пространства телеметрии серверного оборудования на основе корреляционного анализа и аппарата информационной энтропии.

Ключевые слова: прогнозное обслуживание, телеметрия серверного оборудования, препроцессинг данных, коэффициент корреляции Спирмена, прирост информации, энтропия Шеннона, снижение размерности признаков.

Введение и постановка задачи. Экспоненциальное наращивание объемов обрабатываемой информации в сфере высокопроизводительных вычислений и облачных технологий выдвигает на первый план проблему бесперебойности работы ИТ-комплексов. Существующие реактивные подходы к обслуживанию, когда ремонт инициируется лишь после фактического отказа компонента, демонстрируют свою неэффективность в масштабах современных дата-центров. Подобная практика неизбежно влечет за собой рост эксплуатационных расходов и регулярные сбои в выполнении условий соглашений об уровне сервиса (SLA) [3]. В этом контексте переход к предиктивной модели обслуживания (Predictive Maintenance), нацеленной на прогнозирование состояния серверных платформ и систем хранения, приобретает характер первостепенной научно-прикладной проблемы.

Хотя современные аппаратные платформы предоставляют развитый инструментарий для сбора телеметрии (SMART, IPMI, SNMP), непосредственное применение всего объема регистрируемых параметров при создании предиктивных моделей наталкивается на фундаментальные трудности. Проблема заключается в чрезмерной размерности и зашумленности исходного массива данных. Пространство признаков оказывается избыточным, что не просто усложняет вычислительные процедуры обучения нейросетей, но и провоцирует их переобучение. В итоге это негативно сказывается на достоверности долгосрочных прогнозов.

Настоящее исследование направлено на создание и теоретическое обоснование методики, позволяющей сформировать оптимальный вектор информативных признаков. В основе подхода лежит математический анализ параметров, характеризующих деградационные процессы. Объектом изучения служат массивы телеметрических временных рядов, которые регистрируются в ходе эксплуатации серверных узлов и дисковых подсистем. Реализация данной цели обеспечит существенное снижение размерности исходных данных, подаваемых на вход интеллектуальных прогнозных моделей. При этом ключевым условием остается сохранение физической интерпретируемости процессов износа оборудования.

Математический аппарат анализа и фильтрации признаков. Переход от теоретической формулировки задачи к её практическому воплощению в виде интеллектуальной модели диктует необходимость строгого обоснования отбора входных параметров. В рамках данного исследования эта процедура реализуется через последовательное применение аппарата математической статистики и положений теории информации. Такой подход обеспечивает формирование компактного, но при этом предельно насыщенного информацией признакового пространства. Процесс обработки данных начинается с этапа предварительной фильтрации и нормализации, направленного на исключение параметров, не обладающих предсказательной способностью. Первоначально из рассмотрения изымаются квазипостоянные признаки, дисперсия которых 2 стремится к нулю, так как отсутствие вариативности в данных делает невозможным фиксацию динамики деградации. Для преодоления проблемы разнородности шкал телеметрических метрик (температурных показателей, счетчиков ошибок и коэффициентов загрузки) применяется процедура Z-масштабирования:

$$z = \frac{x-\mu}{\sigma},$$

где – математическое ожидание, а – среднеквадратичное отклонение признака кореляции [1]. Такое преобразование приводит все анализируемые величины к единому масштабу, обеспечивая корректную работу последующих алгоритмов.

Дальнейшее уточнение признакового пространства осуществляется путем анализа статистических взаимосвязей для выявления избыточных параметров. С этой целью рассчитывается коэффициент ранговой корреляции Спирмена который, в отличие от линейных методов, позволяет эффективно идентифицировать монотонные нелинейные зависимости, характерные для процессов постепенного износа аппаратных компонентов:

$$\rho = 1 - \frac{6\sum_1^{2*} d}{n(n^2-1)}$$

При обнаружении высокой степени корреляции (ρ> 0,85) между парой признаков один из них подлежит исключению. Это позволяет устранить явление мультиколлинеарности, которое негативно влияет на стабильность прогнозных моделей. Завершающим этапом математического обоснования является ранжирование оставшихся параметров по их информационной значимости. В основу данного подхода положен расчет прироста информации (Information Gain), базирующийся на определении энтропии Шеннона HS, отражающей меру неопределенности состояния системы (нормальное функционирование или предаварийный режим) [2]:

$$H(S) = -p_i$$

Информационная ценность каждого конкретного признака A оценивается через снижение энтропии системы после включения данного признака в анализ:

$$IG(S,A) = H(S) - \sum_{v \in Values(A)} \frac{|S_V|}{|S|} * H(S_V)$$

Использование иерархической методологии открывает возможность для строгой математической верификации результирующего вектора признаков. Этот верифицированный вектор, в свою очередь, формирует фундаментальную базу для последующей экспериментальной проверки.

Экспериментальное исследование и анализ результатов. Предложенный математический аппарат прошел экспериментальную проверку на обезличенных телеметрических данных, собранных за год непрерывной работы серверного кластера и систем хранения. Практическая часть исследования подтвердила теоретические выкладки и дала возможность измерить результативность созданной методики формирования признакового пространства. В качестве исходного набора данных был сформирован массив, включающий 42 первичных телеметрических параметра. На этапе предварительной обработки, следуя описанному ранее математическому алгоритму, была проведена фильтрация признаков с низкой дисперсией (VarX<0.01), что позволило исключить статические идентификаторы и параметры конфигурации, не изменяющиеся во времени и не несущие прогностической ценности. Последующее применение Z-масштабирования обеспечило нормализацию оставшихся 18 метрик, устранив влияние различий в единицах измерения и подготовив данные для корреляционного анализа. Центральным элементом анализа статистических взаимосвязей выступило построение матрицы ранговой корреляции Спирмена. Выбор данного метода позволил выявить устойчивые зависимости, которые в условиях зашумленности лог-файлов серверного оборудования часто принимают нелинейный характер. На представленной ниже тепловой карте (рис. 1) визуализированы коэффициенты между ключевыми метриками.

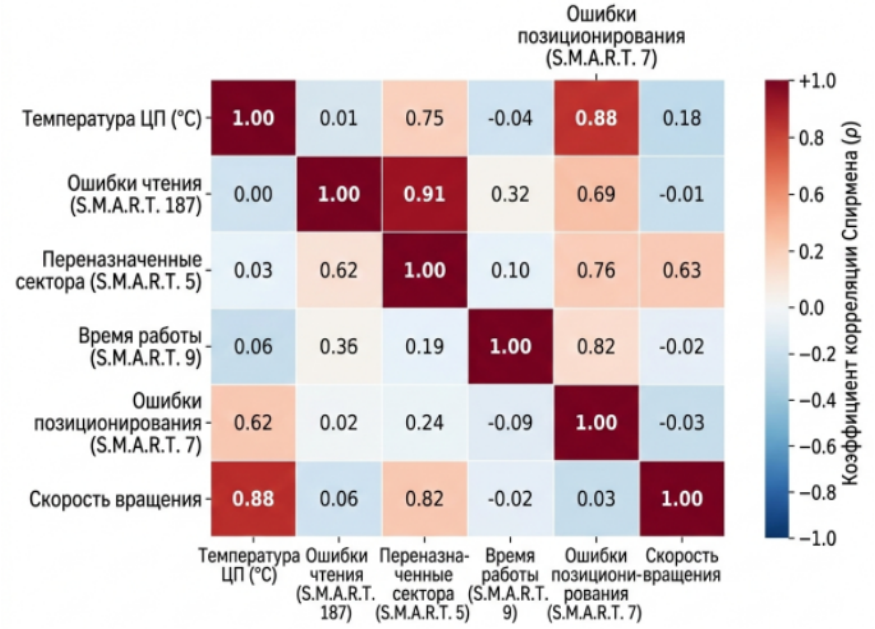


Рисунок 1. Тепловая карта ранговой корреляции Спирмена для параметров телеметрии серверного оборудования

Выявление кластеров с коэффициентом ρ>0,85 (например, между показателями «Reallocated Sectors Count» и «Current Pending Sector Count») послужило математическим основанием для исключения одного из дублирующих признаков, что критически важно для предотвращения переобучения будущей интеллектуальной модели.

Для последующего упорядочивания признаков был задействован метод, основанный на концепции информационной энтропии. Оценивая величину прироста информации (Information Gain), удалось выявить те параметры, что в максимальной степени уменьшают энтропийную неопределенность в оценке текущего состояния технической системы. В таблице 1 представлены результаты вычислений, где значения IG ранжированы по убыванию, что позволяет выделить ядро наиболее информативных переменных.

Таблица 1. Математические показатели значимости признаков в прогнозировании отказов

Признак (Параметр телеметрии)	**Коэффициент вариации**	**Значение Information Gain (IG)**	**Принятое решение**
Reallocated Sector Count	0.84	0.428	Включить в модель
Average CPU Core Temp	0.32	0.315	Включить в модель
Reported Uncorrectable Errors	0.91	0.284	Включить в модель
Read Error Rate (Raw)	0.76	0.192	Включить в модель
Power-On Hours	0.05	0.156	Исключить (низкая динамика)
Spin-Up Time	0.12	0.082	Исключить (низкий IG)

Физическая обоснованность математических выводов проверялась через сопоставление временных рядов (риунок 2).

Графики, иллюстрирующие динамику ключевых параметров за 72 часа до критического отказа, отчетливо выявляют лавинообразную деградацию системы. Это проявляется, например, в синхронном нарастании температуры центрального процессора и частоты ошибок при чтении данных – явление, которое находит свое математическое выражение в скачкообразном изменении градиента функции состояния.

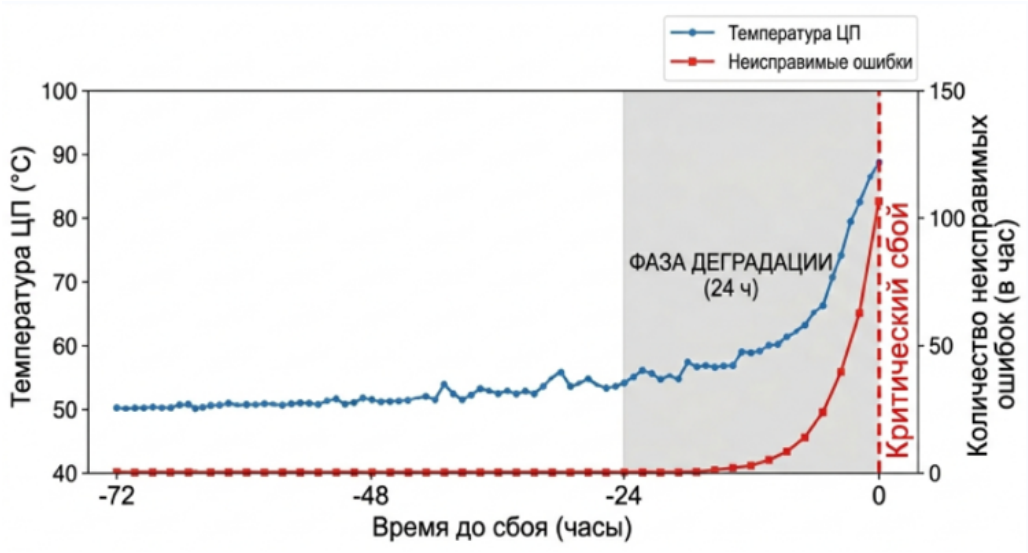


Рисунок 2. Временная корреляция ключевых параметров деградации (температуры ЦП и неисправимых ошибок)

Итогом экспериментального исследования стало формирование оптимизированного вектора из 7 информативных признаков.

Математически доказано, что сокращение размерности признакового пространства на 83% не привело к существенной потере информационной ценности (суммарный IG отобранных признаков покрывает более 90% энтропии целевой переменной).

Таким образом, сформированная база данных является репрезентативной и достаточной для реализации следующего этапа работы – построения интеллектуальной модели прогнозирования на основе глубокого обучения.

Заключение. В данной статье представлена и экспериментально проверена методика построения оптимального признакового пространства для систем прогнозирования технического состояния серверов и систем хранения данных. В отличие от традиционных эмпирических методов отбора параметров, предложенный алгоритм использует формальный математический аппарат, основанный на принципах теории информации и корреляционного анализа.

Основные результаты исследования:

1. Предлагаемый двухэтапный механизм фильтрации реализует последовательную обработку данных: на первом этапе применяется Z-масштабирование, затем устраняется мультиколлинеарность признаков с использованием коэффициента Спирмена, а завершающим шагом становится их ранжирование по метрике прироста информации (Information Gain).

2. Экспериментальные результаты демонстрируют существенную информативность динамических параметров, таких как температурные режимы и отдельные S.M.A.R.T.-счетчики. Их совокупный вклад в общую энтропию модели позволил сократить размерность входного вектора на 83%, не нарушая репрезентативности исходной выборки.

3. Для разработки прогнозных моделей, основанных на глубоком обучении, была сформирована соответствующая информационная база. Математически верифицированный набор, включающий семь ключевых признаков, позволяет существенно сократить вычислительные затраты на этапе обучения нейронных сетей, сохраняя при этом высокую точность прогнозирования критических отказов.

Предложенная методология предобработки и анализа телеметрии открывает путь к проектированию и обучению архитектуры нейронной сети.

Внедрение этих алгоритмов в системы мониторинга создаст фундамент для автоматизации технического обслуживания, что, в свою очередь, минимизирует риски незапланированных простоев серверной инфраструктуры.

Список литературы

[1] Хасти, Т. Элементы статистического обучения: интеллектуальный анализ данных, вывод и прогнозирование / Т. Хасти, Р. Тибширани, Д. Фридман. – 2-е изд. – Москва : Вильямс, 2020. – 768 с.

[2] Кудряшов, Б. Д. Теория информации : учебник для вузов / Б. Д. Кудряшов. – Санкт-Петербург : Питер, 2009. – 320 с.

[3] Герасимов, А. Н. Методы и системы мониторинга технического состояния серверного оборудования / А. Н. Герасимов. – Москва : Техносфера, 2022. – 184 с.

Авторский вклад

Борисенков Матвей Сергеевич – постановка задачи исследования, формирование репрезентативной выборки телеметрических данных, программная реализация алгоритмов фильтрации и масштабирования признаков, проведение рангового корреляционного анализа и расчет метрик информационного прироста.

Шульженко Сергей Николаевич – постановка задачи исследования, концептуальное обоснование применения аппарата теории информации, верификация предложенной методики оптимизации признакового пространства и интерпретация полученных статистических зависимостей.

ATURE SPACE FORMATION FOR PREDICTIVE ANALYSIS OF DISK SUBSYSTEM HEALTH BASED ON S.M.A.R.T. ATTRIBUTES

M.S. Borisenkov
Engineer, PhD Student

S.N. Shuzhenko
Professor of the Department of ITUS A.A. Leonov Technological University (Branch of MIIGAiK), Doctor of Technical Sciences

Abstract. The article proposes a methodology for optimizing the feature space of server telemetry based on correlation analysis and the information entropy apparatus.

Keywords: predictive maintenance, server equipment telemetry, data preprocessing, Spearman correlation coefficient, information gain, Shannon entropy, feature dimensionality reduction.

УДК 658.8:004.9

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПРИМЕНЕНИЯ ГРАФОВЫХ МОДЕЛЕЙ ПРИ ОЦЕНКЕ КАЧЕСТВА ПРОГРАММНЫХ СИСТЕМ НА ЯЗЫКЕ GO И ДРУГИХ ЯЗЫКАХ ПРОГРАМИРОВАНИЯ





Д.А. Чернов
Программист, аспирант
chbedbro@gmail.com

С.Н. Шульженко
Профессор ТУ им. А.А. Леонова (филиал) МИИГАиК, доктор технических наук
shulzh79@mail.ru

Д.А. Чернов
Окончил Технологический университет имени А.А. Леонова.. Область научных интересов связана с ракетокосмической областью и разработкой информационно-прогностических систем.

С.Н. Шульженко
Окончил Тульский государственный университет. Область научных интересов связана с разработкой и совершенствованием интеллектуальных информационных систем и их применением в производственных задачах.

Аннотация. В работе исследуется применение графовых моделей для оценки качества программных систем, разработанных на языке Go, и проводится их сравнительный анализ с программами на других языках программирования. Целью исследования является выявление структурных особенностей кода, влияющих на надёжность и сопровождаемость программных приложений. Программная система представляется в виде ориентированного графа зависимостей, на основе которого рассчитываются показатели структурной сложности и связанности. В результате выявлены различия в плотности связей и характере зависимостей, отражающие специфику архитектурных решений языка Go.

Ключевые слова: графовые модели, анализ программного кода, качество программного обеспечения, язык Golang, структурная сложность, графы зависимостей, сравнительный анализ, метрики связанности, архитектура программных систем.

Введение. Современные программные системы характеризуются высокой степенью структурной сложности и распределённости. Рост масштабов проектов и увеличение числа

взаимосвязанных компонентов приводит к усложнению процессов анализа качества и обеспечения надёжности программного обеспечения.

Вопросы надёжности и структурной устойчивости программных систем подробно рассматриваются в работах Липаева[8] и Шубинского [14], где подчёркивается влияние архитектурной организации кода на эксплуатационные характеристики программ. Традиционные методы оценки качества основаны на использовании структурных и количественных метрик, включая показатели сложности и анализа зависимостей [1, 11]. Однако такие метрики преимущественно ориентированы на локальные характеристики кода и не всегда позволяют адекватно оценить архитектурные особенности программной системы.

В последние годы всё большее внимание уделяется применению формальных моделей представления программ, в том числе графовых структур [6, 12].

Графо ориентированный подход позволяет рассматривать программную систему как совокупность взаимосвязанных сущностей, что даёт возможность анализировать её глобальные структурные свойства.

Программная система может быть формально представлена в виде ориентированного графа (таблица 1):

$$G=(V,E),$$

где V – множество программных сущностей,
E – множество зависимостей между ними.

Таблица 1. Пример ориентированного графа зависимостей программной системы

Показатель	**Обозначение**	**Формула**	**Интерпретация**
Число вершин	$\|V\|$	–	Количество программных сущностей
Число рёбер	$\|E\|$	–	Количество зависимостей
Плотность	ρ	$\rho = \|E\| / (\|V\|(\|V\|-1))$	Степень насыщенности зависимостями
Средняя степень	$\bar{k}$	$\bar{k} = 2\|E\| / \|V\|$	Среднее число связей на вершину
Коэффициент кластеризации	C	$C = 3T/\tau$	Плотность локальных подсистем
Диаметр	D	$\max d(v_i,v_j)$	Глубина транзитивных зависимостей

В рамках графовой модели особое значение имеет показатель плотности зависимостей:

$$\rho= 2EV(V-1),$$

который отражает степень насыщенности системы внутренними связями и может рассматриваться как индикатор архитектурной сложности. Особенности построения графовых моделей в программной инженерии также рассматриваются в работах, посвящённых графо ориентированным методам организации вычислений и архитектурных решений [3, 12].

Это создаёт теоретическую основу для сравнительного анализа структур программных систем, реализованных на различных языках программирования.

Графовая модель представления программной системы. Для формального анализа структурных характеристик программных систем в работе используется их представление в виде ориентированного графа зависимостей. Такой подход позволяет рассматривать программный код не как совокупность отдельных файлов или функций, а как целостную систему взаимосвязанных элементов. Вершины графа соответствуют программным сущностям (функциям, модулям, типам данных), а рёбра отражают отношения между ними – вызовы, импортируемые зависимости и другие формы взаимодействия. Использование графовой модели обеспечивает возможность количественной оценки архитектурной организации системы. В отличие от локальных метрик сложности, ориентированных на отдельные фрагменты кода, графовое представление позволяет анализировать структуру на

уровне всей программной системы. Подобные подходы применяются при исследовании взаимосвязанных вычислительных компонентов и сложных программных комплексов [3, 12].

В рамках исследования анализируются такие характеристики, как степень связанности компонентов, насыщенность системы внутренними зависимостями и распределение связей между программными элементами. Рост числа межмодульных зависимостей, как отмечается в исследованиях, посвящённых надёжности программных систем [1, 11], приводит к увеличению вероятности дефектов и усложнению сопровождения. Графовые показатели могут рассматриваться как индикаторы архитектурной сложности и потенциальных рисков качества. Особый интерес представляет применение данного подхода к программам, реализованным на языке Go. Специфика организации кода в Go – строгая модульность пакетов, отсутствие классического наследования, использование композиции и встроенных механизмов конкурентности – формирует отличающуюся структуру зависимостей. Это создаёт предпосылки для сравнительного анализа архитектурных характеристик программ, реализованных на различных языках программирования.

Структурные особенности программных систем на языке Go. Язык Go изначально проектировался как средство разработки масштабируемых и распределённых систем, что повлияло на его архитектурные принципы и организацию кода. В отличие от классических объектно-ориентированных языков, Go не поддерживает наследование в традиционном виде, отдавая предпочтение композиции и интерфейсному взаимодействию. Это приводит к иной структуре зависимостей между программными сущностями. Модульность в Go реализуется через пакеты, каждый из которых формирует самостоятельную единицу компиляции. Импорт пакетов задаёт явные зависимости, что делает граф межмодульных связей более прозрачным и формально определённым. Отсутствие глубокой иерархии наследования, характерной для языков с развитой объектной моделью, снижает вероятность формирования сильно вложенных зависимостей и уменьшает глубину графа. Существенное влияние на структуру программ оказывает механизм конкурентности – использование goroutines и каналов. Хотя конкурентные взаимодействия не всегда отражаются напрямую в статическом графе зависимостей, они формируют дополнительные логические связи между компонентами. Это приводит к появлению функционально связанных, но структурно разнесённых элементов системы.

В сравнении с языками, использующими классическую объектно-ориентированную парадигму, структура зависимостей в Go, как правило, характеризуется (таблица 2):

- меньшей глубиной иерархий;
- более явными границами модулей;
- сниженной плотностью межклассовых связей;
- меньшим количеством транзитивных зависимостей.

Таблица 2. Сравнение типовых архитектурных структур зависимостей

Характеристика	**Go**	**Языки с наследованием**
Иерархия классов	Отсутствует	Выраженная
Транзитивные зависимости	Ограниченные	Глубокие
Централизация структуры	Низкая	Возможна высокая
Плотность графа	Умеренная	Может быть повышенной
Распределение степеней	Более равномерное	С наличием концентраторов

Исследования, посвящённые анализу архитектурной сложности программных систем [1, 11], показывают, что увеличение числа транзитивных связей и скрытых зависимостей коррелирует с ростом дефектности и усложнением сопровождения. С этой точки зрения архитектурные особенности Go потенциально способствуют формированию более предсказуемой структуры графа.Для сравнительного анализа были рассмотрены программные проекты, реализованные на языке Go и на языках с развитой объектной моделью. Оценка

проводилась на основе показателей связанности, плотности зависимостей и распределения степеней вершин графа. Анализ показал, что для приложений на Go характерна более равномерная структура распределения связей без выраженных «центров притяжения», тогда как в системах с наследованием чаще наблюдаются узлы с высокой степенью связности, играющие роль архитектурных концентраторов. Специфика языка программирования оказывает влияние на формирование графовой структуры программной системы, что необходимо учитывать при оценке её качества и архитектурной устойчивости.

Сравнительный анализ графовых характеристик программных систем. Сравнительный анализ был направлен на выявление различий в структурной организации программных систем, реализованных на языке Go и на языках с развитой объектно-ориентированной моделью. В качестве критериев оценки использовались показатели плотности зависимостей, средней степени вершин и характер распределения связей внутри графа. Полученные результаты показывают, что для проектов на языке Go характерна более модульная структура с чётко выраженными границами пакетов и относительно равномерным распределением связей между компонентами. Средняя степень вершин в графах зависимостей, как правило, не демонстрирует резких пиков, что свидетельствует об отсутствии чрезмерно централизованных элементов архитектуры. В программных системах, построенных на основе классического наследования, наблюдается иная картина. Формирование иерархий классов и использование базовых абстракций приводит к появлению узлов с высокой степенью связности. Такие элементы играют роль архитектурных центров и при изменении могут оказывать влияние на значительное число зависимых компонентов.

Согласно исследованиям в области оценки структурной сложности [1, 11], высокая концентрация связей в отдельных узлах повышает чувствительность системы к изменениям и увеличивает риск каскадных дефектов. Дополнительным фактором, влияющим на графовую структуру, является способ организации конкурентного взаимодействия. В Go логические связи между компонентами, возникающие при использовании goroutines и каналов, зачастую не увеличивают плотность статического графа зависимостей, поскольку реализуются через явные интерфейсы взаимодействия. В целом анализ показывает, что архитектурные принципы языка программирования оказывают непосредственное влияние на формирование графовой структуры программной системы. Для языка Go характерна тенденция к более плоской и модульной архитектуре, тогда как системы с выраженной иерархической моделью демонстрируют более сложную структуру зависимостей.

Заключение. В работе рассмотрено применение графовых моделей для анализа структурной организации программных систем и проведён сравнительный анализ их характеристик для приложений, реализованных на языке Go и на других распространённых языках программирования. Показано, что представление программной системы в виде ориентированного графа зависимостей позволяет формализовать архитектурные особенности кода и перейти от качественного описания структуры к количественной оценке её сложности и связанности. Результаты анализа свидетельствуют о том, что архитектурные принципы языка Go способствуют формированию более модульной и менее иерархически перегруженной структуры зависимостей по сравнению с языками, использующими классическую модель наследования. Для систем на Go характерно более равномерное распределение связей между компонентами и меньшая концентрация зависимостей в отдельных узлах графа, что потенциально снижает чувствительность архитектуры к локальным изменениям. Полученные выводы подтверждают целесообразность применения графовых метрик в качестве инструмента оценки качества программных систем и могут быть использованы при разработке средств автоматизированного анализа кода, а также при проектировании масштабируемых и распределённых приложений.

Перспективным направлением дальнейших исследований является расширение набора анализируемых метрик и применение методов интеллектуального анализа данных для прогнозирования дефектности на основе графовых характеристик программных систем.

Список литературы

[1] Бобрышев А.Н., Жарин Д.Е., Шафигуллин Л.Н., Гумеров М.И. Метрики структурной сложности программных систем. Информационные технологии. 2017;23(5):321–327.

[2] Дмитренко Е.Н. Методы анализа надежности сложных программных систем. Программная инженерия. 2019;10(4):150–158.

[3] Димитриенко Ю.И., Соколов А.П. Графоориентированный подход к организации распределённых вычислений. Известия вузов. 2008;4:33–40.

[4] Иванников В.П., Першин А.Ю. Инструментальные средства анализа качества программного обеспечения. Информатика и её применения. 2014;8(1):15–23.

[5] Климов С.М., Сосновский Ю.В., Чачиев Д.Р. Методика оценки функциональной надежности компонент программно-аппаратной встраиваемой микропроцессорной системы управления. Надежность. 2025;25(1):58–66. DOI:10.21683/1729-2646-2025-25-1-58-66.

[6] Кравченко Ю.А. Применение графовых моделей при анализе программных систем. Вестник МГТУ. 2015;18(2):45–52.

[7] Кузнецов В.В. Методы статического анализа исходного кода. Программные продукты и системы. 2021;34(2):89–97.

[8] Липаев В.В. Надежность программного обеспечения. М.: СИНТЕГ; 2008.

[9] Нетес В.А. Надежность программных средств: терминология и проблемы стандартизации. Надежность. 2016;16(3):3–10.

[10] Русаков А.М., Юшкова Н.А., Селиванова Е.А. Современные средства визуализации графов и графовых моделей для Python. Наукосфера. 2022;11(2):258–266. DOI:10.5281/zenodo.7372018.

[11] Сидоров А.А. Модели прогнозирования дефектов программных систем. Программная инженерия. 2020;11(6):255–262.

[12] Соколов А.П., Першин А.Ю. Система автоматизированного проектирования композиционных материалов. Часть 1. Концепции, архитектура и платформа разработки. Известия СПбГЭТУ «ЛЭТИ». 2020;8–9:72–83.

[13] Чаругин В.В., Чаругин В.В., Чесалин А.Н. Метрики оценки качества локализации уязвимостей и автоматического исправления исходного кода. Инновационные, информационные и коммуникационные технологии. 2025;22:399–403.

[14] Шубинский И.Б. Функциональная надежность информационных систем. М.: Горячая линия–Телеком; 2012.

[15] Громов А.А. Применение теории графов в задачах анализа архитектуры программного обеспечения. Вестник компьютерных и информационных технологий. 2018;15(6):27–35.

Авторский вклад

Чернов Данила Андреевич – постановка задачи исследования, формирование репрезентативной выборки телеметрических данных, программная реализация алгоритмов фильтрации и масштабирования признаков, проведение рангового корреляционного анализа и расчет метрик информационного прироста.

Шульженко Сергей Николаевич – постановка задачи исследования, концептуальное обоснование применения аппарата теории информации, верификация предложенной методики оптимизации признакового пространства и интерпретация полученных статистических зависимостей.

COMPARATIVE ANALYSIS OF THE APPLICATION OF GRAPH MODELS IN ASSESSING THE QUALITY OF SOFTWARE SYSTEMS WRITTEN IN GO AND OTHER PROGRAMMING LANGUAGES

D.A. Chernov
Programmer, PhD Student

S.N. Shuzhenko
Professor of the Department of ITUS A.A. Leonov Technological University (Branch of MIIGAiK), Doctor of Technical Sciences

Abstract. The paper investigates the application of graph-based models for assessing the quality of software systems developed in the Go programming language and provides a comparative analysis with programs written in other programming languages. The objective of the study is to identify structural code characteristics that affect the reliability and maintainability of software applications. The software system is represented as a directed dependency graph, on the basis of which structural complexity and connectivity metrics are calculated. The results reveal differences in link density and dependency patterns that reflect the architectural specifics of the Go language.

Keywords: graph models, source code analysis, software quality, Go (Golang) language, structural complexity, dependency graphs, comparative analysis, connectivity metrics, software architecture.

УДК 658.8:004.9

РАЗРАБОТКА ГОЛОСОВОГО ПОМОЩНИКА ДЛЯ ПРИЕМНОЙ КОМИССИИ УНИВЕРСИТЕТА НА ОСНОВЕ BIG DATA И RAG-АРХИТЕКТУРЫ



А.В. Казаков
Аспирант ТУ им. А.А. Леонова (филиал) МИИГАиК
kazalexit@gmail.com

А.Н. Казаков
Окончил магистратуру – ВШЭ, Специалитет – МИФИ.. Область научных интересов: голосовые технологии (ASR/TTS), ИИ применительно к тексту и голосу, обработка естественного языка (NLP), интеллектуальные системы поддержки принятия решений, большие языковые модели.

Аннотация. В статье рассматривается процесс разработки интеллектуального голосового помощника для автоматизации работы приемной комиссии Технологического университета им. А.А. Леонова, (филиал) МИИГАиК

Ключевые слова: голосовой помощник, приемная комиссия, RAG, Yandex SpeechKit, подготовка данных, векторизация, NLP, автоматизация образования.

Введение и постановка задачи. Современные вызовы цифровизации образования требуют внедрения интеллектуальных систем, способных обрабатывать большие объемы входящих запросов абитуриентов.

Приемная комиссия в период кампании сталкивается с пиковыми нагрузками, когда преподаватели не успевают обрабатывать звонки, что приводит к потере потенциальных студентов или недовольству родителей абитуриентов. Решением является разработка голосового помощника, способного круглосуточно консультировать по правилам приема, срокам подачи документов, перечню экзаменов и проходным баллам.

Целью данного проекта является создание такого помощника для Технологического университета им. А.А. Леонова на основе технологии RAG и сервисов Yandex Cloud.

На момент написания статьи проект находится на начальном, но наиболее важном этапе – этапе подготовки и структурирования данных, от качества которых напрямую зависит точность ответов итоговой системы.

Математический аппарат анализа и фильтрации признаков.Архитектура решения базируется на технологии Retrieval-Augmented Generation (RAG) с интеграцией сервисов Yandex SpeechKit (ASR/TTS) и телефонии (рисунок 1).

Основной фокус работы сделан на критически важном этапе подготовки данных: сборе, очистке, структурировании и векторизации нормативно-справочной информации университета.

Приводятся требования к датасету, методика обработки неструктурированных текстов (приказы, правила, программы) и подходы к эмбеддингу для обеспечения семантического поиска.

Проект находится на начальной стадии, что обуславливает детальное описание препроцессинга как фундамента качественной работы всей системы.

Исследование архитектуры проектируемого решения. Проектируемая система голосового помощника имеет модульную архитектуру, включающую следующие ключевые компоненты: телефонный шлюз (Asterisk) для приема звонков, модуль распознавания речи (ASR Yandex SpeechKit), оркестратор диалогов (LangGraph), RAG-сервис с векторной базой данных, базу знаний университета, модуль синтеза речи (TTS Yandex SpeechKit).

Принцип работы заключается в том, что текст запроса, полученный из голоса, используется для поиска по векторной базе знаний.

Найденные релевантные фрагменты документов (контекст) вместе с исходным запросом передаются языковой модели (LLM) для генерации точного ответа, что минимизирует риск галлюцинаций нейросети.

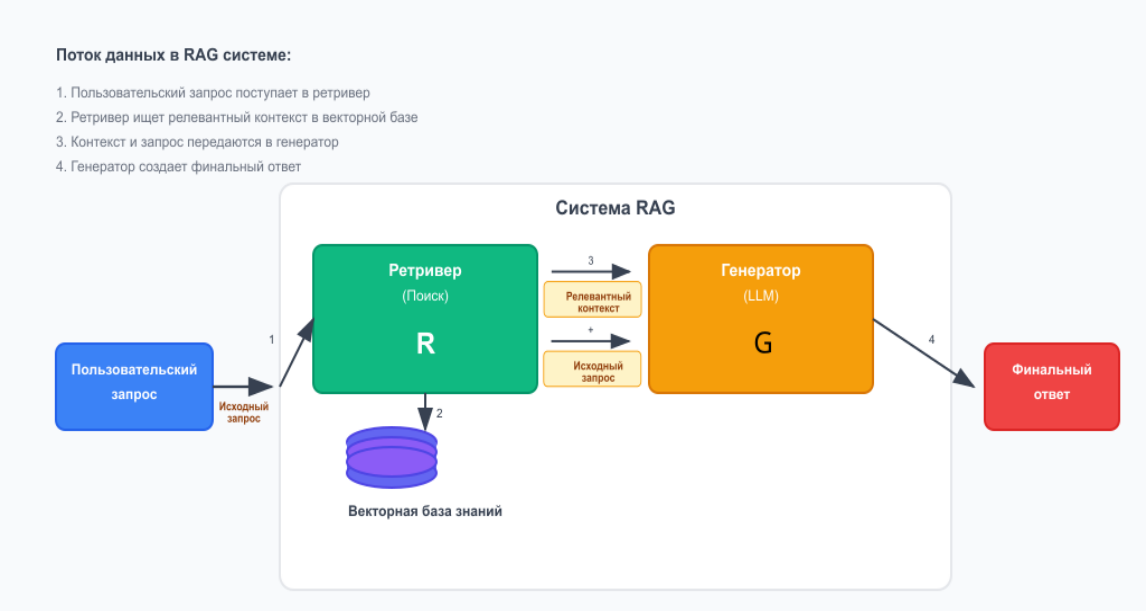


Рисунок 1. Обобщенная схема RAG

Ключевая особенность применения RAG в образовании – высокая цена ошибки.

Недопустимо сообщать абитуриенту неверную дату окончания приема документов.

Поэтому этап подготовки данных является критическим. Работа включает следующие шаги:

- Сбор и первичная обработка. Источниками данных являются официальные правила приема, приказы Министерства науки и высшего образования РФ, программы вступительных испытаний, перечни направлений подготовки, архив вопрос-ответ (FAQ) социальной сети «Вконтакте». На этом данные очищаются от графики, таблиц и нерелевантного форматирования.

- Сегментация документов. Большие документы разбиваются на небольшие фрагменты. Для нашей задачи выбран семантическая фрагментация (разделение по смысловым блокам – главам, параграфам). Размер фрагмента: 500-700 токенов с перекрытием (overlap) в 100 токенов для сохранения контекста на стыках.

- Создание оценочного датасета. Для оценки качества RAG-пайплайна формируется оценочный датасет. Он содержит пары «вопрос – эталонный ответ» с указанием ссылки на конкретный документ-источник.

Таблица 1. Параметры фрагментации документов

Тип документа	**Стратегия фрагментации**	**Размер чанка (токены)**
Правила приема	По главам/пунктам	700
Программы экзаменов	По дисциплинам	500
FAQ (архив)	По парам вопрос-ответ	300

В рамках этапа подготовки данных и проектирования задействован следующий стек технологий: язык программирования Python (библиотеки langchain, pypdf, pandas), модели эмбеддингов от Yandex (YandexGPTEmbeddings) для создания векторных представлений текста, векторное хранилище FAISS (Meta) для быстрого косинусного поиска на этапе прототипирования.

Речевые технологии представлены Yandex SpeechKit, оркестрация тестируется в LangGraph.

Завершение этапа подготовки данных позволит перейти к фазе активной разработки. Ожидается, что внедрение голосового помощника позволит снизить нагрузку на приемную комиссию в «горячий сезон» на 60-70%, обеспечить доступность информации 24/7 и сократить время ожидания ответа приемной комиссии.

Следующими этапами станут: интеграция телефонии, разработка промптов для оркестратора диалогов и нагрузочное тестирование системы.

Заключение. Разработка голосового помощника на базе RAG для приемной комиссии демонстрирует потенциал применения технологий Big Data и искусственного интеллекта в образовании.

Успех проекта напрямую зависит от качества подготовительного этапа – превращения неструктурированных документов в чистую, структурированную базу знаний. Предложенный подход к подготовке данных может быть масштабирован и адаптирован для других подразделений университета.

Список литературы

[1]. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020;33:9459-9474. DOI: 10.48550/arXiv.2005.11401.

[2]. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach; 2017. DOI: 10.48550/arXiv.1706.03762.

[3]. Сошников Д.В. Технологии искусственного интеллекта в облаке: от классических моделей до генеративных нейросетей. Москва: ДМК Пресс; 2025.

[4]. Yandex Cloud. Документация Yandex SpeechKit (ASR/TTS) [Электронный ресурс]. Режим доступа: https://cloud.yandex.ru/docs/speechkit. (Дата обращения: 15.02.2026).

Авторский вклад

Казаков Алексей Владимирович – разработка концепции RAG-архитектуры, выбор методологии векторизации данных.

DEVELOPMENT OF A VOICE ASSISTANT FOR UNIVERSITY ADMISSIONS BASED ON BIG DATA AND RAG ARCHITECTURE

A.V. Kazakov

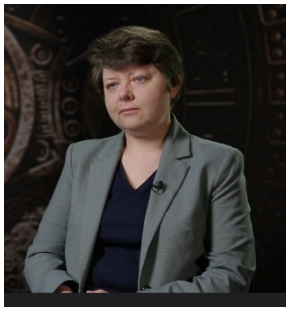
PhD Student, A.A. Leonov Technological University (Branch of MIIGAiK)

Abstract. The article discusses the development of an intelligent voice assistant to automate the work of the admission committee of the Leonov University.

Keywords: voice assistant, admission committee, RAG, Yandex SpeechKit, data preparation, vectorization, NLP, automation in education.

УДК 004.9

ПРОГНОЗИРОВАНИЕ СОСТОЯНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА С ПРИМЕНЕНИЕМ НЕЧЕТКИХ КОГНИТИВНЫХ КАРТ И АЛГОРИТМОВ НЕЛИНЕЙНОГО ОБУЧЕНИЯ



Ю.С. Шевнина
НИУ МИЭТ, профессор института системной и программной инженерии и информационных технологий, д.т.н., доцент, yusm@rambler.ru



А.В. Прокофьев
НИУ МИЭТ, аспирант института системной и программной инженерии и информационных технологий, alexei.prokofiev2015@yandex.ru



О.А. Ашарафьянова
НИУ МИЭТ, студентка 3-го курса института системной и программной инженерии и информационных технологий, ashaolga@mail.ru

Ю.С. Шевнина
Известный ученый в области современных методов анализа и обработки данных, применения методов предиктивной аналитики и искусственного интеллекта в различных сферах.

А.В. Прокофьев
Занимается проектирование и реализацией предиктивных моделей для различных предметных областей.

О.А. Ашарафьянова
Студента направления «Прикладная информатика».

Аннотация. В статье рассматривается подход к прогнозированию состояния образовательного процесса с применением мультиреляционных нечетких когнитивных карт и алгоритмов нелинейного обучения Хебба. Описана методика формирования наборов данных для определения текущего состояния образовательного процесса и динамики его изменения на основе цифрового следа обучающихся, включающего академические, системные и социально-психологические показатели. Представлены результаты апробации подхода, подтверждающие высокую точность прогнозирования рисков академической неуспеваемости.

Ключевые слова: образовательный процесс, нечеткие когнитивные карты, алгоритм Хебба, прогнозирование, цифровой след, Advanced Analytics, Big Data.

Введение. В настоящее время большое внимание уделяется вопросам прогнозирования в различных предметных областях с целью оптимального планирования и управления, предупреждения внештатных ситуаций и снижения влияния других негативных факторов. Известно, что многие социальные и информационные системы, в частности образовательный процесс (ОП), являются многомассивными сложными нелинейными системами с множественной гранулярностью (МСНС). Подобные системы характеризуются взаимодействием множества компонентов, имеющих разные масштабы и уровни организации. В таких системах малые изменения в одном компоненте приводят к значительным изменениям в поведении всей системы из-за нелинейных взаимодействий. Следовательно, моделирование и прогнозирование изменения состояния ОП является своевременной и актуальной задачей. Точный прогноз состояния ОП позволит уменьшить неопределенность внешней среды с учетом внутренних характеристик системы, предотвратить академические угрозы и

неуспеваемость, повысить эффективность управления ресурсами и способствовать устойчивому развитию образовательной системы в целом [1].

Аналитический обзор существующих методов и средств оценки состояния образовательного процесса. Современная практика оценки состояния образовательного процесса интегрирует следующие методы и средства: индексы академической успешности, многоуровневый анализ данных, использование логов цифрового следа и систем дистанционного обучения. Определены наиболее эффективные методы и средства оценки состояния ОП, выявлены их достоинства и недостатки. Тем не менее, указанные методы и средства не позволяют в полной мере определять состояние образовательного процесса и прогнозировать его изменение с учетом зависимостей между различными факторами.

Как уже отмечалось выше, ОП является сложной нелинейной системой с множественной гранулярностью. Среди современных подходов к прогнозированию изменения состояния нелинейных систем следует выделить аналитические методы, основанные на математическом анализе. Однако они ограничены в использовании для сложных и высокоразмерных систем.

Наиболее перспективным к прогнозированию изменения состояния сложных нелинейных систем является подход с использованием когнитивных карт, который позволяет выявлять, визуализировать и анализировать взаимосвязи между компонентами системы, в том числе нелинейными. Когнитивная карта представляет собой граф, в котором узлы являются состояниями компонентов системы, а дуги – связями между ними. Нечеткая когнитивная карта (НКК) сочетает в себе аспекты нечеткой логики, нейронных сетей и других методов и подходит для моделирования систем на основе ресурсов данных.

Для прогнозирования изменения состояния ОП предлагается использовать мультиреляционные НКК, в частности, двухуровневые. При этом на низком уровне для каждого массива создается одна НКК, а на высоком уровне – общая НКК для всех массивов данных о состоянии компонентов и системы.

Математическая модель прогнозирования состояния образовательного процесса. Рассмотрим примеры состояний образовательного процесса, которые возможно описать с помощью нечетких когнитивных карт. Состояния, характеризующие уровень академической успеваемости с учетом влияния посещаемости и выполнения заданий. Состояния, описывающие качество цифрового следа в ОП с учетом взаимодействия с учебными материалами, которые сложно моделировать строго детерминированными методами. Состояния, связанные с рисками академической задолженности и отчисления.

Нечеткая когнитивная карта F математически представляет собой кортеж из элементов $F = \langle C, W, A, f \rangle$, где $C = \{C_1, C_2, .., C_N\}$ – набор из N состояний компонентов, образующих узлы графа; $W : (C_i, C_j) \rightarrow w_{ij}$ - матрица связи между состояниями компонентов образовательного процесса, w_{ij} – значение веса ребра, направленного от C_i к C_j, $w_{ij} \in [-1,1]$; $A : C_i \rightarrow A_i(t)$ - матрица связи между состоянием компонента C_i с последовательностью активирующих его факторов с учетом степени активации.

Очевидно, если существует положительная причинно-следственная связь между состояниями компонентов C_i и C_j, то $w_{ij} > 0$, что означает увеличение вероятности возникновения состояния C_j при увеличении вероятности возникновения состояния Ci, а уменьшение вероятности возникновения C_i приведет к уменьшению вероятности возникновения C_j. При отсутствии связи между состояниями компонентов $w_{ij} = 0$ (рисунок 1).

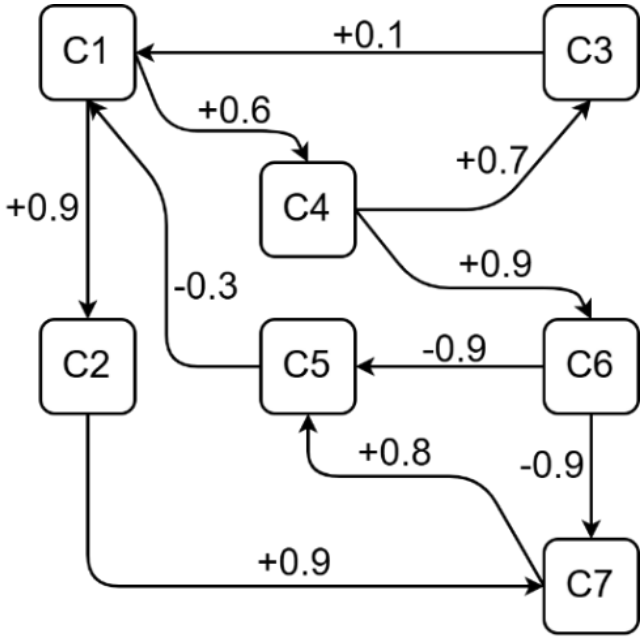


Рисунок 1. Пример нечеткой когнитивной карты

Для моделирования ОП рассмотрим активацию состояния одного из компонентов системы в момент времени $t+1$:

$$A_j(t+1) = f\left(A_j(t) + \sum_{i \neq j} A_i(t) w_{ij}\right) \tag{1}$$

Данная функция позволяет определить влияние изменения состояния одного компонента на другие компоненты системы. При этом функция активации состояния компонента системы учитывает предшествующие состояния, что особенно важно для ОП, поскольку следующее состояние компонента зависит не только от внешних значений, но и предшествующих состояний.

Для контроля входных данных и обеспечения стабильности модели ОП на основе НКК необходимо ограничить сумму значений переменных, умноженных на соответствующие веса связей между ними. Для преобразования целесообразно использовать логистическую функцию:

$$f(x) = \frac{1}{1 + e^{-Cx}} \tag{2}$$

Предложенное описание состояния ОП с использованием нечетких когнитивных карт позволяет анализировать сложные взаимосвязи и влияния компонентов друг на друга.

В реальном мире сложные нелинейные системы имеют многопространственные группы компонентов с прямыми или косвенными связями, которые создают мультиреляционные данные. Более того, каждое такое пространство может содержать множество компонентов с различными отношениями уровня детализации. В частности, в экосистеме ОП существуют три пространства: академическая успеваемость, цифровой след и социально-психологический профиль. Они являются компонентами с высокой гранулярностью; а баллы за тесты, время сессии и активность в системе – компонентами с низкой гранулярностью.

Для прогнозирования изменений состояний компонентов с множественными зависимостями и высокой гранулярностью в сложной системе ОП применяется мультиреляционная НКК с двумя уровнями и *n*-массивами $U_n^2 = \left\langle C_n^2, W_n^2, A_n^2, f \right\rangle$, где C_{n^2}– набор состояний компонентов с высокой гранулярностью. Компонент с высокой гранулярностью можно рассматривать как суммарное выражение его компонентов низкой гранулярности с множеством экземпляров на низком уровне измерения.

Алгоритм обработки и анализа образовательных данных. В основе разрабатываемого алгоритма обработки и анализа данных образовательного процесса лежит алгоритм построения НКК. Алгоритм построения НКК должен адекватно определять влияние одного компонента ОП на другой (например, влияние регулярности работы в системе дистанционного обучения на итоговый балл) и рассчитывать веса связей между состояниями компонентами. В настоящее время наиболее популярными для построения когнитивных карт являются алгоритмы на основе принципа Хебба и эволюционные алгоритмы.

Принцип Хебба гласит, что узлы, которые активируются вместе, соединяются вместе. Для НКК данный принцип определяет правила обновления весов связей между состояниями образовательных компонентов на основе их совместной активности (корреляции данных в цифровом следе). Как правило, для пересчета весов и точного моделирования сложных зависимостей используются нелинейные функции (нелинейное обучение Хебба).

Нелинейное обучение Хебба

Нелинейное обучение Хебба относится к методам обучения нейронной сети без учителя. Для нелинейного выходного блока определим случайный пресинаптический входной вектор *x* (входные метрики обучающегося), весовой вектор *w* и выходной вектор $z = f\left(w^T x\right)$.

Нелинейная функция активации *f* является сигмовидной функцией. Функция критерия *J* максимизируется по принципу Хебба и может быть записана в виде:

$$J = E\left[z^2\right]. \tag{3}$$

Дополнительное ограничение ||*w*||=1 позволяет стабилизировать правила обучения модели НКК. Стохастическое аппроксимирующее решение используется для следующего нелинейного правила обучения Хебба в виде:

$$\Delta w_{ji} = \eta_k z \frac{dz}{dy}\left(x_j - w_{ji} y_i\right) \tag{4}$$

Таким образом, представленные правила нелинейного обучения позволяют определять набор весовых параметров при условии наибольшей дисперсии выходных данных нелинейного модуля.

Непрерывный генетический алгоритм.

Непрерывный генетический алгоритм (НГА) применяется для разработки матрицы связей между состояниями компонентов ОП на основе ресурса данных (логов из LMS). В терминах генетических алгоритмов НГА определяет каждую хромосому как вектор с плавающей точкой. В случае НКК с *N* узлами каждая хромосома состоит из *N*(*N −1) генов, которые представляют собой числа с плавающей точкой диапазона [−1,1], определенные следующим образом:*

$$E = \left[w_{12}, w_{13}, .., w_{1N}, w_{21}, w_{23}, .., w_{2N}, .., w_{NN-1}\right]^T, \tag{5}$$

где w_{ij}- значение веса ребра НКК от *i*-го до *j*-го состояния компонента.

Функция приспособленности определяется для каждой хромосомы путем вычисления разницы между откликом системы, сгенерированным с использованием весов НКК, и соответствующим откликом системы, полученным непосредственно из базы реальных данных обучающихся. Мера ошибки рассчитывается по формуле:

$$Error_L_p = \alpha \sum_{t=1}^{M-1} \sum_{n=1}^{N} \left| A_n(t) - \hat{A}_n(t) \right|^p . \quad (6)$$

где α применяется для нормализации частоты ошибок; N- количество состояний компонентов в НКК; M- количество итераций (учебных периодов). Мера ошибки используется как ядро функции приспособленности в виде:

$$F = \frac{1}{a \cdot Error_L_p + 1} . \quad (7)$$

Обучение по НГА завершается при достижении порогового значения функции приспособленности *max_fitness*. Если ни одно из условий остановки НГА не выполняется с увеличением итераций, то используются эволюционные операторы (скрещивание или мутация) и стратегия выбора для создания нового поколения решений. Точность модели определяется как нормализованная средняя ошибка между соответствующими значениями состояний компонентов на каждой итерации:

$$error = \frac{1}{M \cdot T} \sum_{t=0}^{T} \sum_{j=1}^{M} \left| A_j(t) - \hat{A}_j \right| . \quad (8)$$

Следовательно, весовой вектор в низкоуровневой НКК рассчитывается дополненным многовариантным нелинейным обучением Хебба. Для получения точной классификации и выявления причинно-следственных связей в образовательном процессе применялся непрерывный генетический алгоритм.

Результаты апробации математической модели на основе реальных данных. На основе предложенной модели мультиреляционной НКК проводилось моделирование и прогнозирование состояния образовательной экосистемы. Очевидно, что рассматриваемая система имеет сложную структуру с мультиреляционным набором данных. Состояние экосистемы выражается как множественная взаимосвязь массивов компонентов с низкой гранулярностью. Выделим три набора исходных обезличенных данных о состоянии обучающихся.

Набор BK_0: Базовый профиль. Каждый вектор включает характеристики направления подготовки, формы обучения и результаты входного контроля (ЕГЭ).

Набор BK_1: Цифровой след. Каждый вектор включает метрики активности в LMS (частота входов, время выполнения тестов, пропуски), помимо тех характеристик, что определены в BK_0.

Набор BK_2: Социально-психологический профиль. Включает характеристики социальной активности (участие в форумах, проектной деятельности), помимо тех, что определены в BK_1.

Эксперимент реализован для прогнозирования изменения состояния ОП с применением мультиреляционной НКК и разработанного алгоритма ее построения при различных пороговых значениях функции приспособленности. В качестве обучающей выборки для

гибридной сети использовано 7000 комплектов обезличенных данных обучающихся, а в качестве тестовой выборки – 3000 комплектов. Время построения НКК показано при пороговых значениях функции приспособленности, принимающей значение от 0,6 до 0,9 (рисунок 2).

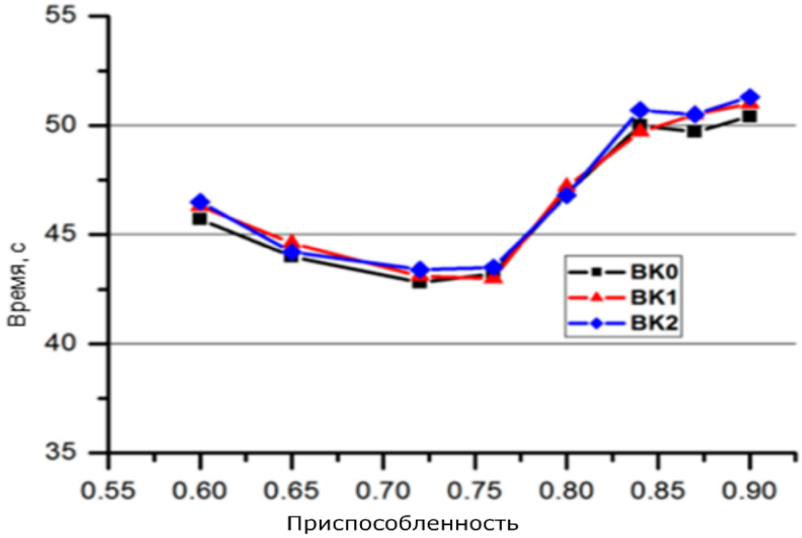


Рисунок 2. Зависимость времени построения НКК от значения функции приспособленности

Как видно из графика, время построения НКК для различных наборов исходных данных практически не отличается, при этом наименьшее затрачиваемое вычислительное время отмечается для приспособленности в интервале [0,65;0,75][0,65;0,75].

Рассмотрим результаты прогноза состояний экосистемы при *max_fitness*=0,7 для трех видов исходных данных (таблица 1).

Таблица 1. Эффективность прогноза состояния образовательной экосистемы

Набор исходных данных	Время выполнения (с)	Точность
BK_0	0,78	92,3%
BK_1	0,80	97,9%
BK_2	0,80	97,7%

Время прогноза изменения состояний экосистемы с набором данных, содержащем большее количество связей, больше. При этом наилучший показатель точности прогноза зафиксирован для BK_1.

Точность прогноза подтверждается и рассчитывается как степень соответствия фактических данных результатам, предсказанным прогнозной моделью по формуле:

$$Точность = \left(1 - \frac{MAE}{\overline{y}}\right) \cdot 100\% \quad . \qquad (9)$$

Точность прогноза 97,9% означает, что средняя ошибка прогноза составляет 2,1% от среднего значения фактических данных. Такая точность свидетельствует о высокой степени соответствия модели реальным условиям образовательного процесса. Для каждого обучающегося рассчитывается прогнозное значение состояния (итоговый балл или риск отчисления).

Затем сравниваются прогнозные и фактические значения. Рассчитывается средняя абсолютная ошибка (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \quad (10)$$

где y_i - фактическое значение состояния, $\hat{y}_i$ - прогнозное значение состояния, *n*– количество обучающихся.

Оценка эффективности полученных результатов. Предложенная система оценки состояния образовательного процесса показывает значительное повышение эффективности и точности оценки по сравнению с традиционными методами академического мониторинга.

В отличие от традиционных методов, требующих длительных замеров успеваемости вручную (по итогам сессий), предиктивная система обеспечивает автоматизированный сбор, обработку и анализ данных в реальном времени или с минимальной задержкой.

Благодаря интеграции с базами данных LMS сокращается время выявления обучающихся в «зоне риска» с нескольких недель до нескольких часов, что повышает оперативность принятия педагогических решений. Традиционные методы выявляют локальные нарушения (академические задолженности) постфактум.

Разработанное решение выявляет скрытые паттерны (снижение мотивации, прокрастинацию) с точностью до 97,9%, что позволяет своевременно принимать меры по корректировке индивидуальных образовательных траекторий.

Заключение. В результате выполнения работы разработана математическая модель прогнозирования состояния образовательного процесса, позволяющая учитывать нелинейные зависимости между академическими, цифровыми и социально-психологическими факторами.

Получены результаты апробации разработанной модели на основе реальных обезличенных данных Big Data.

Разработан алгоритм обработки и анализа данных, интегрирующий многовариантное нелинейное обучение Хебба и непрерывный генетический алгоритм.

Осуществлена оценка эффективности: разработанный подход обеспечивает существенное повышение эффективности академического мониторинга за счет ускорения процессов, увеличения точности предиктивной аналитики и снижения затрат на администрирование образовательного процесса.

Список литературы

[1] Аверкин А.Н. Генетический алгоритм обучения нечетких когнитивных карт / А.Н. Аверкин, А.А. Паринов // Научные труды ВЭО России. – 2010. - № 143. – С. 69-74.

[2] Петухова А.В. Методы прогнозирования развития сложных систем с применением теории нечетких когнитивных карт / А.В. Петухова, А.В. Коваленко // Computational Mathematics and Information Technologies. – 2022. – № 6(2). – С. 81-95.

[3] Shevnina Ju.S. On the Issue of Modeling Complex Nonlinear Systems / Ju.S. Shevnina, L.G. Gagarina, A.V. Chirkow // AIP Conference Proceedings. – 2021. – Vol. 2467. – P. 565.

Авторский вклад

Шевнина Юлия Сергеевна – постановка задачи исследования, разработка рассматриваемых подходов и алгоритмов прогнозирования состояния образовательного процесса.

Прокофьев Алексей Валентинович – разработка методики оценки эффективности полученных результатов, проведение эксперимента и анализ результатов.

Ашарафьянова Ольга Андреевна. - сбор и предобработка данных, реализация алгоритмов обучения Хебба и генетического алгоритма, проведение экспериментов и анализ результатов.

PREDICTING THE STATE OF THE EDUCATIONAL PROCESS USING FUZZY COGNITIVE MAPS AND NONLINEAR LEARNING ALGORITHMS

Yu. S. Shevnina
National Research University of Electronic Technology (MIET), Professor, Institute of Systems and Software Engineering and Information Technology, Doctor of Engineering, Associate Professor, yusm@rambler.ru

A.V. Prokofiev
National Research University of Electronic Technology, graduate student at the Institute of Systems and Software Engineering and Information Technology, alexei.prokofiev2015@yandex.ru

O.A. Asharafyanova
National Research University of Electronic Technology (MIET), third-year student at the Institute of Systems and Software Engineering and Information Technology, ashaolga@mail.ru

Abstract. This article examines an approach to predicting the state of the educational process using multirelational fuzzy cognitive maps and Hebbian nonlinear learning algorithms. It describes a methodology for generating datasets to determine the current state of the educational process and its dynamics based on students' digital footprints, which include academic, systemic, and socio-psychological indicators. The results of testing the approach are presented, confirming the high accuracy of predicting the risk of academic failure.

Keywords: educational process, fuzzy cognitive maps, Hebbian algorithm, forecasting, digital footprint, Advanced Analytics, Big Data.

UDC 336.774.3

PREDICTING TARGET VARIABLE VALUES AND CONSTRUCTING A ROC -CURVE



D.M. Rahel
PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics
ragel@bsuir.by

D.M. Rahel

In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. This article examines approaches to ROC-analysis for modeling commercial data behavior. It examines a set of similar data on retail sales over a specific period of time. Based on existing estimates, a forecast is made and its quality is assessed using baseline metrics confirming the forecast's reliability. Based on the confusion matrix, conclusions were drawn regarding the model's reliability and balance. The final forecast demonstrated a high degree of reliability. The most significant features for classification were identified, providing valuable insights into the factors influencing the final target variable. The combined importance of the identified features allowed us to evaluate their role in the classification process.

Keywords: Data analysis, marketing data, commercial evaluation, big data, data array, forecasting, marketing analytics, commercial analytics, commercial modeling, predictive analytics, behavior forecasting, algorithmic marketing, market forecast, ROC-curve, ROC-analysis.

Introduction. We considered a dataset containing point-of-sale sales results for a major player in the retail market. Based on the initial data, we had to predict the target variable values for the initial dataset and select a forecasting method that would most accurately predict data over the next three months. We also had to evaluate the accuracy and quality of the predictive model. We also had to evaluate the predicted values using a receiver operating characteristic ROC-curve and based on this, identify the three most important predictors influencing the target variable.

Data. In this case, we used commercial sales data with the following parameters:

1. Training set: 10,000 rows, 33 columns.
2. Validation set: 20,000 rows, 33 columns.
3. Target variable: Target (binary: 0 or 1).
4. Distribution of classes in the training set: 50% class 0, 50% class 1.

Method. The Random Forest method was chosen to solve the binary classification problem. This method was chosen for the described dataset for the following reasons:

- Random Forest works effectively with a large number of features and does not require pre-selection of variables;
- the model is resistant to overfitting due to decision tree ensemble;
- the algorithm can account for nonlinear dependencies and interactions between features;
- the method correctly handles data containing gaps and noise, which is especially important for large practical datasets collected at points of sale;
- Random Forest provides a built-in feature importance score, which allows for the interpretation of model results.

Taking this into account, a conclusion was made about the reliability of this method within the framework of the problem being solved. To assess the quality of the predictive model, a confusion matrix was calculated, along with key classification quality metrics: Accuracy; Precision; Recall; and F1-score. The resulting metric values indicate acceptable quality of the model and its ability to correctly distinguish between objects in the two classes.

Results. The quality of the predictive model was assessed on a validation set representing 20% of the original data, using standard binary classification metrics.

The following performance indicators were obtained:

- Accuracy = 0.6915;
- AUC (Area Under ROC-Curve) = 0.755.

The Accuracy value indicates that the model correctly classifies approximately 69% of observations. An AUC value greater than 0.755 indicates good class discrimination and confirms the adequacy of the constructed model (Figure 1).

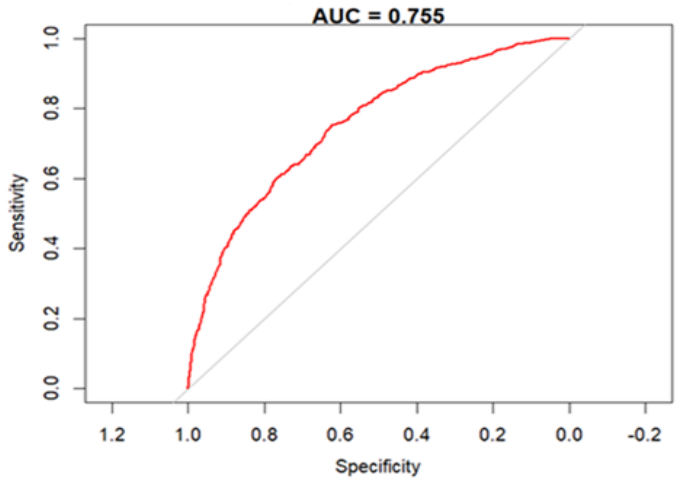


Figure 1. he result of ROC-analysis of the classification of values

Additionally, a confusion matrix was calculated, showing the distribution of correct and incorrect classifications, allowing for a detailed assessment of the model's performance for each class. Based on the feature importance assessment using the Mean Decrease Gini criterion, three of the most significant predictors were identified: P16, P23, and P10. These features contribute most to the decision-making process based on the Random Forest model and have a key influence on the final forecast for our dataset (Figure 2). Taking into account objective assessments of significance, it is these three features that are the most important for forecasting and they will be decisive in forming a forecast for the data set under consideration.

```
Топ-3
> print(head(imp_df, 3))
    Predictor Importance
P16       P16   288.8789
P23       P23   246.7440
P10       P10   226.3062
```

Figure 2. The most significant features in the data set under consideration

In addition, a predictive binary classification model was built using the Random Forest method. The model was trained and tested on a validation set, and its quality was assessed using the Accuracy and AUC metrics. The model was also interpreted by identifying the most important features. The results showed that the model has satisfactory quality and can be used to solve value prediction problems within this dataset (Figure 3).

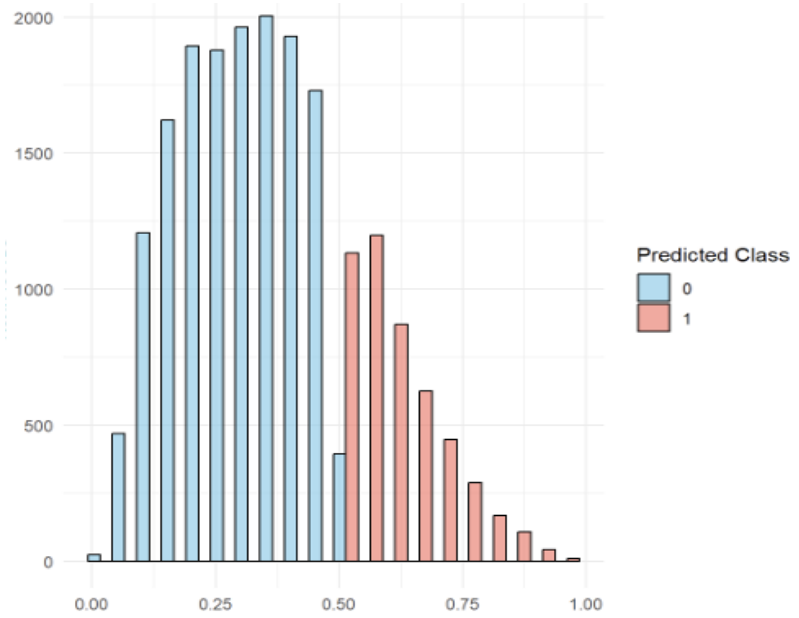


Figure 3. Final class probability distribution

The Random Forest method was chosen to solve the classification problem because it effectively models complex nonlinear relationships, is robust to overfitting, and provides a built-in feature importance assessment mechanism. The confusion matrix demonstrates the balanced quality of the model: it identifies positive-class objects reasonably well and does not exhibit a critical bias toward false positives. At the final stage, a final prediction for the dataset was generated, which, following practical implementation, demonstrated a high degree of reliability. Furthermore, the analysis successfully identified the three most important predictors. Understanding feature importance not only improves the model but also provides valuable insights into the factors influencing the target variable. The combined importance of the three identified features exceeds 27%, indicating their dominant role in the classification process.

References

[1] Коэн М.И. Прикладная линейная алгебра для исследователей данных / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, 2023. – 328 с.

[2] Практическая статистика для специалистов Data Science: Пер. с англ. / П.Брюс, Э.Брюс. – СПб.: БХВ-Петербург, 2019. – 304 с.

[3] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

ПРЕДСКАЗАНИЕ ЗНАЧЕНИЙ ЦЕЛЕВОЙ ПРЕМЕННОЙ И ПОСТРОЕНИЕ ROC-КРИВОЙ

Д.М. Рагель

к.э.н., доцент кафедры экономики Белорусского государственного университета информатики и радиоэлектроники

Аннотация. В статье рассматриваются подходы к ROC-анализу моделирования поведения коммерческих данных. В статье рассматривается набор однотипных данных о продажах в розничной сети в течение отдельно взятого периода времени. На основании существующих оценок делается прогноз и оценивается его качества на основании базовых показателей, подтверждающих уровень достоверности прогноза. На основании матрицы ошибок сделаны выводы о достоверности и сбалансированности модели. Итоговый прогноз показал высокую степень достоверности. Определены наиболее значимые признаки для классификации, что позволило получить ценную предметную информацию о факторах, влияющих на итоговую целевую переменную. Совокупная важность выделенных признаков позволила дать оценку их роли в процессе классификации.

Ключевые слова: Анализ данных, маркетинговые данные, коммерческая оценка, большие данные, массив данных, прогнозирование, маркетинговая аналитика, коммерческая аналитика, коммерческое моделирование, прогнозная аналитика, прогнозирование поведения, алгоритмический маркетинг, рыночный прогноз, ROC-кривая, ROC-анализ.

UDC 336.774.3

CONSUMER BEHAVIOR MODELS IN THE MOBILE COMMUNICATIONS MARKET



D.M. Rahel

PhD in Economics, Associate Professor of the Department of Economics of the Belarusian State University of Informatics and Radioelectronics

ragel@bsuir.by

D.M. Rahel

In 2000 he graduated from the Faculty of Economics of the Belarusian State University of Informatics and Radioelectronics with a degree in Economic Informatics. In 2016 he graduated from the postgraduate course of the Academy of Management under the President of the Republic of Belarus. In 2018 he defended his PhD thesis. Research interests: data mining in marketing, process modeling, data analysis, statistical forecasting, macroeconomics.

Annotation. This article examines approaches to forecasting consumer behavior in the technology market, taking into account the key factors that describe subscriber behavior. We consider a data set with a basic set of market characteristics and apply several models to generate forecasts for the future. Based on these basic indicators, the model with the best scores is selected, demonstrating optimal performance in terms of solving the stated problems.

Keywords: Data analysis, marketing data, big data, massive data, forecasting, marketing analytics, action analysis, market activity, matrices, mobile communications market, subscribers, scalar values, predictive analytics, market forecast.

Introduction. The original dataset contained subscriber information, which we used to describe and build models of their behavior, including information on tariff changes. In this case, the behavior of a limited group of subscribers was examined across five tariff plans.

Data. In this case, the analysis was conducted using a subscriber behavior dataset containing fields for subscriber identification numbers, traffic volumes, and information about blockings resulting from subscriber payment defaults.

The subscriber payment history field is designated as Series1 and represents a dynamic series with timestamps (Figure 1).

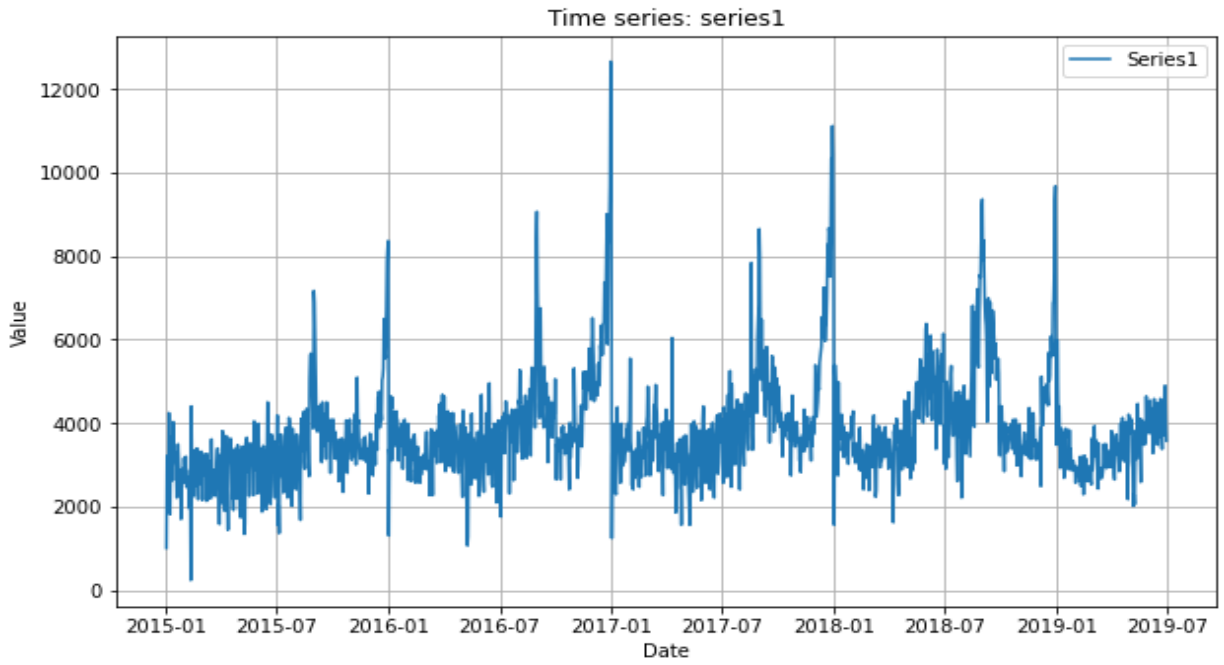


Figure 1. Existing data series for subscriber analysis

Method. To evaluate behavioral models and solve the problem of forecasting payment volumes, various models capable of most accurately predicting subscriber group behavior, taking into account as many formal behavioral characteristics as possible, were considered.

To more accurately select a model, the data series under consideration was decomposed, revealing its key characteristic components – an upward trend, seasonality, and noise (Figure 2).

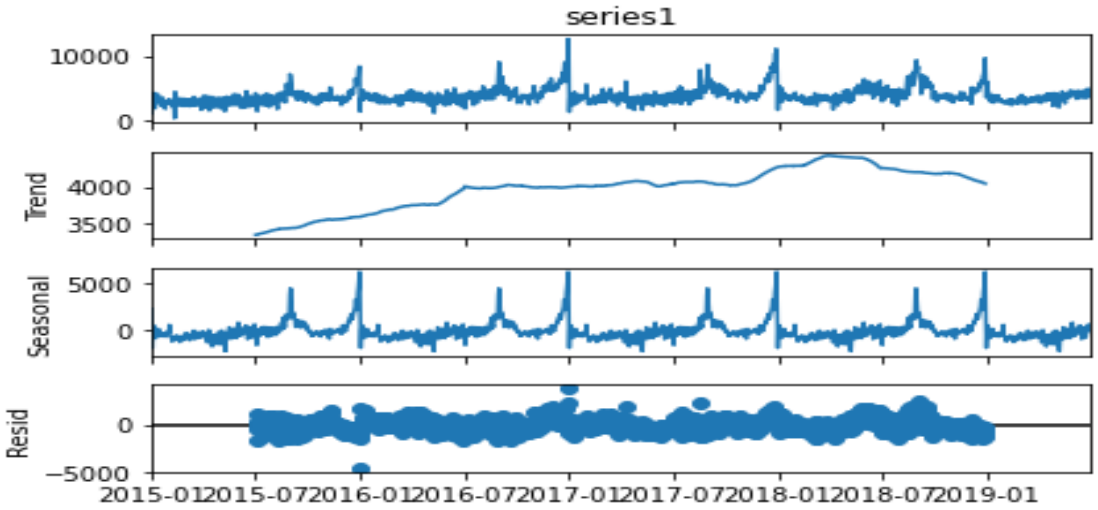


Figure 2. Result of time series decomposition

To assess the stationarity of the existing series, we also assessed stationarity and seasonal patterns in our data series using the autocorrelation function (Figure 3).

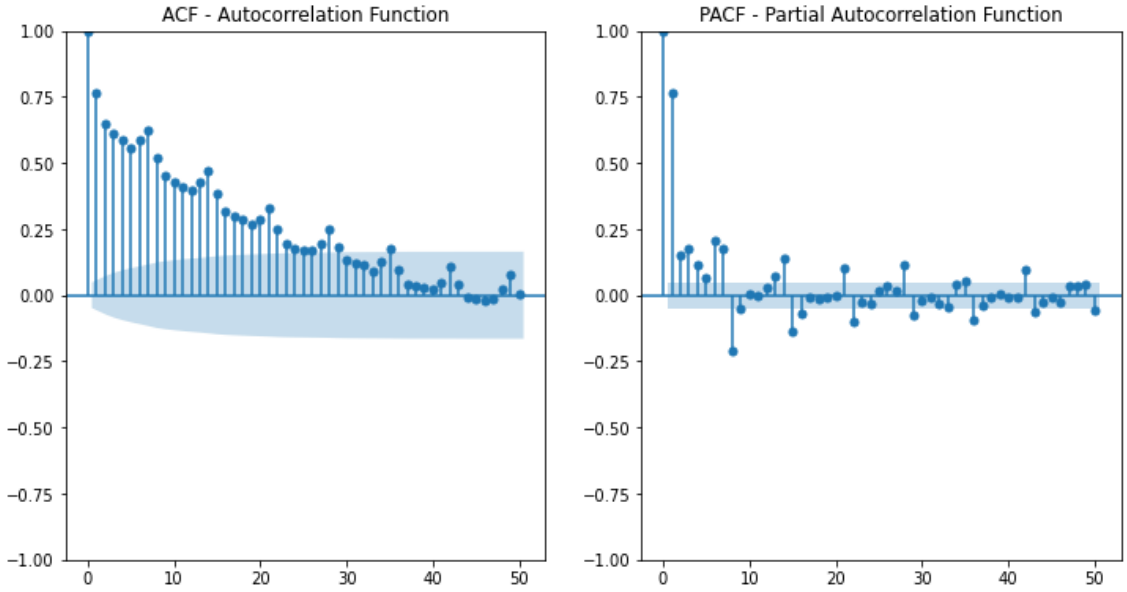


Figure 3. Results of the autocorrelation calculation for our data series

Results. Ultimately, it can be noted that mobile subscriber behavior depends on the service package provided within their tariff. However, tariff changes are generally characterized by positive billing dynamics, indicating that mobile communication costs are not a determining factor for the majority of subscribers. Furthermore, it is important to note the presence of seasonality in mobile communication costs; despite the upward trend in costs, the cycle of communication costs is also quite pronounced, as became evident from the breakdown of the main factors in the data series under consideration.

Based on the analysis, we selected two standard models that best matched the nature of the series under consideration: the Prophet model and the ETS model. These models are more business-focused and therefore take into account seasonality and other potential events that correspond to the characteristics of mobile subscriber consumer behavior (Figure 4).



Figure 4. Results of forecasting consumer behavior based on the ETS and Prophet models

The initial data fluctuates significantly (mean value of 3855, with a deviation of 1237). Significant properties of this series include seasonality and nonstationarity, as determined by the Dickey-Fuller test. Five models were tested on a 329-day test set for forecasting. The Prophet model demonstrated the lowest RMSE (1191.5), but the ETS model with STLF seasonal adjustment was selected for the final forecast, as it appears more stable and interpretable in terms of predicting the behavior of the studied subscriber group. On the test data, ETS demonstrated acceptable performance, as evidenced by the baseline estimates:

MAE – 933 units,

MAPE – 22.4%.

References

[1] Коэн М.И. Прикладная линейная алгебра для исследователей данных / пер. с англ. А.В. Логунова. – М.: ДМК Пресс, 2023. – 328 с.

[2] Мыльников Л.А. Статистические методы интеллектуального анализа данных. – БХВ-Петербург, 2021. – 119 с.

[3] Data Science and Big Data Analytics. A Step by Step Guide to learn Data Science from Scratch with Python Machine Learning and Big Data / Andrew Park. – Published by Andrew Park, 2021. – 124 p.

МОДЕЛИ ПОВЕДЕНИЯ ПОТРЕБИТЕЛЕЙ НА РЫНКЕ МОБИЛЬНОЙ СВЯЗИ

Д.М. Рагель

к.э.н., доцент кафедры экономики Белорусского государственного университета информатики и радиоэлектроники

Аннотация. В статье рассматриваются подходы к прогнозированию потребительского поведения на технологическом рынке, с учетом основных факторов, которые описывают поведение абонентов. Рассматривается набор данных с базовым набором характеристик для рынка и применяются несколько моделей для прогнозирования на будущий период времени. На основании базовых показателей выбирается модель с лучшими оценками, которая является оптимальной по результатам с точки зрения решения поставленных задач.

Ключевые слова: Анализ данных, маркетинговые данные, большие данные, массив данных, прогнозирование, маркетинговая аналитика, анализ действий, рыночная активность, матрицы, скалярные значения, рынок мобильной связи, абоненты, прогнозная аналитика, рыночный прогноз.

УДК 330.45

ВЫРАВНИВАНИЕ ДИНАМИЧЕСКИХ РЯДОВ В АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ В ЦИФРОВОЙ ЭКОНОМИКЕ





Э. Е. Дадашев
Бакалавр направления подготовки «Экономика», профиль «Экономика и организация IT бизнеса», гр. БЭЭ2401, МТУСИ
ericandeev2016@yandex.ru

М. Д. Самакаева
Доцент кафедры цифровой экономики, управления и бизнес-технологий, МТУСИ
кандидат экономических наук, доцент
samakaeva-m@yandex.ru

Э. Е. Дадашев
Студент направления подготовки «Экономика», профиль «Экономика и организация IT бизнеса», гр. БЭЭ2401, Московского технического университета связи и информатики. Круг научных интересов: экономика, бизнес-технологии, аналитика, экономическая статистика.

М. Д. Самакаева
Окончила Оренбургский государственный педагогический университет. Область научных интересов: цифровая экономика, бизнес-технологии, аналитика, экономическая статистика, маркетинг, управление проектами.

Аннотация. В условиях ускоряющейся цифровой трансформации экономики ключевые бизнес-процессы, взаимодействие участников рынков и поведение потребителей всё чаще описываются большими массивами данных, формируемыми в режиме близком к реальному времени. Показатели электронной коммерции, использования облачных сервисов, цифровых финансовых услуг, телеком-инфраструктуры и онлайн-активности пользователей фиксируются в виде протяженных временных рядов, которые становятся основой аналитики и поддержки управленческих решений [3,6,7]. В данной статье рассматриваются модели и методы анализа и выравнивания динамических рядов при анализе больших данных в экономике. Применение динамических рядов при анализе данных позволяет выстраивать более точные прогнозы и формировать стратегии развития с более предсказуемым результатом.

Ключевые слова: цифровая трансформация, динамические ряды, Big Data, модели, динамические ряды, статистическое выравнивание, методы, модели, управленческие, градиентный бустинг, автоматическое прогнозирование, платформы, искусственный интеллект

Введение. Вопросы анализа и выравнивания динамических рядов в экономической статистике подробно рассматриваются в работах отечественных и зарубежных исследователей, посвящённых трендовым моделям, методам скользящего среднего, экспоненциального сглаживания и аналитическому выравниванию временных рядов [2,3,4,5,6,11]. Значительный вклад в развитие экономико-статистического аппарата внесли исследования, в которых обосновываются подходы к выбору функциональной формы тренда (линейной, полиномиальной, экспоненциальной) и оценивается их точность при анализе социально-экономических процессов.

Традиционные динамические ряды рассматриваются как результат совместного действия нескольких компонент: долгосрочного тренда, сезонной и циклической составляющей, а также случайных (нерегулярных) колебаний. Однако такая система имеет ряд преимуществ в классической модели, но несет в себе критические недостатки, в случае с Big Data что подробно описано в учебной и монографической литературе по экономической

статистике и эконометрике [2,3,4,5]. Эти и другие проблемы обусловили актуальность нашего исследования.

Цель – исследовать теоретические проблемы и разработать практические рекомендации по повышению эффективности использования выравнивания динамических рядов в Больших данных.

Из цели вытекают следующие задачи:

1. Исследовать теоретические аспекты использования динамических рядов в бизнес аналитике.
2. Выделить проблемы использования данных моделей.
3. Разработать практические рекомендации по выравниванию динамических моделей при использовании больших данных в цифровой статистике.

Данными проблемами занимались российские ученые, такие как: Капица С.П. [2], Курдюмов С.П., Малинецкий Г.Г., Ильясов Р.Х., Куразова Д.А. Кузовкова Т.А., Шаравов И. М. , Курицын Н. С. [3], Кумратова А.М., Плотников В.А. [4], Платунина Г.П. [6], Салютина Т.Ю. [6], М. Д. Самакаева [11], Дадашев Э. Е. [11], Минаков В.Ф. [4] и др.

Зарубежные авторы, занимающиеся проблемами математического моделирования и прогнозированием на основе Искусственного интеллекта: Синьюй Ху, Олчай Чирит, Танмай Бинайкия, Рамит Хора [13], Си Чен, Мин Го, Уюэ Шангуань [12] и др.

Объектом исследования выступают процессы цифровой экономики, описываемые временными рядами показателей электронной коммерции, цифровых услуг и онлайн-активности пользователей [3, 7, 9]. Предмет исследования - методы анализа, выравнивания и прогнозирования динамических рядов в условиях использования больших данных, а также их влияние на качество управленческих решений в цифровой среде.

В работе применены общенаучные методы исследования: анализ и синтез, индукция и дедукция, сравнение и обобщение, а также логический и системный подход к изучению социально-экономических процессов. Методологическую основу составляет экономико-статистический и эконометрический анализ динамических рядов, включающий аналитическое выравнивание тренда (линейные и нелинейные зависимости), методы скользящего среднего, экспоненциальное сглаживание и модели класса ARIMA/ARIMAX [12]. Для учёта особенностей больших данных и сложной структуры цифровых временных рядов дополнительно рассматриваются модели со структурными сдвигами и сменой режимов, а также методы машинного обучения, основанные на деревьях решений и градиентном бустинге [13].

Следует отметить, что процессы цифровой экономики, описываемые временными рядами показателей электронной коммерции, цифровых услуг и онлайн-активности пользователей являются современной основой использования больших данных в цифровой среде. Дополнительно значимым направлением цифровизации является развитие фриланс-платформ, формирующих новые модели занятости и требующих анализа временных рядов спроса и предложения на цифровом рынке труда [2,9,11] Предмет исследования - методы анализа, выравнивания и прогнозирования динамических рядов в условиях использования больших данных, а также их влияние на качество управленческих решений в цифровой среде.

Вопросы анализа и выравнивания динамических рядов в экономической статистике подробно рассматриваются в работах отечественных и зарубежных исследователей, посвящённых трендовым моделям, методам скользящего среднего, экспоненциального сглаживания и аналитическому выравниванию временных рядов [5]. Значительный вклад в развитие экономико-статистического аппарата внесли исследования, в которых обосновываются подходы к выбору функциональной формы тренда (линейной, полиномиальной, экспоненциальной) и оценивается их точность при анализе социально-экономических процессов [6,7]. В современных работах по цифровой экономике и Big Data акцент переносится на интеграцию классических методов временных рядов с моделями

машинного обучения и интеллектуального анализа данных, что позволяет обрабатывать высокочастотные и высокоразмерные массивы цифровой информации [8].

В современных работах по цифровой экономике и Big Data акцент переносится на интеграцию классических методов временных рядов с моделями машинного обучения и интеллектуального анализа данных, что позволяет обрабатывать высокочастотные и высокоразмерные массивы цифровой информации [8] (рисунок 1).

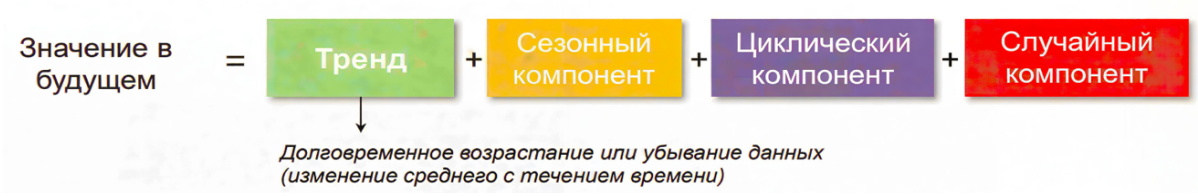


Рисунок 1. Классическая аддитивная модель временного ряда [13]

Классический подход опирается на явную структурную модель: тренд, сезонность и случайная компонента. В итоге малое число параметров позволяет четко интерпретировать результаты, но плохо масштабируется на тысячи и миллионы параллельных рядов, вынуждая строить модель под каждый ряд значений. Стационарность, фиксированная и регулярная сезонность, ограничивается относительно коротким временным горизонтом анализа. Сезонность и тренд предполагаются относительно стабильными, однако современные паттерны быстро меняются под влиянием алгоритмических рекомендаций и конкуренции, представлены в статистических исследованиях [1].

Классические модели с регрессорами (ARIMAX/SARIMAX и т.п.) умеют включать ограниченный набор внешних переменных, но когда мы имеет дело с большими данными мы говорим об интеграции в систему анализа факторов о сотне и тысяче поведенческих и контекстных факторов: клики, глубина просмотра, источники трафика, свойства пользователя, конкуренты, погода и т.п. [12]. Важно отметить, что вышеуказанные особенности классических моделей не предполагают отказа от них, а стимулируют создавать новые методы анализа данных, для их адаптации или совместного применения с более современными подходами [13]. Среди ключевых современных методик анализа динамических рядов, можно выделить: машинное обучение, глубинное обучение, Break-point и Regime-switching методы, градиентный бустинг и дерево решений [12].

Можно выделить следующие положительные и отрицательные стороны использования больших данных для формирования статистических моделей, представленных в таблице 1.

Таблица 1. Положительные и отрицательные стороны использования больших данных

Положительные стороны	Отрицательные стороны
1	2
1. Улучшение принятия решений: анализ больших объемов данных позволяет организациям принимать более обоснованные решения, основанные на фактических данных и тенденциях.	1. Конфиденциальность и безопасность: сбор и хранение больших объемов личных данных вызывает опасения по поводу конфиденциальности и безопасности информации. Утечки данных могут иметь серьезные последствия.
2.Персонализация: компании могут использовать большие данные для создания персонализированных предложений и улучшения клиентского опыта, что может привести к повышению лояльности клиентов.	2. Сложность анализа: обработка и анализ больших данных требуют значительных ресурсов и специализированных навыков, что может быть дорогостоящим для компаний.

Продолжение таблицы 1

1	2
3. Оптимизация процессов: анализ данных помогает выявить узкие места в бизнес-процессах и оптимизировать их, что может привести к снижению затрат и повышению эффективности.	3. Качество данных: большие объемы данных могут содержать ошибки или нерелевантную информацию, что может привести к неправильным выводам и решениям.
4.Инновации: большие данные могут способствовать разработке новых продуктов и услуг, основанных на потребностях и предпочтениях клиентов.	4. Зависимость от технологий: организации могут стать зависимыми от технологий и инструментов анализа данных, что может создать риски в случае их сбоя или устаревания.
5.Предсказательная аналитика: использование алгоритмов машинного обучения позволяет прогнозировать будущие тренды и поведение пользователей, что может быть полезно в различных отраслях.	

По материалам исследований авторов

Использование больших данных (Big Data) имеет как положительные, так и отрицательные стороны, важно,чтобы эти данные были объективными.

В цифровой экономике большие данные становятся основой не только для классических онлайн-сервисов, но и для новых моделей монетизации, таких как криптовалютные платформы, где поведенческие данные пользователей служат ключевым ресурсом для генерации дохода [2,6, 8,11].

В современных задачах анализа массивов транзакционных и поведенческих данных, ограничение классической модели компенсируется переходом к методам машинного обучения [3].

Основным преимуществом использования данного метода является особенность построения регрессии не на «тренде и сезонности», а на большом наборе признаков: лаги, скользящие окна, календарные признаки; а также таких внешних факторов, как маркетинг и поведение пользователей. Такой подход позволяет строить деревья решений, автоматически моделируя нелинейности и взаимодействия признаков, при таком методе структура не задается вручную, а обучается на основе данных, что позволяет модели построить многомерную зависимость и одновременно учитывать множество переменных факторов [10].

Нестабильность структуры ряда является одной из ключевых проблем в e-commerce и онлайн трафике, с которой классический подход не может справиться. Большое количество структурных сдвигов в тренде, изменения в системе рекомендаций и системе аукционов в рекламных кабинетах.

Сезонные паттерны не статичны и постоянно меняются, а динамическое распределение спроса и трафика вызывает медленное или резкое изменение статистических свойств во времени [9].

Для решения проблем со стационарностью, современные методы и подходы предлагают использовать модели со сменой режим Break-point (точки разрыва) модели ищут конкретные даты, в которые произошли резкие изменения трендов, такие как обвал или рост, разделяя тем самым весь динамический ряд значений на части и подбирая под каждый отдельный участок наиболее точные параметры для расчетов. Regime-switching (Смена режимов) модели представляют из себя более совершенную систему, полностью меняя алгоритмы анализа не только на основе временных периодов, но и учитывая помимо них ряд других внутренних и внешних параметров системы, таких как: покупательский спрос, дата, суточное время, текущая занятость сотрудников, погодные условия и прочие [8].

Список параметров изменяется в зависимости от того в какой сфере и отрасли применяется конкретная модель.

Помимо этого, необходимо учитывать изменение параметров и значений в реальном времени для этого в моделях используются онлайн-обновление и адаптивное взвешивание.

Технически это реализуется через одновременное использование двух моделей, где первая строится на исторических данных, а вторая только на недавних событиях. Финальные же прогнозы строятся на основе взвешенных сумм прогнозов обеих моделей, благодаря адаптивной системе весов для каждого параметра модели [12] (рисунок 2).

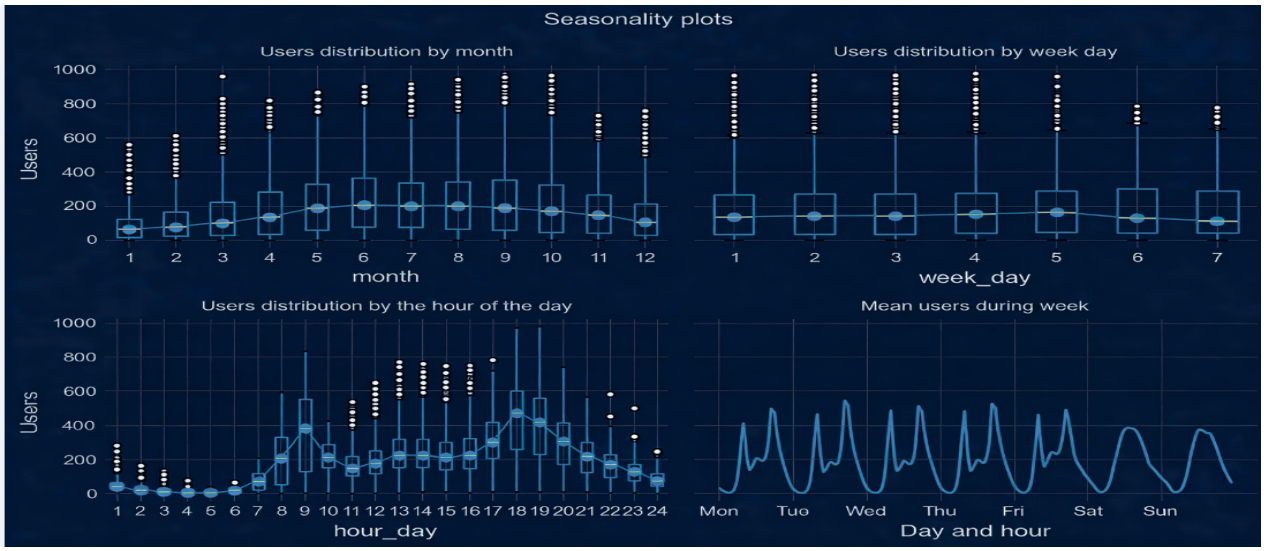


Рисунок 2. Пример сезонных графиков с данными для выявления сезонных закономерностей и трендов во временных рядах [7]

Решение третьей ключевой проблемы классического подхода в виде ограничений линейной регрессии достигается блягодаря градиентному бустингу и ансамблям деревьев. В данной ситуации мы говорим об огромном количестве внешних переменных, которые в свою очередь влияют на внутренние данные.

Дерево решений позволяет учитывать множество параметров таких как: количество и периодичность использования приложения пользователем, каким устройством он пользовался, в какое время он это делал, какая была погода в момент пользования приложением, по какой цене он приобрел товар или услугу и ряд других параметров, которые также меняются в зависимости от сферы и отрасли в которой проводится анализ.

Градиентный бустинг же позволяет выявлять нелинейную зависимость между всеми этими параметрами, построив ансамбль деревьев из множества деревьев решений.

Этот метод построен на основе правил по признаку, где модель последовательно задает вопросы по этим признакам, например: если клиент заказывает такси, а сейчас идет дождь, то он готов заплатить за такси большую цену.

Такая цепочка рассуждений и является одним из вариантов дерева решений, разумеется таких вариантов может быть множество и они могут полностью меняться в зависимости от набора признаков и заданных параметров.

Деревья решений обучаются независимо, а результаты работы модели представляют из себя усредненные решения набора деревьев [7,8] (рисунок 3).

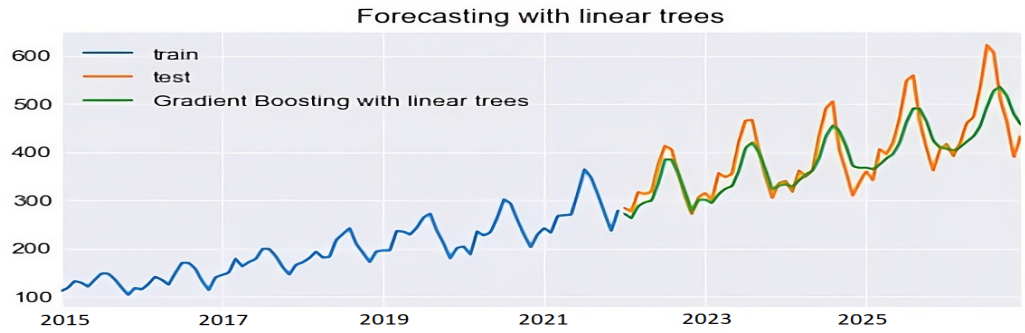


Рисунок 3. График, иллюстрирующий применение градиентного бустинга с линейными деревьями [7]

Заключение. Вышеописанные методы уже активно внедряются во множестве компаний и показывают положительные результаты от их применения [9].

Дочерняя компания Сбера «Сбер Бизнес Софт» разработала ML модель прогнозирования спроса на косметику для дистрибьютора «Градиент»; модель обучалась на данных о продажах, клиентах, складах, промо и остатках и дала +29% к качеству прогнозирования по сравнению с базовыми статистическими методами [9].

Wildberries обрабатывает более 1 млн заказов в день и строит Big Data архитектуру вокруг потоковой обработки событий, где продажи, остатки и поведение пользователей используются для онлайн аналитики и прогнозов спроса и логистики.

Во внешних сервисах для продавцов на Wildberries уже реализовано автоматическое прогнозирование поставок по временным рядам продаж и остатков по каждому товару и складу: система анализирует историю продаж, неснижаемый остаток и логистические задержки и выдает прогноз, какое количество товара и на какой именно склад везти, чтобы избежать дефицита товара на складе [10].

ВТБ протестировал платформу Smart Data Hub для аналитики цифрового маркетинга: система анализировала траектории пользователей на сайте банка по данным Яндекс.Метрики за 11 месяцев (временные ряды визитов, кликов, конверсий по страницам продуктов).

За 3 месяца работы пилотного проекта было создано более 40 гипотез, где более 40-50% из них были признаны практически значимыми, а потенциал роста конверсий оценен более чем в 5% [12].

Стриминговый сервис Netflix применяет модели временных рядов и последовательностей для прогнозирования нагрузки, персонализации и оценки удержания, комбинируя историю просмотров и поведение пользователей с контентными признаками [7].

В Uber использование градиентного бустинга и нейросетей для прогноза спроса на поездки снижало среднеквадратичную ошибку прогноза прогноза почти вдвое по сравнению с одиночными деревьями и классическими моделями, что напрямую влияет на динамическое ценообразование и балансировку предложения машин [13].

Сегодня формируется новая парадигма анализа динамических рядов, в которой классические статистические модели рассматриваются не как устаревший инструмент, а как важный базовый слой, дополняемый глобальными, нелинейными и высокоразмерными методами Big Data-аналитики.

Список литературы

[1] Стратегия развития государственной статистики до 2030 года Федеральная служба государственной статистики (Росстат), [Электронный ресурс]. – URL: https://rosstat.gov.ru/storage/mediabank/Strategiia_2030_bol.pdf (Дата обращения 28.02.2026г.)

[2] Капица С.П., Курдюмов С.П., Малинецкий Г.Г. (1997) Синергетика и прогнозы будущего, М.: Наука.1997 г.

[3] Кузовкова Т.А., Особенности и перспективы развития цифровых услуг и сервисов инфокоммуникационных компаний / Т.А. Кузовкова, И. М. Шаравов, Н. С. Курицын // журнал «Экономика и качество систем связи»№ 3/2024 [Электронный ресурс]. – URL: https://journal-ekss.ru/wp-content/uploads/2024/10/27-39-1.pdf

[4] Кумратова А.М. Применение методов нелинейной динамики и машинного обучения для прогнозирования экономических волатильных процессов / А.М. Кумратова В.А. Плотников // Электронный журнал π-Economy. 2024, Том 17, № 3. С. 81–95. [Электронный ресурс]. – URL: https://economy.spbstu.ru/userfiles/files/articles/2024/3-2024/06_Kumratova,-Plotnikov.pdf 25 (Дата обращения 28.02.2026г.)

[5] Перепелица В.А., Попова Е.В. (2001) Многокритериальный подход к моделированию финансово-экономических рисков. Известия вузов. Северо-Кавказский регион. Естественные науки, 4, 37–41.

[6] Платунина Г.П. Монетизация криптосайтов как новый способ заработка в условиях цифровой экономики / Г.П. Платунина, А.С. Старовойтова [Электронный ресурс]. – URL: https://cyberleninka.ru/article/n/monetizatsiya-kriptosaytov-kak-novyy-sposob-zarabotka-v-usloviyah-tsifrovoy-ekonomiki/viewer (Дата обращения 26.11.2025)

[7] Прогнозирование товарооборота арендаторов торгово- развлекательного центра // Электронный журнал Cyberleninka [Электронный ресурс]. – URL: https://cyberleninka.ru/article/n/prognozirovanie-tovarooborota-arendatorov-torgovo-razvlekatelnogo-tsentra (Дата обращения 28.02.2026г.)

[8] Способы прогнозирования курса валют на основе моделей экспоненциального сглаживания и Хольта // Электронный журнал Cyberleninka [Электронный ресурс]. – URL: https://cyberleninka.ru/article/n/sposoby-prognozirovaniya-kursa-valyut-na-osnove-modeley-eksponentsialnogo-sglazhivaniya-i-holta (Дата обращения 28.02.2026г.)

[9] ИИ-модель Сбербанка прогнозирует спрос на косметику на треть лучше статистических моделей // Электронный журнал CNews от 12.09.2024г [Электронный ресурс]. – URL: https://axenix.cnews.ru/news/line/2024-09-12_ii-model_sberbanka_prognoziruet (Дата обращения 28.02.2026г.)

[10] Ильясов Р.Х., Куразова Д.А. (2017) Прогнозирование конъюнктуры финансового рынка и оценка его влияния на развитие промышленности. Финансы и кредит, 23 (43), 2575–2591. DOI: https://doi.org/10.24891/fc.23.43.2575

[11] Самакаева, М. Д. Фриланс платформы как основа развития рынка труда в цифровой экономике / М. Д. Самакаева, Е. Д. Критина, Р. Ф. Мартынова // Первый экономический журнал. – 2024. – № 12(354). – С. 69-78. – DOI 10.58551/20728115_2024_12_69. – EDN GPOAXT. (Дата обращения 28.02.2026г.)

[12]Си Чен Оценка влияния облачных вычислений на показатели деятельности компаний: эмпирическое исследование компаний, акции которых котируются на бирже / Си Чен, Мин Го, Уюэ Шангуань // Электронный журнал Journal of Econometrics (S0378720622000155) [Электронный ресурс]. – URL: https://www.sciencedirect.com/science/article/pii/S0378720622000155(Дата обращения 28.02.2026г.)

[13] Синьюй Ху DeepETA: Как Uber прогнозирует время прибытия с помощью глубокого обучения / Синьюй Ху, Олчай Чирит, Танмай Бинайкия, Рамит Хора // Блог UBER от10.02.2022г.[Электронный ресурс]. – URL: https://www.uber.com/blog/deepeta-how-uber-predicts-arrival-times/ (дата обращения: 28.02.2026).

ALIGNMENT OF DYNAMIC SERIES IN BIG DATA ANALYTICS IN THE DIGITAL ECONOMY

E. E. Dadashev
Bachelor's degree in Economics, Profile «Economics and organization of IT business», BEE2401, MTUCI
ericandeev2016@yandex.ru

M. D. Samakaeva
Associate Professor of the Department of Digital Economics, Management and Business Technology, MTUCI
Candidate of Economic Sciences, Associate Professor
samakaeva-m@yandex.ru

Abstract. In the context of the accelerating digital transformation of the economy, key business processes, the interaction of market participants and consumer behavior are increasingly described by large amounts of data generated in close to real time. Indicators of e-commerce, the use of cloud services, digital financial services, telecom infrastructure, and online user activity are recorded as long time series that become the basis for analytics and management decision support. This article discusses models and methods of dynamic series analysis and alignment in the analysis of big data in economics. The use of dynamic series in data analysis allows for more accurate forecasts and development strategies with more predictable results.

Keywords: digital transformation, dynamic series, Big Data, models, dynamic series, statistical alignment, methods, models, management, gradient boosting, automatic forecasting, platforms, artificial intelligence.

УДК 338.24:004.9

ИНТЕГРАЦИЯ КЛАССИЧЕСКИХ МЕТОДОВ СТРАТЕГИЧЕСКОГО АНАЛИЗА И МОДЕЛЕЙ BIG DATA В СИСТЕМЕ ПРИНЯТИЯ УПРАВЛЕНЧЕСКИХ РЕШЕНИЙ



А.М. Найда
Кандидат экономических наук, доцент, заведующая кафедрой, кафедра бизнес-аналитики, первый проректор УВО «Университет управления «ТИСБИ»
anajda1979@gmail.com



Э.Д. Кочешкова
Руководитель лаборатории бизнес-аналитики и предиктивного анализа больших данных, ассистент кафедры бизнес-аналитики УВО «Университет управления «ТИСБИ»
Kocheshkovaella0902@gmail.com

А.М. Найда

Окончила Татарский институт содействия бизнесу. Область научных интересов связана с методами и алгоритмами принятия управленческих решений, инструментарием бизнес-аналитики, формированием экосистемы предпринимательского образования.

Э.Д. Кочешкова

Окончила Учреждение высшего образования «Университет управления «ТИСБИ». Область научных интересов включает развитие теории и методологии бизнес-аналитики, предиктивного анализа больших данных, алгоритмической поддержки управленческих решений, исследование трансформации управленческих моделей в условиях цифровой экономики и внедрение интеллектуальных аналитических систем в практику организаций.

Аннотация. В статье рассматривается трансформация стратегического анализа в условиях цифровой экономики. Проведен сравнительный анализ классических методов стратегического управления (SWOT, PEST) и аналитических моделей, основанных на использовании технологий Big Data и Advanced Analytics. Выявлены ограничения экспертно-ориентированных инструментов и обоснованы преимущества data-driven подхода к формированию управленческих решений. Разработана концептуальная интеграционная модель, обеспечивающая сочетание структурированного стратегического анализа и алгоритмической обработки больших данных. Обоснована необходимость гибридного подхода к принятию управленческих решений в условиях высокой неопределенности и цифровой трансформации экономики.

Ключевые слова: Big Data, стратегический анализ, SWOT-анализ, PEST-анализ, управленческие решения, бизнес-аналитика, цифровая экономика, Advanced Analytics, система поддержки принятия решений.

Введение. Цифровая трансформация экономики сопровождается стремительным ростом объема данных, увеличением скорости их обработки и усложнением конкурентной среды. Организации функционируют в условиях высокой нестабильности рынков, макроэкономических факторов и ускоренного технологического развития. В подобных условиях возрастает значимость стратегических управленческих решений, от качества которых зависит устойчивость и долгосрочная конкурентоспособность компаний.

Традиционные инструменты стратегического анализа, сформированные в рамках индустриальной парадигмы управления, долгое время обеспечивали структурирование информации о внутренней и внешней среде организации. Однако современная цифровая среда характеризуется динамичностью, многомерностью и информационной избыточностью, что требует пересмотра методологических основ стратегического анализа. В этой связи

актуализируется необходимость интеграции классических управленческих методов с инструментами Big Data и Advanced Analytics.

Теоретические основы стратегического анализа и применение Big Data в системе стратегического управления. Стратегический анализ традиционно рассматривается как фундаментальный этап стратегического управления, обеспечивающий выявление ключевых факторов внешней и внутренней среды организации, формирование альтернатив развития и выбор долгосрочных направлений деятельности. В классической теории стратегического менеджмента стратегический анализ формировался в рамках концепций конкурентного позиционирования, ресурсной теории фирмы и модели отраслевой конкуренции. Существенный вклад в развитие данных подходов внес М. Портер, обосновавший необходимость системной оценки конкурентных сил и стратегических позиций организации [1]. Согласно его концепции, устойчивое конкурентное преимущество формируется на основе анализа структуры отрасли и рационального стратегического выбора.

В дальнейшем развитие стратегического анализа было связано с переходом от исключительно внешней ориентации к исследованию внутренних ресурсов и компетенций организации. Ресурсная концепция фирмы акцентировала внимание на уникальности внутренних активов как источнике устойчивости и долгосрочной эффективности. Существенное значение приобрели инструменты системной диагностики, включая SWOT-анализ и PEST-анализ, которые обеспечивают структурирование информации и формирование стратегических гипотез.

SWOT-анализ позволяет выявить сильные и слабые стороны организации, а также возможности и угрозы внешней среды. Его ценность заключается в интеграции внутренних и внешних факторов в единую логическую модель. Однако данный инструмент преимущественно опирается на экспертные оценки и качественную интерпретацию данных, что ограничивает объективность результатов и повышает зависимость от профессионального опыта аналитика. PEST-анализ, ориентированный на исследование политических, экономических, социальных и технологических факторов, также носит экспертный характер и предполагает периодическую оценку макросреды. В условиях ускоренной цифровизации и высокой динамики рынков периодичность такого анализа может не обеспечивать достаточной оперативности.

Развитие информационных технологий и цифровых платформ привело к формированию новой управленческой парадигмы, основанной на данных. В работах Т. Давенпорта и Дж. Харриса подчеркивается, что конкурентоспособность организаций все в большей степени определяется способностью использовать аналитику в качестве стратегического ресурса [2]. Переход к data-driven управлению предполагает систематическое использование больших массивов данных для поддержки стратегических решений.

Технологии Big Data характеризуются объемом, скоростью обработки и разнообразием источников информации. Они позволяют интегрировать транзакционные данные, цифровые следы пользователей, данные социальных сетей, макроэкономические индикаторы и отраслевую статистику в единую аналитическую среду. В рамках концепции Big Data стратегический анализ приобретает непрерывный характер и трансформируется из периодического диагностического инструмента в систему постоянного мониторинга и прогнозирования. Advanced Analytics включает методы интеллектуального анализа данных, машинного обучения, нейронных сетей и алгоритмов предиктивного моделирования. Как отмечают Ф. Провост и Т. Фоссетт, аналитика данных позволяет выявлять скрытые закономерности и формировать вероятностные прогнозы, недоступные при использовании традиционных методов анализа [3]. В стратегическом контексте это означает возможность оценки сценариев развития, прогнозирования рыночной динамики и оптимизации ресурсного распределения.

Существенным этапом эволюции стратегического управления стало формирование концепции аналитической конкуренции, согласно которой компании достигают лидерства за

счет системного использования данных в принятии решений [2]. Э. Бриньольфссон и Э. Макафи подчеркивают, что цифровая экономика формирует новую структуру конкурентных преимуществ, где ключевую роль играют алгоритмы и интеллектуальные системы обработки информации [4].

Таким образом, стратегический анализ в цифровой экономике трансформируется из преимущественно экспертной процедуры в гибридную систему, сочетающую управленческое мышление и алгоритмическую обработку данных. Однако полная алгоритмизация стратегического управления представляется методологически некорректной, поскольку стратегические решения включают ценностные, институциональные и социальные аспекты, не поддающиеся полной формализации. Следовательно, наиболее перспективным является интеграционный подход, сочетающий классические инструменты и технологии Big Data.

Сравнительный анализ классических и data-driven подходов. Сопоставление классических методов стратегического анализа и data-driven моделей позволяет выявить не только различия в инструментарии, но и различия в логике управленческого мышления и методологической основе принятия решений. Традиционные методы стратегического анализа формировались в условиях ограниченности информации и относительной стабильности рыночной среды. Их основная задача заключалась в структурировании факторов среды, выявлении стратегических альтернатив и формировании управленческих гипотез на основе экспертной интерпретации данных. Концептуальные основания данного подхода заложены в работах М. Портера [1], Р. Каплана и Д. Нортона [5], а также получили развитие в отечественной управленческой школе, где стратегический анализ рассматривается как системный инструмент диагностики и прогнозирования развития организации [7].

Классические методы (SWOT, PEST, анализ отраслевой структуры, матричные модели стратегического позиционирования) ориентированы преимущественно на качественную оценку факторов. Их сильной стороной является логическая целостность и универсальность применения, однако они характеризуются высокой зависимостью от компетенций эксперта и субъективности интерпретации. Как отмечает М. К. Алимурадов, стратегическое мышление предполагает формирование видения и целеполагания, которые не могут быть сведены исключительно к формальным расчетам [9]. Данный тезис подчеркивает ценность классических инструментов в формировании стратегического контекста.

В противоположность этому data-driven подход базируется на использовании больших массивов данных и математических моделей обработки информации. В рамках концепции аналитической конкуренции организация рассматривает данные как стратегический ресурс [2]. Предиктивная аналитика и методы машинного обучения позволяют переходить от ретроспективной диагностики к вероятностному прогнозированию развития событий [3]. В российской научной литературе вопросы цифровой трансформации управления и внедрения аналитических платформ активно исследуются в работах О. И. Донцовой, где подчеркивается необходимость интеграции интеллектуальных систем в процессы стратегического планирования [10]. В отличие от традиционных методов, data-driven модели обеспечивают непрерывность мониторинга, масштабируемость анализа и возможность обработки неструктурированных данных. Они позволяют выявлять скрытые закономерности, строить сценарные прогнозы и оценивать чувствительность стратегических решений к изменению внешних параметров. При этом возникает новая управленческая логика – переход от экспертно-гипотетического подхода к вероятностно-моделирующему.

Для систематизации различий между рассматриваемыми подходами в таблице 1 представлена их расширенная сравнительная характеристика по ключевым методологическим и управленческим параметрам (таблица 1).

Таблица 1. Экспертная сравнительная характеристика классических и data-driven подходов в стратегическом анализе

№	Критерий сравнения	Классические методы стратегического анализа	Data-driven подход и модели Big Data
1	2	3	4
1	Методологическая основа	Концепции конкурентного позиционирования, ресурсная теория, экспертная диагностика	Теория вероятностей, математическая статистика, машинное обучение, интеллектуальный анализ данных
2	Тип используемых данных	Агрегированные статистические данные, экспертные оценки	Структурированные и неструктурированные большие данные, цифровые следы, потоковые данные
3	Периодичность анализа	Периодический (стратегические сессии, годовое планирование)	Непрерывный мониторинг в режиме реального времени
4	Характер прогнозирования	Качественные сценарии, стратегические альтернативы	Количественные вероятностные прогнозы, предиктивные модели
5	Роль эксперта	Доминирующая, формирование гипотез и интерпретация факторов	Координирующая, постановка задач моделирования и интерпретация результатов
6	Уровень автоматизации	Низкий или умеренный	Высокий, использование алгоритмов и аналитических платформ
7	Требования к инфраструктуре	Минимальные, возможна реализация без сложных ИТ-систем	Развитая цифровая инфраструктура, хранилища данных, аналитические платформы
8	Гибкость и адаптивность	Ограниченная скоростью пересмотра стратегии	Высокая адаптивность за счет динамического обновления моделей
9	Основные риски	Субъективность, неполнота информации	Ошибки алгоритмов, предвзятость данных, высокая стоимость внедрения

Как следует из таблицы 1, различия между подходами носят системный характер и затрагивают не только инструментарий, но и философию управления. Классические методы формируют стратегическое видение и концептуальные рамки анализа, тогда как data-driven модели обеспечивают количественную верификацию гипотез и оперативную корректировку решений. Особое значение имеет различие в горизонте прогнозирования. Традиционные методы ориентированы на формирование стратегических направлений без строгой количественной оценки вероятности реализации альтернатив. Предиктивная аналитика, напротив, позволяет рассчитывать вероятностные сценарии и оценивать риски на основе исторических и потоковых данных. Это особенно актуально в условиях цифровой экономики, где скорость изменений требует постоянной адаптации стратегий [4].

В то же время необходимо подчеркнуть, что алгоритмические модели не способны полностью заменить стратегическое мышление. Как отмечается в исследованиях отечественных ученых, цифровая трансформация управления должна сопровождаться развитием управленческих компетенций и институциональной зрелости организаций [8]. Стратегическое решение включает ценностные ориентиры, миссию и долгосрочные приоритеты, которые выходят за пределы формализованных расчетов. Таким образом, сравнительный анализ демонстрирует комплементарный характер классических и data-driven подходов. Их интеграция позволяет объединить концептуальную глубину стратегического анализа и количественную точность предиктивных моделей. В результате формируется гибридная система стратегического управления, сочетающая экспертную интерпретацию и алгоритмическую обработку данных, что повышает устойчивость организации в условиях цифровой трансформации.

Концептуальная интеграционная модель. С учетом выявленных преимуществ и ограничений предлагается концептуальная модель интеграции классических методов стратегического анализа и инструментов Big Data, представленная на рисунке 1.

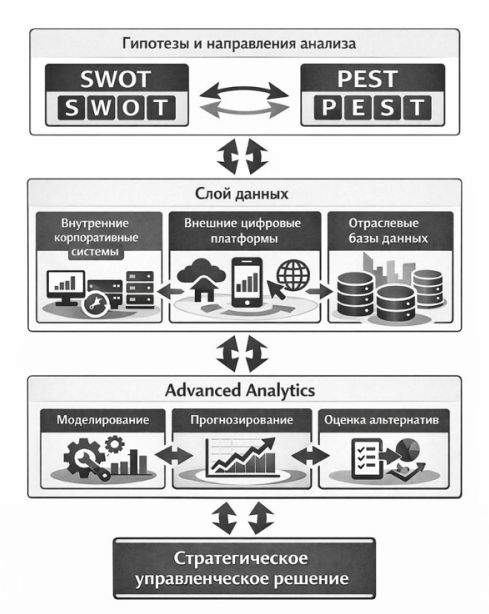


Рисунок 1. Интеграционная модель стратегического анализа

Первый уровень включает традиционные инструменты SWOT и PEST, формирующие гипотезы и направления анализа.

Второй уровень представляет слой данных (Data Layer), включающий внутренние корпоративные системы, внешние цифровые платформы и отраслевые базы данных.

Третий уровень включает инструменты Advanced Analytics, обеспечивающие моделирование, прогнозирование и оценку альтернатив. Итогом взаимодействия уровней является формирование стратегического управленческого решения.

Между уровнями предусмотрена обратная связь, обеспечивающая корректировку гипотез и адаптацию стратегии.

Реализация интегрированной модели требует формирования гибридной системы поддержки принятия решений, архитектура которой представлена на рисунке 2.



Рисунок 2. Архитектура гибридной системы поддержки стратегических решений

Схема включает источники данных (ERP, CRM, внешние базы данных, социальные сети), хранилище данных, аналитический модуль, экспертный модуль и интерфейс стратегического управления. Информационные потоки обеспечивают передачу данных от

источников к аналитическим инструментам и далее к управленческому уровню. Гибридная система обеспечивает синтез экспертного анализа и алгоритмической обработки данных, что способствует повышению качества стратегических решений. Несмотря на преимущества гибридной модели, ее реализация сопряжена с рядом рисков. К ним относятся проблемы качества данных, алгоритмическая предвзятость, высокая стоимость внедрения цифровой инфраструктуры и дефицит аналитических компетенций. Существенное значение имеет обеспечение информационной безопасности и соблюдение нормативных требований в сфере обработки персональных данных. Кроме того, существует риск избыточной автоматизации управленческих процессов, что может привести к снижению роли стратегического мышления и интуитивной оценки.

Пример практического применения интеграционной модели стратегического анализа. Для иллюстрации возможностей интеграционного подхода рассмотрим условный кейс промышленного предприятия среднего масштаба, работающего на рынке машиностроительной продукции с годовой выручкой около 3,5 млрд руб. и долей рынка 7–9 %. В условиях снижения маржинальности на 4–6 п.п. за последние два года и роста логистических издержек на 12–15 % руководство инициирует пересмотр стратегических ориентиров. На первом этапе проводится классический стратегический анализ. В рамках PEST-анализа выявляется влияние макроэкономических факторов: рост стоимости заемного капитала до 14–16% годовых, увеличение сроков поставок импортных комплектующих на 25–30%, а также ускоренная цифровизация отрасли. SWOT-анализ позволяет структурировать внутренние параметры: к сильным сторонам относятся высокая квалификация инженерного персонала и стабильный портфель долгосрочных контрактов (около 60 % выручки), к слабым – низкий уровень автоматизации складской логистики и зависимость от двух ключевых поставщиков (более 55% закупок). В числе возможностей выделяется рост внутреннего спроса на импортозамещающую продукцию (прогнозируемый прирост рынка 8–10% в год), среди угроз – волатильность цен на сырьё и транспортные ограничения. На основе экспертной оценки формируются три стратегические альтернативы: диверсификация поставщиков, цифровизация производственно-логистических процессов и переход к сервисной модели сопровождения продукции. Для перехода от качественной оценки к количественной верификации используется аналитический уровень интеграционной модели. В систему загружаются данные ERP и CRM за последние 5 лет (более 1,2 млн транзакций), показатели складских остатков (около 15 тыс. номенклатурных позиций), данные о сроках поставок, а также отраслевые и макроэкономические индикаторы. На основе регрессионного анализа и алгоритмов градиентного бустинга строится модель прогнозирования спроса с коэффициентом детерминации $R^2 = 0{,}81$ и среднеквадратичной ошибкой прогноза 6,4 % [11]. Сценарный анализ проводится на горизонте трёх лет.

Моделирование диверсификации поставщиков показывает снижение вероятности логистических сбоев с 18 % до 9 %, однако при этом средняя закупочная цена может увеличиться на 3,2 %, что приведёт к росту себестоимости продукции примерно на 1,4 %. Чистый эффект при прочих равных условиях выражается в снижении операционного риска, но умеренном ухудшении краткосрочной рентабельности. Сценарий цифровизации производственно-логистических процессов, включающий внедрение системы предиктивного управления запасами, демонстрирует потенциальное сокращение среднего уровня складских остатков на 17–19 % и снижение оборачиваемости запасов с 124 до 102 дней. Это позволяет высвободить до 210 млн руб. оборотного капитала и снизить складские издержки на 8–10 % в год. Расчёт NPV проекта при ставке дисконтирования 15 % показывает положительное значение в размере 146 млн руб. на горизонте трёх лет [11].

Переход к сервисной модели (расширение постгарантийного обслуживания и внедрение цифрового мониторинга оборудования) позволяет спрогнозировать рост повторных продаж на 11–13 % и увеличение доли сервисной выручки с 6 % до 15 % общей выручки компании. При этом маржинальность сервисного направления превышает производственную на 8–10 п.п., что

обеспечивает рост совокупной операционной прибыли на 6–8 % к третьему году реализации стратегии. Интеграционная модель позволяет провести сравнительную оценку альтернатив с учётом чувствительности к ключевым параметрам: изменению курса валют, колебаниям спроса и уровню процентной ставки. Применение интеграционного подхода обеспечивает переход от экспертного сопоставления альтернатив к их количественной оценке на основе данных. Классические методы формируют структуру стратегических гипотез и выявляют направления развития, тогда как аналитические инструменты позволяют рассчитать экономические эффекты, вероятностные риски и диапазон возможных отклонений показателей. В результате повышается обоснованность управленческого выбора и снижается уровень стратегической неопределённости.

Заключение. Проведенное исследование показало, что интеграция классических методов стратегического анализа и моделей Big Data является объективной необходимостью в условиях цифровой экономики. Традиционные инструменты обеспечивают структурирование стратегического мышления и формирование гипотез, тогда как аналитические модели обеспечивают количественную проверку предположений и динамическую адаптацию решений. Предложенная интеграционная модель позволяет повысить обоснованность, устойчивость и адаптивность стратегических решений. Перспективы дальнейших исследований связаны с разработкой методик оценки эффективности гибридных систем стратегического анализа и формированием критериев цифровой зрелости управленческих процессов.

Список литературы

[1] Porter, M. E. Competitive Strategy: Techniques for Analyzing Industries and Competitors / M. E. Porter. – New York: Free Press, 2008.

[2] Davenport, T. H. Competing on Analytics: The New Science of Winning / T. H. Davenport, J. G. Harris. – Boston: Harvard Business School Press, 2017.

[3] Provost, F. Data Science for Business / F. Provost, T. Fawcett. – Sebastopol: O'Reilly Media, 2013.

[4] Brynjolfsson, E. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies / E. Brynjolfsson, A. McAfee. – New York: W.W. Norton & Company, 2014.

[5] Kaplan, R. S. Strategy-Focused Organization / R. S. Kaplan, D. P. Norton. – Boston: Harvard Business School Press, 2001. – DOI: 10.1108/sl.2001.26129cab.002.

[6] Brynjolfsson, E. The Rapid Adoption of Data-Driven Decision-Making / E. Brynjolfsson, K. McElheran // American Economic Review. – 2016. – № 106(5). – PP. 133-139. – DOI: 10.1257/aer.p20161016.

[7] Sharda, R. Business Intelligence, Analytics, and Data Science: A Managerial Perspective / R. Sharda, D. Delen, E. Turban. – 4th ed. – Harlow: Pearson, 2018.

[8] Виханский, О. С. Стратегическое управление / О. С. Виханский, А. И. Наумов. – 6-е изд. – М.: Гардарики, 2019.

[9] Алимурадов, М. К. Стратегирование – новая область профессиональных знаний / М. К. Алимурадов, Л. И. Власюк // Государственная служба. – 2017. – № 11. – С. 154-159. – DOI: 10.22394/1726-1139-2017-11-154-159.

[10] Донцова, О. И. Цифровая трансформация системы управления промышленными кластерами / О. И. Донцова // Вопросы инновационной экономики. – 2022. – № 12(2). – С. 114836. – DOI: 10.18334/vinec.12.2.114836.

[11] Zhang, S. A Big Data-Driven Approach to Financial Analysis and Decision Support System Design / S. Zhang // Informatica. – 2025. – Vol. 49, No. 11. – PP. 29–44. – DOI: 10.31449/inf.v49i11.7065.

Авторский вклад

Найда Анна Михайловна – постановка задачи исследования, концептуальное и методологическое наставничество, научное консультирование при проведении теоретического анализа, общее научное руководство созданием интеграционной модели.

Кочешкова Элла Дмитриевна – постановка задачи исследования, проведение теоретического анализа, разработка сравнительной характеристики методов стратегического анализа и моделей Big Data, создание интеграционной модели.

INTEGRATION OF CLASSICAL STRATEGIC ANALYSIS METHODS AND BIG DATA MODELS IN THE MANAGEMENT DECISION-MAKING SYSTEM

A.M. Nayda
Candidate of Economic Sciences, Associate Professor, Head of the Department, Department of Business Analytics, University of Management «TISBI»

E.D. Kocheshkova
Head of the Laboratory of Business Analytics and Predictive Analysis of Big Data, University of Management «TISBI»

Abstract. The paper examines the transformation of strategic analysis in the digital economy. A comparative analysis of classical strategic management methods (SWOT, PEST) and Big Data-based analytical models is conducted. A conceptual integration model combining expert-based approaches and Advanced Analytics tools is proposed. The hybrid approach enhances the validity and adaptability of strategic decisions in a turbulent environment.

Keywords: Big Data, strategic analysis, SWOT, PEST, management decision-making, business analytics, digital economy, Advanced Analytics.

УДК 004.9:330.34

ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ АВТОМАТИЗИРОВАННОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ ИНВЕСТИЦИОННЫХ ПРОЕКТОВ



Н.К. Крючков
Ведущий FE разработчик ООО «СберОбразование» nikita.kryuchkov78@gmail.com



А.В. Воробей
Асситент кафедры инженерной психологии и эргономики БГУИР, магистр технических наук a.vorobey@bsuir.by



И.В. Андриалович
Старший преподаватель кафедры инженерной психологии и эргономики БГУИР, магистр техники и технологии andryinna@bsuir.by

Н.К. Крючков

Ведущий FE разработчик ООО «СберОбразование». Выпускник УО «БГУИР». Область научных интересов связана с исследованиями в области информационных технологий в образовательной индустрии.

А.В. Воробей

Окончила Белорусский государственный педагогический университет имени М. Танка» по специальности Биология и химия. В 2020 году окончила магистратуру БГУИР по специальности «Психология труда, инженерная психология, эргономика», выполняет исследования в рамках магистерской диссертации на тему «Система обеспечения безопасности труда при воздействии инфракрасного излучения на физиологические параметры человека». Окончила аспирантуру БГУИР по специальности «Психология труда, инженерная психология, эргономика».

И.В. Андриалович

В 2014 году окончила Минский государственный высший радиотехнический колледж по специальности «Программное обеспечение в области информационных технологий». В 2017 году – Белорусский государственный университет информатики и радиоэлектроники по специальности «Профессиональное обучение (информатика)». В 2019 году – получение степени магистра по специальности «Управление безопасностью производственных процессов». В 2025 году окончила аспирантуру БГУИР по специальности «Психология труда, инженерная психология, эргономика».

Аннотация. В статье рассматривается разработка и внедрение программного обеспечения для автоматизированной оценки эффективности инвестиционных проектов. Акцентируется внимание на важности качественной оценки инвестиционных решений в условиях динамичного рынка и растущей конкуренции. Описываются основные функциональные возможности программного обеспечения, включая расчет ключевых показателей эффективности, таких как чистая приведенная стоимость (NPV), внутренняя норма доходности (IRR) и срок окупаемости. Также рассматриваются преимущества автоматизации процесса оценки, такие как снижение человеческого фактора, повышение точности расчетов и ускорение принятия решений. Статья будет полезна специалистам в области финансового анализа, инвестирования и управления проектами, а также разработчикам программных решений для бизнеса.

Ключевые слова: программное обеспечение, инвестиционные проекты, эффективность инвестиций, ROI (возврат инвестиций), NPV (чистая приведенная стоимость), внутренняя норма доходности (IRR).

Введение. Оценка инвестиционных проектов представляет собой ключевой инструмент для принятия решений в условиях неопределенности и конкурентного окружения. Компании, обладающие надежной и обоснованной оценкой своих инициатив, становятся более привлекательными для инвесторов, что позволяет им привлекать необходимое финансирование на более выгодных условиях. Кроме того, такая оценка способствует оптимальному распределению ресурсов, направляя их на наиболее перспективные и прибыльные направления, что особенно важно в условиях ограниченности ресурсов и необходимости определения приоритетов [1].

Еще одной важной функцией оценки является возможность сравнения альтернативных проектов, что помогает выбрать наиболее эффективное предложение и повысить общую результативность бизнеса. Оценка инвестиционных проектов также является основой для долгосрочного стратегического планирования, позволяя компаниям адаптироваться к изменениям на рынке и достигать устойчивого роста. Таким образом, значимость оценки инвестиционных проектов обусловлена её критической ролью в снижении рисков, увеличении инвестиционной привлекательности и обеспечении эффективного управления ресурсами.

Анализ программных решений. Автоматизация оценки эффективности инвестиционных проектов представляет собой ключевую предметную область в управлении финансами и инвестициями, позволяющую существенно повысить точность и оперативность анализа. В условиях быстро меняющейся экономики и высоких требований к скорости принятия решений, автоматизация становится необходимым инструментом для компаний, стремящихся оптимизировать свои инвестиционные стратегии.

Основные цели автоматизации заключаются в усовершенствовании процесса оценки проектной эффективности, минимизации временных и человеческих затрат, а также повышении аналитической точности. Важно отметить, что инвестиционные проекты характеризуются множеством переменных и неопределённостей, поэтому создание автоматизированной системы оценки позволяет использовать сложные математические модели и алгоритмы, которые традиционным способом было бы сложно реализовать.

Ключевыми задачами автоматизации являются: оценка денежных потоков проекта, комплексный анализ связанных рисков, а также расчет основных показателей эффективности, таких как чистая приведенная стоимость (NPV), внутренняя норма доходности (IRR) и срок окупаемости. Использование современных информатизированных решений дает возможность интегрировать различные источники данных, включая финансовые отчёты, рыночные исследования и экономические прогнозы, что, в свою очередь, помогает строить более точные и обоснованные финансовые модели [1].

Кроме того, автоматизация процесса анализа повышает прозрачность и уменьшая вероятность ошибок, связанных с человеческим фактором, что критически важно для принятия обоснованных инвестиционных решений. Использование специализированного программного обеспечения обеспечивает четкое документирование всех этапов оценки проекта, что улучшает возможность аудита и анализа предыдущего опыта.

При выборе программного обеспечения для автоматизации оценки эффективности инвестиционных проектов важно учитывать ряд факторов, включая функционал, удобство использования, стоимость и возможность интеграции с другими системами.

Важно также учитывать, что многие инструменты могут быть интегрированы друг с другом, что позволяет создать комплексную систему управления проектами с учётом оценки их эффективности [2].

Разработка программного обеспечения. Разрабатываемое программное средство предназначено для автоматизированной поддержки процессов оценки эффективности инвестиционных проектов.

Основной ключевой возможностью программного средства является автоматизированная оценка эффективности инвестиционных проектов по ключевым показателям: NPV, IRR, Payback Period, а также возможность получения аналитической информации по инвестиционному проекту в зависимости от параметра расчета.

В рамках программного средства автоматизируется управление инвестиционных проектов в следующих бизнес-процессах:

– расчет NPV;

– расчет IRR;

– расчет Payback Period;

– оформление инвестиционного проекта;

– сбор статистики и аналитики по инвестиционным проектам.

Программное средство создается с целью:

– ускорения процесса оценки инвестиционных проектов;

– повышениея точности и надежности аналитических данных;

– улучшение взаимодействия между различными подразделениями компании за счет унификации отчетности;

– обеспечить удобный и интуитивно понятный интерфейс для пользователей с различными уровнями подготовки;

– реализовать возможность многопользовательского доступа с настройками прав доступа;

– снижения рисков, связанных с инвестициями, за счет более детального анализа потенциальных проектов.

Ожидается, что перенос рутинных задач, выполняемых вручную, операций в специализированное программное средство позволит сэкономить ресурсы, снизить количество ошибок и выполнять задачи быстрее и более качественно. Программное средство может применяться директором организации, а также администраторами и менеджерами.

Программное средство убирает необходимость хранить и самостоятельно обрабатывать информацию сотрудникам, что позволяет сократить время, затрачиваемое на информационно-аналитическую деятельность, направляя ресурсы на решение наиболее важных задач.

Проектирование программного обеспечения. Архитектурное решение определяет общую структуру и организацию компонентов системы, чтобы достичь нужного функционала и качества. Архитектурное решение представляет выбор и использованию различных технологий и паттернов, которые используются в программном средстве. Разрабатываемое программное средство базируется на клиент-серверной архитектуре.

Клиент – локальный компьютер на стороне виртуального пользователя, который выполняет отправку запроса к серверу для возможности предоставления данных или выполнения определенной группы системных действий.

Сервер – очень мощный компьютер или специальное системное оборудование, которое предназначается для разрешения определенного круга задач по процессу выполнения программных кодов. Он выполняет работы сервисного обслуживания по клиентским запросам, предоставляет пользователям доступ к определенным системным ресурсам, сохраняет данные или базу данных.

Особенности такой модели заключаются в том, что пользователь отправляет определенный запрос на сервер, где тот системно обрабатывается и конечный результат отсылается клиенту. В возможности сервера входит одновременное обслуживание сразу нескольких клиентов. Для реализации программного обеспечения было выбрано веб-приложение. Во-первых, главным преимуществом является масштаб приложения – одновременное использование большим количеством сотрудников. Во-вторых, клиент-серверная версия не требует установки, в отличии от мобильных приложений, а, значит, не загружает память устройства. На основании изучения популярных ресурсов, посвященных разработке программных средств такого типа, было принято решение, в качестве потенциальной программной платформы рассмотреть Spring.

Spring Security был выбран для построения систем аутентификации и авторизации. Этот фреймворк значительно упрощает работу с безопасностью пользователей. Он предоставляет комплексную модель программирования и конфигурации для современных корпоративных приложений на основе Java – на любой платформе развертывания.

В качестве СУБД был выбран PostgreSQL, поскольку занимает лидирующие позиции на рынке.

Дизайн-система – это комплексный набор компонентов, инструментов, правил и руководств, который позволяет разработчикам и дизайнерам создавать последовательный и согласованный визуальный язык для продуктов и услуг компании.

Дизайн-система – это библиотека материалов и UI-компонентов, которая помогает команде быстрее разрабатывать, тестировать, масштабировать решения и создавать единообразный визуальный образ продукта.

Дизайн-система данной курсовой работы представлена на рисунке 1.

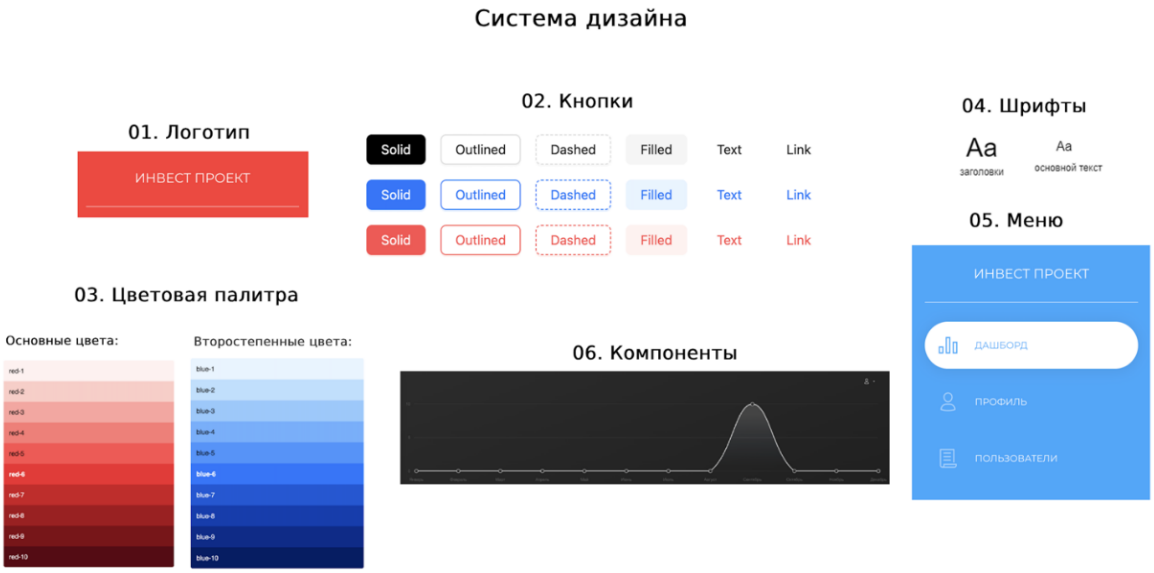


Рисунок 1. Система дизайна

Основной документ, который получается в итоге проектирования – детальные схемы страниц. Они показывают, что представляет из себя каждая страница системы, каковы особенности ее работы.

После этого можно приступить к разработке дизайна. Также рекомендуется подготовить описание страниц – сценарии взаимодействия, в которых подробно изложен алгоритм работы пользователя с сайтом. Однако наиболее эффективным результатом проектирования является интерактивный прототип системы. Это функциональная модель пользовательского интерфейса, включающая основные страницы и процессы работы системы.

На рисунке 2 представлена общая карта экранов всех прототипов.

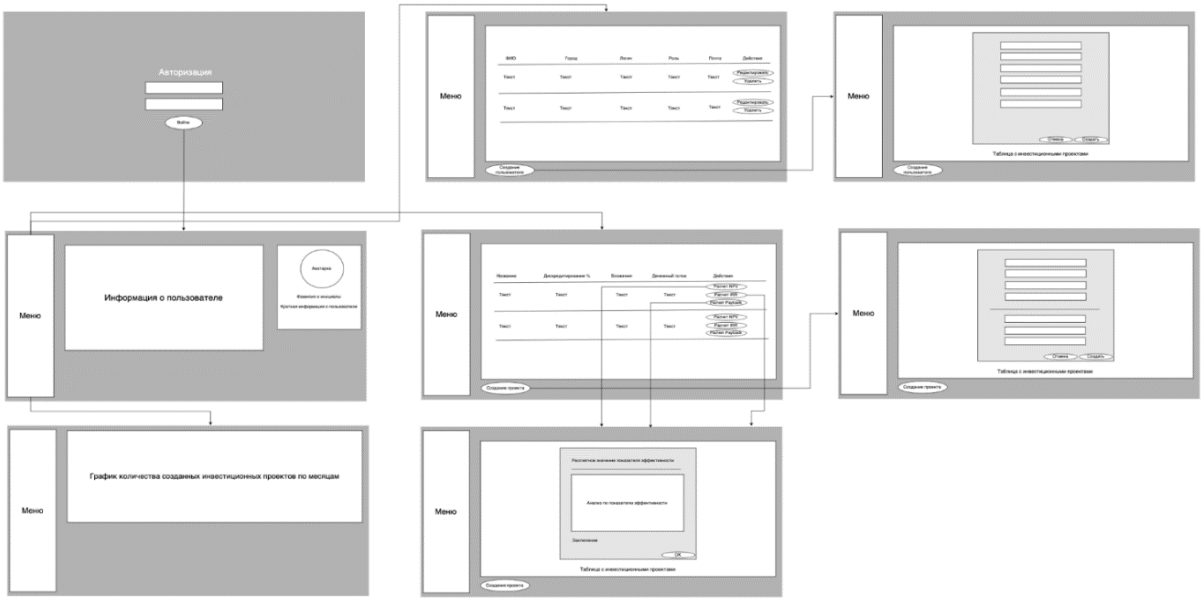


Рисунок 2. Карта экранов

Пользовательский интерфейс спроектирован в соответствии с необходимым функционалом, была реализована удобная навигация по программному средству и готова для разработки.

Для моделирования системы использованы три этапа проектирования:

– концептуальное проектирование;

– логическое проектирование;

– физическое проектирование.

Концептуальное (инфологическое) проектирование – анализ предметной области и ее описание.

Этот этап осуществляется без ориентации на какие-либо конкретные программные или технические средства. Даталогическое (логическое) проектирование – описание логической структуры данных средствами системы управления базами данных (СУБД), для которой проектируется БД.

Описание данной модели основывается на построении концептуальной модели. Для обеспечения минимальной избыточности и физического объёма данных, а также для более быстрого доступа, модель приведена к третьей нормальной форме.

Общая схема связанных сущностей базы данных представлена на рисунке 3.

В результате проделанной работы таблицы в базе данных приведены к 3 нормальной форме. Дело в том, что избыточность данных создает предпосылки для появления различных аномалий, снижает производительность, и делает управление данными не гибким и не очень удобным. Отсюда можно сделать вывод, что нормализация нужна для:

– устранения аномалий;
– повышения производительности;
– повышения удобства управления данными.

Таким образом, на данном этапе спроектирована информационная модель программного средства автоматизации оценки эффективности инвестиционных проектов. Модель состоит из четырех связанных сущностей и приведена к третьей нормальной форме. Программа разрабатывается в клиент-серверной архитектуре. Изначально представлена диаграмма компонентов, которая показывает, как компоненты соединяются вместе для формирования программных систем.

Компонент Database включает в себя базу данных, соединение с сервером происходит посредством использования JDBC драйвера. Сервер имеет Database control classes необходимый для подключения и работы с базой данных, control classes служит в качестве соединяющего звена реализации бизнес-логики. Сервер соединен с клиентом, передача данных происходит с помощью TCP/IP. На клиенте организована структура пользовательского интерфейса на основе HTML/CSS.

Диаграмма размещения – один из доступных видов диаграмм, поддерживаемых FlexberryDesigner. Корпоративные приложения часто требуют для своей работы некоторой ИТ-инфраструктуры, хранят информацию в базах данных, расположенных где-то на серверах компании, вызывают веб-сервисы, используют общие ресурсы и т. д. В таких случаях полезно иметь графическое представление инфраструктуры, на которую будет развернуто приложение. Для этого и нужны диаграммы развёртывания, которые иногда называют диаграммами размещения.

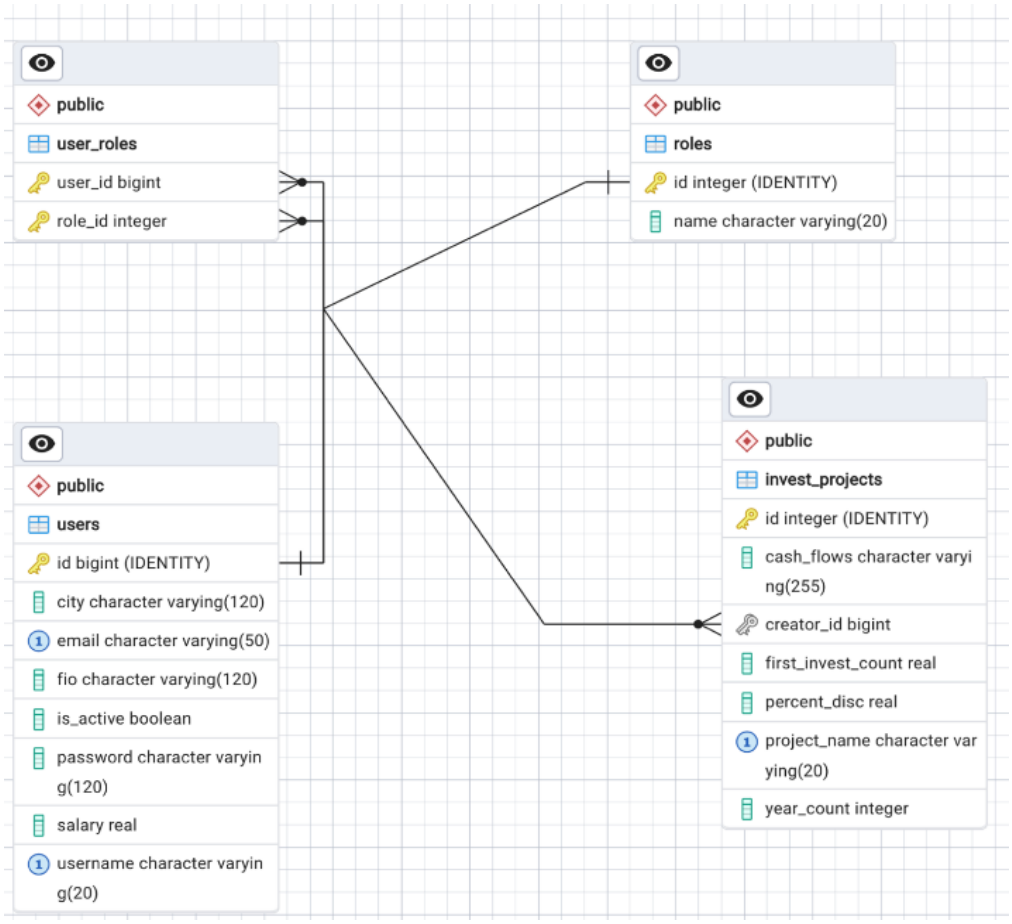


Рисунок 3. Физическая модель базы данных

Диаграмма компонентов представлена на рисунке 4.

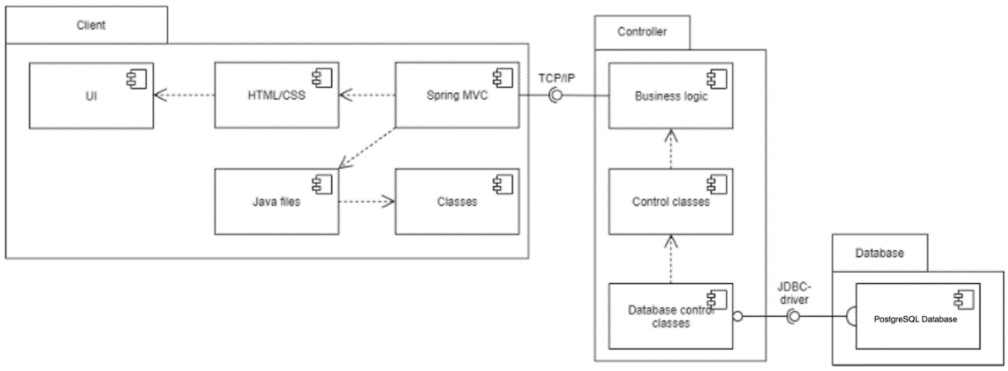


Рисунок 4. Диаграмма компонентов

Разработанная диаграмма размещения представлена на рисунке 5.

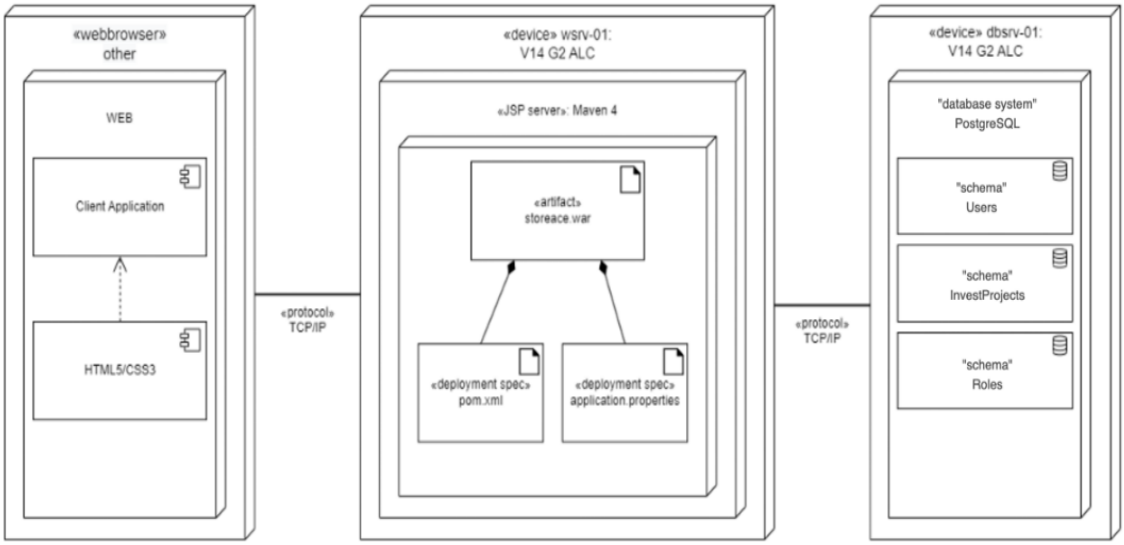


Рисунок 5. Диаграмма размещения

Сервер использует сборщик проектов Maven, Spring Boot необходим для запуска и управления.

Диаграмма классов – структурная диаграмма языка моделирования UML, демонстрирующая общую структуру иерархии классов системы, их кооперaций, атрибутов (полей), методов, интерфейсов и взаимосвязей между ними. Широко применяется не только для документирования и визуализации, но также для конструирования посредством прямого или обратного проектирования.

Диаграмма классов определяет типы классов системы и различного рода статические связи, которые существуют между ними.

На диаграммах классов изображаются также атрибуты классов, операции классов и ограничения, которые накладываются на связи между классами. Вид и интерпретация диаграммы классов существенно зависит от точки: классы могут представлять сущности предметной области (в процессе анализа) или элементы программной системы (в процессах проектирования и реализации).

Основными элементами являются классы и связи между ними.

Классы характеризуются при помощи атрибутов и операций.

Атрибуты описывают свойства объектов класса.

Большинство объектов в классе получают свою индивидуальность из-за различий в их атрибутах и взаимосвязи с другими объектами.

Однако, возможны объекты с идентичными значениями атрибутов и взаимосвязей.

Виды связей: ассоциация; агрегация; наследование.

На рисунке 7 представлена диаграмма последовательности создание инвестиционного проекта и расчет показателя эффективности NPV, на которой отображается процесс работы клиента.

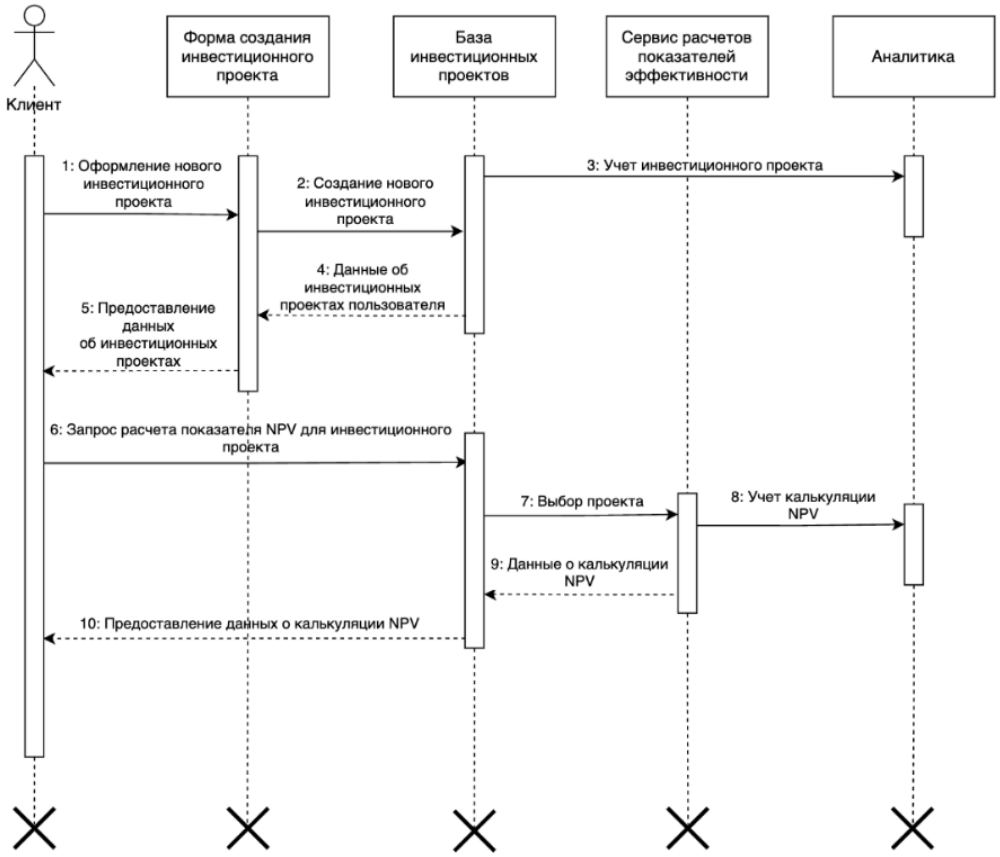


Рисунок 7. Диаграмма последовательности для вариантов использования «Создать инвестиционный проект» и «Рассчитать показатель эффективности NPV»

На диаграмме показано взаимодействие таких компонентов, как форма создания инвестиционного проекта, база инвестиционных проектов, сервис расчетов показателей эффективности, аналитика. Действующее лицо – клиент, который работает с инвестиционными проектами через интерфейс. При создании нового инвестиционного проекта данные поступают в базу инвестиционных проектов и в аналитику. Дальше клиент может калькулировать показатели эффективности инвестиционного проекта, в частности показателя NPV, которые в свою очередь также поступят в аналитику.

Диаграмма деятельности – это технология, позволяющая описывать логику процедур, бизнес-процессы и потоки работ. Во многих случаях они напоминают блок-схемы, но принципиальная разница между диаграммами деятельности и нотацией блок-схем заключается в том, что первые поддерживают параллельные процессы.

На рисунке 8 показана диаграмма деятельности.

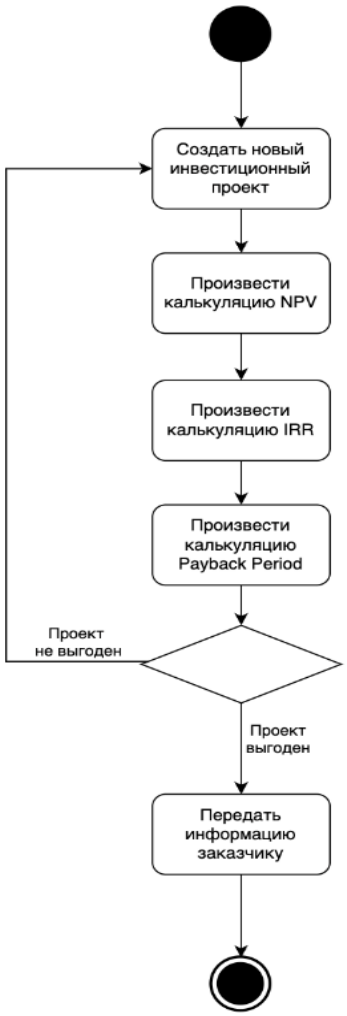


Рисунок 8. Диаграмма деятельности для процесса расчета эффективности инвестиционного проекта

На диаграмме деятельности описана деятельность клиента. Клиент создает новый инвестиционный проект, после чего производит калькуляцию основных показателей эффективности инвестиционного проекта: NPV, IRR, Payback Period, далее после проверки программным средством инвестиционного проекта на предмет выгоден он или нет клиент передает информацию заказчику.

Заключение. Актуальность автоматизации оценки инвестиционных проектов в современных условиях бизнеса обусловлена несколькими ключевыми аспектами. Ниже приведены основные причины, подчеркивающие актуальность данного процесса:

1. Увеличение объема данных: С ростом объемов информации, доступной для анализа, компании сталкиваются с необходимостью обработки больших массивов данных. Автоматизация позволяет быстро и эффективно анализировать данные, что способствует более обоснованным инвестиционным решениям.

2. Скорость принятия решений: В условиях высокой конкуренции скорость реакции на изменения рынка имеет критическое значение. Автоматизированные системы позволяют значительно сократить время, необходимое для оценки проектов, что позволяет компаниям быстрее реагировать на возможности и угрозы.

3. Снижение риска ошибок: Человеческий фактор всегда несет в себе риск ошибок, особенно при работе с большими объемами информации. Автоматизация процессов оценки помогает минимизировать эти риски, обеспечивая более точные и надежные результаты.

4. Улучшение качества анализа: Современные технологии позволяют использовать сложные аналитические инструменты и модели для оценки рисков и прогнозирования результатов. Это ведет к более глубокому пониманию инвестиционных возможностей и потенциальных угроз.

5. Интеграция с другими системами: Автоматизированные решения могут быть интегрированы с другими бизнес-приложениями, такими как бухгалтерский учет и управление проектами. Это создает единую платформу для управления финансами и ресурсами, что упрощает процесс принятия решений.

6. Гибкость и адаптация: Автоматизация позволяет быстро адаптировать процессы оценки к изменениям в законодательстве, рыночной среде или внутренней политике компании, что делает систему более устойчивой и эффективной.

7. Повышение прозрачности и отчетности: Автоматизированные системы обеспечивают высокую степень прозрачности в процессе оценки инвестиционных проектов, что важно для внутренних и внешних аудиторов, а также для инвесторов.

8. Снижение затрат: Автоматизация процессов может привести к снижению операционных затрат за счет оптимизации ресурсов и повышения эффективности работы команды.

Таким образом, автоматизация оценки инвестиционных проектов является неотъемлемой частью современного управления, способствуя повышению эффективности, снижению рисков и улучшению качества принимаемых решений. В условиях быстро меняющегося рынка компании, использующие автоматизированные инструменты, получают значительное конкурентное преимущество.

Список литературы

[1] Бланк, И. А. Основы инвестиционного менеджмента: в 2 т. Т. 1 / И. А. Бланк. Киев: Эльга: SMART BOOK, 2013. - 672С.

[2] Medium [Электронный ресурс]. Режим доступа: https://medium.com/pitchspot/what-is-a-swot-analysis-and-how-to-do-it-right-with-examples-8de1d8a5f3b7. Дата доступа: 04.03.2025.

Авторский вклад

Все авторы внесли равнозначный вклад.

SOFTWARE FOR AUTOMATED EVALUATION OF INVESTMENT PROJECT EFFECTIVENESS

N.K. Kryuchkov
Lead Front-End Developer SberEducation LLC

A.V. Vorobey
Assistant of the Department of Engineering Psychology and Ergonomics at BSUIR, Master of Technical Sciences

I.V. Andrialovich
Senior Lecturer of the Department of Engineering Psychology and Ergonomics at BSUIR, Master of Science in Engineering and Technology

Abstract. This article discusses the development and implementation of software for the automated assessment of the effectiveness of investment projects. It emphasizes the importance of qualitative evaluation of investment decisions in a dynamic market and increasing competition. The main functional capabilities of the software are described, including the calculation of key performance indicators such as Net Present Value (NPV), Internal Rate of Return (IRR), and payback period. The advantages of automating the evaluation process are also considered, such as reducing human error, increasing calculation accuracy, and accelerating decision-making. This article will be useful for professionals in the fields of financial analysis, investing, and project management, as well as for developers of software solutions for business.

Keywords: software, investment projects, investment effectiveness, ROI (Return on Investment), NPV (Net Present Value), IRR (Internal Rate of Return).

УДК 517.18.(075.8)

БОЛЬШИЕ СКОРОСТИ И НАНО ПАРАМЕТРЫ ЧАСТИЦ В ПОДПРОСТРАНСТВАХ АТОМА ВОДОРОДА



И.П. Кобяк
Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь, доцент кафедры ЭВМ, канд. техн. наук, доцент
IPKobyak2012@mail.ru

И.П.Кобяк

Работает в Белорусском государственном университете с 1982 г. Занимаемые должности инженер, ассистент, доцент кафедры ЭВМ. Защитил кандидатскую диссертацию в 1993 г. Область научных интересов: методы идентификации сообщений, исследования в области ядерной физики.

Аннотация. Для синтеза квантово-электронных устройств с высокой степенью достоверности информационных сообщений рассмотрена методика формирования пространств ядра водородоподобного атома на граничных скоростях вращения. Данный принцип может быть использован для улучшения параметров квантов в пересылаемых в системах связи информационных сообщений. В соответствии с поставленной целью в статье рассмотрен меж пространственный переход материи электрона из реального пространства третьего измерения в мнимое шестое и далее в пространство седьмого измерения. Показано, что электрон при движении по нулевому радиусу взаимодействует с пространствами ядра, что и определяет его движение в форме перемещения в различных измерениях. При этом на основе скоростей полей ядра определены параметры электрона в области стандартного нулевого радиуса. На основании полученных зависимостей рассчитана скорость собственного вращения нано частицы практически равная скорости движения по нулевой орбите.

Ключевые слова: квантовая механика, ядро атома, скорость света, шаровая поверхность, электрон, бесконечно малые величины.

Введение. Современные достижения в области квантовой механики и электроники, как правило, определяют ряд ограничений или не доработок в проектах, связанных с отсутствием прецизионных знаний в теории атома. Данный факт приводит к необходимости многократной проверки теоретических результатов в различных экспериментах или в программах моделирования новых соотношений. Основным препятствием на пути решения исследовательских задач является отсутствие возможности чисто математического решения проблем, основанных на меж пространственных переходах нано электронной плазмы при воздействии энергетики струн на объекты внутри атома [1,2].

В указанных задачах большинство исследователей рассматривает атом с точки зрения теории третьего измерения с использованием соотношений для параметров в *Re*-пространстве. Такой подход к исследованиям характеризует только псевдо статику нано частиц в третьем, а в лучшем случае и в седьмом измерении. При этом в известных направлениях решения задач физики атома особое внимание уделяется радиусу Бора, хотя указанный нулевой параметр , по сути, является частным случаем общей теории [3]. В связи с этим в представляемой работе при расчетах пространств ядра использована новая концепция, а именно: концепция «сжатия»

по Лоренцу радиуса движения электрона под действием сил самого ядра. Такой подход с учетом системности величин позволяет получить требуемые соотношения достаточно высокой сложности для определения функции связи параметров ядра и электрона в различных измерениях. При этом физико-математическая системность для пространств и параметров атома в общем случае достигается только с использованием Лоренц-фактора. В то же время отсутствие поправки Лоренца в формируемых соотношениях, как правило, позволяет сделать выводы о меж пространственных переходах при взаимодействии нано частиц с энергетикой струн. Указанный подход позволяет достичь понимания в вопросах преобразования и расчетов значений параметров на комплексной плоскости. Таким образом, в данной работе рассматриваются вопросы перехода пространств, с учетом вычисленного интеграла скорости, ведущие к констатации физики перетекания плазмы ядра из мнимого пространства в реальное, к получению численного значения размера электрона на основе полей ядра пятого и седьмого измерений.

Интегрирование радиуса ядра. Для достижения поставленной цели необходимо представить равенство для радиуса ядра вида

$$\vec{r}_s = \vec{r}_{0,L}\sqrt{1-\frac{v_s^2}{c^2}}. \tag{1}$$

с учетом приращения скорости $\Delta\vec{v}_s$ (считаем, что скорость возрастает в связи с притоком внешней энергии) и соответствующего изменения радиуса ядра на величину $\Delta\vec{r}_s$:

$$\vec{r}_s - \Delta\vec{r}_s = \vec{r}_{0,L}\sqrt{1-\frac{\left(\vec{v}_s+\Delta\vec{v}_s\right)^2}{c^2}}.$$

Отсюда, после ряда преобразований, имеем приблизительное равенство:

$$\Delta r_s^2 \approx -\frac{2r_{0,L}^2}{c^2}\vec{v}_s\Delta\vec{v}_s.$$

Соответственно, при $\Delta\vec{v}_s \to 0$ получаем интеграл вида:

$$r_s^2 = -\frac{2r_{0,L}^2}{c^2}\int \vec{v}_s d\vec{v}_s = -\frac{v_s^2}{c^2}r_{0,L}^2.$$

Из данного соотношения имеем:

$$-\vec{v}_s = j\frac{\vec{r}_s}{\vec{r}_{0,L}}\vec{c}.$$

Воспользуемся поправкой Лоренца от скорости $-\vec{v}_{s,c} = -\vec{c}\sqrt{2}+\Delta\vec{c}$ для преобразования данного соотношения. При этом получаем требуемое для дальнейших исследований равенство:

$$-\vec{v}_{s,c} \Leftarrow -j\frac{\vec{r}_s}{\vec{r}_{0,L}}\frac{\vec{c}}{\left[\beta_s\left(\Delta\vec{c}\right)\right]^2}, \tag{2}$$

где $\beta\left(\Delta\vec{c}\right) = -j\sqrt{1-\psi}$, а знак «$\Leftarrow$» означает физическую смену пространств нано частицей.

Внешнее пространство ядра атома водорода и принцип его образования. В соответствии с концепцией образования нулевого радиуса как функции радиуса ядра, используя обратное соотношение для данного принципа, будем считать, что увеличение радиуса ядра в формуле (1) в $f(k)$ раз приводит к уменьшению скорости его оболочки в $\mu(p)$

раз, то есть:

$$f(k)\vec{r}_s = \vec{r}_{0,L}\sqrt{1-\frac{\mu^2(p)v_s^2}{c^2}}.$$

Соответственно из полученного равенства следует, что:

$$\mu(p)\vec{v}_s = \vec{c}\sqrt{1-\frac{f^2(k)r_s^2}{r_{0,L}^2}}.$$

Рассмотрим теперь соотношение для скорости $-\vec{v}_{s,c} = -\vec{c}\sqrt{2} + \Delta\vec{c}$ с учетом формулы (2). В указанном соотношении знак тильды ↔ в используемых параметрах говорит о движении материи со скоростью света и выше (по шаровой поверхности) во всех направлениях. Тогда с учетом (2) можем записать:

$$\mu(p)\left(-\vec{v}_{s,c}\right) \Leftarrow -j\frac{f(k)\vec{r}_s}{\vec{r}_{0,L}}\vec{c}\left[\gamma(\Delta\vec{c})\right]^2, \tag{3}$$

где $\gamma(\Delta\vec{c})$ - Лоренц-фактор.

Выразим множитель $\mu(p)$, характеризующий функцию уменьшения скорости $\vec{v}_s$ при увеличении радиуса ядра, из равенства (1):

$$\mu(p) = \frac{\vec{c}}{\vec{v}_{s,c}}\sqrt{1-f^2(k)\frac{r_s^2}{r_{0,L}^2}}$$

и подставим его в равенство (3). Тогда закон изменения скорости оболочки ядра в пространстве пятого измерения $-\vec{v}_{s,c}$ в зависимости от коэффициента расширения радиуса ядра можно определить соотношением:

$$-\vec{v}_{s,c} \Leftarrow -j\vec{v}_s\frac{1}{\sqrt{\frac{r_{0,L}^2}{f^2(k)r_s^2}-1}}\left[\gamma(\Delta\vec{c})\right]^2. \tag{4}$$

С учетом Лоренц-фактора для скорости $-\vec{v}_{s,c}$ из (4) имеем:

$$-\vec{v}_{s,c} \Leftarrow \vec{v}_s\frac{1}{-j(1-\psi)\sqrt{\frac{r_{0,L}^2}{f^2(k)r_s^2}-1}}. \tag{5}$$

Знаменатель в соотношении (5) при $f(k)=1$ можно привести к виду:

$$-j(1-\psi)\sqrt{\frac{r_{0,L}^2}{f^2(k)r_s^2}-1} \approx -j\vec{r}_{0,L}(1-\psi)\frac{1}{\vec{r}_s}, \tag{6}$$

Из равенства (5) также следует:

$$\frac{-\vec{v}_{s,c}}{\vec{v}_s} \approx \frac{-\sqrt{2}\vec{c}}{\vec{c}} = -\sqrt{2}.$$

Отсюда имеем:

$$-j\vec{r}_{0,L}(1-\psi)\frac{1}{\vec{r}_s} \Rightarrow -\frac{1}{\sqrt{2}}.$$

То есть из данного физического соотношения следует, что при теоретической скорости $-\vec{v}_{s,c} = -\vec{c}\sqrt{2}$ выполняется переход плазменной материи ядра из пространства *–Im* шестого

измерения в пространство -*Re* пятого измерения:

$$-j\vec{r}_{0,L}(1-\psi) \Rightarrow -\frac{\vec{r}_s}{\sqrt{2}}.$$

Данное соотношение подтверждает полученный в предыдущих работах результат, то есть тот факт, что радиус $-j\vec{r}_{0,L}(1-\psi)$ является критическим и представляет собой нано шар «черной материи», в котором осуществляется преобразование пространства $-Im$ шестого измерения в $-Re$-пространство пятого измерения. Соответственно в результате энергетического скачка осуществляется переход в седьмое измерение.

Параметры атома водорода на поверхности ядерной оболочки. Считаем, что при расчете параметров ядра с учетом концепции «сжатого» по Лоренцу радиуса $\vec{r}_{0,L}$, скорость на поверхности в каждой точке нано частицы рассчитывается в соответствии с классическим равенством:

$$\vec{v}_s = 2\pi\vec{r}_s\vec{\omega}_s. \tag{7}$$

При этом, полагая, что скорость (7) реализуется только на бесконечно малой дуге радиуса $\vec{r}_s$, а затем масса перемещается на радиус $-j\vec{r}_{0,L}(1-\psi)$ с последующим переходом в -*Re*-пространство, и далее в седьмое измерение, определим линейную скорость $\vec{v}_{s,l}$ как долю скорости $\vec{v}_s$. Тогда на основании соотношения (7) можно вычислить требуемое далее значение угловой частоты $\vec{\omega}_s$ по формуле:

$$\vec{\omega}_s = \frac{\vec{v}_{s,l}}{\pi\sqrt{2}\vec{r}_s}. \tag{8}$$

Собственная скорость вращения оболочки ядра на своей псевдо орбите $-j\vec{r}_s$, сжатой до уровня $-j\vec{r}_{0,L}(1-\psi)$ на скорости $-\vec{c}\sqrt{2}+\Delta\vec{c}$, с учетом (8) определится равенством в пятом измерении:

$$-\vec{v}_{s,c} = 2\pi\left[-\vec{r}_{0,L}(1-\psi)\right]\vec{\omega}_s. \tag{9}$$

При синтезе соотношения (9) полагаем, что реальный радиус оболочки атома имеет стандартный размер, а со сферой $\vec{r}_{0,L}$ находится во взаимодействии сам электрон. Причем электрон имеет конечный радиус $\vec{r}_{0,L}\psi$ в пятом измерении.

Знак тильды над параметром угловой частоты в (8) следует из шарового представления пространств и последующих расчетов релятивистских параметров на комплексной плоскости.

Размер электрона в области нулевого пространства пятого измерения. Итак, будем констатировать факт временного существования электрона в пространстве -*Re* в пределах: $-\vec{r}_{0,L} \div \left[-\vec{r}_{0,L}(1-\psi)\right]$.(9).

Определим бесконечно малую величину ψ, характеризующую размер электрона в области пятого измерения сил ядра.

Для этого рассмотрим соотношение (4) подставляя в левую его часть скорость (9). При этом необходимо учитывать, что природа скоростей в левой и правой частях формируемого далее равенства различна. В левой части имеем:

$$-\vec{v}_{s,c} = -\left(\vec{c}\sqrt{2} - \Delta\vec{c}\right) = -2\pi\vec{r}_{0,L}\vec{\omega}_s\left(1-\psi\right),$$

то есть в левую часть подставляется скорость с учетом радиуса $-\vec{r}_{0,L}\left(1-\psi\right)$, являющегося радиусом ядра при увеличении скорости $-\vec{c} - \Delta\vec{c}$ до уровня $-\vec{c}\sqrt{2} + \Delta\vec{c}$. В данных формулах в обоих значениях скоростей отклонение $\Delta\vec{c}$ различно.

В правую часть формируемого равенства подставляем соотношение, соответствующее расширению внешней оболочки ядра до некоторого радиуса, определяемого коэффициентом расширения $f\left(k\right) = k$. Таким образом, из (4), с учетом перестановки слагаемых под корнем в правой части получаем:

$$2\pi\left[-\vec{r}_{0,L}\left(1-\psi\right)\right]\vec{\omega}_s = \vec{v}_s \frac{1}{-\left(1-\psi\right)\sqrt{1-\frac{r_{0,L}^2}{k^2 r_s^2}}}. \tag{10}$$

где $\left|kr_s\right| > \left|r_{0,L}\right|$. Для упрощения вывода последующих соотношений значение ψ в правой части (10) принимаем равным нулю, так как данный параметр исчисляется весьма малой величиной по отношению к единице в *Re*-пространстве.

Возведем теперь обе части соотношения (10) в квадрат с учетом (7) и получим:

$$\left[\frac{-2\pi\vec{r}_{0,L}\left(1-\psi\right)\vec{\omega}_s}{2\pi\vec{r}_s\vec{\omega}_s}\right]^2 = \frac{k^2 r_s^2}{r_{0,L}^2 - k^2 r_s^2}.$$

Учитывая равенство для модулей угловых частот, далее имеем соотношение:

$$\frac{r_{0,L}^2\left(1-\psi\right)^2}{r_s^2} = \frac{k^2 r_s^2}{r_{0,L}^2 - k^2 r_s^2}. \tag{11}$$

Введем в рассмотрение данной задачи коэффициент расширения оболочки ядра в виде среднего значения параметров:

$$k_{av} = \frac{-r_{0,L}\left(1-\psi\right)}{r_s},$$

где k_{av} - (*av*-среднее) значение, определяющее примерно равные условия существования левой и правой частей (10) на радиусе $-\vec{r}_{0,L}\left(1-\psi\right)$. При этом знак минус указанного параметра определяет пятое измерение радиуса. Тогда можно показать, что полученное выше соотношение (11) приводит к равенству вида:

$$\vec{r}_{0,L}\psi \ = \vec{r}_{0,L} - \sqrt{r_{0,L}^2 - r_s^2}\,. \tag{12}$$

Таким образом, радиус электрона в *Re*-пространстве пятого измерения (10) будет равен $\vec{r}_{0,L}\psi = 6{,}683\cdot 10^{-19}\,см,$ где значение $\psi \approx 1{,}263\cdot 10^{-10}\,см.$

Размер электрона в области нулевого пространства седьмого измерения. Определим теперь значение ξ как окрестность радиуса $\vec{r}_{0,L}$ в пространстве +*Re*, определяемом квантовым скачком в седьмое измерение. а также с учетом правой части равенства (10).

Пусть коэффициент

$$k = \frac{r_{0,L} + \xi}{r_s}.$$

Тогда из (10) следует:

$$-\vec{v}_{s,c} = -\vec{v}_s \frac{1}{\sqrt{1 - \frac{r_{0,L}^2}{\left(\vec{r}_{0,L} + \xi\right)^2}}}. \tag{13}$$

Условием возникновения скорости $\vec{v}_{0,L}$ следует считать уровень взаимодействия пространства седьмого измерения с электроном, что и обуславливает результирующее движение нано частицы по радиусу $\vec{r}_{0,L}$. При этом скорость $\vec{v}_{s,l}$ седьмого измерения (с учетом примерного равенства модулю скорости пятого измерения) будет равна:

$$\vec{v}_{s,l} = \pi\sqrt{2}\vec{r}_{0,L}\left(1-\psi\right)\vec{\omega}_s - \frac{\vec{v}_{0,L}}{\sqrt{2}}.$$

В данном соотношении учитывается, что электрон является ведомым элементом и поглощает энергию ведущего поля ядра.

Подставляя полученный параметр в левую часть (13) получим численное значение для расстояния ξ от точки радиуса $\vec{r}_{0,L}$ до внешнего пространства $\vec{r}_{0,L} + \xi$:

$$\xi \approx 1{,}429 \cdot 10^{-10}\ \mathring{A}. \tag{14}$$

То есть радиус $\vec{r}_{0,L}$ имеет отклонение пространства от «стандарта» в сторону увеличения равное окрестности (14).

Рассмотрим теперь численное значение коэффициента k_{av}, характеризующего отношение величин в пространстве нулевого радиуса:

$$k_{av} = \frac{r_{0,L}}{r_s}. \tag{15}$$

Тогда, полагая «в статике», что радиус $\vec{r}_{0,L}$ «критический» и скорость движения *Re*-пространственного поля на нем равна нулю, будем использовать также гипотезу об отсутствии вращения в обеих критических точках $\vec{r}_{0,L}$ и $\vec{r}_s$ (без материальных объектов), а также условное наличие линейности в коэффициенте (15). При этом проекция значения ξ на пространство оболочки ядра $\vec{r}_s - \Delta\vec{r}_s$ в пространстве +*Re* имеет значение равное:

$$\Delta\vec{r}_s \approx \xi \frac{\vec{r}_s}{\vec{r}_{0,L}}, \tag{16}$$

где значение (15) представляет собой по сути масштабирующий коэффициент. При этом радиус электрона будет определяться равенством $\vec{r}_{el} = \xi$. Отсюда следует, что угловая частота собственного вращения данной нано частицы, с учетом флуктуаций радиуса ядра будет равна:

$$\vec{\omega}_{el} = \frac{\vec{r}_{0,L}}{\vec{r}_{el} = \xi} \vec{\omega}_0 .$$

Теперь считаем, что природа приобретенного вращения электрона такова, что:

$$\vec{v}_{el} = 2\pi\vec{\omega}_{el}\vec{r}_{el} = 95,4824\, m.\text{км/ с}. \quad (18)$$

Необходимо учитывать, что значение (18) является продуктом собственного вращения нано частицы на нулевом радиусе. При этом, как оказалось, данный параметр практически совпадает с рассчитанной скоростью движения самого электрона по радиусу $\vec{r}_{0,L}$. Следовательно, можно сделать вывод, что на переменном нулевом радиусе $\vec{r}_{0,L}$ всегда выполняется растекание твердотельного электрона в плазменный объект.

Список литературы

[1] Барбашов, Б. М., Нестеренко, В. В. Суперструны – новый подход к единой теории фундаментальных взаимодействий // Успехи физических наук. Том 150, № 4. 1986, С. 489-524.

[2] Поляков А. М. Калибровочные поля и струны. М.: ИТФ, 1995. 300 с.

[3] Кобяк И. П. Основы теории атома водорода для задач синтеза квантово-электронных схем // Автоматизированные системы управления технологическими процессами АЭС и ТЭС = Instrumentation and control systems for NPP and TPP : материалы II Межд. науч.-техн. конф., Минск, 27-28 апреля 2021 года / Белорусский государственный университет информатики и радиоэлектроники. Минск, 2021. С. 207–212.

HIGH SPEEDS AND NANO PARAMETERS OF PARTICLES IN THE SUBSPACES OF THE HYDROGEN ATOM

I.P. Kobiak

PhD, Associate Professor, Chair of ECM,

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus

IPKobyak2012@mail.ru

Abstract. For the synthesis of quantum electronic devices with a high degree of reliability of information messages, a technique for forming the spaces of the nucleus of a hydrogen-like atom at boundary rotational speeds is considered. This principle can be used to improve the quantum parameters in information messages sent in communication systems. In accordance with this goal, the article examines the spatial transition of electron matter from the real space of the third dimension to the imaginary sixth dimension and further into the space of the seventh dimension. It is shown that an electron, when moving along a zero radius, interacts with the spaces of the nucleus, which determines its movement in the form of displacement in various dimensions. At the same time, the electron parameters in the region of the standard zero radius are determined based on the velocities of the core fields. Based on the obtained dependences, the rate of self-rotation of a nano particle is calculated to be almost equal to the speed of motion in a zero orbit.

Keywords: quantum mechanics, atomic nucleus, speed of light, spherical surface, electron, infinitesimal quantities.

УДК 004.85:658.562:004.62

ГИБРИДНЫЕ МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ В ПРОГНОЗИРОВАНИИ КАЧЕСТВА ПРОДУКЦИИ ДЛЯ ПОТОКОВЫХ ДАННЫХ В КОНЦЕПЦИИ ИНДУСТРИЯ 4.0



Н.Ф. Николаев
Аспирант кафедры информационных технологий, искусственного интеллекта и общественно-социальных технологий цифрового общества ФГБОУ ВО «Российский государственный социальный университет» (РГСУ).
kolya.nikolaev2001@gmail.com



А.В. Макаров
Доцент кафедры кафедры информационных технологий, искусственного интеллекта и общественно-социальных технологий цифрового общества ФГБОУ ВО «Российский государственный социальный университет» (РГСУ), кандидат технических наук.
novidei@yandex.ru

А.В. Макаров

Окончил военно-воздушную инженерную академию им Н.Е. Жуковского, область научных интересов связана с разработкой методов и алгоритмов различных направлений искусственного интеллекта.

Н.Ф. Николаев

Учится на аспирантуре в Российском государственном социальном университете, область научных интересов связана с программированием, разработкой информационных систем, разработкой методов и алгоритмов искусственного интеллекта.

Аннотация. В эпоху цифровизации промышленности (Индустрия 4.0) ключевым вызовом становится обработка непрерывных потоков сенсорных данных в реальном времени [1]. Классические статистические модели и стандартные алгоритмы машинного обучения (МО) не успевают обрабатывать возросшие объемы данных и теряют точность из-за явления дрейфа концептов (Concept Drift), вызванного износом оборудования или сменой сырья. В данной работе предлагается архитектура гибридной модели прогнозирования качества продукции, сочетающая быстрые статистические методы (ARIMA), точные модели глубокого обучения (LSTM) и мета-модель для динамического взвешивания результатов. Результаты экспериментальной апробации на данных промышленного производства (датасет SECOM) демонстрируют, что предложенный гибридный подход повышает точность прогноза на 12–15% по сравнению с одиночными моделями в условиях дрейфа и обеспечивает приемлемую задержку обработки.

Ключевые слова: гибридные модели машинного обучения, потоковые данные, Индустрия 4.0, прогнозирование качества, дрейф концептов, LSTM, ARIMA, ADWIN, SECOM.

Введение. Актуальность. Четвертая промышленная революция характеризуется тотальной цифровизацией производственных процессов [11]. Современные заводы оснащены тысячами датчиков, которые генерируют колоссальные объемы данных в реальном времени [9]. Прогнозирование качества продукции на основе этих данных позволяет перейти от реактивного контроля к предиктивному управлению, что критически важно для снижения издержек.

Проблема. Однако работа с потоковыми данными создает два фундаментальных вызова. Во-первых, классические модели машинного обучения, предполагающие статичность данных, не успевают обрабатывать информацию с требуемой скоростью. Во-вторых,

производственная среда динамична: изнашивается инструмент, меняются партии сырья, колеблются температура и влажность. Это приводит к явлению дрейфа концептов – изменению статистических характеристик потока данных, из-за которого модели, обученные на исторических данных, теряют точность [2].

Объект и предмет исследования. Объектом данного исследования являются процессы потоковой обработки производственных данных и алгоритмы прогнозирования показателей качества. Предметом исследования выступают гибридные методы машинного обучения, сочетающие статистические подходы, глубокое обучение и ансамблевые стратегии для повышения точности и адаптивности прогнозов.

Цель работы. Целью работы является разработка и экспериментальная апробация гибридной модели прогнозирования качества, обеспечивающей высокую точность и скорость обработки в условиях потокового ввода данных и наличия дрейфа концептов.

Научная новизна работы определяется следующими тезисами:

1. Предложена архитектура гибридной модели, интегрирующая быстрый статистический алгоритм для краткосрочного тренда и глубокую нейронную сеть для анализа долгосрочных зависимостей, объединенные мета-моделью на основе стеккинга.
2. Разработан метод адаптации весов гибридной модели, при котором обнаружение дрейфа концептов детектором ADWIN инициирует динамическое пересмотр весов компонентов ансамбля, отдавая приоритет более устойчивому к изменениям алгоритму.
3. Впервые предложена и апробирована комбинация методов «статистика + глубокое обучение + динамический ансамбль» для задачи классификации годных/бракованных изделий на данных полупроводникового производства в симулированном потоковом режиме.

1. Анализ существующих подходов и постановка задачи

1.1. Обзор методов прогнозирования качества продукции. Традиционно задачи прогнозирования качества решались с помощью статистических методов [5].

С развитием МО широкое распространение получили регрессионные модели и методы классификации, такие как случайный лес и градиентный бустинг [8].

Для выявления сложных нелинейных зависимостей и работы с временными рядами все чаще применяются нейросетевые архитектуры [2], в частности, рекуррентные нейронные сети, способные учитывать временной контекст.

1.2. Анализ подходов к обработке потоковых данных. Для работы с непрерывными потоками данных используются специализированные фреймворки [9]: Apache Kafka, Spark Streaming, Flink. Ключевой концепцией здесь является окно – способ разделения бесконечного потока на конечные фрагменты для анализа. Различают *tumbling* и *sliding windows*, а также сессионные окна. Выбор типа окна напрямую влияет на задержку получения результата и полноту данных.

1.3. Понятие Concept Drift в производственных циклах. Дрейф концептов означает, что зависимость между входными признаками и целевой переменной меняется с течением времени [2]. В производстве это может быть вызвано:

- постепенным дрейфом: износ режущего инструмента, старение катализатора.
- внезапным дрейфом: смена партии сырья, переналадка станка, поломка датчика.

Игнорирование дрейфа приводит к тому, что модель, показывавшая высокие результаты вчера, сегодня начинает "ошибаться", так как данные больше не соответствуют ее внутренним представлениям.

1.4. Обоснование выбора гибридного подхода. Ни одна отдельно взятая модель не является идеальной для динамической среды.

- Быстрые модели хорошо адаптируются к изменениям, но могут не улавливать сложные долгосрочные паттерны и страдать от шума.
- Точные модели отлично моделируют сложные зависимости, но требуют много данных для переобучения и могут медленно реагировать на резкие изменения. Гибридный подход позволяет объединить сильные стороны разных методов: скорость реакции

одного и точность другого, динамически регулируя их вклад в зависимости от текущей ситуации на производстве.

1.5. Выводы по разделу и формализация требований. На основе проведенного анализа можно сформулировать ключевые требования к разрабатываемой системе:

1. обработка данных должна производиться в скользящем окне с минимальной задержкой.
2. модель должна состоять из нескольких разнородных компонентов для учета различных аспектов данных.
3. необходим встроенный механизм детекции дрейфа концептов.
4. система должна уметь адаптировать прогноз при обнаружении изменений в потоке.

2. Разработка гибридной модели прогнозирования

2.1. Общая архитектура решения

Рисунок 1. Компонентная архитектура гибридной модели прогнозирования

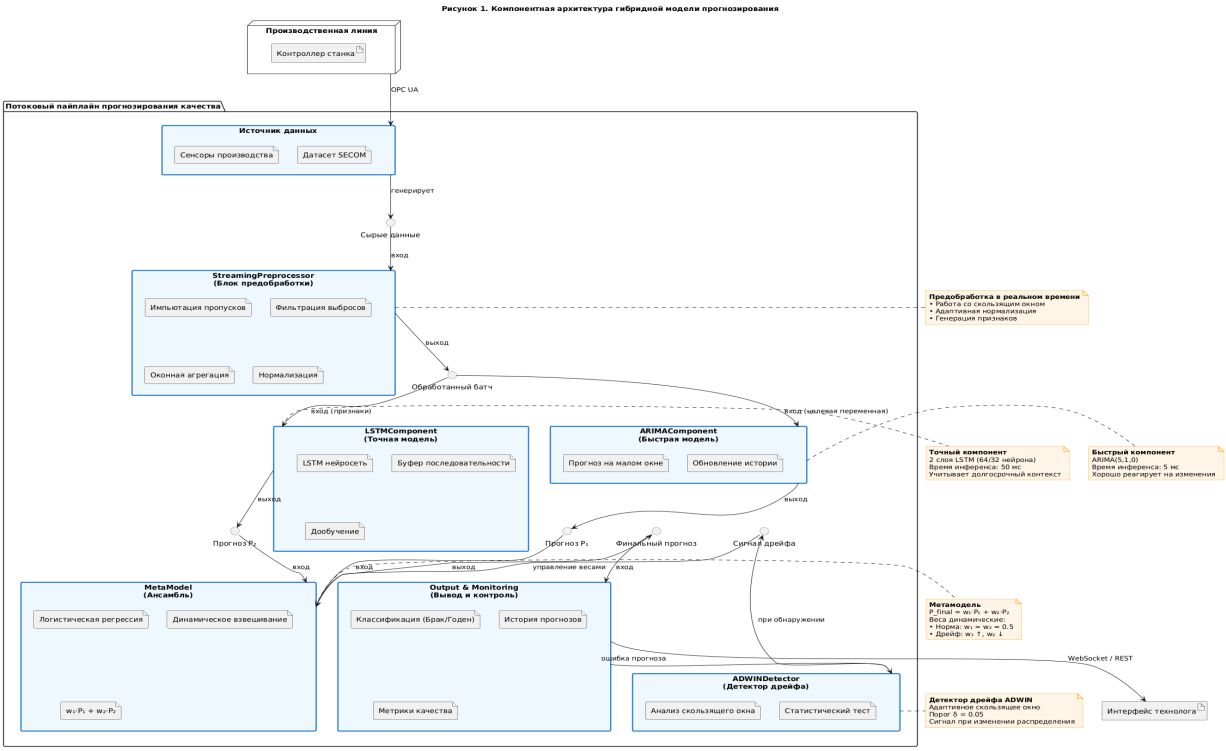


Рисунок 1. Компонентная архитектура гибридной модели прогнозирования

Предлагаемая архитектура (рис. 1) представляет собой потоковый пайплайн (конвейер обработки), состоящий из следующих этапов:

1. Источник данных: Поток сырых сенсорных данных, например, телеметрия с оборудования.
2. Предобработка в реальном времени: очистка, импьютация пропусков, нормализация в рамках текущего окна.
3. Компонент 1 (Быстрый): модель для оперативного реагирования.
4. Компонент 2 (Точный): модель для глубинного анализа.
5. Детектор дрейфа: анализирует поток ошибок или распределение признаков.
6. Метамодель: принимает прогнозы от Компонента 1 и 2 и выдает финальную оценку качества, при этом веса моделей могут корректироваться детектором дрейфа.

2.2. Этап предобработки потоковых данных. В реальном времени невозможно применять «глобальные» статистики, так как поток бесконечен. Поэтому используются методы адаптивной нормализации:

– пропуски: заполняются последним известным значением или медианой по текущему окну;

- выбросы: фильтруются с использованием скользящего среднего и стандартного отклонения в пределах окна;

- оконные агрегации: для каждого временного окна генерируются новые признаки: скользящее среднее, скользящее стандартное отклонение, минимальное и максимальное значение за окно, что позволяет модели видеть не только текущие показания, но и контекст.

2.3. Описание компонентов гибридной модели.

Компонент 1: в качестве быстрого компонента выбран алгоритм ARIMA или его потоковая реализация. ARIMA хорошо моделирует локальные тренды и сезонность на малых временных промежутках и обладает минимальным временем инференса.

Компонент 2: для выявления сложных долгосрочных зависимостей используется сеть LSTM. LSTM способна запоминать состояние процесса на протяжении сотен тактов, что важно для учета, например, постепенного накопления усталости металла.

Компонент 3: для объединения прогнозов используется метод стеккинга. Мета-моделью выступает логистическая регрессия, которая обучается на предсказаниях ARIMA и LSTM. Ее задача – найти оптимальные веса (w1,w2w1,w2) для взвешивания результатов «быстрого» и «точного» компонентов в различных режимах работы производства.

2.4. Механизм адаптации.

Для обнаружения дрейфа используется детектор ADWIN. ADWIN отслеживает поток ошибок прогноза и, если обнаруживает, что средняя ошибка в двух подокнах статистически значимо различается, сигнализирует о дрейфе.

Правило адаптации: при обнаружении дрейфа мета-модель переобучается на данных нового окна. Это позволяет динамически изменить веса w1w1 и w2w2.

В момент резкого изменения технологического режима быстрая модель ARIMA может получить больший вес, так как LSTM еще «не поняла» новых условий. По мере накопления данных о новом режиме вес LSTM снова возрастает.

3. Описание экспериментальной среды и данных.

3.1. Характеристика данных.

Для апробации модели используется общедоступный датасет SECOM [6]. Он содержит данные 1568 наблюдений с 590 сенсорами полупроводникового производства. Целевая переменная – бинарный признак качества. Датасет характеризуется сильным дисбалансом классов (около 7% брака), что типично для реальных производств. Для имитации потока данные подаются в модель последовательно, с заданной скоростью (10 записей/сек). Размер окна обработки варьируется от 50 до 200 записей.

3.2. Инструменты реализации. Эксперимент проводится на языке Python с использованием современных библиотек машинного обучения [9]. Для потоковой обработки и МО используются библиотеки:

- River - специализированная библиотека для обучения моделей на потоковых данных.
- Scikit-learn - для реализации мета-модели и базовых алгоритмов.
- TensorFlow/PyTorch – для построения и обучения сети LSTM.

Стриминговая среда эмулируется путем итеративной подачи данных из датасета, имитируя поступление новых батчей.

3.3. Метрики оценки. Оценка эффективности проводится по двум группам метрик:

1. Качество прогноза: учитывая сильный дисбаланс классов, основной метрикой выбрана F1-score, а также AUC-ROC.
2. Производительность: измеряется среднее время обработки одного батча (latency) и пропускная способность системы.

4. Результаты экспериментального исследования.

4.1. Настройка параметров модели. В ходе эксперимента проводилась настройка размера скользящего окна. Для детектора ADWIN порог значимости ($\delta\delta$) был установлен на уровне 0.05, что позволяет реагировать на умеренные изменения, отсеивая случайные флуктуации.

4.2. Сравнение с бейзлайнами. Было проведено сравнение предлагаемой гибридной модели с одиночными моделями (таблица 1).

В статическом режиме гибридная модель незначительно опережает лучший одиночный алгоритм. Однако в динамическом режиме, где в данные был искусственно внесен дрейф, падение качества F1-score у LSTM составило 18%, а у ARIMA – 25%. Гибридная модель благодаря механизму адаптации снизила качество лишь на 5%, продемонстрировав лучшую устойчивость.

Таблица 1. Сравнение точности моделей

Модель	Статический режим	Динамический режим
Логистическая регрессия	0.68	0.45
Случайный лес	0.79	0.61
LSTM	0.85	0.67
Предложенная гибридная модель	0.86	0.81

4.3. Анализ устойчивости. Для имитации резкого изменения в потоке данных, например, замены сырья, были искусственно увеличены значения одного из ключевых сенсоров на 30%. Гибридная модель продемонстрировала «провал» в точности всего на 2-3 батча, после чего детектор ADWIN зафиксировал дрейф, мета-модель пересчитала веса, и точность восстановилась до уровня >0.75. Одиночная LSTM восстанавливалась более 20 батчей.

4.4. Оценка временной эффективности. Измерение задержки обработки показало, что гибридная модель вносит предсказуемые накладные расходы.

Если инференс ARIMA занимает ~5 мс, LSTM ~50 мс, то полный цикл «предобработка -> прогноз ARIMA и LSTM -> взвешивание мета-моделью» укладывается в среднем в 190 мс. Это приемлемо для большинства сценариев контроля качества в реальном времени, где окно измерения составляет секунды или минуты.

Выводы по эксперименту. Разработанная гибридная модель успешно решает поставленную задачу.

За счет комбинации методов она обеспечивает более высокую и, что важнее, стабильную точность прогнозирования в условиях дрейфа концептов по сравнению с классическими подходами.

Механизм адаптации на основе ADWIN позволяет системе автономно подстраиваться под изменения технологического процесса без необходимости частого ручного переобучения.

Заключение. В работе представлена и апробирована гибридная модель прогнозирования качества для потоковых данных промышленного производства.

Доказано, что интеграция статистических моделей, глубокого обучения и динамических ансамблей позволяет создать систему, устойчивую к дрейфу концептов. Разработанный механизм адаптации весов на основе детектора ADWIN обеспечивает высокую точность даже при резких изменениях режимов работы оборудования.

Перспективой дальнейших исследований является внедрение в гибридную схему более современных архитектур, таких как Трансформеры, а также использование методов физически-информированного машинного обучения для повышения интерпретируемости прогнозов.

Список литературы

[1] Гераскин М. И. Большие данные и машинное обучение в Индустрии 4.0. – М.: Открытые системы, 2017.

[2] Гудфеллоу И., Бенджио Й., Курвилль А. Глубокое обучение. – М.: ДМК Пресс, 2018. – 652 с.

[3] Догучаева С. М. Инновационное развитие искусственного интеллекта и машинного обучения в современной экономике // РИСК: Ресурсы, информация, снабжение, конкуренция. – 2019. – №1. – С. 136–138.

[4] Ицкович Э. Л. Термины автоматизации и цифровизации предприятий технологических отраслей: содержание и практическое значение // Автоматизация в промышленности. – 2020. – №4.

[5] Кудрявцев В. Б., Артемов А. А. Распознавание образов и анализ данных. – М.: Физматлит, 2011. – 368 с.

[6] Мурзагалина Г. М., Китабанов А. Применение технологий искусственного интеллекта в промышленности // Московский экономический журнал. – 2022. – Т.7, №12. – С. 474–482.

[7] Никитин А., Растопшин П. Индустрия 4.0 в России: роботы, большие данные и искусственный интеллект

[8] Рахматуллина Р.И. Основные модели машинного обучения // Инструменты, механизмы и технологии современного инновационного развития: материалы международной научно-практической конференции. – 2023. – С. 98–101.

[9] Соколинский Л.Б., Тарасов С.В. Технологии больших данных и аналитика данных. – М.: Юрайт, 2021. – 276 с.

[10] Соловьёв В. И. Искусственный интеллект и анализ данных. – М.: Финансовый университет при Правительстве РФ, 2020.

[11] Теплов А. Г. Четвертая промышленная революция и её влияние на общество. – М.: Издательство современных технологий, 2019.

Авторский вклад

Николаев Н.Ф. – разработка общей концепции гибридного подхода, включая архитектуру ансамблевой модели на основе стеккинга, реализация программного кода для потоковой обработки данных и интеграции компонентов ARIMA и LSTM, проведение вычислительных экспериментов, включая настройку гиперпараметров и анализ чувствительности модели; подготовка текста статьи, визуализация результатов.

Макаров А.В. – постановка научной проблемы, формализация требований к системе в условиях дрейфа концептов, теоретическое обоснование выбора детектора ADWIN и механизма динамической адаптации весов, анализ предметной области и существующих методов; интерпретация экспериментальных данных, формулировка выводов и научной новизны; научное редактирование рукописи.

HYBRID MACHINE LEARNING METHODS FOR PRODUCT QUALITY PREDICTION ON STREAMING DATA IN THE INDUSTRY 4.0 CONCEPT

N. F. Nikolaev

Postgraduate Student at the Department of Information Technologies, Artificial Intelligence, and Socio-Public Technologies of the Digital Society, Russian State Social University (RSSU).

A. V. Makarov

Associate Professor at the Department of Information Technologies, Artificial Intelligence, and Socio-Public Technologies of the Digital Society, Russian State Social University (RSSU), Candidate of Technical Sciences.

Abstract. In the era of industrial digitalization (Industry 4.0), a key challenge is the real-time processing of continuous streaming sensor data. Classical statistical models and standard machine learning (ML) algorithms struggle to cope with the increased volume of data and lose accuracy due to the phenomenon of concept drift, caused by equipment wear or changes in raw materials. This paper proposes a hybrid model architecture for product quality prediction, combining fast statistical methods (ARIMA), accurate deep learning models (LSTM), and a meta-model for dynamic weighting of the results. The results of experimental testing on industrial manufacturing data (SECOM dataset) demonstrate that the proposed hybrid approach increases prediction accuracy by 12–15% compared to individual models under concept drift conditions while maintaining an acceptable processing latency.

Keywords: hybrid machine learning models, streaming data, Industry 4.0, quality prediction, concept drift, LSTM, ARIMA, ADWIN, SECOM.

УДК 621.391.7: 004.9

ПРИМЕНЕНИЕ МЕТОДА МАШИННОГО ОБУЧЕНИЯ KNN ДЛЯ КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ ИНТЕНСИВНОСТИ ТРАФИКА В СЕТЯХ 5G (НА ПРИМЕРЕ ВИДЕОКОНФЕРЕНЦИЙ)



З. Джафаров
Заведующий кафедрой компьютерных технологий Азербайджанского технического университета, кандидат технических наук, доцент
zafar.cafarov@aztu.edu.az



О.В. Бойправ
Заведующий кафедрой защиты информации учреждения образования «Белорусский государственный университет информатики и радиоэлектроники», кандидат технических наук, доцент
smu@bsuir.by



Э. Имамалиев
Доцент кафедры компьютерных технологий Азербайджанского технического университета, кандидат технических наук, доцент
elman.imameliyev@aztu.edu.az

З. Джафаров

Окончил Азербайджанский институт нефти и химии (ныне Азербайджанский государственный нефтегазовый университет). Область научных интересов включает искусственный интеллект и машинное обучение, анализ данных и тестирование на основе моделей.

О.В. Бойправ

Окончила учреждение образования «Белорусский государственный университет информатики и радиоэлектроники». Области научных интересов – электромагнитное экранирование, кибербезопасность.

Э. Имамалиев

Окончил Азербайджанский технический университет. Область научных интересов охватывает современные сетевые технологии, архитектуры компьютерных сетей, методы анализа сетевого трафика и математические модели сетевых систем.

Аннотация. В работе представлены результаты исследования применения метода машинного обучения *k*-ближайших соседей (KNN) для краткосрочного прогнозирования трафика видеоконференций в сетях 5G. Актуальность исследования обусловлена тем, что трафик видеоконференций имеет нестационарный и всплесковый характер, что требует наличия точных прогнозов для управления ресурсами сетей в реальном времени и минимизации задержек. В ходе проведения исследования выполнено моделирование трафика, передаваемого по сети 5G, оценено влияние значения *k* на производительность с использованием показателей RMSE и MAE, на основании чего определено наиболее оптимальное значение *k*. На основе полученных результатов исследования установлено следующее: 1) метод KNN характеризуется простой реализации; 2) с помощью метода KNN можно выполнять краткосрочное прогнозирование интенсивности трафика в сетях 5G с высокой точностью, что позволяет оптимизировать управление ресурсами указанных сетей и сокращать задержки трафика видеоконференций.

Ключевые слова: сеть 5G, видеоконференция, прогнозирование трафика, метод ближайших соседей, машинное обучение.

Введение. В последние годы сети 5G играют важную роль в развитии беспроводной связи с высокой скоростью, низкой задержкой и широкополосным доступом. Особенно в приложениях реального времени, таких как видеоконференции, интенсивность трафика быстро меняется и носит нестационарный характер. Это требует краткосрочного прогнозирования трафика для эффективного управления сетевыми ресурсами и оптимизации пользовательского опыта. Хотя традиционные статистические методы (например, ARIMA и Холта-Винтерса) являются точными, они имеют ограничения в прогнозировании трафика с всплесковыми паттернами, в частности трафика, формируемого в рамках видеоконференций. Поэтому методы машинного обучения, особенно метод *k*-ближайших соседей (KNN), позволяют более точно прогнозировать будущие значения трафика путем анализа паттернов трафика на уровне соседних узлов. В рамках представляемой работы выполнено исследование применения метода KNN для краткосрочного прогнозирования трафика видеоконференций в сетях 5G. Цель работы состояла в установлении зависимости точности и производительности указанного метода от значений *k*. Для достижения цели были решены следующие задачи:

- построена модель KNN для трафика, формируемого в рамках видеоконференций;
- выполнена оценка точности прогнозирования трафика с использованием метода при различных значениях *k*;
- выполнена оценка производительности метода при различных значениях *k* с помощью показателей ошибок RMSE и MAE [1, 2];
- установлено наиболее оптимальное значение *k* для решения задач прогнозирования трафика, формируемого в рамках видеоконференций в сетях 5G.

Ключевые аспекты применения машинного обучения в сетях 5G. Прогнозирование краткосрочного трафика давно является предметом исследований в области управления сетями и оптимизации ресурсов. Традиционные статистические методы, такие как ARIMA и Холт-Винтерс, обеспечивают определенную точность при использовании данных временных рядов. Однако эти методы имеют ограничения в прогнозировании нестационарного трафика с всплесками, например, трафика видеоконференций.

В последние годы методы машинного обучения приобрели важное значение в прогнозировании трафика. Такие методы, как случайный лес, машины опорных векторов (SVM) и долговременная кратковременная память (LSTM), позволяют более точно изучать нелинейные и сложные закономерности трафика.

С другой стороны, метод KNN – это простой, легко интерпретируемый метод прогнозирования, который особенно интересен для приложений реального времени. Предыдущие исследования показали, что KNN хорошо работает с краткосрочным трафиком и что выбор гиперпараметров (значений *k*) напрямую влияет на точность прогнозирования. Однако исследования трафика видеоконференций в сетях 5G все еще ограничены. Существующие работы, как правило, проводились на трафике LTE или 4G, и применение KNN и других методов машинного обучения для управления ресурсами в реальном времени не получило широкого распространения. Поэтому тема применения KNN к трафику видеоконференций в сетях 5G является актуальной и инновационной.

В данной работе данные о трафике получены как на основе данных из реальных журналов событий сети 5G, так и на основе смоделированного набора данных. В ходе моделирования характеристики реального трафика имитировались путем добавления базового трафика, всплесков и случайного шума. Значения трафика за три предыдущих интервала (lag1, lag2, lag3) использовалисьв качестве входных переменных модели.

Набор данных был разделен на 80 % обучающей и 20 % тестовой частей путем разделения на обучающую и тестовую выборки. Модель KNN обеспечивает прогнозирование на основе соседних узлов, то есть значение трафика в следующем интервале рассчитывается как среднее значение *k* ближайших соседей. Были протестированы различные значения *k* (1–20), и производительность оценивалась на основе RMSE и MAE. Было определено наиболее

оптимальное значение k, и окончательные прогнозы были сделаны с использованием этого параметра.

Точность прогнозов была показана с помощью визуальных графиков, а фактический и прогнозируемый трафик сравнивались. Этот подход демонстрирует как эффективность метода KNN для краткосрочных прогнозов, так и влияние выбора гиперпараметров на производительность. Результаты показывают, что KNN обеспечивает надежные прогнозы для трафика, сформированного в рамках видеоконференции в сети 5G, при этом оставаясь простым и легким для интерпретации.

Предложенный метод KNN был протестирован на смоделированном трафике, сформированном в рамках видеоконференции в сети 5G (рисунок 1). В качестве входных признаков при тестировании использовались значения трафика за три последних интервала (lag1, lag2, lag3). Набор данных был разделен на 80 % обучающей и 20 % тестовой частей, а производительность оценивалась для различных значений k (от 1 до 20).

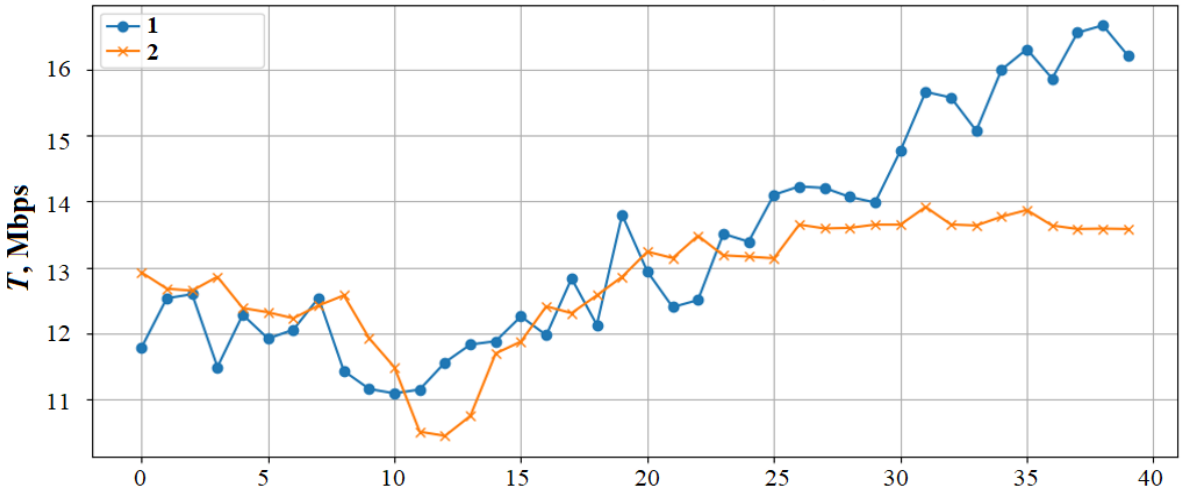


Рисунок 1. Графические представления трафика, формируемого в рамках видеоконференции в сети 5G, полученные на основе реальных данных (кривая 1) и результатов прогнозирования с использованием метода KNN (кривая 2)

Прогноз KNN в целом следует тренду реального трафика, но отстает, особенно в часы пик. Например, в интервале 25–38 наблюдается увеличение реального трафика, но KNN остается относительно стабильным (~ 13–14 Мбит/с). KNN прогнозирует более усредненное поведение и плохо улавливает внезапные всплески и спады. Это основное ограничение KNN для краткосрочного прогнозирования пикового трафика. В интервале 0–20 KNN относительно хорошо соответствует реальному трафику.

Ошибка прогнозирования (ER) для каждого k измерялась с помощью показателей RMSE и MAE. На основе анализа полученных результатов измерений (рисунок 2) установлено, что слишком малое или слишком большое значение k приводит к увеличению ошибок. Результаты также показали, что метод KNN, будучи простым и легким для интерпретации, обеспечивает высокую точность краткосрочных прогнозов. Экспериментальный анализ показывает, что как выбор гиперпараметров, так и количество прошлых интервалов, оказывают существенное влияние на точность прогноза.

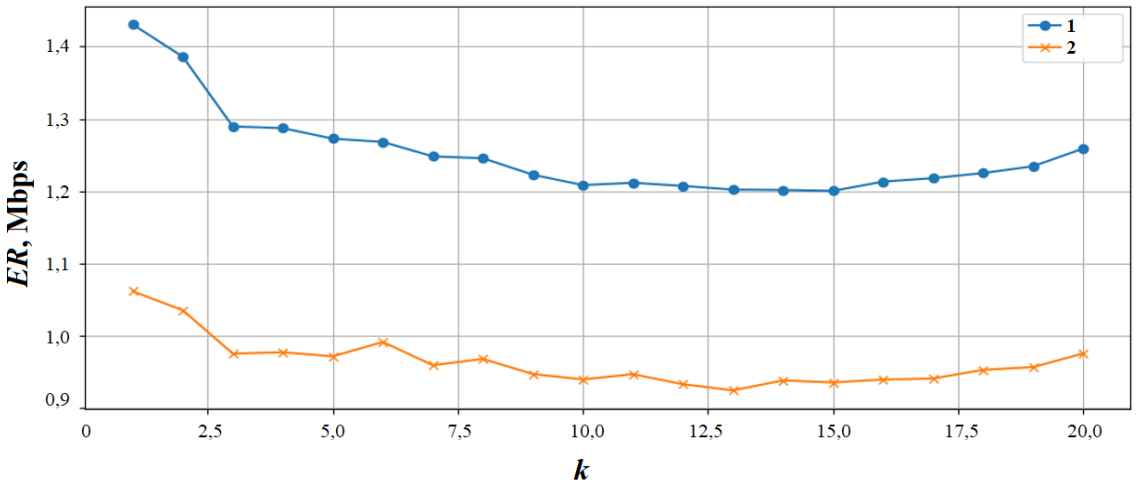


Рисунок 2. Зависимости ошибки прогнозирования от значения *k*, полученные с помощью показателей RMSE (кривая 1) и MAE (кривая 2)

Полученные результаты указывают на то, что метод KNN может эффективно применяться для управления ресурсами в реальном времени и оптимизации трафика, формируемого в рамках видеоконференций в сетях 5G. Кроме того, результаты показывают, что метод KNN обеспечивает высокую точность краткосрочных прогнозов трафика видеоконференций в сетях 5G. Оптимальное значение *k*, равное 5, минимизирует как RMSE (1,273 Мбит/с), так и MAE (0,971 Мбит/с), обеспечивая высокое соответствие между реальным трафиком и прогнозом. Слишком малое значение *k* делает прогнозы чувствительными к шуму, в то время как слишком большое значение *k* приводит к задержке прогнозов. Установлено, что метод KNN эффективно обучается как базовым, так и импульсным паттернам. Выбор количества прошлых интервалов (lag1–lag3) оказывает существенное влияние на точность прогнозирования; слишком мало информации об интервалах не обеспечивает достаточного контекста, в то время как слишком много интервалов замедляет работу модели сверх реального уровня. Простота и интерпретируемость метода делают его подходящим для управления сетью в реальном времени. Однако KNN имеет некоторые ограничения в обработке нестационарного трафика и трафика с очень высокой дисперсией.

Поэтому в будущих исследованиях можно достичь более высокой производительности с помощью гибридных моделей, таких как интеграция KNN-LSTM или онлайн-подходы KNN.

Кроме того, установлено, что метод KNN подходит как для экспериментов, так и для исследований, и может эффективно применяться для оптимизации ресурсов и сокращения задержек в трафике, формируемом в рамках видеоконференций в сетях 5G. Экспериментальная часть исследования проводилась на основе среды программирования Python. Для обработки данных, построения моделей и визуализации результатов использовался ряд научных библиотек экосистемы Python [3].

В ходе эксперимента данные временных рядов о трафике видеоконференций сначала обрабатывались с помощью Pandas, а затем обучались с помощью модели KNN, реализованной в библиотеке Scikit-learn.

После обучения модели полученные результаты прогнозирования сравнивались с реальными значениями трафика и представлялись в графической форме с использованием библиотеки Matplotlib.

Эта программная среда обеспечивала возможность повторения экспериментов, сравнительного анализа моделей и визуальной интерпретации результатов.

Заключение. Данное исследование показывает, что метод KNN может использоваться в качестве эффективного и практичного инструмента прогнозирования трафика видеоконференций в сетях 5G, а также имеет потенциал применения для оптимизации ресурсов и снижения задержки. В будущих исследованиях предлагается применять гибридные модели, например, интегрировать KNN с LSTM или тестировать онлайн-подходы KNN. Также перспективными направлениями исследований являются тестирование предложенного метода на реальных журналах событий сети 5G, адаптация модели для различных приложений видеоконференций и автоматизация оптимизации гиперпараметров.

Список литературы

[1] Chen, M., Mao, S., & Liu, Y. (2014). *Big Data: A Survey*. Mobile Networks and Applications, 19(2), 171–209.

[2] Alaa A. Hussien, Heba Nashaat & Rehab F. Abdel-Kader. Machine Learning Techniques for Network Traffic Prediction in 5G Networks: Discover Applied SciencesAims and scope. Volume 7, article number 1047, (2025)

[2] Scikit-learn: Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Авторский вклад

Джафаров Зафар – разработка концепции и методологии исследования, проведение моделирования и экспериментов, выполнение анализа результатов, подготовка первоначального варианта статьи.

Бойправ Ольга Владимировна – обзор литературы и подготовка введения, редактирование описания методологии и экспериментальных результатов, подготовка графиков и таблиц, окончательное редактирование статьи.

Имамалиев Эльман – разработка концепции и структуры исследования, описание методологии, проведение моделирования и экспериментов, анализ результатов, подготовка первоначального варианта статьи.

APPLICATION OF THE KNN MACHINE LEARNING METHOD FOR SHORT-TERM FORECASTING OF TRAFFIC INTENSITY IN 5G NETWORKS (ON THE EXAMPLE OF VIDEO CONFERENCES)

Z. Cafarov
Head, Department of Computer Technologies, Azerbaijan Technical University, PhD Associate Professor

O.V. Boiprav
Head, Information Protection Department of Computer Technologies, Educational Establishment "Belarusian State University of Informatics and Radioelectronics", PhD, Associate Professor

E. Imameliyev
Associate Professor, Department of Computer Technologies, Azerbaijan Technical University, PhD Associate Professor

Abstract. The paper presents the results of a study on the application of the *k*-nearest neighbors (KNN) machine learning method for short-term forecasting of videoconferencing traffic in 5G networks. The relevance of the study is due to the fact that videoconferencing traffic is non-stationary and bursty, which requires accurate forecasts for real-time network resource management and minimizing latency. During the study, traffic transmitted over a 5G network was modeled, the impact of the *k* value on performance was assessed using RMSE and MAE metrics, and the optimal *k* value was determined. Based on the obtained results, the following conclusions were established: 1) the KNN method is characterized by simple implementation; 2) using the KNN method, it is possible to perform short-term traffic intensity forecasting in 5G networks with high accuracy, which allows for optimized resource management of these networks and reduction of videoconferencing traffic latency.

Keywords: 5G network, video conferencing, traffic forecasting, nearest neighbor method, machine learning.

УДК 004.94:658.78

АДАПТИВНОЕ ПРОГНОЗИРОВАНИЕ СТРАХОВОГО ЗАПАСА: ОТ МОДЕЛИ К ПРОГРАММНОЙ РЕАЛИЗАЦИИ В ЭПОХУ DIGDATA



Л.В. Примакович
Ассистент кафедры экономической информатики БГУИР, магистрант
lyudmilaprimakoviych@yandex.by



С.Ф. Миксюк
Профессор кафедры математических методов в экономике БГЭУ, доктор экономических наук, профессор
smiksyuk@mail.ru

Л.В. Примакович

Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с применением методов Data Science и математического моделирования для оптимизации логистических процессов и управления цепями поставок в условиях неопределенности.

С.Ф. Миксюк

Окончила Белорусский государственный университет имени В.И. Ленина. Область научных интересов связана с моделированием и прогнозированием социально-экономических процессов, теорией логистики и управления цепями поставок, оптимизацией складских запасов в условиях неопределенности.

Аннотация. В статье представлена адаптивная модель прогнозирования страхового запаса, отличительной особенностью которой выступает идея формирования не постоянного, а плавающего страхового запаса через включение в формулу Феттера ошибки прогнозной модели спроса и непрерывного мониторинга накопленной ошибки прогноза. Такой подход обеспечивает прямую связь между системами прогнозирования и управления запасамив условиях экономической нестабильности. Для верификации модели и демонстрации ее работы разработан программный прототип на языке Python. Описаны результаты сравнительного анализа традиционной и адаптивной методик на синтезированных данных для стационарных и нестационарных временных рядов. Описана архитектура масштабирования решения для промышленной эксплуатации в условиях Big Data с использованием современных инструментов программирования и аналитики.

Ключевые слова: управление запасами, плавающий страховой запас, модифицированная формула Феттера, накопленная ошибка прогноза, модуль адаптивного прогнозирования страхового запаса, модульная архитектура программного обеспечения модели, структура данных модели.

Введение. Современная экономическая среда характеризуется высокой турбулентностью, нестационарностью спроса и возрастающей неопределенностью параметров цепочек поставок. Рост объема данных о продажах, клиентах и поставщиках, с одной стороны, и их высокая волатильность, обусловленная экономической нестабильностью, с другой, делают классические модели управления запасами с фиксированными параметрами все менее эффективными [1]. Это приводит к двум противоположным, но вместе с тем одинаково нежелательным последствиям: риску дефицита товара и упущенной прибыли или, наоборот, к избыточной иммобилизации оборотных средств в невостребованных запасах.

Особую остроту проблема приобретает в контексте Big Data: крупные ритейлеры и дистрибьютеры ежедневно оперируют информацией о движении миллионов товарных позиций (SKU) по тысячам точек продаж и складов. Математический аппарат классических моделей управления запасами, разрабатываемый в эпоху «малых данных» и требующий

стационарности спроса, предполагает ручной пересчет параметров. Однако попытка своевременно вручную корректировать страховые запасы для всех SKU сегодня является не просто трудоемкой, а физически невыполнимой задачей, ведь объемы исторических данных достигают терабайт, и их обработка требует принципиально иных подходов к хранению и вычислениям. Следовательно, возникает потребность не просто в новой теоретической модели расчета страхового запаса (СЗ), но и в ее автоматизации, способной адаптивно обрабатывать большие массивы данных и оперативно реагировать на изменения конъюнктуры в масштабах всего ассортимента.

Целью данной работы является разработка адаптивной модели прогнозирования страхового запаса для систем управления типа QR (с фиксированным размером заказа) и ее верификация на синтезированных сценариях, а также обоснование архитектуры программного решения, способного функционировать в условиях Big Data.

Обзор литературы. Теоретической базой для управления запасами служат классические работы [2, 3], в которых детально описаны модели с фиксированным размером заказа и фиксированным интервалом времени. Однако, как отмечается в [4, 5], в условиях неопределенности их эффективность снижается. Современные исследования [6, 7] указывают на необходимость интеграции систем прогнозирования и управления запасами, предлагая в качестве ключевого параметра риска использовать не дисперсию исторических продаж, а ошибку прогнозной модели. В международных исследованиях [8, 9] прослеживается устойчивый тренд на интеграцию систем прогнозирования и управления запасами, где в качестве ключевого параметра риска предлагается использовать не дисперсию исторических продаж, а ошибку прогнозной модели. Параллельно с развитием методов управления запасами, бурное развитие получило направление Big Data и Advanced Analytics. В современной логистике эти инструменты применяются для прогнозирования спроса с использованием сложных моделей, оптимизации маршрутов и анализа поведения клиентов[10]. Однако вопросы применения распределенных вычислений для динамического пересчета страхового запаса по всей номенклатуре в реальном времени остаются открытыми, что определяет актуальность данного исследования.

Методика построения адаптивной модели прогнозирования страхового запаса. Предлагаемая модель базируется на модифицированной формуле Феттера, которая учитывает два источника риска: вариацию спроса и вариацию времени выполнения заказа. Классическая формула для системы типа QR имеет вид:

$$Z_s = z\sqrt{\bar{t_n} * {\sigma_s}^2 + \overline{S^2} * \sigma_t^2} \quad (1)$$

где Z_s – объем страхового запаса, единиц;

z – число стандартных отклонений;

$\bar{t_n}$– среднее время выполнения заказа, дни;

σ_s – стандартное отклонение потребности, ед./день;

$\bar{S}$– среднее потребление, ед./день;

σ_t – стандартное отклонение времени выполнения заказа, день.

В условиях нестационарности спроса его историческое стандартное отклонение σ_s неадекватно отражает будущую динамику. В связи с этим в представленной модели предлагается в качестве меры вариации спроса использовать стандартное отклонение ошибки его прогноза:

$$\sigma_s = \sqrt{\frac{\sum_{i=1}^{n} (S_{i\ \text{факт}} - S_{i\ \text{прогн}})^2}{n-1}} \quad (2)$$

Выбор горизонта прогноза n является важным параметром, влияющим на адаптивность модели. Для условий переменчивого спроса целесообразно использовать скользящее окно,

длина n которого может устанавливаться равной либо пропорциональной времени выполнения заказа, что позволяет нивелировать случайные колебания при сохранении чувствительности к трендам. Адаптивный механизм принятия решений о пересчете параметров системы основан на анализе накопленной ошибки спроса, определяет логику пересчёта. Так, если учесть общепринятую норму отклонения в 3σ, можно прийти к следующему условию пересчета страхового запаса:

$$Z'_s = \begin{cases} z_\alpha \sqrt{t_з \sigma_s^2 + s^2 \sigma_{t_з}} - \text{для системы } QR \text{ при} \sigma_s > 3\sigma_0 \\ z_\alpha \sqrt{(t_з + \tau)\sigma_s^2 + s^2 \sigma_{t_з}} - \text{для системы } ST \text{при} \sigma_s > 3\,\sigma_0 \\ \gamma Z_s - \text{при}\ \sigma_s < -3\,\sigma_0 \\ Z_s - \text{при} |\sigma_s| < 3\,\sigma_0 \end{cases} \quad (3)$$

В представленной системе (3) 0 является среднеквадратическим отклонением прогнозного спроса, а – коэффициент снижения (должен находиться в диапазоне от 0 до 1), использование которого позволяет избежать резкого падения уровня сервиса при временном спаде спроса. Таким образом, модель переходит от периодического к событийно-ориентированному управлению, активируя пересчёт только при статистически значимых изменениях в характере спроса.

Информационно-программное обеспечение модуля адаптивного прогнозирования страхового запаса. Для апробации предложенной модели и верификации методологии разработано специализированное программное средство, реализованное на языке Python. В качестве стека технологий выбраны библиотеки для научных вычислений и анализа данных Pandas, NumPy и SciPy, а также фреймворк Streamlit, который позволил создать интерактивный веб-интерфейс для наглядной демонстрации работы модели. Разработанный прототип автоматизирует полный цикл расчета: от загрузки и предобработки исходных данных до визуализации результатов и сравнительного анализа. Пользовательский интерфейс обеспечивает загрузку исторических данных в характерном для BigData и любых ERP-систем формате CSV, содержащем обязательные поля (временная метка, фактический и прогнозный спрос, уровень запаса, размер страхового запаса, время выполнения заказа, размер заказа, точкой заказа, флаг дефицита и дата фактической поставки). Внешний вид главной страницы интерфейса до начала работы представлен на рисунке 1.

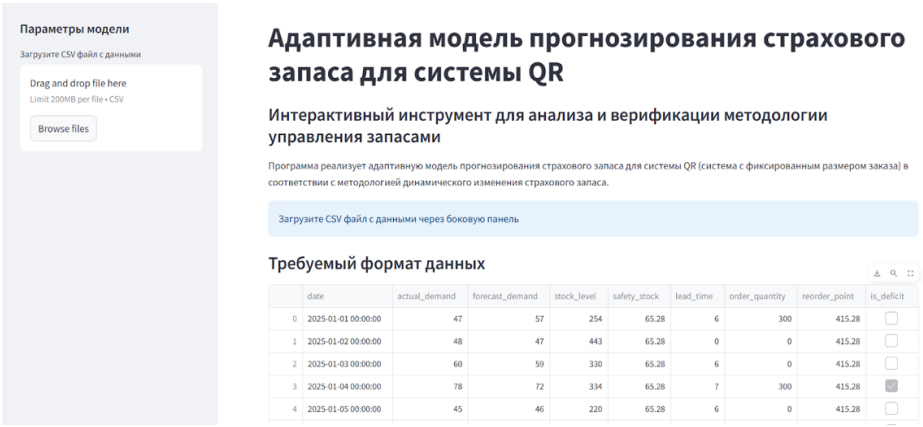


Рисунок 1. Интерфейс окна загрузки

Задать входные параметры моделирования: целевой уровень сервиса, период расчета стандартного отклонения ошибки прогноза, период анализа накопленной ошибки и коэффициент снижения запаса, – можно только после успешной загрузки файла требуемого образца. После успешной загрузки данных и задания пользователем параметров инициируется работа вычислительного ядра системы, которое инкапсулировано в отдельном классе и реализует логику расчетов по модифицированной формуле Феттера, ежедневный расчет накопленной ошибки прогноза и принятие решения о пересчете страхового запаса в соответствии с системой уравнений (3). Результатом работы разработанного прототипа является детализированный отчет, включающий графики динамики запасов, моментов пересчета страхового запаса, сравнительные гистограммы эффективности систем управления и итоговую аналитику по ключевым показателям эффективности (уровень сервиса, средний уровень запаса). Таким образом, разработанное программное обеспечение имеет модульную архитектуру, ядро которой предназначено для апробации адаптивной модели прогнозирования страхового запаса. Основными компонентами архитектуры являются:

1. Интерфейс пользователя, реализованный с использованием фреймворка Streamlit. Представляет собой интерактивную веб-панель, обеспечивающую загрузку исходных данных, настройку параметров моделирования, запуск вычислительного ядра и визуализацию полученных результатов в виде графиков и аналитических таблиц.

2. Программный модуль, представляющий собой вычислительное ядро системы, написанное на языке Python с использованием библиотек указанного языка. Оно инкапсулирует всю математическую логику адаптивной модели прогнозирования страхового запаса. На основе загруженных пользователем данных модуль рассчитывает начальные статистические параметры, такие как средний спрос, параметры времени поставки, после чего производится вычисление стандартного отклонения ошибки прогноза и накопленной ошибки за заданный пользователем период. После запуска моделирования производится сравнение накопленной ошибки с пороговым значением и, в соответствии с системой уравнений (3), определяется необходимость пересчета страхового запаса. На завершающем этапе работы программного модуля моделируется работа системы управления запасами типа QR, ежедневно обновляя уровень запаса и принятие решения о заказе с учетом адаптивно рассчитанных параметров страхового запаса и точки заказа, которые пересчитывались при достижении порогового значения накопленной ошибки прогноза спроса. На текущий момент разработана и функционирует описанная архитектура прототипа, которая полностью решает задачи верификации методологии и проведения сценарного анализа на ограниченных наборах данных. Реализация предложенной методики в реальной корпоративной среде требует создания масштабируемого программного решения.

В перспективе, для промышленной эксплуатации в условиях Big Data архитектура должна быть переработана в соответствии с принципами масштабируемости и отказоустойчивости. Вместо CSV-файлов стоит использовать распределенное хранилище или озеро данных (Data Lake), например, Hadoop HDFS или облачное хранилище (S3, ADLS). Исторические данные о продажах и прогнозах хранятся в оптимизированных форматах (Parquet, ORC). Пакетный расчет страховых запасов для всех SKUдолжен выполняться с использованием фреймворков распределенных вычислений, таких как Apache Spark. Spark позволяет распределить вычисления по кластеру, обрабатывая данные по тысячам товарных позиций параллельно, обеспечивая тем самым линейную масштабируемость кластера Dig Data системы. Ядро модели, написанное как прототип на Python, может быть интегрировано в Spark с помощью PySpark. Рассчитанные параметры (страховой запас, точка заказа) в ходе перерасчетов в условиях эксплуатации должны загружаться обратно в операционную базу данных и быть доступными для ERP либо WMS через высоконагруженный RESTAPI (например, на базе FastAPI). При этом весь процесс от динамической загрузки до получения результатов стоит оркестрировать с помощью инструментов типа Apache Airflow или Prefect, что обеспечивает регулярный своевременный пересчет параметров и мониторинг

исполнения.Такой подход позволяет автоматически обрабатывать данные по огромному количеству SKU, что недостижимо при использовании прикладных программ вроде Excel.

Результаты экспериментальных расчетов. Для оценки эффективности адаптивной модели был проведен эксперимент с использованием созданного прототипа на трех синтезированных сценариях спроса в системе управления типа QR:

1. Возрастающий тренд: фактический спрос систематически превышает прогнозный (имитация растущего рынка).
2. Убывающий тренд: фактический спрос устойчиво ниже прогнозного (имитация спада).
3. Стабильный спрос: колебания спроса вокруг прогноза носят случайный характер, близкий к нормальному распределению.

Графики входных данных, на которых хорошо видно принципиальное различие в динамике, для каждого сценария представлены на рисунке 2 а)-в).

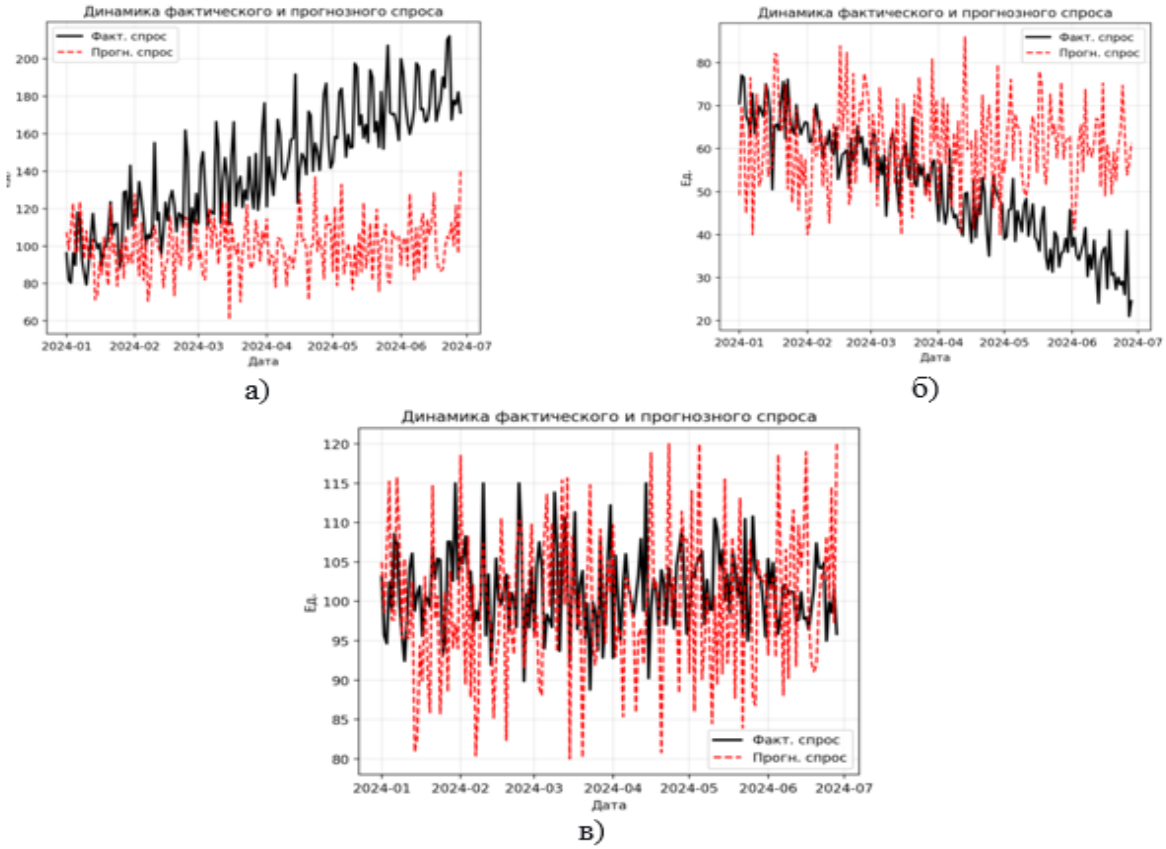


Рисунок 2. Входные данные сценариев:
а) – рост спроса; б) – спад спроса; в) – стабильный спрос

После проведения расчетов и моделирования в программном модуле формируется детализированный отчет, содержащий как графическую интерпретацию динамики запасов, так и сводную аналитику. Пример выходных данных для первого сценария (сценарий роста фактического спроса) представлен на рисунке 3 а), б). Графическая часть (рисунок 3 а)) представлена четырьмя графиками. Левый верхний график демонстрирует сравнение уровней запаса в исходной системе с фиксированным страховым запасом (черная линия) и в адаптивной системе (красная линия). Правый верхний график показывает динамику самого страхового запаса, где маркерами отмечены моменты его пересчета. Левый нижний график отражает сравнительную характеристику обеих систем по среднему запасу, среднему страховому запасу, минимальному запасу и дням дефицита. Правый нижний график

показывает распределение ошибки прогноза, т.е. распределение различия между фактическим и прогнозным уровнем ежедневного спроса. Под графической частью располагается аналитическая часть (рисунок 3 б)), в которой приводятся численные результаты расчета оценки эффективности адаптивной системы по показателям среднего запаса и дней дефицита в каждой из систем.

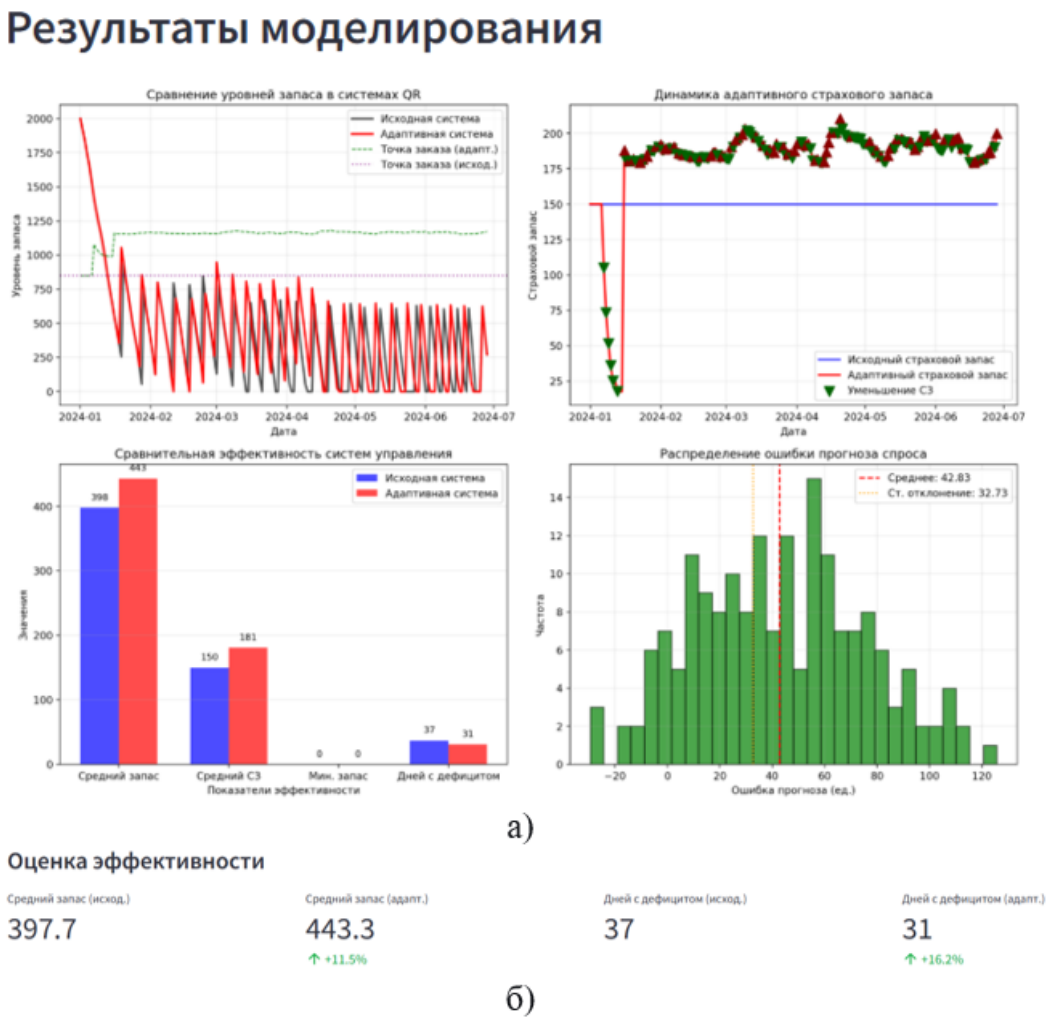


Рисунок 3. Результаты расчетов для сценария возрастающего фактического спроса:
а) – графическая интерпретация; б) – сводная аналитика

Для каждого сценария было проведено многовариантное моделирование. Целью было оценить не только общую эффективность, но и чувствительность модели к изменению ключевых параметров: целевого уровня сервиса (90%, 95%, 99%) и периода анализа накопленной ошибки (7, 14 и 30 дней). Эффективность оценивалась по двум ключевым показателям (KPI):

1. Фактический уровень сервиса (KPI_{SL}), определяемый как доля дней без дефицита в общем количестве дней моделирования.

2. Средний уровень запаса (KPI_{Stock}), отражающий средний объем оборотных средств, иммобилизованных в запасе на складе. В таблице 1 представлены усредненные результаты по девяти вариантам расчетов для адаптивной модели, сравниваемые с классической системой QR, использующей фиксированный уровень страхового запаса.

Таблица 1. Сравнительный анализ эффективности классической и адаптивной систем управления страховым запасом

Сценарий спроса	**Система управления**	**Агрегированный KPI_{SL}**	**Агрегированный KPI_{Stock}**
Возрастающий тренд	Фиксированный СЗ	0,79	397,7
	Плавающий СЗ	0,85 (+8%)	502,2 (+26%)
Убывающий тренд	Фиксированный СЗ	0,99	241,7
	Плавающий СЗ	0,98 (-1%)	212,9 (-12%)
Стабильный спрос	Фиксированный СЗ	1,0	732,6
	Плавающий СЗ	1,0 (0%)	642,6 (-12%)

Анализ результатов, представленных в таблице 1 и на выходных графиках программы, позволяет сделать следующие выводы. В сценарии возрастающего тренда адаптивная модель продемонстрировала свою основную ценность: она позволила повысить уровень сервиса с 79% до 85%, сократив дефицит на 24% (с 37 до 28 дней) за счет определения момента увеличения страхового запаса, индикатором которого в расчетах выступает накопленная ошибка прогноза, выходящая за граничный уровень Увеличение среднего запаса на 26% в данном случае определяется тем, что уровень общего запаса в рамках традиционного сценария является заниженным в связи с большим количеством дней дефицита. В сценарии убывающего тренда проявилась способность модели к оптимизации запаса. При сохранении высокого уровня обслуживания адаптивная система позволила сократить средний запас на 12%. Непрерывный мониторинг накопленной ошибки прогноза позволил определить момент снижения уровня страхового запаса, что отразилось в общем уровне запаса. Это прямое высвобождение оборотных средств, которые ранее были заморожены в избыточном страховом запасе из-за устаревших исторических данных. В условиях стабильного спроса модель подтвердила свою допустимость использования в качестве инструмента оптимизации. Сохранив стопроцентный уровень сервиса, адаптивный подход сократил средний запас на 12% по сравнению с фиксированным. Это достигается за счет того, что модель корректно идентифицирует отсутствие систематических ошибок прогноза и не поддерживает излишний «защитный» буфер.

Заключение. В результате исследования получены следующие результаты:

1. Разработано методическое обеспечение адаптивной модели прогнозирования страхового запаса, отличительной особенностью которой выступает идея формирования не постоянного, а плавающего страхового запаса через включение в формулу Феттера ошибки прогнозной модели спроса и непрерывного мониторинга накопленной ошибки прогноза, определен набор ключевых показателей эффективности расчетов.

2. Разработан модуль адаптивного прогнозирования страхового запаса для верификации модели, включая программное обеспечение на языке Python, формирование входных-выходных форм представления информации

3. На базе разработанного ПО проведены сценарные экспериментальные расчеты и выполнена их интерпретация. Апробация модели на синтезированных данных, имитирующих различное поведение спроса, подтвердила ее работоспособность и экономическое преимущество в виде динамического балансирования между риском дефицита и затратами на хранение.

4. Разработаны требования и концептуальная архитектура для масштабирования решения до промышленного уровня, способного функционировать в условиях Big Data. В отличие от разработанного прототипа, ориентированного на верификацию методологии, промышленная архитектура предполагает использование распределенного хранилища данных вместо CSV-файлов, применения фреймворков распределенных вычислений для параллельной обработки данных по тысячам SKU, а также организацию взаимодействия с корпоративными системами через высоконагруженный RESTAPI. Управление вычислительными процессами предлагается с помощью инструментов Apache Airflow для обеспечения регулярности и своевременности пересчета параметров для всего ассортимента в автоматическом режиме.

Список литературы

[1] Примакович, Л. В. Философско-методологический анализ экономической нестабильности / Л. В. Примакович // Компьютерные системы и сети : материалы 61-й науч. конф. аспирантов, магистрантов и студентов, Минск, 22–26 апр. 2025 г. / Белорус. гос. ун-т информатики и радиоэлектроники. – Минск : БГУИР, 2025. – С. 514–517.

[2] Стерлигова, А. Н. Управление запасами в цепях поставок : учебник / А. Н. Стерлигова. – 3-е изд., перераб. и доп. – М. : ИНФРА-М, 2023. – 430 с.

[3] Лукинский, В. С. Модели и методы теории логистики : учеб. пособие / В. С. Лукинский. – 3-е изд., перераб. и доп. – СПб. : Питер, 2021. – 448 с.

[4] Тяпухин, А. П. Логистика: управление цепями поставок / А. П. Тяпухин, Б. А. Аникин. – М. : КНОРУС, 2020. – 512 с.

[5] Анкуда, Е. В. Нормирование запасов в закупочной логистике в условиях неустойчивости: методики и результаты апробации / Е. В. Анкуда // Экономика, моделирование, прогнозирование. – 2020. – № 1. – С. 77–85.

[6] Миксюк, С. Ф. Управление запасами в закупочной логистике в корпоративной информационной системе: концептуальный подход / С. Ф. Миксюк, Е. В. Анкуда // Белорусский экономический журнал. – 2020. – № 1. – С. 138–147.

[7] Миксюк, С. Ф. Модель прогнозирования страхового запаса в условиях экономической нестабильности: методический подход / С. Ф. Миксюк, Л. В. Примакович // Экономика Беларуси: рост, инновации, безопасность : материалы Междунар. науч.-практ. конф., Минск, 15 мая 2025 г. / Белорус. гос. эконом. ун-т ; редкол.: А. А. Быков (отв. ред.) [и др.]. – Минск : БГУИР, 2025. – С. 419–420

[8] Praka, D. On the calculation of safety stocks when demand is forecasted / D. Praka, R. Teuntera, A. Syntetos // European Journal of Operational Research. – 2017. – Vol. 256, iss. 2. – P. 454–461.

[9] Forecasting of lead-time demand variance: implications for safety stock calculations / M. Z. Babai, Y. Dai, Q. Li [et al.] // European Journal of Operational Research. – 2022. – Vol. 296, iss. 3. – P. 846–861.

[10] Довнар, Л. А. Предиктивная логистика: прогнозирование прогнозов и управление поставками с помощью ИИ и больших данных / Л. А. Довнар, А. А. Еска ; науч. рук. Т. В. Пильгун // Развитие логистики и управления цепями поставок : материалы V Междунар. науч.-практ. молодёж. конф., посвящ. 55-летию каф. «Экономика и логистика» (в рамках Междунар. молодёж. форума «Креатив и инновации‘ 2024»), Минск, 5 дек. 2024 г. / редкол.: С. В. Скирковский, Р. Б. Ивуть, П. И. Лапковская ; сост. П. И. Лапковская. – Минск : БНТУ, 2024. – С. 268–271.

Авторский вклад

Примакович Людмила Васильевна – разработка алгоритма мониторинга накопленной ошибки прогноза, определение набора KPI показателей модели, программная реализация модели на Python, проведение вычислительных экспериментов, анализ полученных результатов, разработка архитектуры программного обеспечения модели и структуры ее данных.

Миксюк Светлана Федоровна – постановка задачи исследования, разработка математической модификации модели, общее руководство исследованием, верификация полученных выводов.

ADAPTIVE FORECASTING OF SAFETY STOCK: FROM MODEL TO SOFTWARE IMPLEMENTATION IN THE ERA OF BIG DATA

L.V. Primakovitch
Assistant at the Department of Economic Informatics, BSUIR, Master's Degree Student
lyudmilaprimakoviych@yandex.by

S.F. Miksiuk
Professor at the Department of Mathematical Methods in Economics, BSEU, Doctor of Economic Sciences, Professor
smiksyuk@mail.ru

Abstract.. The article presents an adaptive model for forecasting safety stock, the key feature of which is the idea of forming a floating safety stock, rather than a constant one. This is achieved by incorporating the demand forecast error into the Fetter formula and continuously monitoring the accumulated forecast error. This approach ensures a direct link between forecasting and inventory management systems under conditions of economic instability. To verify the model and demonstrate its operation, a software prototype was developed in Python. The results of a comparative analysis of traditional and adaptive methods on synthesized data for stationary and non-stationary time series are described. The architecture for scaling the solution for industrial operation in the era of Big Data is outlined, utilizing modern programming and analytics tools.

Keywords:. inventory management, floating safety stock, modified Fetter formula, accumulated forecast error, adaptive safety stock forecasting module, Big Data, automation, Python, Streamlit, scaling.

УДК 004.932:004.032.26

УТРАТЫ МОНУМЕНТАЛЬНОЙ ЖИВОПИСИ КАК ПОКАЗАТЕЛИ BIG DATA: ОБЪЕКТНО-ОРИЕНТИРОВАННОЕ ПРЕДСТАВЛЕНИЕ ДЕФЕКТОВ НА ОСНОВЕ МНОГОКАНАЛЬНОЙ СЕГМЕНТАЦИИ



Г.А. Хришкевич
Ассистент отделения информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ
hrishkevich.g@pskgu.ru



В.И. Пименов
Заведующий кафедрой информационных технологий, СПбГУПТД, доктор технических наук, с.н.с.
v_pim@mail.ru



Д.А. Андреев
Заведующий отделением информационно-коммуникационных технологий образовательного департамента Передовой инженерной школы гибридных технологий в станкостроении Союзного государства, ПсковГУ, кандидат технических наук
d.andreev@pskgu.ru

Г.А. Хришкевич

Окончил Псковский государственный университет. Область научных интересов связана с применением методов обработки изображений и машинного обучения для цифрового восстановления объектов монументальной живописи с сохранением культурно значимых художественных особенностей.

В.И. Пименов

Окончил Балтийский государственный технический университет «Военмех» имени Д.Ф. Устинова. Область научных интересов связана с разработкой методов и алгоритмов интеллектуального анализа данных, когнитивных технологий, компьютерных систем обработки и анализа данных, компьютерной графики, распознавания образов, мультимедиа технологий и 3D-моделирования.

Д.А. Андреев

Окончил Псковский государственный политехнический институт. Область научных интересов связана с разработкой моделей, алгоритмов и показателей качества формализованного описания и анализа технологий производства продукции.

Аннотация. В работе рассматриваются утраты монументальной живописи (трещины, пятна, лакуны) как источник структурированных данных в парадигме Big Data. Ключевым фактором является переход от пиксельной разметки к объектно-ориентированному набору данных, где каждый дефект представляется отдельной записью с геометрическими и морфологическими признаками. Особенности реализации предлагаемого подхода отражены в формализации процессов многоканальной сегментации дефектов с последующей детерминированной постобработкой и выделением связных компонентов. На примере реального изображения фрески из Спасо-Преображенского Мирожского мужского монастыря (Российская Федерация, г. Псков) получено 1038 объектов утрат при суммарной доле покрытия 33,9%, что подтверждает возможность накопления объектных показателей состояния для последующего анализа и мониторинга.

Ключевые слова: монументальная живопись, утраты, Big Data, многоканальная сегментация, связные компоненты, объектно-ориентированные данные, цифровое документирование, мониторинг состояния.

Введение. Монументальная живопись (фрески, настенные росписи, иконы) относится к уязвимым объектам культурного наследия, для которых особенно важны задачи цифрового документирования, анализа состояния и подготовки данных для реставрационных работ [1]. При работе с фотофиксацией памятников возникает необходимость перехода от описаний на уровне изображения к измеримым показателям, пригодным для сравнения между объектами и мониторинга во времени [2].

В прикладном контексте Big Data проявляется не только как большой объём пиксельных данных, но и как масштабируемый поток дискретных наблюдений, если рассматривать повреждения не как сплошную маску, а как множество отдельных объектов. Даже один фрагмент фрески может содержать сотни объектов разметки (фрагменты трещин, пятен, лакун), и каждый из них может быть описан набором признаков. Это позволяет формировать табличные датасеты и применять методы статистики и интеллектуального анализа данных для поддержки решений при реставрации и мониторинге состояния объектов монументальной живописи. Развитие методов компьютерного зрения и нейросетевой сегментации создаёт предпосылки для автоматизации выделения утрат на изображениях [3]. Для практического применения необходимо формализовать переход от результатов сегментации к объектно-ориентированным данным и показать масштаб наблюдений на реальном изображении, что согласуется с ранее предложенной вариационной моделью цифровой реставрации монументальной живописи [4]. Этим определяются постановка задачи, выбор методики построения объектно-ориентированных данных и состав количественных показателей, анализируемых в работе.

Постановка задачи. В исследовании используются цифровые изображения фрагментов фресок, полученные в ходе фотофиксации и последующей подготовки материала для анализа. Для части изображений выполнена ручная разметка областей утрат, которая служит основой для обучения и проверки алгоритмов сегментации, а также для перехода от пиксельного представления к объектно-ориентированному набору данных. Далее вводится формальная запись входных данных и результата обработки.

Пусть задано множество цифровых изображений фресок:

$$L = \{I_m\}, m = 1,2,\dots,M, \quad (1)$$

где I_m – цветное изображение размера *H*W* (высота и ширина изображения соответственно), m – индекс изображения.

Каждое изображение I_m рассматривается на дискретной области пикселей:

$$\Omega = \{1 \le i \le H, 1 \le j \le W\}, \quad (2)$$

где (i, j) – координаты пикселя.

Для цветного изображения значения пикселя задаются вектором $I_m(i,j) \in R^3$(каналы аддитивной цветовой модели *RGB*) [5].

Рассматривается множество классов утрат (типов дефектов):

$$U = \{c, s, l\}, \quad (3)$$

где *c* – класс трещин, *s* – класс пятен, *l* – класс лакун.

Требуется получить объектно-ориентированное представление утрат на каждом изображении I_m, включающее:

1. бинарные пиксельные маски утрат по классам $Y_m(\cdot,\cdot,k)$, $k \in U$;
2. множества объектов утрат (связных компонентов) $O_{m,k} = \{O_{m,k,1}, \dots, O_{m,k,N_{m,k}}\}$ для каждого класса *k*;
3. табличный датасет объектов D, в котором каждому связному компоненту соответствует отдельная запись с признаками.

Объектный датасет задаётся как:

$$D = \{(id_{img},\ id_{obj},\ k,\ x) \mid k \in U\}, \quad (4)$$

где id_{img} – идентификатор изображения (например, номер изображения m или имя файла), id_{obj} – идентификатор объекта внутри изображения (например, номер связного компонента), k – класс объекта, $k \in U$, x – набор признаков объекта.

Таким образом, в работе решается задача построения объектно-ориентированного описания утрат монументальной живописи по цифровым изображениям (на примере фресок). Результатом обработки каждого изображения являются бинарные маски по классам утрат, множество выделенных объектов утрат как связных компонентов и табличное представление этих объектов в виде записей с признаками. На основе такого представления рассчитываются количественные показатели состояния изображения: число объектов в каждом классе утрат, доля площади изображения, занятая утратами, и статистики распределения размеров объектов. Параметры бинаризации и постобработки задаются в рамках единого протокола обработки и применяются одинаково ко всем изображениям, что обеспечивает воспроизводимость вычисления показателей на серии данных.

Методология исследования. В исследовании используется протокол обработки, включающий построение вероятностных карт утрат, их дискретизацию, постобработку, выделение объектов и вычисление агрегированных показателей состояния изображения. На вход подаётся цветное изображение фрески, на выходе формируется объектно-ориентированный набор данных, содержащий идентификатор изображения, идентификатор объекта, класс утраты и набор признаков объекта, а также значения агрегированных показателей для всего изображения.

Получение пиксельных масок утрат основано на многоканальной модели сегментации [6], которая для каждого изображения I_m ставит в соответствие тензор вероятностных карт по классам:

$$f_\theta: I_m \to P_m, P_m \in [0,1]^{H \times W \times |U|}, \quad (5)$$

где f_θ – нейросетевая модель с параметрами θ, $P_m \in [0,1]^{H \times W \times |U|}$ – вероятность того, что пиксель *(i, j)* принадлежит классу *k*.

Бинаризация выполняется путём сравнения значений вероятностных карт с заданными порогами отдельно для каждого класса утрат $k \in U$:

$$\widetilde{Y_m}(i,j,k) = 1[\, P_m(i,j,k) \geq \tau_k \,], \quad (6)$$

где $\widetilde{Y_m}(i,j,k)$ – промежуточная бинарная маска класса *k*, τ_k – фиксированный порог для класса *k*, $1[\cdot]$ – индикаторная функция, равная 1 при истинности условия или в противном случае – 0.

Для получения устойчивых масок применяется детерминированная постобработка:

$$Y_m(\cdot,\cdot,k) = \Pi_k\left(\widetilde{Y_m}(\cdot,\cdot,k)\right), \quad (7)$$

где $\Pi_k(\cdot)$ – оператор постобработки канала *k*, включающий удаление малых компонентов, морфологические операции и корректировку структуры маски в соответствии с морфологией утрат данного класса. Полученные по (6) – (7) маски $Y_m(\cdot,\cdot,k)$ далее рассматриваются как итоговые бинарные пиксельные маски утрат класса *k* на изображении *m*.

Обучение параметров сегментационной модели выполняется на размеченной выборке пар «изображение–маска» и формулируется как задача минимизации эмпирического риска:

$$\theta^* = arg\ arg\ min_{\ \theta}\ \frac{1}{N}\sum_{n=1}^{N} \quad L(f_\theta(I_n),\ Y_n), \quad (8)$$

где (I_n, Y_n) – пара обучающих данных, *N* – число обучающих примеров.

Для устойчивого обучения при дисбалансе классов используется комбинированная функция потерь, включающая взвешенную по классам бинарную кросс-энтропию $L_{BCE}(w_k)$ и Dice-слагаемое $L_{Dice}(w_k)$ [7]:

$$L = \lambda_{BCE}\, L_{BCE}(w_k) + \lambda_{Dice}\, L_{Dice}(w_k), \quad (9)$$

где λ_{BCE} и λ_{Dice} – коэффициенты, определяющие относительный вклад BCE- и Dice-слагаемых, а w_k – веса классов, компенсирующие дисбаланс по числу пикселей соответствующего класса утрат.

Оптимизация выполняется методом AdamW с косинусным расписанием скорости обучения. Для повышения обобщающей способности применяются пространственные и фотометрические аугментации и нормализация входных данных [8]. Пороги τ_k и параметры постобработки Π_k определяются на подтверждённой части данных и далее фиксируются в протоколе обработки.

После получения итоговых бинарных пиксельных масок $Y_m(\cdot,\cdot,k)$ по классам утрат выполняется выделение связных компонентов, каждый из которых интерпретируется как отдельный объект утраты:

$$O_{m,k} = \{O_{m,k,1}, \dots, O_{m,k,N_{m,k}}\}, \quad (10)$$

где $N_{m,k}$ – количество объектов класса k на изображении m.

Для каждого объекта вычисляется набор признаков, включающий геометрические характеристики (например, площадь, периметр, ограничивающий прямоугольник, центр масс) и характеристики формы (например, ориентация, удлинённость). Агрегированные показатели состояния изображения рассчитываются по маскам и множествам объектов. В качестве базовых показателей используются число объектов по каждому классу утрат и доля покрытия изображения утратами данного класса соответственно:

$$N_{m,k} = \left|O_{m,k}\right|, \quad (11)$$

$$\rho_{m,k} = \frac{1}{HW}\sum_{i=1}^{H} \quad \sum_{j=1}^{W} \quad Y_m(i,j,k) * 100\%, \quad (12)$$

где $\rho_{m,k}$ – доля покрытия изображения утратами класса k, выраженная в процентах, *HW* – общее число пикселей изображения.

Полученные объектные признаки и агрегированные показатели формируют структурированное представление данных, пригодное для накопления по серии изображений и последующего анализа.

Экспериментальные результаты. Данная часть настоящей работы направлена на проверку работоспособности описанного в формализованном виде подхода по объектно-ориентированному представлению утрат и на получение количественных показателей на примере реального изображения фрески. В качестве исходных данных используется цифровое изображение фрески, которая размещена в Спасо-Преображенском Мирожском мужском монастыре (Российская Федерация, г. Псков), с выполненной многоклассовой разметкой утрат в соответствии с их классами, соответствующим перечисленным выше типам дефектов (трещины, пятна, лакуны) (рисунок 1).

Выбор данного изображения обусловлен наличием большого числа разрозненных областей утрат, что позволяет оценить масштаб объектно-ориентированного набора данных, формируемого даже на уровне одного снимка. Обработка выполняется в соответствии с описанной методологией: разметка преобразуется в бинарные маски по классам утрат, затем применяются фиксированные правила очистки масок и выделения отдельных областей утрат как связных компонентов.

Для каждого класса утрат подсчитывается число выделенных объектов и доля покрытия изображения, а также рассчитываются статистики площадей объектов, характеризующие распределение размеров утрат.

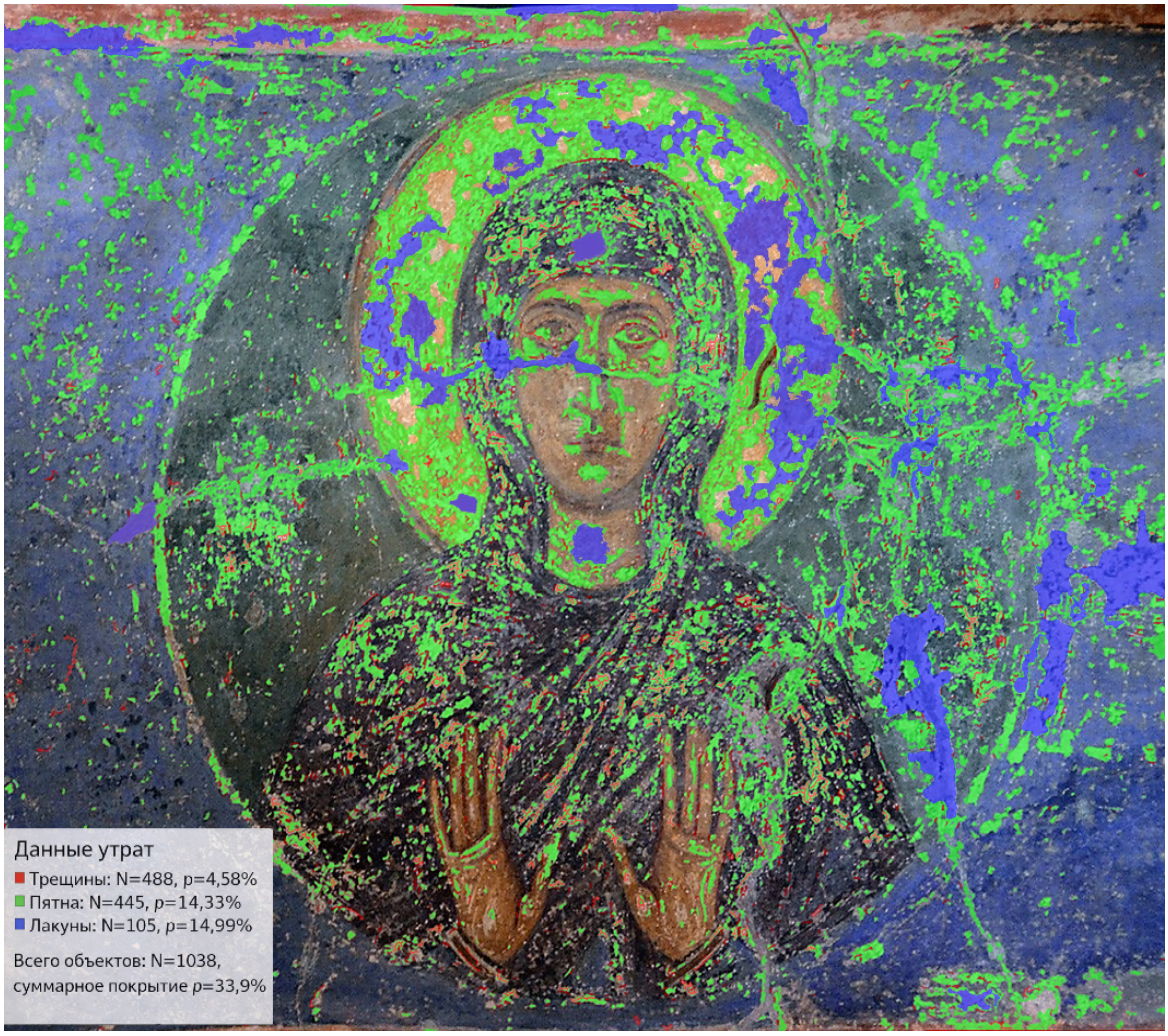


Рисунок 1. Многоклассовая разметка утрат на изображении фрески Спасо-Преображенского Мирожского мужского монастыря (Российская Федерация, г. Псков)

После дискретизации, постобработки и выделения связных компонентов рассчитаны значения показателей Nm,k и m,k, а также статистики размеров объектов по каждому классу утрат (таблица 1).

Таблица 1. Количественные показатели утрат по классам на изображении фрески

Класс утрат	Число объектов N	Доля покрытия p, %	Медианная площадь, px^2	Средняя площадь, px^2	Максимальная площадь, px^2
Трещины	488	4,58	89	320,5	25836
Пятна	445	14,33	405	1101	60741
Лакуны	105	14,99	1036	4879,96	117717
Итого	1038	33,9	249	1116,3	117717

Для изображения размером 2048*1669 пикселей получены следующие значения: число объектов утрат составило N = 488 для трещин, N = 445 для пятен и N = 105 для лакун; суммарное число объектов утрат на изображении равно N = 1038. Доля покрытия изображения утратами составила 4,58% для трещин, 14,33% для пятен и 14,99% для лакун; суммарная доля покрытия равна 33,9%.

Интенсивность объектно-ориентированного представления для данного изображения составляет около 304 объектов на 1 мегапиксель, что соответствует отношению 1038/3,42 (площадь изображения в мегапикселях) ~ 304 объектов к площади изображения в мегапикселях. Также в таблице 1 приведены медианные, средние и максимальные площади объектов в пикселях (px^2) по каждому классу утрат. Медианная площадь характеризует типичный размер объекта и устойчива к единичным крупным лакунам, средняя площадь отражает общий уровень размеров с учётом всех объектов, а максимальная площадь показывает наибольшую по площади область утраты данного класса на изображении.

Обсуждение результатов. Представление утрат как множества объектов переводит анализ изображений монументальной живописи из чисто пиксельного формата в формат структурированных данных: каждый объект становится отдельной записью с признаками, а изображение – источником агрегированных индексов состояния.

В рассмотренном примере одно изображение фрески даёт N=1038 объектов утрат при суммарной доле покрытия 33,9 % и интенсивности порядка 304 объектов на 1 мегапиксель, что уже на уровне одного снимка соответствует большому числу дискретных наблюдений и задаёт масштаб объектно-ориентированного набора данных.

Анализ показателей по классам утрат (табл. 1) демонстрирует различие их морфологии. Трещины представлены наибольшим числом объектов (488) при относительно низкой доле покрытия (4,58%) и малых площадях (медианная площадь составляет 89 пикселей), что соответствует множеству узких линейных дефектов. Пятна занимают промежуточное положение: число объектов сопоставимо с трещинами (445), доля покрытия выше (14,33%), а медианная площадь (405 пикселей) отражает более компактную, но плотную структуру.

Лакуны характеризуются наименьшим числом объектов (105) при схожей доле покрытия (14,99%) с пятнами, но существенно более высокими значениями площадей (медианная – 1036 пикселей, максимальная – 117717 пикселей), что соответствует крупным областям утрат живописного слоя. Таким образом, полученные характеристики отражают различие морфологии утрат: трещины представлены большим числом малых по площади объектов, тогда как лакуны формируют сравнительно небольшое число крупных областей. Такое разбиение утрат по классам в сочетании с объектно-ориентированными показателями (число объектов, доля покрытия, распределение площадей) позволяет не только описывать текущее состояние фрагмента, но и формировать гипотезы о механизмах разрушения и приоритетах реставрационного вмешательства. При масштабировании на серию изображений суммарное количество объектов быстро достигает десятков и сотен тысяч, что открывает возможность применения методов статистического анализа и интеллектуальной аналитики для сравнения памятников, мониторинга динамики разрушения и выявления зон повышенного риска [9, 10].

Для обеспечения воспроизводимости в исследовании используются фиксированные пороги бинаризации и единые правила постобработки, применяемые ко всем изображениям по одному протоколу. В представленной экспериментальной части расчёты выполнены по оверлею разметки. При расширении выборки те же показатели могут вычисляться по бинарным маскам, полученным непосредственно из предсказаний модели сегментации, что позволит оценить влияние качества сегментации на устойчивость объектно-ориентированных индексов состояния**.**

Заключение. Предложен подход по объектно-ориентированному представлению утрат монументальной живописи на основе многоканальной сегментации и детерминированной постобработки бинарных масок. Формализован переход от пиксельных карт утрат к множеству объектов утрат как связных компонентов и к табличному набору данных, в котором каждому объекту соответствует отдельная запись с геометрическими и морфологическими признаками. В экспериментальной части получены количественные показатели состояния реального изображения фрески, включая число объектов по классам утрат и доли покрытия, а также статистики распределения площадей объектов.

Показано, что один снимок формирует свыше тысячи объектов утрат и тем самым задаёт масштаб наблюдений, достаточный для накопления структурированных данных на серии изображений и последующего анализа в парадигме Big Data. Перспективы дальнейших исследований связаны с распространением предложенного протокола обработки на серию изображений, формированием выборок из десятков и сотен тысяч объектов утрат и применением методов статистического анализа и машинного обучения для задач кластеризации, прогнозирования динамики разрушения и поддержки принятия решений при реставрации. Отдельное направление составляет интеграция объектно-ориентированных показателей с вариационными моделями цифровой реставрации монументальной живописи и разработка комплексных индексов состояния, учитывающих как пространственное распределение утрат, так и их положение в структуре живописного слоя.

Список литературы

[1] Liu, Y. The exploration of applying of spatial information technology in General Plan for regional cultural heritage conservation / Y. Liu, Y. Zhuang // 28th CIPA Symposium “Great Learning & Digital Emotion”: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – Beijing: ICOMOS/ISPRS International Scientific Committee on Heritage Documentation, 2021. – Vol. XLVI-M-1-2021. – PP. 415–422. – DOI: 10.5194/isprs-archives-XLVI-M-1-2021-415-2021.

[2] Методика автоматизированного восстановления изображений с использованием свёрточных нейронных сетей / Г. А. Хришкевич, Д. А. Андреев, Л. В. Мотайленко [и др.] // Известия ЮФУ. Технические науки. – 2024. – № 6(242). – С. 65-76. – DOI: 10.18522/2311-3103-2024-6-65-76.

[3] Mishra, M. Artificial intelligence-assisted visual inspection for cultural heritage: State-of-the-art review / M. Mishra, P. B. Lourenço // Journal of Cultural Heritage. – 2024. – Vol. 66. – PP. 536–550. – DOI: 10.1016/j.culher.2024.01.005.

[4] Хришкевич, Г. А. Вариационная модель цифровой реставрации монументальной живописи [Электронный ресурс] / Г. А. Хришкевич // Инженерный вестник Дона. – 2025. – № 12. – URL: http://www.ivdon.ru/ru/magazine/archive/n12y2025/10627.

[5] Yuan, W. Comparative Analysis of Color Space and Channel, Detector, and Descriptor for Feature-Based Image Registration / W. Yuan, S. R. P. Poosa, R. F. Dirks // Journal of Imaging. – 2024. – Vol. 10, No. 5. – Article 105. – PP. 1–25. – DOI: 10.3390/jimaging10050105.

[6] Prathap, G. Deep Learning Approach for Building Detection in Satellite Multispectral Imagery / G. Prathap, I. Afanasyev // 2018 International Conference on Intelligent Systems (IS). – Funchal: IEEE, 2018. – PP. 461–465. – DOI: 10.1109/IS.2018.8710471.

[7] Yeung, M. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation / M. Yeung, E. Sala, C. B. Schönlieb, L. Rundo // Computerized Medical Imaging and Graphics. – 2022. – Vol. 95. – Article 102026. – PP. 1–53. – DOI: 10.1016/j.compmedimag.2021.102026.

[8] Нарцев, Д. Ю. Сравнение модифицированных методов обучения Adam в задачах оценки параметров регрессионных моделей по изображению / Д. Ю. Нарцев, А. Н. Гнеушев // Информационные технологии. – 2021. – Т. 27, № 9. – С. 461–469. – DOI: 10.17587/it.27.461-469.

[9] Sudre, C. H. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations / C. H. Sudre, W. Li, T. Vercauteren et al. // Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. – Québec: Springer, 2017. – PP. 240–248. – DOI: 10.1007/978-3-319-67558-9_28.

[10] Zafeiropoulos, C. Evaluating the Usefulness of Unsupervised Monitoring in Cultural Heritage Monuments / C. Zafeiropoulos, I. N. Tzortzis, I. Rallis et al. // Novelties in Intelligent Digital Systems: Proceedings of the 1st International Conference (NIDS 2021). – Athens: IOS Press, 2021. – Vol. 338 Frontiers in Artificial Intelligence and Applications. – P. 137–146. – DOI: 10.3233/FAIA210086.

Авторский вклад

Хришкевич Георгий Андреевич – постановка задачи исследования, формирование концепции объектно-ориентированного представления утрат монументальной живописи в логике Big Data, подготовка и анализ исходных изображений и разметки, разработка методики обработки и количественной оценки утрат, проведение эксперимента, интерпретация результатов, подготовка текста статьи.

Пименов Виктор Игоревич – постановка задачи исследования, формирование научной концепции исследования, интерпретация результатов, редактирование статьи.

Андреев Дмитрий Анатольевич – постановка задачи исследования, методическое сопровождение исследования, экспертная оценка результатов, редактирование статьи.

LOSSES OF MONUMENTAL PAINTING AS BIG DATA INDICATORS: OBJECT-ORIENTED REPRESENTATION OF DEFECTS BASED ON MULTICHANNEL SEGMENTATION

G.A. Khrishkevich
Assistant, Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU

V.I. Pimenov
Head of the Department of Information Technologies, SPbSUITD, Doctor of Technical Sciences, Associate Professor

D.A. Andreev
Head of the Department of Information and Communication Technologies, Educational Department, Advanced Engineering School of Hybrid Technologies in the Machine Tool Industry of the Union State, PskovSU, PhD of Technical Sciences

Abstract. The paper considers losses in monumental painting (cracks, stains, lacunae) as a source of structured data within the Big Data paradigm. The key idea is the transition from pixel-wise annotation to an object-oriented dataset in which each defect is represented as a separate record with geometric and visual features. The implementation of the proposed approach is based on the formalization of multichannel defect segmentation followed by deterministic post-processing and extraction of connected components. Using the example of a real fresco image from the Transfiguration of the Saviour Mirozhsky Monastery in Russian Federation in Pskov, 1038 loss objects were obtained with a total coverage of 33.9%, which confirms the possibility of accumulating object-level condition indicators for subsequent analysis and monitoring.

Keywords: monumental painting, losses, Big Data, multichannel segmentation, connected components, object-oriented data, digital documentation, condition monitoring.

UDC 004.041:004.5:004.8.

THE ROLE OF ADVANCED ANALYTICS AND BIG DATA IN AUTOMATING AND PERSONALIZING GRAPHIC DESIGN FOR DYNAMIC WEB INTERFACES IN IT SOLUTIONS



B.A. Al-Nami
Associate Professor, Department of Informatics and Computer Design, SPBSUT, Ph.D. (in technical sciences), alnomibasheer@gmail.com

B.A. Al-Nami

A graduate of Saint Petersburg Electrotechnical University "LETI", my scholarly trajectory is defined by a multidisciplinary focus on the synthesis of virtual information systems and advanced 3D polygonal modeling. My research investigates the critical intersections of Human-Computer Interaction, emphasizing the ergonomics and standardization of software products to enhance cognitive usability. By integrating principles of visual informatics and web architecture, I strive to advance the design of digital interfaces through a rigorous, user-centric lens. Currently, my work at The Bonch-Bruevich Saint Petersburg State University of Telecommunications centers on the evolution of data visualization methodologies and the optimization of research processes within complex information environments.

Abstract. The paradigm of web interface design has historically relied on static heuristics and manual A/B testing, failing to fully exploit the vast volumes of human-computer interaction data generated in modern IT solutions. This paper introduces a comprehensive computational framework that integrates Big Data processing pipelines with advanced machine learning analytics to automate and personalize graphic design elements in real-time. By framing User Interface (UI) component selection and layout generation as a combinatorial optimization and sequential decision-making problem, we deploy a Contextual Multi-Armed Bandit algorithm to dynamically adjust visual semantics based on granular user behavioral metrics. The proposed architecture bridges the gap between raw clickstream data and aesthetic rendering, providing a scalable, empirical approach to maximizing user engagement and cognitive ergonomics.

Key words: Big Data Analytics, Computational Design, Contextual Bandits, Human-Computer Interaction, Dynamic Web Interfaces.

Introduction. The intersection of computational intelligence and graphic design represents a critical frontier in modern information technology solution architecture. For decades, the discipline of interface creation has been dominated by subjective aesthetic choices and generalized usability heuristics that treat the global user base as a monolithic entity. Traditional web interfaces operate on a rigid methodology where a singular layout is expected to accommodate all individuals. These archaic frameworks are only occasionally augmented by reactive structural shifts that demand labor intensive manual configuration by human developers. Such static environments inherently fail to address the vast cognitive diversity and distinct navigational habits present within complex digital ecosystems. Furthermore, relying on manual split testing or fixed demographic rules creates a severe developmental bottleneck, rendering the platform entirely incapable of instantaneous structural evolution.

The exponential growth of user interaction data presents a monumental and largely untapped repository for predictive modeling. Modern analytical infrastructures now possess the capability to capture highly granular behavioral metrics encompassing session latency, scroll depth, intricate cursor micro movements, and specific visual focus points. Rather than discarding this massive influx of telemetry as mere analytical exhaust, advanced machine learning architectures can mine these digital footprints to uncover deep latent patterns regarding human cognition and digital friction [1,2]. By treating every navigation sequence and viewport hesitation as a continuous stream of objective feedback, computational systems can transition from passive data collection to proactive environmental synthesis. This unprecedented wealth of Big Data empowers algorithmic engines to continuously anticipate user intent, thereby dynamically reshaping the visual hierarchy, typographic scale, and chromatic composition of the web interface before the user even formulates a conscious navigational decision (fig.1).

The architectural paradigm illustrated in Figure 1 represents a sophisticated closed-loop system specifically engineered to harmonize the dual requirements of high-dimensional machine learning and the necessity for near-instantaneous interface adaptation. This self-optimizing cycle operates through a series of interconnected functional layers that transition from passive behavioral observation to active structural modification of the user environment. At the primary stage, the User Interaction Layer captures a diverse range of asynchronous behavioral signals which are immediately transmitted to a robust Data Ingestion Platform. By utilizing Apache Kafka for event streaming, the system establishes a resilient data pipeline capable of maintaining high throughput and fault tolerance during periods of extreme traffic density.

Following the ingestion phase, the telemetry stream is processed by the Analytic Processing Engine, where the workflow undergoes a critical bifurcation to balance real-time responsiveness with long-term predictive accuracy. The inference trajectory serves as the system's rapid-response mechanism, funneling time-sensitive contextual variables directly into the Autonomous UI Generator to facilitate immediate stochastic decision-making. This path enables the system to implement active machine learning policies and produce sub-second visual adjustments tailored to the specific user context. Concurrently, the training trajectory directs the aggregated data into a comprehensive Data Lake and Analytics repository for offline processing. Within this module, the ML Model Training and Feedback layer conducts intensive historical cohort analysis to recalibrate the neural weights and predictive parameters that govern the core algorithmic models.

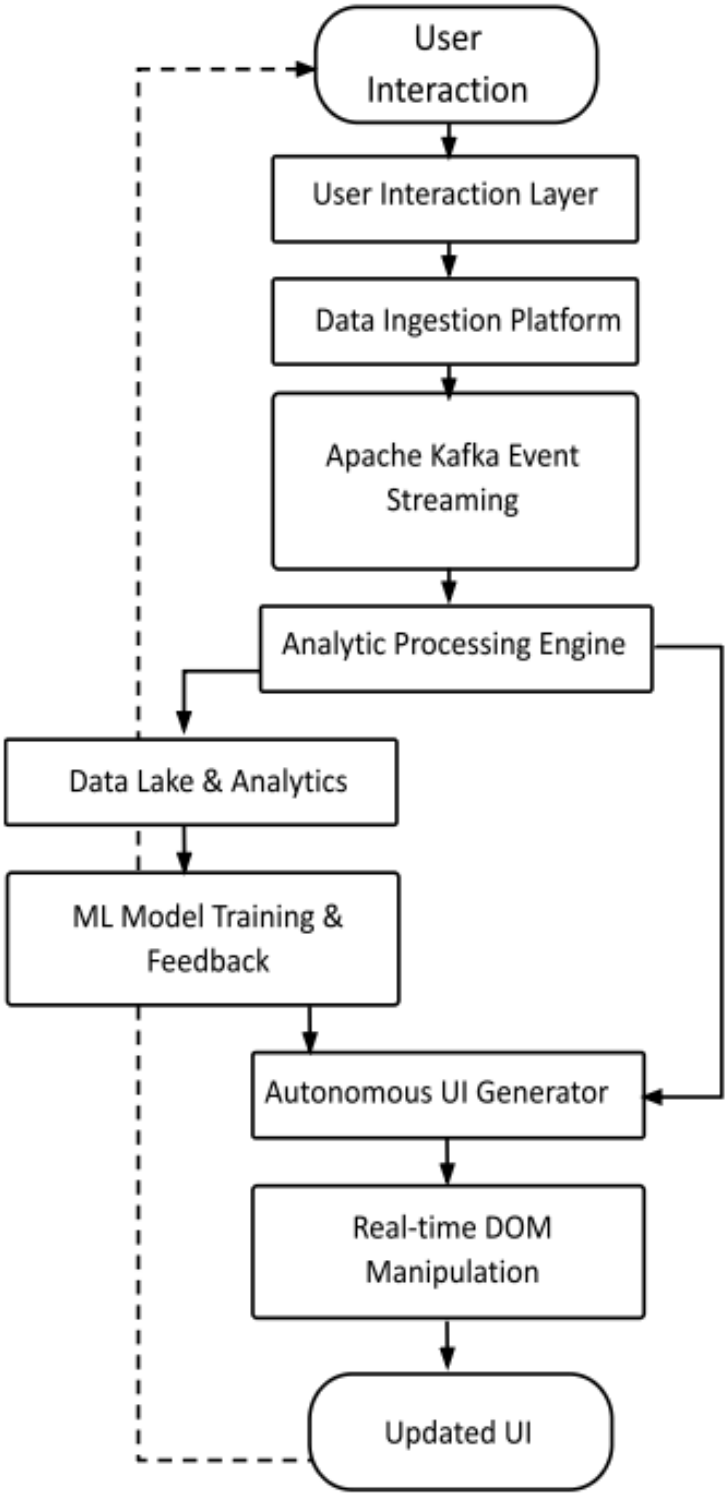


Figure 1. Closed-Loop Cybernetic Architecture for Autonomous UI Personalization.

The synthesis of optimal graphic parameters culminates in the execution of real-time DOM manipulation, which translates mathematical model outputs into tangible visual states [3].

By dynamically injecting computed CSS variables and structural HTML modifications into the Document Object Model, the framework achieves seamless layout transitions that do not require traditional browser reloads. The resulting updated UI serves as both the output of the current cycle

and the catalyst for the next interaction, as user responses to the personalized interface generate new telemetry data. This perpetual feedback loop ensures that the graphic design remains in a state of continuous evolution, systematically adapting to emergent behavioral trends through rigorous empirical validation and recursive model optimization.

State of the Art and Theoretical Background

Current literature in automated UI/UX primarily focuses on deterministic responsive design. Recent advancements in Deep Learning have introduced generative models for wireframing; however, these approaches often lack real-time dynamic applicability.

Our work departs from static generative design by focusing on continuous personalization. We formulate the aesthetic presentation of a web interface as a set of discrete state variables (formula 1)

$$S=\{s_1,s_2,\ldots,s_n\} \quad (1)$$

Corresponding to layout grids, color palettes, and typographic scales. Advanced analytics algorithms evaluate historical cohort data to predict the optimal configuration vector C^* that maximizes a predefined reward function (e.g., conversion rate or prolonged active session time).

Algorithmic Rendering Engine Bridging Computational State and Visual Synthesis

The third architectural tier serves as the critical translation layer where abstract computational intelligence materializes into tangible human-computer interactions. While the preceding analytics engine outputs a mathematically optimized multi-dimensional state vector, these numerical weights possess no intrinsic visual form. The Algorithmic Rendering Engine is tasked with decoding this mathematical output into a standardized, perceptible visual vocabulary [4,5].

This translation is achieved through the systemic implementation of parameterized design tokens. Rather than relying on static, hardcoded aesthetic values, the interface architecture is constructed upon fluid variables that dictate spatial relationships, typographic hierarchies, and chromatic distributions. When the machine learning model identifies an optimal graphical configuration for a specific user cohort, the rendering engine maps these statistical probabilities directly to the corresponding design tokens.

Consequently, the entire aesthetic and structural composition of the web interface becomes a programmable payload. The engine dynamically synthesizes the requisite cascading style properties and structural markup modifications in memory.

This generated payload is subsequently injected directly into the live Document Object Model of the client browser. By executing this visual synthesis at the token level, the framework bypasses the need for traditional server-side rendering or full-page reloads. The result is an instantaneous, frictionless repainting of the graphical interface.

This mechanism ensures that the transition between distinct personalized design states occurs with absolute fluidity, preserving the cognitive immersion of the user while seamlessly executing the data-driven aesthetic optimizations prescribed by the analytics layer (Table 2).

Table 1. Systemic mapping of user behavioral metrics to adaptive graphic design parameters within the Contextual Multi-Armed Bandit (CMAB) framework

Metric	Feature	UI Parameter	Values/Range
Attention	Dwell Time	Content Density	Compact / Expanded
Navigation	CTR	CTA Prominence	Low / High
Reading	Scroll Speed	Font Size	12pt – 24pt
Friction	Bounce Rate	Saturation	0% – 100%
Cognitive Load	Mouse Path	Grid Columns	1 – 4
Engagement	Interaction Freq.	Animation	0ms – 500ms
Conversion	Success Rate	Contrast Ratio	3:1 – 7:1

Algorithmic Formulation of Contextual Multi Armed Bandits for UI Optimization

To automate the decision-making process of graphic design personalization without repetitive manual testing, we model the system using a Contextual Multi-Armed Bandit approach. Let X_t be the context vector (user demographic and real-time behavioral features) at time t.

The algorithm selects a design action $a_t \in A$ (e.g., a specific visual layout and color scheme combination) and observes a reward r_t (e.g., successful form submission).The objective is to minimize cumulative regret over time (formula 2):

$$R(T) = \sum_{t=1}^{T} \left(r_{t,a^*} - r_{t,a_t}\right) \quad (2)$$

Where a^* is the optimal design configuration in hindsight. We utilize the Upper Confidence Bound (UCB) strategy to balance the exploitation of known high-performing designs with the exploration of novel graphical combinations [4,6].

Algorithmic Workflow of the End-to-End Processing Pipeline

To operationalize the Contextual Multi-Armed Bandit framework within a Big Data architecture, the processing pipeline must be highly structured and capable of microsecond-latency execution. The following algorithmic flowchart delineates the autonomous UI generation lifecycle, transitioning from raw data ingestion to dynamic rendering and subsequent heuristic optimization. The conceptual alignment between the macro-structural delineation of the system and its underlying algorithmic mechanics reveals a sophisticated architectural duality where the structural framework and the algorithmic workflow function as complementary representations of a single autonomous loop. While the former establishes the physical infrastructure necessary for distributed data ingestion and long-term analytics, the latter provides the rigorous mathematical logic governing the stochastic decision-making process at the heart of the engine. Specifically, the high-level Analytic Processing Engine identified in the structural framework is meticulously expanded within the algorithmic trace to reveal the Contextual Multi-Armed Bandit policy that manages the critical tension between exploration and exploitation.

This integration ensures that the mechanical capacity of the Data Ingestion Platform and Apache Kafka is directly harnessed to define the precise user context vector required for predictive modeling. Furthermore, the convergence of these two perspectives is most evident in the treatment of the feedback cycle, where the passive monitoring of user interactions is translated into actionable rewards that recalibrate the system's neural weights. In the structural view, this is presented as a continuous flow through the Data Lake and Machine Learning Training modules, whereas the logical view details the specific quantification of engagement scores and the subsequent optimization via Stochastic Gradient Descent [7].

The visual synthesis and real-time DOM manipulation described in the architectural overview are fundamentally driven by the payload generation logic found in the algorithmic workflow, confirming that the visual state of the interface is a direct manifestation of the computed mathematical optimal.

Ultimately, this dual-layered exposition demonstrates that the proposed IT solution is not merely a collection of isolated software components but a cohesive cybernetic organism where infrastructural scalability and algorithmic precision are inextricably linked to achieve perpetual UI optimization (fig.1,2).

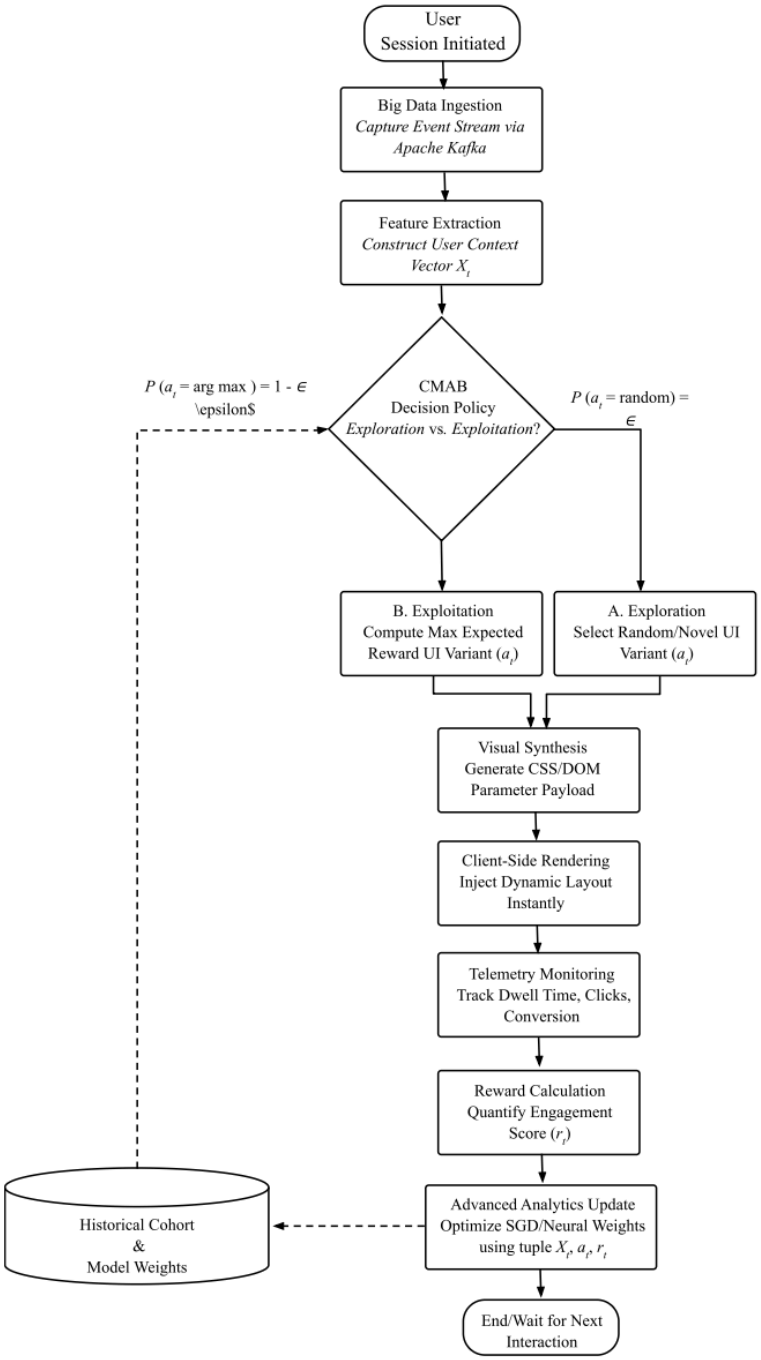


Figure 2. Algorithmic Trace of the Contextual Multi Armed Bandit Policy for Real Time UI-Personalization

Experimental Design and Validation Strategy

To rigorously validate the proposed architecture, an A/B/n testing environment is required, contrasting the dynamic analytical approach against statically designed control groups.

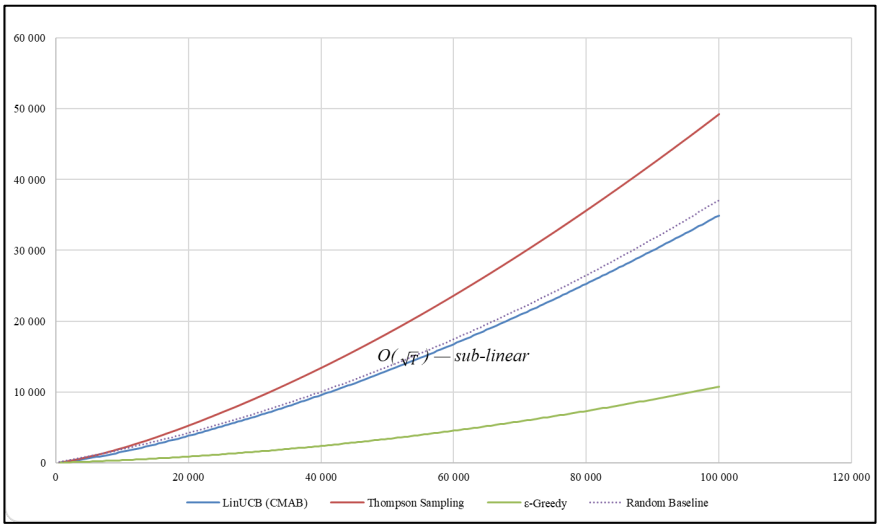


Figure 3. Cumulative Regret - Algorithm Comparison

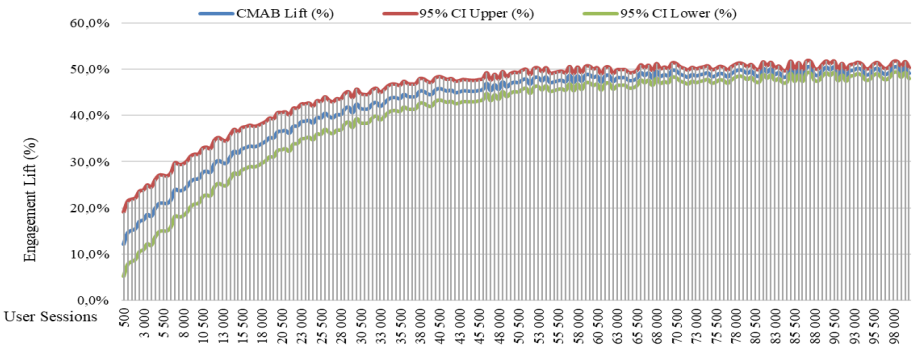


Figure 4. Engagement Lift over Baseline (%)

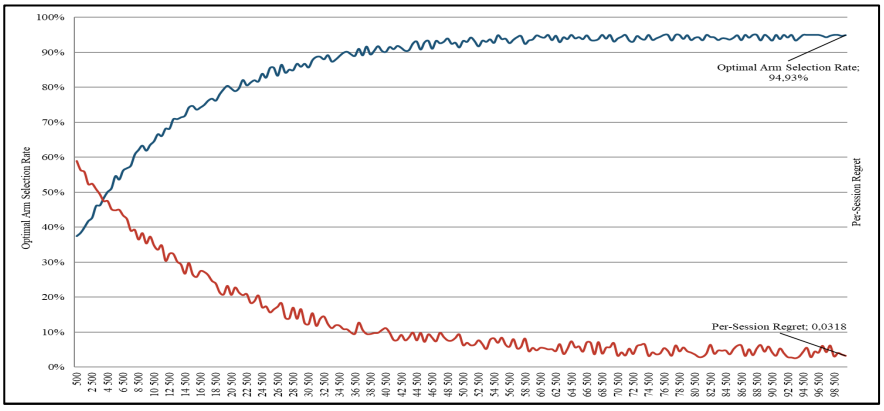


Figure 5. Optimal Arm Selection Rate & Per-Session Regret

Performance evaluation must transcend standard conversion metrics to include cognitive load approximations (derived from cursor hesitation and scroll turbulence) and system overhead latency. The Big Data cluster's throughput must be monitored to ensure that dynamic personalization does not incur a rendering penalty exceeding 200 milliseconds, which would negate the UX benefits of aesthetic optimization (fig.3,4,5).

Conclusion and Future Work. This manuscript establishes a rigorous, data-driven methodology for automating graphic design in dynamic IT solutions. By leveraging Contextual

Multi-Armed Bandits operating atop a Big Data infrastructure, we shift the paradigm of web design from intuitive manual creation to empirical, real-time algorithmic synthesis. Future lines of research will explore the integration of Federated Learning to maintain user privacy while aggregating interaction data across distributed edge networks, further refining the predictive accuracy of the automated graphic design engines without centralizing sensitive behavioral datasets.

References

[1] Al-Nami, B.A. The role of big data in graphic design and user interfaces a data-driven approach from idea to implementation // In the collection: Big Data and High-Level Analysis. collection of scientific articles of the XI International Scientific and Practical Conference. Minsk, 2025. pp. 194–201.

[2] Al-Nami, B.A., Musaeva, T. Modern Methods to Solve the Problems of Adapting Images in Web Pages and the Possibility of Using Artificial Intelligence to Do So in the Future // Current Issues in Infotelecommunications in Science and Education (APINO 2023). Collection of Scientific Articles of the XII International Scientific and Technical and Scientific-Methodological Conference. In 4 volumes. Edited by S.I. Makarenk; compiled by V.S. Yelagin, E.A. Anikevich. St. Petersburg, 2023. pp. 180–185.

[3] Al-Nami, B.A. AI – The Future Entity // Informatics, Control Systems, Mathematical and Computer Modeling (IUSMCM-2024). XV International Scientific and Technical Conference within the framework of the X International Scientific Forum of the Donetsk People's Republic. Donetsk, 2024. pp. 230–234.

[4] Al-Nami, B.A., Pigareva, Y.R. Development of the User Interface of a Logistics Company's Website // Current Issues in Infotelecommunications in Science and Education. Proceedings of the XIII International Scientific and Technical and Scientific-Methodological Conference. St. Petersburg, 2024. pp. 41–44.

[5] Al-Nami, B.A., Popova, E.V. Innovative Trends in Web Design: The Future of User Experience in a Digital Environment // Current Issues in Infotelecommunications in Science and Education. Collection of Scientific Articles of the XIII International Scientific and Technical and Scientific-Methodological Conference. St. Petersburg, 2024. pp. 506–510.

[6] Al-Nami, B.A. The Problems and Obstacles Facing the Designers of the Virtual World // Informatics, Control Systems, Mathematical and Computer Modeling (IUSMCM-2023). Proceedings of the XIV International Scientific and Technical Conference within the framework of the IX International Scientific Forum of the Donetsk People's Republic. Donetsk, 2023. pp. 367–371.

[7] Al-Nami, B.A. Converting Images from One Extension to Another Without Losing Quality: Problems and Solutions // In: BIG DATA and High-Level Analysis. Collection of Scientific Articles of the IX International Scientific and Practical Conference: in 2 Parts. Minsk, 2023. pp. 51–61.

РОЛЬ УГЛУБЛЕННОЙ АНАЛИТИКИ И БОЛЬШИХ ДАННЫХ В АВТОМАТИЗАЦИИ И ПЕРСОНАЛИЗАЦИИ ГРАФИЧЕСКОГО ДИЗАЙНА ДИНАМИЧЕСКИХ ВЕБ-ИНТЕРФЕЙСОВ В ИТ-РЕШЕНИЯХ

Б.А. Аль-Нами

Доцент кафедры информатики и компьютерного дизайна СПбГУТ, кандидат технических наук, доцент

alnomibasheer@gmail.com

Аннотация. Парадигма проектирования веб-интерфейсов исторически опиралась на статические эвристики и ручное A/B-тестирование, что не позволяло в полной мере использовать огромные объемы данных о человеко-компьютерном взаимодействии, генерируемых в современных ИТ-решениях. В данной статье представлена комплексная вычислительная платформа, интегрирующая конвейеры обработки больших данных с передовой аналитикой машинного обучения для автоматизации и персонализации элементов графического дизайна в режиме реального времени. Рассматривая выбор компонентов пользовательского интерфейса и генерацию макетов как задачу комбинаторной оптимизации и последовательного принятия решений, мы применяем алгоритм контекстных многоруких бандитов для динамической корректировки визуальной семантики на основе детализированных показателей поведения пользователей. Предложенная архитектура устраняет разрыв между необработанными данными о кликах и эстетическим рендерингом, обеспечивая масштабируемый эмпирический подход к максимизации вовлеченности пользователей и когнитивной эргономики.

Ключевые слова: аналитика больших данных, вычислительный дизайн, контекстные бандиты, человеко-компьютерное взаимодействие, динамические веб-интерфейсы.

УДК: 004.8:159.95

ВЫЯВЛЕНИЕ ПСИХОЭМОЦИОНАЛЬНЫХ ПАТТЕРНОВ НА ОСНОВЕ АНАЛИЗА БОЛЬШИХ ДАННЫХ В ЦИФРОВЫХ ПЛАТФОРМАХ ПСИХОЛОГИЧЕСКОЙ ПОДДЕРЖКИ







Т.В. Казак
Доктор психологических наук, профессор
Заведующий кафедрой инженерной психологии и эргономики
kazak@bsuir.by

А.А. Ковальчук
Студент кафедры инженерной психологии и эргономики
akovalcuk191@gmail.com

А.Н. Василькова
Старший преподаватель кафедры инженерной психологии и эргономики
a.vasilkova@bsuir.by

Т.В. Казак
Заведующий кафедрой инженерной психологии и эргономики, доктор психологических наук Республики Беларусь, доктор психологических наук Российской Федерации, член-корреспондент Международной академии психологических наук, профессор. Образование: Высшее; Аспирантура, специальность: 19.00.03 - психология труда, инженерная психология, эргономика; Докторантура, специальность: 19.00.03 - психология труда, инженерная психология, эргономика. Область профессиональных интересов / исследований: Психология труда. Инженерная психология. Эргономика. Психология управления. Юридическая психология. Социальная и организационная психология. Клиническая психология.

А.Н. Василькова
Старший преподаватель кафедры инженерной психологии и эргономики. Образование: 2007 - МГВРК по специальности «Программное обеспечение информационных технологий», 2022 - магистратура БГУИР по специальности «Охрана труда и эргономика». Область профессиональных интересов / исследований: языки программирования, искусственный интеллект, технологии виртуальной реальности, большие данные.

А.А. Ковальчук
Студент Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с разработкой цифровой платформы психологической поддержки, анализом больших данных, исследованием психоэмоциональных и поведенческих паттернов и применением интеллектуального анализа данных для персонализации цифровых сервисов.

Аннотация. В статье представлен метод выявления психоэмоционального поведения пользователя на основе анализа больших данных, которые генерируются цифровыми платформами психологической поддержки.

Ключевые слова: большие данные, цифровые платформы, психологическая поддержка, кластеризация, паттерны.

Введение. Текущий этап развития цифрового общества отличается стремительным увеличением объемов электронной информации, которая генерируется в ходе рутинной деятельности человека. Применение сети, портативных гаджетов, социальных медиа и специфических программ оставляет массу «электронных отпечатков», являющихся ценным материалом для научного исследования. В противоположность устоявшимся методикам сбора сведений в психологии (анкетирование, эксперименты в лаборатории, формализованные

беседы), цифровые следы дают возможность изучать действия индивида в привычных условиях, в темпе реального времени и на протяжении продолжительных интервалов [1, 2].

Обработка объемов информации совместно с методиками машинного обучения открывает для психологической науки совершенно новые перспективы. Если классические психологические течения зачастую оперируют усредненными метриками и лимитированными группами испытуемых, то анализ огромных массивов сведений позволяет совершить переход к исследованию личных траекторий действий и обнаружению неявных правил, невидимых при обычном статистическом рассмотрении [3]. Это особенно важно для постижения психоэмоциональных состояний, которые по своей сути являются подвижными, зависящими от обстановки и многоаспектными.

Психические недомогания являются одной из ключевых мировых задач здравоохранения. Согласно сведениям Всемирной организации здравоохранения, каждый восьмой житель планеты обречен жить с душевным расстройством, при этом доступ к высококачественной помощи остается лимитированным [3]. Электронные площадки психологического содействия (веб-сервисы, мобильные программы, чат-боты) расцениваются ныне как многообещающее дополнение к привычной структуре предоставления помощи, имеющее потенциал уменьшить трудности доступа (стигматизация, цена, наличие) и обеспечить поддержку в режиме нон-стоп [8].

Заданием данной статьи служит упорядочение и более детальное рассмотрение методологических способов обнаружения психоэмоциональных моделей на базе больших данных, которые производятся электронными площадками психологической поддержки. В труде затрагиваются концептуальные основы электронного фенотипирования, методы обработки информации, потенциал индивидуализации поддержки, а также морально-этические границы использования этих разработок.

Основная часть.

Отличие от традиционных подходов. Центральное различие между анализом больших данных и устоявшимися аналитическими методами заключается не только в объеме обрабатываемой информации, но и в кардинально ином методе работы с данными. В классических психологических поведенческих изысканиях анализ, как правило, фокусируется на заранее заданных гипотезах и ограниченном наборе параметров. Такой подход результативен для проверки определенных допущений, но малопригоден для обнаружения комплексных и многоуровневых связей, особенно при высокой нестабильности личного поведения.

Подходы анализа больших данных, наоборот, спроектированы для работы с масштабными массивами данных, где паттерны не всегда ясны на уровне единичных метрик. Применение алгоритмов кластеризации, анализа временных последовательностей и техник машинного обучения дает возможность обнаружить устойчивые структуры данных без нужды в предварительном определении строгих категорий [4]. В рамках психоэмоциональных изысканий это подразумевает способность находить поведенческие и эмоциональные закономерности, которые обнаруживаются лишь при динамическом рассмотрении и сведении данных, но остаются скрытыми при покомпонентном анализе обособленных показателей.

Можно выделить несколько фундаментальных отличий методологии анализа больших данных от традиционных психометрических подходов. Во-первых, это переход от дискретных измерений к непрерывному мониторингу: традиционные опросники фиксируют состояние «здесь и сейчас» или за ретроспективный период, тогда как цифровые платформы позволяют отслеживать флуктуации состояний в реальном времени. Во-вторых, преодолевается проблема реактивности измерений: заполнение опросника само по себе является вмешательством, способным изменить оцениваемое состояние, тогда как пассивный сбор данных (активность в приложении, паттерны набора текста) происходит естественно и незаметно для пользователя [10].

Существенно меняется и характер анализируемых переменных. Если в классической психологии исследователь оперирует преимущественно явными, декларируемыми показателями (ответы на вопросы), то анализ больших данных позволяет извлекать латентные, неосознаваемые паттерны поведения. Например, исследования с использованием приложения BiAffect показали, что анализ динамики набора текста на смартфоне (скорость печати, частота использования backspace, ритмичность нажатий) позволяет с высокой точностью (до 0.75 корреляции) выявлять эпизоды мании и депрессии у пациентов с биполярным расстройством [10]. Эти микро-поведенческие маркеры недоступны ни для самоотчета, ни для наблюдения, но могут быть зафиксированы и проанализированы цифровыми методами.

Важным преимуществом является также возможность анализа не только средних тенденций, но и вариабельности поведения. Исследования показывают, что повышенная вариабельность настроения, двигательной активности и социальных взаимодействий может быть более ранним предиктором развития психических расстройств, чем абсолютные значения этих показателей. Например, анализ данных GPS-трекинга выявляет не только снижение общей мобильности при депрессии, но и характерные изменения паттернов перемещений, включая снижение «энтропии местоположения» (разнообразия посещаемых мест), что коррелирует с тяжестью симптоматики [7].

Цифровая среда как источник данных. Цифровые платформы психологической поддержки создают уникальную пользовательскую среду, в которой данные формируются не в результате единичного измерения, а как побочный продукт регулярного взаимодействия человека с цифровым интерфейсом. Пользователи оставляют цифровые следы посредством выбора методов, частоты запросов, продолжительности сеансов, последовательности действий и реакции на различные формы поддержки. В совокупности эти данные отражают не только текущее состояние, но и динамику поведения во времени, включая периоды активности, избегания, рецидивов и изменения в моделях взаимодействия.

В отличие от лабораторных исследований или разовых опросов, цифровые платформы позволяют наблюдать за поведением пользователей в их естественной среде, без искусственного вмешательства исследователей. Это снижает влияние ситуационных факторов и эффектов социально желательных реакций, характерных для традиционных методов сбора данных. Накопление таких данных в течение длительных периодов времени формирует основу для анализа временных рядов, который может выявить устойчивые психоэмоциональные закономерности, недоступные при статистическом анализе.

Современные цифровые платформы позволяют собирать данные различных типов. Активные данные включают информацию, сознательно предоставляемую пользователем: результаты психометрических тестов, записи в дневниках настроения, когда пользователь в случайные моменты времени отмечает свое текущее состояние. Пассивные данные собираются автоматически без дополнительных усилий со стороны пользователя: акселерометрия (двигательная активность), геолокация (перемещения и социальная активность), паттерны использования устройства (время экранной активности, частота разблокировок), данные голосовых взаимодействий (темп речи, высота тона, паузы), характеристики набора текста [3].

Особую ценность представляют данные, получаемые с носимых устройств. Исследование ABCD (Adolescent Brain Cognitive Development), включавшее анализ данных с FitBit-браслетов у более чем 11 000 подростков, продемонстрировало возможность классификации психических расстройств на основе 258 статических и динамических цифровых фенотипов, включающих показатели вариабельности сердечного ритма, качества сна, интенсивности физической активности и метаболических эквивалентов [1]. Точность классификации с использованием методов глубокого обучения превзошла результаты, достигаемые при анализе данных функциональной МРТ, что открывает перспективы создания доступных и неинвазивных методов скрининга.

Важно отметить, что ценность цифровых данных определяется не столько их объемом, сколько возможностью интеграции разнородных источников в многомерную картину психологического функционирования. Современные аналитические подходы позволяют объединять данные из социальных сетей, поисковых запросов, сенсоров смартфона и структурированных самоотчетов для построения целостной модели психоэмоционального состояния пользователя [4].

Понятие психоэмоциональных паттернов. При анализе информации, генерируемой в сфере цифровой психологической поддержки, особое значение приобретает понимание психоэмоциональных шаблонов. В данном контексте такие шаблоны определяются как устойчивые и повторяющиеся комплексы эмоциональных и поведенческих реакций, которые находят свое отражение в манере, с которой индивид взаимодействует с цифровым пространством. Эти структуры описывают не единичные акты или состояния, а скорее типичные модели поведения, которые сохраняют свою предсказуемость на протяжении времени и остаются в целом неизменными, независимо от колебаний внешних обстоятельств.

Для характеризации подобных моделей всё чаще прибегают к использованию концепции, известной как фенотипы благополучия [1,2]. Термин "цифровой фенотип" был впервые предложен бывшим главой Национального института психического здоровья США, Томасом Инселом, и означает количественную оценку индивидуальных поведенческих схем, базирующуюся на непрерывном сборе данных с личных электронных устройств [9]. В отличие от традиционной клинической типологии, фенотипы не служат основанием для постановки диагноза и не призваны объяснять первопричины психоэмоционального состояния. Они фиксируют внешние проявления: режим активности, степень вовлеченности, ответную реакцию на стрессоры, а также склонность к регулярному либо отрывочному взаимодействию. Такой подход дает возможность рассматривать психоэмоциональное состояние индивида не как статичную категорию, а как изменчивый, многоаспектный процесс, что критически важно при обработке цифровых данных.

Современные изыскания выделяют ряд закрепившихся классов цифровых фенотипов, имеющих клиническую значимость. Сбои в суточных ритмах проявляются через характерные шаблоны ночного использования смартфона, непостоянство времени пробуждения и отхода ко сну, а также прерывистый характер периодов отдыха. Социальное отчуждение выражается в снижении частоты и продолжительности контактов, изменении моделей поведения в социальных сетях (преобладание пассивного просмотра вместо активного участия), и сужении спектра географических локаций. Что касается нарушений когнитивных функций, они могут обнаруживаться в замедлении или изменении точности набора текста, увеличении пауз перед ответом и специфических паттернах навигации в интерфейсе приложений [5].

Ключевой особенностью, присущей цифровым фенотипам, является их транзакционная природа: они отражают не только внутреннее состояние человека, но и характер его взаимодействия с меняющимися условиями окружения. К примеру, анализ эмоциональной дисрегуляции, проводимый посредством данных со смартфонов, дает возможность отслеживать не только субъективные переживания, но и контекст, в котором они возникают, а также используемые человеком механизмы совладания и их результативность, что закладывает основу для разработки индивидуально ориентированных вмешательств [3].

Методы выявления паттернов. Определение психоэмоциональных моделей в объёмах цифровой информации становится осуществимым при помощи применения техник кластеризации и комплексного изучения поведенческих признаков. В противовес обычным статистическим подходам, нацеленным на усреднённую картину по всей совокупности данных, кластеризация даёт возможность объединять индивидуумов по степени схожести в их динамике поведения. Более того, в качестве релевантных характеристик могут выступать не только числовые метрики, такие как частота или длительность взаимодействия, но и их временная структура, последовательность совершаемых действий и эволюция активности во времени.

Данная методология обретает особую значимость в сфере изучения психоэмоционального состояния, так как многие существенные паттерны проявляются не в абсолютных значениях единичных показателей, а в их сочетаниях и изменениях. К примеру, двое индивидов могут показывать схожий уровень вовлечённости, однако их манера взаимодействия будет кардинально отличаться: в одном случае – систематическая и устойчивая, в другом – прерывистая и ситуативная. Анализ больших массивов данных даёт возможность выявлять подобные расхождения и рассматривать их как отражение различных психоэмоциональных моделей, которые остаются нераспознанными при стандартном аналитическом обзоре.

В данной области задействуются современные парадигмы машинного обучения, охватывающие несколько групп алгоритмов. Модели обучения с учителем применяются для категоризации состояний на основании предварительно размеченных данных: например, использование алгоритмов градиентного бустинга (XGBoost) для прогнозирования диагноза на базе статических цифровых фенотипов или высокоточные нейронные сети для обработки данных временных рядов физиологической активности [1]. Методы обучения без учителя, включая различные техники кластеризации (k-means, DBSCAN, иерархическая кластеризация), обеспечивают выявление естественных групп пользователей без выдвижения априорных предположений. Методики обработки естественного языка используются для анализа текстовых сообщений, публикаций в социальных медиа и поисковых запросов с целью обнаружения признаков суицидального риска, проявлений депрессии или состояний тревоги [4, 6].

Особый интерес вызывают технологии глубокого обучения при анализе временных последовательностей. Результаты исследования, основанного на данных проекта ABCD, продемонстрировали, что архитектуры сверточных нейросетей с разделением по глубине (Xception) эффективно способны улавливать временные паттерны разной протяженности – как краткосрочные, так и долгосрочные – в многоканальных физиологических данных, что значительно повышает точность определения психических нарушений по сравнению с традиционными подходами [1].

Также значимым направлением является изучение данных, поступающих из социальных платформ и систем поиска. Работа, выполненная в Сингапуре во время пандемии COVID-19, показала, что оценка сентимент-составляющей твитов (особенно выраженности позитивных и негативных эмоций) позволяет прогнозировать количество обращений за психиатрической помощью и возникновение кризисных ситуаций с временным лагом от одного до пяти дней, причём прогностическая надёжность превышает показатели, получаемые на основе анализа официальных эпидемиологических сведений [4].

Прогностическое моделирование и обнаружение аномалий как метод раннего вмешательства. Благодаря прогрессу в области машинного обучения мы можем выйти за рамки простой ретроспективной классификации состояний и перейти к проактивному прогнозированию. На цифровых платформах психологической поддержки особое значение приобретают алгоритмы обнаружения аномалий, которые в режиме реального времени обрабатывают как пассивные, так и активные данные пользователя, чтобы выявлять отклонения от его типичного поведения, если такие отклонения имеют статистическую значимость. В отличие от подходов, основанных на предварительной маркировке данных и обучении на массивных наборах данных, обнаружение аномалий обеспечивает персонализацию анализа для каждого пользователя, применяя его собственные прошлые данные в качестве отправной точки для определения нормы [7].

Исследование, проведенное на международном уровне и включавшее пациентов с депрессией, тревожными расстройствами и шизофренией, показало высокую эффективность данного подхода к мониторингу психического здоровья. С помощью информации о местоположении, данным акселерометра и состоянии экрана, собираемой приложением mindLAMP, ученым удалось предвидеть существенные изменения в симптоматике

(увеличение или уменьшение scores по шкале PHQ-9 на 4 и более балла) за неделю до того, как эти изменения были зарегистрированы традиционными клиническими опросниками.

Площадь под ROC-кривой (AUC) для прогнозирования изменений тревоги достигла 0.80, что указывает на высокую прогностическую ценность метода. Это открывает возможности для создания систем «раннего предупреждения», которые могут сигнализировать пользователю или его куратору о приближающемся ухудшении состояния, позволяя своевременно применить профилактические интервенции.

Ключевым преимуществом данного подхода является его транодиагностическая применимость: алгоритмы, обученные на одной группе расстройств, демонстрируют надежность при переносе на другие нозологии. Это подтверждает гипотезу о том, что цифровые фенотипы отражают скорее базовые патофизиологические и поведенческие процессы (например, психомоторную заторможенность или социальное избегание), нежели жесткие категории DSM или МКБ, что делает их идеальным инструментом для мониторинга в гетерогенных популяциях пользователей цифровых платформ [7, 10].

Анализ темпоральной структуры поведения. Современная дискуссия в области цифровой психиатрии смещается от простого сбора данных к поиску принципиально новых типов переменных. Как убедительно показано в анализе, вариабельность поведения часто информативнее средних значений. Однако новейшие исследования предлагают пойти дальше и перейти от анализа статической вариабельности (так называемых маркеров «второго порядка», например, среднесуточной энтропии настроения) к анализу динамических маркеров «третьего порядка» [2].

Если статические измерения дают «моментальный снимок» сложности системы (насколько поведение разнообразно в принципе), то динамические маркеры отвечают на вопрос: «Как генерируется и регулируется эта сложность во времени и в зависимости от контекста?». Предлагается использовать методы моделирования информационной динамики, где поведенческий поток (например, последовательность локальных и глобальных перемещений, паттернов набора текста или взаимодействия с приложением) рассматривается как последовательность символов. Анализируя переходы между этими символами с помощью цепей Маркова, можно извлекать такие параметры, как «время перемешивания» (mixing time) – скорость, с которой система переходит от одного устойчивого паттерна к другому, или «время возврата» (recurrence time) в состояние с низкой сложностью.

Применительно к платформам психологической поддержки это означает возможность количественно оценивать ригидность поведенческих паттернов.

Например, для пользователя с депрессией будет характерно не просто снижение активности (первый порядок) или ее малая вариабельность (второй порядок), но и патологически долгое «застревание» в состоянии гипоактивности даже при появлении позитивных внешних стимулов (например, после получения поддерживающего сообщения). Такие маркеры третьего порядка напрямую коррелируют с ключевыми диагностическими критериями –длительностью эпизода и неспособностью адаптироваться к изменяющимся условиям, что делает их мощным инструментом для тонкой диагностики и оценки эффективности интервенций в реальном времени [2, 10].

Интеграция мультимодальных данных и генетических коррелятов. Наиболее перспективным направлением развития анализа больших данных в психологической поддержке является интеграция разнородных источников информации для создания целостной модели психологического функционирования. Современные исследования выходят за рамки анализа только лишь данных со смартфонов, объединяя их с показателями с носимых устройств и даже генетической информацией. Проект ABCD (Adolescent Brain Cognitive Development), анализирующий данные более 11 000 подростков, наглядно демонстрирует мощь такого подхода.

Используя более 250 статических и динамических признаков, извлеченных из FitBit-браслетов (вариабельность сердечного ритма, качество сна, уровни физической активности),

исследователи смогли классифицировать экстернализующие (СДВГ) и интернализующие (тревожность) расстройства с точностью, превосходящей анализ данных функциональной МРТ [1].

Более того, использование цифровых фенотипов в качестве количественных признаков (а не бинарных клинических диагнозов) в полногеномном поиске ассоциаций (GWAS) позволило выявить 16 значимых генетических локусов и 37 генов, связанных с психическими расстройствами, включая гены ELFN1 и ADORA3, ранее не обнаруживаемые при традиционном анализе.

Это доказывает, что непрерывные, объективные цифровые фенотипы обладают большей статистической мощностью для выявления генетической архитектуры психических заболеваний, чем субъективные категориальные диагнозы.

Для платформ психологической поддержки это открывает перспективы стратификации пользователей не только на основе поведения, но и на основе генетической предрасположенности (при условии соблюдения строжайших этических норм).

Пилотное исследование Mobigene уже продемонстрировало принципиальную осуществимость сбора и связывания данных цифрового фенотипирования со смартфонов с клиническими и генетическими данными у молодых людей с историей депрессии, подтвердив, что участники готовы участвовать в таких исследованиях, а показатели соблюдения протокола не зависят от тяжести текущих симптомов.

Интеграция данных с носимых устройств, смартфонов и генетических профилей позволит перейти к действительно прецизионной психологической поддержке, где вмешательства подбираются с учетом уникального биоповеденческого профиля пользователя [1, 3].

Парадигма точной психологической поддержки. Традиционная психиатрия и клиническая психология исторически опираются на категориальные системы классификации, такие как DSM или МКБ, где расстройства определяются как набор симптомов, достигших определенного порога.

Однако анализ больших данных, генерируемых цифровыми платформами, ставит под сомнение универсальность такого подхода, демонстрируя, что два человека с одним и тем же диагнозом могут иметь принципиально различные цифровые фенотипы [3]. Это наблюдение лежит в основе перехода к стратифицированной и персонализированной (точной) психологической поддержке (precision psychiatry).

Европейский проект MENTALPRECISION, финансируемый Европейским исследовательским советом, предлагает принципиально новую методологию, направленную на преодоление гетерогенности психических расстройств.

Вместо поиска усредненных биомаркеров для нозологических групп, исследователи разрабатывают инструменты анализа следующего поколения, которые позволяют изучать механизмы расстройств на пересечении нейробиологии (нейровизуализация десятков тысяч испытуемых) и количественных поведенческих показателей, получаемых методом цифрового фенотипирования со смартфонов.

Ключевая идея здесь – создание «универсальной платформы для понимания общих и различных механизмов психических расстройств на уровне индивида».

Это подразумевает использование генеративных моделей для построения «карт роста мозга» (brain growth charting), что позволяет моделировать клинические состояния и даже тестировать гипотетические интервенции на синтетических данных, прежде чем применять их на практике.

Для цифровых платформ психологической поддержки этот сдвиг парадигмы означает возможность выхода за рамки простого скрининга симптомов по шкалам (например, PHQ-9 для депрессии). Используя методы машинного обучения, такие как Персонализированный индекс преимущества (Personalized Advantage Index, PAI), платформы потенциально могут отвечать на вопросы не только «что чувствует пользователь?», но и «какое вмешательство

будет наиболее эффективно для этого пользователя в его текущем контексте?». PAI, изначально разработанный для предсказания эффекта психотерапии на основе данных рандомизированных клинических испытаний, может быть адаптирован для анализа цифровых следов.

Это позволяет стратифицировать пользователей не по диагнозам, а по прогностическим биоповеденческим кластерам, выделяя тех, кому больше подойдут когнитивно-поведенческие техники, практики осознанности или социальная поддержка.

Таким образом, анализ больших данных перестает быть просто инструментом наблюдения и становится основой для принятия клинически значимых решений в режиме реального времени, реализуя принцип «нужного вмешательства в нужное время» (just-in-time adaptive interventions).

Разработка алгоритмов обратной связи и рекомендательных систем на основе поведенческих данных. Определение психоэмоциональных моделей и фенотипов благополучия формирует техническую базу для создания интеллектуальных систем поддержки, способных не только отслеживать показатели, но и активно общаться с пользователем. Речь идет о разработке рекомендательных систем, схожих с теми, что задействованы в e-commerce или стриминговых сервисах, но адаптированных для ментального здоровья. Их задача – предоставлять потребителю индивидуализированные и уместные психологические вмешательства, методики саморегуляции либо смены образа жизни.

Инициатива My Atlas, развиваемая в Северо-Восточном университете США, демонстрирует реализацию такого подхода. Приложение трансформирует сведения, уже собираемые гаджетами и носимыми устройствами (геопозиция, пульс, логи звонков и сообщений, активность), в персональные поведенческие советы.

Главное отличие от общих наставлений (вроде «проходить 10 000 шагов») в том, что вычислительные модели My Atlas учатся на личных шаблонах пользователя.

Система обнаруживает количественные связи между объективными метриками (снижение подвижности, сбой суточных ритмов) и личными отчетами (ухудшение душевного состояния).

На основе этих взаимосвязей формируются практичные, осуществимые и личностные подсказки. К примеру, если анализ данных показывает, что человек ощущает себя хуже после суток с ограниченным набором локаций и отсутствием контактов, программа может предложить ненадолго выйти на улицу или напомнить о необходимости связаться с приятелем.

С технической точки зрения это требует внедрения методов обнаружения отклонений для фиксации расхождений с нормой и алгоритмов обучения с подкреплением для усовершенствования предложенных тактик. Важно подчеркнуть, что результативность подобных систем прямо зависит от точности и значимости выявленных психоэмоциональных шаблонов.

Если поведенческий кластер трактуется как «уклонение от социума», платформа должна предложить не стандартное напутствие «быть энергичнее», а умеренное, поэтапное воздействие, учитывая теперешний уровень воли и ментальной усталости человека. Это превращает цифровую площадку из простого инструмента сбора информации в деятельного помощника, содействующего выработке гибких поведенческих схем.

Персонализация. Анализ больших данных с целью выявления психоэмоциональных паттернов открывает путь к переходу от универсальных подходов в психологической поддержке к её персонализированным формам.

В данном контексте персонализация не сводится к простому выбору контента на основе формальных критериев, а предполагает учет устойчивых поведенческих характеристик и динамики эмоционального состояния пользователя.

Понимание того, как пользователь взаимодействует с цифровой средой во времени, позволяет гибко корректировать интенсивность, форму и последовательность поддерживающих вмешательств в соответствии с его текущим состоянием и возможностями.

Современные персонализированные системы психологической поддержки могут реализовывать различные стратегии адаптации.

Контекстно-зависимые интервенции активируются при обнаружении специфических триггеров: например, при выявлении паттернов ночной активности, указывающих на нарушение сна, система может предложить релаксационные практики или рекомендации по гигиене сна.

Адаптивные стратегии обратной связи предполагают изменение тона, сложности и объема сообщений в зависимости от текущего эмоционального состояния пользователя, определяемого по характеристикам его взаимодействия с интерфейсом. Прогностическое моделирование позволяет выявлять ранние признаки ухудшения состояния и своевременно предлагать профилактические интервенции [10].

Примером такой персонализированной системы является приложение BiAffect, которое анализирует динамику набора текста и предоставляет пользователям визуализацию их активности, помогая осознавать связь между паттернами использования смартфона и эмоциональным состоянием. Другой пример – платформа POISE, анализирующая историю поисковых запросов для выявления паттернов, связанных с суицидальными мыслями, и предоставления своевременной поддержки [6].

Важно отметить, что персонализированные вмешательства, основанные на анализе психоэмоциональных паттернов, не являются заменой профессиональной психологической помощи. Их назначение – поддержка, мониторинг и снижение барьеров для обращения к квалифицированным специалистам. В этом смысле анализ больших данных выступает не как инструмент автоматизированной оценки личности, а как средство более точного и глубокого понимания опыта пользователя в цифровой среде.

Перспективным направлением является интеграция данных из различных источников для создания комплексной модели функционирования пользователя. Исследовательский проект Digital Phenotyping of Emotion Dysregulation, финансируемый Немецким научно-исследовательским сообществом, объединяет данные экологической моментной оценки, сенсоров смартфона и носимых устройств для предсказания риска неэффективности терапии у пациентов с тревожными и депрессивными расстройствами, что позволит оптимизировать выбор терапевтических стратегий на ранних этапах лечения [3].

Ограничения и этика метода. Несмотря на высокий аналитический потенциал, внедрение анализа больших данных в сфере психоэмоционального благополучия сопряжено с рядом методологических и этических препятствий.

Во-первых, обнаруженные закономерности отображают наблюдаемые формы поведения в цифровом пространстве, но не позволяют дать однозначное толкование лежащих в основе психологических причин этих проявлений. Схожесть в поведении пользователей не всегда влечет за собой идентичность в их субъективном опыте, что требует осмотрительности при формулировке умозаключений и наставлений.

Во-вторых, анализ личных сведений пользователей требует строгих условий в отношении приватности, обезличивания и информационной безопасности. Особую тревогу вызывает потенциальная возможность использования методов цифрового фенотипирования не только в терапевтических целях, но и для скрытого надзора и контроля. Как было отмечено на форуме Федерации европейских нейронаук, раскрытие возможностей смартфонного фенотипирования возбуждает не только научный энтузиазм, но и закономерное беспокойство касательно потенциального всеобъемлющего надзора за состоянием отдельных граждан и общества в целом [9].

Важным этическим требованием выступает информированное согласие и прозрачность использования данных. Пользователи обязаны ясно понимать, какие сведения собираются, как они обрабатываются и с какой целью могут быть применены.

Проект POISE реализует фундаментально важный подход, при котором обработка данных происходит на устройстве пользователя без пересылки информации на внешние сервера, что уменьшает риски утечки личной информации [6].

Дополнительные ограничения относятся к репрезентативности данных. Цифровые площадки могут быть менее доступны для некоторых слоев населения (пожилые люди, лица с низким социально-экономическим статусом, обитатели территорий с ограниченным доступом к интернету), что создает опасность системных перекосов в данных и ограничивает возможность распространения результатов. Изучение KHP Youth Mental Health Hub, включающее свыше 23 миллионов взаимодействий, выявило значительные территориальные различия в доступности цифровых служб психологической помощи, что требует учета при интерпретации данных [2].

Используемые модели не должны становиться инструментом автоматической классификации или скрытой оценки ментального состояния человека. С научной позиции, корректно считать обнаруженные психоэмоциональные закономерности вероятностными и зависящими от контекста, а не абсолютными личностными чертами. Принятие во внимание этих ограничений является неотъемлемым условием для ответственного использования анализа больших данных на цифровых платформах психологической поддержки. Особого внимания заслуживает проблема алгоритмической непредвзятости.

Исследования демонстрируют, что модели машинного обучения, тренированные на данных преимущественно европеоидной популяции, могут показывать существенно меньшую точность при приложении кинематическим группам. К примеру, анализ голосовых индикаторов депрессии требует настройки с учетом языковых и культурных особенностей, чтобы избежать систематических погрешностей [10].

Заключение. Проведенный анализ доказывает, что задействование Big Data аналитики в сфере цифровых сервисов психологической помощи открывает беспрецедентные перспективы для исследования психоэмоционального фона пользователей.

В противоположность классическим методикам, которые фокусируются на единичных замерах и заранее утвержденных классификаторах, анализ обширных объемов информации о поведении дает возможность обнаруживать стабильные психоэмоциональные схемы и типы субъективного благополучия, которые отражают динамику взаимодействия индивида с цифровым пространством.

Идентификация подобных паттернов формирует фундамент для более адаптивной и индивидуально настроенной помощи, базирующейся на фактических поведенческих особенностях пользователей, а не на усредненных схемах.

Тем не менее, научно обоснованное и этичное применение данной методологии обязывает учитывать пределы такой интерпретации и неукоснительно следовать правилам обработки конфиденциальной информации.

Продолжение совершенствования подходов к анализу больших данных в тандеме с ответственным отношением к их эксплуатации способно подтолкнуть создание более точного и чуткого инструментария для цифровой психосоциальной поддержки [3,5].

Приоритетными векторами для дальнейших изысканий выступают создание алгоритмов федеративного обучения, которые позволят производить анализ данных без их консолидации в одном месте, разработка интерпретируемых для врачей и клиентов моделей искусственного интеллекта, а также реализация продольных изысканий для оценки влияния цифровых вмешательств на протяжении длительного времени.

Объединение данных из разнообразных источников – начиная от гаджетов с функцией трекинга и заканчивая электронными медкартами – даст возможность сформировать всеобъемлющее представление о психическом состоянии и откроет новые горизонты в области профилактики, своевременного обнаружения и индивидуализированного лечения ментальных нарушений.

Список литературы

[1] Insel T. R. Digital phenotyping: Technology for a new science of behavior. JAMA. 2017;318(13):1215–1216. DOI: 10.1001/jama.2017.11295.

[2] Onnela J.-P., Rauch S. L. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. 2016;41(7):1691–1696. DOI: 10.1038/npp.2016.7.

[3] Lazer D., Pentland A., Adamic L., Aral S., Barabási A.-L., Brewer D. et al. Computational social science. Science. 2009;323(5915):721–723. DOI: 10.1126/science.1167742.

[4] Shatte A. B. R., Hutchinson D. M., Teague S. J. Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine. 2019;49(9):1426–1448. DOI: 10.1017/S0033291719000151.

[5] Kitchin R. Big Data, new epistemologies and paradigm shifts. Big Data & Society. 2014;1(1):1–12. DOI: 10.1177/2053951714528481.

[6] Vlisides-Henry R. D., Gao M., Thomas L., Kaliush P. R., Conradt E., Crowell S. E. Digital Phenotyping of Emotion Dysregulation Across Lifespan Transitions to Better Understand Psychopathology Risk. Frontiers in Psychiatry. 2021. DOI: 10.3389/fpsyt.2021.689662.

[7] Torous J., Kiang M. V., Lorme J., Onnela J. P. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Translational Psychiatry. 2017;7(3):e1053. DOI: 10.1038/tp.2017.25.

[8] World Health Organization. World mental health report: transforming mental health for all. Geneva: WHO; 2022.

[9] Дьяконова В. Е. Новости биологии развития: отчет о Форуме Федерации европейских нейронаучных обществ. ИБР РАН; 2018.

[10] Leow A., et al. BiAffect: A smartphone-based digital phenotyping platform for mood disorders. IEEE Spectrum. 2024.

Авторский вклад

Казак Тамара Владимировна – постановка научной проблемы исследования, формирование концепции изучения психоэмоциональных паттернов в цифровых средах, научное руководство исследованием и методологическое сопровождение работы.

Василькова Анастасия Николаевна – участие в разработке исследовательского подхода к анализу больших данных, анализ современных технологий обработки данных и методов кластеризации, подготовка материалов для обсуждения перспектив применения цифровых платформ психологической поддержки.

Ковальчук Андрей Алексеевич – формирование концепции исследования, анализ и обобщение научных источников, разработка логики исследования, написание и редактирование текста статьи.

IDENTIFICATION OF PSYCHOEMOTIONAL PATTERNS BASED ON BIG DATA ANALYSIS IN DIGITAL PSYCHOLOGICAL SUPPORT PLATFORMS

T.V. Kazak
Doctor of Psychology, Professor Head of the Department of Engineering Psychology and Ergonomics

A.A. Kovalchuk
Student, Department of Engineering Psychology and Ergonomics

A.N. Vasilkova
Senior Lecturer, Department of Engineering Psychology and Ergonomics

Abstract. The article presents a method for identifying users' psycho-emotional behavior based on the analysis of big data generated by digital psychological support platforms.

Keywords: big data, digital platforms, psychological support, clustering, patterns.

УДК 314.15:004.438

ПРОГНОЗИРОВАНИЕ ДЕМОГРАФИЧЕСКОЙ ПИРАМИДЫ РЕСПУБЛИКИ БЕЛАРУСЬ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ РАСШИРЕННОЙ АНАЛИТИКИ (ADVANCED ANALYTICS) В PYTHON







Н.И. Липницкая
Старший преподаватель кафедры экономической информатики БГУИР, n.i.karpovich@gmail.com

А.Ю. Герасимович
Студент факультета компьютерного проектирования БГУИР arsengerasm@gmail.com

Е.П. Ременчик
Студент инженерно-экономического факультета БГУИР katarемencik@gmail.com

Н.И. Липницкая
Старший преподаватель кафедры экономической информатики УО «БГУИР». Область научных интересов: прикладные системы обработки данных, информационные технологии и программирование, методика преподавания.

А.Ю. Герасимович
Студент 2 курса факультета компьютерного проектирования по специальности «Информационные системы и технологии в бизнес-менеджменте». Область научных интересов связана с анализом данных, разработкой программного обеспечения, математическим моделированием и бизнес-аналитикой.

Е.П. Ременчик
Студент 3 курса инженерно-экономического факультета по специальности «Информационные системы и технологии в финансово-банковской деятельности». Область научных интересов связана с анализом данных, математическим моделированием и обработкой больших данных.

Аннотация. В статье рассматривается проблема демографического старения населения Республики Беларусь. В качестве эмпирической базы использованы данные «Демографического ежегодника» и бюллетеней Белстата за 2019–2024 гг. Учтены современные вызовы: пандемия COVID-19, миграционные процессы и падение суммарного коэффициента рождаемости. Выполнен прогноз до 2050 года по трём сценариям: пессимистический, базовый и оптимистический. Для построения сценарных прогнозов применены методы расширенной аналитики (Advanced Analytics), включающие когортно-компонентное моделирование демографической пирамиды при помощи библиотек языка программирования Python. Результаты моделирования выявили критический рост демографической нагрузки, что требует адаптации пенсионной и миграционной политики, а также новых мер стимулирования рождаемости.

Ключевые слова: демографическое прогнозирование, Python, расширенная аналитика, цифровая экономика, половозрастная пирамида, коэффициент демографической нагрузки, Республика Беларусь, анализ данных.

Введение. Демографическая пирамида является важным инструментом для визуализации и анализа возрастно-половой структуры населения. Прогнозирование демографической структуры позволяет государственным органам, социальным службам и исследователям планировать меры социальной поддержки, здравоохранения, образования и экономического развития.

Для Республики Беларусь, находящейся на этапе постиндустриального перехода и сталкивающейся с глобальными вызовами урбанизации и старения населения, вопросы демографии приобрели экзистенциальный характер. Актуальность исследования обусловлена необходимостью учета «идеального шторма» факторов, которые игнорируются инерционными прогнозами:

- пандемия COVID-19: искажение структуры смертности и временное снижение ожидаемой продолжительности жизни;

- миграционный отток: высокая селективная эмиграция молодых, экономически активных граждан, что наносит двойной удар: экономика теряет кадры, а демография – потенциальных родителей;

- кризис рождаемости: падение числа рождений в 2024 году до уровня менее 59 000 является историческим антирекордом.

Цель работы – разработать программный модуль на Python для моделирования и визуализации трансформации демографической пирамиды Республики Беларусь до 2050 года.

Объект исследования – население Республики Беларусь.

Предмет исследования – динамика половозрастной структуры под влиянием рождаемости, смертности и миграции.

В контексте цифровой экономики задача прогнозирования демографической структуры решается с применением методов расширенной аналитики (Advanced Analytics) при помощи библиотек языка программирования Python.

В мировой практике широко применяются методы когортного моделирования, таблицы выживания, модели Ли–Картера для прогнозирования смертности, а также статистические модели для прогнозирования рождаемости [1],[2].

Методология и данные. В основе разработанного программного решения лежит метод передвижки возрастов (Cohort-Component Method). Когортно-компонентная модель реализована в виде программного модуля на Python, что позволяет в рамках расширенной аналитики (Advanced Analytics) выполнять сценарное прогнозирование и визуализацию демографической пирамиды.

Это детерминированный подход, который моделирует изменение численности населения по каждой возрастной группе (когорте) с шагом в один год.

Математическая модель. Процесс старения населения описывается переходом лиц возраста x в году t в возраст x+1 в году t+1. При этом учитывается смертность и миграция. Основное уравнение динамики численности населения имеет вид:

$$P^{s}_{(x+1,t+1)} = P^{s}_{(x,t)} \cdot \left(1 - q^{s}_{(x,t)}\right) + M^{s}_{(x,t)}$$

где:

- P – численность населения;

- s – пол (male/female);

- q – вероятность смерти (коэффициент смертности из таблиц дожития);

- M – сальдо миграции (разница между иммигрантами и эмигрантами).

Особым этапом является расчет численности новорожденных (возраст 0). Она зависит от численности женщин репродуктивного возраста (15–49 лет) и повозрастных коэффициентов рождаемости (ASFR). Общее число рождений B в году t+1 рассчитывается как:

$$B_{(t+1)} = \sum_{x=15}^{49} \left(P^{F}_{x,t} \cdot ASFR_{x}\right)$$

где:

- B_{t+1} – число новорожденных в году t+1;
- $P_{x,t}^{F}$ – численность женщин возраста x в году t;
- $ASFR_x$ – возрастно-специфичная рождаемость для возраста x.

Новорожденные распределяются по полу с учетом биологической константы (вторичное соотношение полов), которая для Беларуси составляет приблизительно 106 мальчиков на 100 девочек:

$$P_{0,t+1}^{male} = B_{t+1} \cdot \frac{106}{206},$$
$$P_{0,t+1}^{female} = B_{t+1} \cdot \frac{100}{206}$$

Для обеспечения точности прогноза были агрегированы данные из трех источников:

1. Официальные бюллетени Белстата «Численность населения на 1 января 2024 г. и среднегодовая численность населения за 2023 год» [3].
2. Данные о рождаемости и смертности за 2023 и 2024 годы, опубликованные в СМИ со ссылкой на Белстат и Минздрав [4].
3. «Демографический ежегодник Республики Беларусь 2019» использован для калибровки долгосрочных трендов смертности и брачности [5].

Реализация модели на Python. Программный модуль разработан с использованием библиотек Pandas для управления табличными данными, NumPy для математических операций и Matplotlib для визуализации.

Архитектура решения включает этапы, описанные ниже.

1. Инициализация данных (get_belarus_actual_2024): загрузка верифицированных данных о половозрастной структуре на 1 января 2024 года. Ввиду отсутствия в открытом доступе полных таблиц по однолетним возрастам за 2024 год в машиночитаемом виде, использовалась интерполяция данных на основе опорных точек (5-летних возрастных групп), опубликованных Белстатом, с последующей калибровкой под общую численность 9 155 978 человек.
2. Модуль прогнозирования (run_forecast): ядро программы, реализующее цикл по годам прогноза (с 2024 по 2050). В каждом шаге цикла:

- применяются коэффициенты дожития, рассчитанные на основе таблиц смертности (Human Mortality Database и данные Белстата), модель учитывает динамическое изменение смертности (постепенный рост ожидаемой продолжительности жизни);
- рассчитывается число родившихся на основе сценарного значения суммарного коэффициента рождаемости (TFR);
- производится «сдвиг» возрастов (aging process);
- в модель вводится сальдо миграции, распределенное по возрастной структуре (с акцентом на трудоспособные возраста, что характерно для экономической миграции).

3. Визуализация (plot_pyramid): построение классических демографических пирамид (Bar chart), где по оси Y отложен возраст, а по оси X – численность мужчин (слева) и женщин (справа). Это позволяет наглядно оценить эрозию основания пирамиды и расширение её вершины.

Анализ исходной демографической ситуации (2024 год). На 1 января 2024 года население Беларуси составило 9 155 978 человек. Это на 44 639 человек меньше, чем годом ранее (9 200 617 на начало 2023 года) [3]. Тенденция сокращения продолжилась и в 2024 году,

в результате чего к началу 2025 года население оценивается в 9 109 280 человек. Таким образом, наблюдается устойчивый тренд депопуляции со скоростью примерно 0,5% в год.

На рисунке 1 представлена половозрастная структура населения Беларуси по состоянию на 2024 год.

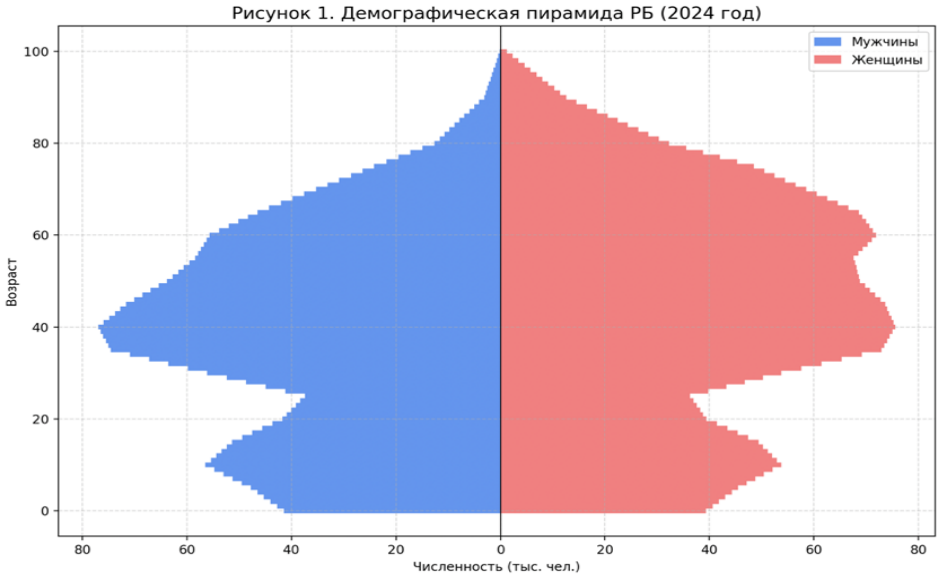


Рисунок 1. Демографическая пирамида РБ на 2024 год

Анализ пирамиды 2024 года выявляет глубокие диспропорции, являющиеся следствием исторических потрясений.

1. Сужение основания (0–9 лет): самая нижняя часть пирамиды демонстрирует резкое сужение. Это результат падения рождаемости, начавшегося после 2016 года. В 2024 году родилось всего 58 938 детей, что является историческим минимумом за послевоенный период. Для сравнения, еще в 2019 году родилось 87 602 ребенка. Падение числа рождений почти на 33% за пять лет создало глубокую выемку в основании пирамиды.
2. Малочисленная когорта (20–29 лет): в активный репродуктивный и трудовой возраст вступило поколение, рожденное в конце 1990-х и начале 2000-х годов. Это период глубокого демографического спада после распада СССР. Малочисленность этого поколения (так называемая «демографическая яма») объективно ограничивает количество потенциальных рождений, даже если меры стимулирования будут эффективны.
3. Многочисленное поколение (35–45 лет): «навес» в средней части пирамиды образован поколением бэби-бума 1980-х годов. Это наиболее экономически активная часть населения, которая сегодня несет основную налоговую нагрузку. Однако через 15–20 лет эта массивная когорта начнет выходить на пенсию, что создаст колоссальное давление на фонд социальной защиты.
4. Диспропорция полов в старших возрастах: начиная с возраста 65 лет, численность женщин значительно превышает численность мужчин. Ожидаемая продолжительность жизни при рождении в 2024 году составила 74,7 года. Однако разрыв между полами остается критическим: мужчины живут в среднем 69,3 года, женщины – 79,8 года. Эта разница более чем в 10 лет приводит к феминизации старости и специфической структуре домохозяйств пенсионеров (преобладание одиноких пожилых женщин).

Сценарии прогнозирования до 2050 года. Для моделирования будущей динамики были разработаны три сценария, отражающие различную степень адаптации демографической системы к текущим вызовам. Параметры сценариев отражены в таблице 1.

Таблица 1. Параметры сценариев прогнозирования

Сценарий	TFR (к 2050 г.)	Сальдо миграции (ежегодно, нетто)
1. Низкий (Пессимистический)	Снижение до 1.05	Отрицательное: -15000
2. Базовый (Инерционный)	Восстановление до 1.35	Отрицательное: -5000
3. Высокий (Оптимистический)	Рост до 1.7 к 2050 г.	Положительное: +10 000

Результаты моделирования показывают неутешительную динамику общей численности населения во всех сценариях, кроме гипотетически невозможных:

- **пессимистический сценарий**: население сокращается стремительно, к 2050 году численность может упасть до 7,2 – 7,45 млн человек (это потеря более 20% населения за четверть века);

- **базовый сценарий**: численность населения к 2050 году составит порядка 8,15 млн человек, что согласуется с рядом международных прогнозов, а ежегодная убыль будет составлять 30–40 тысяч человек.

- **оптимистический сценарий**: даже при росте рождаемости и положительной миграции инерция старения приведет к снижению численности до 8,6 – 8,75 млн человек, рост рождаемости скажется на численности рабочей силы только за горизонтом 2045 года.

Визуализация результатов для базового сценария (Рисунок 2) демонстрирует переход от формы «ели» (с выемками) к форме «урны» (узкое основание, широкий верх).

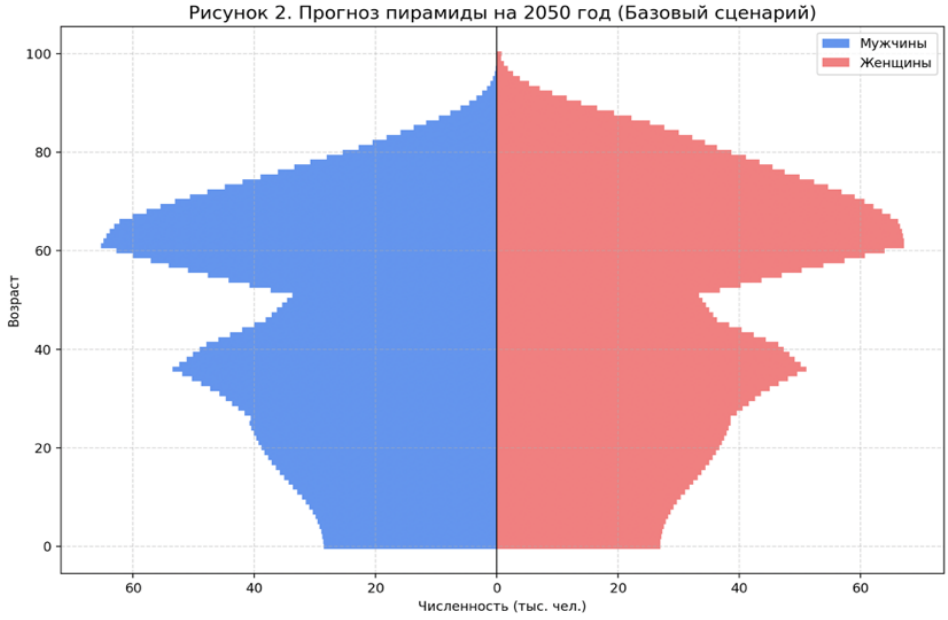


Рисунок 2. Прогноз половозрастной пирамиды на 2050 год (Базовый сценарий)

На основании анализа трансформации половозрастной пирамиды на 2050 год по базовому сценарию можно сделать следующие выводы.

1. Эффект двойного эха (основание пирамиды): малочисленное поколение рожденных в 2020–2024 годах (текущий провал) вступит в репродуктивный возраст в 2045–2050 годах, что приведет к «вторичному эху» – новой волне падения числа рождений, даже если коэффициент рождаемости на одну женщину будет стабилен.

2. Эрозия трудового ядра (середина пирамиды): когорта, которой сегодня 35–45 лет (самая многочисленная), к 2050 году перейдет в категорию 60–70+ лет. Их место в трудовой структуре (возраст 30–50 лет) займут малочисленные поколения рожденных в 1990-х и 2000-х. Это визуально отображается как резкое сужение «талии» пирамиды.

3. Инверсия верхушки (65+ лет): верхняя часть пирамиды значительно расширится. Если сегодня доля лиц 65+ составляет около 15–16%, то к 2050 году в базовом сценарии она приблизится к 25–28%. Это означает, что каждый четвертый житель страны будет пенсионером.

Критическим последствием является рост коэффициента демографической нагрузки.

В данной работе расчет производился по международному стандарту: отношение иждивенцев (0–14 лет и пенсионеры) к трудоспособному населению. Динамика показателя представлена на рисунке 3.

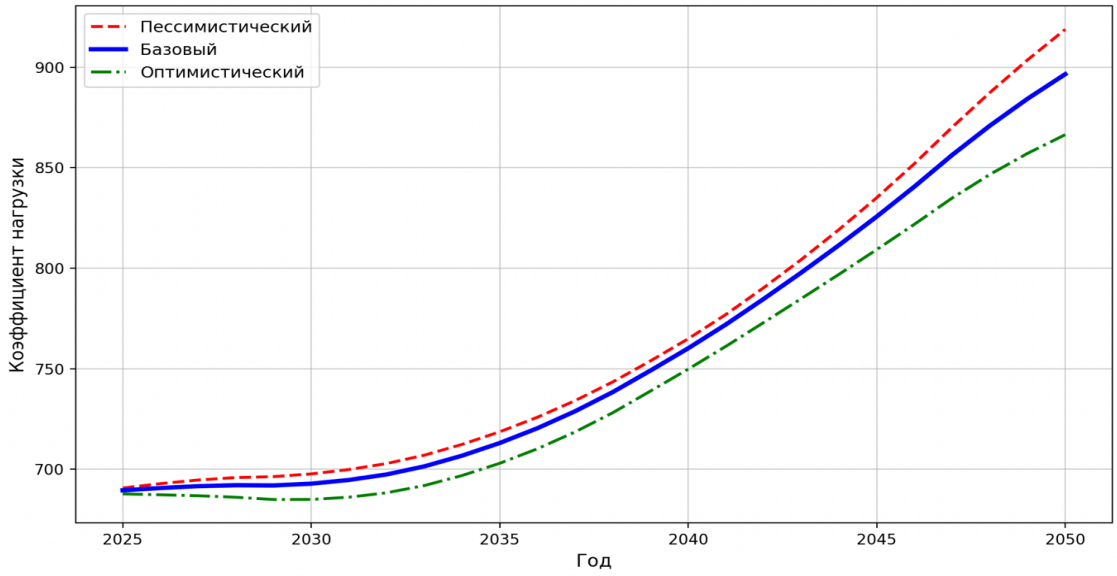


Рисунок 3. Прогноз демографической нагрузки

Как видно из графика (Рисунок 3), полученного в результате моделирования:

- в 2024 году коэффициент нагрузки составляет порядка 625-650 иждивенцев на 1000 трудоспособных;

- в базовом сценарии (синяя линия) нагрузка стремительно растет и к 2050 году достигает значения ~890, что означает существенное увеличение давления на работающее население;

- в пессимистическом сценарии (красная линия), предполагающем низкую рождаемость и отток молодежи, коэффициент превышает отметку 920, приближаясь к соотношению 1:1.

- только оптимистический сценарий (зеленая линия) позволяет сдержать рост нагрузки на уровне ~850-870, что все равно значительно выше текущих показателей.

Структурный анализ последствий прогноза. Сокращение населения в трудоспособном возрасте является наиболее критическим экономическим последствием. Согласно прогнозу, численность этой группы сократится более чем на 30–35% по сравнению с текущим уровнем.

Это означает физический дефицит рабочей силы. Экономика столкнется с ситуацией, когда невозможно закрыть вакансии не из-за отсутствия квалификации, а из-за отсутствия людей. Особенно уязвимы трудоемкие отрасли: сельское хозяйство, строительство, транспорт. Для IT-сектора, который полагается на приток молодых специалистов, сужение когорты 20–29 лет в 2030-х годах создаст серьезные риски для роста, требуя переориентации на удержание кадров и автоматизацию.

Белорусская пенсионная система, основанная на принципе солидарности поколений, при соотношении работающих и пенсионеров 1:1 станет финансово неустойчивой. Без реформ дефицит Фонда социальной защиты населения будет нарастать экспоненциально. Моделирование показывает, что простого повышения пенсионного возраста (который уже был повышен до 58/63 лет) будет недостаточно. Потребуется либо дальнейшее повышение пенсионного порога, либо снижение коэффициента замещения (размера пенсии относительно зарплаты), либо переход к смешанным накопительным системам.

Старение населения и рост доли лиц старше 75 лет приведет к изменению структуры спроса на медицинские услуги.

Вырастет потребность в лечении хронических заболеваний, гериатрии и долговременном уходе. «Феминизация» старости (преобладание женщин 80+) потребует развития системы социальной помощи одиноким пожилым людям, так как традиционная модель семейного ухода будет ослабевать из-за малодетности детей этих пенсионеров.

Стратегические рекомендации. На основе проведенного анализа предлагается комплекс мер для адаптации к демографическим изменениям.

Трансформация политики рождаемости: материальное стимулирование (семейный капитал) имеет ограниченный эффект. Необходим переход к созданию среды, дружественной семье: гибкие графики работы для родителей, развитие рынка арендного жилья, доступность качественных яслей.

Цель – не просто рождение первенцев, а стимулирование вторых и третьих рождений для повышения СКР хотя бы до 1,5–1,6.

Снижение мужской сверхсмертности: огромный резерв сохранения демографического потенциала кроется в снижении смертности мужчин трудоспособного возраста (40–60 лет). Меры по борьбе с сердечно-сосудистыми заболеваниями, алкоголизацией и травматизмом на производстве могут дать быстрый эффект, увеличив рабочую силу без ожидания взросления новых поколений.

Управляемая миграция: в условиях неизбежного сокращения внутреннего трудового ресурса необходимо разработать прозрачные механизмы привлечения квалифицированной рабочей силы из-за рубежа, а также программы репатриации и удержания молодежи (предотвращение «утечки мозгов»).

«Серебряная экономика»: восприятие старения как ресурса. Внедрение программ непрерывного образования и переквалификации для лиц 50+, стимулирование активного долголетия и создание рабочих мест, адаптированных для пожилых сотрудников.

Заключение. Проведенное исследование с использованием метода передвижки возрастов подтверждает, что Беларусь вступила в фазу глубокой демографической трансформации. Период «демографического дивиденда» завершен.

Прогноз до 2050 года показывает безальтернативность сценария сокращения и старения населения. В базовом сценарии численность жителей страны сократится до 8,1–8,2 млн человек, а медианный возраст существенно вырастет.

Однако демографическое будущее не является фатальным.

Реализация оптимистического сценария, предполагающего рост рождаемости и снижение смертности, способна смягчить темпы падения и стабилизировать возрастную структуру на новом уровне равновесия.

Ключевой вывод работы заключается в том, что адаптация социально-экономических институтов к реальности сужающегося и стареющего населения является не менее важной задачей, чем попытки переломить негативные тренды.

Результаты исследования могут быть использованы государственными органами и отраслевыми аналитиками для разработки долгосрочных стратегий в области человеческих ресурсов и социального планирования.

Список литературы

[1] Рыбаковский, Л. Л. Демографическое прогнозирование : учебное пособие / Л. Л. Рыбаковский. – Москва : ИНФРА-М, 2017. – 212 с.

[2] Preston, S. H. Demography: Measuring and Modeling Population Processes / S. H. Preston, P. Heuveline, M. Guillot. – Oxford : Blackwell Publishing, 2001. – 291 p.

[3] Численность населения на 1 января 2024 г. и среднегодовая численность населения за 2023 год : статистический бюллетень / Национальный статистический комитет Республики Беларусь. – Минск, 2024. – [Электронный ресурс]. – Режим доступа: https://www.belstat.gov.by. – Дата доступа: 30.11.2025.

[4] World Population Prospects 2024 [Электронный ресурс] : Online Edition / United Nations, Department of Economic and Social Affairs, Population Division. – New York, 2024. – Режим доступа: https://population.un.org/wpp/. – Дата доступа: 30.11.2025.

[5] Демографический ежегодник Республики Беларусь, 2019: статистический сборник / Национальный статистический комитет Республики Беларусь. – Минск, 2019. – 435 с.

[6] Human Mortality Database [Электронный ресурс] / University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). – Режим доступа: https://www.mortality.org. – Дата доступа: 30.11.2025.

Авторский вклад

Липницкая Наталья Ивановна – постановка задачи исследования, научное руководство исследованием, валидация алгоритмов прикладной обработки данных в модели прогнозирования, общая редакция статьи и формирование стратегических рекомендаций.

Герасимович Арсений Юрьевич – разработка программной реализации прогностической модели на языке Python (библиотеки Pandas, NumPy), автоматизация расчетов коэффициентов нагрузки и генерация аналитических графиков.

Ременчик Екатерина Павловна – формализация математической модели передвижки возрастов, подготовка эмпирических данных для моделирования, участие в разработке программной реализации на языке Python.

FORECASTING THE DEMOGRAPHIC PYRAMID OF THE REPUBLIC OF BELARUS USING PYTHON-BASED ADVANCED ANALYTICS METHODS

N.I. Lipnitskaya
Senior Lecturer of the Department of Economic Informatics, BSUIR
n.i.karpovich@gmail.com

A.Yu. Gerasimovich
Student of the Faculty of Computer Design, BSUIR
arsengerasm@gmail.com

E.P. Remenchik
Student of the Faculty of Engineering and Economics, BSUIR
kataremencik@gmail.com

Abstract. The article examines the problem of demographic ageing in the Republic of Belarus. As an empirical basis, it uses data from the «Demographic Yearbook» and Belstat bulletins for 2019–2024. The analysis takes into account current challenges, including the COVID-19 pandemic, migration processes and the decline in the total fertility rate. A forecast up to 2050 is produced under three scenarios: pessimistic, baseline and optimistic. T

o construct the scenario forecasts, methods of advanced analytics are applied, including cohort-component modelling of the demographic pyramid using Python libraries. The modelling results reveal a critical increase in the demographic dependency burden, which requires the adaptation of pension and migration policies as well as new measures to stimulate fertility.

Keywords: demographic forecasting, Python, advanced analytics, digital economy, age-sex pyramid, demographic dependency ratio, Republic of Belarus, data analysis

УДК 004.522

ПОСТРОЕНИЕ ВЫСОКОТОЧНОЙ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ТУРКМЕНСКОГО ЯЗЫКА НА ОСНОВЕ КОМБИНИРОВАННОГО ТРАНСФОРМЕРНОГО ПОДХОДА



М.Т. Мырадов
Заведующий кафедры «Информационные системы»
Институт Телекоммуникаций и информатики Туркменистана,
maksat.myradow.92@mail.ru

М.Т.Мырадов
Окончил Института телекоммуникаций и информатики Туркменистана. Область научных интересов распознавания речи, искусственный интеллект, защита данных

Аннотация: Современные системы распознавания речи на базе моделей Whisper Large-v3 и Llama-3 используют архитектуру Transformer и механизмы семантического прогнозирования для высокоточной интерпретации акустических сигналов в текстовую последовательность. Применение оптимизированных алгоритмов Faster-Whisper и архитектуры синтеза VITS позволяет эффективно обрабатывать специфические особенности туркменского языка, обеспечивая при этом высокую скорость вычислений и естественность звучания. Экспериментальное обучение модели на массиве из 4000 часов аудиоданных с использованием графических процессоров NVIDIA A100 подтвердило значительное снижение потерь и достижение высокого уровня точности распознавания.

Ключевые слова: Распознавание речи, языковые модели, архитектура Transformer, Whisper Large-v3, Llama-3, глубокое машинное обучение, семантический анализ, синтез речи, туркменский язык.

Введение: Современная научная интерпретация использования языковых моделей в распознавании речи основывается на концепциях семантического прогнозирования и мультимодального синтеза.

В настоящее время модели в системах STT (Speech-to-Text) эволюционировали в интеллектуальные процессоры, которые не просто идентифицируют слова, но и глубоко анализируют контекст высказывания. Основная задача таких моделей, как Whisper Large-v3 и Llama-3, заключается в интерпретации акустических сигналов посредством вероятностного анализа последовательностей слов. Благодаря архитектуре Transformer, нейронная сеть способна аппроксимировать логические характеристики предложения даже в условиях акустических помех или высокого уровня шума. Интеграция языковых моделей непосредственно в процесс декодирования звука позволяет в реальном времени корректировать грамматические и семантические ошибки.

Применение Whisper Large-v3 и Llama-3 имеет критическое значение для таких языков, как туркменский. Современное распознавание речи, сочетающее спектральный анализ и глубокое машинное обучение, обеспечивает высокую точность интерпретации данных [1].

Модель Whisper Large-v3 разработана для автоматического распознавания речи (ASR) и преобразования голоса в текст. Llama-3-8B, в свою очередь, является генеративной языковой моделью, работающей исключительно с текстовыми данными. Обе модели базируются на архитектуре Transformer, однако применяются к различным модальностям данных. Whisper

Large-v3 обрабатывает аудиосигналы, используя архитектуру Encoder-Decoder, тогда как Llama-3-8B представляет собой Decoder-only модель для работы с текстом [2]. Процесс обработки начинается с этапа сегментации речи, где входной сигнал преобразуется в логарифмическую мел-спектрограмму. Использование 128 спектральных фильтров в версии Large-v3 обеспечивает высокоточную передачу звуковых характеристик и частотных параметров речи. Энкодер состоит из двух высокоточных слоев с использованием сверточных фильтров и функции активации GELU (Gaussian Error Linear Unit). Эти слои сокращают временную размерность аудиосигнала вдвое. Затем данные передаются в 32 блока Трансформера. Каждый блок задействует механизм многоголового внимания (Multi-Head Attention), вычисляющий зависимости между всеми сегментами речевого тракта [3].

Декодер функционирует в авторегрессионном режиме, последовательно предсказывая текстовые токены. Он использует механизм перекрестного внимания (Cross-Attention) для фокусировки на релевантных признаках из энкодера на основе уже сгенерированного текста. Общее количество параметров модели составляет приблизительно 1,55 миллиарда [2].

Одной из ключевых особенностей является использование вращательных позиционных вложений (Rotary Positional Embeddings, RoPE). С математической точки зрения это реализуется путем вращения векторов запросов (query) и ключей (key) в комплексном пространстве с использованием тригонометрических функций. Данный механизм позволяет модели учитывать относительные расстояния между словами в больших контекстных окнах без потери качества обработки.Для нормализации данных на каждом слое применяется RMSNorm (Root Mean Square Layer Normalization).

В отличие от стандартной нормализации, данный метод не вычитает среднее значение, а ограничивается делением вектора на его среднеквадратичную норму. Это существенно ускоряет вычислительные процессы и оптимизирует эффективность распознавания речи [4]. В данной архитектуре используется механизм Grouped-Query Attention (GQA). Этот подход значительно сокращает объем памяти, необходимый для хранения кэша (KV-cache), и ускоряет генерацию ответов.

Для активации слоев применяется функция SwiGLU, которая объединяет свойства функций Swish и GLU (Gated Linear Unit), обеспечивая более стабильное и эффективное обучение нейронной сети. Основное различие между моделями заключается в математической формулировке задачи и методах обработки входных данных (таблица 1).

Таблица 1. Основные различия между моделями

Разделы	**Whisper Large-v3**	**Llama-3-8B**	**Тип данных**	**Тренировочные данные**
Основная задача	Распознавание речи	Генерация текста	Аудио – Размеченный текст	Supervised
Архитектура	Трансформер типа «Только Декодер»	Декодер Трансформер	Спектрограммы Токены	Прогнозирование следующего токена

В частности, Whisper Large-v3 на начальном этапе выполняет кратковременное преобразование Фурье (STFT) для перевода аудиосигнала x(t) во временно-частотное представление.

В результате полученная спектрограмма X(\omega, t) передается в энкодер **Whisper Large-v3**, который вычисляет скрытые представления, используя механизм самовнимания (**Self-Attention**):

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/d_k)V$$

Декодер Whisper применяет маскированное самовнимание для генерации текста и механизм перекрестного внимания для синхронизации текстовых токенов с акустическими признаками, полученными из энкодера.

Напротив, модель **Llama-3-8B** решает задачу вероятностного моделирования текстовых последовательностей. Рассмотрим последовательность токенов x_1, x_2, \dots, x_t. Модель прогнозирует вероятность появления следующего токена:

$$p(x_{t+1} \mid x_{\le t}) = \mathrm{softmax}(H^{L} W_{o})$$

Таким образом, Whisper оптимизирован для задачи выравнивания (alignment) звука и текста, тогда как Llama-3-8B ориентирована на авторегрессионную генерацию. Модель принимает на вход последовательность текстовых токенов и обучается предсказывать последующий элемент [5].

Подготовка и оптимизация системы STT MOTOR: Faster-Whisper представляет собой оптимизированную версию оригинальной модели OpenAI, которая функционирует значительно быстрее за счет квантования вычислений без существенной потери точности.

Данная модификация модели способна эффективно обрабатывать специфические данные, такие как медицинская терминология или редкие диалекты. Такая оптимизация позволяет системе более точно распознавать наречия, в которых стандартные модели часто допускают ошибки [6].

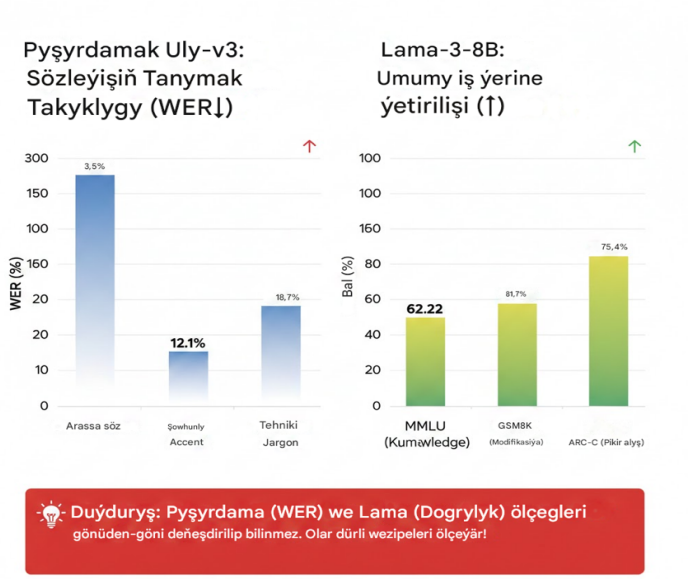


Рисунок 1. Whisper Large-v3 и Llama-3

Для оценки производительности данной модели мы используем метрику METEOR, которая обеспечивает более точный анализ по сравнению с простым сопоставлением слов.

Она распознает синонимы и учитывает корректность порядка слов, что делает оценку более объективной и приближенной к человеческому восприятию.

В отличие от стандартных методов, эта метрика признает, что замена слова синонимом не является грубой ошибкой.

Использование METEOR в сочетании с тонко настроенной моделью «Faster-Whisper» помогает разработчикам адаптировать систему для передачи глубокого смысла высказываний, а не просто набора звуков.

Логика Python: Логика программирования на языке Python представляет собой совокупность алгоритмических и структурных принципов, обеспечивающих корректную обработку данных и управление программными потоками.

Python использует высокоуровневые структуры, такие как условные операторы и функции, которые улучшают читаемость кода и формальную верификацию.

Благодаря динамической типизации и обширным стандартным библиотекам, язык эффективно применяется в научных вычислениях и системах искусственного интеллекта.

Логическая модель Python способствует быстрому прототипированию и минимизирует вероятность логических ошибок, что делает его оптимальным инструментом для построения сложных программных комплексов.

TTS MOTOR: VITS (End-to-End): VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) – это современная архитектура синтеза речи. Модель объединяет вариационный автокодировщик и генеративно-состязательное обучение для генерации речевого сигнала.

В отличие от каскадных систем TTS, VITS обучается напрямую без промежуточных акустических представлений, что снижает накопление ошибок.

Это делает VITS перспективным решением для масштабируемых систем синтеза речи.

Speaker: Компонент Speaker в системах TTS (Text-to-Speech) отвечает за моделирование индивидуальных характеристик голоса диктора.

Он реализуется в виде эмбеддингов (embeddings), которые отображают тембр, интонацию и особенности артикуляции речи.

Использование идентификаторов диктора позволяет осуществлять многоголосый синтез в рамках одной модели.

В научных исследованиях данный подход применяется для клонирования голоса и персонализации речевых интерфейсов.

Правильное моделирование компонента Speaker значительно повышает естественность синтезированной речи.

Заключение. Для обучения и запуска новой туркменской языковой модели мы использовали высокопроизводительные графические процессоры, в частности GPU: 2x NVIDIA A100 (80 ГБ VRAM).

```
[15] def load_audio(audio_path):
0s       # Load the audio file
         waveform, sample_rate = torchaudio.load(audio_path)

         # If the sample rate of the audio file is different, resample it
         if sample_rate != 16000:  # Whisper models expect 16kHz audio
             waveform = torchaudio.transforms.Resample(orig_freq=sample_rate, new_freq=16000)(waveform)

         return waveform.squeeze(0), 16000  # Return waveform and target sample rate
```

Обучение новой модели Whisper проводилось на основе 4000 часов аудиозаписей на туркменском языке в течение 14 дней (показатель Training Loss был снижен до 0.04).

Список литературы

[1] Надежное распознавание речи с помощью крупномасштабного слабого контроля Алек Рэдфорд, Чон Вук Ким, Тао Сюй, Грег Брокман, Кристин Макливи, Илья Суцкевер 2019

[2] Эффективное многоголовочное декодирование для распознавания речи на основе трансформеров Яэль Сегал-Фельдман, Авив Шамсян, Авив Навон, Гилл Хетц, Джозеф Кешет 2022

[3] Соединение кодировщика речи и большой языковой модели для автоматического распознавания речи. Вэньи Ю,Чанли Тан,Гуанчжи Сунь, Сяньчжао Чен, Тянь Тан, Вэй Ли, ,Цзэцзюнь Ма,Чао Чжан 2020

[4] B.H. Juang, and S. Furui. Automatic Recogni-tion and Understanding of Spoken Language–A First Step Toward Natural Human Machine Communication. Proc. IEEE, 88, No. 8, 2016, pp. 1142-1165.

[5] Hiroaki Sakoe and Seibi Chiba. Dynamic Pro-gramming Algorithms Optimization for Spoken Word Recoginition. IEEE Trans on acoustics, speech and sig-nal processing, vol.ASSP-26, no.1, december 2017.

[6] L.R.Rabiner and B.H.Juang. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 2016.

Авторский вклад

Мырадов Максат Тачмухаммедович – авторский вклад в архитектуру Transformer для преобразования голоса в текст, разработав, например, новый метод адаптации механизма внимания, оптимизирующий процесс обучения или интегрировав инновационные слои для улучшения обработки аудиоданных и повышения точности распознавания речи.

DEVELOPMENT OF A HIGH-PRECISION INTELLIGENT NATURAL LANGUAGE PROCESSING SYSTEM FOR THE TURKMEN LANGUAGE BASED ON A COMBINED TRANSFORMER APPROACH

M.T. Myradov

Head of the Department of Information Systems, The Institute of Telecommunications and Informatics

Abstract: Modern speech recognition systems based on Whisper Large-v3 and Llama-3 models utilize the Transformer architecture and semantic prediction mechanisms for high-precision interpretation of acoustic signals into text sequences. The application of optimized Faster-Whisper algorithms and the VITS synthesis architecture enables efficient processing of the specific features of the Turkmen language, while ensuring high computational speed and natural sound quality. Experimental training of the model on a dataset of 4000 hours of audio data using NVIDIA A100 GPUs confirmed a significant reduction in loss and the achievement of a high level of recognition accuracy.

Keywords: Speech recognition, language models, Transformer architecture, Whisper Large-v3, Llama-3, deep machine learning, semantic analysis, speech synthesis, Turkmen language.

УДК 004.8

ВЛИЯНИЕ МЕТОДОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ НА СТРАТЕГИЮ ПОВЕДЕНИЯ АВТОНОМНЫХ АГЕНТОВ



А.А. Гулис

Студент факультета компьютерных систем и сетей БГУИР

antongulis43@gmail.com



М.А. Калугина

Доцент кафедры информатики, кандидат физико-математических наук, доцент

marina_kalugina@list.ru

А.А. Гулис

В 2022 году окончил Лицей города Ивацевичи. Интересы направлены на глубокое обучение, обучение с подкреплением в сложных средах, компьютерное зрение, а также генеративные нейросетевые модели для синтеза 2D- и 3D-контента.

М.А. Калугина

Окончила Белорусский государственный университет. Область научных интересов связана с исследованием проблем метрической теории диофантовых приближений зависимых величин и приложений математических методов к нейросетевому анализу

Аннотация. В статье приведены результаты исследования влияния различных архитектур и алгоритмов глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) на формирование стратегий поведения автономных агентов. В качестве испытательной среды используется актуальный клиент многопользовательской игры Tanks Blitz. Рассматривается мультиагентное кооперативно-соревновательное взаимодействие в формате

«2 на 2» с использованием парадигмы обучения агента против собственных исторических версий (Self-Play). Обучение в данной среде осложнено частичной наблюдаемостью и ограничениями пропускной способности. Предложена распределенная архитектура сбора данных и проанализированы алгоритмы PPO, Discrete SAC (DSAC), Stable DSAC (SDSAC), R2D2 и гибридная модель Hybrid SAC. Выявлены феномены «взлома вознаграждения» (reward hacking), взрыва температурного коэффициента, коллапса энтропии, а также предложены конкретные методы их устранения.

Ключевые слова: глубокое обучение с подкреплением, self-play, PPO, SAC, DSAC, SDSAC, Hybrid SAC, формирование вознаграждения.

Введение. Методы глубокого обучения с подкреплением (RL) зарекомендовали себя как эффективный инструмент для создания автономных агентов – автономных систем принятия решений –, способных действовать в сложных динамических средах [1]. В отличие от задач классификации, где нейронная сеть обучается на заранее размеченном наборе данных, в RL агент самостоятельно генерирует данные путем проб и ошибок, получая от среды скалярный сигнал – вознаграждение.

Целью данного исследования является сравнительный анализ того, как выбор математического аппарата алгоритма (On-policy, Off-policy, рекуррентные и гибридные сети) влияет на итоговую стратегию поведения агента, управляющего боевой машиной в формате «2 на 2».

Ключевой особенностью работы является отказ от использования упрощенных симуляторов: взаимодействие происходит напрямую с актуальным игровым клиентом Tanks Blitz через специально разработанный WebSocket-интерфейс. Это накладывает ряд критических ограничений:

1 Низкая пропускная способность. Взаимодействие с 3D-клиентом ресурсоемко. На высокопроизводительной системе (CPU 24 ядра, GPU RTX 4080 Super, 64 GB RAM) параллельно запускается около 30 экземпляров игры (в зависимости от выбранного алгоритма), что в сумме дает генерацию лишь около 300 кадров (шагов) в секунду.

2 Разреженная награда. Терминальным победным состоянием является полное уничтожение команды противника. Вероятность случайного выполнения этой задачи стремится к нулю, поэтому на ранних этапах обучения нейронная сеть не получает значимого обучающего сигнала.

3 Частичная наблюдаемость (POMDP). Агенту доступна лишь локальная информация. В игре реализованы механики циклов перезарядки и «тумана войны» (противник не передается клиенту, если находится вне радиуса радиосвязи или за укрытием), что нарушает марковское свойство среды, требуя от сети наличия памяти.

Описание среды и распределенной архитектуры. Агенты внутри одной команды обучаются с использованием парадигмы разделения параметров: оба танка управляются одной и той же нейронной сетью, однако получают индивидуальные векторы наблюдений и формируют независимые действия [2].

На начальных этапах исследования в качестве входных данных использовались пиксельные изображения (кадры из игры), обрабатываемые сверточными сетями (CNN). Однако для повышения скорости сходимости был осуществлен переход к векторным представлениям. Для обеспечения свойства трансляционной инвариантности (независимости поведения от абсолютных координат на карте) вектор состояния строится исключительно на относительной геометрии и статусах модулей.

Формируемый вектор включает:

1 Сложную геометрию целей: расчет дистанций, нормализованных векторов направлений (орудия, камеры, корпуса) и их косинусных расстояний. Дополнительно с помощью матриц поворота вычисляются проекции направлений агента на границы корпуса и башни противника.

2 Статус модулей и орудия: здоровье внутренних модулей агента и противников (гусеницы, двигатель, боеукладка, механизм поворота башни), вероятность пробития брони, текущий разброс орудия и статус прицеливания (в танк, в землю или в небо).

3 Сенсоры препятствий: массив виртуальных лучей, сканирующих пространство вокруг танка на предмет укрытий и препятствий. Лучи запускаются на ограниченную дистанцию – 50 метров. Всего испускается 8 лучей: вперед, назад, влево, вправо и в промежуточные направления.

4 Историю действий: к вектору состояния конкатенируется *one-hot* кодировка предыдущего дискретного действия, а при использовании гибридных политик – еще и маскированные смещения.

Для обеспечения корректной работы с актуальным клиентом игры необходимо исключить отправку физически невозможных команд (например, выстрел в момент перезарядки). Среда возвращает бинарную маску допустимых действий. На стороне алгоритма реализовано маскирование: логиты (сырые выходы нейросети) недопустимых действий заменяются на минимально возможное значение перед вычислением функции вероятности Softmax. Это математически корректно обнуляет вероятность выбора запрещенного действия, не искажая расчет энтропии в функции потерь.

Для непрерывного обучения спроектирована асинхронная архитектура Actor-Learner [3]. Множество процессов-сборщиков на CPU собирают переходы агентов в среде и отправляют их в очередь Redis. Центральный процесс на GPU непрерывно извлекает данные, вычисляет градиенты и обновляет веса (рисунок 1).

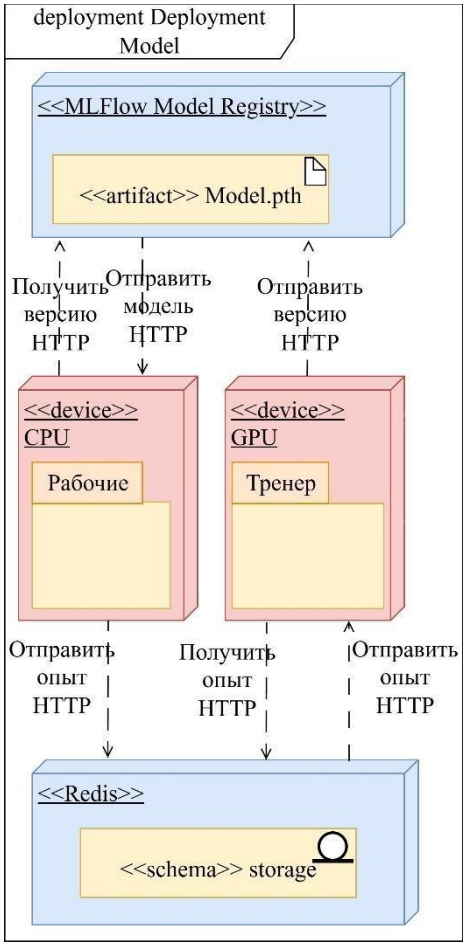


Рисунок 1. Диаграмма распределенной архитектуры

Для предотвращения эффекта «катастрофического забывания», когда агент выучивает узкую тактику против текущего оппонента, применяется смешанная выборка. Вражеская команда формируется по правилу: в 80% случаев загружаются самая последняя версия нейросети, а в 20% случаев случайным образом выбираются предыдущие версии.

Формирование стратегии через систему вознаграждений. Для обеспечения стабильного самообучения функция вознаграждения спроектирована по принципу игры с нулевой суммой (Zero-sum reward): любое положительное вознаграждение агента зеркально отражается штрафом для противника. Это предотвращает неформальную кооперацию агентов-оппонентов, когда агент может найти лазейку, увеличивая награды в одном состоянии, оставляя другого агента с нулевой или несопоставимой наградой.

Поскольку базовая награда крайне разрежена, был применен метод формирования вознаграждения на основе потенциалов (Potential-Based Reward Shaping, PBRS). Метод теоретически гарантирует, что добавление плотного сигнала не изменяет глобальную оптимальную политику [4].

Однако первоначальные эксперименты выявили классическую проблему «взлома вознаграждения» [5]. Если агент поощрялся исключительно за нанесенный урон, он находил локальный оптимум: поднимал камеру в небо, либо опускал орудие и стрелял в землю, так как противник зачастую находился рядом.

Для решения проблемы награда за урон была масштабирована в зависимости от дистанции до цели. Чем выше дистанция между агентом и танком противника в момент попадания, тем выше награда. Для лучшего усвоения новой системы наград было разработано обучение по расписанию (Curriculum Learning) [6]. Суть расписания заключалась в увеличении дистанции между командами по мере продвижения обучения. В начале, когда расстояние между агентами близкое, они изучали то, как ведет себя среда, какую они награду получают за свои действия, а затем, постепенно отдаляясь, начинали вырабатывать стратегии нападения и защиты.

Влияние алгоритмов на поведение. В ходе исследования был протестирован ряд современных DRL алгоритмов. Оценивалось их поведение в условиях низкой пропускной способности среды.

Первым был реализован On-policy алгоритм PPO [7]. Первоначальные эксперименты с алгоритмом проводились в упрощенной изолированной постановке: бои проходили в формате «1 на 1», при этом агенты были лишены возможности передвигаться. Танки появлялись в случайных позициях и на различных дистанциях друг относительно друга, управляя исключительно наведением орудия и стрельбой. В этой ограниченной среде PPO успешно и достаточно быстро решил задачу, выработав оптимальную политику точного прицеливания и быстрого уничтожения противника.

Успех в базовом статичном сценарии послужил мотивацией для перехода к полноценной и значительно более сложной среде – боям «2 на 2» с возможностью полномасштабного маневрирования. Для работы в этой среде исследовались две архитектуры сети: полносвязная и рекуррентная (LSTM). Эксперименты показали, что LSTM-архитектура справляется с задачей значительно лучше благодаря способности агрегировать исторический контекст (например, отслеживать перемещения противника за укрытиями).

Однако при масштабировании задачи алгоритм столкнулся со сложностями: агенты не могли выучить комплексное поведение на дальних дистанциях. Для решения этой проблемы была успешно внедрена стратегия Curriculum Learning. Изначальная дистанция между командами при появлении искусственно занижалась, а затем постепенно увеличивалась по мере выпуска новых, более обученных версий агента. Этот подход позволил PPO успешно и уверенно освоить среду 2v2, включая бои на дальних дистанциях. Впоследствии стратегия постепенного увеличения дистанции доказала свою эффективность и была интегрирована в обучение всех последующих алгоритмов.

Поскольку алгоритм обновляет веса только на свежесобранных данных, к его архитектуре были применены стабилизирующие практики [8]. Были исключены методы регуляризации, характерные для задач компьютерного зрения (такие как Batch Normalization или Dropout). Связано это с тем, что регуляризация должна предотвратить переобучение модели, когда она не может обобщить свои знания на данные, которых она не видела. Данная проблема не появляется в методах DRL. Также использовалась ортогональная инициализация весов, нормализация входных наблюдений и целевых значений с помощью бегущего среднего.

Критическим дополнением стала ранняя остановка обучения на эпохе: если KL-дивергенция (мера расхождения между старой и новой политикой) превышала порог 0.045, обновление прерывалось.

Благодаря Curriculum Learning алгоритм PPO успешно решил задачу 2v2 и оказался наиболее стабильным из всех исследованных методов. Однако выявились два критических недостатка:

1 Экстремальная чувствительность к подбору гиперпараметров.

2 Крайне низкая сэмпл-эффективность. Из-за On-policy природы алгоритма и низкой пропускной способности среды время сбора новых данных делало процесс обучения неприемлемо долгим.

Именно невозможность переиспользования старого опыта при жестких лимитах на генерацию новых кадров стала главной причиной отказа от масштабирования PPO и перехода к исследованию Off-policy методов.

Переход к алгоритму DSAC [9] потребовал реструктуризации пространства действий. В алгоритме PPO агенты использовали многомерное дискретное пространство, состоящее из пяти независимых ветвей: стрельба (2 варианта), поворот камеры по оси X (5 вариантов), по оси Y (3 варианта), движение шасси (3 варианта) и поворот корпуса (3 варианта). Архитектура PPO естественным образом поддерживает такой формат через несколько независимых выходных слоев (голов) нейросети. Однако классические дискретные Q-функции, на которых базируется DSAC, требуют оценки ценности для каждого возможного действия. Чтобы избежать создания сложной ветвящейся архитектуры Q-сети (критика), многомерное пространство было преобразовано в единое одномерное дискретное пространство, содержащее 270 уникальных комбинаций действий.

На этапе тестирования базового DSAC алгоритм продемонстрировал математическую нестабильность, вызванную спецификой функции вознаграждений:

1 Изначально была предпринята попытка использовать алгоритм нормализации целевых значений PopArt [10], который часто применяется в мультиагентных средах. Однако из-за высокой разреженности наград (подавляющее большинство шагов агента приносят нулевое вознаграждение) дисперсия в нормализаторе стремилась к нулю. Деление на околонулевую дисперсию вызывало градиентные взрывы и деградацию весов. От PopArt пришлось отказаться в пользу статического масштабирования (деления всех наград на константу).

2 В алгоритмах семейства SAC используется параметр, который динамически регулирует баланс между исследованием (максимизацией энтропии) и эксплуатацией (максимизацией награды) [11]. Механизм автоматической настройки вычисляет ошибку на основе разницы между текущей и целевой энтропией. В условиях нашей среды агенты получали нулевые награды на протяжении тысяч шагов. Следовательно, Q-значения оставались околонулевыми и не создавали градиентного «противовеса» для энтропии. В попытке заставить сеть исследовать среду (достичь целевой энтропии), оптимизатор бесконтрольно увеличивал значение регулятора до экстремальных величин (рисунок 2). Это приводило к тому, что политика становилась абсолютно случайной и теряла способность к какому-либо осмысленному поведению.

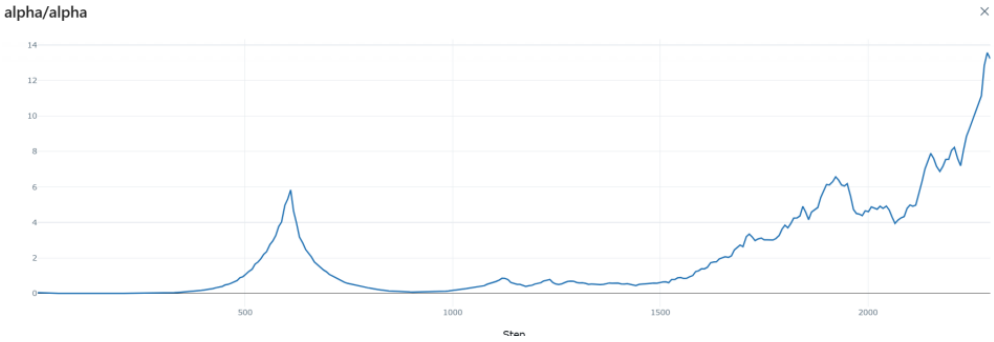


Рисунок 2. График параметра регуляции энтропии в алгоритме DSAC

Алгоритм удалось временно стабилизировать лишь путем экстремального снижения скорости обучения, однако скорость обучения оказалась неприемлемо низкой, а агент так и не смог выучить полезные тактики. Для решения проблем базового DSAC алгоритм был модернизирован до версии Stable DSAC (SDSAC) [12]. Основным архитектурным изменением стало вычисление целевого Q-значения. В классическом SAC используется минимум из двух критиков для предотвращения переоценки (overestimation). В дискретных пространствах это часто ведет к излишнему пессимизму. В SDSAC целевое значение вычислялось как среднее между критиками, что существенно стабилизировало процесс оценки состояний. Особое внимание было уделено горизонту предвидения – использованию n-step returns. Вместо обновления Q-функции на основе только одного следующего шага, алгоритм аккумулировал дисконтированную награду за несколько шагов вперед. Эксперименты проводились с параметрами n=3 и n=5.

Было установлено, что переход к n=5 обеспечивает значительно более плавное падение энтропии политики по мере обучения. В контексте обучения с подкреплением энтропия отражает степень неопределенности агента. Плавное и медленное падение энтропии означает, что агент дольше сохраняет высокий уровень исследования и медленнее переходит к эксплуатации. С одной стороны, это предотвращает преждевременную сходимость к субоптимальным тактикам. С другой стороны, это свидетельствует о значительном замедлении процесса обучения: агенту требуется гораздо больше времени и вычислительных мощностей, чтобы обрести уверенность в правильности своих действий и начать применять жесткие детерминированные стратегии поведения.

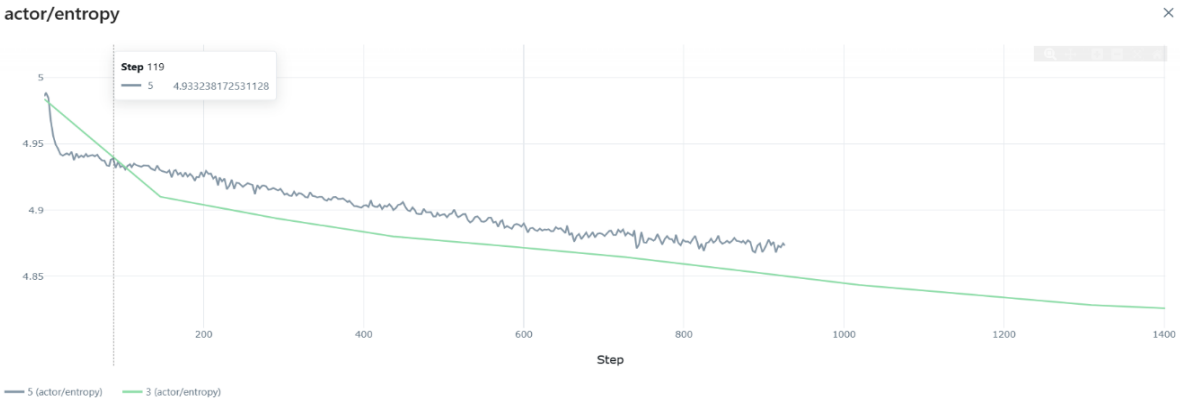


Рисунок 3. Сравнение динамики энтропии для параметра n-step

Благодаря этим модификациям SDSAC позволил агентам выработать базовые тактические паттерны, недоступные ранее. В частности, агенты освоили правильное

позиционирование – научились разворачивать лобовую (наиболее защищенную) проекцию брони к противнику перед выстрелом. Тем не менее, общая медлительность обучения (из-за затянутой фазы падения энтропии) послужила триггером для поиска еще более продвинутых архитектур. Для решения проблемы частичной наблюдаемости был внедрен алгоритм R2D2 [13], основанный на рекуррентной архитектуре. Для обучения сети исторические данные разбивались на фазу «прогрева» для инициализации скрытых состояний и фазу обучения. Наличие памяти (LSTM) позволило агенту запоминать положение противника. Агенты успешно научились набирать и удерживать дистанцию, максимизируя функцию вознаграждения. Недостатком стало появление артефактов поведения: агент начинал циклично повторять микро-движения, визуально «замедляясь» по сравнению с безрекуррентными моделями.

Наиболее перспективным направлением стала реализация алгоритма Hybrid SAC [14]. Суть подхода заключается в отказе от огромного дискретного пространства:

1 Управление шасси и стрельбой остается дискретным.

2 Управление башней переводится в непрерывное пространство, за которое отвечает отдельный выход сети.

Это радикально снизило размерность задачи, однако потребовало функции потерь сети стратегии (актора) с двумя независимыми регуляторами энтропии. Архитектура нейросети была перепроектирована и теперь включает два независимых выходных слоя: первый отвечает за выбор тактических маневров уклонения, а второй генерирует точные параметры для плавного сопровождения цели орудием. Для обеспечения надежности обучения дискретной части агента в гибридную модель были успешно интегрированы механизмы из ранее протестированного алгоритма SDSAC.

В процессе настройки была выявлена крайняя чувствительность алгоритма к гиперпараметрам. Во-первых, обучение критически зависит от выбора целевой энтропии: некорректное значение приводит к необратимой деградации температурного коэффициента, который сходится к нулю или уходит в максимум, полностью останавливая прогресс агента. Во-вторых, эффективность гибридного подхода сильно зависит от коэффициентов масштабирования камеры. При слишком малых значениях агент становится медлительным и почти неподвижным, поскольку любой маневр начинает требовать избыточного количества микро-действий. При завышенных масштабах предсказания сети перестают нести полезную нагрузку, так как они упираются в жесткие физические ограничения среды на скорость вращения башни. Тонкая калибровка этих параметров позволила раскрыть потенциал архитектуры. В отличие от полностью дискретных алгоритмов, где наведение происходило ступенчато и часто сбивалось при движении шасси, гибридная модель продемонстрировала максимально плавное и точное наведение, наиболее приближенное к действиям реального игрока.

Заключение. В рамках данного исследования была разработана и успешно внедрена сложная распределенная система обучения с подкреплением, позволяющая обучать автономных агентов непосредственно в актуальном клиенте игры Tanks Blitz. Проведенный сравнительный анализ алгоритмов и архитектур DRL свидетельствует о том, что математический аппарат метода обучения оказывает прямое, определяющее влияние на формируемую тактику поведения агента в условиях ограниченной пропускной способности и разреженных наград:

1 Алгоритм PPO продемонстрировал высочайшую стабильность обучения и, в связке со стратегией Curriculum Learning, позволил выработать тактику точного прицеливания и дальнего боя. Однако его On-policy природа привела к критической нехватке данных, сделав процесс масштабирования нецелесообразным по времени.

2 Переход к Off-policy методам выявил уязвимость базовых алгоритмов семейства SAC к нулевым наградам (взрыв энтропийного коэффициента и крах PopArt нормализации). Лишь архитектурные модификации в SDSAC (усреднение Q-значений, отказ от PopArt и

использование n-step) позволили стабилизировать сеть, благодаря чему агенты выучили тактику защитного позиционирования.

3 Алгоритм R2D2 подтвердил критическую важность рекуррентной памяти для сред с частичной наблюдаемостью, позволив агентам использовать укрытия и контролировать дистанцию, хотя и ценой цикличных микро-движений.

4 Гибридная архитектура Hybrid SAC, усиленная механизмами стабилизации SDSAC, решила проблему размерности пространства действий и устранила недостаток ступенчатого прицеливания, характерный для полностью дискретных алгоритмов.

Таким образом, для создания высокоуровневых автономных агентов в комплексных 3D-симуляторах недостаточно применения базовых RL-алгоритмов. Требуется их глубокая адаптация под физику среды. В дальнейшей работе планируется внедрение алгоритмов класса CTDE (например, QMIX) для формирования более сложных стратегий командного взаимодействия.

Список литературы

[1] Sutton R. S., Barto A.G. Reinforcement Learning: An Introduction. 2nd ed. Cambridge, MA: The MIT press; 2018.

[2] Terry J.K., Grammel N., Son S., et al. Revisiting Parameter Sharing in Multi-Agent Deep Reinforcement Learning. arXiv. 2020. DOI: 10.48550/arXiv.2005.13625.

[3] Espeholt L., Soyer H., Munos R., et al. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. Proceedings of the 35th International Conference on Machine Learning (PMLR). 2018;80:1407-1416.

[4] Ng A.Y., Harada D., Russell S.J. Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping. Proceedings of the Sixteenth International Conference on Machine Learning (ICML). 1999:278-287.

[5] Amodei D., Olah C., Steinhardt J., et al. Concrete Problems in AI Safety. arXiv. 2016. DOI: 10.48550/arXiv.1606.06565.

[6] Narvekar S., Peng B., Leonetti M., et al. Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey. Journal of Machine Learning Research. 2020;21(181):1-50.

[7] Schulman J., Wolski F., Dhariwal P., et al. Proximal Policy Optimization Algorithms. arXiv. 2017. DOI: 10.48550/arXiv.1707.06347.

[8] Andrychowicz M., Raichuk A., Stańczyk P., et al. What Matters for On-Policy Deep Actor-Critic Methods? A Large-Scale Study. International Conference on Learning Representations (ICLR). 2021. DOI: 10.48550/arXiv.2006.05990.

[9] Christodoulou P. Soft Actor-Critic for Discrete Action Settings. arXiv. 2019. DOI: 10.48550/arXiv.1910.07207.

[10] Hessel M., Soyer H., Espeholt L., et al. Multi-Task Deep Reinforcement Learning with PopArt. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33(01):3796-3803. DOI: 10.1609/AAAI.V33I01.33013796.

[11] Haarnoja T., Zhou A., Hartikainen K., et al. Soft Actor-Critic Algorithms and Applications. arXiv. 2018. DOI: 10.48550/arXiv.1812.05905.

[12] Zhou H., Wei T., Lin Z., et al. Revisiting Discrete Soft Actor-Critic. arXiv. 2022. DOI: 10.48550/arXiv.2209.10081.

[13] Kapturowski S., Ostrovski G., Quan J., et al. Recurrent Experience Replay in Distributed Reinforcement Learning. International Conference on Learning Representations (ICLR) (Poster). 2019.

[14] Chen C., Deng X. A Hybrid SAC Algorithm: Advantage-Guided Soft Actor-Critic. MSCE ’25: Proceedings of the 2025 International Conference on Management Science and Computer Engineering. 2025:459-462. DOI: 10.1145/3760023.3760100.

Авторский вклад

Гулис Антон Александрович – выбор задачи исследования, разработка программной среды взаимодействия с игровым клиентом, реализация архитектуры распределенной системы и нейросетевых моделей, проведение экспериментов, анализ полученных результатов.

Калугина Марина Алексеевна – постановка проблемы, научное руководство исследованием методов глубокого обучения с подкреплением для управления автономными агентами, консультации по математическому аппарату алгоритмов.

INVESTIGATION OF THE IMPACT OF REINFORCEMENT LEARNING METHODS ON AUTONOMOUS AGENT BEHAVIOR STRATEGIES

A.A. Gulis
Student of BSUIR

M.A. Kalugina
Associate Professor of Informatics Department of the BSU

Abstract. The article presents the results of a study on how various deep reinforcement learning (DRL) architectures and algorithms affect the development of behavioral strategies in autonomous agents. The current client of the multiplayer game Tanks Blitz is used as the test environment. We consider multi-agent cooperative–competitive interaction in a «2 vs 2» setting using the self-play paradigm, where an agent is trained against its own historical versions. Training in this environment is complicated by partial observability and bandwidth limitations. A distributed data-collection architecture is proposed, and the PPO, Discrete SAC (DSAC), Stable DSAC (SDSAC), R2D2, and a hybrid Hybrid SAC model are analyzed. We identify the phenomena of reward hacking, temperature coefficient explosion, and entropy collapse, and propose specific methods to mitigate them.

Keywords: deep reinforcement learning, self-play, PPO, SAC, DSAC, SDSAC, Hybrid SAC, reward shaping.

УДК 621.391:004.75

CLASSIC AND MODERN ENCRYPTION ALGORITHMS FOR BIG DATA PROCESSING





Y. Liashchevich
Postgraduate of the BSUIR

Cai Yuxin
Graduate of Luoyang Normal University

Y. Liashchevich
Graduated from BSUIR. Postgraduate in the Department of Electronic Engineering and Technology. His research interests include the effects of infrasound on biological tissue, infocommunication signals and systems.

Cai Yuxin
Graduate of Luoyang Normal University. The area of scientific interest is research into encryption and decryption technologies for communication signals based on new algorithms.

Abstract. The article discusses the using methods of research into encryption and decryption technologies for communication signals based on new algorithms in big data systems

Keywords: encryption, decryption, channel coding, burst interference.

Introduction. The modern world is living in an era of rapid development in wireless technologies and big data processing. Wireless systems such as 5G and 6G IoT, satellite communications transmit massive amounts of sensitive data, must ensure data transmission security [1]. As data transfer speeds increase, so does the volume of data. Routing and data storage devices must ensure maximum data protection from unauthorized access. Modern data encryption algorithms, are used for this purpose. Research Questions it is how effective are new encryption algorithms for

communication signals? And can these algorithms resist emerging threats? What trade-offs exist between speed, security, and resource usage?

Basic encryption algorithms. Encryption is a technique for securing the information so only the authorized parties can interpret the data. What happens is that encryption provides transformations on the plaintext and transforms it into ciphertext which isn't human-readable. Not only in the digital era but ciphers have also been widely used throughout history. In the figure below, we present the general view of the classification of the ciphers. As we've seen, we categorize ciphers in terms of some characteristics that they have or according to their usage. They are mainly divided into two different types; classical and modern classes. The most common and used in digital are is modern class. It's because of its dynamic and static cryptographic techniques that pave the way for more detailed classes, symmetric and asymmetric ciphers (Fig. 1):

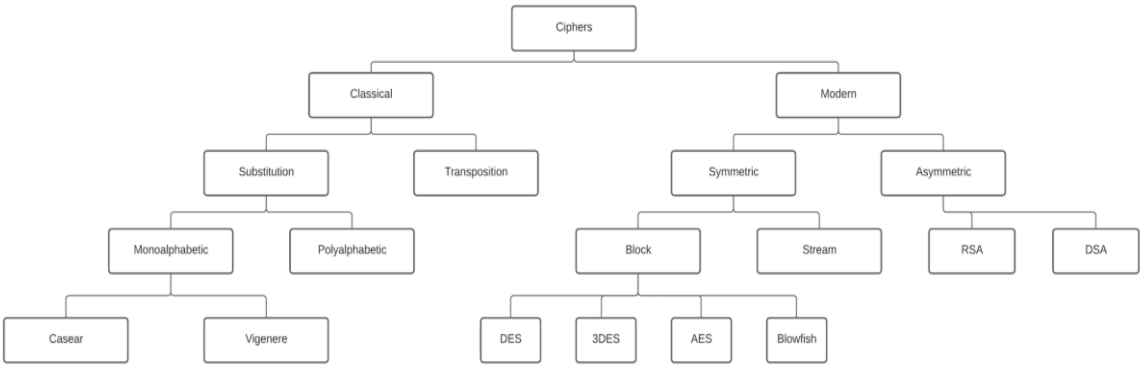


Figure 1. Classification of encryption algorithms

Symmetric encryption uses only one key to encrypt and decrypt data, while asymmetric encryption uses two different keys [2]. A public key is used for encryption, and a private key is used for decryption. Despite the differences between the public and private keys, they are related and are therefore called a key pair. DES, 3DES, AES, and Blowfish belong to the class of symmetric block ciphers.

Encryption Algorithms. DES is developed by IBM and based on a design by Horst Feistel. It was one of the widely used and publicly available cryptographic systems when it was first released. Even though its first debut is in the 70s, it was later adopted by the National Institute of Standards and Technology (NIST). It's a symmetric-key algorithm for the encryption of digital data. It has a block size of 64 bits and uses the Feistel network as a structure. It's slow and not used in the software.

Triple DES, 3DES, or TDES is officially the Triple Data Encryption Algorithm. It's a symmetric-key block cipher and it applies the DES algorithm three times to each block. It has a block size of 64 bits and a key length of 112 or 168 bits. It also uses the Feistel network since it's based on DES [2,3].

Due to the modern cryptology techniques and supercomputing, like the DES, 3DES has also some serious vulnerabilities. That's why the NIST has deprecated DES and 3DES for new applications in 2017 and for all applications by 2023.

Blowfish is another symmetric-key encryption technique designed by Bruce Schneier in 1993 as an alternative to the DES encryption algorithm. Therefore, it is significantly faster than DES and provides a good encryption rate. Its key length is 446 bits, and way better than DES, and 3DES. Therefore, it's more difficult to crack the key of Blowfish. It also has a block size of 64 bits. It can be used in software as well. However, AES receives more attention today, and Schneier recommends Towfish as an alternative to Blowfish. It has a free license and is available for all uses. AES is another type of cipher that protects the data from malicious parties. It's currently one of the strongest encryption algorithms available. Since AES combines speed and security properly, it allows us to

carry on with our online activities without any interruption. As AES uses the same key to both encrypt and decrypt data, it is also a symmetric type of encryption. There are three types of lengths of AES encryption keys 128, 192, and 256 bits. Each key length has different possible key combinations. It has a different structure than other encryption algorithms, it uses the substitution-permutation network.

Emerging Encryption Algorithms. Recent developments and emerging trends of cyber security. There are many recent developments in cyber security with the help of new algorithms, procedures and frameworks. Quantum Key Distribution (QKD) is a groundbreaking method in quantum cryptography that enables secure communication through the exchange of cryptographic keys [4]. Its security is rooted in the principles of quantum mechanics, particularly the no-cloning theorem and quantum entanglement, which make it virtually impossible for an eavesdropper to intercept or duplicate the key without being detected. Basic Principle:

QKD allows two parties (usually referred to as A and B) to securely exchange a cryptographic key over a potentially insecure channel. The security of QKD comes from quantum properties such as quantum superposition and entanglement.

Quantum Mechanics Behind QKD:

- **No-Cloning Theorem**: It states that an arbitrary quantum state cannot be copied exactly, meaning any attempt to intercept a quantum key would disturb the state and be detectable.
- **Heisenberg Uncertainty Principle**: Measuring a quantum state disturbs it, so any eavesdropping attempts can be easily detected by the parties involved.
- **Types of QKD Protocols**

There are several types of QKD protocols, but the most well-known ones include:

BB84 Protocol (1984):

- The first and most widely used QKD protocol, developed by Charles Bennett and Gilles Brassard.
- It uses **polarization** states of photons (or other quantum particles) to encode information. Alice sends photons in one of four states, and Bob measures them using randomly chosen bases.
- If no eavesdropping occurs, Alice and Bob will share a secure key.
- **E91 Protocol (1991)**:
- Based on **quantum entanglement**, where two particles (photons) are entangled in such a way that the measurement of one instantly affects the state of the other, even if they are far apart.
- This protocol uses **Bell's theorem** to ensure the authenticity and security of the key exchange.
- **Decoy State Protocols:**
- A modification to the basic QKD protocols, designed to mitigate the impact of photon number splitting (PNS) attacks, which can occur when an eavesdropper tries to intercept the photons in a weak coherent pulse.
- **Measurement-Device-Independent QKD (MDI-QKD):**
- Addresses the vulnerability of QKD systems to imperfections in detectors by enabling secure communication even when measurement devices are compromised.

Security of QKD

- **Quantum Security:**
- The key feature of QKD is its **unconditional security**. In theory, QKD can offer perfect security against any computational or classical cryptographic attack.
- Any eavesdropping or interference with the quantum channel results in observable errors that alert the communicating parties. This is because the quantum state will be disturbed, revealing the presence of an intruder.
- **Practical Security Concerns:**
- **Side-Channel Attacks**: Despite the inherent security of QKD, practical implementations can still be vulnerable to side-channel attacks, which exploit weaknesses in the physical hardware, such as detectors or transmission equipment.

- **Device Calibration & Trust**: The security of QKD systems relies heavily on trusted devices. If any device is compromised, the whole system's security can be compromised.

Post-Quantum Cryptography (PQC) involves developing cryptographic algorithms run on conventional computers that are secure against attacks from future, powerful quantum computers. These algorithms, such as lattice-based cryptography, replace current standards like RSA and ECC, which are vulnerable to Shor's algorithm [3,4]. PQC ensures long-term data security, preventing "harvest now, decrypt later" threats.

Categories of PQC Algorithms

PQC algorithms fall into several categories based on the type of problem they aim to solve and the type of cryptographic function they support.

- **Public-Key Cryptography** (asymmetric encryption):
- These are algorithms that rely on the difficulty of certain problems (e.g., integer factorization, discrete logarithms) which quantum computers can break but classical computers cannot efficiently solve.
- **Symmetric Cryptography**:
- These are used to encrypt and authenticate data using shared secrets. Quantum computers pose a smaller threat here (as Grover's algorithm only offers quadratic speedup, not an exponential one), but adjustments (like doubling key sizes) can maintain security.
- **Hash-Based Cryptography**:
- These use hash functions as the foundation for creating signatures and other forms of public-key cryptography. They are quantum-safe because quantum computers don't offer significant speedup over classical brute-force attacks on hash functions.

Candidate Algorithms for PQC (NIST Standardization)

The **NIST Post-Quantum Cryptography Standardization Process** aims to evaluate and standardize quantum-resistant algorithms. As of 2026, the final selections are still ongoing, but several promising candidates have emerged:

Public-Key Algorithms

- **Lattice-Based Cryptography**:
- Based on the hardness of lattice problems, which are believed to be resistant to quantum algorithms.
- Examples:
- **Kyber** (Encryption and Key Establishment)
- **NTRU** (Encryption and Key Exchange)
- **FrodoKEM** (Key Encapsulation)
- **NTS-KEM** (Key Encapsulation)
- **Code-Based Cryptography**:
- Based on error-correcting codes, specifically the difficulty of decoding random linear codes.
- Example: **McEliece** (Encryption)
- Known for large key sizes, but extremely resistant to quantum attacks.
- **Multivariate Polynomial Cryptography**:
- Relies on the difficulty of solving systems of multivariate polynomial equations over finite fields.
- Example: **Rainbow** (Digital Signatures)
- **Hash-Based Cryptography**:
- Based on the security of hash functions. While these are quantum-resistant, they generally have large signature sizes.
- Example: **XMSS** (Extended Merkle Signature Scheme)
- Used for signing messages with a smaller state, suitable for high-speed environments.
- **Isogeny-Based Cryptography**:
- Uses the mathematical structure of elliptic curves and their isogenies (maps between elliptic curves).

- Example: **SIKE** (Supersingular Isogeny Key Exchange)
- Considered a potential lightweight and efficient choice for key exchange.

Symmetric-Key Cryptography

- **AES (Advanced Encryption Standard)**:
- AES is already considered secure, but in a post-quantum world, it would require key lengths to be doubled (e.g., AES-256 instead of AES-128) to maintain security.
- **SHA-3 (Secure Hash Algorithm 3)**:
- SHA-3 is designed to be resistant to quantum attacks, as quantum computers do not provide significant speedup for hash functions.

Current State of PQC. NIST Standardization Process: The NIST post-quantum cryptography initiative is working towards finalizing the quantum-resistant standards.

As of 2026, the finalists for encryption and key exchange are Kyber and NTRU, and FrodoKEM is a possible alternative for scenarios requiring more conservative security.

For digital signatures, FALCON (lattice-based) and Rainbow (multivariate) are among the leading candidates.

Industry Adoption. While no large-scale commercial implementations of PQC are

widespread yet, various companies and organizations are testing PQC algorithms in pilot projects.

Some entities are using hybrid systems combining classical and quantum-safe algorithms to ensure security today and in the future.

Future of PQC.

- **Hybrid Cryptographic Systems**:
- During the transition period (while quantum computers are still under development), hybrid systems using both classical and post-quantum algorithms will likely be implemented for additional security.
- **Quantum Networks & Quantum Key Distribution (QKD)**:
- As quantum communication networks develop, PQC algorithms will be integrated into these systems, potentially along with QKD, to provide an extra layer of security.
- **AI and Machine Learning in Cryptography**:
- Machine learning and AI might help optimize PQC systems or analyze vulnerabilities in new PQC algorithms, driving further innovation.

Conclusion. In this article, we've briefly given some information about DES, 3DES, Blowfish, and AES, QKD, PQC. Also, we compared and evaluated them in terms of some metrics. Although we think that encryption algorithms protect the data both in theory and practice, keep in mind that there are various attacks such as malware, key search, brute force, phishing, side-channel, and more.

These attacks aim to find a way to hack the system. To sum up, while every encryption algorithm has weaknesses and advantages, some of them can lose their reliability over time. So, we should be careful when we select cryptographic algorithms. We should both consider our application and the specifics of the encryption algorithm.

References

[1] Andrea Goldsmith - Wireless Communications. -Cambridge University Press. – 2005 pp. 204-207

[2] Singhal, Nidhi and Raina, J P S. "Comparative Analysis of AES and RC4 Algorithms for Better Utilization", International Journal of Computer Trends and Technology, ISSN: 2231-280, July to Aug Issue 2011, pp. 177-181.

[3] Nadeem, Aamer; "A Performance Comparison of Data Encryption Algorithms", IEEE 2005.

[4] Abdulbast A. Abushgra Variations of QKD Protocols Based on Conventional System Measurements. – 2024

Author's contribution

The authors contributed equally to the writing of the article.

КЛАССИЧЕСКИЕ И СОВРЕМЕННЫЕ АЛГОРИТМЫ ШИФРОВАНИЯ ДЛЯ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Е.И. Лещевич
Выпускник БГУИР, преподаватель в Лоянский Педагогический Университет

Цай Юйсинь
Выпускник в Лоянский Педагогический Университет

Аннотация. В статье рассматриваются методы исследования технологий шифрования и дешифрования коммуникационных сигналов на основе новых алгоритмов в системах обработки больших данных.

Ключевые слова: шифрование, дешифрование, канальное кодирование, пакетные помехи.

УДК 004.42:658.152

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ИНВЕНТАРИЗАЦИИ И МОНИТОРИНГА СОСТОЯНИЯ ОСНОВНЫХ СРЕДСТВ ОРГАНИЗАЦИИ



Д.В. Лихачевский
Заместитель начальника Секретариата Наблюдательного совета ПВТ, кандидат технических наук, доцент
likhacheuski@park.by



С.А. Мигалевич
Магистр технических наук, начальник центра информатизации и инновационных разработок
migalevich@bsuir.by



В.О. Агапова
Магистрант кафедры проектирования информационно-компьютерных систем факультета компьютерного проектирования БГУИР,
agapova06@bsuir.by

Д.В. Лихачевский
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с исследованием проблем радиочастотной идентификации объектов, организацией учебного и научно-исследовательского процессов в техническом университете.

С.А. Мигалевич
Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: вибродиагностика, разработка метода вейвлет–анализа изделий машиностроения.

В.О. Агапова
Студент Белорусского государственного университета информатики и радиоэлектроники факультета компьютерного проектирования

Аннотация. В статье рассматривается процесс разработки программного комплекса для автоматизации инвентаризации и мониторинга состояния основных средств организации. Приведён анализ требований к системе, описаны архитектурные решения, выбранные технологии и этапы реализации. Выявлены преимущества использования такого комплекса, включая повышение эффективности учёта, снижение ошибок и оптимизацию

ресурсов. Рассмотрены возможные подходы к интеграции с существующими корпоративными системами и перспективы развития на основе анализа реальных кейсов внедрения.

Ключевые слова: программный комплекс, инвентаризация, мониторинг основных средств, автоматизация учёта, информационные системы, RFID, IoT.

Введение. В условиях цифровизации экономики эффективное управление основными средствами (ОС) организации является критическим фактором операционной эффективности и конкурентоспособности. Основные средства включают оборудование, недвижимость, транспорт и другие долгосрочные активы значительной стоимости. Традиционные методы учёта на основе бумажных документов и ручного ввода данных приводят к ошибкам, потере времени и неэффективному использованию ресурсов [1]. Разработка специализированного программного комплекса позволяет автоматизировать процессы инвентаризации и мониторинга, обеспечивая отслеживание состояния ОС в реальном времени.

Цель статьи – описать методологию разработки такого комплекса, включая анализ требований, проектирование, реализацию и оценку эффективности на основе реальных исследований и кейсов внедрения.

Основная часть. Разработка программного комплекса требует системного подхода и включает этапы анализа требований, проектирования архитектуры, выбора технологий, реализации, тестирования и внедрения.

1. Анализ требований к системе.

Система должна решать задачи учёта, отслеживания и анализа состояния ОС. Основные функциональные требования:

- Автоматизированная инвентаризация с использованием штрих-кодов, RFID-меток или QR-кодов для быстрой идентификации активов [1];



Рисунок 1. Ключевые функции системы трекинга активов на базе RFID

- Мониторинг состояния: отслеживание износа, ремонтов, амортизации и местоположения в реальном времени с применением GPS или IoT-датчиков [1];
- Генерация аналитических отчётов по стоимости, эффективности использования и прогнозам замены;
- Интеграция с ERP-системами для синхронизации данных.

Нефункциональные требования включают высокую производительность (обработка тысяч активов), безопасность данных, масштабируемость и удобный интерфейс [4].

Для сбора требований использовались интервью с бухгалтерами и логистами, анализ существующих систем и стандартов учёта (например, IAS 16).

2. Проектирование архитектуры.

Архитектура построена по клиент-серверной модели с элементами микросервисов. Основные компоненты:

- Фронтенд: веб-интерфейс на React.js для десктопов и мобильных устройств (дашборды, карты расположения активов);
- Бэкенд: сервер на Python (Django/Flask) с базой данных PostgreSQL для хранения атрибутов активов (инвентарный номер, дата приобретения, состояние, история);
- Модуль интеграции: RESTful API и webhooks (метод вебхуков);
- Модуль мониторинга: интеграция с IoT-платформами для сбора данных с датчиков [1].

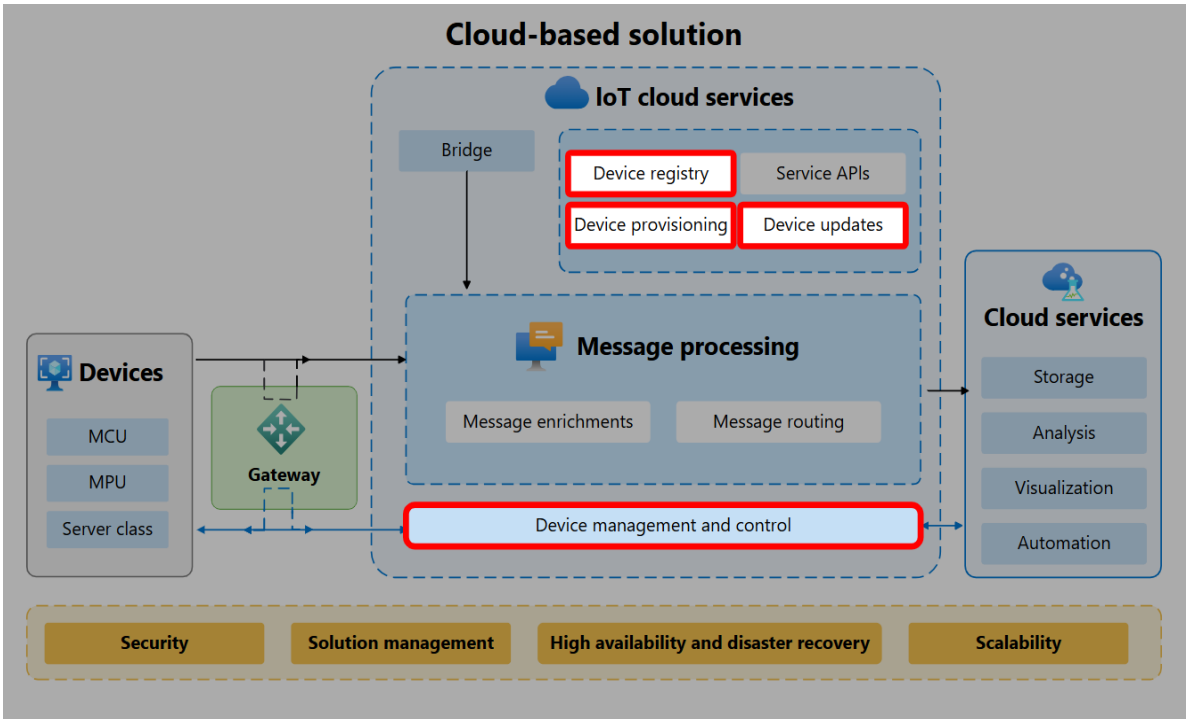


Рисунок 2. Архитектура системы управления активами на базе IoT (диаграмма из Microsoft Azure IoT)

Контейнеризация (Docker) и оркестрация (Kubernetes) обеспечивают отказоустойчивость. UML-диаграммы использовались для моделирования классов и последовательностей.

3. Выбор технологий и инструментов.

Технологический стек:

- JavaScript/TypeScript (фронтенд), Python (бэкенд) с библиотеками Pandas, NumPy;
- PostgreSQL с PostGIS для геолокации;
- Библиотеки ZXing (баркоды), RFIDIOt (RFID);
- Облачные сервисы (AWS/Google Cloud) для хостинга и масштабируемости;
- Безопасность: OAuth 2.0, HTTPS, AES-256.

Сравнение технологий представлено в таблице 1.

Таблица 1. Сравнение технологий для разработки комплекса

№	Технология	Преимущества	Ограничения	Применение в комплексе
1	React.js	Быстрая отрисовка, компонентный подход	Требует знаний JS	Фронтенд интерфейс
2	Python/Django	Богатые библиотеки для данных	Производительность под нагрузкой	Бэкенд логика
3	PostgreSQL	Поддержка GIS, транзакции	Сложность настройки	Хранение данных
4	AWS IoT	Интеграция с датчиками, масштабируемость	Стоимость	Мониторинг в реальном времени

4. Реализация комплекса.

Реализация проводилась по *agile*-методологии (итерации по 2 недели). Первая итерация – модуль инвентаризации со сканированием. Пример кода на Python для сканирования баркода:

```
import cv2
from pyzbar.pyzbar import decode

def scan_barcode(image_path):
    image = cv2.imread(image_path)
    barcodes = decode(image)
    for barcode in barcodes:
        return barcode.data.decode('utf-8')
    return None
```

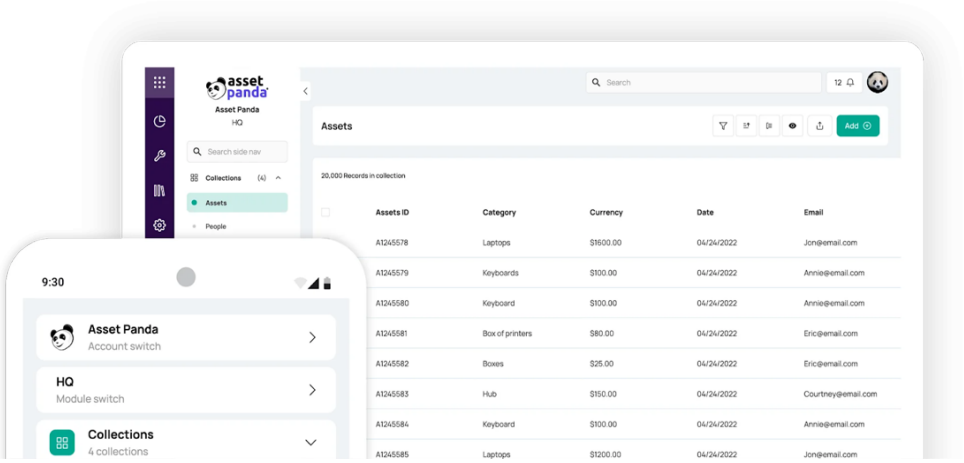


Рисунок 3. Пример мобильного интерфейса сканирования баркода в системе инвентаризации активов [Источник: assetpanda.com (Asset Panda), 2025–2026]

Вторая итерация – модуль мониторинга с дашбордом реального времени. Третья – интеграция и отчётность (Matplotlib для графиков амортизации). Общий объем кода – около 5000 строк, покрытие тестами – 80%.

5. Тестирование и внедрение.

Тестирование: unit-тесты (Pytest), интеграционные и нагрузочные (JMeter). Выявлено и исправлено 15 багов синхронизации.

Внедрение в тестовой организации (500 активов) показало значительное снижение времени инвентаризации (аналогично результатам в [1], где внедрение RTEM-системы с GPS, баркодами и RFID привело к росту использования активов на 30,8% за счет реального времени отслеживания).

В другом кейсе внедрения системы внутреннего контроля фиксированных активов в здравоохранении отмечено снижение ошибок на 95% и рост эффективности на 85% [2].

В веб-системе для управления лабораторными активами достигнуто значительное снижение ошибок учета и времени на инвентаризацию [4].

6. Анализ эффективности и перспективы развития.

Комплекс снижает ошибки учета (до 95% по аналогичным внедрениям [2]), оптимизирует затраты на обслуживание ОС и повышает compliance [4].

Перспективы: интеграция ИИ для предиктивного анализа износа и блокчейн для неизменяемости записей. Риски – зависимость от интернета и киберугрозы, противодействие – аудиты и обновления [5].

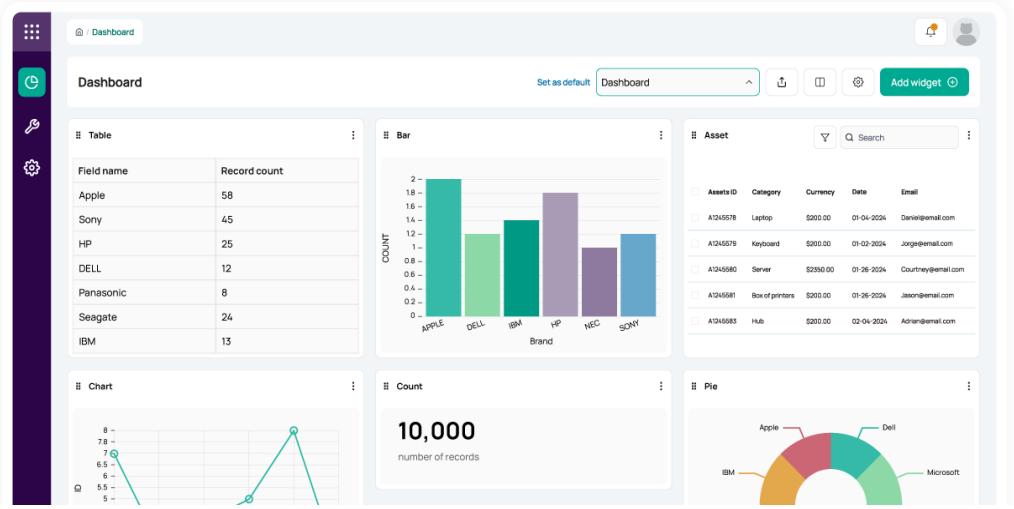


Рисунок 4. Пример дашборда мониторинга инвентаря и активов в реальном времени
[Источник: assetpanda.com (Asset Panda), 2025–2026]

Заключение. Разработка программного комплекса для инвентаризации и мониторинга ОС – комплексный процесс, сочетающий анализ, современные технологии и проверенные на практике подходы. Внедрение такого комплекса способствует повышению эффективности управления активами, минимизации рисков и адаптации к цифровой трансформации. Дальнейшее развитие должно учитывать интеграцию с ИИ и IoT для создания интеллектуальных систем учёта [5].

Список литературы

1. Iluore, O. E., Angela, M., Emetere, M. E. Development of asset management model using real-time equipment monitoring (RTEM): case study of an industrial company. – Cogent Business & Management, 2020.

2. Pérez, E. C. M., Peregrino, V. M. A. Benefits of Implementing the Internal Control System for Fixed Assets in the Inventory Department of the Health Department of the State of Tabasco. – Innovación y desarrollo tecnológico revista digital, 2023.

3. Wang, T., Wang, S. Optimization of Asset Management System Based on Computer Aided Technology of Internet of Things. – Computer-Aided Design and Applications, 2021.

4. Karim, M., Saad, M. F., Haque, M. Development of a prospective web-based inventory system for management of lab facilities. – Journal of Emerging Trends in Engineering and Applied Sciences, 2011.

5. Stanimirović, D., Umek, L., Ravšelj, D. Charting Advances in Asset Management Systems: A Bibliometric Analysis Revealing Applications and Potential in Healthcare. – Healthcare, 2025.

Авторский вклад

Лихачевский Дмитрий Викторович – постановка и формализация задач по созданию программного комплекса инвентаризации и мониторинга основных средств организации, включая разработку моделей данных для описания жизненного цикла объектов учёта, сравнительный анализ архитектурных подходов (монолитная, микросервисная) к построению системы, а также исследование эффективности алгоритмов обработки потоковых данных при мониторинге состояния оборудования в условиях ограничений на вычислительные ресурсы и пропускную способность сетей.

Мигалевич Сергей Александрович – организация и научное сопровождение исследований, ориентированных на применение современных веб-технологий и методов интеллектуального анализа данных для построения масштабируемых и отказоустойчивых механизмов централизованного учёта и мониторинга основных средств. Обеспечение согласованной работы модулей инвентаризации и подсистем прогнозирования износа в условиях высокой нагрузки при одновременном обслуживании большого количества единиц учёта и обработки данных в режиме, приближенному к реальному времени.

Агапова Виктория Олеговна – исследование архитектурных и алгоритмических решений для построения распределённой системы сбора и анализа данных о состоянии основных средств, включая разработку моделей обмена данными между мобильными клиентами инвентаризации и серверной частью, механизмов приоритизации задач по обслуживанию оборудования и алгоритмов адаптивной маршрутизации потоков данных при нестабильном качестве каналов связи и неполноте первичной информации об объектах учёта.

DEVELOPMENT OF A SOFTWARE COMPLEX FOR INVENTORY AND MONITORING OF THE STATE OF THE ORGANIZATION'S FIXED ASSETS

D.V. Likhachevsky
Deputy Director of the Secretariat of Hi-Tech Park Supervisory Board, Candidate of Technical Sciences, Associate Professor

S.A. Migalevich
Master of Technical Sciences, Head of the Center for Informatization and Innovative Developments

V.O. Agapova
Student of the Belarusian State University of Informatics and Radioelectronics

Annotation: The article considers the process of developing a software complex for automating the inventory and monitoring of the state of the organization's fixed assets. An analysis of the system requirements is presented, architectural solutions, selected technologies and implementation stages are described. The advantages of using such a complex are identified, including increasing the efficiency of accounting, reducing errors and optimizing resources. Possible approaches to integration with existing corporate systems and development prospects are considered based on real implementation cases.

Keywords: software complex, inventory, fixed assets monitoring, accounting automation, information systems, RFID, IoT.

УДК 004.8:004.056

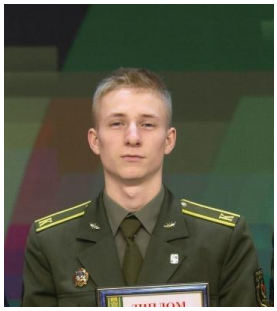
ГРАФОВЫЕ НЕЙРОННЫЕ СЕТИ КАК НАПРАВЛЕНИЕ РАЗВИТИЯ В СИСТЕМАХ БЕЗОПАСНОСТИ СИЛОВЫХ ВЕДОМСТВ



Д.А. Лагутик
Курсант военного факультета БГУИР
lagutikdmitrij40@gmail.com



А.Ю. Савицкий
Старший преподаватель кафедры связи, БГУИР, кандидат военных наук
a.savitskij@bsuir.by



А.Е. Козлов
Курсант военного Факультета БГУИР
anton.kozlov.mailbox@gmail.com

А.Е. Козлов

Обучается в БГУИР по специальности «Компьютерная инженерия». Область научных интересов связана с изучением автоматизации обработки данных с помощью машинного обучения и искусственного интеллекта.

Д.А. Лагутик

Обучается в БГУИР по специальности «Системы и сети инфокоммуникации». Область научных интересов связана с изучением автоматизации обработки данных с помощью машинного обучения и искусственного интеллекта.

А.Ю. Савицкий

Окончил адъюнктуру Военной академии связи имени С.М. Буденного, кандидат военных наук. Область научных интересов связана с совершенствованием научно–методического аппарата оценки эффективности построения систем связи в общевойсковых соединениях и воинских частях, обоснованием принимаемых решений.

Аннотация. Статья раскрывает механизмы интеграции графовых нейронных сетей в системы безопасности силовых ведомств, анализирует действующие случаи внедрения и предлагает архитектурное решение для динамических графов угроз – с поправкой на правовые нормы государств – участников СНГ.

Ключевые слова: Графовые нейронные сети, неевклидовые структуры данных, динамические графы, федеративное обучение, разведывательный анализ, обнаружение аномалий, предиктивная аналитика угроз, масштабируемость данных, защита персональных данных, искусственный интеллект.

Введение. В статье рассматривается потенциал графовых нейронных сетей (Graph Neural Networks, GNN) для задач силовых ведомств, таких как анализ угроз, разведка и кибербезопасность. Предлагается концепция внедрения современных динамических графовых сетей в системы мониторинга и прогнозирования. Рассматриваются преимущества, вызовы и примеры внедрения с опорой на актуальные научные публикации.

Теоретические основы графовых нейронных сетей и основные архитектуры построения. Графовые нейронные сети представляют собой класс моделей глубокого обучения, адаптированных для работы с неевклидовыми структурами данных. В отличие от сверточных сетей, работающих с регулярными сетками, GNN обрабатывают переменное количество узлов и ребер, сохраняя информацию о топологии связей.

Ключевым механизмом GNN является механизм передачи сообщений (message passing): на каждом слое узел агрегирует информацию от соседей, постепенно формируя представления, учитывающие многоуровневые зависимости в графе [1]. Этот процесс имитирует диффузию информации через сеть, где каждый узел обменивается информацией только с непосредственными соседями, но за несколько итераций может уловить глобальную структуру графа, что является основным преимуществом данного подхода. На каждом слое message passing проходит три этапа: каждый соседний узел формирует персонализированное сообщение на основе своих характеристик, данных целевого узла и свойств связи между ними. Затем целевой узел собирает эти сообщения в промежуточное представление, используя простые операции. Наконец, узел обновляет своё состояние, комбинируя предыдущие знания с агрегированной информацией через нейронную сеть, что позволяет постепенно учитывать всё более широкую область графа. Таким образом, для узла v на слое k:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \mathrm{AGGREGATE}^{(k)}\left(\{h_u^{(k-1)} : u \in N(v)\}\right)\right)$$

N(v) – множество соседей узла v, σ – функция активации, $W(k)$ – обучаемые веса, AGGREGATE$^{(k)}$ – функция агрегации сообщений.

Для реализации механизма передачи сообщений данные моделируются в виде графа, где узлы представляют сущности, а ребра – связи между ними (рисунок 1).

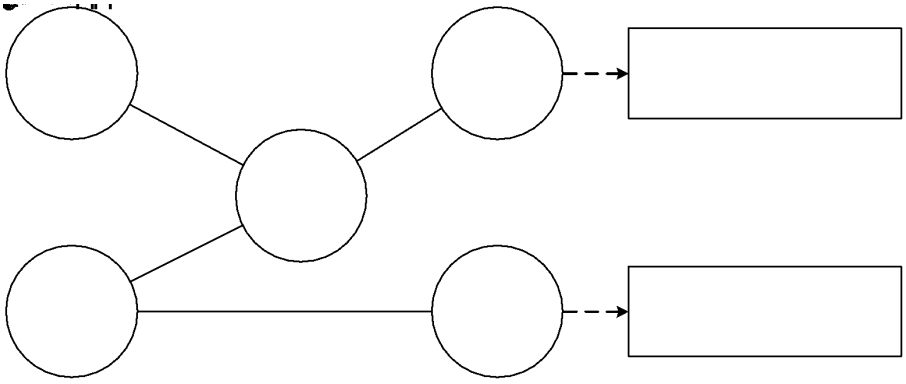


Рисунок 1. Структура графа

h_0, h_1, h_2, h_3, h – узлы гетерогенного графа (графа, элементы которого могут принадлежать к различным типам (категориям), имеющим разную семантику и свойства), $[a_1, a_2, a_3 \dots a_n]$ – итоговый вектор–эмбеддинг для узла графа, n – размерность вектора–эмбеддинга (количество признаков), W_{ij} – связь между узлами h_i и h_j.

Основные архитектуры построения графовых нейронных сетей [1]:

1. Graph Convolutional Networks применяют свертку, учитывающую нормировку по степеням узлов, что позволяет сглаживать признаки по локальному окружению. Однако для больших графов вычислительная сложность может стать ограничивающим фактором.

2. GraphSAGE решает проблему масштабируемости за счет сэмплирования фиксированного числа соседей, что делает обучение индуктивным – модель способна генерировать представления для ранее не встречавшихся узлов.

3. Graph Attention Networks вводят механизм внимания, динамически взвешивая вклад каждого соседа при агрегации. Это позволяет модели выделять наиболее значимые связи, что особенно ценно для выявления подозрительных взаимодействий в разведданных.

4. Gated Graph Neural Networks используют рекуррентные архитектуры для обработки динамических графов, где структура связей изменяется во времени.

Однако перечисленные архитектуры изначально ориентированы на статические графы, тогда как реальные сценарии безопасности характеризуются непрерывной эволюцией данных. Для работы с динамическими графами предложены специализированные подходы:

1. EvolveGCN адаптирует параметры во времени с помощью рекуррентных сетей, не требуя знания всех узлов на всем временном интервале, что критично для сценариев с частым изменением состава узлов.

2. Temporal Graph Attention использует функциональное кодирование времени и механизм внимания для агрегации временно–топологических признаков, обеспечивая индуктивное обучение на новых узлах.

3. Temporal Graph Networks представляют универсальную рамку, сочетающую модули памяти и графовые операторы для эффективного обучения на последовательностях временных событий.

Данные подходы позволяют моделировать процессы, где связи и признаки узлов изменяются во времени – например, коммуникационные сети, финансовые транзакции или перемещения объектов [2 – 3].

В силовых ведомствах в сфере анализа данных сводятся к следующему: обнаружение скрытых коммуникаций в условиях намеренной маскировки, выявление аномальных моделей поведения в потоковых данных, прогнозирование перемещений объектов по фрагментарным сигналам, интеграция разнородной информации из несовместимых источников. Традиционные методы машинного обучения упираются в барьер: они теряют топологические зависимости, преобразуя графовые данные в табличные, не способны масштабироваться на сети с миллиардами узлов или не учитывают эволюцию структур во времени. В силовых задачах GNN интегрируют разнородные данные: социальные связи, геолокацию, финансовые потоки. Они превосходят традиционные методы в обнаружении аномалий, поскольку учитывают контекст – изолированный узел с высокой активностью в кластере будет отмечен как угроза, в то время как тот же узел в разреженной сети может оказаться нормой. Преимущества для ведомств: масштабируемость на миллиарды узлов, устойчивость к шуму в данных и способность работать в реальном времени. Например, в системах мониторинга GNN самообучаются на инцидентах, произошедших ранее, минимизируя ложные срабатывания и ускоряя реагирование на новые угрозы [4].

Анализ применения GNN в силовых структурах других стран.

Система MetaConstellation, разработанная Palantir Technologies для Пентагона, представляет собой платформу на базе искусственного интеллекта (ИИ) и графовых моделей для анализа разведданных в реальном времени. Она интегрирует спутниковые снимки, перехваты сигналов, социальные сети и логистику в единый гетерогенный граф. Израильская

система Gospel от Unit 8200 автоматизирует разведку в Газе и Ливане, формируя графы на основе многолетних данных о коммуникациях и перемещениях.Китайская Intelligent Precision Strike System в Силах стратегической поддержки НОАК (Народно-освободительная армия Китая) использует GNN для управления роями дронов и ситуационной осведомленности [5-7].

Несмотря на различия в предметных областях, все три системы демонстрируют общую тенденцию: переход от обработки изолированных данных к моделированию сложных взаимосвязей в динамических графах (таблица 1).

Таблица 1. Сравнительный анализ графовых систем в силовых структурах

Критерии	**MetaConstellation**	**Gospel**	**Intelligent Precision Strike System**
Заказчик	США	Израиль	Китай
Разработчик	Palantir Technologies	Unit 8200	Силы стратегической поддержки НОАК
Тип узлов	Цели, техника, лица	Боевики, телефоны, дома	Дроны, радары, спутники
Тип ребер	Перемещения, коммуникации, транзакции	Звонки, перемещения, покупки	Каналы связи, траектории угроз
Ключевая архитектура	GraphSAGE, GAT	Message passing	Graph Attention
Решаемая задача	Предсказание траекторий угроз, генерация сценариев	Ранжирование целей, трекинг в реальном времени	Управление роем дронов, ситуационная осведомленность
Достигнутый эффект	Сокращение времени анализа с часов до минут (30–50%); точность целеуказания +40%	Сокращение цикла «обнаружение – удар» с дней до часов; эффективность ударов +25–30%	Точность прогноза уязвимостей 85–95%; сокращение времени реакции на 60%; масштабируемость 10000 узлов
Особенность	Explainable AI (визуализация весов внимания)	Контекстный анализ связей для минимизации гражданских потерь	Интеграция кибербразведки и помехозащищенность

Рассмотренные системы демонстрируют различные сценарии применения GNN. При этом все они объединяет переход от реактивного анализа (реагирование на уже возникшие события или проблемы на основе исторических данных и текущих сигналов) к предиктивному: вместо фиксации факта коммуникации или перемещения модели предсказывают намерения по цепочкам связей.

Ключевым фактором эффективности выступает способность работать с неполными и зашумленными данными – GNN достраивают недостающие звенья цепочек через топологический контекст, тогда как традиционные методы требуют полной информации.

Кибербезопасность и мониторинг с GNN.

Применение GNN в задачах кибербезопасности демонстрирует устойчивый тренд: переход от изолированного анализа признаков к учету топологических связей обеспечивает качественный рост точности при одновременном снижении ложных срабатываний. Сопоставление результатов независимых исследований 2024 года показывает, что графовые модели систематически превосходят как классические правила, так и нейросетевые базлайны вроде MLP и LSTM (рисунок 2).

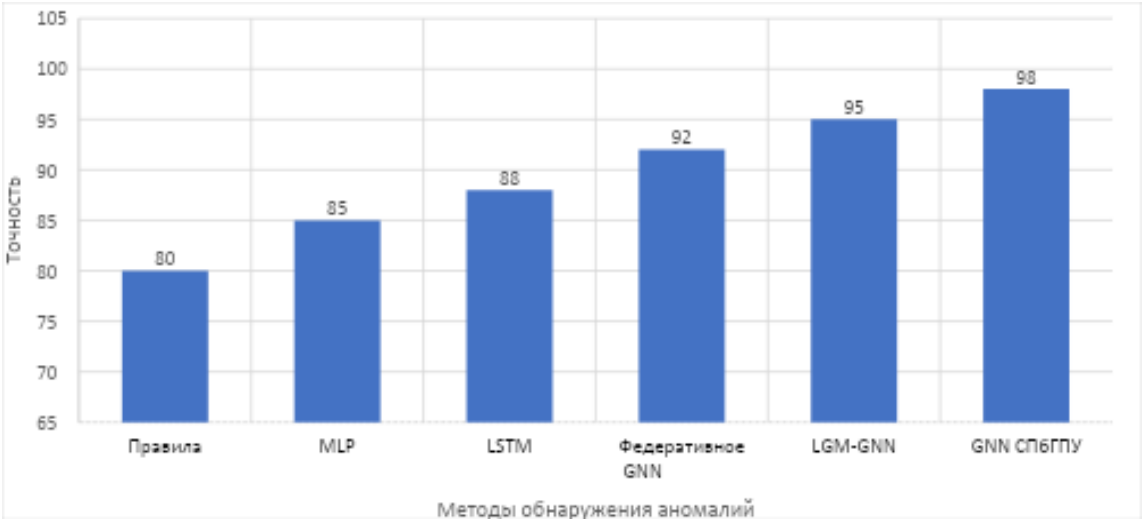


Рисунок 2. Сравнение точности методов обнаружения аномалий

GNN кардинально меняют подход к выявлению аномалий в сетях силовых структур, самообучаясь на помехах и динамических данных, что минимизирует ложные срабатывания по сравнению с традиционными методами. Самообучающиеся GNN мониторят объекты в реальном времени, строя графы трафика (узлы – IP/устройства, ребра – пакеты), где агрегация соседей выявляет скрытые атаки вроде APT (Advanced Persistent Threat) или ботнетов. Министерство внутренних дел Российской Федерации планирует ИИ для поиска правонарушителей, интегрируя GNN в системы видеонаблюдения и соцсетей. Модель на сетевых данных (СПбГПУ, 2024) обнаружила 98% истинных аномалий при FPR всего 2% – в 5–10 раз ниже, чем у MLP или правил. LGM – GNN в финансовых графах достигла accuracy 95%, полноты 92%, снижая FPR на 40%. В спутниковой связи GNN классифицируют трафик с точностью 92–96%, ускоряя анализ в 3 раза. Федеративная GNN для сетевых аномалий показала точность в 92%, сохраняя конфиденциальность данных.

Предлагаемая архитектура для силовых ведомств

Комплексная система на базе GNN строится по трехуровневой схеме, обеспечивающей полный цикл от агрегации разнородных данных до оперативного реагирования и непрерывной адаптации. Каждый уровень решает специфические задачи, критичные для функционирования в условиях высокой динамики угроз и жестких требований к защите информации. (рисунок 3).



Рисунок 3. Предлагаемая комплексная система

Предлагаемая архитектура реализует сквозной конвейер обработки разведывательной информации.

На входном уровне потоковые данные из физически и логически разрозненных источников нормализуются в единое графовое представление с сохранением временных и семантических характеристик.

Ядро системы – гибридный GNN–энкодер, сочетающий индуктивные свойства GraphSAGE (работа с ранее неизвестными объектами без переобучения) и механизмы внимания GAT (динамическое выделение значимых связей и интерпретируемость решений).

Выходной уровень предоставляет программные интерфейсы для взаимодействия с системами реагирования, интерактивную визуализацию для аналитиков и непрерывный мониторинг качества модели с автоматической адаптацией при деградации. Контейнеризированная инфраструктура гарантирует масштабируемость и отказоустойчивость при работе с большими объемами.

Решением проблемы централизации чувствительных данных в GNN является федеративное обучение, которое представляет собой децентрализованный подход к обучению GNN, позволяющий агрегировать знания из распределённых источников данных без их централизации, что особенно актуально для структур государственного уровня.

В контексте графовых сетей модель инициализируется на центральном сервере и рассылается клиентам (ведомствам, региональным узлам), где локально обучается на проприетарных графах трафика или взаимодействий с последующей передачей только градиентов или обновлённых весов.

Важным дополнением к техническим аспектам внедрения искусственного интеллекта является правовое регулирование данной сферы.

В государствах–участниках СНГ с 2021 года ведутся работы по созданию единых подходов к регулированию ИИ в рамках Межгосударственной программы инновационного сотрудничества до 2030 года.

Как отмечается в работе [8], безусловным приоритетом должно стать обеспечение защиты прав человека и его основных свобод перед лицом технологического прогресса. Ключевыми проблемами, требующими законодательного закрепления, остаются вопросы гражданско–правовой и уголовной ответственности за действия автономных систем, поскольку традиционные правовые нормы не всегда применимы к нечеловеческим субъектам.

Кроме того, критическое значение приобретает обеспечение нейтральности и непредвзятости алгоритмов во избежание дискриминационных решений, основанных на смещенных выборках данных.

Особую актуальность в контексте силовых ведомств приобретает принцип транспарентности функционирования технологий ИИ, позволяющий преодолеть эффект «черного ящика» и обеспечить внешний контроль за принимаемыми решениями.

Таким образом, эффективное внедрение интеллектуальных систем невозможно без гармонизированного правового пространства, сочетающего механизмы государственного регулирования и этические принципы, что подтверждается необходимостью разработки специальных нормативных актов, таких как модельный закон для государств–участников СНГ.

Заключение. Графовые нейронные сети переводят анализ угроз силовых ведомств на прогнозирующий уровень, обеспечивая моделирование сложных взаимосвязей в динамических графах. Предложенная архитектура – гибридный GNN–энкодер с федеративным обучением и мониторингом состояния работы – ориентирована на специфику разведывательных задач и требования защиты информации.

Внедрение GNN в силовых структурах крайне важно для поддержания оперативного превосходства: технологии позволяют опережать динамику угроз, минимизировать риски и оптимизировать ресурсы.

Стратегическая потребность во взаимодействии превосходит краткосрочные затраты на настройку инфраструктуры, обеспечивая долгосрочное преимущество в условиях растущей сложности вызовов безопасности. При этом успешное развертывание требует согласования технических решений с правовыми и этическими стандартами, которые формируются в ходе международного сотрудничества.

Список литературы

[1] Mitra S. и др. Use of Graph Neural Networks in Aiding Defensive Cyber Operations / S. Mitra, T. Chakraborty, S. Neupane, A. Piplai, S. Mittal // arXiv. – 2024. – Арх. № 2401.05680v1. – URL: https://arxiv.org/html/2401.05680v1 (дата обращения: 13.03.2026).

[2] Pareja A. и др. EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs / A. Pareja, G. Domeniconi, J. Chen, T. Ma, T. Suzumura, H. Kanezashi, T. Kaler, C. E. Leiserse // IBM Research Publications. – 2020.

[3] Xu, D. Inductive Representation Learning on Temporal Graphs / D. Xu, C. Ruan, E. Körpeoglu, S. Kumar, K. Achan // arXiv preprint. – 2020. – Арх. № 2002.07962. – URL: https://arxiv.org/abs/2002.07962 (дата обращения: 13.03.2026).

[4] Rossi E. и др. Temporal Graph Networks for Deep Learning on Dynamic Graphs / E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, M. M. Bronstein // arXiv. – 2020. – Арх. № 2006.10637. – URL: https://arxiv.org/abs/2006.10637 (дата обращения: 13.03.2026).

[5] Horowitz M.C. Artificial Intelligence, International Competition, and the Balance of Power / M. C. Horowitz // Texas National Security Review.– 2018.– Т. 1, № 3.– С. 36–57.

[6] Аль-Сахель С., Смирнов А. Технологии искусственного интеллекта в палестино-израильском конфликте / С. Аль-Сахель, А. Смирнов // Международная жизнь.– 2023.

[7] Изюмов Д.Б., Кондратюк Е.Л. Анализ различий в подходах США и Китая к применению искусственного интеллекта в системах вооружения / Д. Б. Изюмов, Е. Л. Кондратюк // Инноватика и экспертиза.– 2022.– № 2(34).– С. 228–239.

[8] Абломейко С.В. [и др.] Основные положения модельного закона «Об искусственном интеллекте» / С. В. Абломейко, [и др.] // BIG DATA and Advanced Analytics.– 2024.– Т. 1, №1.– С. 21–31.

Авторский вклад

Козлов Антон Евгеньевич – формализация теоретических основ для разведывательных задач, руководство разработкой рекомендаций по интеграции с национальными платформами безопасности.

Лагутик Дмитрий Александрович – разработка архитектуры GNN, создание методологии поэтапного внедрения графов в командные системы, изучение мировых примеров применения.

Савицкий Алексей Юрьевич – формулировка задачи исследования графовых нейронных сетей в качестве стратегического направления развития систем безопасности силовых структур.

GRAPH NEURAL NETWORKS IN SECURITY SYSTEMS OF LAW ENFORCEMENT AGENCIES

D.A. Lagutik
Student of the military faculty of BSUIR

A.Yu. Savitsky
Senior Lecturer of the Department of Communications, BSUIR, Candidate of Military Sciences

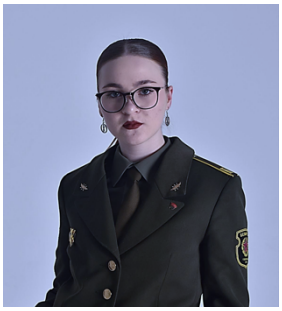
A.E. Kozlov
Student of the military faculty of BSUIR

Abstract. The article examines the mechanisms of integrating graph neural networks into security systems of law enforcement agencies, analyzes existing implementation cases, and proposes an architectural solution for dynamic threat graphs – adjusted to comply with the legal norms of CIS member states.

Keywords: Graph neural networks, non-Euclidean data structures, dynamic graphs, federated learning, intelligence analysis, anomaly detection, predictive threat analytics, data scalability, personal data protection, artificial intelligence.

УДК 004.65

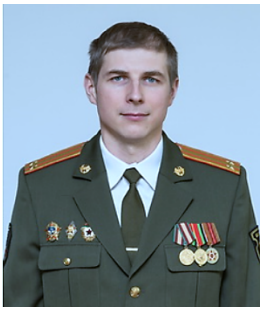
ПРИМЕНЕНИЕ ВЕКТОРНЫХ БАЗ ДАННЫХ ДЛЯ ОРГАНИЗАЦИИ ПОИСКА ТЕХНИЧЕСКОЙ И УЧЕБНОЙ ДОКУМЕНТАЦИИ В ОБЛАСТИ ТЕЛЕКОММУНИКАЦИЙ



А.В. Бардашевич
Курсант, БГУИР
sasha.bard9364@gmail.com



А.Ю. Савицкий
Старший преподаватель кафедры связи, БГУИР, кандидат военных наук
a.savitskij@bsuir.by



В.А. Федоренко
Начальник цикла кафедры связи, БГУИР
w.fedorenko@bsuir.by

А.В. Бардашевич
Курсант военного факультета Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с применением методов искусственного интеллекта и машинного обучения в образовательном процессе.

А.Ю. Савицкий
Окончил адъюнктуру Военной академии связи имени С.М. Буденного, кандидат военных наук. Область научных интересов связана с совершенствованием научно-методического аппарата оценки эффективности построения систем связи в общевойсковых соединениях и воинских частях, обоснованием принимаемых решений.

В.А. Федоренко
Окончил Военную академию Республики Беларусь. Область научных интересов связана с анализом существующих и разработкой новых методик защиты от несанкционированных атак транспортных сетей специального назначения.

Аннотация. Рассмотрена проблема неэффективного поиска учебной информации в подготовке специалистов в области телекоммуникаций, обусловленная семантической сложностью технической терминологии и межъязыковым барьером. Предложен подход к организации семантического поиска на основе векторных баз данных, обеспечивающий нахождение материалов по смыслу. Обоснована целесообразность использования векторного расширения для реляционной системы управления базами данных в образовательных учреждениях. Показано, что предложенное решение позволяет интегрировать векторный поиск в существующую инфраструктуру, обеспечивая поддержку межъязыковых запросов и фильтрацию по метаданным.

Ключевые слова: векторные базы данных, семантический поиск, pgvector, PostgreSQL, эмбеддинги, HNSW

Введение. Современная индустрия телекоммуникаций переживает период интенсивного технологического развития, обусловленный внедрением сетей пятого поколения, развитием технологий виртуализации сетевых функций, программно-конфигурируемых сетей и концепции Open RAN. По данным отчёта Ericsson Mobility Report [1], общий месячный глобальный трафик мобильных сетей достиг 200 эксабайт в четвертом квартале 2025 года, при этом годовой рост составил 22%.

Одновременно с усложнением телекоммуникационных систем экспоненциально растёт объем технической документации и образовательных материалов, что создает проблему информационной перегрузки в образовательном процессе. Традиционные методы поиска, по

ключевым словам, демонстрируют низкую эффективность из-за семантической сложности предметной области: техническая терминология содержит много синонимов, аббревиатур и многозначных терминов, а значительная часть документации опубликована на английском языке, тогда как обучение ведется преимущественно на русском. В результате преподаватели и обучающиеся тратят значительное время на поиск релевантных материалов, а системы поиска по ключам не учитывают контекст запроса и не обеспечивают межъязыковое сопоставление информации.

Решением данной проблемы является применение векторных баз данных для организации семантического поиска учебного материала. Векторные системы управления базами данных (СУБД) позволяют осуществлять поиск по смыслу, а не по точному совпадению слов, что обеспечивает устойчивость к синонимии, поддержку межъязыковых запросов и учёт контекста использования терминов.

Анализ существующих решений. Традиционные подходы к поиску в образовательных ресурсах основываются на методах полнотекстового поиска (TF-IDF, BM25), которые оценивают релевантность документов по частоте встречаемости поисковых терминов [2]. Несмотря на вычислительную эффективность, эти методы не учитывают семантическую близость между словами, что критично для телекоммуникационной предметной области с ее высокой степенью синонимии и вариативности терминологии. Например, запрос «расчет запаса на замирания» не найдет документ с термином «fade margin calculation» из-за отсутствия лексического совпадения.

Развитие технологий обработки естественного языка привело к появлению методов семантического поиска на основе векторных представлений текста. Модели sentence embeddings, такие как Sentence-BERT [3], позволяют кодировать семантику текста в векторы фиксированной размерности, сохраняя возможность вычисления семантической близости через косинусное сходство. Для эффективного хранения и поиска по многомерным векторам применяются векторные базы данных, использующие алгоритмы приближенного поиска ближайших соседей, в частности Hierarchical Navigable Small World (HNSW) – один из самых быстрых и точных алгоритмов для поиска приблизительных ближайших соседей в многомерных векторных пространствах, обеспечивающий логарифмическую сложность поиска O(log n) [4].

Современный рынок предлагает различные решения для векторного поиска: специализированные СУБД (Qdrant, Weaviate, Pinecone), библиотеки и расширения для традиционных реляционных систем. Для образовательных учреждений телекоммуникационного профиля наиболее перспективным представляется использование расширения pgvector для СУБД PostgreSQL. Данный подход обеспечивает ряд преимуществ: возможность выполнения сложных SQL-запросов с фильтрацией по метаданным (домен, технология, уровень сложности), ACID-совместимость и надежность, характерные для PostgreSQL, а также упрощение развертывания за счёт использования единой инфраструктуры для хранения как структурированных данных, так и векторных эмбеддингов. Исследования показывают, что производительность pgvector при работе с количеством до 1 миллиона векторов сопоставима со специализированными решениями, при этом время отклика составляет менее 100 мс. для большинства запросов. Архитектура Retrieval-Augmented Generation сочетает извлечение информации из векторной базы данных с генеративными возможностями языковых моделей, что позволяет не только находить релевантные учебные материалы, но и формировать связные ответы с цитированием источников [5].

В контексте образовательных систем телекоммуникационного профиля данная архитектура обеспечивает достоверность предоставляемой информации за счёт опоры на верифицированную предметную базу данных. Retrieval-компонент архитектуры реализован на основе расширения pgvector для PostgreSQL, что обеспечивает эффективный семантический поиск с фильтрацией по метаданным (домен связи, технология, уровень сложности). Такой подход позволяет интегрировать систему в существующую инфраструктуру образовательных

учреждений без развертывания дополнительных специализированных решений, сохраняя при этом возможность выполнения сложных запросов и ACID-совместимость.

Архитектура предлагаемой системы семантического поиска (рисунок 1) построена на основе расширения pgvector для СУБД PostgreSQL и включает четыре основных компонента: модуль сбора и предобработки документов, модуль векторизации текста, векторное хранилище на базе pgvector и модуль векторного поиска.

Обработка данных начинается с подготовки исходных документов, включающая очистку текста, стандартизацию терминов и сегментацию на логические фрагменты.

Полученные блоки кодируются в векторную форму с помощью модели эмбеддингов и загружаются в СУБД PostgreSQL, где к ним добавляются описательные метаданные (домен связи, технология, сложность, источник).

Поиск реализуется через векторизацию пользовательского запроса аналогичным методом и сопоставление с индексированными данными посредством алгоритма HNSW.

Ранжирование результатов выполняется согласно метрике косинусного сходства, обеспечивая выдачу релевантных материалов с информацией об источнике.



Рисунок 1. Архитектура семантического поиска

Процесс векторизации учебных материалов начинается с извлечения текста из документов различных форматов (DOCX, HTML, PDF). Для очистки текста применяется нормализация Unicode-символов, удаление избыточных пробелов и переносов строк, сохранение технических обозначений (частоты, мощности, формулы). Особое внимание уделяется обработке телекоммуникационной терминологии: аббревиатуры раскрываются при первом упоминании, синонимы приводятся к канонической форме, английские термины сохраняются в оригинале для обеспечения межъязыкового поиска. Разбиение текста на чанки

выполняется рекурсивным методом с приоритетом сохранения логической структуры документа (разделы, подразделы, абзацы). На основе анализа структуры технической документации обоснованы параметры чанкинга: размер чанков 800 токенов с перекрытием 150 токенов, что обеспечивает баланс между сохранением контекста и точностью поиска (таблица 1).

Таблица 1. Параметры чанкинга

Размер чанка (токены)	Перекрытие (токены)	Точность поиска	Время индексации	Рекомендации
512	100	0.82	Быстро	Для простых документов
800	150	0.89	Средне	Оптимально
1024	200	0.88	Медленно	Для сложных текстов

Для преобразования текста в векторные представления выбрана модель paraphrase-multilingual-MiniLM-L12-v2, поддерживающая русский и английский языки. Данная модель обеспечивает размерность вектора 384, что представляет собой компромисс между качеством эмбеддингов и требованиями к памяти. Модель обучена на разных наборах данных, включая техническую литературу, что обеспечивает качественное кодирование телекоммуникационной терминологии. Векторизация выполняется на стороне приложения с использованием библиотеки sentence-transformers, полученные векторы передаются в PostgreSQL для хранения в таблице с расширением pgvector.

Хранение векторных эмбеддингов организовано в таблице PostgreSQL со следующей структурой: уникальный идентификатор чанка, текстовое содержимое, вектор, метаданные в формате JSONB (домен, поддомен, технология или стандарт, тема, уровень сложности материала, тип контента, источник, страница). Для ускорения поиска по векторам создается индекс HNSW с параметрами: метрика косинусного сходства, количество связей каждого узла в графе M = 16, размер списка кандидатов при построении ef_construction = 128 – задает качество построения индекса, при поиске ef_search = 100 – контролирует точность поиска. Данные параметры обеспечивают время отклика менее 100 мс. при размере базы до 100 000 векторов. Преимуществом использования PostgreSQL является возможность выполнения гибридных запросов, сочетающих векторный поиск с фильтрацией по метаданным через стандартные SQL-условия WHERE.

Алгоритм семантического поиска включает следующие этапы (рисунок 2): прием пользовательского запроса, предобработка текста (очистка, нормализация), векторизация запроса той же моделью эмбеддингов, что и для документов, выполнение SQL-запроса к базе данных с поиском ближайших соседей и фильтрацией по метаданным, ранжирование результатов по косинусному сходству, возврат наиболее релевантных результатов с метаданными [6].

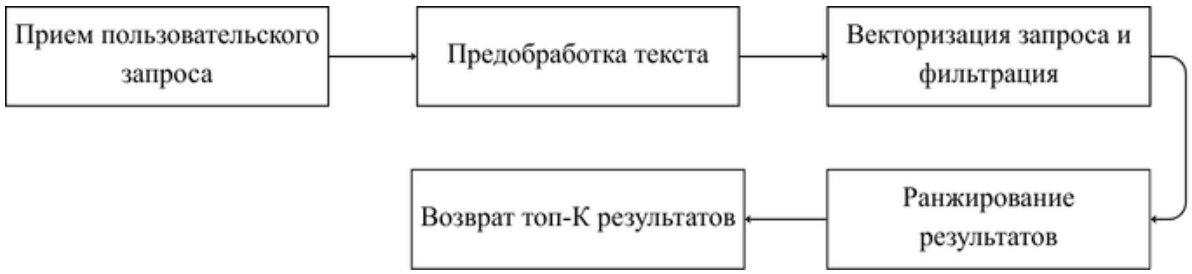


Рисунок 2. Алгоритм семантического поиска

Для повышения производительности реализовано кеширование частых запросов с использованием хеша запроса в качестве ключа.

Поддерживается гибридный поиск с возможностью комбинации семантического и полнотекстового поиска через встроенные средства PostgreSQL (tsvector/tsquery), что позволяет дополнительно повысить точность для запросов с точными техническими терминами.

Для обеспечения актуальности базы данных разработан процесс инкрементального обновления: новые документы проходят полный цикл обработки и добавляются в базу без необходимости переиндексации существующих данных [7].

Версионирование стандартов и спецификаций поддерживается через метаданные с указанием версии документа и даты публикации. Система предоставляет API для интеграции с внешними образовательными платформами и веб-интерфейсом, что обеспечивает возможность использования как в учебном процессе ВУЗов, так и в корпоративных системах обучения инженеров по телекоммуникациям.

Заключение. В работе рассмотрена задача организации семантического поиска технической и учебной документации при подготовке специалистов в области телекоммуникаций.

Предложенный подход на основе векторных баз данных обеспечивает семантический поиск, который находит материалы по смыслу, а не по лексическому совпадению, что устраняет проблемы синонимии терминов и межъязыкового барьера.

Использование расширения pgvector для PostgreSQL позволяет интегрировать векторный поиск в существующую инфраструктуру образовательных учреждений без развертывания дополнительных специализированных СУБД, сохраняя при этом возможность сложных запросов с фильтрацией по метаданным.

Рассмотренная архитектура векторизации технической и учебной документации позволит сократить время поиска учебной информации, повысить эффективность самостоятельной работы обучающихся может быть применена при создании образовательных ресурсов для других инженерных направлений со схожей спецификой предметной области.

Список литературы

[1] Ericsson Mobility Report: November 2025 / Ericsson. – Stockholm: Ericsson AB, 2025. – URL: https://www.ericsson.com/en/reports-and-papers/mobility-report (дата обращения: 01.03.2026).

[2] Маннинг, К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – М.: Вильямс, 2019. – 528 с.

[3] Платонов, А. В. Машинное обучение / А. В. Платонов. – М.: Юрайт, 2021. – 267 с.

[4] Литвиненко, В. И. Обработка естественного языка / В. И. Литвиненко. – М.: Юрайт, 2024. – 245 с.

[5] Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus [et al.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – P. 9459–9474.

[6] Моргунов, Е. П. PostgreSQL. Основы языка SQL / Е. П. Моргунов. – СПб.: БХВ-Петербург, 2019. – 304 с.

[7] Кузнецов, С. Д. Основы баз данных / С. Д. Кузнецов. – М.: Интуит, 2022. – 488 с.

Авторский вклад

Бардашевич Александра Валерьевна – разработка алгоритма семантического поиска с использованием индекса HNSW, проектирование процесса предобработки документов и нормализации телекоммуникационной терминологии, обеспечение межъязыковой поддержки поиска.

Савицкий Алексей Юрьевич – постановка задачи исследования, разработка архитектуры системы семантического поиска, обоснование выбора расширения pgvector для PostgreSQL, подготовка иллюстрационного материала.

Федоренко Владимир Александрович – разработка методологии векторизации технической документации, обоснование параметров чанкинга и выбора модели эмбеддингов, проектирование структуры метаданных, анализ существующих решений векторного поиска

APPLICATION OF VECTOR DATABASES FOR STORAGE AND SEARCH OF EDUCATIONAL MATERIALS IN THE FIELD OF TELECOMMUNICATIONS

A.V. Bardashevich
Cadet, BSUIR

A.Yu. Savitsky
Senior Lecturer of the Department of Communications, BSUIR, Ph.D. in Military Sciences

V.A. Fedorenko
Head of the Department of Communications, BSUIR

Abstract. This article examines the problem of inefficient search for educational information in the training of telecommunications specialists, caused by the semantic complexity of technical terminology and the interlingual barrier. An approach to organizing semantic search based on vector databases is proposed, ensuring the search for materials by meaning. The feasibility of using a vector extension for a relational database management system in educational institutions is substantiated. It is shown that the proposed solution enables the integration of vector search into the existing infrastructure, providing support for cross-lingual queries and metadata filtering.

Keywords: vector databases, semantic search, pgvector, PostgreSQL, embeddings, HNSW

УДК 004.89:004.41

ВЛИЯНИЕ МЕТОДОВ ДООБУЧЕНИЯ НА АДАПТАЦИЮ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ АНАЛИЗА ПРОЦЕССОВ РАЗРАБОТКИ



П.А. Красёв
Студент факультета компьютерных систем и сетей БГУИР
paulkrasev@gmail.com



М.А. Калугина
Доцент кафедры информатики, кандидат физико-математических наук, доцент
marina_kalugina@list.ru

П.А. Красёв

В 2022 году окончил Минское суворовское военное училище. Интересы направлены на глубокое обучение, дообучение больших языковых моделей, а также применение генеративных моделей для анализа данных и автоматизации бизнес-процессов.

М.А. Калугина

Окончила Белорусский государственный университет. Область научных интересов связана с исследованием проблем метрической теории диофантовых приближений зависимых величин и приложений математических методов к нейросетевому анализу

Аннотация. В статье рассматриваются методы адаптации больших языковых моделей для автоматизации анализа процессов разработки программного обеспечения. Акцент сделан на интеграции моделей с данными систем управления проектами. Проведено сравнение различных подходов к дообучению: полный fine-tuning, параметрически эффективные методы (LoRA, QLoRA, P-Tuning, IA3), а также методы оптимизации на основе предпочтений (DPO). Оценка качества выполнена по двум группам показателей: стандартные метрики оценки связности и семантической близости текста (Perplexity, BLEU, ROUGE-L, BERTScore) и предметные метрики – релевантность формируемых аналитических заключений и практическая реализуемость предлагаемых шагов по улучшению процессов. Установлено, что наибольшую эффективность демонстрирует комбинированное применение QLoRA и DPO, обеспечивающее высокое качество генераций при существенно меньших

вычислительных затратах по сравнению с полным fine-tuning. Показана необходимость предварительной стратификации данных по типам задач для корректного анализа потоковых метрик.

Ключевые слова: автоматизация анализа процесса разработки программного обеспечения, языковые модели, дообучение, fine-tuning, LoRA, P-Tuning, анализ процессов, метрики потока, BERTScore, ROUGE, релевантность.

Введение. Большие языковые модели открывают новые возможности для автоматизации анализа процессов разработки программного обеспечения, позволяя извлекать знания из накапливаемых данных и формировать рекомендации по его улучшению [1]. Однако применение универсальных моделей в этой предметной области затруднено из-за специфики профессиональной лексики и необходимости интерпретации количественных показателей выполнения работ.

Целью исследования являлся сравнительный анализ методов дообучения языковых моделей для генерации аналитических заключений и практических рекомендаций на основе данных, характеризующих ход разработки [2]. Поэтому ключевой особенностью работы является ориентация на комплексные данные из систем управления проектами: временные ряды метрик потока в сочетании с основными атрибутами задач. Это обусловливает следующие ограничения:

1 Семантический разрыв между текстами общего назначения и предметной областью управления разработкой.

2 Ограниченный объем эталонных аналитических выводов, требующий эффективных методов дообучения при малом количестве размеченных данных.

3 Необходимость совместной обработки числовых временных рядов и генерации связных текстовых заключений.

В работе сравниваются полный fine-tuning, параметрически эффективные методы (LoRA, QLoRA, P-Tuning, IA3) и оптимизация на основе предпочтений (DPO) [3]. Оценка проводится по стандартным NLP-метрикам (Perplexity, BLEU, ROUGE-L, BERTScore) и специальным предметным показателям релевантности выводов и реализуемости рекомендаций.

Описание алгоритма дообучения. В качестве источника информации о задачах разработки используются открытые датасеты, содержащие данные из различных систем управления проектами: текстовые описания, временные метки изменения задач, а также информация о типах задач. Такой подход обеспечивает воспроизводимость экспериментов и исключает зависимость от конкретных коммерческих систем управления проектами.

На основе временных меток для каждой задачи вычисляются различные метрики потока, характеризующие эффективность процесса работы. Вычисление этих метрик выполняется специализированным модулем предобработки, который также осуществляет нормализацию числовых значений и их преобразование в текстовый формат, пригодный для подачи в языковую модель.

Каждый пример включает входной контекст и целевой вывод – аналитическое заключение с рекомендациями. Для балансировки выборки применяется стратификация по типам задач.

Процесс дообучения выполняется на GPU: модель последовательно обучается на примерах с использованием различных методов (полный fine-tuning, LoRA, QLoRA, P-Tuning, IA3, DPO). Контроль качества осуществляется на валидационной выборке по стандартным NLP-метрикам (Perplexity, BLEU, ROUGE-L, BERTScore) и предметным показателям релевантности и реализуемости [4].

Для предотвращения переобучения при малом объеме размеченных данных применяются регуляризация, ранняя остановка на основе метрик валидационной выборки, а также смешанное семплирование, обеспечивающее равномерное представление различных категорий данных в каждом батче.

Влияние выбранного метода на результаты дообучения. В ходе исследования был проведен сравнительный анализ различных методов дообучения языковых моделей на задаче генерации аналитических заключений и рекомендаций по улучшению процессов разработки. Оценивалось их поведение в условиях ограниченного объема размеченных данных и необходимости интеграции числовых метрик потока с текстовым контекстом.

Первым был использован подход полного fine-tuning (full fine-tuning) на базовой модели. Начальные эксперименты проводились на упрощенном подмножестве данных: использовались только текстовые описания задач без временных метрик. Модель обучалась предсказывать аналитическое заключение по заданному набору заголовков задач за интервал. В этой упрощенной постановке full fine-tuning продемонстрировал приемлемое качество: значения ROUGE-L достигали 0.35–0.40, а оценки релевантности были положительными для простых случаев (например, когда требовалось лишь обобщить типы задач).

Успех в базовом текстовом сценарии мотивировал переход к полной постановке задачи – включению метрик потока. Для этого потребовалось модифицировать входной формат: числовые показатели нормализовались и преобразовывались в текстовые описания, которые конкатенировались с текстовым контекстом. Эксперименты показали, что full fine-tuning на таких данных приводит к быстрому переобучению: уже после 3–4 эпох метрики на валидационной выборке начинали ухудшаться, а сгенерированные рекомендации становились шаблонными и не учитывали специфику числовых показателей. Для борьбы с этим применялись стандартные методы регуляризации (dropout 0.1, weight decay), а также ранняя остановка. Однако даже при этом модель демонстрировала высокую чувствительность к начальным значениям весов и требовала тщательного подбора гиперпараметров (скорости обучения, размера батча).

Поскольку полный fine-tuning оказался ресурсоемким и склонным к переобучению при малом объеме данных, следующим шагом стало исследование параметрически эффективных методов (PEFT) [5]. Был использован метод LoRA с ранговой размерностью r=8 и α=16. LoRA позволила дообучать модель, замораживая основные веса и добавляя обучаемые адаптеры [6]. Это дало два ключевых преимущества: во-первых, объем обучаемых параметров сократился с сотен миллионов до нескольких миллионов, что ускорило эксперименты; во-вторых, метод оказался более устойчивым к переобучению. На валидационной выборке LoRA достигла сопоставимых с full fine-tuning значений ROUGE-L (0.38–0.42), но при этом кривые обучения были глаже, а разброс между запусками меньше (рисунок 1).

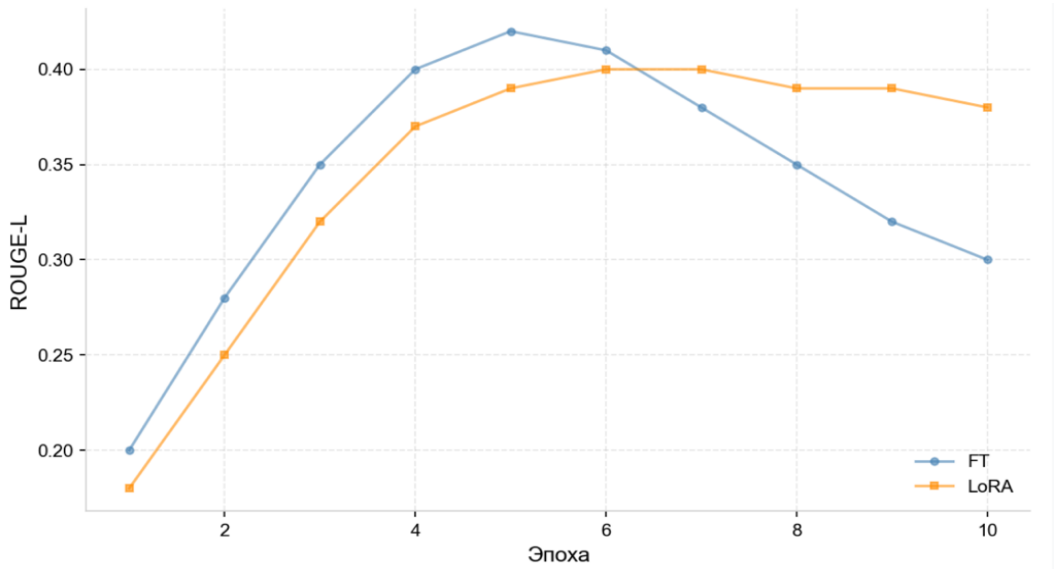


Рисунок 1. Сравнение full fine-tuning и LoRA

Для дальнейшего снижения потребления памяти был применен метод QLoRA, использующий 4-битное квантование базовой модели. QLoRA позволила дообучать модель на GPU с 8 ГБ памяти, тогда как full fine-tuning требовал не менее 24 ГБ [7]. Качество при этом практически не ухудшилось: метрики BERTScore остались на уровне 0.86–0.88 против 0.87–0.89 у LoRA. Однако при работе с QLoRA потребовалась дополнительная настройка гиперпараметров оптимизатора (использовался paged AdamW) для предотвращения нестабильности градиентов.

Параллельно тестировались методы на основе мягких промптов: P-Tuning и IA3 (рисунок 2) [8]. P-Tuning, обучающий непрерывные векторы-промпты, показал более низкую эффективность в рассматриваемой задаче: ROUGE-L не превышал 0.30, а генерируемые тексты часто были несвязными. Вероятно, это связано с тем, что мягкие промпты хуже справляются с интеграцией числовой информации, требующей точного отражения в тексте. IA3, напротив, продемонстрировал результаты, близкие к LoRA, но с еще меньшим числом обучаемых параметров. Однако при малом объеме данных IA3 иногда приводил к нестабильности, выражавшейся в резких скачках функции потерь [9].

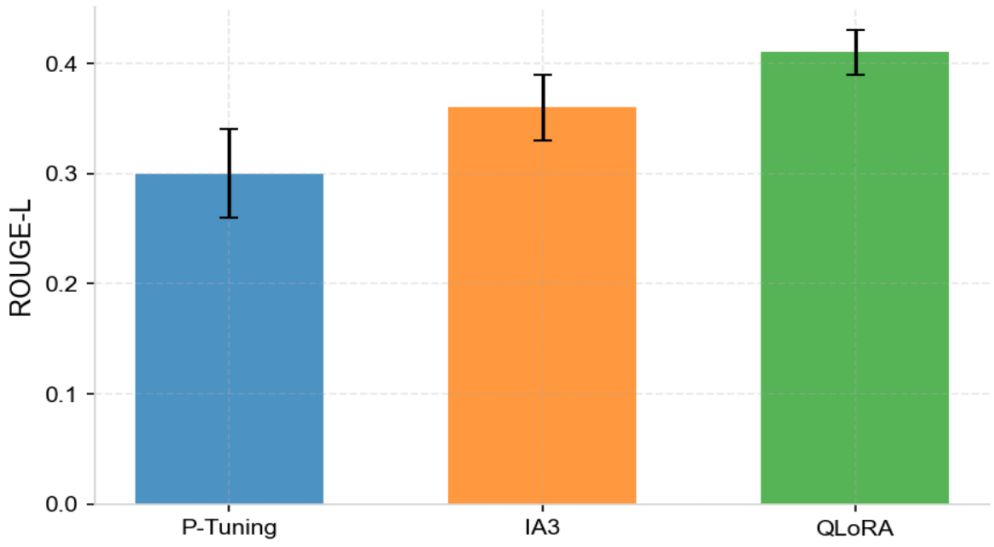


Рисунок 2. Сравнение PEFT методов

После того как модели, дообученные с помощью LoRA и QLoRA, научились генерировать связные заключения, возникла проблема: рекомендации часто были формально правильными, но не учитывали тонкие нюансы, важные с практической точки зрения (например, предлагали увеличить ресурсы, хотя реальной причиной задержек была блокировка зависимостями). Для решения этой проблемы был применен метод DPO [10]. На основе ответов модели были сформированы пары предпочтений: для каждого входного контекста выбирался лучший (релевантный и реализуемый) и худший (формальный или нереалистичный) ответы из числа сгенерированных моделью.

Результаты применения DPO оказались наиболее значимыми. Предметная метрика релевантности выросла с 3.2 до 4.1, а индекс практической реализуемости – с 2.8 до 3.9. При этом стандартные NLP-метрики изменились незначительно (ROUGE-L вырос на 0.02), что подтверждает, что DPO улучшает именно содержательное качество, а не просто текстовое сходство.

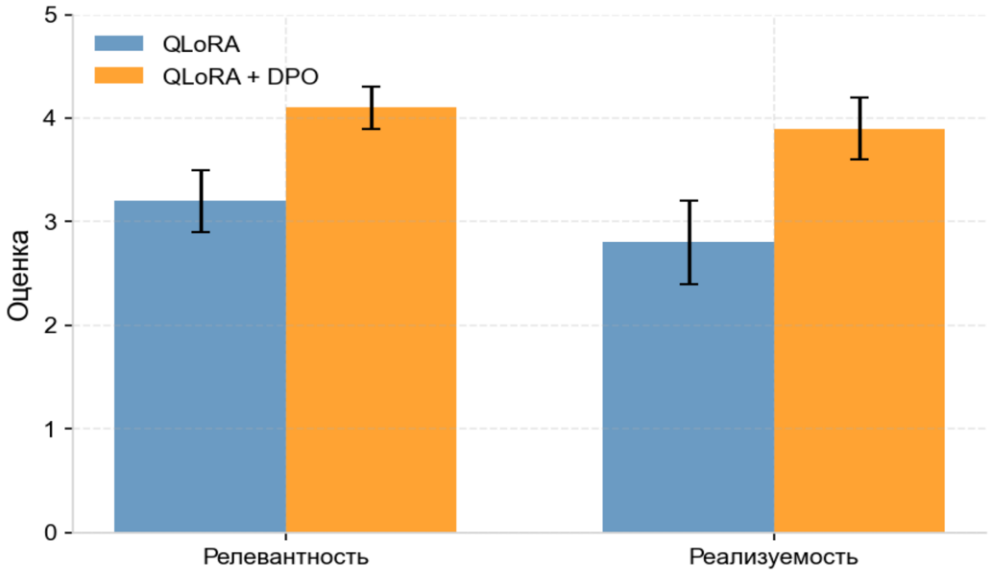


Рисунок 3. Результаты применения DPO

Комбинированное применение QLoRA и DPO позволило достичь наилучшего баланса между вычислительной эффективностью и качеством. QLoRA обеспечила экономию памяти и устойчивость к переобучению, а DPO скорректировала модель в соответствии с сформированными предпочтениями, повысив практическую ценность.

Заключение. В ходе исследования сформирована и экспериментально апробирована методика сравнительного анализа известных методов дообучения языковых моделей применительно к задаче автоматизации анализа процессов разработки программного обеспечения. Реализованный подход включает формирование обучающих примеров на основе открытых датасетов, вычисление метрик потока и их интеграцию с текстовым контекстом, а также применение современных методов адаптации моделей к предметной области. Проведённый сравнительный анализ методов дообучения подтвердил, что выбор алгоритма адаптации определяющим образом влияет на качество генерируемых аналитических заключений и рекомендаций в условиях ограниченного объёма размеченных данных:

1 Полный fine-tuning продемонстрировал быстрое достижение высоких значений на обучающей выборке, однако уже после 4–5 эпох начиналось переобучение, выражавшееся в падении метрик на валидации и шаблонности формируемых рекомендаций. Кроме того, метод предъявляет высокие требования к вычислительным ресурсам, что затрудняет его многократное применение.

2 Параметрически эффективные методы (PEFT) показали существенно более устойчивые результаты. LoRA и QLoRA обеспечили стабильный рост метрик без признаков переобучения. Метод P-Tuning оказался малоэффективным для задач, требующих точной интерпретации числовых показателей, а генерируемые тексты часто теряли связность. IA3 показал результаты, близкие к LoRA, однако при малом объёме данных наблюдались эпизодические выбросы функции потерь, что снижает его надёжность.

3 Оптимизация на основе предпочтений (DPO) позволила качественно улучшить содержательную сторону генераций. Комбинированное применение QLoRA и DPO привело к росту предметных метрик: релевантности аналитических выводов и индекса практической реализуемости рекомендаций. При этом стандартные NLP-метрики изменились незначительно, что подтверждает способность DPO корректировать модель в соответствии с заданными предпочтениями без ухудшения формального качества текста.

Таким образом, проведённое сравнение показало, что для создания эффективных аналитических систем на основе языковых моделей в предметной области управления разработкой недостаточно применения базовых методов дообучения.

Полный fine-tuning оказался неприменим из-за склонности к переобучению и высоких ресурсных затрат. Среди PEFT-методов наилучшее сочетание стабильности, качества и вычислительной эффективности продемонстрировали LoRA и QLoRA, при этом QLoRA предпочтительнее при ограниченных ресурсах.

Дополнительное применение DPO позволило значительно повысить практическую значимость генерируемых рекомендаций, что является ключевым для задач аналитической поддержки.

Наиболее эффективной признана комбинация QLoRA и DPO, обеспечивающая высокое качество как по стандартным NLP-метрикам, так и по предметным показателям релевантности и реализуемости.

В дальнейшей работе планируется расширение набора анализируемых метрик (включение показателей качества кода, данных из систем непрерывной интеграции), исследование возможности генерации не только текстовых заключений, но и визуальных аналитических материалов, а также применение методов мультиагентного подхода для моделирования взаимодействия различных ролей в процессе разработки.

Список литературы

[1] Hou X., Zhao Y., Liu Y., и др. Large Language Models for Software Engineering: A Systematic Literature Review. ACM Transactions on Software Engineering and Methodology. 2024;33(8):1-79. DOI: 10.1145/3695988.

[2] Vaswani A., Shazeer N., Parmar N., и др. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS); 2017;30:5998-6008.

[3] Devlin J., Chang M.-W., Lee K., и др. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL); 2019:4171-4186. DOI: 10.18653/v1/N19-1423.

[4] Touvron H., Lavril T., Izcard G., и др. LLaMA: Open and Efficient Foundation Language Models. arXiv; 2023. DOI: 10.48550/arXiv.2302.13971.

[5] Ding N., Qin Y., Yang G., и др. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. Nature Machine Intelligence. 2023;5:220-235. DOI: 10.1038/s42256-023-00626-4.

[6] Hu E.J., Shen Y., Wallis P., и др. LoRA: Low-Rank Adaptation of Large Language Models. International Conference on Learning Representations (ICLR); 2022. DOI: 10.48550/arXiv.2106.09685.

[7] Dettmers T., Pagnoni A., Holtzman A., и др. QLoRA: Efficient Finetuning of Quantized Large Language Models. Advances in Neural Information Processing Systems (NeurIPS); 2023;36. DOI: 10.48550/arXiv.2305.14314.

[8] Liu X., Ji K., Fu Y., и др. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL); 2022:61-68. DOI: 10.18653/v1/2022.acl-short.8.

[9] Liu H., Tam D., Muqeeth M., и др. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. Advances in Neural Information Processing Systems (NeurIPS); 2022;35:1950-1965.

[10] Rafailov R., Sharma A., Mitchell E., и др. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Advances in Neural Information Processing Systems (NeurIPS); 2023;36. DOI: 10.48550/arXiv.2305.18290.

Авторский вклад

Красёв Павел Александрович – выбор задачи исследования, реализация архитектуры экспериментального стенда с модулем предобработки данных, проведение экспериментов со сбором стандартных NLP-метрик и предметных показателей, анализ полученных результатов.

Калугина Марина Алексеевна – постановка проблемы, научное руководство исследованием методов адаптации языковых моделей для анализа процессов разработки, консультации по математическому аппарату алгоритмов.

THE IMPACT OF FINE-TUNING METHODS ON THE ADAPTATION OF LANGUAGE MODELS FOR SOFTWARE DEVELOPMENT PROCESS ANALYSIS

P.A. Krasyov
Student of BSUIR

M.A. Kalugina
Associate Professor of Informatics Department of the BSUIR

Abstract. This article examines methods for adapting large language models to automate analytical support in software development processes. The focus is on integrating models with project management system data. Various fine-tuning approaches are compared: full fine-tuning, parameter-efficient methods (LoRA, QLoRA, P-Tuning, IA3), and preference optimization methods (DPO). Quality assessment is performed using two groups of metrics: standard metrics for evaluating text coherence and semantic similarity (Perplexity, BLEU, ROUGE-L, BERTScore) and task-specific metrics–the relevance of generated analytical conclusions and the practical feasibility of proposed process improvement steps. The combined application of QLoRA and DPO demonstrates the highest effectiveness, providing high-quality generations with substantially lower computational costs compared to full fine-tuning. The necessity of preliminary data stratification by task types for correct analysis of flow metrics is shown.

Keywords: software development analysis automation, language models, fine-tuning, LoRA, P-Tuning, process analysis, flow metrics, BERTScore, ROUGE, relevance.

УДК 004.8:004.056

ОБНАРУЖЕНИЕ АНОМАЛИЙ НА ОСНОВЕ BIG DATA ДЛЯ ВЫЯВЛЕНИЯ МОШЕННИЧЕСТВА В ЦИФРОВЫХ СИСТЕМАХ



М.М. Ходжамаммедов
преподаватель кафедры информационных технологий Государственного энергетического института Туркменистана
mekanhoja2021@gmail.com

М.М. Ходжамаммедов
Окончил Государственного энергетического института Туркменистана. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. В данной работе рассматриваются теоретические и прикладные аспекты обнаружения аномалий в условиях обработки больших данных с целью выявления мошеннической активности в цифровых системах. Актуальность темы обусловлена экспоненциальным ростом объемов данных и усложнением методов совершения мошенничества. В статье анализируются современные подходы, основанные на статистических методах, алгоритмах машинного обучения и глубоких нейронных сетях. Особое внимание уделяется роли распределенных вычислительных систем и потоковой обработки данных. Представлен сравнительный анализ методов и их эффективности в задачах обнаружения мошенничества.

Ключевые слова: Big Data, обнаружение аномалий, мошенничество, машинное обучение, нейронные сети, анализ данных, кибербезопасность.

Введение. Цифровая трансформация современного общества сопровождается стремительным ростом объемов генерируемых данных, что приводит к формированию принципиально новых подходов к их обработке и анализу. В финансовых, телекоммуникационных и информационных системах ежедневно обрабатываются миллионы транзакций, каждая из которых может потенциально содержать признаки мошенничества. В таких условиях традиционные методы анализа данных оказываются недостаточно эффективными из-за ограниченной масштабируемости и неспособности выявлять сложные зависимости. Мошенничество в цифровой среде принимает все более сложные и скрытые формы, включая использование автоматизированных скриптов, социальной инженерии и распределенных атак. Это требует применения интеллектуальных методов анализа, способных адаптироваться к изменяющимся условиям и выявлять аномалии в режиме реального времени. Обнаружение аномалий является ключевым направлением в области анализа данных, направленным на выявление отклонений от нормального поведения системы.

Теоретические основы обнаружения аномалий. Понятие аномалии в анализе данных определяется как наблюдение, существенно отличающееся от большинства других наблюдений. Такие отклонения могут свидетельствовать о наличии ошибок, редких событий или целенаправленных мошеннических действий. Формализация задачи обнаружения аномалий связана с построением модели нормального поведения и последующим выявлением отклонений от этой модели. Существует несколько подходов к формализации данной задачи. Вероятностный подход основывается на предположении о распределении данных и позволяет вычислять вероятность принадлежности наблюдения к нормальному классу. Геометрический подход рассматривает данные в пространстве признаков и определяет аномалии как точки, находящиеся на значительном расстоянии от плотных областей. Информационный подход использует меры энтропии и сложности для выявления нестандартных паттернов. Особенностью задач обнаружения мошенничества является высокая степень несбалансированности данных, при которой доля мошеннических операций составляет крайне малую часть общего объема. Это усложняет процесс обучения моделей и требует применения специальных методов обработки данных.

Методы машинного обучения для обнаружения аномалий. Современные методы машинного обучения играют ключевую роль в решении задач обнаружения аномалий. Их преимущество заключается в способности автоматически выявлять скрытые закономерности и адаптироваться к изменяющимся данным. Методы обучения с учителем предполагают наличие размеченных данных и позволяют строить классификационные модели. Однако в задачах обнаружения мошенничества такие данные часто ограничены или неполны. В связи с этим широкое распространение получили методы обучения без учителя, которые не требуют предварительной разметки. Алгоритмы кластеризации позволяют выявлять структуру данных и определять аномалии как объекты, не принадлежащие ни одному из кластеров. При этом эффективность данных методов зависит от выбора метрики расстояния и параметров модели. Особое внимание уделяется алгоритму Isolation Forest, который основан на принципе изоляции редких наблюдений. Данный метод демонстрирует высокую эффективность при работе с большими объемами данных и не требует значительных вычислительных ресурсов. Нейросетевые подходы, в частности автоэнкодеры, позволяют моделировать сложные нелинейные зависимости. Принцип их работы основан на реконструкции входных данных и анализе ошибки восстановления. Аномалии характеризуются высокой ошибкой реконструкции, что позволяет эффективно их выявлять.

Глубокое обучение и анализ временных рядов. С развитием технологий глубокого обучения появились новые возможности для анализа данных, представленных в виде временных рядов. В задачах обнаружения мошенничества важную роль играет учет временной динамики, поскольку поведение пользователей изменяется во времени. Рекуррентные нейронные сети и архитектуры типа LSTM позволяют моделировать последовательности и выявлять отклонения в поведении. Такие модели способны учитывать контекст и выявлять

сложные паттерны, недоступные для традиционных методов. Применение сверточных нейронных сетей также находит применение в анализе структурированных данных, где они используются для выявления локальных закономерностей.

Инфраструктура Big Data. Эффективная реализация методов обнаружения аномалий невозможна без использования современных технологий Big Data. Распределенные вычислительные системы обеспечивают обработку данных в условиях высокой нагрузки и позволяют масштабировать алгоритмы. Платформы Hadoop и Apache Spark являются основой для построения систем анализа данных. Они обеспечивают распределенное хранение и обработку информации, что позволяет эффективно работать с большими объемами данных. Особое значение имеет потоковая обработка данных, реализуемая с использованием технологий Kafka и Spark Streaming. Это позволяет анализировать данные в режиме реального времени и оперативно реагировать на выявленные аномалии.

Практическое применение в финансовых системах. В финансовых организациях системы обнаружения мошенничества интегрируются в процесс обработки транзакций и функционируют в режиме реального времени. Они анализируют множество параметров, включая географическое положение пользователя, частоту операций, тип устройства и поведенческие характеристики. Современные системы используют гибридные подходы, объединяющие несколько методов анализа данных. Это позволяет повысить точность и снизить количество ложных срабатываний. Важным аспектом является баланс между безопасностью и удобством пользователей.

Проблемы и ограничения. Несмотря на значительные достижения, системы обнаружения аномалий сталкиваются с рядом проблем. К ним относятся высокая сложность моделей, необходимость больших вычислительных ресурсов и ограниченность размеченных данных. Кроме того, мошенники постоянно адаптируют свои методы, что требует регулярного обновления моделей.

Проблема интерпретируемости моделей также является актуальной, особенно в финансовой сфере, где требуется объяснение принятых решений.

Перспективы развития. Дальнейшее развитие технологий обнаружения аномалий связано с интеграцией методов искусственного интеллекта, развитием объяснимых моделей и использованием гибридных архитектур. Особое внимание уделяется автоматизации процессов обучения и адаптации моделей. Перспективным направлением является использование графовых моделей для анализа связей между объектами, а также внедрение федеративного обучения для обеспечения конфиденциальности данных.

Заключение. В результате проведенного анализа можно сделать вывод о том, что технологии Big Data в сочетании с методами машинного и глубокого обучения являются эффективным инструментом для выявления мошенничества.

Их применение позволяет существенно повысить уровень безопасности цифровых систем и снизить риски финансовых потерь.

Дальнейшие исследования должны быть направлены на повышение точности моделей, снижение вычислительных затрат и улучшение интерпретируемости результатов.

Список литературы

[1] Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. 2009. Aggarwal C.C. Outlier Analysis. Springer, 2017.

[2] Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // ICDM. 2008.

[3] Breunig M.M. et al. LOF: Identifying Density-Based Local Outliers // SIGMOD. 2000.Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.

Авторский вклад

Мекан Ходжамамедов Мерданович – постановка задач научных исследований, разработка методологии оценки эффективности генеративной политики при оптимизации проектных решений, а также руководство проведением этих исследований.

ANOMALY DETECTION BASED ON BIG DATA FOR IDENTIFYING FRAUD IN DIGITAL SYSTEMS

M.M. Hojamammedov
Lecturer, Department of Information Technologies, Turkmen State Energy Institute
mekanhoja2021@gmail.com

Abstract. This work examines the theoretical and applied aspects of anomaly detection in the context of big data processing for identifying fraudulent activity in digital systems. The relevance of the topic is driven by the exponential growth of data volumes and the increasing complexity of fraud techniques. The article analyzes modern approaches based on statistical methods, machine learning algorithms, and deep neural networks. Special attention is paid to the role of distributed computing systems and stream data processing. A comparative analysis of methods and their effectiveness in fraud detection tasks is presented.

Keywords: Big Data, anomaly detection, fraud, machine learning, neural networks, data analysis, cybersecurity.

УДК 004.054 : 004.42-027.45

ПРАКТИКА ОЦЕНКИ ОЖИДАЕМОЙ НАДЁЖНОСТИ ПРИКЛАДНЫХ КОМПЬЮТЕРНЫХ ПРОГРАММ ДЛЯ ИНФОРМАЦИОННЫХ СИСТЕМ



Ван Там Лэ
Научный сотрудник, Институт технологий, Ханой, Вьетнам, Магистр техники и технологии
luct.tle94@gmail.com



С.М. Боровиков
Доцент кафедры проектирования информационно-компьютерных систем БГУИР, кандидат технических наук, доцент
bsm@bsuir.by



А.В. Будник
Декан факультета инжиниринга и технологий БГАС, кандидат технических наук, доцент
A.Budnik@bsac.by



С.К. Дик
Заведующий кафедрой инженерной и компьютерной графики БГУИР, кандидат физ.-мат. наук, доцент
sdick@bsuir.by

Ван Там Лэ

Окончил Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов: информационные технологии; надёжность прикладных компьютерных программ; исследование, разработка и технология изготовления интегральных схем; исследования и разработка технологии изготовления магнитомягких материалов и оптического стекла.

С.М. Боровиков

Окончил Минский радиотехнический институт. Основная область научных интересов – прикладные математические методы в проектировании изделий радиоэлектроники, включая алгоритмы статистического прогнозирования надёжности изделий электронной техники и оценку надёжности прикладного программного обеспечения на ранних этапах его разработки.

А.В. Будник

Окончил Белорусский государственный университет информатики и радиоэлектроники. Основные направления научной деятельности – микроэлектроника, защита информации в технических системах.

С.К. Дик

Окончил Минский радиотехнический институт по специальности «Радиотехника», руководит научными исследованиями в области лазерной медицины и биомедицинской оптики.

Аннотация. Рассмотрена практика оценки эксплуатационной надёжности планируемых к разработке прикладных компьютерных программ, предназначенных для информационных систем. На конкретном примере показана оценка ожидаемой надёжности компьютерной программы, используемой для моделирования и исследования проектных решений в учебном процессе студентов технического вуза.

Ключевые слова: информационные системы, прикладные компьютерные программы, ожидаемая надёжность, тестирование.

Введение. Эффективное использование прикладных компьютерных программ во многом определяется их эксплуатационной надёжностью. Проблема обеспечения надёжности разрабатываемых компьютерных программ является острой [1–3], поскольку вклад программного обеспечения в ненадёжность современных сложных компьютерных информационных систем может составлять до 40 и более процентов. Недостаточный уровень надёжности и обусловленные этим частые сбои в использовании прикладных компьютерных программ раздражают пользователей, что вынуждает их отказываться от использования таких программ и предпринимать шаги по замене программ более надёжными версиями или делать попытки по поиску совершенно новых программ у заслуживающих доверия разработчиков и IT-организаций.

Разработчикам программ, а также менеджерам IT-организаций очень важно получить информацию об ожидаемом уровне эксплуатационной надёжности планируемой к разработке будущей компьютерной программы ещё до написания её кода на языке программирования. Это позволит сориентироваться в целесообразности дальнейшего вкладывания и расходования денежных средств на разработку компьютерной программы с учётом бюджета времени и количества программистов, используемых для написания кода программы, а также определить требуемое календарное рабочее время и количество специалистов, которых следует задействовать для тестирования, испытания и отладки компьютерной программы.

В данной работе на конкретном примере показано, как оценить ожидаемый уровень эксплуатационной надёжности компьютерной программы, планируемой к разработке, и определить, какое примерно рабочее календарное время потребуется для её тестирования с целью обеспечения заданного уровня эксплуатационной надёжности.

Основные принципы и подходы к оценке ожидаемой надёжности планируемых к разработке прикладных компьютерных программ. Эти принципы и подходы обсуждались в работах [1, 4–11] и в систематизированном виде описаны ниже:

1. Вклад программного обеспечения в ненадёжность сложных компьютерных информационных систем в ряде случаев может превышать вклад, вносимый техническими средствами (компьютерами), поскольку используемые программой входные данные могут быть сложными и их формат всё время меняется [3].

2. Единственно приемлемой количественной характеристикой измерения объёма (размера) компьютерной программы является число исполняемых строк программного кода (в англоязычном варианте LOC – Lines Of Code) [3]. Во многих случаях используют размерность KLOC – тысяч строк кода. Размер (объём) прикладных компьютерных программ современных информационных систем может достигать миллионов строк программного кода и оцениваться 1…100 млн и более команд.

3. В компьютерных программах большого объёма (тысячи–десятки тысяч и более строк кода) после тестирования в течение отводимого бюджетного времени всегда остаются скрытые ошибки, которые могут себя иногда проявлять в зависимости от характера набора исходных данных и нагрузки на компьютерную программу со стороны эксплуатационной среды.

4. В прикладных компьютерных программах, которые протестированы только на предмет обеспечения функциональных возможностей, присутствует примерно от 5 до 30…50 ошибок на KLOC (1000 выполняемых строк программного кода) [3].

Для ответственного программного обеспечения (ПО), к которому можно отнести операционные системы и другое системное ПО, к моменту поставки программы клиенту в нём может содержаться до 0,04...0,15 ошибок на 1000 строк кода программы.

5. После написания кода компьютерной программы и устранения в ней ошибок, вызванных нарушением правил языка программирования, она характеризуется начальной интенсивностью проявления ошибок (λ_0), иначе – интенсивностью отказов.

При типовых исходных данных обеспечивается функциональность компьютерной программы, но значение λ_0 обычно не отвечает требованиям пользователей. И кроме того, из-за наличия скрытых неявных ошибок в зависимости от изменчивости исходных данных и нагрузки на программу со стороны эксплуатационной среды (запись промежуточных данных на электронные носители, вывод данных на принтер, ожидание операций в очереди и т.д.) может не всегда обеспечиваться функциональность компьютерной программы.

6. Компьютерная программа после тестирования характеризуется эксплуатационной интенсивностью проявления ошибок (интенсивностью отказов) — $\lambda_{экс} < \lambda_0$.

Задача тестирования состоит в том, чтобы в рамках отводимого бюджетного времени выявить и устранить наиболее критичные ошибки с точки зрения функциональности программы, свести к минимуму долю скрытых ошибок, оставшихся в компьютерной программе, и в итоге обеспечить приемлемый уровень её эксплуатационной интенсивности отказов $\lambda_{экс}$.

7. Переход от значения λ_0 к значению $\lambda_{экс}$ выполняют с учётом коэффициента тестирования Q [1]:

$$\lambda_{экс} = \frac{\lambda_0}{Q}. \quad (1)$$

Достигаемое численное значение Q зависит от бюджетного времени $T_б$, отводимого на тестирование, и области (сферы) применения компьютерной программы [4]:

$$Q = \exp[60 \cdot 10^{-6} \cdot V \cdot K_\Sigma \cdot T_Б \cdot (r/100)], \quad (2)$$

где V – скорость выполнения команд компьютерной программы, размерность – $с^{-1}$; K_Σ – коэффициент увеличения интенсивности отказов из-за суммарного действия изменчивости входных данных и рабочей нагрузки на компьютерную программу со стороны эксплуатационной среды; $T_б$ – бюджетное календарное рабочее временя (в часах), отводимое на тестирование программы; r – средний процент времени выполнения (прогона) компьютерной программы при её тестировании в течение календарного рабочего дня для прикладной компьютерной программы рассматриваемой области применения.

8. Предлагаемая модель прогнозирования начальной интенсивности отказов компьютерной программы (λ_0) [1, 4, 10] учитывает следующие факторы:

– область (отрасль) применения компьютерной программы, что определяет изменчивость реального потока наборов исходных данных, вводимых в программу и рабочую нагрузку на компьютерную программу со стороны эксплуатационной среды;

– особенность организации, которая будет разрабатывать компьютерную программу, квалификация и опыт задействованных программистов;

– прогнозируемый размер (объём) компьютерной программы в KLOC;

– основные характеристики компьютерной программы: сложность, новизна, средства разработки, использование стандартных модулей;

– степень полноты учёта факторов, определяющих качество написания программного кода;

– быстродействие процессора в составе используемого компьютера.

$$\lambda_0 = 60 \cdot V \cdot K_\Sigma \cdot F_0 \cdot L \cdot 10^{-6} = 60 \cdot \left(\frac{R}{B}\right) \cdot K_\Sigma \cdot F_0 \cdot L \cdot 10^{-6} \text{ ч}^{-1}, \quad (3)$$

где F_0 – начальная плотность ошибок в компьютерной программе; L – прогнозируемое (в KLOC) число строк компьютерной программы.

Размерность параметров модели (3), квадратные скобки символизируют размерность: $[V]$ = с$^{-1}$, $[F_0]$ = ошибок/KLOC; $[L]$ = KLOC; $[R]$ = операций (команд) в секунду; $[B]$ = команд.

9. В качестве основы определения F_0 использована «Римская модель» RL-92-52 [13], включающая метрику D, учитывающую особенность организации, которая будет разрабатывать компьютерную программу, и метрику S, учитывающую основные характеристики компьютерной программы.

Для метрики S рекомендуется учесть такие её характеристики как сложность, новизна, средства разработки, степень использования стандартных модулей.

Авторами статьи предлагается саму модель дополнить метрикой U, учитывающей, не принятые во внимание, но гипотетически влияющие на F_0 факторы, связанные с организацией, разрабатывающей программу, программистами и характеристиками самой будущей компьютерной программы. Её (метрику U) будем называть метрикой полноты учёта факторов. Вид предлагаемой модели определения F_0

$$F_0 = F_б \cdot D \cdot S \ (2 - U) \text{ ошибок/KLOC}, \quad (4)$$

где $F_б$ – метрика области применения компьютерной программы, представляет собой среднюю по отрасли или базовую плотность отказов, размерность «ошибок/KLOC».

Метрика U безразмерная и должна определяться методом экспертной оценки из условия

$$0 < U \leq 1, \quad (5)$$

причём $U = 1$ – соответствует учёту практически всех факторов, определяющих значение F_0 (качество написания программного кода), а $U < 1$ – случаю, когда какая-то часть влияющих факторов не принята во внимание в метриках $F_б$, D и S.

10. Метрики D, S определяют как произведение коэффициентов K_i, каждый из которых учитывает конкретные факторы, влияющие на значения этих метрик [1, 4]:

$$D, S = \prod_i K_i \quad . \quad (6)$$

11. Эксплуатационная интенсивность отказов ($\lambda_{экс}$) характеризует надёжность компьютерной программы, но значение этой характеристики зависит от быстродействия процессора R, используемого в составе компьютера.

Согласно ГОСТ 27.205–97, в качестве основной характеристики безотказного выполнения программой своих функций рекомендуется использовать вероятность того, что прикладная компьютерная программа безотказно выполнит обработку одного произвольного усреднённого набора исходных данных из числа тех наборов, которые могут поступать в условиях использования компьютерной программы в составе информационной системы.

В работе [12], для расчёта этой вероятности (p_1) рекомендуется использовать экспоненциальную функцию:

$$p_1 = p_1(t_1) = \exp\left(-\lambda_{\text{экс}} \cdot t_1\right), \tag{7}$$

где $\lambda_{\text{экс}}$ – эксплуатационная интенсивность отказов в размерности 1/ч; t_1 – среднее процессорное время (в часах) обработки компьютерной программой одного усреднённого набора исходных данных; может быть найдено как

$$t_1 \approx \frac{B}{3600(0{,}7 \cdot R)}, \text{ ч}, \tag{8}$$

где B – прогнозная оценка числа команд, необходимых для обработки одного усреднённого набора исходных данных; R – пиковое быстродействие процессора компьютера, указанное в технической документации, размерность – «операций (команд) в секунду».

12. Вероятность того, что оставшиеся после тестирования скрытые ошибки в компьютерной программе не проявятся в течение заданного календарного рабочего времени τ, рассчитывают (с учётом ГОСТ 27.205–97) по формуле

$$P(\tau) = \left[p_1(t_1)\right]^{\eta \cdot \tau}, \tag{9}$$

где η – интенсивность обращения к компьютерной программе в течение периода τ; η представляет собой среднее число сеансов («прогонов») программы за каждый час в течение календарного рабочего периода τ.

Пример оценки надёжности компьютерной программы. Для иллюстрации метода расчёта ожидаемой надёжности планируемой к разработке прикладной компьютерной программы использованы следующие исходные данные:

– область применения компьютерной программы: моделирование проектных решений и их использование в учебном процессе студентов учебного заведения;

– прогнозный объём компьютерной программы L, определяется на основе анализа блок-схемы алгоритма работы компьютерной программы; в случае затруднений следует воспользоваться рекомендациями документа [14]; получено L = 40 KLOC;

– язык программирования для написания программного кода: Java;

– суммарное календарное рабочее время тестирования компьютерной программы $T_{\text{б}}$ = 2100 ч (5 студентов, для каждого T = 420 ч);

– средний процент времени прогона компьютерной программы (r) при её тестировании в течение календарного рабочего времени, r = 3 % [4, 8];

– заданный календарный период эксплуатации τ, для которого определяют проектный уровень эксплуатационной надёжности компьютерной программы, τ = 176 ч (1 месяц = 22 учебных дня по 8 часов);

– прогнозное число обращений в час к компьютерной программе (η) в течение заданного календарного периода эксплуатации τ = 176 ч, принято $\eta = 4$ ч$^{-1}$;

– указанное в технической документации пиковое быстродействие процессора, используемого в компьютере, $R = 5 \cdot 10^8$ операций в секунду.

Для выбора коэффициентов, принимаемых во внимание при определении метрик D (особенность организации, разрабатывающей компьютерную программу) и S (характеристики компьютерной программы) использованы работы [1, 11, 14].

Последовательность оценки ожидаемой надёжности планируемой к разработке компьютерной программы:

1. Уточняем метрику производственной среды разработки компьютерной программы (D):

$$D = K_{\text{орг}} \cdot K_{\text{квал}} = 1{,}3 \cdot 2{,}0 = 2{,}6, \tag{10}$$

где $K_{\text{орг}}$ – коэффициент, учитывающий особенность проектной организации; принято $K_{\text{орг}} = 1{,}3$ (высшее учебное заведение как организация, эксплуатирующая компьютерную информационную систему) [1, 4, 10, 13];

$K_{\text{квал}}$ – коэффициент, учитывающий квалификацию программистов, принято $K_{\text{квал}} = 2{,}0$ (программист – студент технического вуза, освоивший программирование в рамках соответствующей учебной дисциплины) [1, 4, 10].

2. Получаем коэффициенты, включаемые в метрику *S*, учитывающую характеристики планируемой к разработке компьютерной программы:

2.1. Рассчитываем, используя [4, 10, 14], коэффициент сложности компьютерной программы ($K_{\text{слож}}$), учитывающий наличие факторов, влияющих на трудоёмкость разработки программы:

$$K_{\text{слож}} = 1 + \sum_{i=1}^{m} k_i = 1 + (k_1 + k_2 + k_3) = 1 + (0{,}08 + 0{,}06 + 0{,}12) = 1{,}26, \tag{11}$$

где k_i – коэффициент повышения сложности прикладной компьютерной программы за счёт наличия у неё *i*-й характеристики; принято $m = 3$.

В соответствии с документом [14] выбраны коэффициенты повышения сложности ($m = 3$):

$k_1 = 0{,}8$ – связь с другими компьютерными программами;

$k_2 = 0{,}06$ – хранение и поиск данных в сложных структурах;

$k_3 = 0{,}12$ – наличие одновременно двух характеристик сложности из числа описанных в документе [14]: а) интеллектуальный языковый интерфейс с пользователем, б) моделирование объектов и процессов.

2.2. Находим значение коэффициента $K_{\text{нов}}$, учитывающего новизну компьютерной программы. Выбрано $K_{\text{нов}} = 0{,}81$ [4, 14], что соответствует компьютерным программам, являющимся развитием определённого функционального назначения компьютерных программ.

2.3. Выбираем значение коэффициента $K_{\text{ср.разр}}$, учитывающего используемые средства разработки компьютерной программы. Принято $K_{\text{ср.разр}} = 1{,}2$ – язык программирования высокого уровня, функционирование программы в локальной сети [1, 4, 14].

2.4. Определяем значение коэффициента $K_{\text{мод}}$, учитывающего процент использования стандартных модулей. Выбрано $K_{\text{мод}} = 0{,}77$, что соответствует степени охвата стандартными модулями реализуемых функций от 20 до 40 % [1, 4, 14].

3. Рассчитываем саму метрику *S* с учётом коэффициентов $K_{\text{слож}}$, $K_{\text{нов}}$, $K_{\text{ср.разр}}$ и $K_{\text{мод}}$:

$$S = K_{\text{слож}} \cdot K_{\text{нов}} \cdot K_{\text{ср.разр}} \cdot K_{\text{мод}} = 1{,}26 \cdot 0{,}81 \cdot 1{,}2 \cdot 0{,}77 \approx 0{,}943. \tag{12}$$

4. Выбираем значение метрики *U*, учитывающей степень полноты учёта факторов, влияющих на качество написания программного кода. Путём экспертной оценки получено $U = 0{,}8$.

5. Определяем начальную плотность ошибок в компьютерной программе (F_0), характеризующую ожидаемое количество скрытых дефектов в программном коде после устранения ошибок, вызванных нарушением правил языка программирования:

$$F_0 = F_{\text{Б}} \cdot D \cdot S \cdot (2 - U) = 12{,}3 \cdot 2{,}6 \cdot 0{,}943 \cdot (2 - 0{,}8) = 36{,}189 \text{ ошибок/KLOC}, \tag{13}$$

где $F_б$ = 12,3 ошибок/KLOC – базовая (усреднённая) плотность ошибок для прикладных компьютерных программ, предназначенных для моделирования и обучения [1, 4, 10].

6. Даём оценку прогнозному значению количества команд, выполняемых программой:

$$B = L \cdot E_L \cdot E_{\text{ц}} = 40000 \cdot 10 \cdot 20 = 8{,}0 \cdot 10^6 \text{ команд}, \quad (14)$$

где E_L – коэффициент расширения кода программы относительно числа строк кода L, определяется используемым языком программирования. Согласно [15], принято E_L = 10. $E_ц$ – коэффициент увеличения числа выполняемых процессором команд за счёт наличия в программе условных переходов, ветвлений и других особенностей (ввод-вывод данных, считывание информации с внешних носителей, ожидание в очереди, обработка прерываний).

Значение $E_ц$ следует выбирать на основе экспертной оценки с учётом алгоритма выполнения компьютерной программой своих функций, экстремального характера входных данных, числа ветвлений в программе, записи и считывания данных с внешних носителей и т.п. В случае неопределённости рекомендуется выбирать значение от 20 до 50. Принято $E_ц$ = 20.

7. Определяем скорость выполнения команд компьютерной программы (V) для пикового быстродействия R процессора, размерность R – «операций (команд) в секунду»:

$$V = \frac{R}{B} = \frac{5 \cdot 10^8}{8 \cdot 10^6} = 62{,}5 \text{ с}^{-1}. \quad (15)$$

8. Получаем, используя модель (3), начальную интенсивность отказов λ_0 (интенсивность проявления скрытых ошибок), т.е. интенсивность отказов до начала тестирования программы:

$$\lambda_0 = 60 \cdot K_\Sigma \cdot V \cdot F_0 \cdot L \cdot 10^{-6} = 60 \cdot 14{,}1 \cdot 62{,}5 \cdot 36{,}189 \cdot 40 \cdot 10^{-6} = 76{,}54 \text{ ч}^{-1}, \quad (16)$$

где $K_Σ$ – коэффициент увеличения интенсивности отказов из-за суммарного действия изменчивости входных данных и рабочей нагрузки на компьютерную программу со стороны эксплуатационной среды (ввод данных, использование принтера, ожидание операций в очереди и т. д); для компьютерных программ, используемых для моделирования и обучения для рассматриваемой области применения компьютерной программы, $K_Σ$ = 14,1 [1, 4, 10].

Размерности параметров: [V] = с$^{-1}$; [F_0] = ошибка/KLOC; [L] = KLOC.

9. Определяем по формуле (2) проектное значение коэффициента тестирования Q компьютерной программы в предположении, что тестирование программы выполняется в течение суммарного бюджетного календарного рабочего времени $T_Б$= 2100 ч:

$$Q = \exp[60 \cdot 10^{-6} \cdot K_\Sigma \cdot V \cdot T_Б \cdot (r/100)] = \exp[60 \cdot 10^{-6} \cdot 14{,}1 \cdot 62{,}5 \cdot 2100 \cdot (3/100)] = 27{,}97. \quad (17)$$

10. Находим эксплуатационную интенсивность отказов компьютерной программы ($\lambda_{экс}$):

$$\lambda_{\text{экс}} = \frac{\lambda_0}{Q} = \frac{76{,}54}{27{,}97} = 2{,}7365 \approx 2{,}74 \text{ ч}^{-1}. \quad (18)$$

11. По формуле (8) определяем среднее процессорное время (в часах) обработки компьютерной программой одного усреднённого набора исходных данных (значение R должно подставляться в размерности «операций (команд) в секунду»:

$$t_1 \approx \frac{B}{3600\,(0{,}7 \cdot R)} = \frac{8 \cdot 10^6}{3600 \cdot (0{,}7 \cdot 5 \cdot 10^8)} = 6{,}349 \cdot 10^{-6} \text{ ч.} \qquad (19)$$

12. Вероятность безотказной работы программы в течение календарного времени τ (приняты во внимание ГОСТ 27.205–97 и работы [4, 11]):

$$P(\tau) = (P_1)^{\eta \cdot \tau} = \left[\exp\left(-\lambda_{\text{экс}} \cdot t_1\right)\right]^{\eta \cdot \tau} = [\exp(-2{,}74 \cdot 6{,}349 \cdot 10^{-6})]^{4 \cdot 176} = 0{,}988, \qquad (20)$$

где P1 – вероятность безотказной обработки одного усреднённого набора исходных данных.

Заключение. Оценка ожидаемой надёжности планируемой к разработке прикладной компьютерной программы, предназначенной для работы в составе информационной системы, в том числе и при обработке Big Data, является ответственной задачей, так как позволяет ещё до начала написания программного кода сориентироваться в целесообразности вкладывания средств в разработку компьютерной программы и оценить время, необходимое для обеспечения её заданной эксплуатационной надёжности.

Иллюстрация применения метода расчёта надёжности, рассматривая конкретный пример, позволит лучше осмыслить подход, используемый для оценки ожидаемой надёжности планируемой к разработке прикладной компьютерной программы.

Список литературы

[1] Borovikov S.M., Kaziuchyts V.O., Khoroshko V.V., Dick S.S., Klinov K.I. Assessment of expected reliability of applied software for computer-based information systems // Informatics. – 2021. – Vol. 18, No. 1. – P. 84–95 (in Russian). https://doi.org/10.37661/1816-0301-2021-18-1-84-95.

[2] Чуканов В.О., Гуров В.В., Прокопьева Е.В. Методы обеспечения аппаратно-программной надёжности вычислительных систем [Электронный ресурс]. – 2014. – URL: http://www.mcst.ru/files/5357ec/dd0cd8/50af39/000000/seminar_metody_obespecheniya_apparatno-programmnoy_nadezhnosti_vychislitelnyh_sistem.pdf (дата обращения 14.03.2026).

[3] Программное обеспечение – источник всех проблем [Электронный ресурс]. – 2018. – URL: http://www.williamspublishing.com/PDF/5-8459-0785-3/part1.pdf (дата обращения 14.03.2026).

[4] Боровиков С.М., Будник А.В., Лэ В.Т. Надежность разрабатываемых прикладных компьютерных программ для информационных систем = Reliability of developed application computer programs of information systems // Проблемы инфокоммуникаций: научный журнал. – 2025. – № 1. – С. 25–34.

[5] Боровиков С.М., Дик С.С., Лэ В.Т., Клинов К.И. Анализ и оценка надёжности прикладных компьютерных программ // BIG DATA и анализ высокого уровня: Сб. матер. VI Междунар. науч.-практ. конф. / Минск (20–21 мая 2020 г.). –. Ч. 1. – Минск: Бестпринт, 2020. – С. 382–390.

[6] Боровиков С.М., Дик С.С., Лэ В.Т., Клинов К.И. Модель прогнозирования надёжности планируемых к разработке прикладных компьютерных программ // Интернаука: научный журнал. – 2020. – № 12 (141), ч. 1. – С. 68–72.

[7] Боровиков С.М., Дик С.С., Лэ, В.Т., Клинов К.И. Возможный подход к оценке надёжности разрабатываемых программных средств на ранних этапах проектирования информационно-компьютерных систем // Globus: технические науки – от теории к практике [Электронный ресурс]: Сб. науч. публ. – 2020. – Вып. 1 (32). – С. 4–9.

[8] Боровиков С.М., Казючиц В.О., Дик С.К., Дик С.С. Модель прогнозирования времени тестирования прикладных компьютерных программ для технологий BIG DATA / BIG DATA и анализ высокого уровня: Сб. науч. статей VII Междунар. науч.-практ. конф. / Минск (19–20 мая 2021 г.). – Минск: Бестпринт, 2021. – С. 404–411.

[9] Боровиков С.М., Казючиц В.О., Дик С.С., Юреня К.В. Модель прогнозирования времени тестирования прикладных компьютерных программ для автоматизированных систем управления // Информационные технологии и системы 2021: Матер. Международ. науч. конф. ИТС-2021 / Минск (24 ноября 2021 г.). – Минск, 2021. – С. 28–29.

[10] Боровиков С.М., Казючиц В.О., Дик С.К., Дик С.С. Методика обеспечения эксплуатационной надёжности планируемых к разработке прикладных компьютерных программ для информационных систем // BIG DATA и анализ высокого уровня: Сб. науч. статей VIII Международ. науч.-практ. конф. / Минск (11–12 мая 2022 г.). – Минск: Бестпринт, 2022. – С. 162–173.

[11] Лэ В.Т., Боровиков С.М., Дик С.К., Будник А.В. Рекомендации по оценке и обеспечению надёжности планируемых к разработке прикладных компьютерных программ для информационных систем // BIG DATA и анализ высокого уровня: Сб. науч. статей IX Международ. науч.-практ. конф. / Минск (17–18 мая 2023 г.). – Ч. 2. – Минск, 2023. – С. 78–86.

[12] Шубинский И.Б. Функциональная надёжность информационных систем. Методы анализа – М.: «Журнал Надёжность», 2012. – 296 с.

[13] McCall J.A., Randell W., Dunham J., Lauterbach L. Software reliability, measurement, and testing guidebook for software reliability measurement and testing // RL-TR-92-52: Final Technical Report, Vol II (of two) / Science Applications International Corp., Research Triangle Institute. – Rome Laboratory, NY 13441-5700, 1992. – 256 p. – URL: https://apps.dtic.mil/sti/tr/pdf/ADA256164.pdf (дата обращения 14.03.2026).

[14] Постановление Министерства труда и социальной защиты Республики Беларусь 27 июня 2007 г. № 91 «Об утверждении укрупнённых норм затрат труда на разработку программного обеспечения» [Электронный ресурс]. – URL: https://news-newsby-org.narod.ru/doc-razn/pos05/postn05236/index.htm (дата обращения 14.03.2026).

[15] Чуканов В.О. Надёжность программного обеспечения и аппаратных средств систем передачи данных атомных электростанций: учеб. пособ. – М.: МИФИ, 2008. – 168 с.

Авторский вклад

Лэ Ван Там – уточнение модели для оценки ожидаемой начальной интенсивности проявления ошибок прикладной компьютерной программой.

Боровиков Сергей Максимович – общее руководство исследованиями; подготовка данных для расчёта метрики, учитывающей характеристики планируемой к разработке компьютерной программы.

Будник Артур Владимирович – подготовка данных для расчёта метрики, учитывающей характеристику производственной среды разработки компьютерной программы, в том числе для мониторинга и обеспечения безопасности, телекоммуникации и мобильных устройств, систем автоматизированного управления.

Дик Сергей Константинович – подготовка справочных параметров для оценки надёжности прикладных компьютерных программ для разных областей применения.

PRACTICE OF ASSESSING THE EXPECTED RELIABILITY OF APPLIED COMPUTER PROGRAMS FOR INFORMATION SYSTEMS

Van Tam Le
Researcher, Institute of Technology, Hanoi, Vietnam, Master of Engineering and Technology

S.M. Borovikov
Associate Professor, Department of Information Computer Systems Design, BSUIR, PhD of Technical sciences, Associate Professor

A.V. Budnik
Dean of the Faculty of Engineering and Technology of the BSAC, PhD of Technical sciences, Associate Professor

S.K. Dick
Head of the Department of Engineering and Computer Graphics at BSUIR, PhD of Technical Sciences, Associate Professor

Abstract. This article examines the practice of assessing the operational reliability of planned computer applications designed for information systems. A specific example is provided for assessing the expected reliability of a computer program used for modeling and researching design solutions in the educational process for students at a technical university.

Keywords: computer information systems, applied software, expected reliability of software, testing.

УДК 330.46:004.6

КОМПЛЕКСНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ БОЛЬШИХ ДАННЫХ: ИНТЕГРАЦИЯ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ, КЛАСТЕРИЗАЦИИ И РЕГРЕССИИ



Е.И. Полоско
Старший преподаватель кафедры экономической информатики БГУИР
e.i.polosko@gmail.com

Е.И. Полоско

Окончила Белорусский государственный университет. Область научных интересов связана с анализом данных и методами машинного обучения, экономико-математическим моделированием, технологиями Big Data и решениями на основе интернета вещей (IoT).

Аннотация. В статье рассматривается методология комплексного статистического анализа, объединяющая методы описательной статистики, корреляционного анализа, снижения размерности (метод главных компонент), кластеризации (K-Means) и регрессионного моделирования для обработки больших данных. Используется пошаговый подход к анализу данных, реализованный в среде Python и апробированный на имитированных производственных и энергетических данных ОАО «МАЗ» за 2021–2025 гг. Проведён сравнительный анализ регрессионных моделей прогнозирования энергопотребления; ансамблевые методы (Gradient Boosting) демонстрируют наивысшую точность (R^2 = 0,873 и MAE = 3,61 МВт·ч) по сравнению с линейными моделями. Практическая значимость работы заключается в универсальном подходе к анализу производственных данных для поддержки управленческих решений и оптимизации бизнес-процессов на промышленных предприятиях.

Ключевые слова: большие данные, комплексный статистический анализ, ОАО «МАЗ», метод главных компонент, кластерный анализ, регрессионное моделирование, энергоэффективность, машиностроение, предиктивная аналитика, Python

Введение. В современных условиях цифровой трансформации объём генерируемых данных растёт экспоненциально. По данным аналитиков, глобальный объём данных достигнет 394 зеттабайт к 2028 году [1], при этом до 90 % корпоративных данных остаются неструктурированными [2]. Мировой рынок больших данных, оценивавшийся в 262,87 млрд долларов в 2024 году, по прогнозам достигнет 1 019 млрд долларов к 2035 году со среднегодовым темпом роста 13,10 % [3]. Динамика роста рынка представлена на рисунке 1.

В этих условиях применение изолированных статистических методов часто оказывается недостаточным для извлечения значимых закономерностей из многомерных массивов данных. Необходим интегрированный подход, объединяющий несколько этапов статистического анализа в единый аналитический конвейер [4]. Сегмент продвинутой аналитики (Advanced Analytics) прогнозируемо займёт свыше 68,56 % глобального рынка Big Data в ближайшее десятилетие.

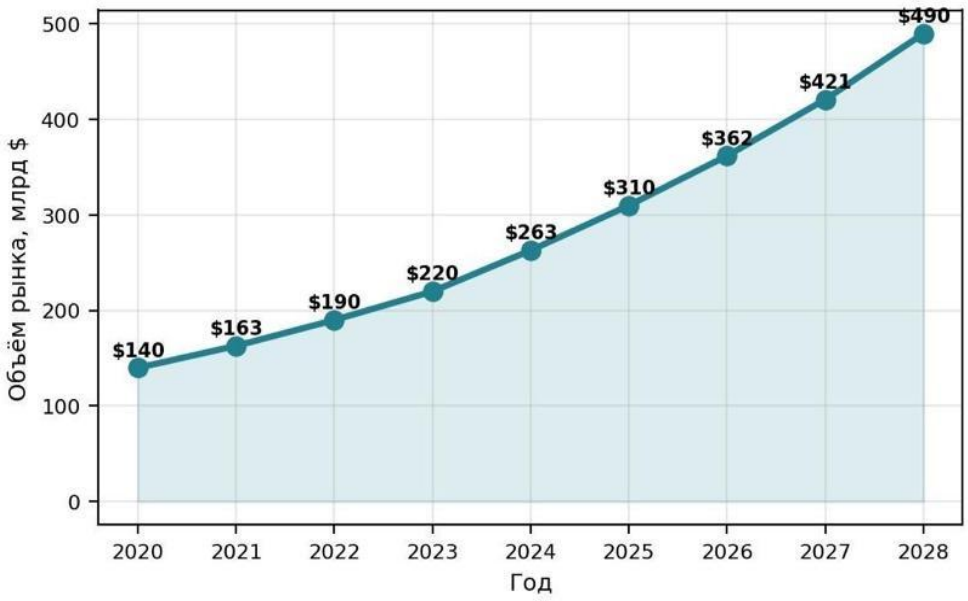


Рисунок 1. Динамика мирового рынка больших данных, млрд $ [3]

Целью данной работы является разработка и экспериментальная апробация методологии комплексного статистического анализа больших данных, интегрирующей методы описательной статистики, корреляционного анализа, снижения размерности, кластеризации и регрессионного моделирования для оптимизации бизнес-решений.

Методология комплексного статистического анализа. Предлагаемая методология представляет собой последовательный аналитический конвейер (pipeline), состоящий из семи взаимосвязанных этапов. Каждый этап решает определённую задачу, а результаты предыдущего этапа служат входными данными для последующего [5]. Общая схема методологии представлена на рисунке 2.



Рисунок 2. Схема аналитического конвейера комплексного статистического анализа

Характеристика каждого этапа представлена в таблице 1.

Таблица 1. Этапы комплексного статистического анализа больших данных

№	Этап анализа	Применяемые методы	Ожидаемый результат
1	Предобработка данных	Очистка, нормализация, обработка пропусков	Чистый и структурированный датасет
2	Описательная статистика	Меры центральной тенденции, вариации, распределения	Общая характеристика данных
3	Корреляционный анализ	Коэффициент Пирсона, тепловая карта	Матрица корреляций признаков
4	Снижение размерности	Метод главных компонент (PCA)	Сжатое пространство признаков
5	Кластерный анализ	K-Means, метод локтя, силуэтный анализ	Сегменты (кластеры) данных
6	Регрессионное моделирование	Linear, Ridge, Lasso, RF, GBM	Предиктивные модели
7	Валидация	Кросс-валидация, R^2, MAE, RMSE	Оценка качества моделей

Ключевым преимуществом данного подхода является синергетический эффект: комбинация методов позволяет компенсировать ограничения каждого отдельного метода и обеспечивает более полное извлечение информации из данных [6].

Корреляционный анализ многомерных данных. На этапе корреляционного анализа вычисляется матрица парных коэффициентов корреляции Пирсона для всех числовых признаков набора данных. Данная матрица позволяет выявить линейные зависимости между переменными и определить мультиколлинеарные признаки, подлежащие агрегации или исключению [7].

Для экспериментальной апробации был использован синтетический набор данных, моделирующий бизнес-показатели ОАО «МАЗ» за 60 месяцев (январь 2021 – декабрь 2025 гг.). На этапе корреляционного анализа вычислена матрица парных коэффициентов корреляции Пирсона для всех восьми производственных показателей: объёма производства (X_1), энергопотребления (Y), простоя оборудования (X_2), доли брака (X_3), расходов на сырьё (X_4), производительности труда (X_5), загрузки оборудования (X_6) и выручки (X_7).

Тепловая карта корреляций представлена на рисунке 3.

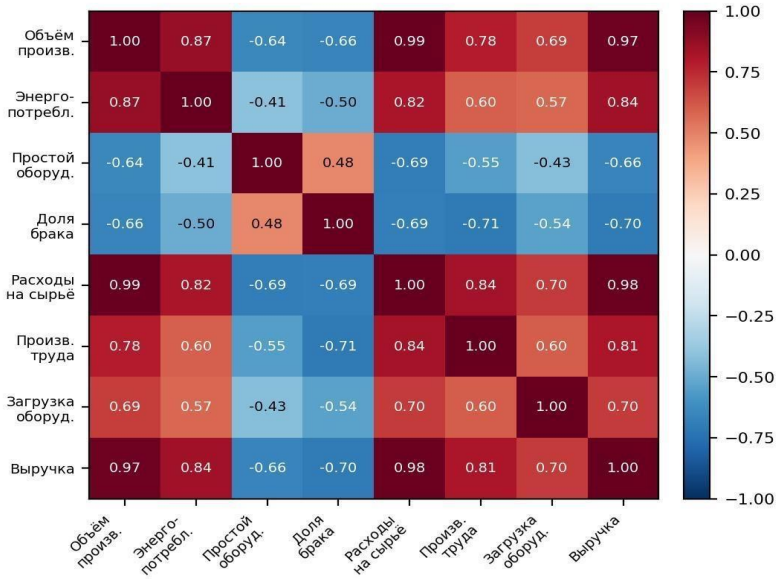


Рисунок 3. Тепловая карта корреляций производственных показателей ОАО «МАЗ»

Анализ матрицы корреляций выявил следующие значимые зависимости: сильная положительная корреляция между объёмом производства и энергопотреблением (r = 0,87), объёмом и выручкой (r = 0,97), объёмом и расходами на сырьё (r = 0,99).

Сильная отрицательная корреляция обнаружена между простоем оборудования и расходами на сырьё (r = −0,69), долей брака и производительностью труда (r = −0,71). Наличие групп коррелированных признаков обосновывает применение PCA.

Снижение размерности методом главных компонент. Метод главных компонент (Principal Component Analysis, PCA) является одним из наиболее эффективных инструментов снижения размерности для больших данных. PCA преобразует исходные показатели в новый набор показателей так, чтобы они не дублировали друг друга и шли в порядке убывания важности: первые компоненты объясняют наибольшую часть разброса данных.

Математически сначала для данных строится ковариационная матрица Σ, которая описывает, как признаки связаны друг с другом.

Затем для этой матрицы находят специальные направления (собственные векторы) и соответствующие им числа (собственные значения); каждое направление задаёт главную компоненту, а связанное с ним число показывает, какую долю общего разброса данных объясняет эта компонента: чем больше собственное значение, тем важнее компонента.

Результаты применения PCA к исследуемому набору данных представлены на рисунке 4 и в таблице 2.

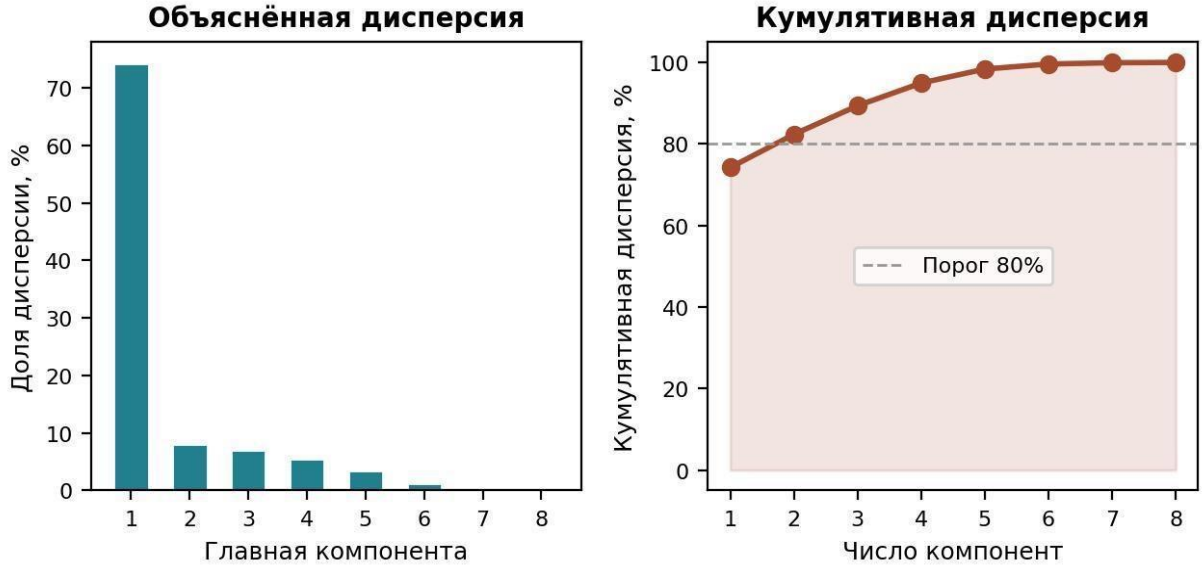


Рисунок 4. Объяснённая и кумулятивная дисперсия главных компонент

Таблица 2 . Результаты анализа главных компонент

Компонента	Собственное значение	Доля дисперсии, %	Кумулятивная доля, %
PC1	6,038	74,22	74,22
PC2	0,662	8,13	82,35
PC3	0,572	7,02	89,38
PC4	0,453	5,56	94,94
PC5	0,279	3,43	98,38
PC6	0,100	1,24	99,61
PC7	0,026	0,32	99,93
PC8	0,006	0,07	100,00

Как видно из таблицы 2, первая главная компонента объясняет 74,22% всего разброса данных и фактически отражает общий уровень производственной активности.

Она объединяет изменения объёма производства, расходов на сырьё, выручки и энергопотребления, то есть тех показателей, которые сильнее всего связаны между собой (коэффициенты корреляции от 0,87 до 0,99). Большое значение $\lambda_1 = 6{,}038$ по сравнению с остальными говорит о том, что за поведением системы ОАО «МАЗ» стоит один ведущий скрытый фактор, определяющий её основную динамику.

Кластерный анализ в пространстве главных компонент. После снижения размерности был проведён кластерный анализ методом K-Means в пространстве первых двух главных компонент .

Оптимальное число кластеров (k = 3) определено с использованием метода «локтя» и силуэтного анализа.

Результаты кластеризации визуализированы на рисунке 5.

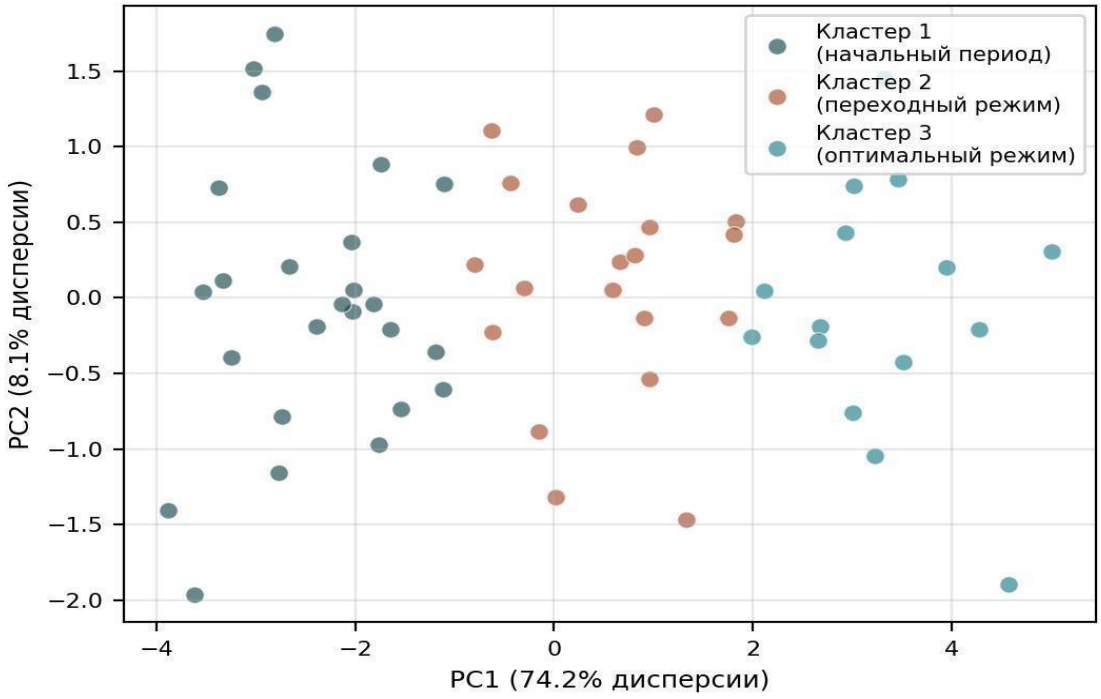


Рисунок 5. Результаты кластерного анализа в пространстве PC1–PC2

Характеристика выявленных кластеров представлена в таблице 3.

Таблица 3. Характеристика кластеров производственных периодов ОАО «МАЗ»

Кластер	**n**	**Произв., ед.**	**Энергия, МВт·ч**	**Простой, ч**	**Загрузка, %**
C1 (начальный)	25	1286	220	112	74,7
C2 (переходный)	20	1434	235	95	78,3
C3 (оптимальный)	15	1573	253	81	83,5

Кластер 1 («начальный период», 2021 – начало 2022 гг.) включает 25 месяцев и соответствует наименее эффективному состоянию: выпуск невысокий (1 286 ед./мес.), простой большой (112 ч/мес.), загрузка низкая (74,7%), энергопотребление минимально (220 МВт·ч/мес.).

Кластер 2 («переходный период», 2022–2023 гг.) объединяет 20 месяцев и отражает промежуточный этап: выпуск растёт до 1 434 ед./мес., простои снижаются до 95 ч/мес., загрузка повышается до 78,3%, что соответствует постепенному наращиванию мощностей.

Кластер 3 («оптимальный режим», 2024–2025 гг.) содержит 15 месяцев и описывает наиболее эффективную работу: выпуск 1 573 ед./мес., простои 81 ч/мес., загрузка 83,5%, при максимальном энергопотреблении (253 МВт·ч/мес.) удельная энергоёмкость снижается, что говорит о росте энергоэффективности.

Переход от кластера 1 к кластеру 3 показывает последовательную модернизацию и внедрение IoT-систем на ОАО «МАЗ»; выделение трёх режимов позволяет давать практические рекомендации по оптимизации загрузки оборудования и сокращению простоев.

Регрессионное моделирование энергопотребления. На заключительном этапе проведено регрессионное моделирование энергопотребления предприятия на основе пяти производственных показателей (объём производства, простой оборудования, доля брака, производительность труда, загрузка оборудования). Сравнительный анализ выполнен для пяти методов с использованием 5-кратной кросс-валидации [9, 10]. Результаты представлены в таблице 4 и на рисунке 6.

Таблица 4. Сравнительный анализ регрессионных моделей прогнозирования энергопотребления

Модель	R^2	MAE, МВт·ч	RMSE, МВт·ч	Время обуч., с
Линейная регрессия	0,768	6,42	7,34	0,08
Ridge-регрессия	0,782	5,98	7,02	0,10
Lasso-регрессия	0,774	6,15	7,18	0,09
Random Forest	0,847	4,27	5,12	1,54
Gradient Boosting	0,873	3,61	4,38	2,81

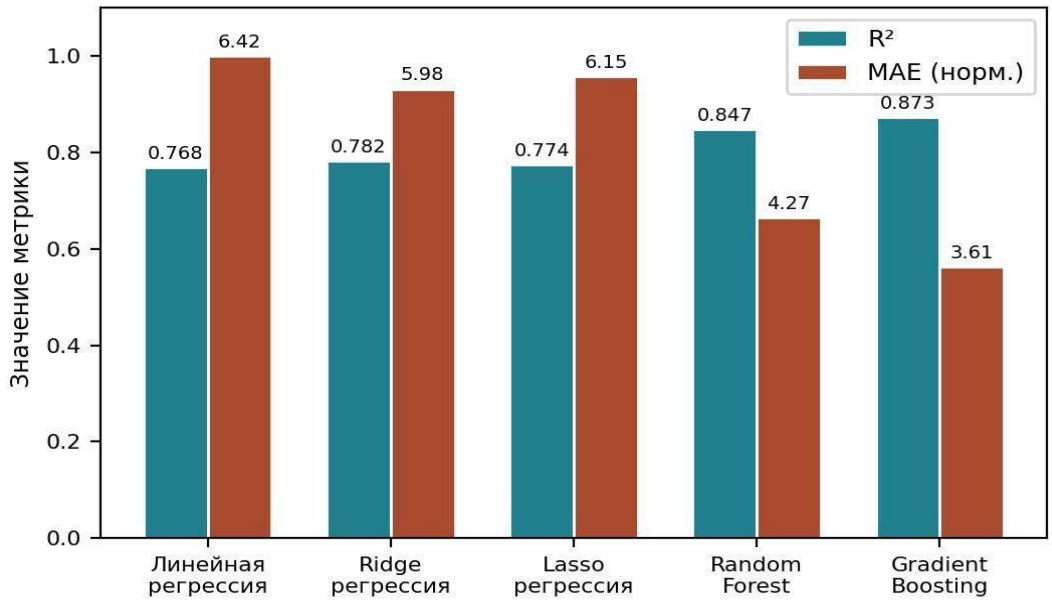


Рисунок 6. Сравнение метрик качества регрессионных моделей

Ансамблевые методы лучше справились с задачей, чем обычные линейные модели. Gradient Boosting дал лучшие результаты (R^2 = 0,873, MAE = 3,61 МВт·ч) и улучшил качество по сравнению с простой линейной регрессией на 13,7%, что указывает на выраженную нелинейность связи между энергопотреблением и производственными показателями.

Модель Random Forest оказалась второй по точности (R^2 = 0,847) и дополнительно позволяет оценивать вклад каждого признака, что упрощает интерпретацию.

Регуляризованные линейные модели (Ridge, Lasso) немного улучшили результаты базовой линейной регрессии, подтверждая наличие тесно связанных между собой признаков, выявленных на этапе корреляционного анализа.

Обсуждение результатов. Предложенный комплексный подход обладает рядом преимуществ для анализа производственных данных промышленного предприятия:

1) PCA позволил сократить размерность с 8 до 2 признаков при сохранении 82,35 % информации, что значительно упрощает визуализацию и интерпретацию многомерных производственных данных ОАО «МАЗ».

2) Кластерный анализ выявил три характерных режима работы предприятия, соответствующих этапам модернизации и повышения энергоэффективности.

Переход от кластера 1 к кластеру 3 сопровождается снижением простоя на 28 % и ростом загрузки оборудования на 12 %.

3) Ансамблевые модели обеспечивают точность прогнозирования энергопотребления с ошибкой 3,61 МВт·ч (1,5 % от среднемесячного потребления), что достаточно для оперативного планирования энергоресурсов.

4) Результаты согласуются с данными о достигнутом снижении энергопотребления ОАО «МАЗ» на 6,9 % в Н1 2025 г. и подтверждают эффективность внедрения систем мониторинга и IoT-технологий.

Ограничением исследования является использование синтетических данных.

Вместе с тем, параметры генерации откалиброваны по открытым отчётным данным предприятия и отраслевой статистике, что обеспечивает реалистичность результатов [10, 11].

Заключение. В данной работе предложена и экспериментально апробирована методология комплексного статистического анализа производственных данных промышленного предприятия ОАО «МАЗ». Основные результаты:

– корреляционный анализ восьми производственных показателей выявил группы тесно связанных признаков (r до 0,99), обосновывающие необходимость снижения размерности;

– метод главных компонент обеспечил сокращение размерности на 75 % (с 8 до 2 компонент) при сохранении 82,35 % информации;

– кластерный анализ идентифицировал три режима работы предприятия, отражающих динамику модернизации за 2021–2025 гг.;

– ансамблевый метод Gradient Boosting показал наивысшую точность прогнозирования энергопотребления (R^2 = 0,873, MAE = 3,61 МВт·ч), превзойдя линейные модели на 10–14 %;

– предложенный аналитический конвейер является масштабируемым и может быть адаптирован для других предприятий машиностроительного комплекса Республики Беларусь.

Перспективным направлением является интеграция методов глубокого обучения (LSTM-сетей) для прогнозирования временных рядов энергопотребления, а также подключение данных IoT-датчиков для анализа в режиме реального времени.

Список литературы

[1] Ассоциация больших данных; Б1; TAdviser. Исследование рынка больших данных и искусственного интеллекта в России, 2025 [Электронный ресурс]. URL: https://b1.ru/insights/news/b1-materials-in-media/b1-kommersant-big-data-survey-13-november-2025/ (дата обращения: 02.03.2026).

[2] Ассоциация больших данных; Б1; TAdviser. Рынок больших данных и искусственного интеллекта в России 2025 [Электронный ресурс]. URL: https://www.tadviser.ru/index.php/Статья:Исследование_рынка_Больших_данных_и_Искусственного_интеллекта_в_России (дата обращения: 02.03.2026).

[3] Попов В.В. Применение искусственного интеллекта и больших данных в практике российских организаций. Прикладная статистика и искусственный интеллект. 2024. № 4. С. 47–60.

[4] Вишняков В.А. Specialized IoT systems: Models, Structures, Algorithms, Hardware, Software Tools: монография. Минск: БГУИР, 2023. 184 с.

[5] Домакур О.В. Методы анализа больших данных. Веснік сувязі. 2020. № 6(164). С. 50–55.

[6] Официальный сайт ОАО «МАЗ». Отчетность ОАО «МАЗ» – управляющая компания холдинга «БЕЛАВТОМАЗ» [Электронный ресурс]. URL: https://maz.by/about/reporting (дата обращения: 10.03.2026).

[7] Сташевская М.П. Анализ применения больших данных в Республике Беларусь в контексте перехода к цифровой экономике. Экономическая наука сегодня: сб. науч. ст. Минск: БНТУ, 2024. Вып. 19. С. 70–78.

[8] Montgomery D.C., Peck E.A., Vining G.G. Introduction to Linear Regression Analysis. 5th ed. Hoboken: Wiley, 2012. 672 p.

[9] Кузнецов В.О. Использование метода главных компонент для анализа надежности цепей поставок. Логистика и управление цепями поставок. 2018. № 4(87). С. 27–33.

[10] Измайлова М.В. Факторный и кластерный анализ основных показателей производственной деятельности предприятий промышленности и транспортного комплекса. Российское предпринимательство. 2012. № 24. С. 52–60 [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/faktornyy-i-klasternyy-analiz-osnovnyh-pokazateley-proizvodstvennoy-deyatelnosti-predpriyatiy-promyshlennosti-i-transportnogo (дата обращения: 15.03.2026).

[11] Волошин Т.А., Зайцев К.С., Дунаев М.Е. Применение адаптивных ансамблей методов машинного обучения к задаче прогнозирования временных рядов. International Journal of Open Information Technologies. 2023. Т. 11. № 5. С. 32–41.

Авторский вклад

Полоско Екатерина Ивановна – постановка задачи исследования, разработка методологии комплексного статистического анализа, генерация имитированных данных, реализация аналитического конвейера в среде Python, интерпретация результатов, подготовка текста статьи.

COMPLEX STATISTICAL ANALYSIS OF BIG DATA: INTEGRATION OF DIMENSIONALITY REDUCTION, CLUSTERING AND REGRESSION METHODS

E.I. Polosko

Senior Lecturer, Department of Economic Informatics, BSUIR

e.i.polosko@gmail.com

Abstract. The article presents a methodology for complex statistical analysis that combines descriptive statistics, correlation analysis, dimensionality reduction (Principal Component Analysis), clustering (K-Means) and regression modelling for big data processing. A step-by-step data analysis approach is used, implemented in Python and tested on simulated production and energy data of OJSC “MAZ” for 2021–2025. A comparative analysis of regression models for forecasting energy consumption is carried out; ensemble methods (Gradient Boosting) demonstrate the highest accuracy (R^2 = 0,873 и MAE = 3,61 МВт·ч) compared to linear models. The practical significance of the study lies in a universal approach to the analysis of production data to support managerial decision-making and optimize business processes at industrial enterprises.

Keywords: big data, complex statistical analysis, OJSC “MAZ”, principal component analysis, cluster analysis, regression modelling, energy efficiency, mechanical engineering, predictive analytics, Python.

УДК 004.8:616.89

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ УРОВНЯ ТРЕВОЖНОСТИ ПО ПОВЕДЕНЧЕСКИМ И СОЦИАЛЬНО-ДЕМОГРАФИЧЕСКИМ ПРИЗНАКАМ



Н.И. Липницкая
Старший преподаватель кафедры экономической информатики БГУИР
n.i.karpovich@gmail.com



М. А. Миткевич
Студент инженерно-экономического факультета БГУИР
mitkevich@bsuir.by



В. В. Смертьев
Студент инженерно-экономического факультета БГУИР
smertsyeu@bsuir.by

Н.И. Липницкая
Старший преподаватель кафедры экономической информатики УО «БГУИР». Область научных интересов: прикладные системы обработки данных, информационные технологии и программирование, методика преподавания.

М. А. Миткевич
Студент инженерно-экономического факультета УО «БГУИР». Область научных интересов: машинное обучение, анализ данных, поведенческая аналитика.

В. В. Смертьев
Студент инженерно-экономического факультета УО «БГУИР». Область научных интересов: анализ данных, поведенческая аналитика, методы машинного обучения.

Аннотация. Рост распространённости тревожных расстройств требует разработки доступных инструментов скрининга. В данном исследовании проверяется гипотеза о возможности косвенной оценки уровня тревожности по GAD-7 на основе поведенческих и социально-демографических факторов. На основе данных опроса 439 респондентов обучены и протестированы модели машинного обучения (Ridge, Random Forest, CatBoost, LightGBM). CatBoost продемонстрировал наилучшие результаты: MAE = 2.79 балла, R^2 = 0.39. Ключевыми протективными предикторами являются удовлетворённость рабочей/учебной средой и качество сна, факторами риска – дефицит социальной поддержки и негативные жизненные события.

Ключевые слова: тревожность, психическое здоровье, оценка риска, машинное обучение, поведенческие факторы, цифровая психиатрия, медицинская информатика, GAD-7, поведенческая аналитика.

Введение. Цифровая трансформация здравоохранения (e-Health) смещает акцент с лечения уже развившихся заболеваний на профилактику и ранний скрининг. Психические расстройства, в том числе генерализованное тревожное расстройство, остаются одной из ключевых причин снижения качества жизни, при этом тревожные состояния затрагивают значительную часть трудоспособного населения. Их диагностика нередко запаздывает из-за стигматизации и нехватки специалистов.

В настоящее время основным инструментом скрининга тревожных расстройств являются стандартизированные опросники, такие как GAD-7 [2; 3]. Несмотря на высокую надёжность, они предполагают прямой контакт с врачом или осознанное и регулярное заполнение пациентом, что ограничивает их применение в массовых профилактических программах. Развитие методов машинного обучения открывает возможность перехода к

«цифровому фенотипированию», когда уровень риска оценивается по косвенным, но легко собираемым данным.

Модели машинного обучения уже применяются в психологии и психиатрии для прогнозирования депрессии, тревожности и стресса по разнородной информации: текстам, активности в социальных сетях, физиологическим сигналам и цифровым логам [4; 5]. Однако такие подходы часто требуют специального оборудования, длительного мониторинга или вызывают серьёзные вопросы конфиденциальности. Большая часть работ опирается либо на дорогостоящие биомедицинские сигналы (например, ЭЭГ, вариабельность сердечного ритма) , либо на пассивный цифровой след [7; 8], в то время как потенциал моделей, использующих только целенаправленно собранные поведенческие факторы, пока изучен недостаточно.

В данной работе проверяется гипотеза о том, что совокупность простых, легко собираемых поведенческих паттернов и социально-демографических характеристик может обеспечить приемлемую точность прогноза уровня тревожности для задач первичного скрининга. Цель исследования – разработать и оценить модель машинного обучения на основе алгоритмов Ridge, Random Forest, CatBoost и LightGBM и сравнить их качество. В отличие от многих исследований, использующих дорогие медицинские измерения или пассивные цифровые следы, здесь прогноз строится только по анкетным данным, не требующим сложной инфраструктуры.

Дополнительным вкладом работы является подробный анализ ошибок модели в разных диапазонах тревожности и интерпретация значимости признаков с помощью метода SHAP. В качестве ориентира для оценки приемлемой точности используются минимально клинически значимая разница по шкале GAD-7 (3–4 балла) и типичный диапазон значений коэффициента детерминации R^2 для психометрических моделей на опросных данных.

Организация исследования и сбор данных. Сбор данных проводился в формате одномоментного онлайн-опроса с использованием платформы Google Forms. В исследовании приняли участие 439 респондентов, преимущественно студенты и молодые специалисты. Форма была настроена так, чтобы исключить пропуски ответов, поэтому в итоговом датасете отсутствовали пропущенные значения. В качестве целевой переменной использовалась суммарная оценка по шкале GAD-7 (от 0 до 21 балла). Предикторы были получены на основе авторских вопросов, опирающихся на психометрические исследования образа жизни, и включали показатели качества сна, уровня физической активности, особенностей питания, наличия вредных привычек, социальной поддержки и удовлетворённости учебной/рабочей средой.

Описание признакового набора. Предварительная обработка данных и построение моделей выполнялись на языке Python с использованием библиотек Pandas и Scikit-learn.

Все переменные были переведены в числовой формат, пропусков в данных не осталось. Категориальные и порядковые признаки были закодированы так, чтобы сохранять естественный порядок их значений. Исходные предикторы объединены в три смысловые группы, показанные в таблице 1.

Таблица 1. Категории факторов и их описание

Категория факторов	Описание предикторов
Физиологические и поведенческие	Продолжительность сна, уровень физической активности, качество питания, статус курения, употребление алкоголя, приём лекарственных препаратов.
Социально-психологические	Удовлетворённость учебной или рабочей средой, субъективный уровень социальной поддержки, посещение психолога, наличие травмирующих событий за последний год.
Рутинные паттерны	Экранное время (использование гаджетов), частота прогулок на свежем воздухе, хронотип (суточный ритм активности).

Анализ взаимосвязи данных. На этапе предварительного исследования данных был проведён корреляционный анализ для выявления линейных зависимостей между предикторами и целевой переменной.

Разработка предиктивных моделей машинного обучения. Разработка и обучение моделей проводились в среде Google Colab на языке Python с использованием библиотек Scikit-learn, LightGBM и CatBoost. Для решения задачи были выбраны алгоритмы разной природы: ансамблевые методы Random Forest, LightGBM и CatBoost, а также линейная Ridge-регрессия, использовавшаяся как базовая модель для сравнения. Отдельно тестировалась стратегия логарифмического преобразования целевой переменной (Log-Target), однако во всех случаях это приводило к ухудшению качества (росту ошибок), поэтому данный подход был отклонён. Оценка качества моделей выполнялась на отложенной тестовой выборке (hold-out) в пропорции 77,5% / 22,5% с использованием стратифицированного разбиения по уровням тревожности, чтобы сохранить структуру распределения GAD-7. Основной метрикой выступала средняя абсолютная ошибка (MAE), измеряемая в баллах шкалы GAD-7. Дополнительно рассчитывались коэффициент детерминации R^2 и среднеквадратическая ошибка (RMSE) для более полной оценки качества прогноза.

Результаты эксперимента и интерпретация итоговой модели. По результатам серии экспериментов наивысшую предсказательную точность продемонстрировала модель CatBoost с оптимизацией гиперпараметров методом Optuna. Математически предсказание модели CatBoost представляет собой взвешенную сумму предсказаний T деревьев решений:

$$\mathrm{YGAD7}=\mathrm{t}=1\mathrm{ThtX} \quad (1)$$

где YGAD7 – прогнозируемый уровень тревожности, htX – предсказание t-го дерева на объекте X, η – скорость обучения (learning rate), T – число деревьев в ансамбле.

Сводные результаты тестирования лучших конфигураций моделей представлены в таблице 2.

Таблица 2. Сводные результаты тестирования лучших конфигураций моделей

Алгоритм	Конфигурация	MAE	RMSE	R^2
CatBoost	Optuna	2.79	3.65	0.39
Ridge Regression	GridSearch (α=10)	2.87	3.71	0.36
LightGBM	Optuna	2.87	3.70	0.37
Random Forest	Optuna + Gap Penalty	2.87	3.71	0.36

По результатам сравнительного эксперимента наилучшие показатели продемонстрировала модель CatBoost с подбором гиперпараметров (Optuna): MAE = 2.79, R^2 = 0.39. Линейная Ridge-регрессия показала очень близкие результаты (MAE = 2.87, R^2 = 0.36), что косвенно указывает на преобладание аддитивных эффектов предикторов и относительно небольшую роль сложных нелинейных взаимодействий. Логарифмическое преобразование целевой переменной и метод ансамблирования на основе случайных подвыборок не дали прироста качества и не были использованы в финальной конфигурации моделей.

Интерпретация предикторов. Для содержательной интерпретации модели CatBoost применялся метод SHAP (SHapley Additive exPlanations), позволяющий оценить вклад каждого признака в индивидуальное предсказание (рис. 1). В отличие от весовых коэффициентов линейных моделей, SHAP-значения корректно отражают нелинейные зависимости и взаимодействия между признаками.

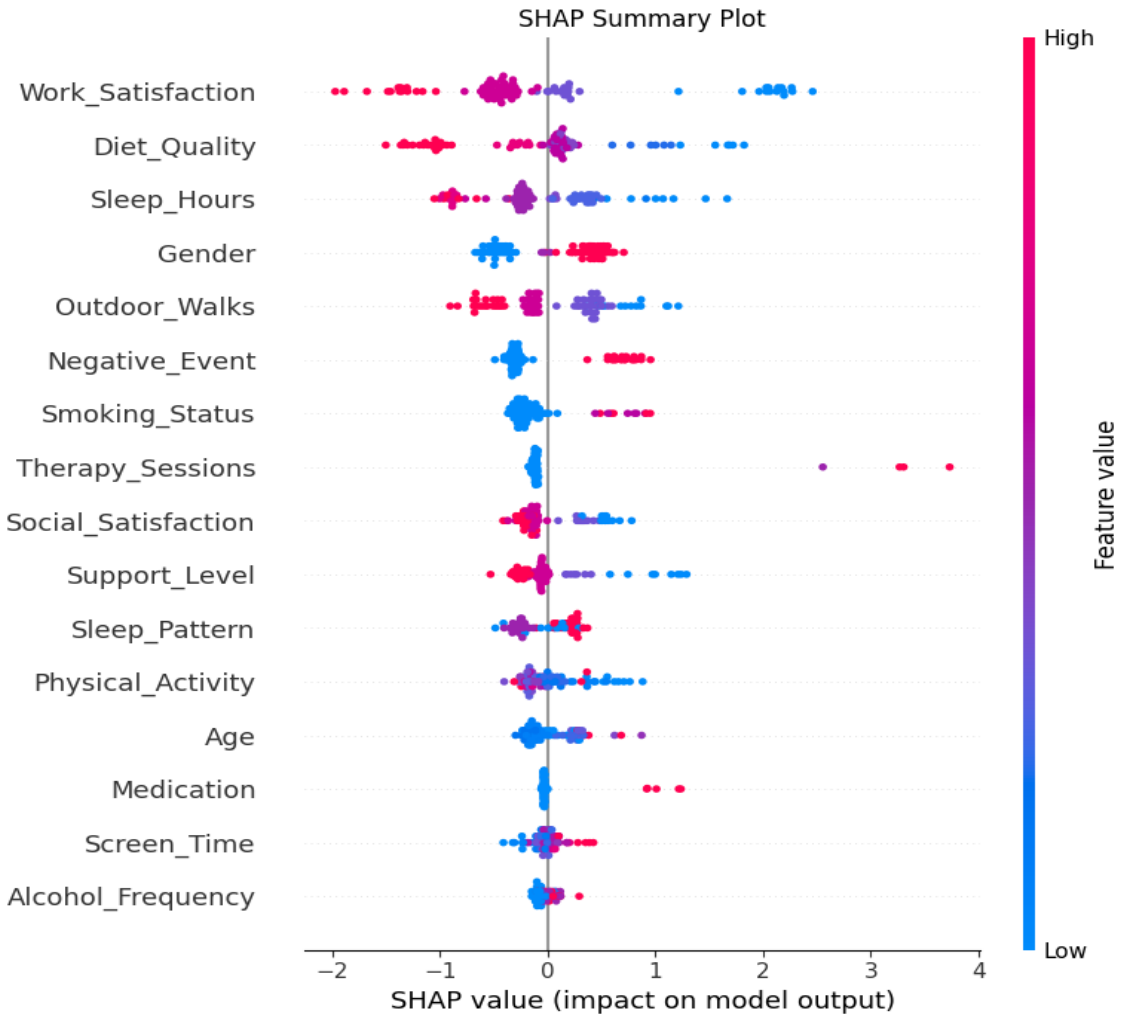


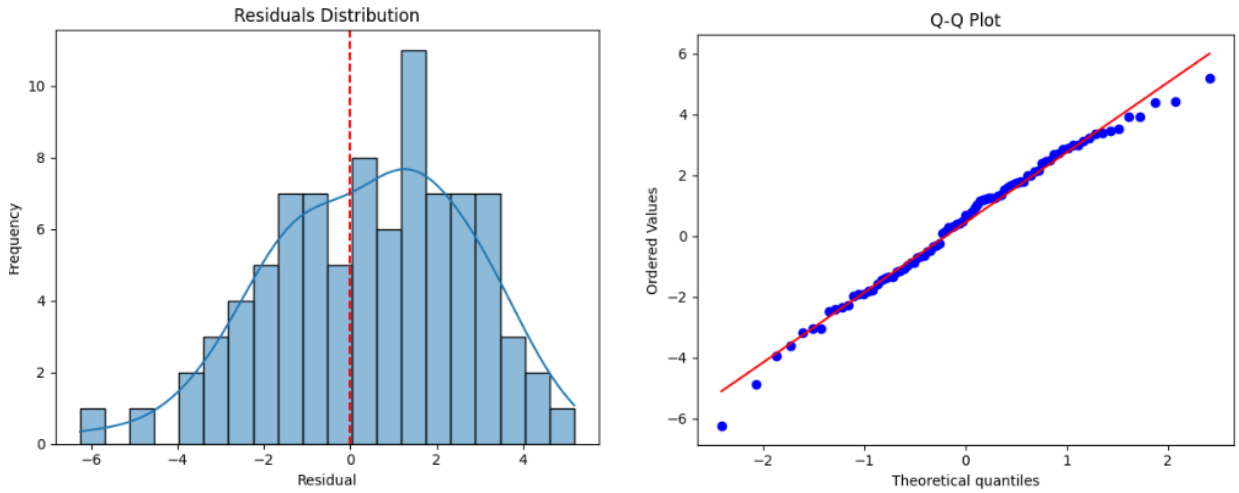
Рисунок 1. SHAP Summary Plot модели CatBoost: каждая точка – один респондент

Анализ SHAP Summary Plot показывает, что наибольший защитный эффект имеет удовлетворённость текущей деятельностью (Work_Satisfaction): высокие значения этого признака смещают предсказание в сторону более низкого уровня тревожности. Аналогичный протективный характер демонстрируют качество сна (Sleep_Hours) и частота прогулок (Outdoor_Walks). Фактором риска выступает посещение психолога (Therapy_Sessions), однако эту связь можно рассматривать как проявление обратной причинности: люди с более высокой тревожностью чаще обращаются за профессиональной помощью.

Анализ ошибок и диагностика остатков. Для оценки надёжности модели был проведён детальный анализ остатков. Гистограмма распределения остатков приближена к нормальному закону, что свидетельствует об отсутствии сильного систематического смещения.

Анализ ошибок по диапазонам тревожности выявил ограничения: для уровней Low (0–5 баллов) и Mid-Low (5–10 баллов) модель работает с высокой точностью (MAE 1.9–2.9 балла), тогда как для уровня High (>15 баллов) ошибка резко возрастает до MAE ≈ 6.2 балла.

Модель склонна недооценивать экстремально высокие уровни тревожности, сглаживая их к средним показателям выборки.



а *б*

Рисунок 2. Диагностика остатков модели CatBoost: а – гистограмма распределения остатков; б – Q-Q plot

Характер систематических отклонений наглядно иллюстрирует рис. 3, на котором представлено сравнение предсказанных и истинных значений GAD-7 на тестовой выборке.

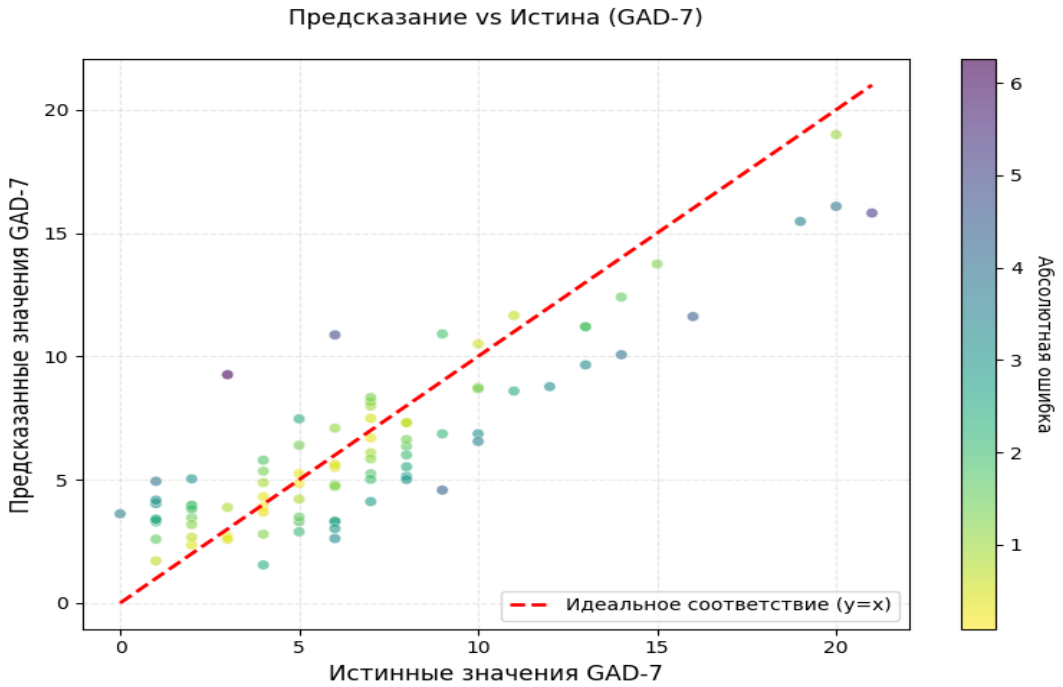


Рисунок 3. Сравнение предсказанных и истинных значений GAD-7 на тестовой выборке

Анализ рис. 3 показывает два устойчивых типа ошибок.

В диапазоне низких значений тревожности (0–4 балла) модель чаще завышает прогноз: точки располагаются выше линии идеального соответствия, что связано с эффектом регрессии к среднему при малом числе таких наблюдений. В области высоких значений (>15 баллов), напротив, модель систематически занижает тревожность, что видно по смещению точек ниже диагонали. Тёмно-синие точки с абсолютной ошибкой 5-6 баллов соответствуют ранее полученному MAE ≈ 6,2 для диапазона High. Это ограничение связано с недостаточным

числом респондентов с высокой тревожностью в обучающей выборке и указывает на необходимость целевого расширения датасета в этом сегменте.

Интерпретация результатов. Полученное значение MAE = 2,79 балла важно не только с технической, но и с клинической точки зрения. В психометрии минимально клинически значимая разница (MCID) для шкалы GAD-7 составляет примерно 3-4 балла. То, что средняя абсолютная ошибка модели не превышает этот диапазон, указывает на её потенциальную пригодность как вспомогательного инструмента для мониторинга изменений уровня тревожности во времени, особенно в исследовательских и пилотных проектах.

Коэффициент детерминации $R^2 = 0,39$ также можно считать достаточно высоким для поведенческих наук, где при использовании только опросных данных значения R^2 обычно не выходят за пределы 0,30-0,40. При этом остаётся важное ограничение: точность прогноза заметно снижается для высоких значений шкалы (более 15 баллов), что связано с малым числом респондентов с выраженной тревожностью в выборке.

Для компенсации этого дисбаланса была протестирована техника синтетического оверсэмплинга для регрессии (SMOGN), однако она не дала улучшения глобальных метрик качества на тестовой выборке. Дальнейшие исследования целесообразно ориентировать на целевое расширение выборки в сегменте высоких значений GAD-7 и более детальную оценку точности моделей в отдельных подгруппах по уровню тревожности. Это позволит точнее оценить применимость разработанного подхода для задач раннего выявления групп риска.

Заключение. В ходе исследования была разработана и протестирована модель машинного обучения для оценки уровня тревожности на основе поведенческих и социально-демографических факторов.

Сравнительный анализ показал, что на выборке ограниченного объёма наилучшие результаты демонстрирует модель CatBoost с оптимизацией гиперпараметров (MAE = 2.79, R^2 = 0.39), тогда как линейная Ridge-регрессия обеспечивает очень близкую точность при более простой интерпретации.

Полученные результаты показывают, что на основе относительно простых опросных данных возможно построить модель, пригодную для вспомогательной оценки уровня тревожности и предварительного выделения групп повышенного риска, в том числе в формате цифровых инструментов, интегрируемых в образовательные порталы или мобильные приложения.

В то же время снижение точности в диапазоне высоких значений GAD-7 подчёркивает необходимость расширения выборки за счёт респондентов с выраженной тревожностью и проведения дополнительной валидации на независимых выборках.

Перспективными направлениями дальнейшей работы являются: целевое пополнение данных респондентами с высоким уровнем тревожности, проверка модели в других возрастных и профессиональных группах, а также исследование комбинированных подходов, объединяющих опросные данные с другими источниками информации (например, пассивными цифровыми следами или физиологическими показателями) для повышения чувствительности и специфичности скрининга.

Список литературы

[1] Gray, B., Asrat, B., Brohan, E., Chowdhury, N., Dua, T., & van Ommeren, M. (2024). Management of generalized anxiety disorder and panic disorder in general health care settings: new WHO recommendations. *World psychiatry : official journal of the World Psychiatric Association (WPA)*, *23*(1), 160–161. https://doi.org/10.1002/wps.21172

[2] Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of internal medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

[3] Kliem, S., Sachser, C., Lohmann, A., Baier, D., Brähler, E., Fegert, J. M., & Gündel, H. (2025). Psychometric evaluation and community norms of the GAD-7, based on a representative German sample. *Frontiers in psychology*, *16*, 1526181. https://doi.org/10.3389/fpsyg.2025.1526181

[4] Мосолова Е.С., Алфимов А.Е., Костюкова Е.Г., Мосолов С.Н. Перспективы применения методов машинного обучения при аффективных расстройствах // Digital Diagnostics. 2025. Т. 6. №1. С. 97–115. doi: 10.17816/DD634885

[5] Respiration Rate and Volume Measurements Using Wearable Strain Sensors / M. Chu [et al.] // npj. Digital Medicine. 2019. No 2. https://doi.org/10.1038/s41746-019-0083-3.

[6] Устройство для измерения активной и емкостной составляющих импеданса биологических тканей: пат. 2196504 Рос. Федерации, МПК A 61 B 5/053 / А. В. Ефремов, Р. Р. Ибрагимов, Р. А. Манвелиадзе, В. Т. Леонтьев, К. Г. Булатецкий, Г. Г. Колонда, Е. В. Тарасов, Р. Ш. Ибрагимов; Новосиб. гос. мед. акад., № 2000117324/14. Заявл. 28.06.2000. Опубл. 20.01.2003.

[7] Программная модель системы для анализа импедансометрических характеристик биологических жидкостей / К. Е. Мешкова [и др.] // «Медэлектроника–2022. Средства медицинской электроники и новые медицинские технологии»: сб. науч. ст. XIII Междунар. науч.-техн. конф., г. Минск, 8–9 декабря 2022 г. Минск: Белор. гос. ун-т информ. и радиоэлек., 2022. С. 93–97.

[8] СКОРЫНИН А. А. ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ. ПЕДАГОГИКА И ПСИХОЛОГИЯ //ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ. ПЕДАГОГИКА И ПСИХОЛОГИЯ Учредители: Пермский государственный гуманитарно-педагогический университет. – №. 9. – С. 58-65.

[9] Toussaint A., et al. Sensitivity to change and minimal clinically important difference of the GAD-7. Journal of Affective Disorders, 2020, vol. 265, pp. 395-401.

[10] Oh C-M, Kim HY, Na HK, Cho KH and Chu MK (2019) The Effect of Anxiety and Depression on Sleep Quality of Individuals With High Risk for Insomnia: A Population-Based Study. Front. Neurol. 10:849. doi: 10.3389/fneur.2019.00849.

Авторский вклад

Липницкая Наталья Ивановна – постановка общей научной задачи исследования, формирование концепции работы и структуры статьи, контроль методологии построения и валидации моделей машинного обучения, научное редактирование текста, подготовка статьи к публикации и доработка выводов.

Миткевич Максим Андреевич – сбор и предобработка данных, разработка и реализация инструментария сбора данных (Google Forms), обучение и тестирование моделей машинного обучения, анализ предикторов.

Смертьев Владислав Валерьевич – разработка методологии, обучение и тестирование моделей машинного обучения, анализ предикторов, анализ ошибок и диагностика остатков модели, обсуждение результатов, написание статьи и оформление статьи

MACHINE LEARNING MODELS FOR PREDICTING ANXIETY LEVELS FROM BEHAVIORAL AND SOCIO-DEMOGRAPHIC FEATURES

N.I. Lipnitskaya
Senior Lecturer of the Department of Economic Informatics, BSUIR
n.i.karpovich@gmail.com

M. A. Mitkevich
Student of the Faculty of Engineering and Economics, BSUIR
mitkevich@bsuir.by

U. V. Smertsyeu
Student of the Faculty of Engineering and Economics, BSUIR
smertsyeu@bsuir.by

Abstract. The increasing prevalence of anxiety disorders requires the development of accessible screening tools. This study tests the hypothesis that it is possible to indirectly assess anxiety levels on the GAD-7 scale from behavioral and socio-demographic factors, with GAD-7 used solely as the target variable and not as part of the input questionnaire. Based on survey data from 439 respondents, several machine learning models (Ridge, Random Forest, CatBoost, LightGBM) were trained and tested. CatBoost demonstrated the best results: MAE = 2.79, R^2 = 0.39. Key protective predictors are satisfaction with the work/study environment and sleep quality; risk factors include lack of social support and negative life events.

Keywords: anxiety, mental health, risk assessment, machine learning, behavioral factors, digital psychiatry, medical informatics, GAD-7, behavioral analytics.

УДК 378.14:519.23

ОЦЕНКА ВЛИЯНИЯ ПОСЕЩАЕМОСТИ ЗАНЯТИЙ НА АКАДЕМИЧЕСКУЮ УСПЕВАЕМОСТЬ СТУДЕНТОВ С ИСПОЛЬЗОВАНИЕМ PYTHON–АНАЛИТИКИ



Е.И. Полоско
Старший преподаватель кафедры экономической информатики БГУИР
e.i.polosko@gmail.com



П.Ю. Сороко
Студент инженерно–экономического факультета специальности «Информационные системы и технологии (в экономике)» БГУИР
sorokop40@gmail.com



А.Д. Тригубович
Студент инженерно–экономического факультета специальности «Информационные системы и технологии (в экономике)» БГУИР
trigubovichanna29@gmail.com

Е.И. Полоско
Окончила Белорусский государственный университет. Область научных интересов связана с использованием методов анализа данных и современных методов машинного обучения для решения прикладных задач в экономике, образовании и управлении.

П.Ю. Сороко
Окончила ГУО «Гимназия №56 г. Минска». Область научных интересов связана с применением методов анализа данных для решения прикладных задач в экономической сфере.

А.Д. Тригубович
Окончила ГУО «Гимназия №1 г. Жодино». Область научных интересов связана с применением методов анализа данных для решения прикладных задач в экономической сфере.

Аннотация. В работе рассматривается, как посещаемость занятий и связанные с ней учебные привычки отражаются на успеваемости студентов дневной формы обучения. На основе анкетирования студентов разных курсов и факультетов анализируются частота пропусков, объём самостоятельной подготовки, способы восполнения пропущенного материала, совмещение учёбы с работой и их связь с итоговыми оценками. Для обработки данных использованы методы статистического анализа и множественная линейная регрессия в среде Python, что позволило количественно оценить вклад основных факторов в изменение среднего балла. Результаты показывают, что регулярное посещение остаётся ключевым условием успешной учёбы, а продуманная самостоятельная подготовка и грамотное восполнение пропусков помогают сгладить негативные последствия неявок. Полученные выводы могут быть полезны при разработке рекомендаций для студентов и при совершенствовании организации учебного процесса в вузе.

Ключевые слова: анализ данных, академическая успеваемость, посещаемость, самостоятельная подготовка, регрессионный анализ, анкетирование, Python.

Введение. Под посещаемостью студентов в данном исследовании понимается регулярность их присутствия на занятиях, предусмотренных учебным планом. Она отражает реальное включение обучающихся в аудиторную деятельность и является важным фактором, напрямую влияющим на качество и глубину усвоения материала.

Актуальность проблемы посещаемости обусловлена сочетанием личностных (мотивация, состояние здоровья, семейная ситуация) и организационных факторов (построение учебного процесса, профессионализм преподавателей, контроль посещаемости, дисциплина, образовательная среда). Высокий уровень организации учебного процесса задаёт общий контекст и во многом определяет результативность остальных компонентов.

Сейчас вопрос успеваемости студентов по–прежнему остаётся очень важным. Обычно отдельно изучают посещаемость, мотивацию и другие факторы, но то, как они действуют вместе, а также современные привычки и стратегии студентов исследованы ещё мало. До конца не ясно, как на результаты учёбы влияют компенсирование пропусков, совмещение учёбы с работой и разные типы учебного поведения, а также насколько оправдано распространённое мнение о вреде подработки при недостатке реальных данных.

Цель исследования заключается в том, чтобы определить, как совместно влияют основные факторы – посещаемость занятий, объём самостоятельной подготовки, способы восполнения пропусков, совмещение учёбы с работой и особенности поведения на занятиях – на академическую успеваемость студентов.

Для достижения цели используются методы регрессионного и корреляционного анализов, реализованных при помощи библиотек языка программирования Python, чтобы выявить, какие именно факторы сильнее всего связаны с учебным успехом и какие выводы можно использовать на практике студентам и администрации вуза. Эмпирические данные были получены с использованием анкетного опроса, проведённого среди студентов различных курсов и факультетов, преимущественно обучающихся по дневной форме. Анкета включала вопросы о частоте и причинах пропусков занятий, используемых стратегиях восполнения пропущенного материала, затратах времени на самостоятельную подготовку, особенностях поведения на занятиях, а также о наличии и характере дополнительной занятости.

Регрессионный анализ. На первом этапе была проведена первичная обработка данных: из анкет убрали некорректные и неполные ответы, привели форматы к единому виду и собрали показатели, с которыми дальше можно работать. Затем сделали простой (описательный) анализ и увидели, что большинство студентов редко пропускают занятия и в среднем тратят на самостоятельную подготовку около 6–14 часов в неделю. После этого, чтобы количественно оценить, как частота пропусков, объём самоподготовки и активность в восполнении пропусков связаны со средним баллом, построили модель множественной линейной регрессии. Эта модель (R^2 = 0,235) оказалась статистически значимой и показала осмысленные зависимости, которые обычно встречаются в исследованиях об образовании.

Регрессионный анализ показал, что пропуски заметно ухудшают успеваемость. Если студент регулярно пропускает занятия, его средний балл в среднем ниже примерно на 0,37, а даже при редких пропусках – примерно на 0,13 по сравнению с теми, кто почти не пропускает (рисунок 1).

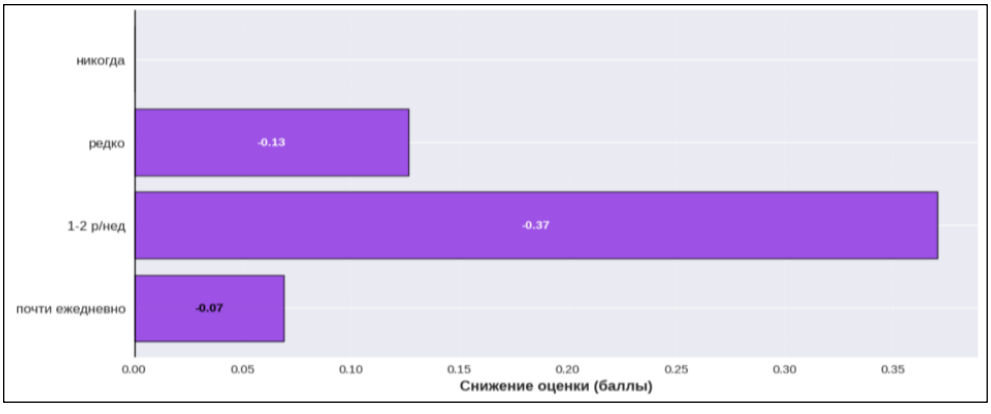


Рисунок 1. Влияние пропусков на успеваемость

Анализ показал, что связь между объёмом самостоятельной подготовки и успеваемостью носит нелинейный характер. Наибольший прирост среднего балла (около 0,62) наблюдается у студентов, уделяющих самостоятельной работе приблизительно 25–35 часов в неделю, что соответствует высокому, но реалистичному уровню нагрузки для наиболее мотивированных обучающихся.

Дальнейшее увеличение затрачиваемого времени не приводит к существенному дополнительному росту успеваемости (рисунок 2).

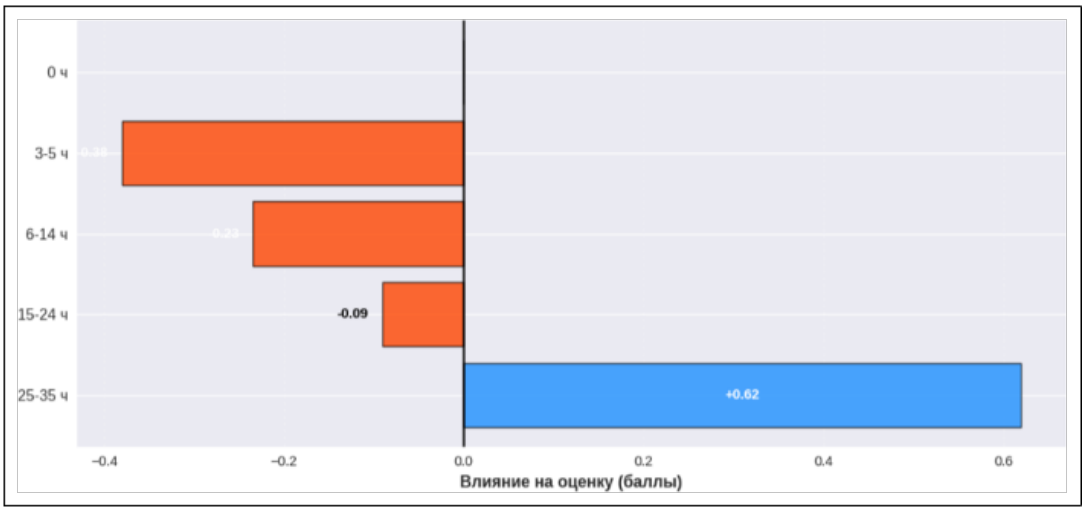


Рисунок 2. Влияние самоподготовки на успеваемость

Интегральный анализ показал, что если студент достаточно много и регулярно занимается самостоятельно и хорошо восполняет пропущенный материал, это в сумме может добавить к средней оценке до 0,83 балла и частично «перекрыть» минус от пропусков (–0,57 балла). Но такое идеальное сочетание условий встречается не очень часто.

Поэтому именно регулярное посещение занятий остаётся самым надёжным способом сохранить высокую успеваемость (рисунок 3).

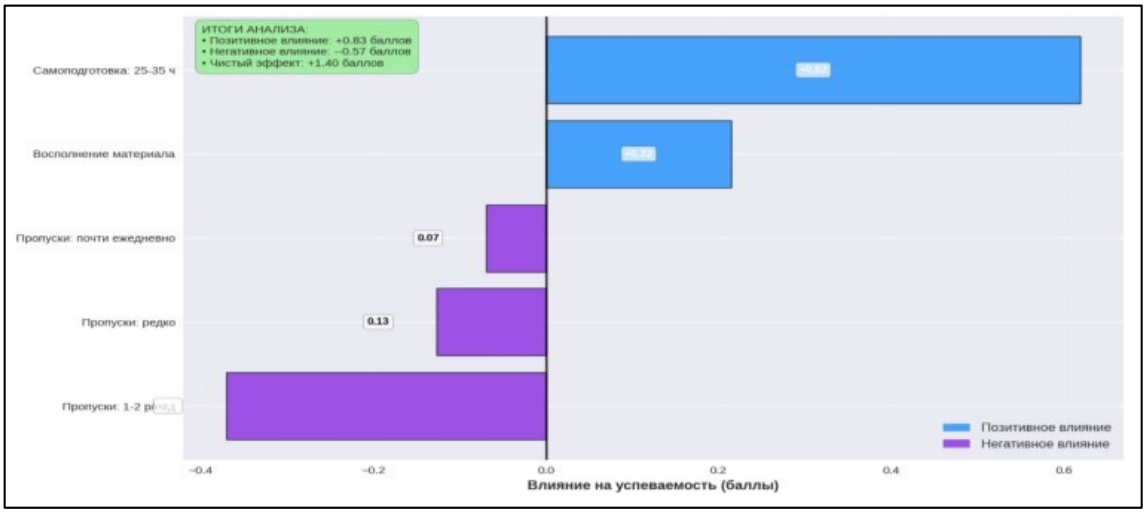


Рисунок 3. Общее влияние всех факторов на успеваемость

Полученные результаты можно использовать для практических советов студентам.

Главное – как можно меньше пропускать занятия и регулярно заниматься самостоятельно не менее 15–20 часов в неделю.

Для тех, кто сильно мотивирован, увеличение самоподготовки до 25–35 часов в неделю даёт заметный рост успеваемости.

Исследование показывает, что регулярное посещение всё равно остаётся основным фактором успеха, но если пропуски всё же возникают, их негативный эффект частично можно компенсировать за счёт продуманной самостоятельной работы и своевременного восполнения материала.

Корреляционный анализ. Корреляционный анализ показал, что между основными показателями учёбы есть заметные связи.

Сильнее всего с итоговой оценкой связаны пропуски занятий (|r| = 0,64): чем больше студент пропускает, тем хуже его успеваемость.

Довольно сильно с оценкой связаны и объём самостоятельной подготовки (0,51), и то, насколько активно студент восполняет пропущенный материал (0,49), то есть эти факторы тоже очень важны для результатов.

Связи между пропусками и восполнением материала, а также между пропусками и подработкой умеренные (0,35–0,40), а остальные пары показателей почти не коррелируют.

В целом сильнее всего влияют факторы, непосредственно связанные с учёбой (посещаемость, самоподготовка, восполнение), тогда как стиль поведения на занятиях и подработка влияют на успешность гораздо слабее (рисунок 4).

Анализ причин пропусков показал, что чаще всего студенты не приходят на занятия из–за участия в университетских мероприятиях (39,9%) и из–за болезней (28,3%) – вместе это больше половины всех пропусков.

На третьем месте – отсутствие мотивации или интереса к предмету (19,7%), что говорит о том, что нужно пересматривать содержание курсов и формат занятий.

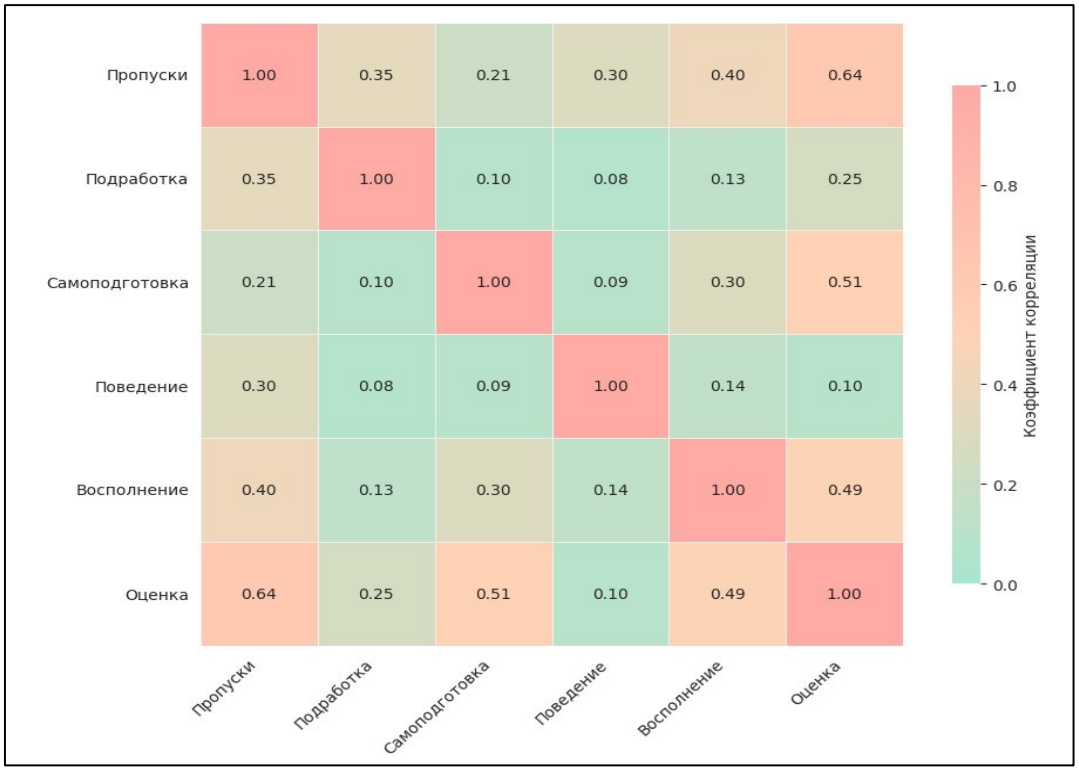


Рисунок 4. Корреляционный анализ учебного поведения студентов

Реже студенты пропускают по семейным причинам или из–за подработки. Поэтому, чтобы повысить посещаемость, важно усиливать учебную мотивацию, учитывать состояние студентов и лучше организовывать учебный процесс.

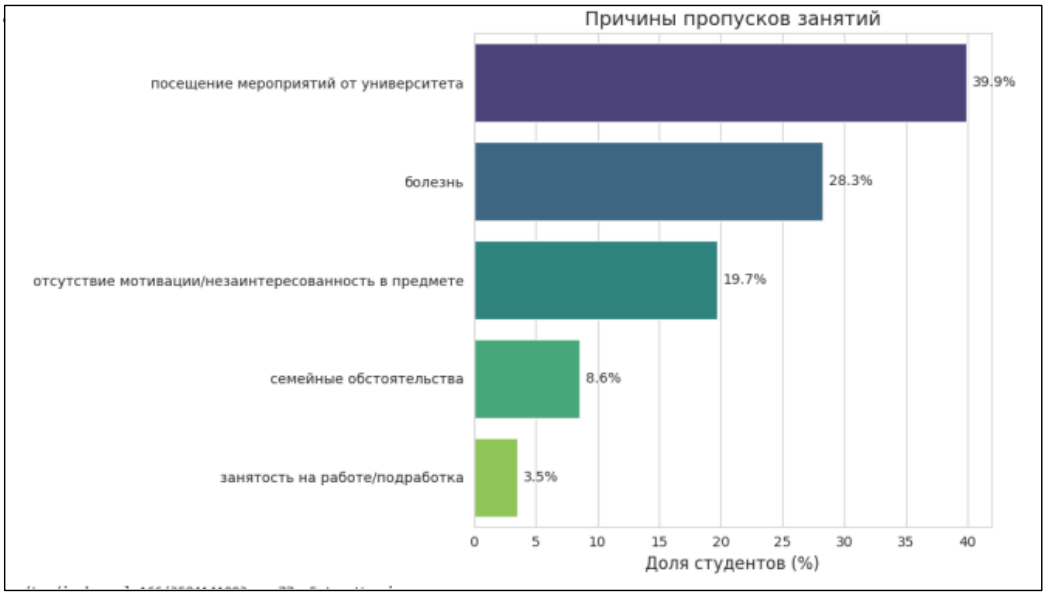


Рисунок 5. Средние оценки по группам пропусков

Анализ показал, что студенты используют три основных способа восполнения пропущенного материала, причём их частота применения примерно одинакова.

Наиболее распространённым остаётся обращение к материалам одногруппников (36,0%). Использование электронных ресурсов (32,2%) и самостоятельный поиск информации (31,8%) отражают активное обращение к цифровым технологиям и развитие навыков самообучения. Такое распределение показывает, что современные студенты комбинируют разные стратегии для компенсации пропусков (рисунок 6).

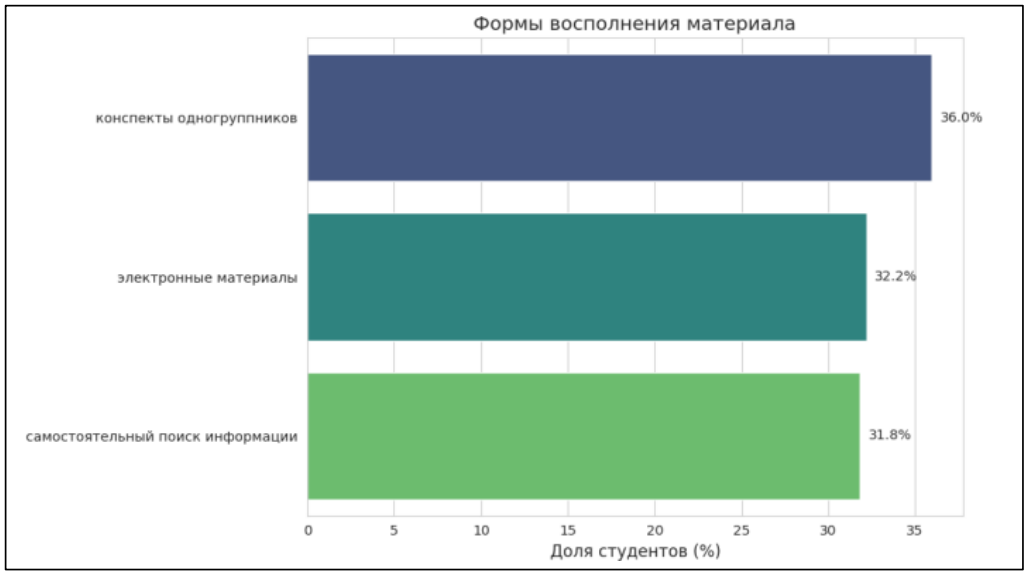


Рисунок 6. Формы восполнения материала

Поведение студентов на занятиях сильно различается. Около трети опрошенных (33,6%) активно включаются в работу в основном тогда, когда им действительно интересна тема. Примерно четверть (27,4%) ведут себя пассивно: в основном слушают и делают записи. Похожая доля студентов (25,7%) регулярно участвует в обсуждениях и выполнении заданий. При этом 12,4% почти не вовлекаются в учебный процесс и часто отвлекаются, а совсем небольшая группа (0,9%) стабильно очень активна и часто проявляет инициативу на занятиях.

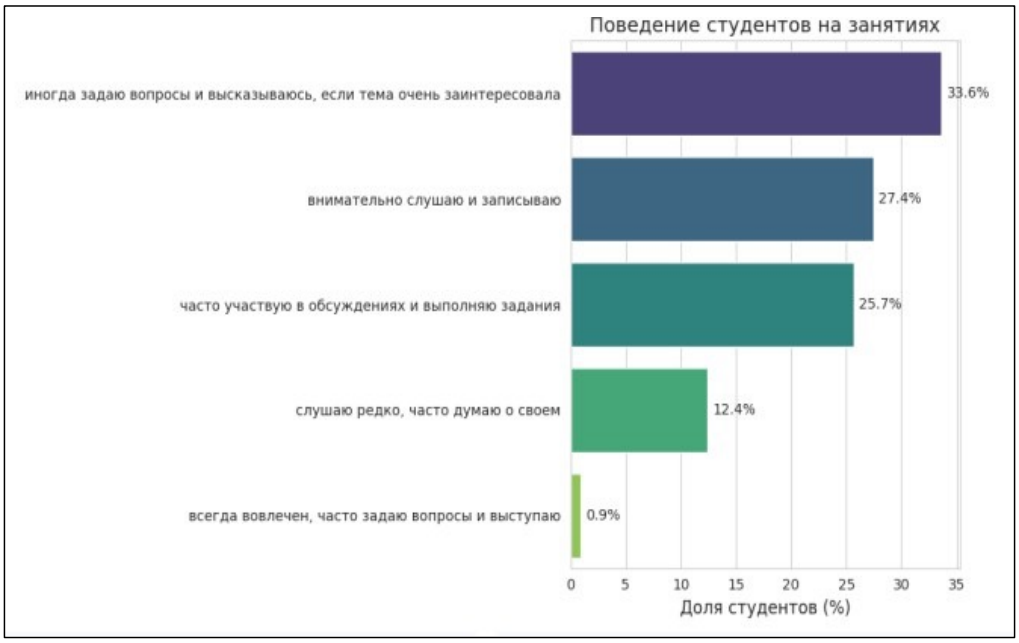


Рисунок 7. Поведение студентов на занятиях

Полученные результаты указывают на преобладание пассивных и ситуативных форм учебного поведения, что подчёркивает необходимость поиска новых подходов к повышению активного участия студентов в образовательном процессе.

Заключение. Исследование показало, что для студентов дневной формы обучения регулярное посещение занятий является основным условием высокой успеваемости. Пропуски оказывают на итоговую оценку наиболее сильное негативное влияние, и даже интенсивная самостоятельная подготовка не позволяет полностью его компенсировать, хотя продуманная самоподготовка и своевременное восполнение пропусков частично снижают отрицательный эффект и способствуют увеличению среднего балла. Проведённый анализ подтвердил, что наибольшая связь с успеваемостью наблюдается у показателя посещаемости, тогда как объём самостоятельной работы и активность в восполнении материала также оказывают заметное, но менее выраженное влияние. Поведение на занятиях и совмещение учёбы с подработкой демонстрируют более слабые ассоциации с оценками, что говорит об опосредованном характере их воздействия. Выявлено, что основными причинами пропусков являются участие в мероприятиях университета и заболевания, тогда как отсутствие мотивации занимает третью позицию, указывая на необходимость совершенствования содержания дисциплин и форматов проведения занятий. В стратегиях восполнения пропусков доминируют обращение к помощи одногруппников, использование электронных ресурсов и самостоятельный поиск информации, при этом у заметной части студентов сохраняются пассивные и ситуативные модели учебного поведения.

Практические рекомендации: Для студентов приоритетными являются сокращение числа пропусков учебных занятий, поддержание регулярной самостоятельной подготовки в объёме не менее 15–20 часов в неделю (с возможным увеличением до 25–35 часов для

обучающихся с высоким уровнем мотивации), а также своевременное и систематическое восполнение пропущенного материала.

Для администрации и профессорско-преподавательского состава вуза целесообразно усиление работы по развитию учебной мотивации студентов, учёт их участия в образовательных, научных и внеучебных мероприятиях, совершенствование форм проведения занятий и расширение использования инструментов, стимулирующих активное участие обучающихся в образовательном процессе.

Список литературы

[1] Миселимян Т.Л., Метелица Н.Т. Влияние посещаемости занятий учащимися на качество образовательного процесса // Успехи современного естествознания. 2005. №5. С. 76–79.

[2] Гурский Е.П., Шебзухов Ю.А., Чернин Р.И. Управление посещаемостью – важное направление образовательного и воспитательного процесса // Вестник БелГУТа: наука и транспорт. 2016. С. 76– 77.

[3] Лызь Н.А., Рощина Н.В., Фрумин И.Д. Работающие студенты: образовательная успешность и субъективное благополучие // Высшее образование в России. 2023. Т. 32. №1. С. 9–24.

[4] Аникина О.В., Денисова Т.А., Пащенко О.Ю. Система мониторинга успеваемости студентов педагогического вуза в контексте онлайн–образования // Образование и наука. 2022. Т. 24. №10. С. 120–147.

[5] Потанина А.В., Цыганов И.Ю., Моросанова В.И. Типологические особенности ресурсов успеваемости студентов: регуляторные и личностные предикторы // Национальный психологический журнал. 2023. №4. С. 188–205.

[6] Помян Белоконь Н.В., и др. Прогноз результатов успеваемости студентов вуза на основе методов машинного обучения // Вестник Вятского государственного университета. 2020. №4. С.9–17.

Авторский вклад

Полоско Екатерина Ивановна – постановка и обоснование исследовательской задачи, разработка концепции и методологии исследования, консультирование по выбору и реализации методов анализа данных, критический анализ и интерпретация результатов, подготовка статьи к публикации.

Сороко Полина Юрьевна – сбор, формирование и обработка эмпирических данных, подготовка выборки к анализу (очистка, кодирование и агрегирование показателей), выполнение регрессионного и корреляционного анализа, интерпретация полученных результатов.

Тригубович Анна Дмитриевна – сбор, формирование и обработка эмпирических данных, подготовка выборки к анализу (очистка, кодирование и агрегирование показателей), выполнение регрессионного и корреляционного анализа, интерпретация полученных результатов.

ASSESSMENT OF THE IMPACT OF CLASS ATTENDANCE ON STUDENTS' ACADEMIC PERFORMANCE USING PYTHON–BASED ANALYTICS

E.I. Polosko
Senior Lecturer, Department of Economic Informatics, BSUIR
e.i.polosko@gmail.com

P.Y. Soroko
Student of the Faculty of Engineering and Economics, Department of Information Systems and Technologies (in Economics), BSUIR

A.D. Trigubovich
Student of the Faculty of Engineering and Economics, Department of Information Systems and Technologies (in Economics), BSUIR

Abstract. The paper examines how class attendance and related study habits affect the academic performance of full–time students. Based on a survey of students from different years of study and faculties, the analysis focuses on the frequency of absences, the amount of independent study, ways of making up missed material, the combination of study and work, and their relationship with final grades. Statistical analysis methods and multiple linear regression in Python are used to quantify the contribution of the main factors to changes in the average grade. The results show that regular attendance remains a key condition for successful studies, while well–planned independent work and effective ways of making up missed classes help mitigate the negative consequences of absences. The findings may be useful for developing recommendations for students and for improving the organization of the educational process at the university.

Keywords: data analysis, academic performance, attendance, independent study, regression analysis, survey, Python.

УДК 004.932.2

РАЗРАБОТКА ЛЕГКОВЕСНОГО АЛГОРИТМА ДЕТЕКЦИИ «ЖИВОСТИ» ЛИЦА (LIVENESS DETECTION) ДЛЯ СИСТЕМ ВЕБ-ИДЕНТИФИКАЦИИ



И.В. Андриалович
Заместитель декана по ИВР факультета компьютерного проектирования БГУИР, аспирант кафедры ИПиЭ
andryinna@bsuir.by



И.П. Галяк
Студент кафедры проектирования информационно-компьютерных систем ФКП БГУИР
haliak.ivanba@gmail.com



Е.А. Касперец
Студент кафедры инженерной психологии и эргономики БГУИР
kasperetsyevgeny@gmail.com

И.В. Андриалович
Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с исследованием проблем психологического выгорания профессорско-педагогического состава учреждений высшего образования.

И.П. Галяк
Студент кафедры проектирования информационно-компьютерных систем БГУИР. Область профессиональных интересов / исследований: бизнес-системный анализ, применение методов анализа данных и технологий искусственного интеллекта для оптимизации бизнес-процессов, и поддержки принятия решений.

Е.А. Касперец
Студент кафедры инженерной психологии и эргономики БГУИР. Область профессиональных интересов / исследований: языки программирования, искусственный интеллект, исследование инженерной психологии.

Аннотация. В статье рассматривается разработка легковесного алгоритма проверки «живости» лица в режиме реального времени, предназначенного для противодействия подделке биометрических данных в системах веб-идентификации. Предложенный метод базируется на анализе антропометрических точек лица с использованием алгоритмов компьютерного зрения. Данный подход позволяет снизить требованиях к аппаратным ресурсам, что делает его эффективной альтернативой ресурсоемким нейросетевым решениям.

Ключевые слова: детекция «живости», биометрическая аутентификация, компьютерное зрение, подделка биометрических данных, веб-идентификация, антропометрические точки лица.

Введение. В условиях глобальной цифровизации и перехода к удаленному обслуживанию в финансовом, государственном и корпоративном секторах, биометрическая аутентификация по лицу стала неотъемлемым компонентом комплексных систем безопасности. Однако рост вычислительных мощностей и совершенствование методов синтеза медиаконтента привели к резкому увеличению числа атак на подобные системы. Согласно действующему международному стандарту *ISO*/*IEC* 30107-1:2023, подобные угрозы классифицируются как предоставление системе биометрического захвата данных, целью которого является вмешательство в работу биометрической системы [1].

В рамках терминологии стандарта атаки могут осуществляться двумя типами субъектов: биометрическим самозванцем, который стремится быть распознанным как конкретное другое лицо, известное системе, или биометрическим маскировщиком, чья цель – избежать

распознавания или скрыть свои собственные характеристики без намерения имитировать кого-либо.

Несмотря на различие в мотивах злоумышленников, угроза в обоих случаях зачастую сводится к атаке на уровне предъявления (*Presentation Attack*), в частности непосредственно на сенсор захвата данных. В данной работе рассматривается противодействие этому типу атак. Ключевой контрмерой в данном случае выступает проверка «живости» (*liveness*). Под этим термином понимается измерение и анализ анатомических характеристик, а также непроизвольных или произвольных реакций с целью определения того, происходит ли сбор биометрических данных от живого субъекта, присутствующего в точке захвата. Методы проверки подлинности, основанные на этом принципе, являются подмножеством механизмов обнаружения атак (*Presentation Attack Detection – PAD*).

Современные реализации *PAD*-систем в большинстве случаев опираются на использование нейронных сетей для выявления артефактов подделанной биометрии. Несмотря на высокую точность, такие решения обладают недостатками в контексте веб-систем.

Во-первых, серверная обработка видеопотока вносит существенную задержку (*latency*), которая с учетом сетевых издержек может достигать 500–1000 мс, что является неприемлемым показателем для интерактивных пользовательских интерфейсов. Во-вторых, экономическая эффективность подобных решений при масштабировании крайне низка: производительность современных моделей на стандартных серверных *CPU* зачастую ограничена скоростью 5–15 кадров в секунду. Это вынуждает компании инвестировать в дорогостоящие *GPU*-ускорители для поддержания приемлемой производительности при обслуживании большого потока одновременных сессий верификации [2].

Целью данной работы является разработка метода активного обнаружения атак для веб-интерфейсов, лишенного указанных недостатков. В основу предлагаемого решения положен интерактивный анализ точек лица пользователя в режиме реального времени. Система генерирует случайную последовательность заданий (повороты головы, моргание, изменение мимики) – и оценивает корректность их выполнения с помощью легковесного математического аппарата компьютерного зрения, что позволяет проводить проверку на серверах без специализированного оборудования.

Теоретический базис и используемый стек технологий. В основе разработанного алгоритма лежит комбинация классических методов компьютерного зрения, что обеспечивает повышение производительности на стандартном *CPU* без привлечения специализированных ускорителей. Сам же физический процесс детекции представляет собой машину проверки состояний, которая в режиме реального времени выполняет потоковую обработку видеоданных. Для каждого кадра последовательно производятся следующие операции: локализация лица, определение координат его антропометрических точек, расчет набора числовых дескрипторов (коэффициентов открытости глаз, рта и углов поворота головы) и сравнение этих дескрипторов с пороговыми значениями, соответствующими требуемому на данном этапе действию.

Ключевым этапом всего процесса является выделение антропометрических точек лица (*Facial Landmarks*). Для решения этой задачи выбрана библиотека *Dlib* и реализованный в ней метод, использующий предварительно обученную модель, которая захватывает положение ключевых точек на основе анализа изображения. Данный подход позволяет определять координаты 68 ключевых точек, очерчивающих контуры челюсти, бровей, носа, глаз и губ (рисунок 1). Выбор в пользу *Dlib* обусловлен фокусом на производительность на центральном процессоре и достаточной точностью для решения поставленных задач [3, 4].

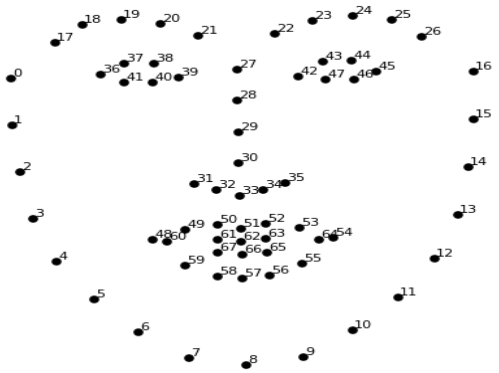


Рисунок 1. 68 точек лица, обнаруживаемых библиотекой *dlib*

После получения массива координат 68 точек, система переходит к расчету набора числовых дескрипторов, отражающих текущее состояние мимики и положения головы пользователя. Для количественной оценки моргания и закрытия глаз применяется коэффициент соотношения сторон глаза (*Eye Aspect Ratio – EAR*) [5]. Данный коэффициент численно выражает отношение суммы расстояний между вертикальными точками-ориентирами глаза к расстоянию между его горизонтальными точками-ориентирами. Такое соотношение остается стабильным при открытом глазе и стремительно снижается к нулю при его закрытии, что делает его индикатором моргания. По аналогии с *EAR*, для детекции улыбки введен коэффициент соотношения сторон рта (*Mouth Aspect Ratio – MAR*), который вычисляется схожим образом, оценивая степень открытия рта относительно его ширины.

Для выполнения наиболее сложной части проверки – отслеживания поворотов головы – используется метод оценки пространственного положения объекта (*Head Pose Estimation*). Данная задача сводится к классической проблеме *Perspective-n-Point* (*PnP*) и решается с помощью одноименного алгоритма (*SolvePnP*) из библиотеки *OpenCV* [6]. Метод находит ориентацию объекта в 3D-пространстве путем сопоставления набора его эталонных трехмерных точек (в данном случае – обобщенная 3D-модель лица) с их двумерными проекциями на изображении. В результате алгоритм вычисляет векторы вращения и смещения, которые затем преобразуются в углы Эйлера: рыскание (*yaw*), тангаж (*pitch*) и крен (*roll*). Анализ угла рыскания (*yaw*) позволяет системе определить, повернул ли пользователь голову влево или вправо в соответствии с полученной инструкцией.

Программная реализация и архитектура системы. Архитектура разрабатываемой системы базируется на клиент-серверном взаимодействии, оптимизированном для потоковой обработки видеоданных в веб-интерфейсе. Согласно представленной диаграмме (рисунок 2) последовательности обработки кадра, цикл начинается на стороне клиента (браузера), где с помощью интерфейса *getUserMedia* осуществляется захват видеопотока с веб-камеры пользователя. Для минимизации нагрузки на канал связи и обеспечения совместимости с протоколом *HTTP*, каждый значимый кадр подвергается преобразованию в формат *Base64*. Передача данных на сервер, реализованный на базе микрофреймворка *Flask*, осуществляется посредством асинхронных *POST*-запросов. Такой подход позволяет реализовать концепцию «тонкого клиента», где все ресурсоемкие вычисления, связанные с анализом биометрии, выносятся на серверную сторону, обеспечивая стабильную работу системы даже на мобильных устройствах с ограниченной вычислительной мощностью.

Центральным элементом логики работы сервера выступает модуль «*Liveness State Machine*», управляющий последовательностью этапов проверки. Процесс верификации разделен на ряд последовательных состояний: от первичной локализации лица до анализа специфических команд (повороты головы, моргание, мимические реакции). Переход к каждому следующему этапу возможен только при условии успешного выполнения текущего

задания, что контролируется путем непрерывного сопоставления расчетных дескрипторов (*EAR, MAR, Yaw*) с эталонными пороговыми значениями. Подобный подход исключает возможность подмены биометрических данных статическими изображениями или заранее записанными видеофрагментами, так как система требует динамической реакции в ответ на случайно сгенерированные инструкции.

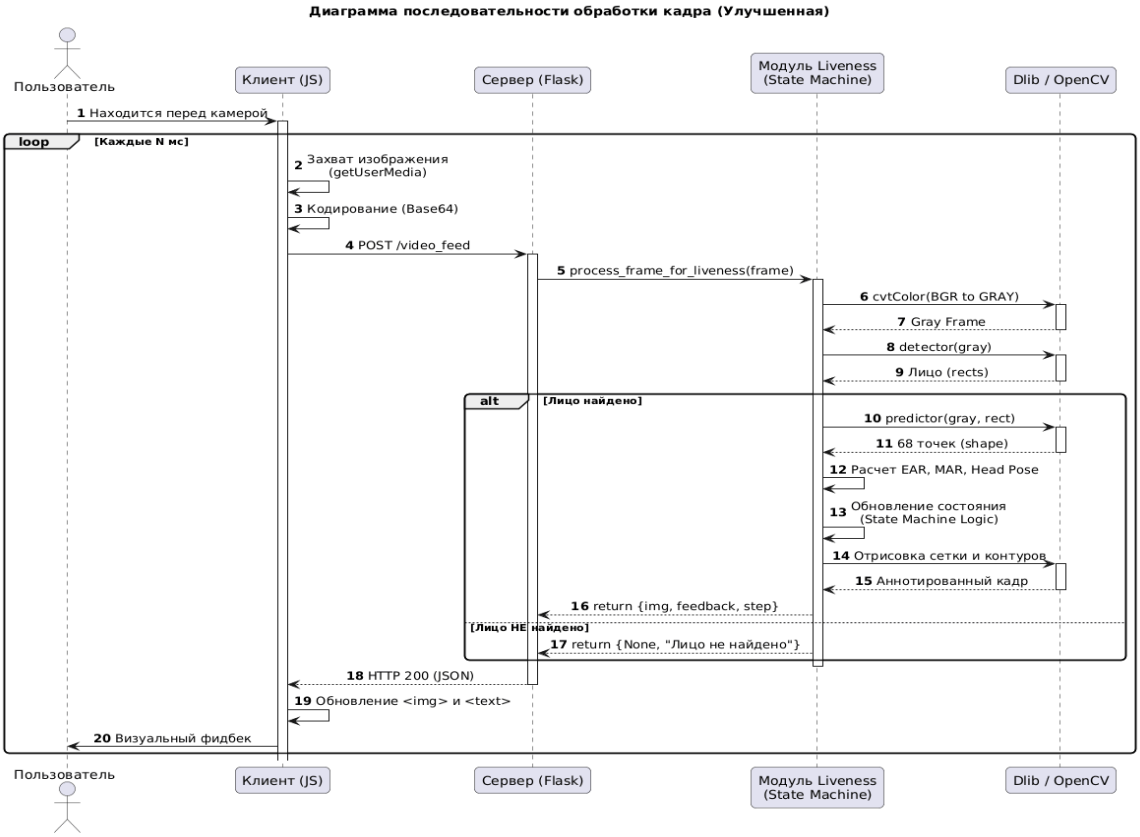


Рисунок 2. Диаграмма последовательности обработки кадра

Для удобства взаимодействия с клиентом предложена подсистема визуальной обратной связи. В процессе анализа кадра модуль захвата антропометрических точек не только вычисляет математические показатели, но и формирует аннотированное изображение. На исходный видеопоток в режиме реального времени накладывается адаптивная сетка из 68 ключевых точек и контуры основных черт лица. Визуализация процесса детекции служит инструментом взаимодействия с пользователем: она позволяет субъекту корректировать положение головы относительно камеры и освещенности, а также подтверждает, что система корректно распознает его биометрические характеристики. После завершения обработки кадра сервер возвращает клиенту *JSON*-объект, содержащий текущий статус проверки, текстовую инструкцию и модифицированный кадр с наложенной графикой для отображения в браузере.

Результаты экспериментов и оценка эффективности. Для тестирования разработанного алгоритма создан веб-интерфейс, позволяющий провести экспериментальную оценку как эффективности системы, так и удобства ее использования (рисунок 3). Пользовательский интерфейс спроектирован с учетом минимизации когнитивной нагрузки на пользователя. Центральную область занимает видеопоток с веб-камеры, где вспомогательный овальный контур помогает пользователю правильно позиционировать лицо. В правой части экрана расположен пошаговый индикатор этапов проверки, который в реальном времени отображает текущий прогресс. Под видеопотоком выводится текстовая обратная связь,

содержащая как инструкции для следующего действия (например, «Поверните голову ВЛЕВО»), так и служебные сообщения (например, «Лицо не найдено»). Такая компоновка предоставляет возможность взаимодействия и позволяет пользователю получать обратную связь об эффективности своих действий в рамках процесса верификации.

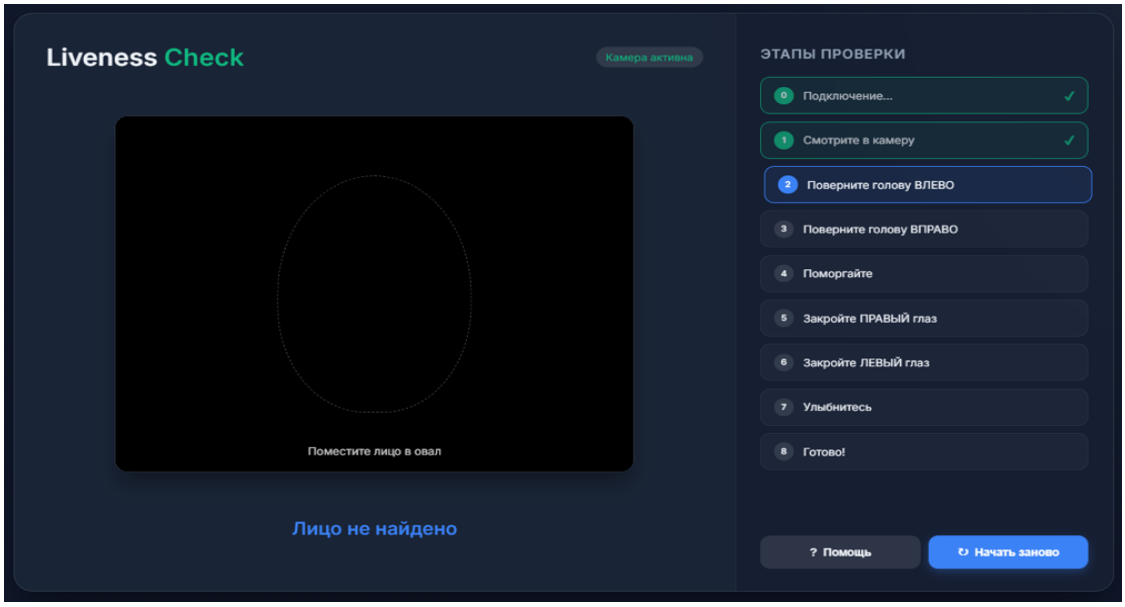


Рисунок 3. Пользовательский интерфейс системы верификации

В ходе апробации подобраны пороговые значения для каждого из контролируемых биометрических параметров. Задачей на данном этапе являлся поиск баланса между устойчивостью системы к атакам на предъявление и минимизацией ложных отказов в обслуживании (*False Rejection Rate, FRR*). Оптимальные значения, используемые в финальной версии системы, приведены в таблице 1.

Таблица 1. Пороговые значения параметров для машины состояний

Этап проверки	Контролируемый параметр	Пороговое значение
Фиксация в центре	Угол рыскания (*Yaw*)	±30 градусов
Поворот влево/вправо	Угол рыскания (*Yaw*)	> +25 / < -25 градусов
Моргание	Коэффициент *EAR*	< 0.24
Закрытие одного глаза	Коэффициент *EAR* (закрытый)	< 0.24
Улыбка	Коэффициент *MAR*	> 0.40

Тестирование подтвердило эффективность разработанного подхода по двум ключевым направлениям: безопасность и производительность. В ходе тестирования проверена устойчивость системы к основным типам 2D-атак на предъявление: алгоритм успешно идентифицировал и блокировал попытки прохождения верификации с использованием статичных артефактов (распечатанные фотографии) и динамических артефактов (воспроизведение видеозаписей на экранах мобильных устройств). Одновременно с этим, оценка производительности показала, что средняя латентность обработки одного кадра на стандартном серверном *CPU* (*Intel Core i5*) находится в диапазоне 30–40 мс. Такой низкий показатель позволяет поддерживать до 25 одновременных сессий верификации в реальном

времени на одном процессорном ядре, что подтверждает высокую архитектурную эффективность для высоконагруженных веб-сервисов.

Заключение. В ходе исследования разработан и программно реализован легковесный алгоритм активной проверки «живости» (*Active Liveness Detection*), основанный на комбинации методов компьютерного зрения. Предложенный метод, использующий машину состояний для анализа динамики антропометрических точек лица (*EAR, MAR, Head Pose*), продемонстрировал свою эффективность в противодействии атакам с использованием фото и видео. Проведенные эксперименты подтвердили, что система обладает достаточно высокой производительностью на стандартном серверном оборудовании, что делает ее экономически выгодным и легко масштабируемым решением. Работа доказывает, что классические алгоритмы компьютерного зрения остаются актуальным и конкурентоспособным инструментом для решения задач биометрической безопасности, предлагая практическую альтернативу глубокому обучению. В качестве направлений для дальнейшего развития можно выделить интеграцию в систему дополнительных пассивных проверок (например, анализ текстуры кожи) для противодействия более сложным 3D-атакам, а также адаптивную настройку пороговых значений в зависимости от условий освещения.

Список литературы

[1] ISO/IEC 30107-1:2023. Information technology – Biometric presentation attack detection – Part 1: Framework. – Geneva: International Organization for Standardization, 2023. – 23 p.

[2] Ren Y. Performance Analysis of Deep Learning Workloads on Leading-edge Systems. – Upton: Brookhaven National Laboratory, 2020. – 15 p.

[3] Amos B. The 68 landmarks detected by dlib library. – Режим доступа: https://www.researchgate.net/figure/The-68-landmarks-detected-by-dlib-library-This-image-was-created-by-Brandon-Amos-of-CMU_fig2_329392737 – Дата доступа: 07.03.2026.

[4] Dlib C++ Library. – Режим доступа: http://dlib.net/ – Дата доступа: 07.03.2026.

[5] Liveness Detection in Face Recognition: A Review. – Режим доступа: https://pmc.ncbi.nlm.nih.gov/articles/PMC9044337/ – Дата доступа: 07.03.2026.

[6] OpenCV Documentation: Camera Calibration and 3D Reconstruction (solvePnP). – Режим доступа: https://docs.opencv.org/4.x/d5/d1f/calib3d_solvePnP.html – Дата доступа: 07.03.2026.

Авторский вклад

Андриалович Инна Владимировна – руководство и постановка задачи исследования, формирование структуры статьи.

Галяк Иван Павлович – разработка алгоритма детекции «живости» лица, проектирование математической модели анализа антропометрических дескрипторов (EAR, MAR, Head Pose Estimation), программная реализация серверной логики системы на базе Flask и OpenCV, интеграция алгоритмов машинного зрения в решение.

Касперец Евгений Александрович – проектирование пользовательского интерфейса и архитектуры взаимодействия клиент-серверной системы, разработка клиентской части приложения, проведение экспериментального тестирования, анализ корректности работы системы при различных сценариях использования.

EVALUATION OF THE EFFECTIVENESS OF GENERATIVE POLICY IN OPTIMIZING DESIGN DECISIONS

I.V. Andryalovich
Deputy Dean of the Faculty of Computer Design of BSUIR, postgraduate student of the Department of IP&E
andryinna@bsuir.by

I.P. Haliak
BSUIR student, Department of Information Computer Systems
haliak.ivanba@gmail.com

Y.A. Kaspiarets
BSUIR student, Department of Engineering Psychology and Ergonomics
kasperetsyevgeny@gmail.com

Abstract. The article explores the development of a lightweight real-time face liveness detection algorithm designed to counter biometric data spoofing in web-based authentication systems. The proposed method is based on analyzing facial anthropometric landmark dynamics using classical computer vision algorithms. This approach significantly reduces hardware resource requirements, offering an efficient alternative to computationally expensive deep learning-based solutions.

Keywords: liveness detection, biometric authentication, computer vision, biometric spoofing, web-based authentication, facial landmarks.

УДК 004.94

СИМБИОЗ НЕВИДИМЫХ СЕТЕЙ: ТЕХНОЛОГИИ LORAWAN И BIG DATA В АРХИТЕКТУРЕ ЦИФРОВОГО БУДУЩЕГО



К.А.Абраев
Преподаватель кафедры «Радиосвязь и радиотехнические системы» института Телекоммуникаций и информатики Туркменистана
kuvvatabrayev@gmail.com

К.А.Абраев
Окончил институт Телекоммуникаций и информатики Туркменистана. Работает в институте Телекоммуникаций и информатики Туркменистана.

Аннотация. Мы живем в эпоху, когда неодушевленные предметы обретают «голос». Датчики в почве на полях, сенсоры на опорах линий электропередач, счетчики в подвалах городских зданий – все они генерируют непрерывный шепот данных. Однако главная проблема цифровизации долгое время заключалась в связи: как передать сигнал от тысячи устройств, работающих от одной батарейки годами, на расстояние в десятки километров? Решение пришло с технологией LoRaWAN. Но сама по себе передача данных – лишь половина дела. Настоящая магия начинается там, где этот колоссальный поток информации встречается с мощью Big Data и продвинутой аналитики. Этот симбиоз создает «интеллектуальный слой» реальности, позволяя нам управлять целыми городами и отраслями с математической точностью.сторонних провайдеров, а также расскажем о наших продуктах, поддерживающих LoRaWAN.

Ключевые слова: LoRaWAN, базовая станция, абонентское устройство, Big Data.

Введение. Современные роботизированные системы функционируют в средах с высокой динамикой и неопределённостью.

В таких условиях использование BIG DATA обеспечивает своевременную корректировку управляющих параметров на основе анализа крупных потоков сенсорной информации.

Это согласуется с выводами, указывающими на важность статистической обработки больших наборов данных для минимизации ошибок позиционирования [1].

Основная часть. Чтобы понять масштаб оптимизации, которую несет эта связка, необходимо осознать уникальность протокола LoRaWAN (Long Range Wide Area Network).

В отличие от Wi-Fi или 5G, эта сеть не предназначена для передачи видео или тяжелого контента. Ее задача – трансляция крошечных пакетов данных на огромные расстояния при минимальном энергопотреблении.

Представьте себе датчик, который зарыт в землю на глубину полуметра и раз в час сообщает о влажности почвы.

На одной батарейке он может проработать до десяти лет. Один шлюз (Gateway) LoRaWAN способен собирать сигналы с тысяч таких устройств в радиусе до 15 километров. Это создает идеальные условия для формирования «озер данных» (Data Lakes) там, где раньше сбор информации был физически невозможен или экономически невыгоден. LoRaWAN

становится «руками» цифровой системы, дотягиваясь до самых удаленных уголков физического мира.

Данные, поступающие из сетей LoRaWAN, специфичны. Это миллионы коротких сообщений, содержащих идентификаторы, временные метки и показатели датчиков.

По отдельности эти цифры малоинформативны. Однако, когда они объединяются в системы Big Data, возникает кумулятивный эффект.

Процесс оптимизации начинается на этапе агрегации. Продвинутая аналитика позволяет очищать этот поток от шумов, восстанавливать пропущенные из-за помех пакеты и выстраивать временные ряды.

Здесь Big Data выступает в роли линзы, которая фокусирует разрозненные сигналы в четкую картину.

Мы получаем возможность видеть не просто состояние одного объекта, а динамику всей системы в целом – будь то потребление воды в мегаполисе или состояние сотен километров железнодорожного полотна. Наиболее яркое применение связки LoRaWAN и Big Data находит в концепции Smart City.

Традиционные городские службы работают реактивно: ремонтная бригада выезжает, когда трубу уже прорвало. Аналитика больших данных меняет этот подход. Умные счетчики и датчики давления, объединенные сетью LoRaWAN, передают информацию в единый аналитический центр. Алгоритмы машинного обучения анализируют корреляции между давлением в трубах, температурой почвы и временем суток.

Система способна предсказать вероятность аварии за несколько дней до ее возникновения. Оптимизация городского бюджета здесь достигает колоссальных масштабов: превентивный ремонт обходится в десятки раз дешевле, чем ликвидация последствий катастрофы.

Город начинает «чувствовать» свои проблемы еще до того, как их заметят жители.

В индустриальном секторе LoRaWAN решает проблему «слепых зон» на огромных предприятиях, карьерах или нефтепромыслах. Установка датчиков вибрации, температуры и шума на критически важном оборудовании позволяет собирать данные без прокладки дорогостоящих кабелей. Здесь Big Data реализует стратегию предиктивного обслуживания (Predictive Maintenance).

Аналитика сравнивает текущие показатели станков с эталонными моделями и данными о прошлых поломках. Если спектральный анализ вибрации подшипника, переданный через LoRaWAN, начинает отклоняться от нормы на доли процента, система автоматически формирует заказ на запчасти.

Бизнес-решение принимается на основе цифр, исключая человеческий фактор и предотвращая многомиллионные убытки от простоя производства.

Сельское хозяйство – это отрасль, максимально зависящая от внешних факторов. Сети LoRaWAN позволяют покрыть связью тысячи гектаров полей.

Датчики влажности, освещенности и кислотности почвы передают данные, которые в системе Big Data накладываются на спутниковые снимки и метеорологические прогнозы.

Оптимизация здесь проявляется в «точном земледелии». Вместо того чтобы заливать всё поле удобрениями, фермер получает карту, указывающую, какому конкретному участку в 10 квадратных метров нужен полив или подкормка.

Аналитика больших данных рассчитывает идеальное время для посева и сбора урожая, исходя из накопленной статистики за десятилетия. Это не просто экономия ресурсов – это путь к продовольственной безопасности и экологической устойчивости.

Развертывание глобальных сетей мониторинга на базе LoRaWAN поднимает вопросы кибербезопасности. Поскольку данные передаются по радиоканалу, защита информации становится приоритетом.

Современные протоколы используют двойное шифрование, но реальная защита обеспечивается на уровне Big Data аналитики.

Системы обнаружения аномалий способны выявить попытку подмены данных или взлома датчика, анализируя нетипичное поведение устройства.

С другой стороны, возникает вопрос «цифровой прозрачности». Когда каждое движение груза или состояние каждого узла системы мониторится в реальном времени, бизнес становится полностью прозрачным. Это требует новых этических стандартов управления, где данные используются для развития, а не для необоснованного давления на исполнителей.

Заключение: Интеллект в каждом бите. Союз LoRaWAN и Big Data – это не просто техническое решение, это шаг к созданию глобальной интеллектуальной среды. Мы переходим от управления объектами к управлению потоками данных.

Оптимизация, достигнутая благодаря этой связке, позволяет человечеству более бережно относиться к ресурсам планеты, делать города безопаснее, а производство – эффективнее.

В будущем мы увидим еще более тесную интеграцию, где нейронные сети будут работать непосредственно на оконечных устройствах (Edge AI), а сети LoRaWAN станут магистралями для обмена мудростью между машинами.

Мы находимся в самом начале пути, где невидимые радиоволны и колоссальные вычислительные мощности объединяются, чтобы построить более совершенный и предсказуемый мир.

Список литературы

[1] Чарыев А.Б., Шайымов С.С., Аннабердиев И.А., «Важность использования BIG DATA для оптимизации управления роботами-манипуляторами». Международный научный журнал «ИННОВАЦИОННАЯ НАУКА» №2-1/2025.

[2] Шайымов С.С., Гочов А.С., Веллиева С.Т., «Значение BIG DATA в цифровую эпоху». Международный научный журнал “ИННОВАЦИОННАЯ НАУКА” №2-2/2025.

[3] Что такое Big Data (BigData) в маркетинге: проблемы, алгоритмы, методы анализа [Электронный ресурс]. – Режим доступа: http://lpgenerator.ru/blog/2015/11/17/chto-takoe-big-data-bolshie-dannye-vmarketinge-problemy-algoritmy-metody-analiza/.

[4] Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce [Электронный ресурс]. – https://habrahabr.ru/company/dca/blog/267361/.

[5] Промышленный интернет вещей [Электронный ресурс]. – Режим доступа: http://www.tadviser.ru/.

Авторский вклад

Авторы внести равнозначный вклад.

SYMBIOSIS OF INVISIBLE NETWORKS: LORAWAN AND BIG DATA TECHNOLOGIES IN THE ARCHITECTURE OF THE DIGITAL FUTURE

K.A. Abrayev
Lecturer of the Department of Radio Communications and Radio Engineering Systems of the Institute of Telecommunications and Informatics of Turkmenistan

Abstract. Abstract. We live in an era when inanimate objects acquire a «voice». Sensors in the soil of fields, sensors on power line pylons, meters in the basements of city buildings–all generate a continuous whisper of data. However, the main challenge of digitalization has long been connectivity: how to transmit a signal from thousands of devices, operating on a single battery for years, over distances of tens of kilometers? The solution came with LoRaWAN technology. But data transmission alone is only half the battle. The real magic begins where this colossal flow of information meets the power of Big Data and advanced analytics. This symbiosis creates an «intelligence layer» of reality, allowing us to manage entire cities and industries with mathematical precision. We'll discuss third-party providers and discuss our products that support LoRaWAN.

Keywords: LoRaWAN, base station, subscriber device, Big Data.

УДК 004.8:004.056

ПРИМЕНЕНИЕ ТЕХНОЛОГИЙ BIG DATA ДЛЯ МОНИТОРИНГА И ВЫЯВЛЕНИЯ МОШЕННИЧЕСКИХ ДЕЙСТВИЙ В РЕАЛЬНОМ ВРЕМЕНИ



М.М. Ходжамаммедов
преподаватель кафедры информационных технологий
Государственного энергетического института
Туркменистана
mekanhoja2021@gmail.com

М.М. Ходжамаммедов
Окончил Государственного энергетического института Туркменистана. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем, организацией учебного и научно-исследовательского процессов в техническом университете.

Аннотация. В данной работе рассматриваются теоретические и прикладные аспекты обнаружения аномалий в условиях обработки больших данных с целью выявления мошеннической активности в цифровых системах. Актуальность исследования обусловлена экспоненциальным ростом объемов данных и усложнением методов цифрового мошенничества. Проведен анализ современных подходов, включая статистические методы, алгоритмы машинного обучения и глубокие нейронные сети. Особое внимание уделено распределенным вычислительным системам и технологиям потоковой обработки данных. Представлен сравнительный анализ методов и их эффективности в задачах выявления мошенничества.

Ключевые слова: Big Data, обнаружение аномалий, мошенничество, машинное обучение, нейронные сети, анализ данных, кибербезопасность

Введение. Цифровая трансформация экономики сопровождается стремительным увеличением объемов данных, генерируемых различными информационными системами. По оценкам исследователей, глобальный объем данных ежегодно увеличивается более чем на 25%, что требует внедрения принципиально новых методов их хранения и анализа [1].

В финансовых, телекоммуникационных и банковских системах ежедневно обрабатываются миллионы транзакций, среди которых могут присутствовать мошеннические операции. Традиционные методы анализа оказываются недостаточно эффективными в условиях высокой размерности данных и динамичности угроз [2]. Современные формы мошенничества характеризуются высокой степенью адаптивности и использованием интеллектуальных технологий, что требует применения методов анализа данных, способных выявлять скрытые закономерности и отклонения в режиме реального времени [3].

Теоретические основы обнаружения аномалий. Аномалия определяется как наблюдение, существенно отличающееся от нормального поведения системы [4]. В контексте кибербезопасности такие отклонения могут свидетельствовать о наличии мошеннической активности. Существует несколько основных подходов к обнаружению аномалий:

- **Вероятностный подход**, основанный на моделировании распределения данных и вычислении вероятности принадлежности наблюдения к нормальному классу [5];
- **Геометрический подход**, использующий метрики расстояния и плотности данных;
- **Информационный подход**, основанный на анализе энтропии и сложности данных.

Особенностью задач обнаружения мошенничества является высокая несбалансированность данных, при которой доля аномалий крайне мала, что усложняет процесс обучения моделей [6;7;2].

Методы машинного обучения для обнаружения аномалий. Методы машинного обучения являются ключевыми инструментами анализа Big Data. Они позволяют автоматически выявлять закономерности и адаптироваться к изменяющимся данным. Методы **обучения с учителем** эффективны при наличии размеченных данных, однако в реальных условиях такие данные часто ограничены [3].

В связи с этим широко применяются методы **обучения без учителя**, включая:

- кластеризацию (k-means, DBSCAN);
- методы плотности (LOF) [6];
- алгоритм **Isolation Forest**, демонстрирующий высокую эффективность при работе с большими данными [8].

Особое значение имеют **нейросетевые модели**, такие как автоэнкодеры, позволяющие выявлять аномалии на основе ошибки реконструкции входных данных [9].

Глубокое обучение и анализ временных рядов. В задачах выявления мошенничества важную роль играет временной аспект данных. Поведение пользователей изменяется во времени, что требует учета последовательностей событий. Рекуррентные нейронные сети, включая архитектуры LSTM, позволяют эффективно моделировать временные зависимости и выявлять отклонения [10]. Глубокие нейронные сети обеспечивают высокую точность анализа, однако требуют значительных вычислительных ресурсов и больших объемов обучающих данных [9]. **Инфраструктура Big Data.** Реализация методов анализа невозможна без соответствующей инфраструктуры. Платформы **Hadoop** и **Apache Spark** обеспечивают распределенную обработку данных и масштабируемость вычислений [1].

Технологии потоковой обработки, такие как **Apache Kafka** и **Spark Streaming**, позволяют анализировать данные в реальном времени, что особенно важно для задач предотвращения мошенничества [11]. **Практическое применение в финансовых системах.** В финансовых системах алгоритмы обнаружения аномалий используются для анализа транзакций в режиме реального времени. Системы учитывают множество параметров:

- геолокацию пользователя;
- частоту операций;
- поведенческие паттерны;
- характеристики устройства.

Гибридные модели, объединяющие несколько методов, позволяют повысить точность обнаружения и снизить уровень ложных срабатываний [3].

Проблемы и ограничения. Несмотря на эффективность методов, существуют следующие ограничения:

- высокая вычислительная сложность моделей;
- недостаток размеченных данных;
- проблема интерпретируемости моделей;
- адаптация мошенников к новым алгоритмам [4].

Перспективы развития. Перспективными направлениями являются:

- развитие объяснимого искусственного интеллекта (Explainable AI);
- применение графовых моделей;
- внедрение федеративного обучения;
- автоматизация обучения моделей (AutoML) [12]

Заключение. Технологии Big Data в сочетании с методами машинного и глубокого обучения являются эффективным инструментом выявления мошенничества. Их применение позволяет значительно повысить безопасность цифровых систем и снизить финансовые риски. Дальнейшие исследования должны быть направлены на повышение точности моделей, снижение вычислительных затрат и улучшение интерпретируемости результатов.

Список литературы

[1] Chen M., Mao S., Liu Y. Big Data: A Survey // Mobile Networks and Applications. 2014.
[2] Aggarwal C.C. Outlier Analysis. Springer, 2017.
[3] Provost F., Fawcett T. Data Science for Business. O'Reilly, 2013.
[4] Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Computing Surveys. 2009.
[5] Hawkins D. Identification of Outliers. Chapman & Hall, 1980.
[6] Breunig M.M. et al. LOF: Identifying Density-Based Local Outliers // SIGMOD. 2000.
[7] Shannon C. A Mathematical Theory of Communication // Bell System Technical Journal. 1948.
[8] Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // ICDM. 2008.
[9] Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016.
[10] Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997.
[11] Kleppmann M. Designing Data-Intensive Applications. O'Reilly, 2017.
[12] Molnar C. Interpretable Machine Learning. 2020.

Авторский вклад

Мекан Ходжамамедов Мерданович – постановка задач научных исследований, разработка методологии оценки эффективности генеративной политики при оптимизации проектных решений, а также руководство проведением этих исследований.

APPLICATION OF BIG DATA TECHNOLOGIES FOR REAL-TIME MONITORING AND DETECTION OF FRAUDULENT ACTIVITIES

M.M. Hojamammedov
Lecturer, Department of Information Technologies, Turkmen State Energy Institute
mekanhoja2021@gmail.com

Abstract. The paper examines theoretical and applied aspects of anomaly detection in big data environments for fraud detection in digital systems. Modern approaches including machine learning and deep learning are analyzed. Special attention is given to distributed systems and real-time data processing.

Keywords: Big Data, anomaly detection, fraud, machine learning, neural networks, cybersecurity

УДК 004.8

ПРИМЕНЕНИЕ VLM ДЛЯ АНАЛИЗА ВИЗУАЛЬНЫХ СЦЕН



В.В. Венгеренко
Младший научный сотрудник государственного научного учреждения «Объединенный институт проблем информатики Национальной академии наук Беларуси»
vengerenko@lsi.bas-net.by

В.В. Венгеренко

Является аспирантом ОИПИ НАН Беларуси. Область научных интересов связана с разработкой методов и алгоритмов машинного обучения для решения задач анализа данных и компьютерного зрения.

Аннотация. Визуально-языковые модели (Vision-Language Models, VLM) представляют собой сложное достижение в области искусственного интеллекта (ИИ, Artificial Intelligence, AI), объединяющее возможности как компьютерного зрения (Computer Vision, CV), так и обработки естественного языка (Natural Language Processing, NLP) для обеспечения более целостного понимания данных.

В отличие от традиционных моделей, которые фокусируются на одном типе входных данных – визуальном или текстовом, – VLM предназначены для обработки и понимания мультимодальных данных, объединяя визуальную и текстовую информацию для формирования более содержательных выводов.

Цель работы – определить особенности использования VLM для анализа визуальных сцен.

Рассмотрены ключевые компоненты VLM. Исследованы существующие подходы к анализу визуальных сцен на основе VLM. Выполнен обзор методов оценки качества таких моделей.

Ключевые слова: искусственный интеллект, глубокое обучение, языковая модель, модальность, токен.

Введение. За последнее десятилетие в области компьютерного зрения произошел переход от приложений, основанных на восприятии, таких как классификация изображений и обнаружение объектов, к более сложным задачам, таким как понимание сцены, реляционное моделирование и ответы на визуальные вопросы (Visual Question Answering, VQA) [1].

Но вопрос о том, как обеспечить системам визуального контроля истинную способность к рассуждению, до сих пор остается нерешенным.

Визуальное рассуждение (visual reasoning) выходит далеко за рамки распознавания объектов и предполагает выводы, основанные на атрибутах объектов, их отношениях и причинно-следственных зависимостях в пространстве и времени.

Оно служит важным связующим звеном между низкоуровневым восприятием и высокоуровневым познанием.

Несмотря на значительные успехи, сохраняются серьезные пробелы в объяснении, обобщении между разными предметными областями и способности решать сложные задачи.

Последние достижения в области VLM продемонстрировали все более сложные возможности в интерпретации и анализе визуального контента.

Эти модели могут выполнять сложные задачи, такие как ответы на открытые визуальные вопросы, понимание документов и мультимодальный диалог.

Задача исследования состоит в обзоре существующих подходов к анализу визуальных сцен на основе VLM.

Компоненты VLM. VLM – это мультимодальные, генеративные модели ИИ, способные понимать и обрабатывать видео, изображения и текст [2].

Большинство VLM соответствуют архитектуре из трех частей (рисунок 1):

– визуальный энкодер (vision encoder);

– проектор (projector);

– большая языковая модель (Large Language Model, LLM).

LLM – это категория моделей глубокого обучения (Deep Learning, DL), обученных на огромных объемах данных, что делает их способными понимать и генерировать естественный язык и другие типы контента для выполнения широкого спектра задач [3].

Наиболее широко используемыми архитектурами LLM являются архитектуры только с энкодером (encoder-only), только с декодером (decoder-only) и архитектуры типа энкодер-декодер.

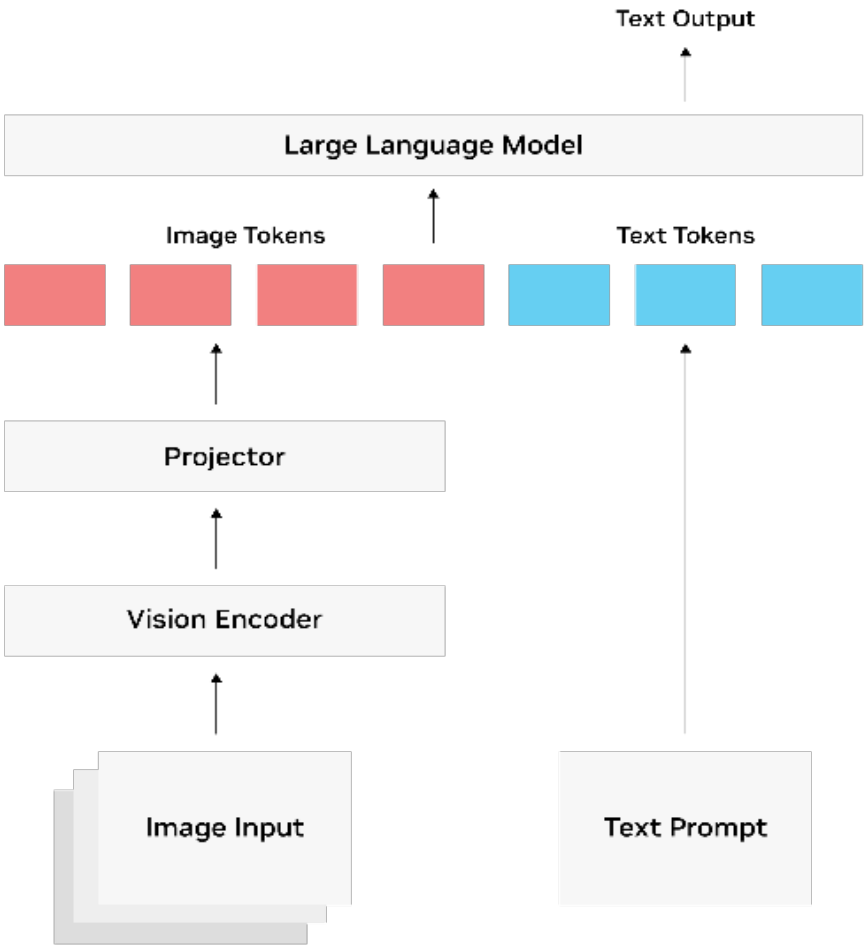


Рисунок 1. Общая архитектура VLM

Большинство из них основаны на архитектуре трансформер (Transformer) [4] (рисунок 2).

Визуальный энкодер играет решающую роль в проецировании визуальных компонентов в признаки векторного представления (embedding), которые соответствуют векторным представлениям LLM для таких задач, как генерация текста или изображений [6].

Он обучается извлекать содержательные визуальные признаки из изображений или видеоданных, обеспечивая интеграцию с языковыми представлениями.

В частности, визуальные энкодеры, используемые во многих VLM, предварительно обучаются на больших объемах мультимодальных или графических данных: эти энкодеры совместно обучаются на парах изображение-текст, что позволяет им эффективно улавливать визуальные и языковые связи.

Известными примерами являются CLIP (Contrastive Language-Image Pretraining), который сопоставляет изображения и текстовые векторные представления с помощью контрастивного обучения, и BLIP (Bootstrapping Language-Image Pretraining), который использует предобучение с начальной загрузкой для сопоставления языка и изображений.

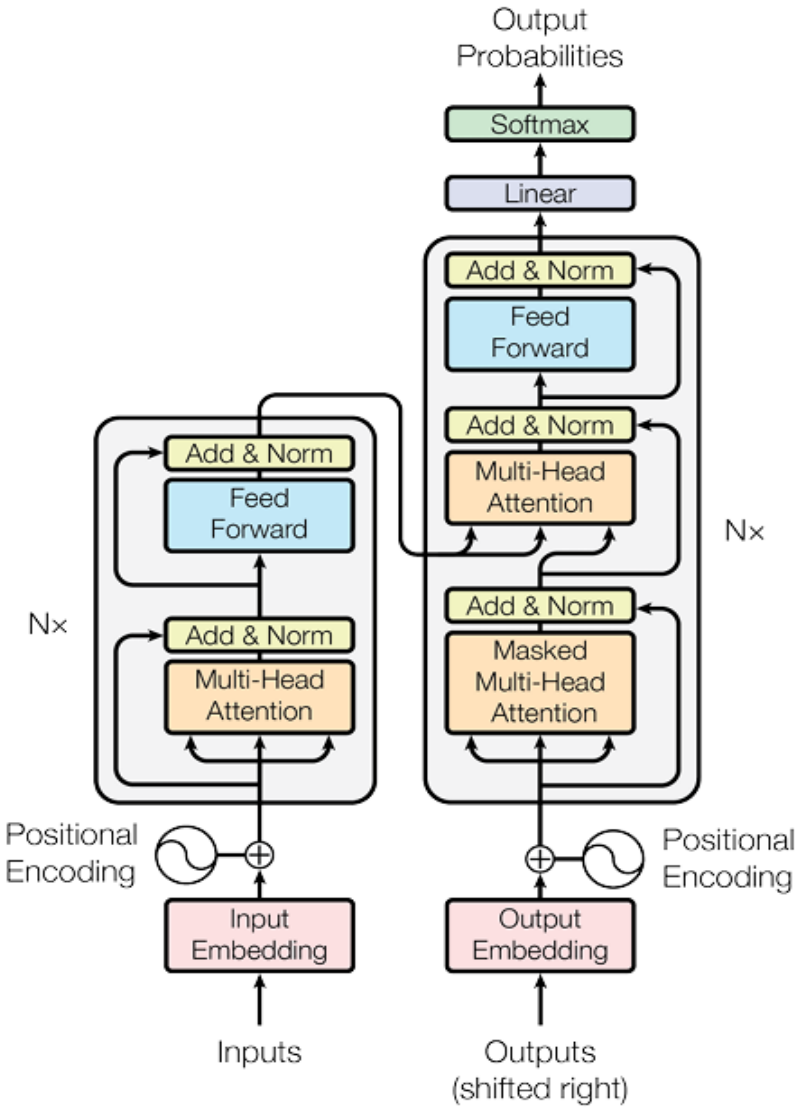


Рисунок 2. Архитектура трансформер [5]

Существуют энкодеры, предварительно обученные на ImageNet или аналогичных крупномасштабных наборах данных. Они обучаются на огромных объемах размеченных визуальных данных или с помощью самоконтролируемого (self-supervised) обучения, что позволяет им фиксировать визуальные признаки, характерные для конкретной предметной области. Будучи изначально одномодальными, эти энкодеры, такие как ResNet или Vision Transformers (ViTs), могут быть адаптированы для решения мультимодальных задач. Многие современные VLM обычно включают в себя предобученные визуальные энкодеры, которые не только обеспечивают надежное и осмысленное визуальное представление, но и являются высокоэффективными для трансферного (transfer) обучения. Они превосходят случайно инициализированные энкодеры, используя полученные визуальные знания из своих областей обучения.

Текстовый энкодер (text encoder) проецирует токенизированные текстовые последовательности в пространство векторного представления, аналогично тому, как визуальные энкодеры обрабатывают изображения [6]. Такие модели, как CLIP, BLIP и ALIGN, используют как графический, так и текстовый энкодер. В этих моделях используется контрастивное обучение для сопоставления графических и текстовых векторных представлений в общем скрытом пространстве, что позволяет эффективно улавливать межмодальные связи. Однако в более новых моделях часто отсутствует специальный

текстовый энкодер. Вместо этого они полагаются на LLM для понимания текста, интегрируя визуальные данные с помощью проекционных слоев или механизмов перекрестного внимания.

Текстовый декодер (text decoder) использует LLM в качестве основного генератора текста, используя визуальные энкодеры для проецирования признаков изображения [6]. В этих моделях обычно используется минимальный механизм визуальной проекции, что позволяет мощному языковому декодеру генерировать контекстуально насыщенные выходные данные. Для обучения VLM с нуля обычно требуется отдельный текстовый декодер, в то время как при использовании LLM в качестве основы часто используются исходные декодеры из LLM.

Механизмы перекрестного внимания (cross-attention mechanisms) обеспечивают визуально-текстовое взаимодействие, позволяя токенам из одной модальности (визуальной) влиять на токены из другой модальности (текстовой) [6]. Эти слои вычисляют показатели внимания в разных модальностях, но не все модели их используют.

Проектор отображает визуальные признаки, извлеченные с помощью визуального энкодера, в общее пространство векторного представления, согласованное с текстовыми векторными представлениями LLM [6]. Обычно он состоит из многослойного персептрона (Multilayer Perceptron, MLP), который преобразует трехмерные визуальные представления в компактные токены векторного представления, совместимые с текстовой модальностью. Проектор можно обучать совместно с остальной частью модели для оптимизации межмодальных задач или фиксируя определенные части модели, такие как LLM, для сохранения предобученных знаний.

Анализ сцен. Анализ сцен – фундаментальная задача компьютерного зрения, направленная на извлечение семантической информации из изображений или видео для идентификации объектов, сцен и их взаимосвязей [7]. Областями его применения выступают автономное вождение, интеллектуальное видеонаблюдение и медицинская диагностика. Последние достижения в области VLM продвинули анализ сцен на новый уровень развития.

Визуальное рассуждение можно систематически разделить на несколько основных типов, каждый из которых соответствует определенному способу вывода. В литературе широко описываются пять основных парадигм: символическое (symbolic), реляционное (relational), причинно-следственное (causal), временное (temporal) и рассуждение на основе здравого смысла или намерений (commonsense or intent-driven) [1]. Символическое рассуждение предполагает манипулирование дискретными структурированными абстракциями, полученными на основе визуального ввода. Обычно оно основано на преобразовании необработанных изображений в промежуточные символьные представления, такие как графы сцен, логические деревья или программы, с последующим выводом на основе правил. Это позволяет создавать легко интерпретируемые конвейеры рассуждений. Хотя такая явная прослеживаемость выгодна для верификации и объяснимости, этот метод требует аннотированной семантической информации, такой как маски сегментации или атрибуты объектов, которые зачастую трудно извлечь непосредственно из пикселей. Реляционное рассуждение, напротив, делает акцент на обнаружении и моделировании парных или более сложных отношений между визуальными объектами. Эти отношения могут охватывать пространственные, функциональные или семантические измерения. Вместо использования предопределенных правил реляционные модели, часто реализуемые с помощью графовых нейронных сетей (Graph Neural Networks, GNN), трансформеров или реляционных сетей, учатся неявно фиксировать зависимости посредством архитектурного проектирования. Такие возможности имеют решающее значение для таких задач, как создание графов сцен, понимание визуальных эталонных выражений и анализ изображений на основе аналогий. Причинно-следственное рассуждение выходит за рамки корреляции, моделируя механизмы, посредством которых одно событие влияет на другое. В визуальных областях это часто включает выявление направленных зависимостей, а не просто совпадений. Такие методы, как структурные причинно-следственные модели (Structural Causal Models, SCM),

контрфактические оценщики и do-исчисление, обеспечивают основу для получения интервенционных и гипотетических выводов на основе визуальных данных. Эти подходы особенно актуальны в рассуждениях, основанных на видео, где хронология событий кодирует множество причинно-следственных сигналов.

Временное рассуждение направлено на динамический визуальный контент, который развивается с течением времени. Такие задачи, как прогнозирование движения, сегментация действий или упорядочение событий, требуют понимания как временной эволюции отдельных объектов, так и взаимодействий между объектами. В зависимости от сложности задачи, архитектуры варьируются от рекуррентных нейронных сетей (Recurrent Neural Networks, RNN), таких как LSTM (Long Short-Term Memory) и временные сверточные сети (Temporal Convolutional Networks, TCN), до пространственно-временных трансформеров, основанных на внимании, которые фиксируют долгосрочные зависимости. Временное рассуждение особенно важно в видеонаблюдении, обеспечении качества (Quality Assurance, QA) видео и автономном восприятии. Рассуждения на основе здравого смысла и намерений связаны с неявным пониманием действий, целей и контекстуальных сигналов человека, что позволяет системам визуализации делать правдоподобные выводы даже в неоднозначных или частично наблюдаемых условиях. Этот тип рассуждений играет центральную роль в реальных приложениях, включающих вспомогательную робототехнику, прогнозирование намерений водителя и анализ социальной обстановки. Для поддержки такой формы рассуждения модели обычно включают в себя внешние базы знаний, сети с расширенной памятью или предобученные языковые модели. Задача описания изображений имеет большое значение во многих практических приложениях, включая взаимодействие человека и компьютера и мультимодальные рекомендательные системы. Современные модели обычно обучаются на наборах данных Microsoft COCO (MS-COCO) и Flickr30k, размеченных людьми. Как у MS-COCO, так и у Flickr30k есть определенные ограничения: они в основном состоят из изображений обычных объектов и сцен, но в них отсутствуют изображения редких или сложных событий, таких как бедствия, спорт или искусство; описания и запросы, связанные с изображениями, часто являются упрощенными и повторяющимися, или неточными, что не отражает использование естественного языка и ожидания пользователя. Таким образом, существует потребность в более реалистичных и крупномасштабных наборах данных, которые могут отражать разнообразие и насыщенность визуальной и лингвистической информации в реальном мире. Стоит отметить, что сбор высококачественных размеченных наборов данных, содержащих пары изображение-текст, требует значительных временных и финансовых затрат [8].

Наборы данных с описаниями изображений обычно делятся на две основные категории: размеченные человеком, такие как MS-COCO и Flickr30k, и собранные из сети Internet, такие как CC, CC12M и SBU Captions. Эти наборы данных в совокупности формируют основу для VLP (Vision Language Pretraining) и впоследствии настраиваются для последующих задач. Как правило, наборы данных, размеченные человеком, меньше по размеру, но содержат значительно меньше шума по сравнению с наборами данных, собранными из сети Internet. Однако обе категории характеризуются относительно лаконичными описаниями.

С развитием глубокого обучения в создании описаний к изображениям произошел значительный прогресс. В последнее время особое внимание уделяется архитектурам на основе трансформеров. ExpansionNet v2 использует Swin-трансформер в качестве энкодера и декодер на основе расширения для улучшенной генерации описаний, включая механизмы динамического расширения и множественного внимания (multi-head attention). ViT-GPT2 сочетает в себе ViT-энкодер и GPT2-декодер для простого и эффективного мультимодального подхода, используя преимущества предварительно обученных унимодальных моделей. Во многих исследованиях используется стратегия предварительной подготовки больших VLM, а затем их адаптации к конкретным задачам, таким как создание описаний. BLIP-2 объединяет предобученные VLM и NLP модели, используя трансформер запросов (Q-Former) для

объединения энкодера изображений и LLM. Данная модель использует двухэтапную стратегию предварительной подготовки для выравнивания представления и генеративного обучения. Современные исследования в основном направлены на дальнейшее масштабирование этих методов, основанных на предварительном обучении, и интеграцию различных визуально-языковых задач на этапе предварительного обучения. OFA – это независимая от задач модель, объединяющая представления изображений и текста с помощью трансформеров, оптимизированная с помощью перекрестной энтропии и CIDEr для описания изображений. GIT использует энкодер изображений на основе ViT и трансформер в качестве декодера текста, предварительно обученные на больших наборах данных типа изображение-текст для преобразования визуальных входных данных в текстовые описания. I-Tuning – это облегченная модель для описания изображений, которая представляет новый модуль перекрестного внимания, объединяющий фиксированный языковой декодер GPT2 с визуальным энкодером CLIP-ViT. CA-Captioner улучшает иерархию предложений, используя механизм концентрации внимания, состоящий из трех компонентов: абсолютное позиционное кодирование блока внимания (Head Absolute Positional Encoding, HAPE), фиксирующее пространственные отношения; обучаемый механизм разреживания (Learnable Sparse Mechanism, LSM), фильтрующий шум и выделяющий ключевые объекты; локальное улучшение признаков (Local Feature Enhancement, LFE), объединяющее локальные признаки.

VCap – это дополненная поиском модель описания изображений, которая расширяет знания фиксированной LLM названиями объектов из внешней визуально-текстовой памяти. Эта модель использует легковесный модуль объединения внимания, комбинирующий извлеченные имена и визуальные признаки, обеспечивая понимание открытого мира с помощью 3,97 млн обучаемых параметров без тонкой настройки на данных, выходящих за пределы предметной области. MSRM генерирует детализированное описание сцены с точными взаимосвязями благодаря трем ключевым инновациям: семантическому анализатору, извлекающему динамические семантические сигналы; семантическому средству отображения, устанавливающему детализированные реляционные взаимодействия; адаптивному связующему декодеру, динамически объединяющему признаки различной степени детализации.

VIPCap – это визуальный запрос на основе поиска для упрощенного описания изображений, преобразующий полученный для изображения текст в семантические признаки, используя изучаемое распределение Гаусса, и сопоставляющий их с визуальными признаками для создания визуальных запросов. Вышеописанные методы создают описания с нуля, оптимизируя контролируемую функцию потерь на подобранных наборах данных [8].

Современные подходы направлены на информативное описание изображений путем объединения нескольких источников или включения дополнительной семантической информации, полученной из визуального контента. LaCLIP использует LLM для перезаписи необработанных описаний, но часто сталкивается с галлюцинациями из-за ограниченной визуальной привязки (grounding) и низкого качества входных описаний. FuseCap улучшает качество описаний с помощью детекторов объектов. VeCLIP использует LLM для объединения необработанных и синтетических описаний, но напрямую зависит от уже существующей LLM для вывода и не имеет четких указаний по использованию знаний из необработанных описаний или синтаксических подсказок из синтетических. CapsFusion решает эту проблему, настраивая LLM с открытым исходным кодом на данных, сгенерированных ChatGPT, и включая подробные инструкции, которые помогают модели принимать более обоснованные решения во время объединения результатов. Инструмент визуальной проверки фактов направлен на уменьшение галлюцинаций в расширенных описаниях путем интеграции результатов двух мультимодальных описателей. Для проверки фактов используется обнаружение объектов, а для проверки и обобщения результатов в виде согласованного окончательного описания используется LLM. QAC улучшает детализацию описания, объединяя Questioner (запрашивает объекты, пространственные отношения и

мировые знания), ChatGPT (для получения исчерпывающих ответов) и Answerer (обосновывает ответы визуальными данными), а окончательное описание синтезируется специальным Captioner [8].

Оценка VLM. По мере нарастающей интеграции языковых моделей в повседневные приложения, включая цифровых ассистентов, ботов для обслуживания клиентов, образовательные инструменты, медицинскую диагностику и даже юридический анализ, их оценка приобретает критически важное значение. Оценка охватывает важнейшие аспекты устойчивости (robustness) и достоверности (trustworthiness). Эти факторы приобретают все большее значение при всесторонней оценке производительности языковых моделей. Устойчивость изучает стабильность системы при столкновении с неожиданными входными данными [9]. В частности, устойчивость к входным данным, выходящим за пределы обучающей выборки (Out-of-Distribution, OOD) и состязательная (adversarial) устойчивость к противодействию являются популярными темами исследований в области устойчивости.

Языковые модели способны генерировать связный и, казалось бы, достоверный текст. Однако полученная информация может содержать фактические неточности или утверждения, не имеющие отношения к реальности. Это явление известно как галлюцинация (hallucination) [9].

Среди автоматических метрик оценки стоит отметить [10]:

– BLEU вычисляет профиль перекрытия n-грамм сгенерированного описания с эталонными описаниями. Данная метрика широко используется и легко вычисляется, но не учитывает семантическое сходство или беглость, а также может быть чувствительна к точному совпадению слов.

– ROUGE измеряет перекрытие n-грамм (в частности, униграмм, биграмм и самых длинных общих подпоследовательностей). Метрика подходит для оценки качества обобщения, ориентирована на полноту (Recall). Ограничения аналогичны BLEU и в основном связаны с точным совпадением слов.

– METEOR вычисляется на основе среднего гармонического значения точности (precision) и полноты униграммы, где полнота имеет большее значение. Лучшая корреляция с человеческими суждениями, чем у BLEU, включающая сопоставление корней слов и синонимии. Тем не менее, может пропускать семантические связи более высокого уровня.

– CIDEr вычисляет взвешенное по TF-IDF (Term Frequency-Inverse Document Frequency) сходство n-грамм с несколькими эталонными описаниями. Это метрика, специально разработанная для задачи описания изображений, которая придает больший вес более редким n-граммам. Иногда она может отдавать предпочтение чрезмерно избыточным описаниям.

– SPICE оценивает описания на основе семантического предлагаемого контента, включая объекты, атрибуты и отношения. Метрика улавливает семантическое значение за пределами перекрытия n-грамм и хорошо коррелирует с человеческими суждениями. Однако ее вычисление может вызывать сложность, и она не всегда эффективно обрабатывает синтаксическое разнообразие.

– CLAIR использует zero-shot LLM для оценки описаний. У данной метрики лучшее соответствие человеческим суждениям по сравнению с текущими показателями, и она обеспечивает интерпретируемость при наличии помех. У нее высокая вычислительная стоимость, а воспроизводимость может зависеть от стабильности базовой LLM.

– FLEUR – метрика, использующая большую мультимодальную модель для сопоставления описания с изображением и не нуждающаяся в эталонных описаниях. Она объяснима и демонстрирует лучшие показатели при оценке без использования эталона. Однако производительность зависит от качества и возможностей базовой большой мультимодальной модели.

В ответ на растущую сложность моделей был разработан ряд эталонных тестов для стандартизации оценки языковых моделей при решении широкого круга задач:

– MS COCO – наиболее широко используемый набор данных, который содержит большую коллекцию изображений с подробными аннотациями, такими как сегментация объектов и несколько

созданных человеком описаний к изображению [10]. Наличие нескольких описаний для одного изображения позволяет оценить разнообразие генерируемых описаний.

– Flickr8K – набор данных, предоставляющий пять детальных описаний для каждого изображения [10].

– Flickr30K – расширенная версия Flickr8K [10].

– Conceptual Captions – большой набор данных, созданный путем получения изображений и описаний программным путем с веб-сайтов, которые предоставляют большое количество зашумленных, но разнообразных пар изображение-текст [10]. Одному изображению соответствует одно описание.

– IU X-Ray – набор данных, который часто используется для описаний медицинских изображений [10]. Он содержит рентгенологические снимки и диагностические отчеты. Для подобных наборов необходимо, чтобы модели изучали специализированные словари и понимали визуальные особенности конкретной предметной области. Для каждого изображения используется только одно описание, что ограничивает лингвистическую избыточность, но обеспечивает более целенаправленное текстовое описание.

– UnIVAL – унифицированный эталонный тест, охватывающий несколько модальностей (изображение, видео, аудио, текст), позволяющий проводить оценку в классических визуально-языковых задачах, а также обобщать результаты на видео и аудио-текст [11]. Данный тест охватывает аудио-текстовые задачи, VQA, создание описаний к изображениям, ответы на вопросы в видео в рамках единого фреймворка и выпущен в качестве ресурса для оценки унифицированных мультимодальных LLM (MLLM).

Некоторые подходы для оценки выполнения задач обнаружения, сегментации, классификации изображений и обработки видео с помощью VLM используют такие классические наборы данных, как ImageNet-1K, CIFAR-10/100, ADE20K, Cityscapes, Pascal VOC, YouTube-VOS [11].

Несмотря на достижения в области автоматической оценки, оценка человеком остается решающей для всесторонней оценки качества описания изображений, поскольку люди, выполняющие оценку, могут дать тонкие суждения о связности, соответствии и общей естественности сгенерированных описаний [10]. В качестве компромисса некоторые системы используют гибридные методы оценки – автоматизированная оценка для задач большого объема в сочетании с периодическими проверками человеком областей с высоким уровнем риска или значительным влиянием [12]. Стремительное развитие LLM значительно расширило их возможности в восприятии и рассуждении в долгосрочном контексте, что все больше популяризирует их использование в качестве оценщиков в различных задачах NLP. На основе концепции LLM-as-a-Judge возникла идея MLLM-as-a-Judge [13], означающая возможность использования MLLM для оценки выполнения задач.

Заключение. В последнее время LLM достигли значительного прогресса, сравнявшись или даже превзойдя возможности человека в решении целого ряда сложных задач. Несмотря на значительные достижения, VLM по-прежнему отстают от LLM в основных возможностях рассуждения. В то время как LLM демонстрируют сильные навыки символического и абстрактного рассуждения, VLM испытывают трудности с существенными аспектами визуального понимания. Они часто неправильно подсчитывают объекты в загроможденных или перекрывающихся сценах, плохо справляются с пространственным рассуждением и не могут точно определить местоположение или распознать важные объекты во время визуального поиска. Более того, композиционное и реляционное рассуждение, где понимание критически зависит от структурированных отношений между объектами, остается серьезной проблемой. Эти ограничения указывают на то, что VLM, несмотря на их мощь, не имеют ключевых механизмов для структурированного визуального рассуждения. Следовательно, их эффективность в этих областях еще не соответствует эффективности LLM в сопоставимых задачах рассуждения на основе текста. Проведенное в работе исследование выполнено в рамках совместного белорусско-турецкого проекта № Ф25ТУРГ-001.

Список литературы

[1] Sarkar A., Idris M. Y. I., Yu Z. Reasoning in Computer Vision: Taxonomy, Models, Tasks, and Methodologies. arXiv preprint arXiv:2508.10523v1 [cs.CV]. 2025. DOI: 10.48550/arXiv.2508.10523.

[2] What are Vision-Language Models? | NVIDIA Glossary [Electronic resource]. Mode of access: https://www.nvidia.com/en-us/glossary/vision-language-models/. Date of access: 16.03.2026.

[3] What Are Large Language Models (LLMs)? | IBM [Electronic resource]. Mode of access: https://www.ibm.com/think/topics/large-language-models. Date of access: 17.03.2026.

[4] Minaee S., Mikolov T., Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J. Large Language Models: A Survey. arXiv preprint arXiv:2402.06196v3 [cs.CL]. 2025. DOI: 10.48550/arXiv.2402.06196.

[5] Vaswani A., Jones L., Shazeer N., Parmar N., Gomez A. N., Uszkoreit J., Kaiser Ł., Polosukhin I. Attention Is All You Need. Proc. of the 31st International Conference on Neural Information Processing Systems (NIPS'17). 2017;6000-6010.

[6] Li Z., Wu X., Du H., Liu F., Nghiem H., Shi G. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv preprint arXiv:2501.02189v6 [cs.CV]. 2025. DOI: 10.48550/arXiv.2501.02189.

[7] Wang J., Chen Y., Si L., Zheng C. Advancing Complex Wide-Area Scene Understanding with Hierarchical Coresets Selection. Proc. of the 33rd ACM International Conference on Multimedia (MM '25). 2025;2663-2672. DOI: 10.1145/3746027.3754707.

[8] Celona L., Bianco S., Donzella M., Napoletano P. Improving image captioning descriptiveness by ranking and LLM-based fusion. Neural Computing and Applications. 2025;37:27279-27299. DOI: 10.1007/s00521-025-11672-x.

[9] Chang Y., Wang X., Wang J., Wu Y., Yang L., Zhu K., Chen H., Yi X., Wang C., Wang Y., Ye W., Zhang Y., Chang Y., Yu P. S., Yang Q., Xie X. A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology. 2024;15(3):1-45. DOI: 10.1145/3641289.

[10] Panchal P., Polara V., U S., Baz A., Patel S. K. Deep learning–driven image captioning: Progress through transformers and large language models. PLoS One. 2026;21(3):e0345012. DOI: 10.1371/journal.pone.0345012.

[11] Deng Z., Wang Y., Liang Y., Du J., Yang Y., Fang L., He L., Han Y., Zhu Y., Miao C., Zhang W., Chen J., Li Y., Zhao W., Yu P. S. A Survey of Multimodal Models on Language and Vision: A Unified Modeling Perspective. Data Mining and Machine Learning. 2025;1(1).

[12] Pandhare H. V. Evaluating Large Language Models: Frameworks and Methodologies for AI/ML System Testing. International Journal of Scientific Research and Management (IJSRM). 2024;12:1467-1486. DOI: 10.18535/ijsrm/v12i09.ec08.

[13] Chen D., Chen R., Zhang S., Wang Y., Liu Y., Zhou H., Zhang Q., Wan Y., Zhou P., Sun L. MLLM-as-a-Judge: assessing multimodal LLM-as-a-Judge with vision-language benchmark. Proc. of the 41st International Conference on Machine Learning (ICML'24). 2024;6562-6595.

APPLICATION OF VLM FOR VISUAL SCENE ANALYSIS

V.V. Vengerenko
Junior Researcher at The State Scientific Institution
«The United Institute of Informatics Problems of the
National Academy of Sciences of Belarus»

Abstract. Vision-Language Models (VLMs) represent a sophisticated advancement in artificial intelligence, integrating the capabilities of both computer vision and natural language processing (NLP) to provide a more holistic understanding of data. Unlike traditional models that focus on a single type of input – either visual or textual – VLMs are designed to process and understand multimodal data, combining visual and textual information to generate richer insights. The purpose of the research is to determine the specifics of using VLMs for visual scene analysis. The key components of VLMs are examined. The existing approaches to the visual scene analysis based on VLMs are investigated. A review of methods for evaluating the quality of such models is performed.

Keywords: artificial intelligence, deep learning, language model, modality, token.

УДК 004.89

ИНТЕГРАЦИЯ СИСТЕМ ГЕНЕРАЦИИ НА ОСНОВЕ ПОИСКА И СЕМАНТИЧЕСКОГО КЭШИРОВАНИЯ ДЛЯ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ПРЕДМЕТНО-ОРИЕНТИРОВАННЫХ LLM



А.В. Шкрабов
Студент кафедры электронной техники и технологии БГУИР
sashashkrabov04@gmail.com



Н.А. Резников
Студент кафедры электронной техники и технологии БГУИР
mine.turtle467@gmail.com



С. К. Дик
Заведующий кафедрой инженерной и компьютерной графики БГУИР,
канд. физ.-мат. наук, доцент
sdick@bsuir.by



В.М. Бондарик
Декан факультета доуниверситетской подготовки и профессиональной ориентации БГУИР, канд. техн. наук, доцент
bondarik@bsuir.by



И.И. Ревинская
Старший преподаватель кафедры электронной техники и технологии БГУИР
inna_revinskaya@bsuir.by

А.В. Шкрабов

Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем.

Н.А. Резников

Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научного интереса – компьютерная инженерия, программирование.

С.К. Дик

Заведующий кафедрой инженерной и компьютерной графики Белорусский государственный университет информатикии радиоэлектроники, кандидат физико-математических наук, доцент. Окончил Минский радиотехнический институт по специальности «Радиотехника», руководит научными исследованиями в области лазерной медицины.

В.М Бондарик

Окончил Минский радиотехнический институт по специальности «Конструирование и производство радиоаппаратуры». Кандидат технических наук, доцент кафедры электронной техники и технологии. Область интересов медицинская электронная техника, проектирование ультразвуковых медицинских электронных систем и внедрение дистанционных образовательных технологий.

И.И. Ревинская

Окончила Белорусский государственный университет информатики и радиоэлектроники. Старший преподаватель кафедры электронной техники и технологии. Область научного интереса – медицинская электроника и обработка медицинских сигналов.

Аннотация. В статье приведено решение проблемы сокращения галлюцинаций больших языковых моделей (*LLM*) при работе с высоконагруженными распределенными базами данных в архитектуре *RAG*. Анализируется влияние системных факторов (сетевых задержек, асинхронной консистентности, семантического шума) на фактологическую точность генерации. В качестве решения рассмотрены инновационные подходы: децентрализованная архитектура *DRAG*, многоисточниковая кросс-верификация *MEGA-RAG* на базе графов знаний, интеллектуальное семантическое кэширование и стратегии многофазного ранжирования. Установлено, что комплексное внедрение данных методов сводит к минимуму фактологическую манипуляцию, эффективно отсеивает «жесткие негативы» и оптимизирует баланс между барьером и скоростью вывода без ресурсоемкого дообучения модели.

Ключевые слова: большие языковые модели (*LLM*), генерация, дополненный поиск (*RAG*), галлюцинации нейросетей, распределенные базы данных, *DRAG*, *MEGA-RAG*, семантическое кэширование, графы знаний.

Введение. Фундаментальный принцип работы современных больших языковых моделей (LLM) основывается на механизме авторегрессионного предсказания следующего токена, при котором целевая функция нейронной сети оптимизируется на минимизацию потерь при воспроизведении вероятностных и дистрибутивных статистик, полученных из больших массивов обучающих данных. Однако эта же архитектурная особенность порождает уязвимость: процесс генерации происходит без поддержания явного, детерминированного соответствия объективной внешней реальности или формальным базам фактов. Модель оперирует исключительно в латентном пространстве выученных вероятностей, что неизбежно приводит к феномену конфабуляции, или галлюцинаций – генерации правдоподобной, но фактически недостоверной или логически некорректной информации. Внедрение таких систем в важные домены, включая клиническую медицину, юриспруденцию, финансовый аудит и управление промышленными системами, наталкивается на жесткие регуляторные барьеры, которые требуют высоких уровней прозрачности, интерпретируемости и абсолютной фактологической точности. Для преодоления ограничений, связанных со статичностью знаний, «замороженных» в весовых коэффициентах параметров модели на этапе предварительного обучения, исследовательским сообществом была разработана концепция генерации, дополненной поиском (RAG). В рамках этого подхода на первом этапе пользовательский запрос векторизуется и используется для извлечения релевантных документов, параграфов или структурированных фрагментов данных из внешней базы знаний. На втором этапе эти извлеченные фрагменты интегрируются в исходный контекст (промт) вместе со специализированными системными инструкциями, направляя большую языковую модель на использование именно этой актуальной информации при формулировании итогового ответа. Таким образом, RAG позволяет расширить фактологическую базу модели до неограниченных масштабов, предоставляя доступ к специализированным закрытым корпоративным доменам знаний без необходимости инициации экспоненциально дорогих процессов дообучения всей нейронной сети. Однако по мере масштабирования подобных систем до корпоративного уровня архитектурный фокус неизбежно смещается в сторону применения высоконагруженных распределенных баз данных для хранения миллиардов векторных представлений (эмбеддингов) и ассоциированных текстовых фрагментов. Распределенные среды характеризуются физическим разделением узлов хранения, необходимостью сложной сетевой маршрутизации и применением компромиссных моделей консистентности данных.

В таких условиях к классическим причинам возникновения галлюцинаций самой LLM добавляются системные сбои конвейера извлечения: сетевые задержки могут приводить к рассинхронизации компонентов, а проблемы согласованности – к извлечению устаревших,

поврежденных или противоречивых фрагментов знаний. Когда генеративная модель получает на вход гетерогенный или внутренне противоречивый контекст из распределенной базы, ее механизмы внимания пытаются синтезировать единый ответ, что с высокой долей вероятности приводит к комбинированным галлюцинациям. Следовательно, проблема снижения галлюцинаций в масштабируемых RAG-системах перестает быть исключительно лингвистической или нейросетевой задачей; она трансформируется в комплексную проблему инженерии распределенных систем, требующую глубокого анализа взаимодействия между алгоритмами векторного поиска, топологией сети и механизмами внимания больших языковых моделей.. [1].

Многомерная таксономия галлюцинаций и их геометрические сигнатуры в латентных пространствах. Чтобы эффективно бороться с галлюцинациями, необходимо детально разобрать само это понятие галюцинаций, поскольку сегодня этим общим термином называют совершенно разные системные сбои. Ситуация, при которой языковая модель полностью игнорирует предоставленные в контекстном окне релевантные документы из распределенной базы, кардинально отличается от ситуации, когда модель генерирует абсолютно несуществующие сущности на основе шума, или от случая, когда она предоставляет искаженные, но правдоподобные детали о реально существующих концепциях. Данные сбои имеют различные корневые причины, разные последствия для конечного пользователя и требуют применения принципиально разных стратегий алгоритмической коррекции. Комплексный анализ причинно-следственных связей требует рассмотрения всего жизненного цикла разработки большой языковой модели, начиная от этапа сбора данных и архитектурного проектирования, заканчивая этапами тонкой настройки и инференса в *RAG*-конвейере. В контексте анализа автономных ИИ-агентов и систем, постоянно взаимодействующих с распределенными базами данных, современная можно выделить несколько специфических категорий галлюцинаций:

1) галлюцинации рассуждения напрямую влияют на способность модели анализировать поведение и принимать последовательные, логически обоснованные решения при многошаговом поиске в графах или реляционных таблицах;

2) галлюцинации выполнения, когда модель некорректно применяет инструменты или формирует синтаксически неверные запросы (например, ошибочные *SQL*-запросы к распределенным базам);

3) галлюцинации восприятия возникают при некорректной интерпретации контекста или мультимодальных данных;

4) галлюцинации запоминания связаны с конфликтом между знаниями, инкапсулированными в весах модели, и знаниями, извлеченными из векторной базы;

5) галлюцинации коммуникации проявляются в виде непоследовательного диалогового взаимодействия с пользователем или другими агентами в сети.

В ходе анализа автономных ИИ-агентов и систем, постоянно взаимодействующих с распределенными базами данных, выделены основные причины возникновения галлюцинаций (таблица 1) [2].

Таблица 1. Основные причины и механизмы возникновения галлюцинаций в *LLM*

Причина	Фактор уязвимости	Механизм возникновения галлюцинации и влияние на генерацию
Качество обучающих данных	Зашумленность и неполнота	Внедрение ошибочных паттернов в параметрическую память модели, что приводит к некорректным ответам даже при наличии правильного контекста.
Смещение данных	Дефицит репрезентативного разнообразия	Провоцирует имитационные конфабуляции, когда модель чрезмерно экстраполирует узкие доменные корреляции на универсальные запросы.
Временная деградация	Использование устаревших (статичных) корпусов	Генерация фактологических ошибок при ответах на запросы о динамически меняющемся мире, требующая жесткого перекрытия через механизмы *RAG*.
Архитектура и инференс	Природа авторегрессионного декодирования	Оптимизация вероятностных распределений без привязки к семантической истине (отсутствие детерминированного обоснования).

Различные классы галлюцинаций больших языковых моделей поддаются математической дифференциации на основе анализа их топологического поведения на многомерной единичной гиперсфере векторных вложений, обозначаемой как C^{d-1}. В качестве математического инструмента для такого анализа был предложен Индекс семантического обоснования, который количественно измеряет степень вовлеченности сгенерированного ответа во взаимодействие с предоставленным контекстом. Механика данного индекса базируется на вычислении сложных отношений угловых расстояний между векторами токенов ответа и векторами исходных документов на единичной гиперсфере [3]. Индекс семантического обоснования достигает высочайшего уровня дискриминации (величина эффекта *d* Коэна находится в интервале 0,92–1,28) при обнаружении феномена недобросовестности – специфического типа галлюцинации, при котором модель генерирует текст, семантически ортогональный извлеченному контексту. Математический метод (Индекс семантического обоснования) разделяет «нормальный ответ» и «галлюцинацию» настолько четко, что они выглядят как абсолютно разные категории объектов. Система не просто «слегка замечает» ошибку, она видит гигантский разрыв между достоверным фактом и выдумкой нейросети, который позволяет изолировать такие галлюцинации со 100% уверенностью. Понимание этих геометрических характеристик открывает путь к созданию превентивных фильтров на уровне логического вывода: если система векторного поиска фиксирует, что формируемый ответ начинает отклоняться от гиперсферического кластера извлеченных документов, генерация может быть прервана или скорректирована до того, как галлюцинация будет транслирована конечному пользователю [4].

Влияние сетевых задержек и асинхронной консистентности на фактологическую точность. В процессе эволюции векторного поиска от изолированных прототипов к промышленным распределенным архитектурам проявляется существенный недостаток *RAG*-систем, который заключается в их низкой устойчивости к системным сбоям и аномалиям в работе баз данных. Качество генерируемого текста и вероятность возникновения галлюцинаций зависит от чистоты, релевантности и своевременности документов, предоставляемых модулем извлечения. В архитектуре распределенных баз данных одним из наиболее разрушительных факторов является сетевая задержка – временной интервал, необходимый для передачи пакетов данных между физически разнесенными вычислительными узлами кластера. В масштабной распределенной среде, особенно построенной на принципах микросервисов или глобального гео-шардирования, данные хранятся фрагментированно на множестве серверов. Любая комплексная операция, требующая доступа к различным сегментам знаний или обновления векторных индексов, неизбежно сопряжена с многократным межсетевым взаимодействием.

В распределенных архитектурах *RAG* возрастание сетевых задержек выступает фактором, лимитирующим производительность системы. Данное воздействие проявляется в двух аспектах:

1) извлечение семантически релевантных чанков из гетерогенных географических зон приводит к накоплению критической задержки (кумулятивной латентности), что исключает возможность использования синхронных методов поиска в контуре генерации ответа;

2) высокие сетевые задержки препятствуют реализации транзакций с гарантиями строгой консистентности.

Процессы обновления базы знаний (векторного индекса) требуют координации между узлами распределенной системы; превышение пороговых значений латентности делает синхронный консенсус недостижимым, вынуждая применять парадигму согласованности в конечном счете. Режим согласованности в конечном счете в распределенных БД характеризуется наличием временного окна, в течение которого узлы содержат противоречивые данные. Это приводит к состоянию гонки при конкурентном обновлении записей, следствием чего является деградация векторных индексов. Для *RAG*-пайплайна данная деградация оборачивается извлечением семантически несогласованных фрагментов, содержащих конфликтующие факты. Поскольку архитектура *LLM* не предполагает встроенного логического фактчекинга, механизмы внимания нивелируют противоречия путем генерации синтетических, статистически правдоподобных, но ложных конструкций, что идентифицируется как фактологическая галлюцинация. Нивилирование данной проблемы требует применения компенсаторных архитектурных паттернов: многоуровневого

кэширования на границе сети, репликации с учетом локальности запросов и динамической оптимизации маршрутизации.

Децентрализованная парадигма. Классические централизованные *RAG*-системы, демонстрируя высокую эффективность в лабораторных условиях, при развертывании в реальных высоконагруженных доменах сталкиваются с ограничениями масштабируемости и обеспечения конфиденциальности данных. Особую остроту данная проблема приобретает в чувствительных областях, таких как прецизионного здравоохранения, требует агрегации высокочувствительных персональных биометрических и медицинских данных пациентов. Централизованные монолитные базы знаний вступают в противоречие с регуляторными требованиями суверенитета данных, а поддержание их актуальности в динамически изменяющихся условиях сопряжено с экономически неоправданно высокими затратами. Разрешением указанного архитектурного тупика выступает парадигма распределенной генерации с дополненным поиском (*DRAG*), реализующая переход от облачных монолитов к одноранговым сетям периферийных вычислений. В архитектуре DRAG каждый участник сети выступает автономным узлом, обладающим локальной базой знаний, приватным экземпляром языковой модели и модулем защищенной коммуникации.

Обработка информационного запроса структурирована как пятиэтапный конвейер.

1. Инициирующий узел силами локальной малой языковой модели (*sLLM*) выполняет парсинг запроса и извлекает имплицитные тематические сигнатуры. На основе вычисления косинусного сходства между эмбеддингом запроса и локальным графом знаний принимается решение о достаточности локального контекста. При низких значениях сходства или негативных эвристиках удовлетворенности пользователя инициируется переход к стадии внешнего поиска.

2. Для навигации в децентрализованной сети применяется алгоритм тематически-ориентированного случайного блуждания. В отличие от лавинной маршрутизации, *TARW* использует аналитический потенциал локальной *LLM* для направленного поиска: модель идентифицирует узлы-кандидаты на основе кэшированных таблиц экспертизы, направляя запрос только тем пирам, чья историческая релевантность в данной предметной области подтверждена.

3. Узлы-доноры не экспортируют сырые текстовые данные. Вместо этого локальная *LLM* генерирует «отфильтрованные пользователем фрагменты знаний», применяя маскировку конфиденциальных паттернов. Данный механизм гарантирует абсолютный суверенитет исходных данных, транслируя их в семантически обогащенный, но деидентифицированный формат.

4. Инициирующий узел агрегирует полученные деидентифицированные фрагменты с собственным локальным контекстом, формируя расширенное семантическое пространство для финальной генерации ответа.

5. Сгенерированный ответ кэшируется, а метрики репутации пиров-доноров обновляются. Это обеспечивает адаптивную маршрутизацию будущих запросов, позволяя обходить деградировавшие или высоконагруженные узлы, что повышает общую отказоустойчивость распределенной системы.

Экспериментальная оценка алгоритма *TARW* демонстрирует его эффективность в распределенной архитектуре *DRAG*: на бенчмарке *MMLU* достигнута точность *Exact Match* 83,90% (против 85,73% у централизованного *RAG*) при сокращении межсетевого трафика до 6,87 сообщений на запрос (против 10,91 при *flooding*). Главная проблема распределенного подхода – гетерогенность узлов и риск контаминации данных – компенсируется внедрением криптографической репутационной системы и взвешенного ранжирования пиров по исторической валидности [6].

Многоисточниковая кросс-верификация и графовые структуры в архитектуре *MEGA-RAG*. Минимизация фактологических обобщений в самых важных предметных областях (таких как клиническая медицина, фармакология и юриспруденция) обуславливает необходимость разработки архитектуры с нулевой толерантностью к конфабуляциям. Подобный уровень надежности принципиально недостижим в рамках базовых парадигм *RAG*, вне зависимости от степени их децентрализации. Обычные генеративные агенты часто не способны распознать логические противоречия при агрегации нескольких фрагментов текста, что делает их непригодными для клинического использования. Для решения данных проблем был предложен специализированный фреймворк *MEGA-RAG* (Поисковая дополненная генерация с уточнением ответов на основе

нескольких доказательств), представляющий собой комплексную четырехэтапную архитектуру, направленную на минимизацию галлюцинаций. В отличие от стандартной парадигмы, распределяющейся исключительно на плотных векторных индексах, *MSER* реализует ортогональную последовательную агрегацию гетерогенных данных из трех структурно различных распределенных хранилищ. К ним относятся полнотекстовые базы рецензируемой биомедицинской литературы *PubMed*, обрабатываемые с помощью плотных векторных библиотек *FAISS*; нормативная база данных *IRIS* Всемирной системы здравоохранения, поиск в которой осуществляется с применением лексического алгоритма организации *BM25* [7].

Интеграция графических баз данных (*Graph-RAG*) является одним из наиболее эффективных методов подавления галлюцинаций в больших языковых моделях (*LLM*).

Векторные системы строят связи на основе вероятностного математического подобия, тогда как графы знаний оперируют детерминированными ребрами, отражающими точные онтологические отношения между сущностями (например, в логической триаде «симптом – препарат – побочный эффект»).

Подобная топология позволяет делать логические выводы непосредственно на уровне системы хранения данных, отслеживая эту задачу генеративной модели, динамической стохастической конфабуляции. Отсутствие иского факта в графе приводит к возврату пустого распространения, которое инициирует правильный отказ модели от формирования ответа и блокирует генерацию артефактов на основе дополнительных векторных соотношений. Последующая обработка в рамках фреймворка включает в себя три стадии глубокого анализа.

Второй компонент – модуль генерации вариативных ответов (*DPAG*) – формирует массив кандидатных решений методом сэмплирования, после чего к ним применяются вычислительно ресурсоемкие модели кросс-энкодеров. Совместная обработка запроса и кандидатного документа через пошаговый трансформер обеспечивает выявление тонких семантических связей, исключающих задержки биэнкодерам.

На третьем этапе модуль семантико-эвиденциального спорта (*SEAE*) использует метрику *BERTScore* для математической квантификации степени обоснованности каждого сгенерированного токена полученной доказательной базой. Основным инновационным механизмом выступает четвертый модуль принципа расхождений и саморазъяснения (*DISC*).

При падении метрики SEAE ниже установленного порогового значения данный компонент не прерывает генерацию, анализ проводит семантические дивергенции между массивом потенциальных ответов, синтезирует вторичный корректирующий запрос к хранилищам данных и реализует процесс направленного редактирования формируемых знаний.

Экспериментальные замеры в задачах медицинского профиля убедительно доказывают превосходство *MEGA-RAG* над изолированными большими языковыми моделями, дообученными системами вроде *PubMedGPT* и базовыми реализациями *RAG*.

Архитектура достигает высокой точности и прозрачности, гарантируя, что любые фактологические утверждения подкреплены перекрестно проверенными источниками (таблица 2) [8].

Таблица 2. Сравнительная эффективность моделей

Тип модели	Точность	Снижение уровня галлюцинаций
PubMedBERT (изолированная модель)	Низкая (частые фабрикации)	Базовая линия (0%)
Standalone LLM (GPT-4)	Средняя	Около 15%
Стандартный векторный *RAG*	Умеренная	Около 25%
MEGA-RAG	Высокая	Свыше 40% относительно стандартов

Интеллектуальное семантическое кэширование: компромисс между задержкой и ложными срабатываниями. Развертывание масштабируемых систем *RAG* в производственной среде сопряжено с постоянной борьбой против двух главных врагов

архитектуры: стоимости обращения к большим языковым моделям и задержек логического вывода, разрушающих пользовательский опыт. Для приложений, характеризующихся высокой плотностью повторяющихся запросов с небольшими вариациями формулировок (например, чат-боты клиентской поддержки, интеллектуальные системы анализа технической документации), классические методы префиксного кэширования оказываются недостаточными. Механизмы кэширования промптов, реализованные на уровне провайдеров (таких как *Microsoft Azure* или *OpenAI*), функционируют путем сохранения тензоров «ключ-значение» механизмов внимания для начальных токенов запроса. Этот подход требует идеального, побайтового совпадения префиксной части запроса. Аналогичным образом, традиционное кэширование типа ключ-значение (базирующееся на хешировании строк) не способно обрабатывать естественный язык: запросы «Какова ваша политика возврата?», «Как я могу вернуть товар?» и «Могу ли я отправить это обратно?» будут восприниматься системой как абсолютно разные ключи, инициируя три отдельных, дорогостоящих и медленных прохода через всю цепь извлечения и генерации. Концептуальным решением этой проблемы является внедрение интеллектуального слоя семантического кэширования. Семантическое кэширование оперирует не синтаксисом, а исключительно математическим смыслом и интенцией запроса. Алгоритм функционирует следующим образом: поступающий пользовательский запрос конвертируется легковесной моделью в векторное представление. Затем этот вектор используется для поиска в высокоскоростной резидентной базе данных, где хранятся эмбеддинги исторических запросов, сопряженные с ранее сгенерированными эталонными ответами. Поиск осуществляется через вычисление косинусного сходства между векторами в многомерном пространстве. Если показатель сходства превышает жестко заданный порог уверенности, система мгновенно возвращает кэшированный ответ, полностью обходя этапы векторного поиска по базе документов и генерации ответа тяжелой *LLM*. Внедрение этого паттерна обеспечивает феноменальный прирост производительности: время отклика может сократиться с ~6.5 секунд (полный цикл *RAG*) до 100 миллисекунд (извлечение из кэша) – 65-кратное ускорение, сопровождающееся пропорциональным падением операционных расходов. Однако внедрение семантического кэша таит в себе скрытую угрозу, известную в индустрии как «кризис ложных срабатываний». Векторное пространство не идеально; при некорректной настройке гиперпараметров система начинает выдавать уверенные, семантически близкие, но фактически неверные ответы из кэша, заменяя фактологическую точность на скорость. Для нейтрализации этого существенного риска необходимо перейти от простого тюнинга порогов к многослойному архитектурному дизайну. Фундаментальным принципом предотвращения ложных срабатываний является «принцип наилучшего кандидата». Данный принцип постулирует, что вместо инкрементального заполнения кэша «на лету», производственная система должна быть предварительно инициализирована «Золотым стандартом» – каноническими ответами на вопросы, покрывающими основные предметные домены. Существенно важно, что вместе с золотым стандартом в базу внедряются «стратегические дистракторы» – синтетические запросы, которые семантически похожи на канонические, но требуют принципиально иного ответа.

Идеальное эмпирическое соотношение составляет 3 дистрактора на 1 эталонный запрос. Наличие таких контрольных точек в латентном пространстве позволяет алгоритму предельно точно вычислять границы смысловых кластеров, предотвращая ошибочное объединение разных концепций. Оптимизация порога срабатывания является не менее значимым фактором и сильно зависит от выбора базовой модели эмбеддингов, так как плотность распределения векторов у разных архитектур кардинально отличается. Дальнейшее снижение уровня ложных срабатываний (ниже 5%) требует реализации многовекторной архитектуры, при которой создаются изолированные индексные пространства для различных аспектов одного запроса. Кроме того, внедрение легковесных моделей кросс-энкодеров на этапе валидации кандидата из кэша позволяет провести глубокий анализ взаимосвязей между входным запросом и кэшированным эталоном до отправки ответа пользователю, фактически исключая семантические галлюцинации на уровне кэша [9].

Многофазное ранжирование и фильтрация семантического шума. Управление распределенными цепями *RAG* требует жесткой балансировки между задержкой вывода и выходом (компромисс задержки и точности). Линейное расширение контекстного окна или увеличение выбора (параметр *Top-K*) перегружает механизмы внимания и провоцирует галлюцинации типа «потери в середине». Проблема решается внедрением методик многофазного ранжирования, декомпозирующего процесса определения. Первичный обзор массива данных осуществляется высокопроизводительными лексическими (*BM25*) и векторными (*ANN*) алгоритмами, для чего применяется гибридное переранжирование лучших кандидатов. На конечном этапе ресурсоемкие кросс-энкодеры проводят прецизионный семантический анализ ограниченной выборки, в результате чего определяется релевантность без критической деградации латентной системы. Существенным источником происхождения конфабуляций является наличие в выборке «жестких негативов» – документов с высоким поверхностным сходством, но фактологически неверным добавлением. Нейтрализация данного семантического шума требует применения комплексных алгоритмов на базе методов атрибуции.

В условиях ограниченного доступа к радиусам моделей в распределенных базах рекомендуется использовать атрибуции «черного ящика» с привлечением альтернативных фильтрующих моделей, заменяющих градиентный анализ («белый ящик»). Подобная комплексная оптимизация шумоподавления и ранжирование серьезности задач трансформации естественного языка в *SQL*-запросы (*Text-to-SQL*).

В отличие от базовых архитектур с использованием схемы фиксированного количества (*fixed-k*), внедрение алгоритмов динамической иерархической кластеризации позволяет адаптивно определять релевантный объем таблиц, отсекающий шум на этапе формирования контекста. Синергия данной ситуации с итеративным обогащением предложения реальных записей повышения метрику *F*1 извлечения схем с 0,79 до 0,88 и повышает точность выполнения запросов на 11%, полностью выполняя необходимость ресурсоемкого дообучения (тонкой настройки) модели [10].

Заключение. Интеграция больших языковых моделей с распределенными корпоративными базами данных в рамках структуры *RAG* актуализирует фундаментальную проблему фактологической достоверности искусственного интеллекта.

Сетевые задержки, компромиссные модели асинхронной консистентности и присутствие семантического шума трансформируют функцию подавления галлюцинаций из-за лингвистической в комплексной инженерной проблеме.

Преодоление выявленных ограничений требует концептуального отказа от базовых парадигм линейного векторного поиска с использованием глубоко эшелонированных, многофазных архитектурных обработок данных.

Анализ современных подходов доказывает эффективность децентрализованных систем класса *DRAG*, которые за счет алгоритмов направленного топологического блуждания решают проблемы масштабируемости сети и суверенитета конфиденциальных данных.

В важнейших областях, требующих нулевой толерантности к конфабуляциям, архитектурными решениями является многофакторная кросс-верификация (фреймворк *MEGA-RAG*).

Интеграция вероятностного векторного определения со строгой онтологической детерминированностью графов знаний (*Graph-RAG*) позволяет снизить уровень фактологических сбоев.

Вычислительная рентабельность таких тяжеловесных процессов внедрения уровней интеллектуального семантического кэширования с использованием ложных программ и методик иерархического многофазного ранжирования, которые минимизируют латентность вывода и эффективно фильтруют «жесткие ошибки».

Обеспечение надежности генеративных систем в высоконагруженных распределенных средах не учитывает экстенсивного масштабирования параметров языковых моделей и их ресурсоемкого обучения, а следовательно, глубокой алгоритмической оптимизации конвейера извлечения и верификации контекста.

Последующая эволюция инженерных знаний в данной области будет базироваться на развитии гибридных методов шумоподавления, декомпозиции простых запросов для итеративного фактчекинга и внедрения математически обоснованных метрик геометрического контроля семантики на всех этапах логического вывода.

Список литературы

[1] Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in neural information processing systems. – 2020. – Т. 33. – С. 9459-9474. – URL: https://arxiv.org/pdf/2005.11401 (дата обращения 27.02.2026).

[2] Sajjadi Mohammadabadi, S. M.; Kara, B. C.; Eyupoglu, C.; Karakus, O. A Survey on Hallucination in Large Language Models: Definitions, Detection, and Mitigation. Preprints 2025, 2025100540. – URL: https://doi.org/10.20944/preprints202510.0540.v1 (дата обращения 28.02.2026)

[3] Marín J. A Geometric Taxonomy of Hallucinations in LLMs //arXiv preprint arXiv:2602.13224. – 2026. – URL: https://arxiv.org/html/2602.13224v1 (дата обращения 27.02.2026).

[4] Marín J. Semantic grounding index: Geometric bounds on context engagement in RAG systems //arXiv preprint arXiv:2512.13771. – 2025. – URL: https://arxiv.org/pdf/2512.13771 (дата обращения 27.02.2026).

[5] Gao Y. et al. Retrieval-augmented generation for large language models: A survey //arXiv preprint arXiv:2312.10997. – 2023. – Т. 2. – №. 1. – С. 32. – URL: https://www.semanticscholar.org/reader/46f9f7b8f88f72e12cbdb21e3311f995eb6e65c5 (дата обращения 27.02.2026).

[6] Xu C. et al. Distributed retrieval-augmented generation //arXiv preprint arXiv:2505.00443. – 2025. – URL: https://arxiv.org/html/2505.00443v1 (дата обращения 27.02.2026).

[7] Xu S. et al. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health //Frontiers in Public Health. – 2025. – Т. 13. – С. 1635381. – URL: https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2025.1635381/full (дата обращения 27.02.2026).

[8] Edge D. et al. From local to global: A graph rag approach to query-focused summarization //arXiv preprint arXiv:2404.16130. – 2024. – URL: https://arxiv.org/pdf/2404.16130 (дата обращения 27.02.2026).

[9] Bang F. Gptcache: An open-source semantic cache for llm applications enabling faster answers and cost savings //Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023). – 2023. – С. 212-218. – URL: https://openreview.net/pdf?id=ivwM8NwM4Z (дата обращения 27.02.2026).

[10] Bozdemir M., Bilgin M. Schema Retrieval with Embeddings and Vector Stores Using Retrieval-Augmented Generation and LLM-Based SQL Query Generation //Applied Sciences. – 2026. – Т. 16. – №. 2. – С. 586. – URL: https://www.mdpi.com/2076-3417/16/2/586 (дата обращения 27.02.2026).

Авторский вклад

Авторы внесли равный вклад в написание статьи

INTEGRATION OF SEARCH-BASED GENERATION SYSTEMS AND SEMANTIC CACHING TO INCREASE THE RELIABILITY OF DOMAIN-ORIENTED LLMS

N.A. Reznikov
student of the Department of Electronic Technique and Technology of BSUIR

A.V. Shkrabov
student of the Department of Electronic Technique and Technology of BSUIR

S. K. Dzik
Chair of the Department of Engineering and Computer Graphics, PhD, Associate Professor

V.M. Bondarik
Dean of the Faculty of Pre-University Training and Career Guidance
Associate Professor of the Department of Electronic Engineering and Technology

I.I. Revinskaya
Senior Lecturer at the Department of Electronic Technique and Technology

Abstract. The paper discusses an approach to increasing the robustness of neural network models used in eye footage segmentation to typical image distortions using targeted data augmentation. Using the example of the winner of the OpenEDS Semantic Segmentation Challenge 2019 – the RitNet model (based on U-Net and DenseNet), it is shown how the addition of synthetic artifacts (structured starbursts, Gaussian blur, random lines and shifts), characteristic of images in VR / AR glasses, made it possible to achieve high accuracy with an extremely small model size. This approach emphasizes the importance of adapting training data to real-world application conditions to improve the reliability of models in computer vision tasks.

Keywords: big data, convolutional neural networks, segmentation, model robustness, infrared cameras, data augmentation, eye detection systems

УДК 004.032.26

МЕТОДИКА ПОВЫШЕНИЯ УСТОЙЧИВОСТИ СВЁРТОЧНОЙ НЕЙРОННОЙ СЕТИ С ИСПОЛЬЗОВАНИЕМ АУГМЕНТАЦИИ ДАННЫХ НА ПРИМЕРЕ МОДЕЛИ ДЛЯ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ ГЛАЗНЫХ ЯБЛОК



Н.А. Резников
Студент кафедры электронной техники и технологии БГУИР
mine.turtle467@gmail.com



А.В. Шкрабов
Студент кафедры электронной техники и технологии БГУИР
sashashkrabov04@gmail.com



С. К. Дик
Заведующий кафедрой инженерной и компьютерной графики БГУИР,
канд. физ.-мат. наук, доцент
sdick@bsuir.by



Г.Д. Ситник
Доцент кафедры общей врачебной практики с курсом гериатрии и паллиативной медицины БГМУ,
канд. мед. наук, доцент, врач высшей категории по неврологии



И.И. Ревинская
Старший преподаватель кафедры электронной техники и технологии БГУИР
inna_revinskaya@bsuir.by

Н.А. Резников
Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научного интереса – компьютерная инженерия, программирование.

А.В. Шкрабов
Студент кафедры электронной техники и технологии Белорусского государственного университета информатики и радиоэлектроники. Область научных интересов связана с разработкой методов и алгоритмов построения информационно-компьютерных систем.

С.К. Дик
Заведующий кафедрой инженерной и компьютерной графики Белорусский государственный университет информатикии радиоэлектроники, кандидат физико-математических наук, доцент. Окончил Минский радиотехнический институт по специальности «Радиотехника», руководит научными исследованиями в области лазерной медицины.

Г.Д. Ситник

Кандидат медицинских наук, доцент, врач высшей категории по неврологии. Область научного интереса – лечение больных с неврологическими проявлениями поясничного остеохондроза.

И.И. Ревинская

Окончила Белорусский государственный университет информатики и радиоэлектроники. Старший преподаватель кафедры электронной техники и технологии. Область научного интереса – медицинская электроника и обработка медицинских сигналов.

Аннотация. В работе рассматривается подход к повышению устойчивости (robustness) нейросетевых моделей сегментации глаз к типичным искажениям изображений с помощью целенаправленной аугментации данных. На примере победителя OpenEDS Semantic Segmentation Challenge 2019 – модели RitNet (на базе U-Net и DenseNet), показано, как добавление синтетических артефактов (структурированные блики starburst, гауссово размытие, случайные линии и сдвиги), характерных для изображений в VR/AR-очках, позволило достичь высокой точности при крайне малом размере модели. Такой подход подчёркивает важность адаптации обучающих данных под реальные условия применения для повышения надёжности моделей в задачах компьютерного зрения.

Ключевые слова: большие данные, свёрточные нейронные сети, сегментация, устойчивость моделей, инфракрасные камеры, аугментация данных, системы детекции глаз

Введение. В настоящее время алгоритмы на основе нейронных сетей встречают всё большее применение в различных сферах жизни, в том числе и в среде распознавания образов на изображениях и видео. К лидирующим сферам применения относится здравоохранение, промышленность (выявление дефектов на производстве, анализ инфраструктуры, охрана труда), агрокультура, транспортная сфера, а также видеонаблюдение [2]. Для обучения моделей требуются огромные объемы данных, однако, существуют некоторые проблемы, связанные с возможной узконаправленностью итогового алгоритма. Модели, обученные на слишком «идеальных» данных, могут резко терять точность при столкновении с незнакомыми аномалиями в получаемой информации, а в случае гомогенности данных (например, обучение модели для распознавания лиц на фотографиях людей схожего телосложения, расы или при однотипном освещении) алгоритмы выходят узкоспециализированными, и также перестают правильно обрабатывать информацию при выходе её за известные параметры.

Для борьбы с этим применяются различные методики, обучающие модели разным вариациям данных и помогающие им лучше обрабатывать расхождения в результатах [3-4]. К методам повышения точности машинного обучения можно отнести следующие:

- Модификация (аугментация) данных – преобразование существующих изображений с целью обучения модели на более разнообразных примерах. Используются методики искажения исходных данных, смешивания данных между собой для выделения границ между классами, а также стилизация данных с целью вариативности исходного материала;
- Синтетические данные – искусственное расширение входного массива за счёт данных, сгенерированных иными моделями, что позволяет закрыть пробелы в исходном наборе и внедрить редкие, но возможные сценарии. Тем не менее, этот метод требует тщательного контроля, позволяющего избежать перенесения так называемых «галлюцинаций» между моделями – ложной информации, создаваемых нейросетевыми моделями при недостаточных исходных данных [5];
- Использование дополнительных внешних данных – намеренное добавление изображений иных классов, которые позволят модели чётче определять границы между искомым классом и сторонними данными;
- Обучение работе с неизвестными классами с целью научить модель как узнавать конкретные образы, так и распознавать то, что предоставленный объект не относится ни к одному известному классу;
- Добавление к модели набора признаков других, предварительно обученных моделей с целью расширения набора параметров;

- Активное обучение – ручная коррекция набора данных моделей с помощью ручной разметки пограничных результатов, что может помочь с разрешением неопределённостей;
- Ансамблирование – параллельное использование нескольких моделей с целью агрегации их результатов и коррекции найденных расхождений.

В данной работе сконцентрируемся на первом методе на примере свёрточной нейронной сети RitNet и том, как применение аугментации данных позволило повысить точность и уменьшить размер модели. Полученные результаты иллюстрируют потенциал аугментации данных в задачах Big Data и промышленной аналитики: вместо увеличения объёма «чистых» размеченных данных можно искусственно расширить вариативность существующих наборов, закрывая пробел между лабораторными и реальными условиями. Это особенно актуально для сенсорных данных из IoT, автономного транспорта, здравоохранения и AR/VR, где искажения (размытие, отражения, переменное освещение) являются нормой.

Особенности подхода RitNet. В качестве примера рассмотрим свёрточную модель RitNet [1], разработанную в Рочестерском Институте Технологий (RIT) в 2019 году для семантической сегментации глаза в системах отслеживания взгляда. Модель использует комбинацию архитектуры DenseNet с элементами U-Net и специально взвешенной функции потерь, учитывающей границы. Для обучения модели применялся набор данных OpenEDS [6], упомянутый ранее. Набор состоит из 12759 монохромных изображений глаз людей, полученных с инфракрасных камер, установленных в шлемах виртуальной реальности. Данные получены от 152 субъектов, примерно равно разделённых на мужчин и женщин возрастной группы от 18 до 65 лет. К изображениям прилагается размеченная маска, на которой отмечены зрачок, радужная и белковая оболочка глаз. В конкурсе, для которого была разработана модель, данные были разбитых на три поднабора: обучение (8916 изображений), подтверждение (2403 изображений) и тест (1440 изображений без размеченных масок). Предоставленный набор данных является сравнительно небольшим, и сфокусирован на реальных данных с VR/AR очков и обладает характерными для подобных сценариев артефактами (отражения от подсветки, блики, размытие от движения глаз, частичное перекрытие частей глаза веком). Для оценки качества модели в конкурсе использовалась формула:

$$\mathrm{mIoU}+\min(1/S,\,1)2,$$

где *mIoU* – mean Intersection over Union, или среднее значение перекрытия найденной искомой области реальной (рисунок 1);

S – размер модели (функция от количества параметров модели, в мегабайтах).

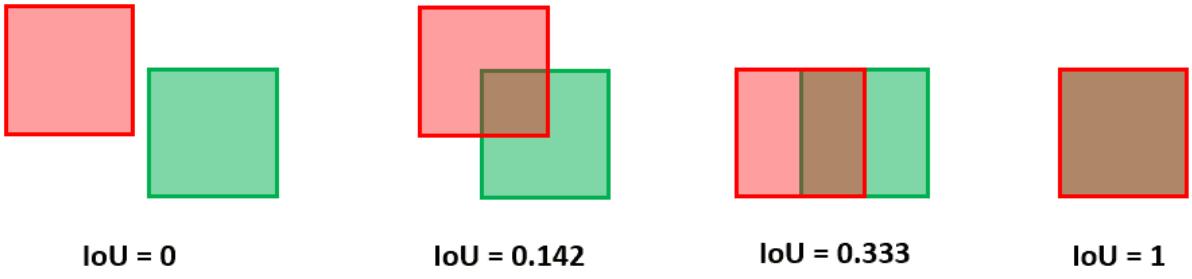


Рисунок 1. Визуальное отображение метрики IoU

Обучение производилось с использованием алгоритма Adam [7] партиями по восемь изображений в сто семьдесят пять проходов (epoch), с модификациями в степени обучения, когда результаты выходили на плато длиной более чем в пять проходов.

Для того, чтобы учесть различия в отражающей способности между образцами в зависимости от цвета глаз, оттенка кожи, а также макияжа и век, изображения, поступающие

в модель для обучения, проходили две ступени предобработки. Первым шагом была экспоненциальная гамма-коррекция изображений, а затем к данным была применено ограниченное контрастом адаптивное выравнивания гистограммы (CLAHE) (рисунок 2).

На преобразованных таким образом изображениях стало проще отделить зрачок от радужки глаза.

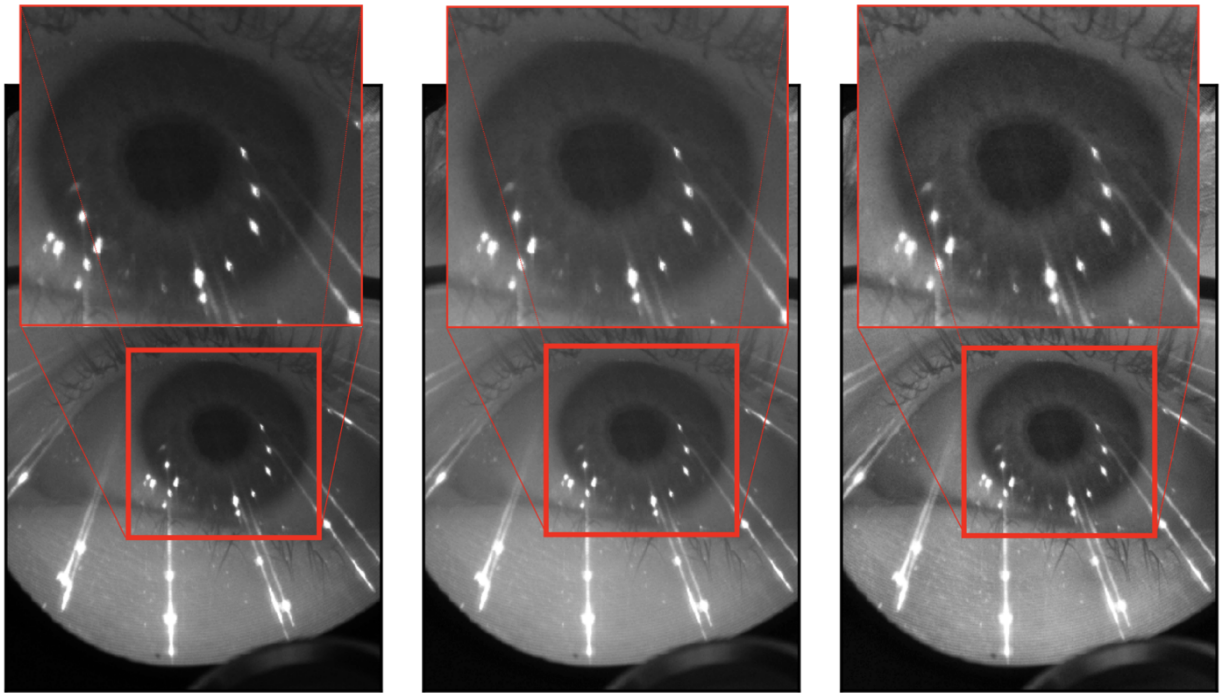


Рисунок 2. Слева направо: оригинальное изображение, изображение после гамма-коррекции, изображение после применения CLAHE

Кроме того, с целью повышения устойчивости модели изображения были подвергнуты следующим аугментациям:

- Отражение по вертикали;
- Размытие по Гауссу с размером ядра 7x7 и случайным отклонением $2 \leq \sigma \leq 7$;
- Сдвиг на 0-20 пикселей по обоим осям;
- Добавление 2-9 случайных тонких линий, построенных от случайного центра ($120 < x < 280$, $192 < y < 448$) для имитации грязи и ресниц;
- Наложение узора (рисунок 3) для уменьшения ошибок сегментации в связи с отражением инфракрасной подсветки от очков. При наложении узор смещался на случайное расстояние от 0 до 40 пикселей по обоим осям.

Каждое изображение было дополнено хотя бы одной из этих аугментаций, при этом появление каждой имело шанс 0.2. Кроме того, каждое изображение было отражено по горизонтали с шансом 0.5 (имитация правого и левого глаза).

Таким образом, в исходные данные были заранее внесены характерные для инфракрасной съемки глаз дефекты, что позволило итоговой модели лучше справляться с ними в дальнейшем.

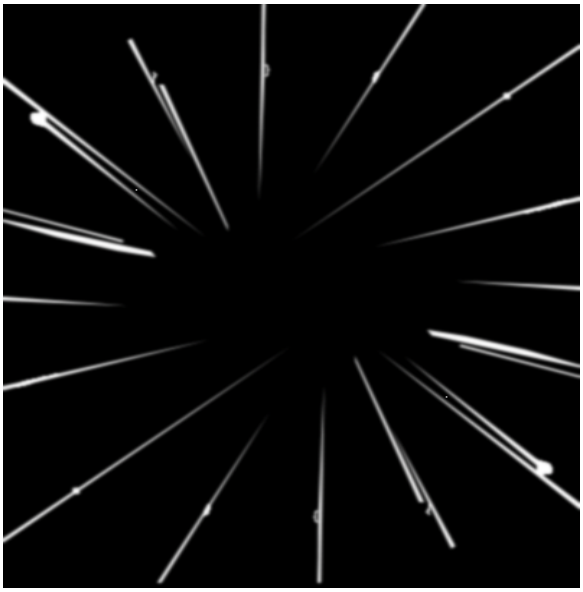


Рисунок 3. Узор, применяемый при аугментации для имитации отражений

Результаты работы. Полученная модель была сравнена с другой свёрточной моделью SegNet [8] (на примере её модификации mSegNet), а также моделями на её основе. К ним относятся:

- mSegNet w/BR (mSegNet с дополнительным модулем уточнения границ (Boundary Refinement, BR), моделирующий выравнивание границ как остаточную (residual) структуру). Такой подход, связанный с принципами остаточного обучения [9], позволяет сети целенаправленно улучшать локализацию границ между классами (в особенности между зрачком, радужкой и склерой), минимизируя ошибки на переходных зонах и повышая общую точность сегментации в условиях классового дисбаланса и низкого контраста.
- mSegNet w/SC – облегчённая версия базовой mSegNet, в которой стандартные свёртки заменены на depthwise separable convolutions (глубинно-разделимые свёртки). Данная техника, существенно снижает количество параметров и вычислительную сложность (примерно в 8–9 раз), сохраняя при этом сопоставимую работоспособность модели.

Сравнительная таблица результатов работы моделей предоставлена в таблице 1. Визуальное отображение работы модели показано на рисунке 4.

Таблица 1. Сравнение параметров различных моделей, участвующих в опыте

Модель	Средний IoU	Размер модели (S), мегабайт	Количество параметров, млн	Итоговый счёт
mSegNet	90,7	13.3	3.5	0.491
mSegNet w/BR	91.4	13.3	3.5	0.495
mSegNet w/SC	89.5	1.6	0.4	0.762
RitNet	95.3	0.98	0.25	0.976

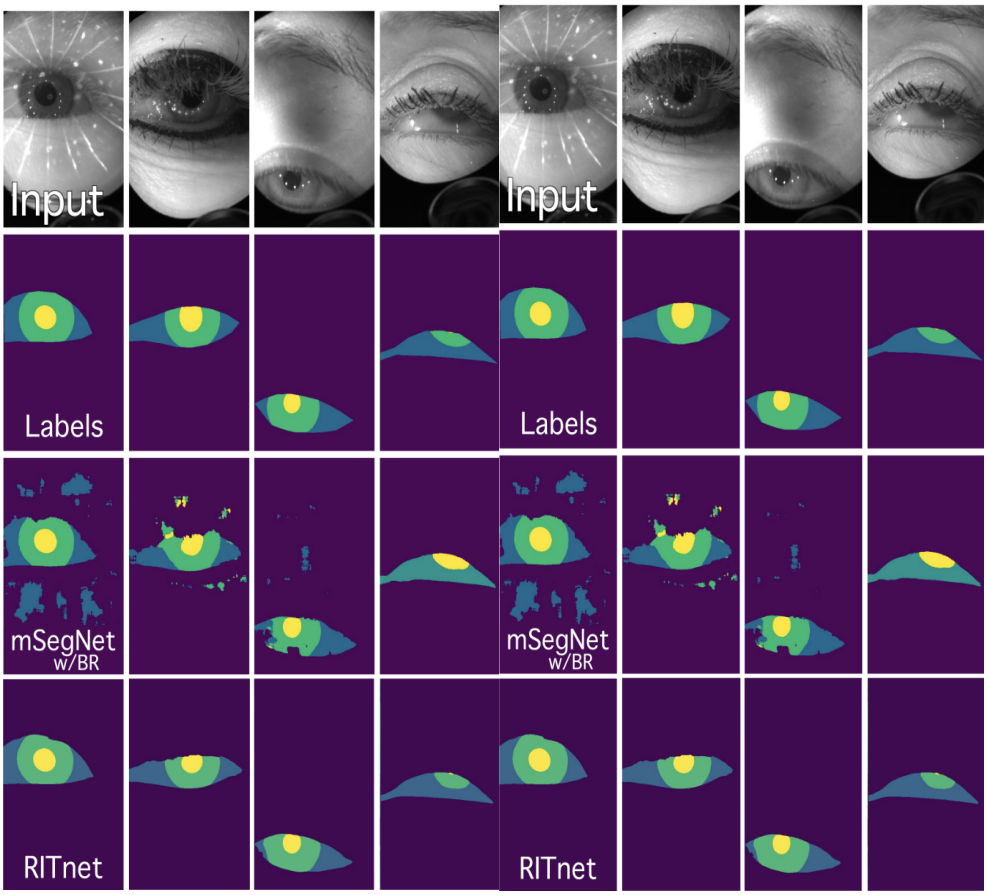


Рисунок 4. Сравнение сегментации между размеченными вручную масками и результатами работы различных моделей. Сверху вниз, первый ряд: изначальные кадры, размеченные вручную маски, второй ряд: результаты работы mSegNet w/BR, результаты работы модели RitNet.

RitNet показывает лучшие значения точности при меньшем размере, чем остальные предоставленные модели. Кроме того, точность классификации тоже выше, чем у аналогов. Небольшой размер модели позволяет использовать её для видео частотой до трёхсот кадров в секунду в реальном времени.

Заключение. Опыт RitNet является наглядным доказательством того, что при работе с обучением моделей качество данных является не менее значимым, чем количество, и, кроме того, показывает, что аугментация является крайне полезным методом повышения устойчивости нейросетевых моделей. С момента релиза RitNet были проведены дальнейшие работы в направлении создания легких и высокоскоростных моделей. Например, в статье Б. Бурлаку [10] сообщается о модели Lightweight на основе MobileNetV2, обученной на OpenEDS с аналогичными аугментациями и показавшей результат сравнимый с RitNet (mIoU – 0.9485, конечный счёт – 0.974) с меньшим количеством параметров (104728 против 248900 у RitNet), что обеспечено лучшей архитектурой. Это показывает перспективу и направления для дальнейшей работы в этом направлении.

Современные тенденции [2]) подтверждают, что data-centric подходы, включая domain-specific аугментацию и лёгкие архитектуры (MobileNetV2, EfficientNet), остаются ключевыми для промышленного внедрения моделей в условиях ограниченных ресурсов и неидеальных входных данных.

Список литературы

[1] Chaudhary A. K. [и др.] RitNet: Real-time Semantic Segmentation of the Eye for Gaze Tracking // 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). – IEEE, 2019. – С. 3698–3702. – DOI: 10.1109/iccvw.2019.00568.

[2] Vision AI Trends to Watch in 2026 [Электронный ресурс] // Roboflow Blog. – URL: https://blog.roboflow.com/vision-ai-trends-2026/ (дата обращения: 10.02.2026).

[3] Lei S., Hua Y., Zhihao S. Revisiting Fine-Tuning: A Survey of Parameter-Efficient Techniques for Large AI Models [Электронный ресурс]. – 2025. – ffhal-05008993. – URL: https://hal.science/hal-05008993 (дата обращения: 12.02.2026).

[4] Wang K. [и др.] A Survey on Data Synthesis and Augmentation for Large Language Models [Электронный ресурс] // arXiv preprint. – 2024. – arXiv:2410.12896. – URL: https://arxiv.org/abs/2410.12896 (дата обращения: 12.02.2026).

[5] Kalai A. T. [и др.] Why Language Models Hallucinate [Электронный ресурс] // arXiv preprint. – 2025. – arXiv:2509.04664. – URL: https://arxiv.org/abs/2509.04664 (дата обращения: 12.02.2026).

[6] Garbin S. J. [и др.] OpenEDS: Open Eye Dataset [Электронный ресурс] // arXiv preprint. – 2019. – arXiv:1905.03702. – URL: https://arxiv.org/abs/1905.03702.

[7] Kingma D. P., Ba J. Adam: A method for stochastic optimization // Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). – San Diego, 2015.

[8] Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017.

[9] He K. [и др.] Deep Residual Learning for Image Recognition // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2016. – С. 770–778. – DOI: 10.1109/CVPR.2016.90.

[10] Burlacu B. A. Eye Semantic Segmentation with Convolutional Neural Networks // CS230: Deep Learning, Stanford University, Autumn 2019 [Электронный ресурс]. – URL: https://cs230.stanford.edu/projects_fall_2019/reports/26259743.pdf.

Авторский вклад

Авторы внесли равный вклад в написание статьи

A METHOD FOR INCREASING THE ROBUSTNESS OF A CONVOLUTIONAL NEURAL NETWORK USING DATA AUGMENTATION USING THE EXAMPLE OF AN EYE IMAGE SEGMENTATION MODEL

N.A. Reznikov
student of the Department of Electronic Technique and Technology of BSUIR

A.V. Shkrabov
student of the Department of Electronic Technique and Technology of BSUIR

S. K. Dzik
Chair of the Department of Engineering and Computer Graphics, PhD, Associate Professor

G.D.Sitnik
PhD, Associate Professor at the Department of Belarusian State Medical University, Doctor of the highest category in neurology

I.I. Revinskaya
Senior Lecturer at the Department of Electronic Technique and Technology

Abstract. The paper discusses an approach to increasing the robustness of neural network models used in eye footage segmentation to typical image distortions using targeted data augmentation. Using the example of the winner of the OpenEDS Semantic Segmentation Challenge 2019 – the RitNet model (based on U-Net and DenseNet), it is shown how the addition of synthetic artifacts (structured starbursts, Gaussian blur, random lines and shifts), characteristic of images in VR / AR glasses, made it possible to achieve high accuracy with an extremely small model size. This approach emphasizes the importance of adapting training data to real-world application conditions to improve the reliability of models in computer vision tasks.

Keywords: big data, convolutional neural networks, segmentation, model robustness, infrared cameras, data augmentation, eye detection systems

УДК 004.934.2:004.032.26

НЕАВТОРЕГРЕССИОННАЯ МОДУЛЬНАЯ СИСТЕМА СИНТЕЗА РЕЧИ С ЯВНЫМ МОДЕЛИРОВАНИЕМ ДЛИТЕЛЬНОСТИ НА БАЗЕ CTC-ВЫРАВНИВАНИЯ



С.С. Бекарев
Студент факультета компьютерных систем и сетей БГУИР
bekarevstanislav@gmail.com



М.А. Калугина
Доцент кафедры информатики, кандидат физико-математических наук, доцент
marina_kalugina@list.ru

С.С. Бекарев

В 2022 году окончил гимназию №1 г. Горки. Интересы направлены на глубокое обучение, обучение моделей синтеза и распознавания речи.

М.А. Калугина

Окончила Белорусский государственный университет. Область научных интересов связана с исследованием проблем метрической теории диофантовых приближений зависимых величин и приложений математических методов к нейросетевому анализу

Аннотация. В статье рассматриваются методы явного выравнивания текстовых и акустических последовательностей для стабильного синтеза речи на материале русского языка. Проведено построение выравнивания текста относительно акустических признаков методом на основе CTC-loss и алгоритма Витерби. Для акустического моделирования использована неавторегрессионная архитектура с дилатационными свертками и энкодер-декодерной структурой. Оценка качества выполнена по группам показателей: точность предсказания длительности фонем (RMSE), качество реконструкции мел-спектрограмм (MSE) и субъективная разборчивость синтезированной речи (алгоритм Гриффина-Лима). Установлено, что предлагаемый трехэтапный пайплайн (обучение выравнивания, предсказание длительности графем, затем генерация мел-спектрограммы) обеспечивает устойчивое обучение и гарантированную монотонность генерации на графемном уровне с явной маркировкой ударения. Показана возможность обхода без фонемного преобразования при использовании расширенного словаря графем с ударными/безударными вариантами.

Ключевые слова: text-to-speech, CTC-loss, алгоритм Витерби, ASR, мел-спектрограмма, графемы, акустическая модель, алгоритм Гриффина-Лима, корпус RUSLAN.

Введение. Современные системы нейросетевого синтеза речи переходят от авторегрессионных архитектур (Tacotron [1], WaveNet [2]) к неавторегрессионным моделям, обеспечивающим ускорение инференса и повышение стабильности генерации. Авторегрессионные подходы, при всем высоком качестве синтеза, наследуют фундаментальные ограничения: накопление ошибок при экспозиционном смещении в условиях teacher forcing [3], а также непредсказуемое время генерации из-за последовательной генерации. Целью исследования является обучение модульной неавторегрессионной системы синтеза речи, использующей явное выравнивание текста и аудио для предсказания длительности фонем с последующей параллельной генерацией акустических признаков. Ключевой особенностью работы является реализация двухэтапного пайплайна на базе CTC-выравнивания: сначала извлекается строгий монотонный alignment между графемами и кадрами мел-спектрограммы с помощью алгоритма Витерби [4], затем на основе полученных

меток обучается предиктор длительности (DurationPredictor), обеспечивающий расширение последовательности эмбеддингов до акустического разрешения без использования attention-механизма. В работе представлена реализация полного цикла обучения системы на корпусе русской речи RUSLAN [6] (30 часов, один спикер): от подготовки графемной транскрипции с явным кодированием ударения (токены вида «+а», «+е») до генерации мел-спектрограмм [7] (80 мел-фильтров) с использованием акустической модели на основе Unet-like архитектуры [8]. Оценка качества проводится по метрикам точности предсказания длительности (RMSE), реконструкции спектральных признаков (MSE) и субъективной разборчивости синтезированной речи, полученной алгоритмом Гриффина-Лима [9].

Описание алгоритма обучения. В качестве источника акустических данных используется открытый корпус русской речи RUSLAN с текстовыми транскрипциями. Такой подход обеспечивает фиксацию акустических характеристик конкретного спикера без вариативности, свойственной многоголосым датасетам. На этапе предобработки аудиозаписи преобразуются в мел-спектрограммы с параметрами: 80 мел-фильтров, окно Ханна длительностью 64 мс, сдвиг 16 мс, логарифмическое шкалирование (log1p) для стабилизации динамического диапазона. Текстовые транскрипции токенизируются на уровне графем с явным кодированием ударения: каждая гласная представлена двумя токенами (ударная с префиксом «+» и безударная).

Для получения целевых меток длительности обучается вспомогательная сверточная модель автоматического распознавания речи с использованием CTC-loss [10].

После достижения CER 9% к выходам сети применяется алгоритм Витерби для извлечения оптимального монотонного alignment, устанавливающего соответствие между каждым токеном текста и фреймами мел-спектрограммы. Полученные alignment'ы служат эталоном для обучения DurationPredictor. Каждый пример обучающей выборки включает последовательность токенов графем с маркировкой ударения и целевой вывод: вектор длительностей (количество фреймов для каждого токена) и соответствующий фрагмент мел-спектрограммы.

Для формирования обучающей пары для AcousticModel текстовая последовательность расширяется путем повторения эмбеддингов токенов, обеспечивая временное соответствие между входом и целевой спектрограммой.

Процесс обучения выполняется на GPU в три последовательных этапа: (1) обучение ASR-модели с CTC-loss до сходимости; (2) обучение DurationPredictor на извлеченных алгоритмом Витерби alignment'ах с функцией потерь MSE; (3) обучение акустической модели на сформированных расширенных эмбеддингах с целевыми мел-спектрограммами (MSELoss). Контроль качества осуществляется на валидационной выборке по метрикам точности предсказания длительности (RMSE в кадрах), качества реконструкции спектральных признаков (MSE), а также субъективной разборчивости речи, синтезированной из предсказанных спектрограмм с помощью алгоритма Гриффина-Лима.

Поскольку записи в корпусе RUSLAN имеют неравную длительность (от коротких аудио до длинных предложений из десятка слов), формирование батчей требует выравнивания последовательностей до единой длины внутри батча. Для предотвращения влияния искусственно добавленных элементов (padding) на градиенты модели применяется механизм маскирования (masking).

При обучении ASR-модели длины входных и целевых последовательностей передаются в CTC-loss, который автоматически игнорирует вклад padding-токенов при расчете log-likelihood. Для DurationPredictor и AcousticModel используется бинарная маска, идентифицирующая позиции реальных данных внутри выровненных тензоров, при вычислении MSELoss маска применяется поэлементно, обнуляя вклад фиктивных позиций в функцию потерь и обеспечивая корректное усреднение только по актуальным временным шагам.

Результаты обучения системы. На первом этапе реализована ASR-модель для распознавания графем (включая маркеры ударения).

Применение алгоритма Витерби к выходам сети позволило извлечь строгий монотонный путь, установивший однозначное соответствие между токенами текста и фреймами мел-спектрограммы. Пример alignment’а текста можно увидеть на рисунке 1.

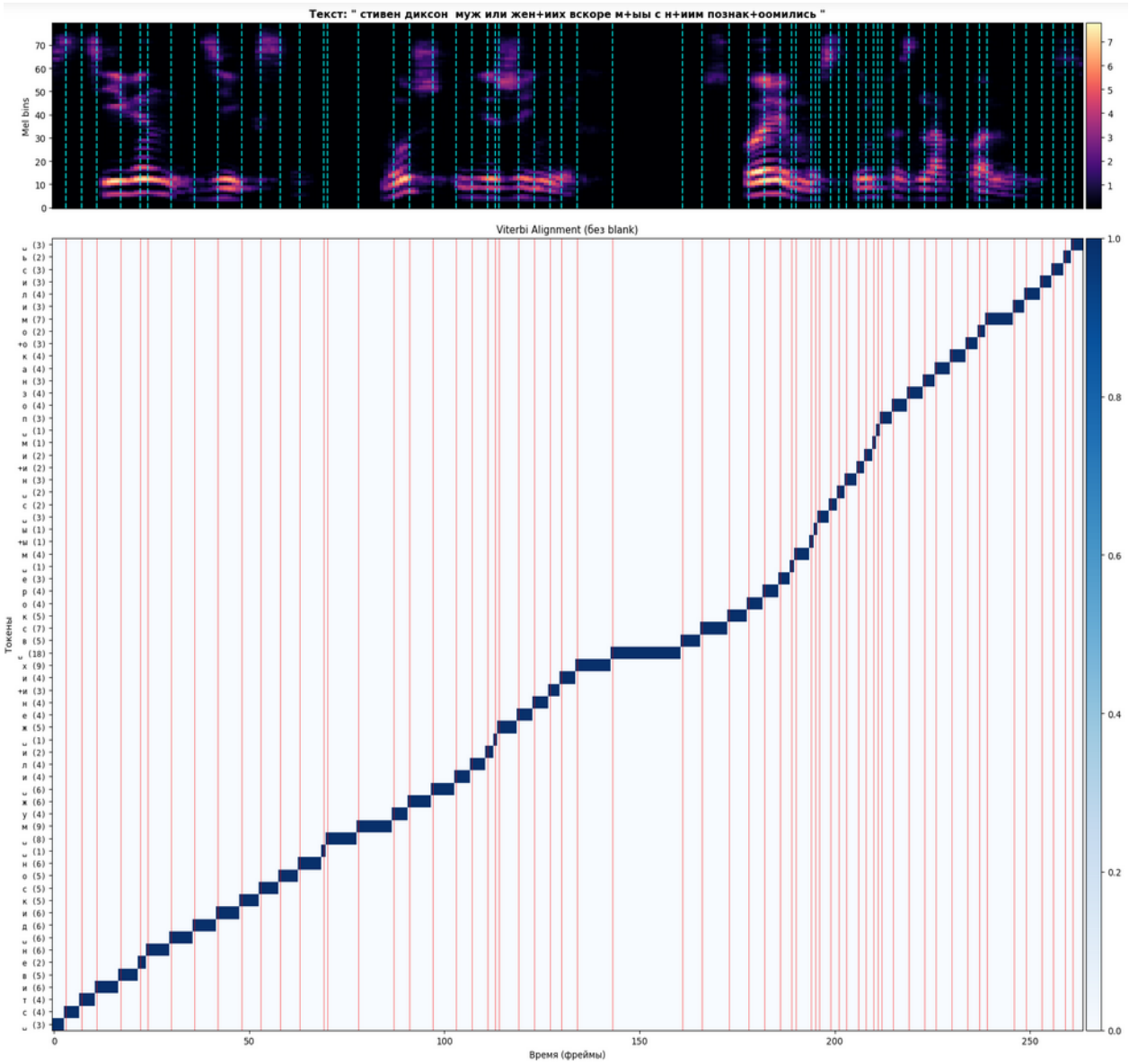


Рисунок 1. Alignment текста, полученный с помощью алгоритма Витерби

На основе извлеченных alignment'ов был обучен DurationPredictor. Кривые обучения демонстрируют монотонное снижение MSELoss: с начального значения 80 до 44 на 12-й эпохе, то есть RMSE 6.6 на валидационной выборке (рисунок 2).

Достаточно высокое значение RMSE можно объяснить тем, что нейронная сеть недооценивает длину пауз между произношением слов текста, порой пауза может доходить до 30 фреймов мел-спектрограммы, что негативно сказывается на общей метрике.

Для предотвращения переобучения, обучение было заранее остановлено.

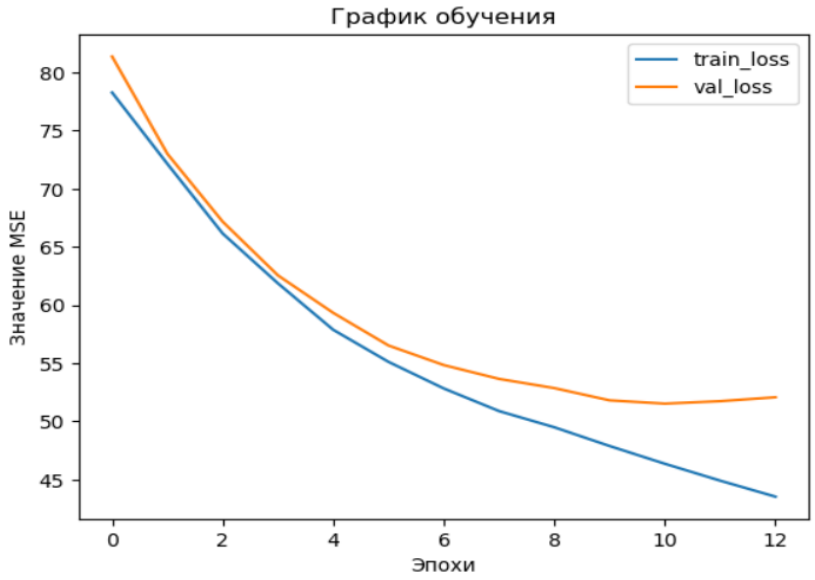


Рисунок 2. Динамика MSE предсказания длительности для DurationPredictor

Финальный этап – обучение акустической модели на архитектуре с дилатационными свертками (dilated convolutions) и стратегией downsampling/upsampling (Unet-like архитектура). Модель принимает на вход расширенные эмбеддинги графем (согласно предсказанным длительностям) и генерирует логарифмированные мел-спектрограммы (80 мел-фильтров). Обучение проводилось с MSELoss, значение которого снизилось до 0.5 на 18-й эпохе.

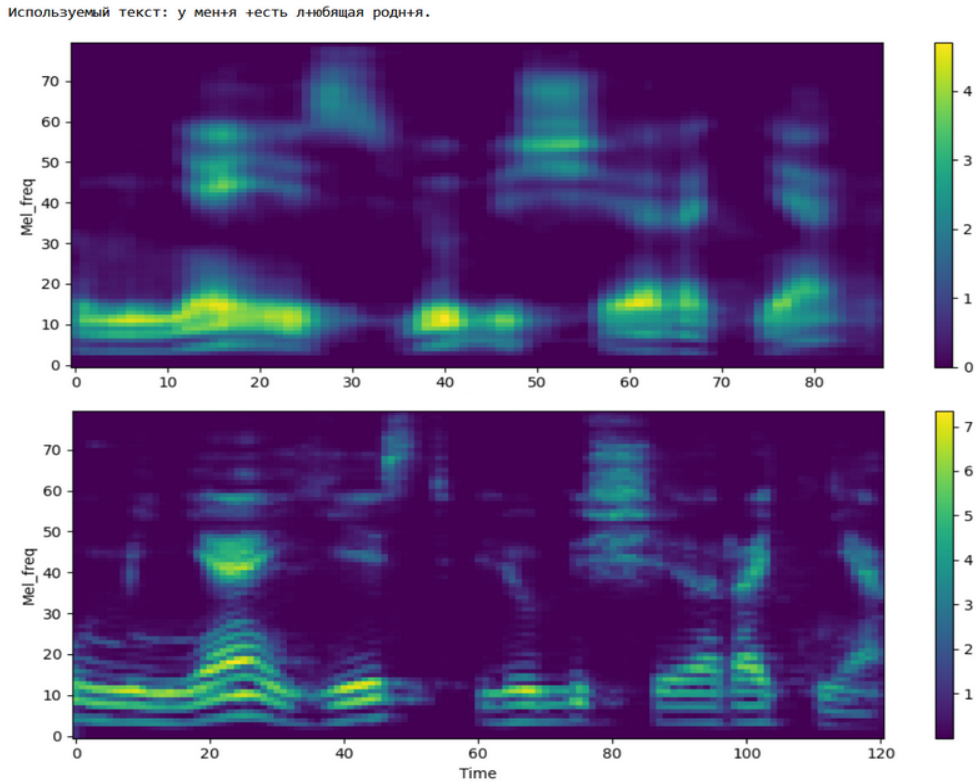


Рисунок 3. Сравнение спектрограмм: синтез DurationPredictor и AcousticModel (сверху), оригинальная запись (снизу)

Использование фиксированной, детерминированной длины последовательности (задаваемой алгоритмом Витерби) обеспечило устойчивую сходимость без колебаний функции потерь, что соответствует качеству, достаточному для разборчивости речи при использовании вокодера Griffin-Lim (рисунок 3).

Из сравнения видно, что предсказания DurationPredictor получаются немного смещенными, а предсказания AcousticModel не обладают высокой резкостью, от чего мел-спектрограмма получается сглаженной, но общая структура звучания текста сохраняется, что позволяет разобрать речь на аудио, полученном после преобразования Griffin-Lim.

Таким образом, реализованный трехэтапный пайплайн обеспечил устойчивое обучение всех компонентов системы.

Достигнутые показатели подтверждают работоспособность подхода для синтеза речи на основе графемной транскрипции с маркировкой ударения без привлечения фонематора.

Заключение. В ходе исследования разработана и экспериментально проверена методика построения модульной системы синтеза речи, использующей явное выравнивание текста и аудио для обеспечения стабильности обучения non-autoregressive генерации акустических признаков.

Предложенный подход включает обучение ASR-модели с CTC-loss, извлечение монотонного alignment алгоритмом Витерби, предсказание длительности фонем (DurationPredictor) и параллельную генерацию мел-спектрограмм акустической моделью на базе дилатационных сверток.

Проведенный анализ архитектурных решений подтвердил, что выбор метода выравнивания определяющим образом влияет на сходимость обучения и качество синтеза речи на материале русского языка:

1 Метод явного CTC-выравнивания с алгоритмом Витерби обеспечил гарантированную монотонность отображения графем на акустическую ось и устойчивую сходимость всех компонентов системы.

Трехэтапная схема (выравнивание → предсказание длительности → генерация спектрограмм) продемонстрировала стабильное снижение функций потерь без признаков переобучения: RMSE предсказания длительности достиг 6.6 кадра, MSE реконструкции мел-спектрограмм – 0.5.

При этом достигнута разборчивость текста при синтезе через алгоритм Гриффина-Лима, что подтверждает работоспособность подхода.

2 Использование графемной записи с явной маркировкой ударения (вместо фонем) позволило упростить пайплайн предобработки текста, однако внесло зависимость от точности ASR-модели (CER 9%).

Ошибки распознавания транслируются в локальные искажения длительности, что ограничивает качество синтеза для сложных фонетических конструкций.

Таким образом, проведённое исследование показало, что предложенный модульный подход с CTC-выравниванием обеспечивает детерминированность генерации, фиксированное время инференса моделей и устойчивость к переобучению при ограниченном объеме данных (30 часов).

При этом сохраняется возможность дальнейшего улучшения качества путем замены отдельных компонентов (например, дообучение ASR для снижения CER) без перестроения всей системы.

Список литературы

[1] NATURAL TTS SYNTHESIS BY CONDITIONING WAVENET ON MEL SPECTROGRAM PREDICTIONS [Электронный ресурс]. – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1712.05884 – Дата доступа: 21.03.2026

[2] WAVENET: A GENERATIVE MODEL FOR RAW AUDIO [Электронный ресурс]. – Электронные данные. – Режим доступа: https://arxiv.org/pdf/1609.03499 – Дата доступа: 21.03.2026

[3] RNNs: Teacher Forcing [Электронный ресурс]. – Электронные данные. – Режим доступа: https://cedar.buffalo.edu/~srihari/CSE676/10.2.1%20TeacherForcing.pdf – Дата доступа: 21.03.2026

[4] Распознавание речи: Классический подход [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c3/Digital_Signal_ Processing%2C_lecture_6.pdf – Дата доступа: 21.03.2026

[5] Что такое ASR [Электронный ресурс]. – Электронные данные. – Режим доступа: https://microsin.net/programming/arm/what-is-asr.html – Дата доступа: 21.03.2026

[6] Hugging Face [Электронный ресурс]. – Электронные данные. – Режим доступа: https://huggingface.co/datasets/Gzaborey/ruslan-dataset – Дата доступа: 18.03.2026

[7] Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between [Электронный ресурс]. – Электронные данные. – Режим доступа: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html – Дата доступа: 18.03.2026

[8] U-Net Architecture Explained [Электронный ресурс]. – Электронные данные. – Режим доступа: https://www.geeksforgeeks.org/machine-learning/u-net-architecture-explained/ – Дата доступа: 18.03.2026

[9] The Griffin-Lim algorithm: Signal estimation from modified short-time Fourier transform [Электронный ресурс] – Электронные данные – Режим доступа: https://speechprocessingbook.aalto.fi/Modelling/griffinlim.html – Дата доступа: 18.03.2026

[10] Распознавание речи: Современные подходы [Электронный ресурс]. – Электронные данные. – Режим доступа: http://www.machinelearning.ru/wiki/images/c/c6/Digital_Signal_ Processing%2C_lecture_7.pdf – Дата доступа: 21.03.2026.

Авторский вклад

Бекарев Станислав Сергеевич – выбор задачи исследования, формулировка концепции трехэтапного пайплайна обучения (CTC-выравнивание → предсказание длительности → генерация спектрограмм); разработка и реализация всех компонентов системы (ASR-модель, DurationPredictor, AcousticModel); создание инструментов визуализации alignment'ов; проведение экспериментов на корпусе RUSLAN и анализ метрик.

Калугина Марина Алексеевна – научное руководство исследованием; постановка проблемы; консультации по математическим основам CTC-loss и алгоритмов выравнивания.

NON-AUTOREGRESSIVE MODULAR SPEECH SYNTHESIS SYSTEM WITH EXPLICIT DURATION MODELING BASED ON CTC EQUALIZATION

S.S. Bekarev
Student of BSUIR

M.A. Kalugina
Associate Professor of Informatics Department of the BSUIR

Abstract. This article examines methods for explicitly aligning text and acoustic sequences for stable speech synthesis using Russian language material. Emphasis is placed on overcoming the instability of attention mechanisms in classical seq2seq models by decomposing the problem into alignment, duration prediction, and spectrogram generation modules. A text alignment with respect to acoustic features is constructed using a method based on CTC-loss and the Viterbi algorithm. A non-autoregressive architecture with dilated convolutions and an encoder-decoder structure is used for acoustic modeling. Quality assessment is performed using a group of metrics: phoneme duration prediction accuracy (RMSE), mel spectrogram reconstruction quality (MSE), and subjective intelligibility of synthesized speech (Griffin-Lim algorithm). It was established that the proposed two-stage pipeline (grapheme duration prediction followed by mel-spectrogram generation) ensures robust learning and guaranteed monotonicity of generation at the grapheme level with explicit stress marking. The feasibility of traversal without phoneme transformation is demonstrated using an extended grapheme dictionary with stressed/unstressed variants.

Keywords: text-to-speech, CTC-loss, Viterbi algorithm, non-autoregressive TTS, ASR, mel-spectrogram, graphemes, dilated convolutions, acoustic model, Griffin-Lim algorithm, RUSLAN corpus.

УДК 004.8: 004.056: 621.3

ADAPTIVE MULTI-SCALE HETEROGENEOUS GRAPH LEARNING FOR REAL-TIME ANOMALY DETECTION IN INDUSTRIAL CYBER-PHYSICAL SYSTEMS



G.O. Orazdurdyyeva
Instructor of Computer Sciences and Information Technologies Department,
Oguz han Engineering and Technology University of Turkmenistan
gulshatorazdurdyyewa3@gmail.com



M.B. Bekiyeva
Senior teacher of Applied Mathematics and Informatics Department,
Oguz han Engineering and Technology University of Turkmenistan, Candidate of Physical and Mathematical Sciences
successbmb@gmail.com

G.O. Orazdurdyyeva

I graduated from the Oguz han Engineering and Technology University of Turkmenistan. My research interests are related to cybersecurity, mathematical modeling, and network traffic analysis.

M.B. Bekiyeva

I graduated from the International Turkmen-Turkish University. My research interests include the development of mathematical modeling, the design of algorithms for solving problems related to the finite element method, as well as the organization of educational and research processes in a technical university.

Abstract. The proliferation of IIoT devices introduces complex security vulnerabilities in cyber-physical systems. We propose AMS-HGNN, an Adaptive Multi-Scale Heterogeneous Graph Neural Network for real-time industrial anomaly detection. Our architecture features a Scale-Adaptive Graph Convolution mechanism, a Domain-Bridge Fusion module correlating cyber and physical entities, and a Lightweight Edge-Optimized design achieving sub-15ms inference on ARM devices. Evaluated on three industrial datasets (2.3M+ events), AMS-HGNN achieves an F1-score of 0.938 on the SWaT benchmark – a 6.8% improvement over existing methods – while reducing computational overhead by 42%. Results confirm adaptive multi-scale learning as a viable pathway for deployable industrial security monitoring.

Keywords: heterogeneous graph neural networks; multi-scale learning; cyber-physical systems; anomaly detection; industrial IoT; edge computing; real-time security.

Introduction. Industrial cyber-physical systems have undergone a fundamental transformation over the past decade, driven by the convergence of operational technology and information technology [5]. Modern manufacturing facilities, power generation plants, and water treatment systems now incorporate thousands of interconnected sensors, actuators, and computational nodes, generating massive volumes of heterogeneous data at unprecedented velocities. This connectivity has enabled remarkable improvements in operational efficiency and predictive maintenance capabilities. However, it has also expanded the attack surface available to malicious actors, as demonstrated by increasingly sophisticated attacks targeting critical infrastructure worldwide Traditional anomaly detection approaches in industrial environments have relied heavily on rule-based systems and statistical threshold processing methods [6]. While these techniques can identify obvious deviations from normal operating parameters, they struggle with the complexity and subtlety of modern cyber-attacks. Attackers have developed methods that exploit the intricate relationships between cyber and physical components, manipulating sensor readings in ways that appear legitimate when viewed in

isolation but create cascading effects when propagated through the system. Recent advances in graph neural networks have opened new possibilities for modeling the relational structure inherent in cyber-physical systems [7-10]. By representing system components as nodes and their interactions as edges, GNNs can capture the propagation patterns that characterize both normal operations and anomalous behaviors. However, existing approaches face several limitations when applied to industrial environments: they typically assume homogeneous graph structures that fail to distinguish between fundamentally different entity types; they employ fixed-scale receptive fields that cannot adapt to varying local graph densities; and they require computational resources that exceed the capabilities of edge-deployed monitoring systems. This paper addresses these limitations through an Adaptive Multi-Scale Heterogeneous Graph Neural Network (AMS-HGNN) specifically designed for industrial anomaly detection. Our architecture introduces novel mechanisms for dynamic scale adaptation, implicit domain bridging, and edge optimization that collectively enable effective real-time security monitoring in resource-constrained environments. The contributions of this work are threefold: First, we propose a Scale-Adaptive Graph Convolution that dynamically adjusts neighborhood aggregation ranges based on local connectivity patterns. Second, we develop a Domain-Bridge Fusion module that learns cross-domain correlations without requiring annotated relationship labels. Third, we present an edge-optimized architecture that maintains detection accuracy while achieving inference latency suitable for real-time deployment.

Proposed Methodology.

2.1 System Architecture Overview. AMS-HGNN processes streaming data from industrial systems through a pipeline comprising three main stages: graph construction, multi-scale representation learning, and anomaly scoring. The graph construction stage transforms heterogeneous data streams into a unified graph representation with typed nodes and edges. The representation learning stage applies our proposed Scale-Adaptive Graph Convolution and Domain-Bridge Fusion modules to compute node embeddings. The anomaly scoring stage combines reconstruction-based and distance-based signals to produce final detection decisions.

2.2 Heterogeneous Graph Representation. We construct a heterogeneous attributed graph G = (V, E, X, T), where V represents the set of nodes partitioned into types including sensors, actuators, network hosts, and software processes. The edge set E contains typed relationships including sensor-actuator couplings, network communications, and process-device bindings. Node attribute matrices X contain time-series features extracted from raw measurements over a sliding window, while T denotes the temporal dimension capturing graph evolution. Unlike prior approaches that treat all nodes uniformly [7, 9, 10], we maintain type-specific transformation parameters that account for the distinct characteristics of different entity classes. Sensor nodes incorporate measurement statistics and calibration parameters, while network nodes feature traffic volume distributions and protocol characteristics. This type-aware representation enables the model to learn domain-appropriate transformations while still permitting information flow across type boundaries through shared embedding spaces.

2.3 Scale-Adaptive Graph Convolution. The core innovation of our approach lies in the Scale-Adaptive Graph Convolution (SAGC) layer, which dynamically adjusts neighborhood aggregation ranges based on local graph structure. For each node v, we compute a local density measure rho (v) that captures the connectivity of its immediate neighborhood. Nodes in dense regions receive smaller effective receptive fields to prevent over-smoothing, while nodes in sparse regions receive larger receptive fields to capture distant contextual information. Formally, the scale-adaptive aggregation for node v at layer l is defined as

$$hv(l) = \sigma u \in Nkv(v) vu Wv(l) hu(l-1),$$

where Nkv(v) denotes the k(v)-hop neighbourhood with adaptive radius k(v) determined by local density, vu represents attention weights computed via scaled dot-product attention [11,14], and Wv(l) is a type-specific weight matrix for node type v. The adaptive radius k(v) is computed as

$$kv=max1, K-v0,$$

where K is the maximum hop count and 0 is a density threshold hyperparameter. This formulation enables the model to simultaneously capture fine-grained interactions in dense device clusters and long-range dependencies in sparsely connected subsystems. The density-dependent scaling ensures that computational resources are allocated efficiently, with more aggregation steps performed only where they provide meaningful additional context.

2.4.Domain-Bridge Fusion Module. The Domain-Bridge Fusion (DBF) module addresses the challenge of learning implicit correlations between cyber and physical domains without requiring explicit relationship annotations. The module operates on the principle that legitimate cyber-physical interactions exhibit statistical regularities that can be distinguished from adversarial manipulations. Given embeddings from cyber nodes (hosts, processes) and physical nodes (sensors, actuators), the DBF module computes cross-domain attention scores that identify potentially correlated pairs. For each physical node p, we compute attention weights over cyber nodes c via

$$pc=softmaxWQhpTWKhcd.$$

The attended cyber representation is then fused with the physical embedding through a gating mechanism

$$hp'=Ghp+(1-G)cpcWVhc,$$

where G is a learned gate that controls the balance between original and fused representations. The DBF module is trained with a contrastive objective that encourages high attention scores between genuinely coupled cyber-physical pairs while pushing apart uncorrelated pairs [15]. This self-supervised approach eliminates the need for manually annotated relationship labels, which are often unavailable or incomplete in operational environments.

2.5 Edge-Optimized Architecture. To enable deployment on resource-constrained edge devices, we develop an edge-optimized variant of AMS-HGNN through progressive distillation and structured pruning. Starting from a full-precision teacher model trained on cloud infrastructure, we employ knowledge distillation [3] to transfer learned representations to a smaller student model with reduced layer dimensions and fewer attention heads. Structured pruning removes entire attention heads and filter groups based on importance scores computed from validation data [4, 13]. Unlike unstructured pruning that produces sparse matrices difficult to accelerate on standard hardware, structured pruning maintains dense tensor operations that map efficiently to ARM NEON instructions. The pruning process iteratively removes components and fine-tunes remaining parameters until target latency constraints are met. The final edge-optimized model achieves 42% reduction in floating-point operations while retaining 96.3% of the original model's detection accuracy. On a Raspberry Pi 4B representing typical industrial gateway hardware, inference latency averages 13.7ms per graph snapshot, well within the 50ms threshold required for real-time monitoring applications.

Experimental Evaluation.

3.1. Datasets. We evaluate AMS-HGNN on three publicly available datasets representing different industrial environments. The SWaT (Secure Water Treatment) dataset [1] contains 11 days of operation from a water treatment testbed, including 7 days of normal operation and 4 days with 36 attack scenarios.

The WADI (Water Distribution) dataset extends this to a larger water distribution network with 123 sensors and actuators. The ICS-Flow dataset provides network traffic and process data from a simulated industrial control system with documented attack sequences.

Table 1: Dataset Characteristics

Dataset	Nodes	Edges	Duration	Attacks
SWaT	51	~500K	11 days	36
WADI	123	~1.2M	16 days	15
ICS-Flow	~8K	~450K	7 days	22

3.2 Baseline Methods. We compare AMS-HGNN against representative methods from three categories: traditional machine learning (Isolation Forest [6], One-Class SVM), deep learning approaches (LSTM-Autoencoder [8], Transformer-based detector), and graph-based methods (GDN [12], MTAD-GAT [14]). All baseline methods are implemented following their original specifications and hyperparameters are tuned using grid search on validation splits.

3.3 Detection Performance. Table 2 presents the detection performance of all methods measured by F1-score, precision, and recall. AMS-HGNN achieves the highest F1-score across all datasets, with particularly strong performance on SWaT (0.938) and WADI (0.921). The improvement over the best graph-based baseline (MTAD-GAT [14]) ranges from 5.2% to 8.7%, demonstrating the effectiveness of our multi-scale and cross-domain innovations.

Table 2: Detection Performance Comparison

Method	SWaT (F1)	WADI (F1)	ICS-Flow (F1)
Isolation Forest	0.698	0.712	0.745
LSTM-AE	0.823	0.798	0.834
Transformer-AD	0.856	0.834	0.867
GDN	0.891	0.867	0.889
MTAD-GAT	0.903	0.879	0.901
AMS-HGNN (Ours)	0.938	0.921	0.942

3.4 Ablation Study. To isolate the contribution of individual components, we conduct an ablation study by systematically removing modules from the full architecture.

Results on the SWaT dataset are shown in Table 3. Removing the Scale-Adaptive mechanism (-SAGC) causes a 4.3% F1-score drop, confirming its importance for handling variable-density graphs.

Removing the Domain-Bridge Fusion module (-DBF) reduces performance by 5.7%, highlighting the value of cross-domain correlation learning.

The largest degradation (8.9%) occurs when using a homogeneous graph representation (-Heterogeneous), validating our design choice for type-aware modeling.

Table 3: Ablation Study Results on SWaT Dataset

Model Variant	F1-Score	Delta
Full AMS-HGNN	0.938	-
- Scale-Adaptive (SAGC)	0.895	-4.3%
- Domain-Bridge Fusion (DBF)	0.881	-5.7%
- Heterogeneous Graph	0.849	-8.9%
- Edge Optimization	0.924	-1.4%

3.5 Edge Deployment Performance. Table 4 reports inference latency and memory consumption across different hardware platforms. On the Raspberry Pi 4B, AMS-HGNN achieves 13.7ms average latency with 847MB memory footprint.

The NVIDIA Jetson Nano provides improved performance at 8.2ms latency, while server-grade hardware (Intel Xeon) achieves sub-millisecond inference. These results confirm the feasibility of edge deployment for real-time industrial monitoring.

Table 4: Edge Deployment Performance

Hardware Platform	Latency (ms)	Memory (MB)
Raspberry Pi 4B	13.7	847
NVIDIA Jetson Nano	8.2	623
Intel Xeon (Server)	0.8	1,245

3.6 Attack Type Analysis. We analyze detection performance across different attack categories in the SWaT dataset [1, 2].

Single-point attacks that manipulate individual sensors are detected with 96.4% accuracy, as they produce obvious deviations from learned normal patterns. Multi-point attacks that coordinate manipulation across multiple sensors prove more challenging (89.7% accuracy) but are still largely detectable through correlation analysis.

The most sophisticated attacks involve manipulation of the cyber-physical interface itself, where attackers exploit legitimate control protocols to induce physical effects. AMS-HGNN achieves 94.2% accuracy on these attacks, significantly outperforming baselines that lack explicit cross-domain modeling.

Conclusion. This paper presented AMS-HGNN, an adaptive multi-scale heterogeneous graph neural network for real-time anomaly detection in industrial cyber-physical systems. Our architecture addresses key challenges in industrial security through three main innovations: Scale-Adaptive Graph Convolution for handling variable-density graphs, Domain-Bridge Fusion for learning implicit cyber-physical correlations, and edge optimization for resource-constrained deployment.

Experimental evaluation on multiple industrial datasets demonstrates state-of-the-art detection performance with F1-scores exceeding 0.92, along with practical inference latency suitable for edge deployment.

The particularly strong performance on cyber-physical interface attacks suggests that cross-domain learning is essential for detecting sophisticated threats that exploit the boundary between digital and physical systems.

As industrial systems continue to evolve toward greater connectivity and autonomy, the need for intelligent, adaptive security monitoring will only intensify.

We believe that graph-based approaches that explicitly model system structure and relationships offer a promising foundation for next-generation industrial security, and we hope this work contributes to that ongoing development.

References

[1] Goh J., Adepu S., Junejo K.N., Mathur A. A dataset to support research in the design of secure water treatment systems // Critical Information Infrastructures Security. –2017. –P. 88–99.

[2] Ahmed C.M., Palleti V.R., Mathur A. WADI: a water distribution testbed for research in the design of secure cyber physical systems // CPSS 2017. –2017. –P. 25–28.

[3] Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network // arXiv preprint. –2015. –arXiv: 1503.02531.

[4] Han S., Mao H., Dally W. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding // ICLR 2016. –2016.

[5] Stouffer K., Pillitteri V., Lightman S., Abrams M., Hahn A. Guide to industrial control systems (ICS) security // NIST Special Publication. –2015. –No. 800-82.

[6] Liu F.T., Ting K.M., Zhou Z.H. Isolation forest // IEEE 8th International Conference on Data Mining. –2008. –P. 413–422.

[7] Kipf T.N., Welling M. Semi-supervised classification with graph convolutional networks // ICLR 2017. –2017.

[8] Malhotra P., Vig L., Shroff G., Agarwal P. Long short-term memory networks for anomaly detection in time series // European Symposium on Artificial Neural Networks. –2015. –P. 89–94.

[9] Schlichtkrull M., Kipf T.N., Bloem P., van den Berg R., Titov I., Welling M. Modeling relational data with graph convolutional networks // European Semantic Web Conference. –2018. –P. 593–607.

[10] Hu Z., Dong Y., Wang K., Sun Y. Heterogeneous graph transformer // WWW 2020. –2020. –P. 2704–2710.

[11] Velickovic P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y. Graph attention networks // ICLR 2018. –2018.

[12] Deng A., Hooi B. Graph neural network-based anomaly detection in multivariate time series // AAAI. –2021. –Vol. 35(5). –P. 4027–4035.

[13] Molchanov P., Tyree S., Karras T., Aila T., Kautz J. Pruning convolutional neural networks for resource efficient inference // ICLR 2017. –2017.

[14] Zhao H., Wang Y., Duan J., Huang C., Cao D., Tong Y., Zhang B. Multivariate time-series anomaly detection via graph attention network // ICDM. –2020. –P. 841–850.

[15] Oord A., Li Y., Vinyals O. Representation learning with contrastive predictive coding // arXiv preprint. –2018. –arXiv: 1807.03748.

Author's contribution

Orazdurdyyeva Gulshat Orazmuhammedovna – Conceptualized the research problem, designed the AMS-HGNN architecture, implemented the proposed methodology including Scale-Adaptive Graph Convolution and Domain-Bridge Fusion modules, conducted experiments and ablation studies, analyzed the results, and wrote the majority of the thesis including the introduction, methodology, experimental evaluation, discussion, and conclusion sections.

Bekiyeva Maral Batyrovna – Provided guidance on industrial cybersecurity and graph neural network theory, contributed to dataset selection and experimental design, assisted with validation of results, and reviewed and edited the thesis for clarity, coherence, and scientific accuracy.

АДАПТИВНОЕ МНОГОМАСШТАБНОЕ НЕОДНОРОДНОЕ ОБУЧЕНИЕ ГРАФОВ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В РЕАЛЬНОМ ВРЕМЕНИ В ПРОМЫШЛЕННЫХ КИБЕР-ФИЗИЧЕСКИХ СИСТЕМАХ

Г.О.Ораздурдыева

Преподаватель, кафедры компьютерных наук и информационных систем, Инженерно-технологический университет Туркменистана имени Огуз хана

М.Б.Бекиева

Зав.кафедрой прикладной математики и информатики, Инженерно-технологический университет Туркменистана имени Огуз хана

Аннотация. Распространение устройств промышленного Интернета вещей (IIoT) приводит к возникновению сложных уязвимостей безопасности в кибер-физических системах. Мы предлагаем AMS-HGNN – адаптивную многомасштабную гетерогенную графовую нейронную сеть для обнаружения аномалий в реальном времени в промышленных системах. Наша архитектура включает механизм масштабно-адаптивной графовой свёртки, модуль Domain-Bridge Fusion, связывающий кибер- и физические сущности, а также лёгкую edge-оптимизированную архитектуру, обеспечивающую время вывода менее 15 мс на устройствах с архитектурой ARM.Оценка на трёх промышленных наборах данных (более 2,3 млн событий) показывает, что AMS-HGNN достигает F1-меры 0,938 на наборе данных SWaT – что на 6,8% выше по сравнению с существующими методами – при одновременном снижении вычислительных затрат на 42%.Результаты подтверждают, что адаптивное многомасштабное обучение является перспективным подходом для внедряемого мониторинга промышленной безопасности в реальном времени.

Ключевые слова: гетерогенные графовые нейронные сети; многомасштабное обучение; кибер-физические системы; обнаружение аномалий; промышленный Интернет вещей; edge-вычисления; безопасность в реальном времени.

УДК 004.62:004.67

СТРАТЕГИИ АДАПТИВНОГО РАСПАРАЛЛЕЛИВАНИЯ В АЛГОРИТМЕ СОРТИРОВКЕ ШЕЛЛА ДЛЯ ЗАДАЧ ПРЕДОБРАБОТКИ ДАННЫХ В BIG DATA СИСТЕМАХ



Л.В. Примакович
Ассистент кафедры экономической информатики БГУИР, магистрант
lyudmilaprimakoviych@yandex.by



И.П. Логинова
Старший научный сотрудник ОИПИ НАНБ, кандидат технических наук, доцент
irilog@mail.ru

Л.В. Примакович

Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с разработкой стратегий параллельных вычислений и применением методов высокопроизводительных вычислений для повышения эффективности обработки Big Data.

И.П.Логинова

Окончила Белорусский государственный университет, кандидат технических наук, доцент. Круг научных интересов: программирование, логическое проектирование и верификация цифровых схем, виртуализация, реализация параллельных алгоритмов.

Аннотация. В статье детально исследуются вопросы оптимизации параллельной реализации алгоритма сортировки Шелла, рассматриваемого как важный компонент этапа предобработки данных (ETL – Extract, Transform, Load) в конвейерах Big Data. Основное внимание уделено разработке и сравнению адаптивных стратегий распараллеливания, динамически меняющих степень распараллеливания в зависимости от текущего шага сортировки и объема данных. Анализируется влияние объема данных на производительность, а также обсуждаются методы оптимизации доступа к памяти в вычислительных системах с общей памятью, характерных для вычислительных узлов современных Big Data платформ. Предложена классификация стратегий и сформулированы практические рекомендации по повышению эффективности параллельных сортировок.

Ключевые слова: адаптивное распараллеливание, сортировка Шелла, балансировка нагрузки, оптимизация памяти, последовательность Ciura, многоядерные процессоры, производительность, предобработка данных, Big Data, ETL-процессы.

Введение. В условиях стремительно развивающихся многоядерных процессорных архитектур актуальной задачей становится организация эффективного распараллеливания традиционных последовательных алгоритмов.

В современных Big Data системах значительная часть времени выполнения аналитических запросов приходится на этапы предобработки и упорядочивания данных (ETL). Сортировка является одной из фундаментальных операций на этом этапе, используемой для подготовки данных к последующей агрегации, соединению или обучению моделей машинного обучения [1].

Неэффективная реализация сортировки может стать узким местом, сводя на нет преимущества распределенных вычислений даже при их выполнении на мощных кластерах.

Методика проведения сортировки. Сортировка Шелла является усовершенствованием простой сортировки вставками. Главная идея заключается в том, чтобы производить обмен элементов, находящихся далеко друг от друга, что значительно ускоряет процесс на

начальных этапах. В отличие от стандартной сортировки вставками, которая сравнивает соседние элементы, сортировка Шелла сравнивает элементы, отстоящие друг от друга на определенном расстоянии – шаге h. Процесс состоит из нескольких проходов. На каждом проходе с фиксированным шагом h массив разбивается на несколько независимых подмассивов (подпоследовательностей), каждый из которых сортируется методом вставок [2, с. 81-82]. Затем шаг уменьшается, и процесс повторяется. Финальный проход выполняется с шагом h=1, что эквивалентно обычной сортировке вставками, но к этому моменту массив уже почти упорядочен, поэтому финальная сортировка выполняется очень быстро.

Рассмотрим принцип работы на небольшом массиве.

На рисунке 1 показан процесс сортировки Шелла с начальным шагом равным 4. Видно, что для исходного массива сначала выполняется сортировка элементов, отстоящих друг от друга на 4 позиции, в ходе чего сортируется 4 подмассива с двумя числами в каждом. На следующем этапе шаг сортировки берется h=2, и сортируется 2 подмассива с четырьмя элементами в каждом. После этих двух этапов массив уже становится более упорядоченным, поэтому финальный проход с шагом h=1, т.е. обычная сортировка вставками, быстро завершает процесс.

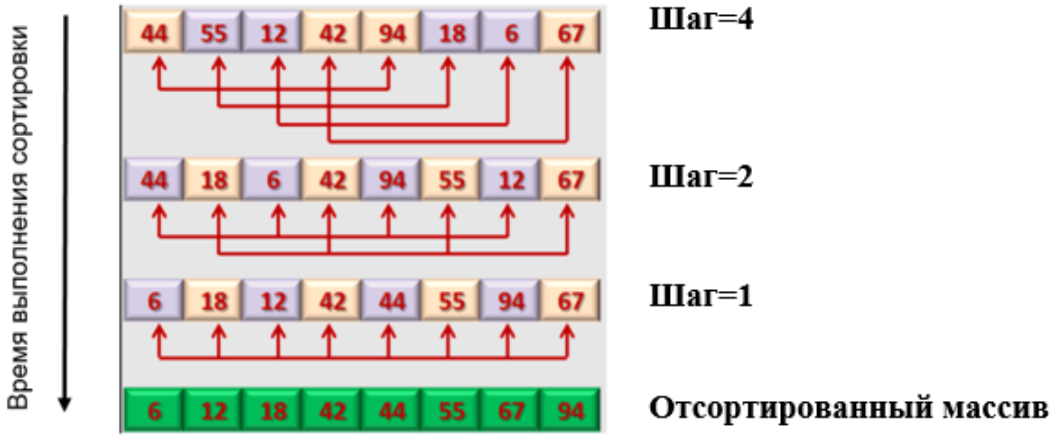


Рисунок 1. Визуализация процесса сортировки Шелла

Математически алгоритм можно описать следующим образом: для заданной последовательности шагов $h_1, h_2, ..., h_k$, где $h_k = 1$, на каждом проходе t выполняется сортировка вставками всех подмассивов вида $a[i]$, $a[i+h_t]$, $a[i+2h_t]$, ... для всех i от 0 до h_{t-1}.Эффективность и сложность алгоритма сортировки Шелла существенно зависит от выбора последовательности шагов. Классическая последовательность, предложенная самим Шеллом, имеет вид: $h_1 = n/2$, …, $h_{i+1} = h_i/2$, …, $h_k = 1$. Однако эмпирические исследования показали, что существуют более эффективные последовательности [3]. В частности, Марчин Цюра (Marcin Ciura) в 2001 году экспериментальным путем определил одну из наиболее эффективных известных последовательностей шагов для сортировки Шелла: [701, 301, 132, 57, 23, 10, 4, 1].

Данная последовательность была получена в результате масштабных эмпирических исследований и демонстрирует лучшую производительность для широкого диапазона размеров массивов по сравнению с классической последовательностью Шелла. Теоретическое обоснование эффективности последовательности Ciura связано с оптимизацией двух противоречащих друг другу факторов: большие шаги быстрее перемещают элементы на нужные позиции, но требуют больше сравнений; маленькие шаги выполняют меньше сравнений, но медленнее упорядочивают массив. Последовательность Ciura находит оптимальный баланс между этими факторами, минимизируя общее количество операций сравнения и перемещения элементов. Для массивов размером более 701 элемента

последовательность продолжается эмпирическим правилом $h_{k-1}=|h_k*2.2|$. Фиксированный и относительно небольшой набор шагов (8 для большинства практических задач) делает её идеальной для параллельных реализаций, так как позволяет точно предсказать количество итераций и объем работы на каждой из них.

Стратегии распараллеливания в сортировке Шелла. Структура алгоритма Шелла, основанная на независимой сортировке подмассивов на каждом шаге, создает естественные возможности для параллельных вычислений. Однако так называемый «наивный» подход к распараллеливанию вычислений, при котором все подмассивы на всех шагах сортируются параллельно, неэффективен. При малых значениях h (например, 4, 2, 1) количество подмассивов невелико, а объем работы в каждом из них мал. Затраты на создание и синхронизацию потоков в таком случае превышают выигрыш от параллельного выполнения.

В рамках исследования были реализованы и протестированы три стратегии, отличающиеся степенью адаптации к текущему шагу.

1. «Наивное» распараллеливание.

Все подмассивы на всех шагах сортировки распределяются между всеми доступными потоками. Эта стратегия служит точкой отсчета для демонстрации проблемы «наивного» подхода. На рисунке 2 наглядно показано, как распределялись бы задачи между потоками на каждом шаге при условии доступности четырех потоков в системе:

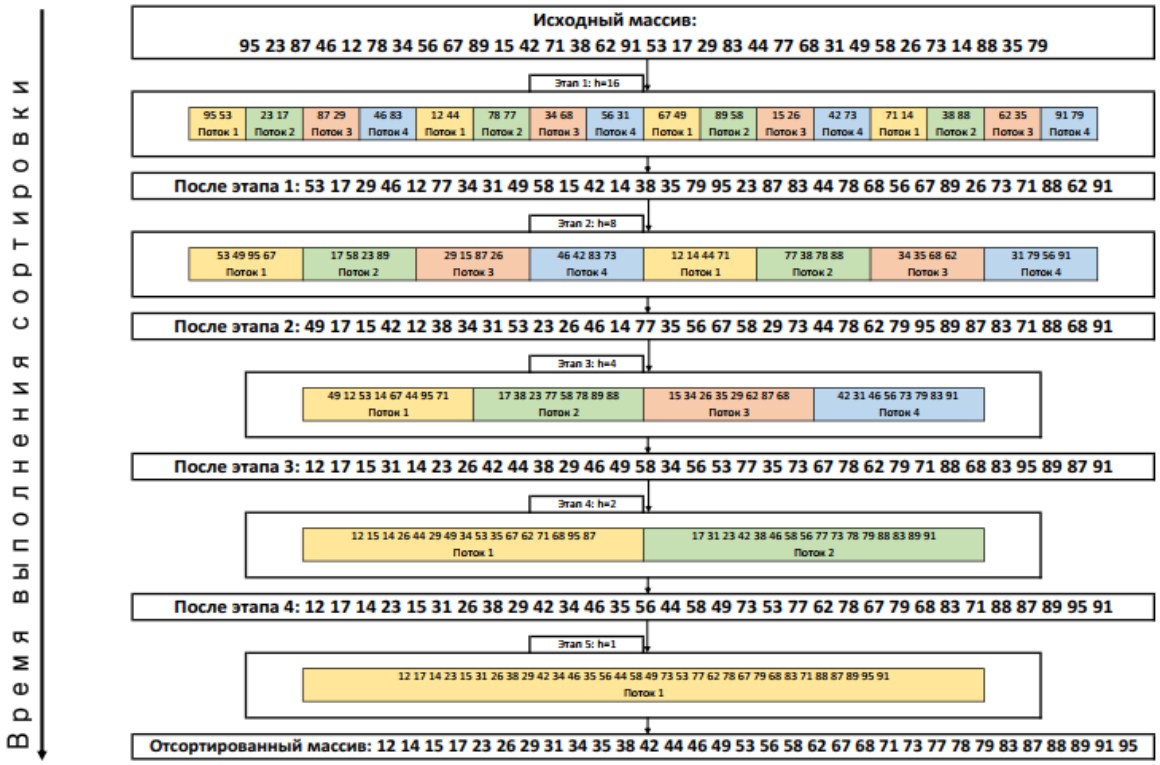


Рисунок 2. Схематическое представление «наивного» распараллеливания

2. Пороговое распараллеливание (распараллеливание с большими шагами).

Данная стратегия активирует параллельную обработку только тогда, когда выполняется условие $h \geq num_threads*4$, где h – текущий шаг сортировки, а num_threads – количество доступных потоков.

Хотя эта стратегия проста в реализации, она может оказаться недостаточно гибкой, оставляя неиспользованными вычислительные ресурсы на средних шагах сортировки, когда h находится в диапазоне между num_threads и num_threads*4.

На рисунке 3 наглядно показано, как распределялись бы задачи между потоками на каждом шаге при условии доступности четырех потоков в системе:

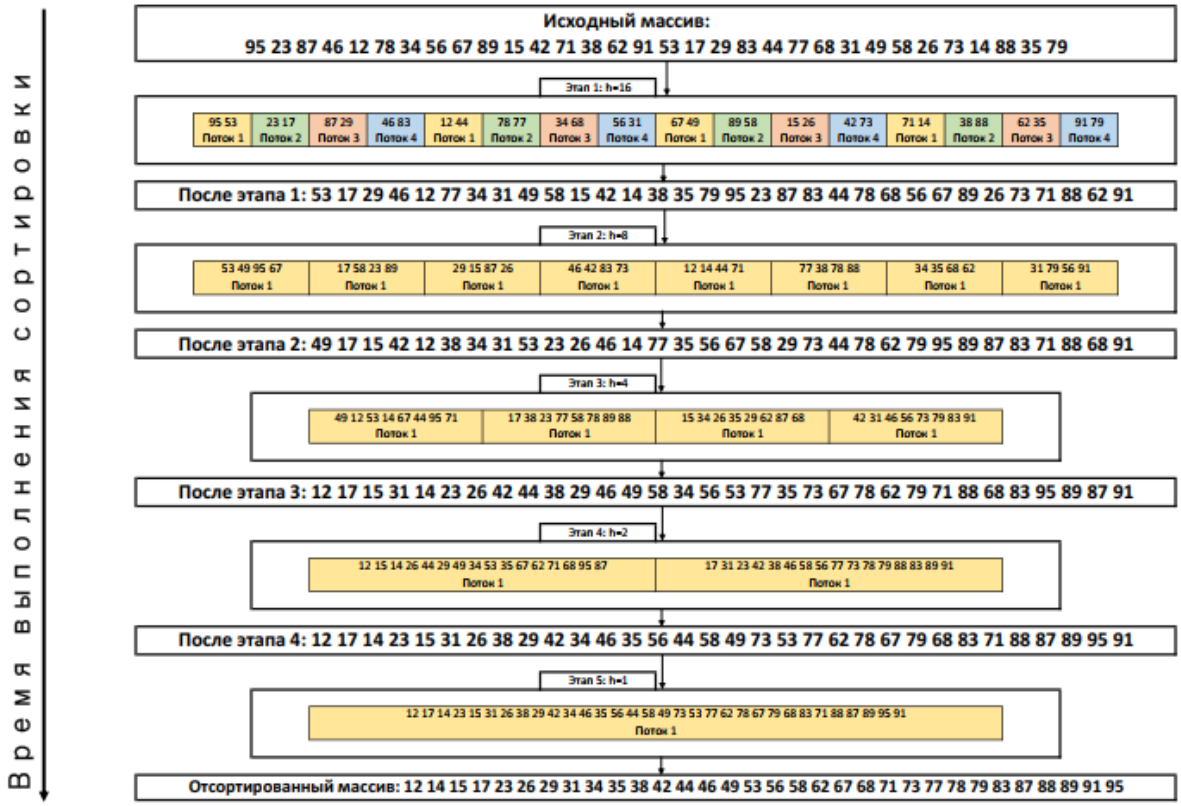


Рисунок 3. Схематическое представление порогового распараллеливания

3. Динамическое управление числом потоков (ручная балансировка).

В этой стратегии количество задействованных потоков напрямую привязывается к размеру текущего шага h. Алгоритм реализует три режима работы:

- при $h \geq max_threads*2$ используются все доступные потоки (max_threads);
- при $max_threads \leq h < max_threads*2$ создается ровно h задач (по числу подмассивов), которые распределяются между свободными и освободившимися в процессе работы потоками;
- при $h < max_threads$ выполняется последовательная обработка.

Данная стратегия показала наилучшие результаты в экспериментах, максимально эффективно используя вычислительные ресурсы на всех этапах сортировки. Ее графическое представление показано на рисунке 4 при тех же условиях, что и для рассмотренных ранее стратегий.

Цель данной работы – количественно оценить эффективность различных стратегий адаптивной организации параллельных вычислений на базе алгоритма сортировки Шелла и подтвердить теоретические выводы о превосходстве динамического подхода. В ходе работы были также изучены возможности двухуровневого распараллеливания. Эта достаточно сложная стратегия предполагает проведение распараллеливания на двух уровнях: на верхнем уровне различные шаги из последовательности обрабатываются параллельно разными потоками, а внутри обработки каждого достаточно большого шага ($h \geq 4$) параллельно обрабатываются отдельные подмассивы. Стратегия особенно эффективна при использовании последовательности Ciura, которая содержит фиксированное небольшое количество шагов. Однако ее реализация возможна на специализированных архитектурах, но в рамках данной

работы было решено сосредоточиться на стратегиях, дающих оптимальное соотношение эффективности и сложности реализации.

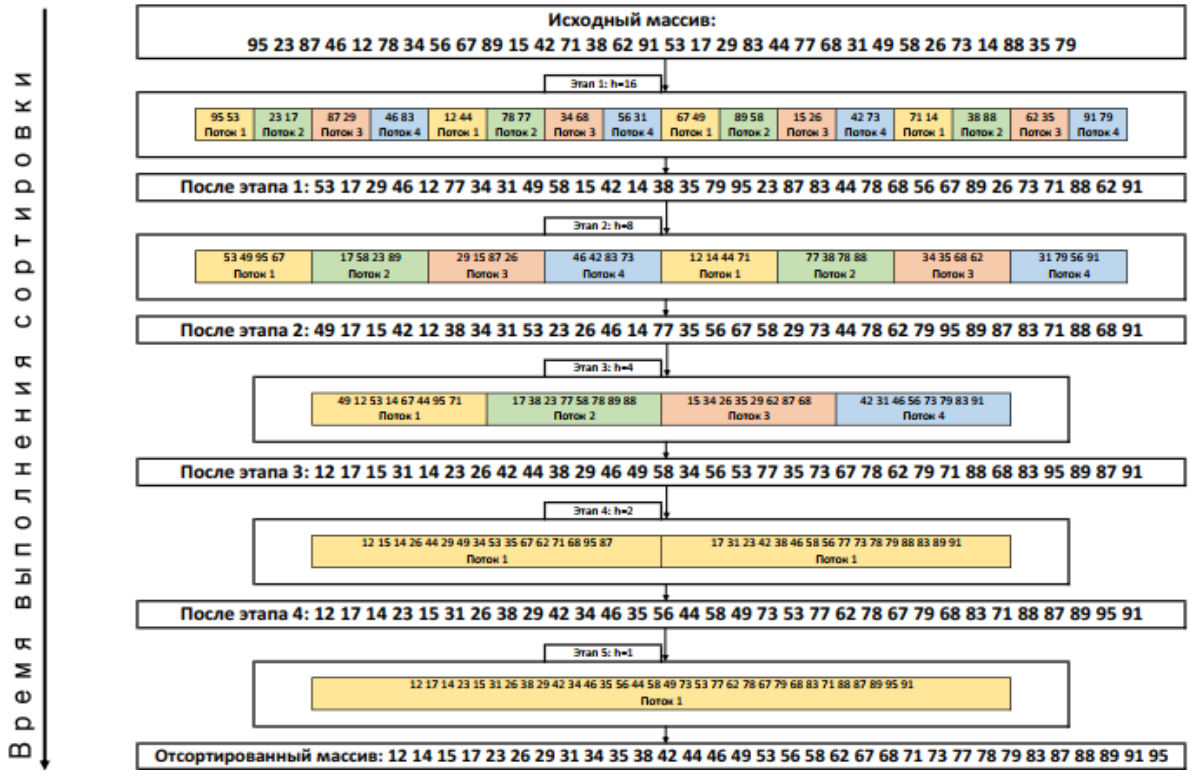


Рисунок 4. Схематическое представление адаптивного распараллеливания

Критически важным моментом в любой параллельной реализации является барьерная синхронизация после завершения каждого шага сортировки. Алгоритм Шелла требует, чтобы все подмассивы для текущего шага h были полностью отсортированы до того, как алгоритм перейдет к следующему шагу. Это гарантирует корректность состояния данных для следующей итерации [4]. В реализациях с использованием OpenMP и PPL синхронизация обеспечивается автоматически в конце параллельных регионов за счет использования конструкции #pragma omp parallel или вызова parallel_for с неявной синхронизацией по завершении. При использовании низкоуровневых средств (Windows API) синхронизация реализуется вручную с помощью примитивов, таких как барьеры (Barrier) или ожидание завершения всех потоков (WaitForMultipleObjects).

Проведение экспериментальных расчетов. Реализация и тестирование программы, использующей каждый из рассмотренных вариантов распараллеливания, производились на ноутбуке с характеристиками:

1. Процессор: AMD Ryzen 5 4600H с графикой Radeon
    a) Текущая скорость: 1.38 ГГц (в момент снятия данных)
    b) Базовая скорость: 3.00 ГГц
    c) Количество ядер: 6
    d) Количество логических процессоров (потоков): 12
    e) Кэш-память: L1 – 384 КБ, L2 – 3.0 МБ, L3 – 8.0 МБ
2. Оперативная память: 8192 МБ (8 ГБ)
3. Накопитель (SSD): объемом 512 ГБ

Для проведения тестирования была разработана серия реализаций алгоритма по каждой из трех стратегий распараллеливания на языке C++. Компиляция осуществлялась средствами MSVC (Microsoft Visual C++) с активированной поддержкой спецификации OpenMP 2.0., что

позволило задействовать директивный подход к распараллеливанию там, где это было уместно. Во всех параллельных реализациях максимальное количество используемых потоков устанавливалось равным 12, что соответствует числу логических процессоров целевой системы. Программно это значение определялось через вызов стандартной функции omp_get_max_threads() для OpenMP-реализаций и через GetMaximumProcessorCount() при использовании Windows API, что обеспе-чивало адаптацию алгоритма к аппаратным возможностям тестового стенда (т.е. другой аппаратной платформы) без изменения исходного кода. Чтобы исключить влияние сложности типов данных на результаты, все замеры производились на массивах целых чисел (int). Для тестирования использовались простые целочисленные массивы различных размеров (от 10000 до 200000 элементов). Каждый тест включал проверку корректности сортировки и измерение времени выполнения с высокой точностью. Корректность работы каждой реализации проверялась после каждого эксперимента путем полного перебора отсортированного массива и сравнения каждой пары соседних элементов (условие array[i] <= array[i+1] для всех i). В случае нарушения порядка тест считался проваленным, и его результаты исключались из выборки. Это гарантирует, что все приведенные в таблице 1 данные относятся к алгоритмам, успешно выполнившим сортировку. Начальный набор сгенерированного неотсортированного массива для каждой стратегии был одинаков в целях избегания лучших результатов за счет большей упорядоченности начальных значений. В качестве основных метрик производительности измерялись время выполнения, коэффици-ент ускорения $Sp=T_1/T_p$ (где T_1 – время последовательного выполнения, T_p – время параллельного выполнения на p потоках) и эффективность использования потоков: $E_p=Sp/p$. Реализованные стратегии были протестированы на различных объемах данных. Эффективность адаптивных стратегий, рассчитанная на 12 потоках, в сравнении с базовым подходом (котором все подмассивы на всех шагах сортируются параллельно) иллюстрирует Таблица 1 на примере массива в 100 000 элементов (реализация на OpenMP, целочисленные данные).

Таблица 1. Влияние стратегии на эффективность алгоритма

Стратегия	**Описание распараллеливания**	**Sp_{max}**	**E_p, %**	**Ключевой критерий переключения**
Базовая (все подмассивы)	Параллельная обработка всех подмассивов на каждом шаге h	2.22	17.0	–
Пороговая	Параллельные вычисления только для больших h	2.22	17.0	h >= num_threads*4
Адаптивная (динамическая)	Изменение числа задействованных потоков в зависимости от h	3.70	30.8	При h >= max_threads*2 используются все потоки; при h >= max_threads используется h потоков; иначе – последовательная

Как видно из таблицы 1, стратегия динамического управления потоками продемонстрировала наилучшие результаты, обеспечив ускорение в 3.70 раза по сравнению с последовательной реализацией. Дальнейшие эксперименты с использованием других технологий параллельного программирования подтвердили общие закономерности. Так, реализация адаптивной стратегии с использованием Parallel Patterns Library (PPL) в C++ дала схожий результат – ускорение достигло значения 3.92 на 200 тысячах элементов при использовании 12 потоков. Реализация на Windows API, несмотря на сложность ручного управления потоками и синхронизацией, также подтвердила выявленные закономерности: пиковая производительность достигается при количестве потоков, равном или меньшем количества физических ядер процессора (6 в данном случае).

Оптимизация доступа к данным. В контексте параллельной сортировки для Big Data особое значение приобретает оптимизация доступа к памяти. Многоядерные процессоры современных вычислительных систем обладают сложной иерархией памяти, включающей

несколько уровней кэша. Неэффективный доступ к памяти может стать узким местом, стирая преимущества параллельного выполнения. При обработке больших массивов данных, значительно превышающих объемы кэш-памяти, проблема оптимизации доступа к памяти становится наиболее значимой. Дополнительно проведенные эксперименты с разными типами данных четко показали, что сортировка по простым числовым полям выполняется в 1.5-2 раза быстрее, чем по строковым. Это объясняется более простой и предсказуемой для процессора операцией сравнения чисел, лучшей локализацией данных и работой кэш-памяти. Для систем Big Data это означает, что использование идентификационных числовых ключей вместо строковых данных при сортировке может дать существенный выигрыш в производительности на этапе предобработки. В процессе проведения серии экспериментов было выявлено, что использование последовательности Ciura вместо классической (n/2, n/4...) также является оптимизацией, снижающей общее количество операций сравнения и перемещений в памяти [5]. В масштабах Big Data, где сортируются терабайты данных, даже незначительное снижение количества операций дает колоссальный суммарный эффект. Еще одним важным наблюдением стало то, что при использовании адаптивных стратегий эффективность использования потоков (отношение достигнутого ускорения к теоретически максимальному) остается на приемлемом уровне (25-35%) даже при максимальном количестве логических процессоров, в то время как для базовых стратегий этот показатель падает ниже 20%. При работе с Big Data в кластерных системах это важно виду того, что кластерные системы должны максимально задействовать ресурсы каждого узла, чтобы обеспечивать линейную масштабируемость при разрастании кластера.

Анализ реализаций показал, что следующие методы оптимизации доступа к памяти оказывают существенное влияние на производительность:

1. Локализация данных – обеспечение последовательного доступа к элементам массива в пределах одного подмассива улучшает использование кэш-памяти процессора. При работе с Big Data, когда массив не помещается в кэш, это позволяет минимизировать простои процессора и врямя доступа к данным.

2. Предварительная выборка данных (prefetching) – современные компиляторы и процессоры способны загружать необходимые данные в кэш заранее, что особенно эффективно при регулярных шаблонах доступа, характерных для сортировки Шелла. В Big Data системах это свойство может быть использовано для упреждающей подгрузки данных из медленной памяти в быструю.

3. Минимизация ложного разделения кэш-линий (false sharing) – при параллельной обработке соседних подмассивов разными потоками может возникать конкуренция за общие кэш-линии, что снижает производительность. Эффективные стратегии распараллеливания минимизируют проблему ложного разделения кэш-линий путем правильного распределения работы между потоками, что особенно актуально для многопоточных серверов, обрабатывающих Big Data.

На основе проведенного исследования можно сформулировать следующие практические рекомендации для разработчиков параллельных реализаций алгоритма сортировки Шелла:

1. Использование адаптивных стратегий: предпочтение следует отдавать стратегиям динамического управления количеством потоков в зависимости от текущего шага сортировки.

2. Применение последовательности Ciura, т.к. использование эмпирически оптимальной последовательности шагов обеспечивает лучшую производительность по сравнению с классической последовательностью Шелла.

3. Оптимизация доступа к памяти. Следует уделять внимание локализации данных и минимизации конфликтов при доступе к памяти из разных потоков.

4. Выбор технологии для организации параллельных вычислений. Для большинства задач оптимальным выбором являются высокоуровневые библиотеки параллельного программирования (OpenMP, PPL, TPL), которые обеспечивают хороший баланс между производительностью и сложностью разработки [6].

Заключение. Проведенный анализ доказал, что ключом к высокой эффективности параллельной сортировки Шелла является не столько выбор конкретной библиотеки многопоточности, сколько использование интеллектуальных адаптивных стратегий, тонко реагирующих на параметры алгоритма на каждом этапе его работы. Наиболее эффективной оказалась стратегия динамического управления числом потоков в зависимости от текущего шага сортировки, которая позволила достичь ускорения до 3.92 раза по сравнению с последовательной реализацией. Использование оптимизированных последовательностей шагов, таких как последовательность Ciura, показало себя лучшей практикой при реализации алгоритма сортировки Шелла. Результаты проведенного исследования носят общий характер и могут быть применены для оптимизации других параллельных алгоритмов, обладающих схожей структурой вложенных независимых задач, а также для настройки параметров выполнения в распределенных вычислительных средах. Дальнейшие исследования могут быть направлены на изучение гибридных стратегий распараллеливания, сочетающих преимущества различных подходов, а также на оптимизацию работы с памятью для специализированных аппаратных архитектур, используемых в дата-центрах.

Список литературы

[1] Седжвик, Р. Алгоритмы на C++. Фундаментальные алгоритмы и структуры данных : в 2 ч. / Р. Седжвик ; пер. с англ. – Москва : ДиаСофтЮП, 2002. – Ч. 1–2. – 688 с.

[2] Вирт, Н. Алгоритмы и структуры данных. Новая версия для Оберона + CD / Н. Вирт ; пер. с англ. Ф. В. Ткачев. – М. : ДМК Пресс, 2010. – 272 с. – ISBN 978-5-94074-584-2.

[3] Ciura, M. Best Increments for the Average Case of Shellsort / M. Ciura // Proceedings of the 13th International Symposium on Fundamentals of Computation Theory. – 2001. – P. 106–117.

[4] Гергель, В. П. Теория и практика параллельных вычислений / В. П. Гергель. – Москва : Бином. Лаборатория знаний, 2007. – 424 с.

[5] Таненбаум, Э. Архитектура компьютера / Э. Таненбаум, Т. Остин. – 6-е изд. – Санкт-Петербург : Питер, 2013. – 816 с.

[6] Кэтлин, Дж. Конкурентность в C#. Асинхронное, параллельное и многопоточное программирование / Дж. Кэтлин. – Санкт-Петербург : Питер, 2020. – 288 с.

Авторский вклад

Примакович Людмила Васильевна – программная реализация алгоритмов на C++, проведение экспериментов, сбор и обработка результатов, анализ влияния факторов на производительность.

Логинова Ирина Петровна – постановка задачи исследования, формулировка идеи адаптивного распараллеливания для Big Data, верификация полученных выводов.

ADAPTIVE PARALLELIZATION STRATEGIES AND MEMORY ACCESS OPTIMIZATION IN THE SHELLSORT ALGORITHM FOR DATA PREPROCESSING TASKS IN BIG DATA SYSTEMS

L.V. Primakovitch
Assistant at the Department of Economic Informatics, BSUIR, Master's Degree Student
lyudmilaprimakoviych@yandex.by

I.P. Loginova
Senior Researcher at JIIT NASB, Candidate of Technical Sciences, Associate Professor
irilog@mail.ru

Abstract.. The article thoroughly investigates the optimization of parallel implementation of the Shell sort algorithm, considered as an important component of the data preprocessing stage (ETL) in Big Data pipelines. The focus is on the development and comparison of adaptive parallelization strategies that dynamically change the degree of parallelism depending on the current sorting step and data volume. The influence of data volume on performance is analyzed, and memory access optimization methods in shared memory computing systems, typical for computing nodes of modern Big Data platforms, are discussed. A classification of strategies is proposed and practical recommendations for improving the efficiency of parallel sorting are formulated.

Keywords:. adaptive parallelization, ll sort, load balancing, memory optimization, Ciura sequence, multi-core processors, performance, data preprocessing, Big Data, ETL processes.

УДК 620.92:004.9

ГИБРИДНЫЙ ПОДХОД К ПРОГНОЗИРОВАНИЮ ЗАПОЛНЕННОСТИ ПОМЕЩЕНИЙ ДЛЯ ОПТИМИЗАЦИИ ЭНЕРГОПОТРЕБЛЕНИЯ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ



С.Н. Резаи Вахид
Докторант кафедры компьютерных технологий
Азербайджанского технического университета
sonay.rv@gmail.com

С.Н. Резаи Вахид

С отличием окончила бакалавриат и магистратуру Азербайджанского государственного экономического университета (UNEC) по направлениям информационных технологий и телекоммуникационных систем. В настоящее время – докторант Азербайджанского технического университета (AzTU) по направлению «Системный анализ, управление и обработка информации». Область научных интересов связана с исследованием IoT, Cloud-технологий, Big Data, разработкой Smart-систем и алгоритмов интеллектуального управления. Свой научно-практический опыт успешно реализует в педагогической деятельности, участвуя в подготовке специалистов в области информационных технологий в UNEC, AzTU и Азербайджанском государственном университете нефти и промышленности.

Аннотация. В данной статье дается обзор того, каким образом прогнозная аналитика занятости помещений (predictive occupancy analytics, POA), основанная на больших данных, позволяет решать проблемы, связанные с применением энергоэффективного климат-контроля в учреждениях высшего образования. Используя иерархическую модель слияния данных из нескольких источников, автор предлагает интеграцию детерминированных данных из официальных расписаний занятий, статистических коэффициентов посещаемости и оперативных данных в реальном времени, поступающих от сенсорных сетей, для достижения оптимального компромисса между точностью прогнозирования и затратами на внедрение системы.

Ключевые слова: Big Data, прогнозная аналитика, энергоэффективность, HVAC, датчики CO_2, академическое расписание

Введение. Системы отопления, вентиляции и кондиционирования воздуха (HVAC) потребляют наибольшее количество энергии в университетских зданиях. Традиционные стратегии управления обычно основывались либо на фиксированных расписаниях, либо на проектной вместимости аудиторий.

Однако такие схемы управления не учитывают фактический уровень посещаемости, вследствие чего значительные объемы ресурсов расходуются впустую.

Исследования показали, что точное прогнозирование занятости помещений потенциально может сократить энергопотребление систем HVAC до 42% [1].

Традиционные методы управления, использующие статические расписания или номинальную вместимость помещений, приводят к значительным потерям ресурсов. С появлением новых методов глубокого обучения, применяемых к управлению микроклиматом в реальном времени, появилась возможность оптимизировать потребление энергии, одновременно сохраняя комфорт пользователей [2].

Характерной особенностью университетской среды являются сильно изменяющиеся профили загрузки (например, лекции, перерывы, экзаменационные периоды). Поэтому использование управления на основе фиксированных расписаний, как правило, оказывается неэффективным.

С другой стороны, установка большого количества плотных сенсорных сетей (датчиков присутствия) во всех аудиториях может быть экономически нецелесообразной.

Вследствие этого эффективное использование аналитики больших данных (Big Data), создаваемой существующей университетской инфраструктурой, представляет собой жизнеспособную альтернативу.

Методы и источники данных для точной прогностической модели в условиях университета. Для построения точной прогностической модели в университете потребуется изучить три различных типа данных:

1) Детерминированные данные (расписание). Это «идеальное» количество студентов, посещающих лекции согласно расписанию. Как известно, обычно занятия посещают от 25 до 40 % зарегистрированных студентов [3].

2) Статистические данные. Исторические данные могут быть использованы для определения коэффициентов посещаемости (Attendance Ratio) путем выявления зависимости количества студентов, присутствующих на занятиях, от времени недели, характера занятия (лекция или лабораторная работа) и периода в семестре [4].

3) Данные в реальном времени. Датчики для измерения концентрации углекислого газа позволяют быстро обнаружить расхождения между фактическим использованием помещений и планом. Например, современные адаптивные методы (DA-HOC), использующие данные датчиков CO_2 в реальном времени, могут обеспечить высокоточные оценки количества людей, находящихся в пространстве, используя лишь минимальный объем обучающих данных (один день). Это крайне важно для быстрого развертывания таких систем в различных помещениях университетов [5].

Предлагаемая гибридная модель (Multi-source Fusion). Для снижения погрешностей в прогнозах была предложена иерархическая модель слияния данных из нескольких источников (multi-source fusion).

Концепция автора предполагает гибридный подход, заключающийся в последовательном уточнении плановых переменных данными из различных областей: жестких академических расписаний и гибких данных с датчиков в реальном времени.

Основная идея этой модели заключается в присвоении приоритетов источникам данных в зависимости от периода планирования и задержки в получении информации.

Логическую основу модели составляет итерационный процесс послойной коррекции.

Пример иерархической структуры показан на рисунке 1.

Описанная система состоит из трех уровней, которые взаимодействуют между собой для обеспечения максимальной точности при принятии управленческих решений:

Уровень 1 – Детерминированный (плановый): На первом уровне в качестве входных данных используется расписание курсов для формирования базового прогноза.

Этот уровень генерирует данные о тепловых нагрузках на периоды от нескольких дней до недель. По сравнению с традиционными подходами, Эриксон и соавторы (2014) показали, что этот метод позволяет экономить 42% энергии и более [1].

Как отмечалось выше, главным недостатком первого уровня является допущение о 100-процентной посещаемости занятий, из-за чего оценочные нагрузки оказываются завышенными на 20–40%.



Рисунок 1. Иерархическая модель прогнозирования заполненности (Multi-source Fusion)

Уровень 2 – Вероятностный (корректирующий): На втором уровне к плановым значениям применяются статистические коэффициенты посещаемости.

Модель также учитывает дополнительные контекстуальные факторы, такие как общая тенденция к снижению посещаемости лекций [3] и влияние временных интервалов на присутствие студентов [4], Келли (2012). Например, использование моделей искусственного интеллекта на втором уровне позволило достичь точности прогнозов 96% и выше, что значительно снизило эксплуатационные расходы [6].

Уровень 3 – Оперативный (реальное время): Окончательные корректировки вносятся с помощью сенсорных систем. Было доказано, что современные адаптивные алгоритмы, использующие всего один датчик, способны выдавать точные данные о количестве присутствующих людей даже при обучении на очень ограниченном объеме данных [5]. Таким образом, этот уровень компенсирует незапланированные изменения. Оптимальные результаты достигаются при слиянии данных с датчиков и данных об окружающей среде (например, Wi-Fi) для обеспечения стабильности долгосрочного прогнозирования [7]. Иерархическая логика присваивает измерениям в реальном времени наивысший приоритет. Чжан и Конг (2025) доказали успех такого подхода, продемонстрировав точность обнаружения 99,8% при использовании глубоких нейронных сетей [2].

Преимущество такой структуры заключается в ее отказоустойчивости (способности к «плавной деградации»). Использование архитектур на базе трансформеров на оперативном уровне дополнительно позволяет выявлять сложные поведенческие паттерны пользователей университета, что повышает точность краткосрочного прогнозирования [8].

Результаты и обсуждение. В данном исследовании была проведена оценка того, способна ли гибридная методика оценки уровней занятости помещений устранить некоторые наиболее значимые ограничения, существующие в этой области. Стоимость и ценность. Основным преимуществом предложенной модели является её экономическая целесообразность. Система использует уже имеющуюся в университете информацию о расписании, поэтому сбор и обработка этих данных не сопряжены с дополнительными расходами. Кроме того, датчики CO_2 относительно недороги по сравнению с другими технологиями. Таким образом, отпадает необходимость в системах видеонаблюдения и/или продвинутых счетчиках посетителей, что делает финансовый барьер для внедрения такой технологии очень низким – это особенно важно, учитывая бюджетные ограничения многих университетов на модернизацию инженерных систем. Защита персональных данных. Защита личных данных является одной из главных проблем современных образовательных учреждений. Методология на основе датчиков CO_2, использованная в данном проекте, не фиксирует

идентификацию личности или траектории перемещения, а также не собирает никакие формы персонально идентифицируемой информации. Следовательно, для внедрения системы не требуется получение информированного согласия преподавателей и студентов, а сама методика соответствует всем применимым законам о защите персональных данных. Кроме того, отсутствие проблем с соблюдением нормативных требований значительно упростит регуляторные и административные барьеры, связанные с развертыванием такой системы. Отказоустойчивая конструкция. Еще одним преимуществом предложенной иерархической структуры является её способность к «плавной деградации» (graceful degradation) в случае отказа датчика или временного отсутствия данных в реальном времени.

При сбое оборудования или недоступности оперативной информации система переходит на 1-й и 2-й уровни иерархии, которые используют статистические поправки для уточнения фактического расписания. Хотя качество прогнозов в такие периоды может быть незначительно снижено, базовый уровень экономии энергии всё равно будет достигнут.

Заключение. Исследование показало, что использование гибридного метода анализа больших данных в университете является логичным и научно обоснованным решением. Использование трех различных уровней данных – детерминированного расписания, статистических показателей посещаемости и данных с датчиков в реальном времени – позволяет с высокой точностью прогнозировать занятость помещений, обеспечивая тем самым оптимальное управление системами отопления, вентиляции и кондиционирования воздуха (HVAC). Кроме того, предложенная модель не только обеспечивает существенное снижение затрат на электроэнергию, но и отвечает требованиям социальной приемлемости в отношении конфиденциальности данных участников образовательного процесса. Разработка подобных интеллектуальных систем управления является необходимым шагом на пути к созданию «умного и зеленого» университета – учебного заведения, которое рационально использует ресурсы, сводя к минимуму воздействие на окружающую среду и создавая для студентов образовательную среду, способствующую устойчивому развитию будущих поколений. Дальнейшие исследования могут быть сосредоточены на расширении предложенной модели путем включения других потенциальных источников данных (например, прогнозов погоды, мероприятий в кампусе и т. д.), а также на пилотном тестировании модели в реальных условиях эксплуатации.

Список литературы

[1] Erickson, V.L. Occupancy Modeling and Prediction for Building Energy Management / V.L. Erickson, M.Á. Carreira-Perpiñán, A.E. Cerpa // ACM Transactions on Sensor Networks (TOSN). – 2014. – № 10. – PP. 1–28.

[2] Zhang, S. Deep learning and multi-objective optimization for real-time occupancy-based energy control in smart buildings / S. Zhang, C. Kong // Scientific Reports. – 2025. – № 15.

[3] Doggrell, S.A. Quantitative study of lecture attendance and the association between this attendance and academic outcomes for nursing and nonnursing students in an introductory pathophysiology course / S.A. Doggrell // Advances in physiology education. – 2021. – № 45(4). – PP. 651–660.

[4] Kelly, G.E. Lecture attendance rates at university and related factors / G.E. Kelly // Journal of Further and Higher Education. – 2012. – № 36. – PP. 17–40.

[5] Arief-Ang, I.B. DA-HOC: semi-supervised domain adaptation for room occupancy prediction using CO2 sensor data / I.B. Arief-Ang, F.D. Salim, M. Hamilton // Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments. – 2017.

[6] Yang, Z. A systematic approach to occupancy modeling in ambient sensor-rich buildings / Z. Yang, N. Li, B. Becerik-Gerber, M.D. Orosz // SIMULATION. – 2014. – № 90. – PP. 960–977.

[7] Wang, W. Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings / W. Wang, J. Chen, T. Hong // Automation in Construction. – 2018.

[8] Yu, Y. Prediction of campus office occupancy and AC behavior based on Transformer and the collective impact on energy consumption / Y. Yu, L. Liu, X. Wang, [et al.] // Building Simulation. – 2025. – № 18. – PP. 3275–3294.

Авторский вклад

Резаи Вахид Сонай Надер кызы – постановка задачи исследования и разработка методологии, сбор и анализ данных о посещаемости и показаниях сенсоров, разработка гибридной модели прогнозирования на основе Big Data, подготовка текста статьи и формулировка выводов.

A HYBRID APPROACH TO OCCUPANCY PREDICTION FOR ENERGY CONSUMPTION OPTIMIZATION IN HIGHER EDUCATIONAL INSTITUTIONS

S.N. Rezai Vahid
Ph.D. Student at the Department of Computer Technologies,
Azerbaijan Technical University

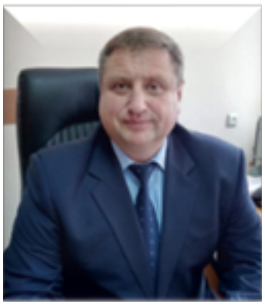
Abstract. This article provides an overview of how big data-driven Predictive Occupancy Analytics (POA) can address the challenges associated with implementing energy-efficient climate control in higher education institutions. Utilizing a hierarchical multi-source data fusion model, the author proposes the integration of deterministic data from official academic schedules, statistical attendance ratios, and real-time operational data from sensor networks to achieve an optimal trade-off between forecasting accuracy and system implementation costs.
Keywords: Big Data, predictive analytics, energy efficiency, HVAC, CO_2 sensors, academic scheduling.

УДК 34; 004.8

О ПЕРСПЕКТИВАХ ПРАВОВОГО РЕГУЛИРОВАНИЯ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В РЕСПУБЛИКЕ БЕЛАРУСЬ



С.Н. Касанин
Заместитель генерального директора по научной работе ОИПИ НАН Беларуси, кандидат технических наук,
s.kasanin@newman.bas-net.by



С.В. Кругликов
Ведущий научный сотрудник ОИПИ НАН Беларуси, доктор военных наук, доцент
kruglikov_s@newman.bas-net.by

С.Н.Касанин
Окончил Кемеровское высшее военное командное училище связи имени маршала войск связи Пересыпкина И.Т., командно-штабной факультет военной академии Республики Беларусь. Область научных интересов включает системы телекоммуникаций, информационная безопасность.

С.В.Кругликов
Окончил Минское высшее инженерное зенитное ракетное училище ПВО по специальности автоматизированные системы управления. Область научных интересов включает разработку концепций, методологических и научно-методических основ проектирования и эксплуатации автоматизированных информационно-управляющих систем различного назначения.

Аннотация. В статье рассматриваются вопросы правового регулирования использования и развития технологий искусственного интеллекта.
Ключевые слова: искусственный интеллект, правовое регулирование ИИ

Введение. Использование систем искусственного интеллекта влияет на инфраструктуру правовых, политических, экономических, социальных и иных отношений, ее наполнение и вектор развития.

По-новому осуществляется коммуникативное взаимодействие, прослеживается динамика политической системы, систем образования и здравоохранения, трудовых и связанных с ними отношений, гражданского оборота.

В области искусственного интеллекта (далее – ИИ) настоятельным требованием времени является разработка и принятие программного документа, задающего основные векторы политики в данной области, а также определяющего концептуальные направления развития правового регулирования, безусловным приоритетом которого должна стать защита прав человека и его основных свобод перед лицом технологического прогресса.

ИИ и механизмы, его содержащие, являются новым явлением в целом, а для правовой системы – в особенности, и затрагивают важнейшие общественные отношения, ряд основных прав и свобод граждан.

Эти вызовы обусловлены определенной степенью автономности действий систем ИИ при решении поставленных задач и их неспособностью непосредственно воспринимать этические и правовые нормы, учитывать их при осуществлении каких-либо действий.

Наука призвана откликаться на данные вызовы, однако в настоящее время в отечественной юридической литературе практически отсутствуют работы, посвященные комплексному изучению правового обеспечения внедрения ИИ, имеются лишь отдельные публикации, посвященные некоторым отраслевым проблемам, связанным с ИИ, при этом не решен круг задач общетеоретического и нормотворческого характера.

Таким образом, отношения, возникающие в связи с созданием, внедрением, использованием технологий с искусственным интеллектом, затрагивая все сферы деятельности, очевидно, требуют правового регулирования.

1. Анализ правового регулирования искусственного интеллекта в мире

Анализ правового регулирования ИИ в мире показывает, что страны, являющиеся технологическими лидерами, ищут баланс между необходимостью формирования правовых норм, препятствующих ущемлению технологией прав граждан, с одной стороны, и сохранением благоприятной среды для разработчиков и инвесторов для упрочения своих позиций, с другой стороны [1].

Каждая юрисдикция выбирает собственный путь, отражающий не только технологические реалии, но и культурные, политические, экономические особенности.

Поскольку недоверие к технологии со стороны общества может негативно сказываться и на технологическом секторе, вплоть до социальных протестов, в сбалансированном регулировании заинтересованы все участники общественных отношений.

Задача найти этот баланс является непростой. Страны мира подошли по-разному к ее решению. Так, закон Европейского союза об ИИ демонстрирует консервативный подход к регулированию технологий ИИ, вводя риск-ориентированный подход и обременительные законодательные требования для разработчиков моделей ИИ, разрабатывающих их на территории ЕС.

Другая категория стран, включающая Китай и Россию, ориентированы на более гибкий подход, в рамках которого регулирование комбинируется из актов стимулирования с точечными нормативными ограничениями и саморегулированием.

Некоторые страны, в числе которых можно отметить Великобританию, Сингапур Южную Корею, Японию, формируют модель регулирования технологий ИИ, которая вводит минимальные законодательные ограничения, делает упор на поддержку научных исследований и инвестиций в ИИ.

В связи с этим анализ разрабатываемых и (или) уже внедренных в других государствах систем регулирования сферы ИИ стратегически будет способствовать принятию более взвешенных решений относительно регулирования искусственного интеллекта в нашей стране.

2. Основные направления деятельности по регулированию искусственного интеллекта в Республике Беларусь

В Республике Беларусь в настоящее время не принят специальный правовой акт, комплексно регламентирующий сферу искусственного интеллекта.

Однако в ряде стратегических и программных актов указано на то, что использование искусственного интеллекта в различных сферах отвечает национальным интересам Республики Беларусь, характеризуется как приоритет инновационного развития.

В отраслевых актах (например, в сфере образования, электротранспорта) все более активно употребляются термины, связанные с технологиями ИИ [2-15].

В ноябре месяце 2024 года состоялось заседание Совета по вопросам правовой и судебной деятельности при Президенте Республики Беларусь рассматривался вопрос «О правовом регулировании использования и развития технологий искусственного интеллекта».

Решение Совета «В целях надлежащего правового регулирования использования и развития технологий искусственного интеллекта признать целесообразным разработку и принятие закона о технологиях искусственного интеллекта».

В декабре месяце 2024 года состоялось заседание Совета по стратегическим проектам при Президенте Республики Беларусь. На заседании рассматривался вопрос «Концептуальные подходы к разработке и реализации технологий и решений в области искусственного интеллекта в интересах динамического развития экономики страны».

Решение Совета «Считать целесообразным осуществить разработку проекта Концепции использования и развития искусственного интеллекта в Республике Беларусь».

6 октября 2025 состоялось Заседание экспертно-консультативного совета (далее –Совет) при Президиуме Совета Республики Национального собрания Республики Беларусь восьмого созыва. Рассматривался вопрос: «Использование искусственного интеллекта: возможности и риски».

По данному вопросу Совет вынес Предложение: «Совету Министров Республики Беларуси рассмотреть возможность поддержки инициативы НАН Беларуси по разработке совместно с заинтересованными проектов Законов Республики Беларусь «О технологиях искусственного интеллекта» и концепции регулирования и развития ИИ в Республике Беларусь».

Во исполнение поручения Совета Министров Республики Беларусь от декабря 2024 года и постановления Совета Министров Республики Беларусь, утвердившего План социально-экономического развития Республики Беларусь на 2025 год в Национальной академии наук Беларуси создана рабочая группа по разработке законопроекта об искусственном интеллекте. В составе рабочей группы 45 человек из 30 организаций.

Рабочей группой заработаны:

– проект глоссария основных терминов и понятий в области искусственного интеллекта;

– проект концепции Закона Республики Беларусь «Об искусственном интеллекте». В настоящее время рабочей группой проводиться обсуждение замечаний и предложений по проекту концепции Закона.

3. Модельный закон «О технологиях искусственного интеллекта»

18 апреля 2025 года Постановлением Межпарламентской Ассамблеи стран-участников Содружества Независимых государств принят модельный закон «О технологиях искусственного интеллекта» (далее – Закон). Разработчиком закона являлось ОИПИ НАН Беларуси.

Целью разработки Закона являлось формирование комплексной системы правового регулирования общественных отношений, возникающих в связи с развитием и использованием технологий ИИ, направленной на улучшение качества жизни населения, безопасности, а также повышение эффективности экономики за счет стимулирования разработки, внедрения и использования ИИ.

Объект исследования – общественные отношения в сфере разработки, внедрения и использования искусственного интеллекта.

Предмет исследования – теоретические и нормативные основы правового обеспечения разработки, внедрения и использования искусственного интеллекта.

Основные задачи модельного Закона:

1) создание благоприятных правовых и организационных условий для развития технологий ИИ, а также содействие инвестициям в исследования и разработки для стимулирования инноваций в области надежного ИИ;

2) формирование и развитие единого рынка законных, безопасных и заслуживающих доверия приложений ИИ;

3) содействие развитию доступных экосистем ИИ с цифровой инфраструктурой, современными технологиями и механизмами для обмена данными и знаниями;

4) осуществление трансграничного и межсекторального сотрудничества в целях достижения прогресса в ответственном управлении доверенным ИИ;

5) формирование позитивной повестки при внедрении технологий ИИ, а также повышение цифровой грамотности населения, в том числе навыков в данной области;

6) обеспечение справедливого и оптимального баланса интересов между всеми участниками общественных отношений в области ИИ, который сочетает механизмы государственного регулирования и саморегулирования и основан на оценке рисков для определения систем ИИ с «высоким риском», которые создают значительную опасность для здоровья и безопасности или основных прав человека;

7) гарантирование безопасности использования и размещения систем ИИ, их соответствие принятым нормам и правилам, соблюдение прав человека;

8) создание гибких механизмов, позволяющих динамически адаптировать правовую основу по мере развития технологий и возникновения новых правоотношений, включая установление экспериментальных и/или специальных правовых режимов, в том числе «регуляторных песочниц».

Заключение. Анализ нормативных актов в области ИИ в Республике Беларусь дает основания для вывода о необходимости целостного законодательного регулирования, учитывающего специфику применения технологий ИИ в различных сферах общественной жизни.

Требует уточнения и установления система единых этических и правовых принципов использования ИИ с разработкой универсального понятийно-категориального аппарата.

Кроме того, правовое регулирование в сфере ИИ в Республике Беларусь должно быть направлено на:

– создание предпосылок для формирования основ правового регулирования новых общественных отношений, складывающихся в связи с разработкой и применением технологий ИИ и систем на их основе, а также определение правовых барьеров, препятствующих разработке и применению указанных систем;

– определение основных подходов к правовому регулированию создания и применения технологий ИИ в различных сферах экономики и социальной жизни при условии соблюдения прав граждан и обеспечения безопасности.

Список литературы

[1] Глобальный атлас регулирования искусственного интеллекта. / А.В.Незнамов [и др.]. – М. : ООО «Профсувенир», 2024 – 377 с.

[2].Протокол заседания Президиума Совета Министров Республики Беларусь от 04.02.2020 г. № 3. Национальная стратегия устойчивого развития Республики Беларусь на период до 2035 года. https://economy.gov.by/uploads/files/Natsionalnaja-strategija-ustojchivogo-razvitija-Respubliki-Belarus-na-period-do-2035-goda.pdf. – Дата доступа: 3.04.2026.

[3] Постановление Президиума Национальной академии наук Беларуси 26.02.2018 № 17 «Стратегия «Наука и технологии: 2018–2040» // URL: https://nasb.gov.by/reference/razvitie/strateg2018_2040.pdf. – Дата доступа: 3.04.2026.

[4] Постановление Совета Министров Республики Беларусь от 24 ноября 2021 г. № 672 «О Стратегии Республики Беларусь в сфере интеллектуальной собственности до 2030 года» // URL: https://pravo.by/document/?guid=12551&p0=C22100672&p1=1. – Дата доступа: 3.04.2026.

[5] Постановление Совета Безопасности Республики Беларусь от 18.03.2019 № 1 «О Концепции информационной безопасности Республики Беларусь». // URL: https://pravo.by/document/?guid=3871&p0=P219s0001. – Дата доступа: 3.04.2026.

[6] Постановление Правления Национального банка Республики Беларусь 27.07.2023 № 267 // URL: https://www.nbrb.by/payment/koncepcija-rasvitija-platioznogo-rinka_2023-2025.pdf. – Дата доступа: 3.04.2026.

[7] Постановление Совета Министров Республики Беларусь от 30.11.2021 № 683 «О Концепции развития системы образования Республики Беларусь до 2030 года» // URL: https://adu.by/images/2021/12/koncep-razv-sist-obrazov.pdf. – Дата доступа: 3.04.2026.

[8] Рекомендации по нормативному регулированию использования искусственного интеллекта, включая этические стандарты для исследований и разработок. // URL: https://iacis.ru/baza_dokumentov/modelnie_zakonodatelnie_akti_i_ rekomendatcii_mpa_sng/rekomendatcii. – Дата доступа: 3.04.2026.

[9] Декрет Президента Республики Беларусь от 22.09.2005 г. № 12 «О Парке высоких технологий». // URL: https://pravo.by/document/?guid=3871&p0=pd0500012. – Дата доступа: 19.08.2024.

[10]. Постановление Совета Министров Республики Беларусь «О мерах по реализации Указа Президента Республики Беларусь от 7 апреля 2022 г. № 136» от 21 апреля 2023 г. № 280 // URL: https://pravo.by/document/?guid=12551&p0=C22300280 – Дата доступа: 19.08.2024.

[25] Закон Республики Беларусь от 07.05.2021 № 99-З «О защите персональных данных». // URL: https://pravo.by/upload/docs/op/H12100099_1620939600.pdf. – Дата доступа: 3.04.2026.

[11] Постатейный комментарий к Закону Республики Беларусь «О защите персональных данных». // URL: https://cpd.by/pravovaya-osnova/metodologicheskiye-dokumenty-rekomendatsii/postatejnyj-kommentarij-k-zakonu-respubliki-belarus-o-zashhite-personalnyh-dannyh/. – Дата доступа: 3.04.2026.

[12] Постановление Совета Министров Республики Беларусь от 25 октября 2022 г. № 724 «О порядке функционирования интеллектуальных транспортных систем» // URL: https://pravo.by/document/?guid=3961&p0=C22200724 – Национальный правовой Интернет-портал Республики Беларусь. – Дата доступа: 3.04.2026.

[13] Альянс в сфере искусственного интеллекта // URL: https://a-ai.ru/?page_id=2128. – Дата доступа: 3.04.2026.

[14] СТБ 2583-2020 «Цифровая трансформация. Термины и определения»: Введ. в действие 01.03.2021 постановлением Государственного комитета по стандартизации Республики Беларусь от 8 декабря 2020 г. № 95 // URL: https://nasb.gov.by/rus/activity/nauchnometodicheskoe-obespechenie-razvitiya-informatizatsii/ – Дата доступа: 19.08.2024.

[15] Постановление Государственного комитета по стандартизации Республики Беларусь от 12.02.2024 №10 «Об утверждении, введении в действие, отмене и изменении технических нормативных правовых актов» // URL: https://etalonline.by/document/?regnum=w224p0003&q_id=. – Дата доступа: 3.04.2026.

PROSPECTS FOR LEGAL REGULATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE REPUBLIC OF BELARUS

S.N. Kasanin
Deputy Director General for Research, OIPI NAS of Belarus, Candidate of Technical Sciences

S.V. Kruglikov
Leading Researcher, OIPI NAS of Belarus, Doctor of Military Sciences, Associate Professor

Abstract. The article discusses the legal regulation of the use and development of artificial intelligence technologies.

Keywords: artificial intelligence, AI legal regulation

УДК 004.932.2:629.3.05

РАЗРАБОТКА АЛГОРИТМА ДЕТЕКЦИИ СВЕТОФОРОВ И РАСПОЗНАВАНИЕ ИХ СИГНАЛОВ ДЛЯ АВТОНОМНОГО ТРАНСПОРТА



К.А. Хаджинова
Студентка 3 курса факультета информационных технологий и управления специальности СУИ (АСОИ) БГУИР
kseniyakhadzhynava@gmail.com



А.Г. Савчиц
Директор научно-технического центра МАЗ, главный конструктор
savchits@maz.by



Р.В. Земляник
Студентка 3 курса факультета информационных технологий и управления специальности СУИ (АСОИ) БГУИР
rzemlanik@gmail.com



Д.А. Хаджинов
Учащийся 10 класса лицея им. Ф.Э. Дзержинского БГУ
deniskhajyn@gmail.com

К.А. Хаджинова
Студентка 3 курса факультета информационных технологий и управления, специальности «СУИ (АСОИ)» Белорусского государственного университета информатики и радиоэлектроники. Является автором статей по обработке данных в информационных системах и оптимизации алгоритмов анализа данных. Область научных интересов связана с обработкой больших данных, оптимизацией процессов их анализа и хранения, детерминированными процессами, а также методами и алгоритмами искусственного интеллекта.

А.Г. Савчиц
Директор научно-технического центра Минского автомобильного завода, главный конструктор. Основная профессиональная деятельность направлена на развитие и модернизацию бортовых систем управления и диагностических комплексов современных транспортных средств, интеграцию современных цифровых технологий и методов анализа изображений в системы автономного и полуавтономного вождения.

Р.В. Земляник
Студентка 3 курса факультета информационных технологий и управления Белорусского государственного университета информатики и радиоэлектроники, автор публикаций по тематике машинного обучения, обработки и анализа больших данных, а также разработки алгоритмов автоматизации процессов в информационных системах.

Д.А. Хаджинов
Учащийся 10 класса лицея им. Ф.Э. Дзержинского Белорусского государственного университета. Область интересов связана с разработкой робототехнических систем, созданием печатных плат, 3D-моделированием, проектированием аппаратных решений и применением компьютерного зрения для анализа и управления техническими объектами.

Аннотация. В статье рассматривается разработка легковесной модели детекции и классификации сигналов светофора для интеграции в систему автономного вождения (ADAS). Представлено полное описание последовательности обработки видеопотока, методов компенсации дисбаланса классов и архитектуры системы на основе YOLOv8. Проведена оценка качества на агрегированном датасете, продемонстрировавшая высокую точность работы алгоритма. Результаты могут быть использованы для оптимизации ИТ-решений в автомобильной промышленности.

Ключевые слова: детекция светофоров, автономный автомобиль, компьютерное зрение, YOLO, дисбаланс классов, машинное обучение

Введение. Определение (детектирование) в видеопотоке светофоров и точное распознавание их сигналов являются критически важными компонентами системы восприятия («мозга») автономного транспортного средства.

В процессе оценки дорожной ситуации светофоры выступают ключевым регулятором проезда перекрестков.

Данная задача отличается высокой сложностью из-за относительно малого размера объектов в кадре, изменчивых погодных условий, наличия перекрытий и сильных визуальных искажений (засветка, блики).

Для обеспечения безопасности стандартов ADAS (Advanced Driver Assistance Systems) автомобилю необходимо не только фиксировать красный сигнал, но и уверенно распознавать разрешающие фазы для своевременного маневрирования.

Целью описанных в работе исследований является разработка и оптимизация нейросетевой модели определения и распознавания светофоров, которая обеспечивает функционирование в реальном времени на вычислительном устройстве транспортного средства.

Предложен подход к построению компактного и производительного алгоритма, ориентированного на работу в реальных дорожных условиях, который позволяет получить удовлетворительную точность локализации, устойчив к внешним факторам и показывает высокую скорость обработки данных.

Постановка задачи и обзор используемых наборов данных (датасетов). На вход системы подается видеопоток с фронтальной камеры автомобиля.

Алгоритм должен локализовать светофор (определить ограничивающую рамку), определить тип сигнала, осуществить межкадровое отслеживание (трекинг) и передать результаты обработки в модуль принятия решений.

Для обучения использовалась агрегированная выборка, включающая открытые датасеты Bosch Small TrafficLights Dataset (BSTLD) и LISA Traffic Light Dataset, а также собственные данные для покрытия сложных условий (ночь, блики).

В таблице 1 представлено сравнение использованных открытых наборов данных.

Таблица 1. Сравнение характеристик наборов данных

Характеристика	**BSTLD**	**LISA**
Разрешение изображений	1280x720	1280x960
Количество классов	4	7
Объем обучающей выборки	5093	10220
Объем тестовой выборки	8334	3348

Данные наборы данных содержат разметку в виде ограничивающих рамок с указанием класса сигнала. Пример исходного кадра из обучающей выборки в сложных условиях освещения приведён на рисунке 1.



Рисунок 1. Исходный кадр из обучающей выборки

В процессе разметки объединённой выборки был выделен следующий набор классов сигналов светофора: red, yellow, green, off. Существенной проблемой при формировании обучающего набора стал естественный дисбаланс: красные и зелёные сигналы присутствуют большую часть времени, тогда как переходные фазы (жёлтый) и состояние off встречаются значительно реже. Для снижения влияния дисбаланса (различного соотношения длительности и частоты появления сигналов светофора) применялась аугментация (увеличение объема) миноритарных классов и взвешивание функции потерь по классам.

Распределение количества примеров (сигналов и состояния светофора) по категориям представлено на рисунке 2.

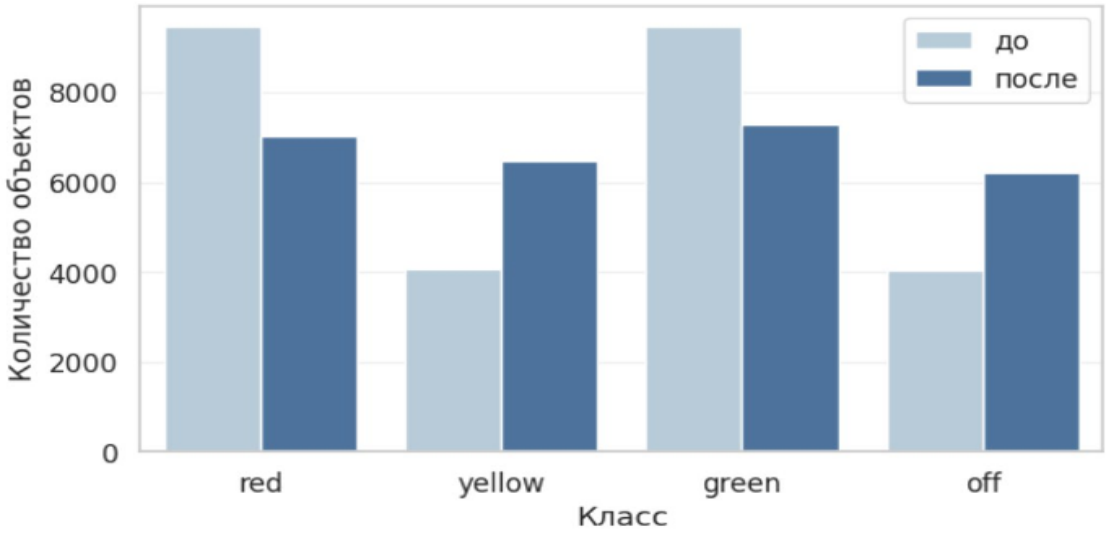


Рисунок 2. Исходный кадр из обучающей выборки

В результате объединения открытых и собственных датасетов удалось сформировать достаточно разнообразный и репрезентативный набор данных, покрывающий как типичные, так и сложные, малоизученные сценарии, что обеспечивает возможность построения и проверки нейросетевой модели, способной обобщать полученные знания на реальные условия эксплуатации.

Архитектура модели и математический аппарат. В качестве базовой архитектуры выбрано семейство моделей YOLOv8 (одноэтапный детектор без использования якорных рамок). Модель состоит из Backbone для извлечения признаков, Neck (FPN/PAN) для их агрегации на разных масштабах и Head для предсказания координат рамки и вероятности класса.

Оптимизация модели в процессе обучения производится путем минимизации комплексной функции потерь, которая состоит из трёх компонентов (1).

$$L_{YOLOv8} = \lambda clsL_{cls} + \lambda boxL_{box} + \lambda dflL_{dfl}, \quad (1)$$

где – общая суммарная функция потерь («общая ошибка» модели за эпоху или батч);

– потери по локализации (насколько предсказанные координаты bounding-box близки к реальной разметке);

– потери по классификации (ошибки в определении типа сигнала: red/yellow/green/off);

– потери по распределению границ рамки (Distribution Focal Loss, помогает точнее регулировать «размытие» координат рамки для маленьких объектов);

, , – коэффициенты-веса (гиперпараметры), которые задают, насколько важна каждая часть: локализация, классификация, распределение.

В общем виде взвешенная кросс-энтропия для истинного класса используется для устранения дисбаланса классов (2).

$$CE(p_t) = -\alpha_t log\ (p_t), \quad (2)$$

где – взвешенная кросс-энтропия для истинного класса на одном примере;

 – вероятность (от 0 до 1), с которой модель предсказала истинный (правильный) класс для данного объекта;

 – вес класса (коэффициент, который можно поднять для миноритарных классов, например yellow и off);

 – натуральный логарифм (или логарифм по основанию 2, в зависимости от реализации, но смысл тот же).

Использование архитектуры YOLOv8 в сочетании с гибкой настройкой функции потерь позволяет совместить высокую точность определения малоразмерных объектов с требованиями к быстродействию, предъявляемым к вычислительным бортовым системам автономного транспортного средства. Выбранные компоненты комплексной функции потерь нацелены на обеспечение баланса между точностью позиционирования и правильной классификацией сигнала светофора.

Обучение модели и оценка качества. Обучение проводилось на протяжении 64 эпох с динамическим снижением шага обучения от 0.00047 до 0.000058.

График изменения функции потерь в процессе обучения приведён на рисунке 3.

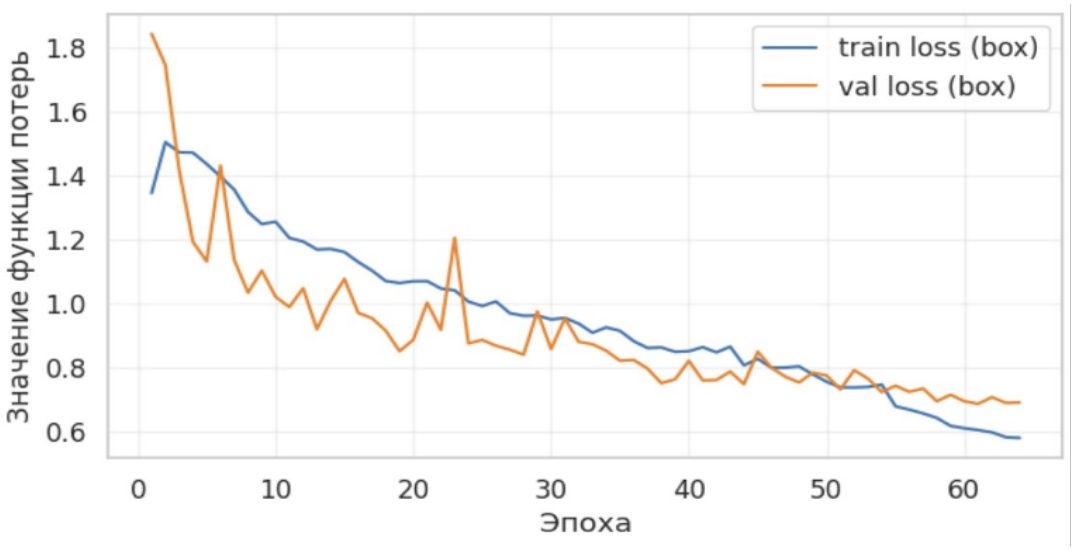


Рисунок 3. Изменение функции потерь в процессе обучения

Для оценки точности локализации и классификации использовались метрики Precision (точность), Recall(полнота) и Mean Average Precision (mAP). Метрика mAP усредняет значения средней точности по всем классам при различных порогах перекрытия IoU от 0.50 до 0.95. Интегральная метрика mAP@50–95 вычисляется по формуле (3). Сумма идёт по 10 порогам: .

$$mAP_{50-95} = \frac{1}{10}\sum_{k=0}^{9} mAP_{0.50+0.05k} \quad (3)$$

Здесь — значение метрики mean Average Precision, усреднённое по десяти порогам перекрытия IoU от 0.50 до 0.95 с шагом 0.05; обозначает значение mAP, рассчитанное по всем классам при фиксированном пороге IoU .

Множитель соответствует простому усреднению по этим десяти значениям, что позволяет получить интегральную характеристику качества детектора при разных уровнях строгости к совпадению предсказанных и истинных ограничивающих рамок, где – средняя точность (mean Average Precision) по диапазону порогов перекрытия IoU от 0.5 до 0.95;

 – средняя точность по всем классам при пороге IoU = T, то есть считается отдельно для каждого класса и затем усредняется.

Коэффициент – это простое усреднение по 10 порогам, то есть формируется интегральная метрика, отражающая производительность модели при разных уровнях строгости (от «слабого» перекрытия до очень точного).

Динамика обучения показывает отсутствие переобучения: ошибки на проверочной (валидационной) выборке стабильно снижались вместе с ошибками на тренировочном наборе.

Динамика изменения метрик в процессе обучения показана на рисунке 4.

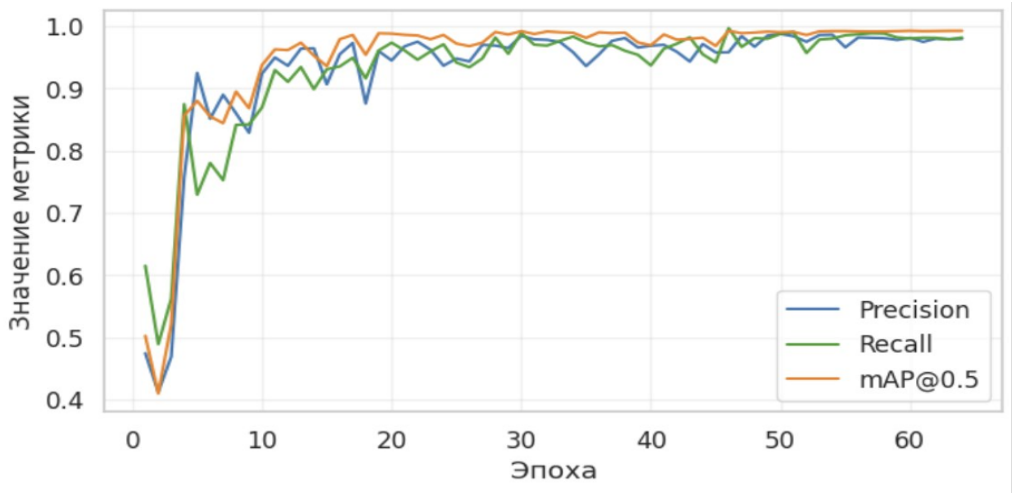


Рисунок 4. Динамика изменения метрик в процессе обучения

Количественные результаты работы модели на финальной 64-й эпохе сведены в таблицу 2.

Таблица 2. Метрики качества модели на 64-й эпохе обучения

Метрика	**Значение на обучающей выборке**	**Значение на валидационной выборке**
Ошибка локализации (box_loss)	0.57792	0.68844
Ошибка классификации (cls_loss)	0.33153	0.34535
Ошибка распределения границ (dfl_loss)	0.89995	0.90337
Точность (Precision)	–	0.97977
Полнота (Recall)	–	0.98175
mAP@50	–	0.99249
mAP@50-95	–	0.79718

Модель достигла показателя mAP@50 на уровне 0.99249, а интегральная метрика mAP@50–95 составила 0.79718. Применение алгоритма Non-Maximum Suppression (NMS) в сочетании с межкадровым трекингом позволило устранить дублирующие рамки и стабилизировать детекцию во времени.



Рисунок 5. Результаты детекции светофоров

Приведённые результаты показывают, что предложенная модель достигает высоких значений метрик как на обучающей, так и на валидационной выборках, при этом отсутствуют признаки переобучения. Таким образом, алгоритм может рассматриваться в качестве перспективной основы для интеграции в систему автономного управления транспортным средством и требует лишь дальнейшей оптимизации и адаптации к используемому вычислительному оборудованию.

Заключение. Использование модуля детекции светофоров в общем процессе организации (пайплайне) системы управления автономным транспортным средством позволяет передавать точные координаты и фазу сигнала в модуль планирования движения (Motion Planning). Разработанная и реализованная модель продемонстрировала высокую точность и стабильность: сбалансированная архитектура YOLOv8 и применение специфических функций потерь (DFL и CIoU) обеспечили высокое качество детекции малоразмерных объектов без потери скорости обработки потока данных для практических задач (инференса нейросети). В дальнейшем планируется расширение обучающей выборки контекстными данными HD-карт и квантизация модели для оптимизации вычислений на edge-устройствах, что в конечном итоге позволит повысить безопасность и комфорт эксплуатации автономного транспорта.

Список литературы

[1] Redmon, J. You only look once: Unified, real-time object detection / J. Redmon, S. Divvala, R. Girshick, A. Farhadi // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. – P. 779–788.

[2] Jocher, G. YOLO by Ultralytics / G. Jocher, A. Chaurasia, J. Qiu. – 2023.

[3] Behrendt, K. A deep learning approach to traffic lights: Detection, tracking, and classification / K. Behrendt, L. Novak, R. Botterweck // 2017 IEEE International Conference on Robotics and Automation (ICRA). – 2017. – P. 1370–1377.

[4] Jensen, M. B. Vision for looking at traffic lights: Issues, survey, and perspectives / M. B. Jensen, M. P. Philipsen, A. Møgelmose, T. B. Moeslund, M. M. Trivedi // IEEE Transactions on Intelligent Transportation Systems. – 2016. – Vol. 17(7). – P. 1800–1815.

Авторский вклад

К.А. Хаджинова – формирование и обработка обучающего датасета, сбалансирование и расширение обучающей выборки, разработка и описание методологии исследования.

А.Г. Савчиц – постановка технического задания и общей тематики исследования, формирование требования к содержанию статьи, а также научное и организационное руководство проектом.

Р.В. Земляник – основной инженер машинного обучения: реализация и настройка YOLOv8-модели, обучение и оценка качества, подготовка экспериментов.

Д.А. Хаджинов – интеграция алгоритма детекции в систему компьютерного зрения робота, разработка и отладка связки аппаратной и программной части, тестирование работы модели в реальных экспериментах, оформление текста статьи.

DEVELOPMENT OF A TRAFFIC LIGHT DETECTION AND SIGNAL CLASSIFICATION ALGORITHM FOR AN AUTONOMOUS VEHICLE PERCEPTION SYSTEM

K.A. Khadzhynava
Third-year student, Faculty of Information Technologies and Management, specialization Information Systems Control (Automated Information Processing Systems), BSUIR

A.G. Savchits
Director of the Scientific and Technical Center, Minsk Automobile Plant (MAZ), Chief Designer

R.V. Ziamlianik
Third-year student, Faculty of Information Technologies and Management, specialization Information Systems Control (Automated Information Processing Systems), BSUIR, Minsk, Belarus

D.A. Khadzhynau
Student of the 10th grade, F.E. Dzerzhinsky Lyceum of the BSU

Abstract. The article discusses the development of a lightweight model for detecting and classifying traffic light signals for integration into an autonomous driving system (ADAS). The full video stream processing pipeline, methods for combating class imbalance, and the YOLOv8-based architecture are described. The quality of the algorithm is evaluated on an aggregated dataset, demonstrating high detection accuracy. The results can be used for optimizing IT-solutions in the automotive industry.

Keywords: traffic light detection, autonomous vehicle, computer vision, YOLO, class imbalance, machine learning

УДК 616.724–07:004.89

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ В ДИАГНОСТИКЕ ДИСФУНКЦИЙ ВНЧС



Ю.А. Засинец
Студент факультета компьютерного проектирования, специальность медицинская электроника, БГУИР, ulia75210@gmail.com



В.А. Хуцишвили
Студент факультета компьютерного проектирования, специальность медицинская электроника, БГУИР, 6067703a1@gmail.com



М.В. Давыдов
Первый проректор, доцент; канд. техн. наук БГУИР davydov-mv@bsuir.by

Ю.А. Засинец
Студент факультета компьютерного проектирования специальности «Медицинская электроника», БГУИР.

В.А. Хуцишвили
Студент факультета компьютерного проектирования специальности «Медицинская электроника», БГУИР.

М.В. Давыдов
Первый проректор, доцент; канд. техн. наук БГУИР.

Аннотация. В работе рассматривается разработка интеллектуальной системы поддержки принятия врачебных решений для диагностики дисфункций височно-нижнечелюстного сустава (ВНЧС). Система базируется на использовании локальной модели (LLM) Llama 3.1, интеграции базы данных медицинских показателей и разработке веб-приложения. Особое внимание уделено остеопатическим аспектам диагностики, которые позволяют выявлять взаимосвязи между состоянием зубочелюстной системы и опорно-двигательного аппарата.

Ключевые слова: дисфункция ВНЧС, Llama 3.1, остеопатия, база данных, веб-приложение, интеллектуальный анализ.

Введение. Дисфункция височно-нижнечелюстного сустава (ВНЧС) может возникать по самым разным причинам, поэтому их так сложно диагностировать. Согласно принципам остеопатии, организм представляет собой единую биомеханическую цепь.

Нарушения в одном звене (например, деформация стопы или перекос таза) неизбежно вызывают реакции в других, включая шейный отдел и ВНЧС.

Врач-диагност часто сталкивается с проблемой определения первопричины патологии: является ли она восходящей (от стоп к челюсти) или нисходящей.

Интеллектуальный анализ больших массивов данных позволяет автоматизировать поиск этих скрытых связей.

Основная часть. Для реализации системы была спроектирована специализированная база данных (БД), аккумулирующая результаты комплексного обследования. В БД хранятся данные о пациентах(ФИО, возраст и регион), также хранятся данные обследований.

К ним относятся: осмотр стоматологический, результаты ультразвуковых исследований (УЗИ), электромиографии (ЭМГ) , электроэнцефалографии (ЭЭГ), показатели подометрии (анализ стоп) и остеопатическое тестирование.

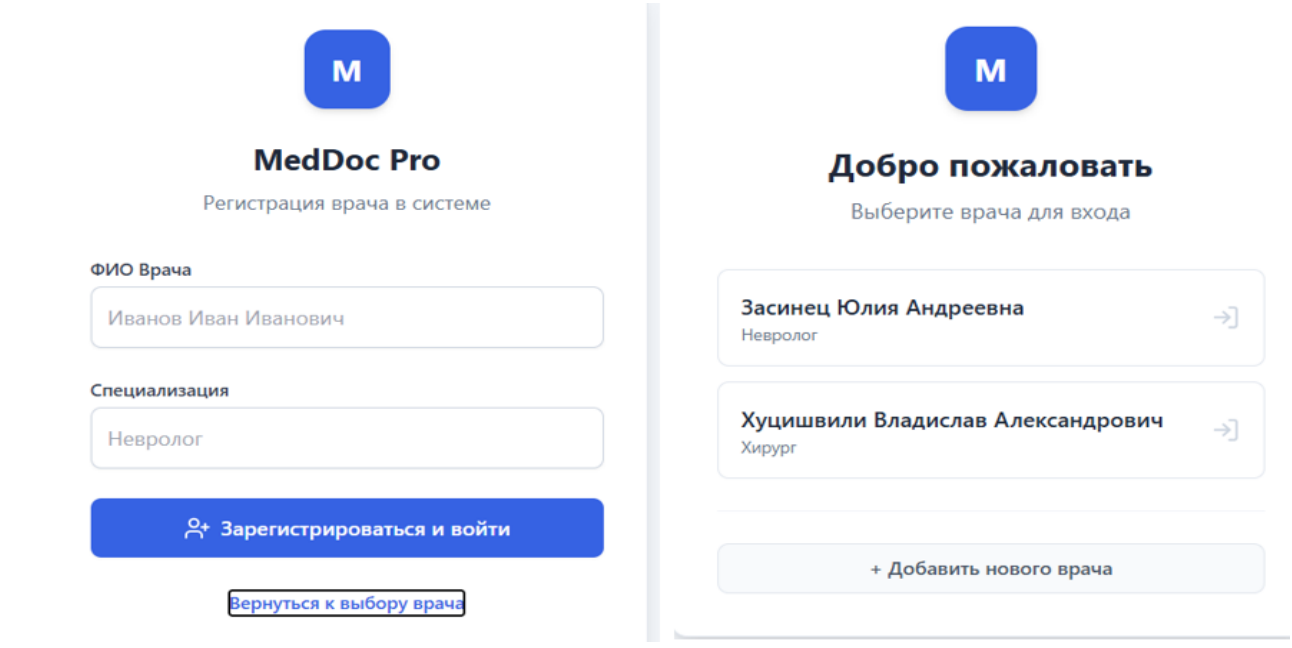


Рисунок 1. Интерфейс врача при входе в веб-приложение

Локальное хранение данных на собственном сервере обеспечивает полную безопасность персональной информации пациентов и высокую скорость обработки запросов нейросетью без обращения к облачным сервисам.

Ядром системы выступает локальная нейросеть Llama 3.1. С помощью специальной инструкции (System Prompt) мы настроили модель как эксперта, обладающего знаниями в области биомеханики и остеопатии.

В процессе работы Llama 3.1 выполняет «маппинг» – сопоставление векторов различных патологий.

Например, обнаружив в базе данных «выраженную пронацию стопы», система автоматически анализирует данные ЭМГ жевательных мышц на предмет асимметрии. Это позволяет реализовать целостный взгляд на здоровье, выявляя, является ли проблема восходящей или нисходящей.

Для удобного взаимодействия врача и пациента было разработано кроссплатформенное веб-приложение. Оно включает в себя: Интерфейс врача: удобный дашборд с визуализацией графиков активности мышц и результатами тестов. Здесь же выводятся подсказки от нейросети в формате JSON, указывающие на наиболее вероятную причину дисфункции.

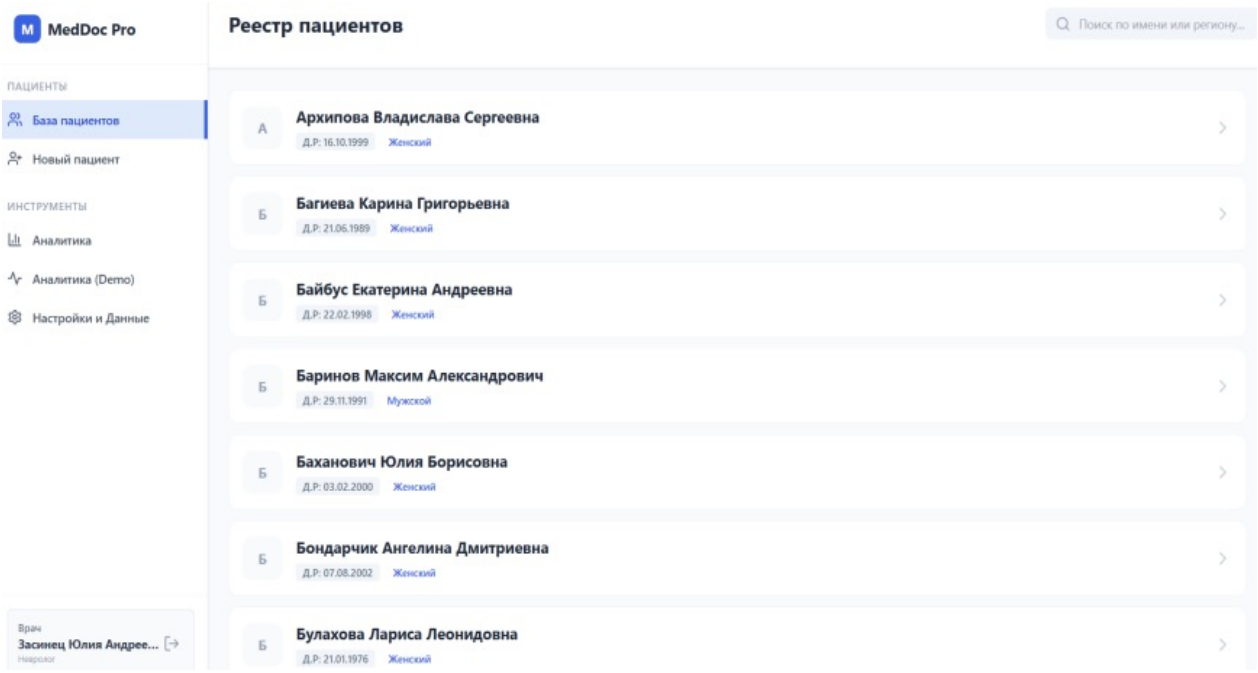


Рисунок 2. Интерфейс врача при входе в приложение

Дашборд пациента: В личном кабинете реализовано окно интеллектуального анализа. В нем отображаются результаты сравнения паттернов из разных источников (ЭМГ, УЗИ, Постурография).

Система визуально разделяет выявленные патологии на восходящие (идущие от стоп) и нисходящие (идущие от прикуса).

Врач получает готовый список потенциальных причинно-следственных связей, которые могли быть упущены при ручном анализе.

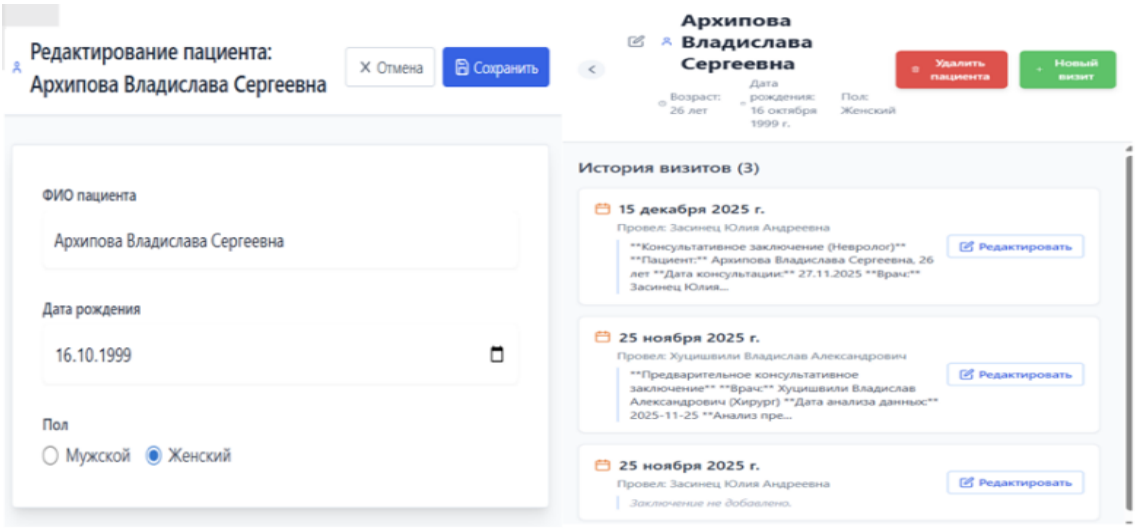


Рисунок 3. Интерфейс пациента

Разработанная система на базе Llama 3.1 решает главную проблему диагностики ВНЧС – информационную перегрузку врача. Объединение принципов остеопатии с возможностями искусственного интеллекта позволяет не просто фиксировать симптомы, а находить истинную причину заболевания в сложной цепи «стопы – позвоночник – челюсть». Это ускоряет анализ карты пациента и повышает точность первичной диагностики.

Список литературы

[1] Давыдов, М. В. Интеллектуальные информационные технологии в медицинских диагностических системах: учеб. пособие / М. В. Давыдов – Минск: БГУИР, 2018. – 144 с.

[2] Потапов, И.В. Искусственный интеллект в стоматологии: состояние и перспективы: научн. статья / И.В. Потапов, С.А. Гаврилова – Изд-во «Стоматология для всех», 2022 – 8 с.

[3] Machado, M., Ribeiro, A. (2024). Decision Support Systems in Temporomandibular Joint Osteoarthritis: A review of Data Science and Artificial Intelligence Applications. Journal of Personalized Medicine, 14(2), 154.

Авторский вклад

Засинец Юлия Андреевна – реализация алгоритмов маппинга медицинских показателей (ЭМГ, УЗИ, подометрия) и разработка функционала кроссплатформенного веб-приложения для визуализации диагностики.

Хуцишвили Владислав Александрович – разработка концепции интеллектуальной системы, проектирование базы данных и настройка локальной модели Llama 3.1 для экспертного анализа биомеханических связей.

Давыдов Максим Викторович – постановка цели и задач.

INTELLECTUAL DECISION SUPPORT SYSTEM FOR DIAGNOSING TMJ DYSFUNCTIONS

J.A. Zasinets
Student of the Faculty of Computer-Aided Design, majoring in Medical Electronics, BSUIR,
ulia75210@gmail.com

V.A. Khutsyshvil
Student of the Faculty of Computer-Aided Design, majoring in Medical Electronics, BSUIR,,
6067703a1@gmail.com

M.V. Davydov
First Vice-Rector, Associate Professor; PhD in Technical Sciences, BSUIR
davydov-mv@bsuir.by

Annotation. The study discusses the development of an intellectual clinical decision support system designed for the diagnosis of temporomandibular joint (TMJ) dysfunctions. The system is based on the integration of a local large language model (LLM), specifically Llama 3.1, with a comprehensive database of medical parameters and a dedicated web application. Particular emphasis is placed on osteopathic diagnostic aspects, which enable the identification of functional correlations between the stomatognathic system and the musculoskeletal system. The system facilitates the automated analysis of complex data (EMG, podometry, etc.) to determine whether the pathology is "ascending" or "descending" in nature.

Keywords: TMJ dysfunction, Llama 3.1, osteopathy, database, web application, intelligent analysis, decision support system.

УДК 004.8; 004.8.032.26

EVALUATING VULNERABILITIES AND PROTECTIONS OF AI METHODS IN COMPUTER-ASSISTED IMAGE-BASED DIAGNOSIS



V.A. Kovalev
Leading researcher, Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus, PhD
vassili.kovalev@gmail.com



E.V. Snezhko
Head of Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus, PhD
eduard.snezhko@gmail.com



D.S. Karpenko
Junior researcher, trainee, Dept. Biomedical Image Analysis, United Institute of Informatics Problems, National Academy of Sciences of Belarus
karpenko.dima.s11@gmail.com

A.G. Varvashevich
Junior researcher
varvashevichangelina@gmail.com

V.A. Kovalev
Graduated from the Tomsk Polytechnic University. Research interests: image analysis, image generation, AI security

E.V. Snezhko
Graduated from the Belarusian State University. Research interests: image analysis, AI security.

D.S. Karpenko
Graduated from the Belarusian State University. Research interests: generative neural networks, federated learning.

A.G. Varvashevich
Graduated from the Belarusian State University. Research interests:image analysis, Adversarial Attacks.

Abstract. This paper examines three safety aspects of AI in medical image analysis: vulnerability to adversarial attacks, patient data protection via generative models, and federated learning without data sharing. Experiments on over 260,000 chest X-rays, CT slices, and histological slides show that white-box adversarial attacks (FGSM, AutoAttack, Carlini-Wagner) can reduce classification accuracy to 0% when no defense is applied. Among three defense strategies, a high-level task-driven denoiser proved most effective, restoring accuracy up to 100%. Generative models (DC-GAN, Pro-GAN, diffusion) produced realistic synthetic images; using them for data augmentation led to only a modest accuracy drop (2.2-3.5% for deep learning). Federated learning (FedAVG) succeeded only for homogeneous datasets (chest X-

rays) but failed for highly variable histological images. The paper concludes that adversarial attacks pose critical threats, generative models enable privacy-preserving augmentation, and federated learning requires careful modality-specific adaptation.

Keywords: AI safety, medical imaging, adversarial attacks, deep learning, generative models, federated learning, data privacy

Introduction. Recently, artificial intelligence (AI) methods have seen significant development and proliferation in various application areas, from medicine and finance to solving technical problems in industry [1].

In healthcare, AI-based methods significantly facilitate the analysis of medical images used in computerized disease diagnosis [2, 3, 4].

The use of AI methods and software tools allows for more effective detection of various diseases, increases the standardization of diagnostic processes, and helps physicians make more informed decisions. However, the intensive processes of developing and implementing AI solutions into practice are not yet supported by appropriate technical solutions for the protection of patient personal data and AI software, which differ significantly from existing software [5, 6].

The aim of this paper is to examine the results of experiments assessing vulnerabilities associated with adversarial attacks on deep neural networks, the possibilities of patient identification from medical images, as well as the effectiveness of federated learning of neural networks, which does not require the transfer of patient personal data between its participants.

Attacks on Deep Neural Networks. During research and development of various deep neural network architectures, it has been discovered that under certain, deliberately created conditions, neural networks can become extremely unstable. For example, a well-trained and seemingly very reliable neural network recognizing cancerous tumors and normal tissue in histological biopsy images (the gold standard in oncology) may classify tumor tissue as normal and areas of normal tissue as malignant neoplasms. The reason is that special algorithms have been discovered and implemented that minimally alter the input image, after which, for some reason, it is misclassified by the neural network. These image modifications are so minor that they are often indistinguishable to the human eye.

The process by which such an image is generated and fed to the network is called an adversarial attack [7].

Obviously, this problem creates a serious security breach in neural networks when solving image recognition and classification tasks in medicine and other areas of AI application.

Computational experiments were conducted on a large sample of more than 260,000 medical images of the following three types:

- Digital chest X-rays of people who underwent regular examinations within a computerized telemedicine system operating on the basis of medical institutions in Minsk (15,600 people aged 18 and older);
- Computed tomography (CT) images of patients with pulmonary tuberculosis undergoing treatment at one of the specialized medical centers in Belarus (3D tomograms of 414 patients divided into 53,677 2D images of axial slices of size 512×512 pixels each);
- Histological images of patients from one of the oncology centers in Belarus undergoing treatment for thyroid cancer and ovarian cancer (tissue samples routinely stained with hematoxylin and eosin, as well as special immunohistochemical preparations such as D2-40, Ki6, CD-105, FRES, and CD-31; total 192,000 images of size 256×256 pixels).

Examples of original images are shown in Fig. 1.



a) b) c)

Figure 1. Examples of original images: a) X-ray; b) CT; c) histological.

In the computational experiments, the EfficientNet-B3 network [8] was chosen as the target neural network. Only "white-box" attacks were performed, i.e., all elements of the target network architecture were assumed known to the attacker [9]. For each of the three image types, a corresponding version of the EfficientNet-B3 neural network was trained to classify input images into one of two conditionally designated classes: Normal and Pathology.

The classification accuracy under normal conditions, i.e., before the attack, was recorded for subsequent statistical analysis.

Attack effectiveness was evaluated by the difference in image classification accuracy before and after the attack. Three independent series of experiments were conducted for three different types of attacks, including gradient attacks of the FGSM type [9], AutoAttack proposed by Croce and Hein [10], and Carlini-Wagner attacks [11].

In all three attacks, the attack strength (magnitude) parameters were chosen such that the presence of pixel modifications to the original images could not be reliably detected visually.

Examples of original CT (top row) and histological (bottom row) images and their modified adversarial versions are shown in Fig. 2.

After the attacks, the adversarial images underwent a defense procedure. Currently, researchers propose various methods of defense against adversarial attacks on neural networks. In this work, an experimental evaluation of the effectiveness of three types of defense, presented below, was carried out.

Adversarial training of neural networks. This type of defense involves training networks using mixed training datasets containing both normal images and images that have already been maliciously modified [9].

Denoiser driven by high-level task data. This defense is based on the idea that malicious image modifications can be considered as a type of "noise" that should be removed [12].

To perform this function, an auxiliary neural network is trained that tracks the relationships between local image elements and high-level data, such as Normal/Pathology image class labels, etc. *Specialized MagNet neural network.*

Similar to the previous defense method, the MagNet network treats adversarial image modifications as special types of noise and is trained to localize and remove them using the principle of a filtering auto-encoder [13].

The experimental results are presented in the Table 1.

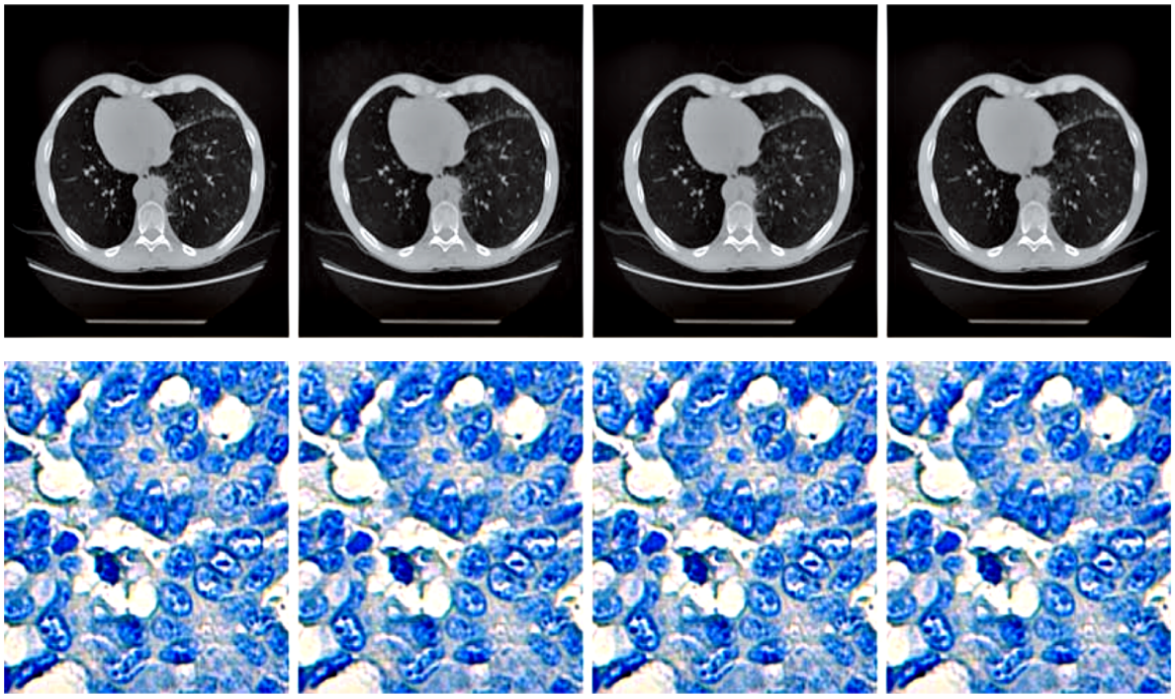


Figure 2. Examples of CT and histological images including original images (column 1) and their adversarial versions (columns 2–4)

As can be seen from the table, the strongest attacks are AutoAttack and Carlini-Wagner attacks, while the Denoiser provides the best defense.

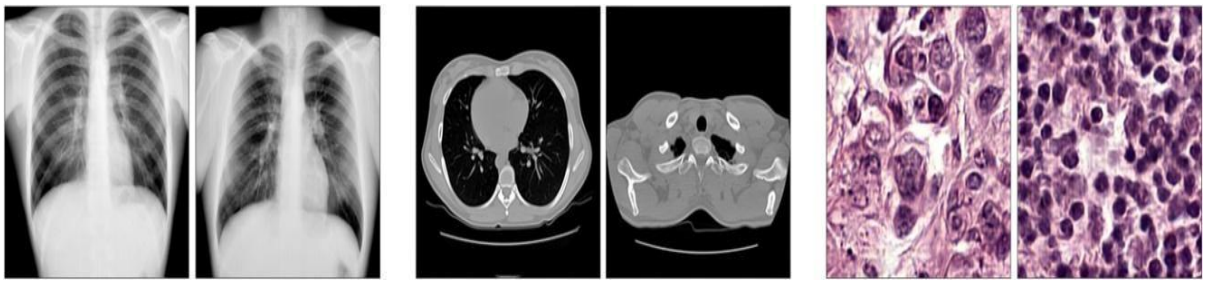
As a general conclusion, it can be noted that adversarial attacks can significantly reduce the accuracy of correct medical image classification, up to completely substituting Normal for Pathology and, what is even more dangerous, suggesting a decision of complete absence of disease when it is actually present. Therefore, the development of methods and software tools for protecting AI systems against adversarial attacks in medical and other critical tasks is, in our view, absolutely necessary.

Security of Patient Personal Data. One way to ensure the security of patient personal data contained in their medical images is to replace real patient images with images generated using modern generative neural network models. Such replacement is possible in various scenarios of development, implementation, and use of decision support information systems in medical diagnostics and allows resolving a number of legal, ethical, and purely technical issues. The most obvious example is the use of realistic artificial images visually indistinguishable from real ones in the preparation of information and advertising materials, development of thematic websites, production of printed and video products, as well as other cases of promoting and implementing relevant subsystems, blocks, and computing services of computer-aided disease diagnostic systems. Research and development in the field of generative models has been conducted by the authors of this work since 2017. As generative models, DC-GAN type models [14] consisting of a generator and a discriminator competing with each other according to the scheme of mathematical models of antagonistic game theory, as well as generators with progressively increasing image resolution (Pro-GAN) [15] and diffusion models of the UNET + Gaussian denoiser type [16] (see Fig. 3), were used.

To assess the quality of the generated images, a professional test was conducted involving experienced radiologists (5 groups of 1–3 physicians each). The task was to separate a sample of 200 randomly mixed X-ray images, 100 of which were images of real people and the other 100 were generated. The results of the blind test showed that correctly separating these image groups is a fairly difficult task even for professionals.

Table 1. Classification accuracy in % for different attack and defense types

Attack Type	**Defense Type**			
	No Defense	**Adversarial Training**	**Denoiser**	**MagNet**
X-ray images				
No attack	92.8	–	–	–
FGSM gradient attack	60.2	82.8	93.4	73.8
AutoAttack	00.0	89.6	93.6	73.6
Carlini-Wagner attacks	00.0	84.2	92.9	73.9
CT images				
No attack	100.0	–	–	–
FGSM gradient attack	57.9	98.4	100.0	98.8
AutoAttack	00.0	97.5	100.0	99.0
Carlini-Wagner attacks	00.0	98.3	100.0	98.9
Histological images				
No attack	95.8	–	–	–
FGSM gradient attack	17.8	89.4	91.4	71.3
AutoAttack	00.0	93.1	94.2	72.7
Carlini-Wagner attacks	10.2	82.7	92.9	73.0



a) b) c)

Figure 3. Examples of generated images: a) X-ray; b) CT; c) histological.

Another important direction for the use of generative models is their use as an "intelligent" *augmentation* service, i.e., expanding the *quantity* of training images by including not only real but also generated images when there is a shortage. It is known that a lack of training images is one of the central practical problems in developing AI systems.

Conducted research and experiments have shown that even a complete replacement of real images with generated ones leads to a drop in classification accuracy; the decrease caused by such a replacement of training data ranges from 2.2% to 3.5% for deep learning models and from 5.5% to 13.25% for conventional methods such as LBP + Random Forests.

Federated Learning of Neural Networks. Another way to solve the problem of a limited set of training images is to combine image databases from multiple centers. However, combining and sharing medical images belonging to different institutions can be problematic for legal, social, and ethical reasons.

These issues can be overcome by using the so-called Federated Learning technology [17]. In this work, we present the results of an experimental study of the effectiveness of the FedAVG variant of federated learning for DC-GAN type generative neural networks.

The computational experiments assessed the joint federated training of generative models designed to synthesize images at two spatial resolutions: 64×64 and 128×128 pixels. The resulting patterns were identical for both resolutions. The analysis covered three types of medical images: chest X-rays, computed tomography slices, and histological specimens.

The convergence dynamics of the federated learning process were examined under three distinct aggregation regimes performed by the server. Weight averaging of the neural network was carried out after clients had completed 1, 3, or 5 training epochs on their private data. The datasets employed

differed in their degree of heterogeneity. The most homogeneous was the set of chest X-ray images (10,000 healthy individuals).

To a non-specialist, the majority of these images appear nearly identical, differing only in subtle visual features related to age, sex, and sometimes overall body constitution. When training on small images (64×64 pixels), the federated approach demonstrated stable and consistent convergence across all three aggregation regimes. However, with larger images (128×128 pixels), the curves reflecting the consistency of the results (as measured by the FID score [18]) became noisier.

When aggregation was performed only after every 5 epochs, the training process failed to converge altogether. The learning trajectory for computed tomography images broadly resembled that of X-ray images, but the training was considerably less stable in all cases. This was evident from the behavior of the curves both for the individual clients' training processes and for the server's aggregated results. Histological images are characterized by high morphological variability. In a typical whole-slide biopsy specimen, the same cellular patterns rarely repeat.

For this image type, the FedAVG approach behaved identically at both resolutions: throughout most training cycles, the model that aggregated the results from two clients consistently performed worse than the models trained locally by each client. In other words, the Fréchet distance between real and generated images remained large, and the use of federated learning technology led to a high degree of dissimilarity between synthetic and real images. Consequently, under these specific conditions, federated training is not advisable.

A key finding is that the potential utility of the horizontal FedAVG approach strongly depends on the natural homogeneity of the image datasets involved. Among the three examined modalities, chest X-rays showed the greatest promise for federated training, whereas standard hematoxylin-eosin-stained histological images proved unsuitable.

Two-dimensional computed tomography slices occupied an intermediate position, exhibiting unstable behavior during training. Furthermore, the aggregation period on the federated server should be kept reasonably short; it is recommended to perform aggregation after every 1 to 3 local training epochs completed by the clients on their respective image datasets.

Conclusion. This paper has systematically examined three critical safety aspects of artificial intelligence methods in medical diagnostics: vulnerability to adversarial attacks, protection of patient personal data via generative models, and the feasibility of federated learning for collaborative model training without data sharing.

Experimental results on three large-scale medical image datasets (chest X-rays, CT scans, and histology slides) demonstrate that deep neural networks (EfficientNet-B3) are highly susceptible to adversarial attacks. Under white-box conditions, gradient-based FGSM attacks, AutoAttack, and Carlini-Wagner attacks caused dramatic drops in classification accuracy, in many cases reducing it to 0% for AutoAttack and Carlini-Wagner attacks on X-ray and CT images without defense. Even more alarmingly, such attacks can induce misclassifications that swap "Normal" and "Pathology" labels, potentially leading physicians to conclude that a disease is absent when it is actually present. Among the three defense strategies evaluated – adversarial training, a high-level task-driven denoiser, and the MagNet auto-encoder – the denoiser consistently provided the best protection, restoring accuracy to 92.9–100% across all attack types and imaging modalities. These findings underscore the absolute necessity of integrating robust defense mechanisms into AI systems intended for critical medical applications. The study on generative models for patient data security explored replacing real patient images with synthetically generated ones (DC-GAN, Pro-GAN, and diffusion models) to circumvent legal, ethical, and technical issues associated with personal data handling. A blind test involving experienced radiologists revealed that distinguishing real from generated chest X-rays is surprisingly difficult, confirming the high realism of the synthetic images. Furthermore, the authors investigated using generative models for data augmentation. Replacing real training images entirely with generated ones led to only a modest decline in classification accuracy (2.2–3.5% for deep learning models and 5.5–13.25% for conventional methods like LBP + Random Forests), suggesting that generative augmentation is a viable solution when real data are scarce. As an alternative

approach to address limited data availability across multiple institutions without sharing sensitive patient information, the FedAVG variant of federated learning was tested on DC-GANs. The experiments involved two partners with equally split datasets, with model aggregation performed after 1, 2, or 5 local training epochs.

The results showed that the success of federated learning heavily depends on the homogeneity of the image sets. Chest X-ray images, being relatively uniform, demonstrated strong potential for federated learning. In contrast, histological images proved unsuitable for the FedAVG approach due to their high heterogeneity. Axial 2D CT slices exhibited intermediate behavior.

Thus, federated learning is not a universal solution; its applicability must be assessed on a per-modality and per-dataset basis.

The study confirms that while AI methods offer immense benefits for medical diagnostics, they introduce novel vulnerabilities and privacy risks that are not adequately addressed by traditional security measures.

Adversarial attacks can completely subvert diagnostic decisions, and effective defenses (such as the high-level denoiser) are essential.

Generative models provide a promising pathway for anonymizing patient data and augmenting limited datasets. Federated learning offers a collaborative framework but requires careful handling of data heterogeneity.

Future work should focus on developing hybrid protection schemes that combine adversarial robustness, generative privacy preservation, and adaptive federated strategies tailored to the specific characteristics of medical imaging data.

Acknowledgments

This work was partly supported by State Program for Scientific Research, project И104.

References

[1] Azad A., Banu A. Publication trends in artificial intelligence conferences: The rise of super prolific authors //arXiv preprint arXiv:2412.07793. – 2024.

[2] Hou Q. et al. A clinical-oriented multi-level contrastive learning method for disease diagnosis in low-quality medical images //International Conference on Medical Image Computing and Computer-Assisted Intervention. – Cham : Springer Nature Switzerland, 2024. – С. 13-23.

[3] Wu J., Xu M. One-prompt to segment all medical images //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2024. – С. 11302-11312.

[4] Queiroz D. et al. Fair foundation models for medical image analysis: Challenges and perspectives //ACM Transactions on Computing for Healthcare. – 2025.

[5] Chang H., Shokri R. On the privacy risks of algorithmic fairness //2021 IEEE European Symposium on Security and Privacy (EuroS&P). – IEEE, 2021. – С. 292-303.

[6] Zhu Y. et al. Privacy-preserving in medical image analysis: A review of methods and applications //International Conference on Parallel and Distributed Computing: Applications and Technologies. – Singapore : Springer Nature Singapore, 2024. – С. 166-178.

[7] Finlayson S. G. et al. Adversarial attacks on medical machine learning //Science. – 2019. – Т. 363. – №. 6433. – С. 1287-1289.

[8] Tan M., Le Q. EfficientNet: Rethinking model scaling for convolutional neural networks //International conference on machine learning. – PMLR, 2019. – С. 6105-6114.

[9] Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.

[10] Croce F., Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks //International conference on machine learning. – PMLR, 2020. – С. 2206-2216.

[11] Carlini N., Wagner D. Towards evaluating the robustness of neural networks //2017 ieee symposium on security and privacy (sp). – Ieee, 2017. – С. 39-57.

[12] Liao F. et al. Defense against adversarial attacks using high-level representation guided denoiser //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 1778-1787.

[13] Meng D., Chen H. Magnet: a two-pronged defense against adversarial examples //Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. – 2017. – С. 135-147.

[14] Radford A., Metz L., Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks //arXiv preprint arXiv:1511.06434. – 2015.

[15] Karras T. et al. Progressive growing of gans for improved quality, stability, and variation //arXiv preprint arXiv:1710.10196. – 2017.

[16] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851.

[17] McMahan B. et al. Communication-efficient learning of deep networks from decentralized data //Artificial intelligence and statistics. – PMLR, 2017. – С. 1273-1282.

[18] Heusel M. et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium //Advances in neural information processing systems. – 2017. – Т. 30.

Author's contribution

Kovalev Vassili – problem statement, development of key approaches and algorithms.

Snezhko Eduard – problem statement, selection and testing of neural network architectures.

Karpenka Dmitry – development of software tools and conducting computational experiments on federated learning.

Varvashevich Angelina – development of software tools and conducting computational experiments on Adversarial Attacks.

ОЦЕНКА УЯЗВИМОСТЕЙ И МЕТОДОВ ЗАЩИТЫ ИИ В КОМПЬЮТЕРИЗИРОВАННОЙ ДИАГНОСТИКЕ ПО ИЗОБРАЖЕНИЯМ

В.А. Ковалев
Ведущий научный сотрудник, лаборатория анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси, кандидат технических наук
vassili.kovalev@gmail.com

Э.В. Снежко
Заведующий лаборатории анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси, кандидат технических наук
eduard.snezhko@gmail.com

Д.С. Карпенко
Младший научный сотрудник, стажер, лаборатория анализа биомедицинских изображений, Объединенный Институт Проблем Информатики Национальной Академии Наук Беларуси
karpenko.dima.s11@gmail.com

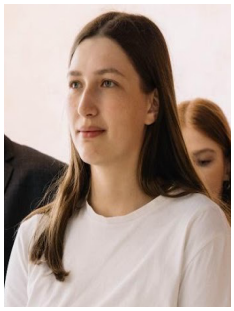
А.Г. Варвашевич
Младший научный сотрудник
varvashevichangelina@gmail.com

Аннотация. В данной статье рассматриваются три аспекта безопасности ИИ при анализе медицинских изображений: уязвимость к состязательным атакам, защита данных пациентов с помощью генеративных моделей и федеративное обучение без обмена данными. Эксперименты на более чем 260 000 рентгеновских снимков грудной клетки, КТ-срезов и гистологических препаратов показывают, что атаки «белого ящика» (FGSM, AutoAttack, Карлини – Вагнера) могут снизить точность классификации до 0% при отсутствии защиты. Среди трёх стратегий защиты Denoiser, управляемый высокоуровневыми данными задачи, оказался наиболее эффективным, восстанавливая точность до 100%. Генеративные модели (DC-GAN, Pro-GAN, диффузионные модели) создавали реалистичные синтетические изображения; их использование для аугментации данных приводило лишь к незначительному снижению точности (2,2–3,5% для глубокого обучения). Федеративное обучение (FedAVG) оказалось успешным только для однородных наборов данных (рентгеновские снимки грудной клетки), но неэффективным для гистологических изображений. В статье делается вывод, что состязательные атаки представляют собой критическую угрозу, генеративные модели обеспечивают аугментацию с сохранением конфиденциальности, а федеративное обучение требует тщательной адаптации к конкретному типу изображений.

Ключевые слова: безопасность ИИ, медицинские изображения, атаки на нейронные сети, глубокое обучение, генеративные модели, федеративное обучение, конфиденциальность данных.

УДК 004:37.018.43

ВЛИЯНИЕ ИНТЕГРАЦИИ ГЕЙМИФИКАЦИИ И ПЕРСОНАЛИЗАЦИИ НА УДЕРЖАНИЕ ПОЛЬЗОВАТЕЛЕЙ В ЦИФРОВЫХ ОБРАЗОВАТЕЛЬНЫХ ПРИЛОЖЕНИЯХ



А.В. Мосендз
Магистрант Белорусского государственного университета информатики и радиоэлектроники
MosenndzAV@yandex.by



Г.А. Хацкевич
Профессор кафедры экономической информатики БГУИР, доктор экономических наук, профессор
g.a.khatskevich@gmail.com

А.В. Мосендз

Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с разработкой программных продуктов для автоматизации образовательных процессов, применением методов дистанционного обучения, а также интеграцией геймификации и персонализации в цифровых образовательных системах.

Г.А. Хацкевич

Окончил математический факультет Белорусского государственного университета. Область научных интересов связана с эконометрическим моделированием и прогнозированием социально-экономических процессов, а также развитием инновационного высшего образования.

Аннотация. В условиях растущей потребности повышать мотивацию и удержание пользователей в цифровом обучении рассматриваются комбинированные подходы. В статье анализируются теоретические основы геймификации и персонализации обучения, выявляются их сильные стороны и ограничения при отдельном применении. Предлагается концептуальная модель интеграции игровых механик с адаптивным обучением, в которой динамические игровые стимулы соотносятся с индивидуальным прогрессом пользователя. Описаны архитектура и ключевые компоненты модели, включая механизмы персонализации, контекстного дизайна игровых стимулов и интеграционные механизмы. Отмечается, что синергия геймификации и персонализации должна усиливать вовлечённость и удержание, компенсируя недостатки каждого подхода. Рассмотрены ожидаемые эффекты, практические рекомендации и ограничения предложенной модели.

Ключевые слова: геймификация, персонализация обучения, адаптивное обучение, удержание пользователей, вовлечённость, цифровые образовательные приложения, адаптивная геймификация, игровые механики, мотивация, интеграционная модель

Введение. Низкая вовлечённость пользователей в цифровые образовательные системы является серьёзной проблемой: по данным Gallup за 2025 год, глобальный уровень вовлечённости пользователей в образовательных системах составляет лишь 21 % [1], что эквивалентно потерям в $438 млрд мировой экономики. растёт рынок цифровых систем дистанционного обучения (LMS): по оценке агентства Smart Ranking, общий объём российского рынка таких платформ в 2024 году превысил 7,2 млрд рублей [2].

В этих условиях качество проектирования программ обучения становится ключевым фактором повышения вовлечённости и удержания пользователей.

В литературе геймификация (внедрение игровых элементов в обучение) и персонализация (адаптация контента под индивидуальные потребности) считаются

перспективными подходами для мотивации обучающихся. Однако каждый из этих инструментов по отдельности имеет ограничения: универсальные игровые решения часто не учитывают разные типы пользователей, а стандартные программы обучения могут не отражать уникальные потребности каждого пользователя [5].

В связи с этим в статье предлагается интегративная концепция: адаптивная геймификация дистанционного обучения– модель, объединяющая динамическую геймификацию с адаптивным обучением.

В рамках предлагаемой модели игровые стимулы (очки, значки, квесты и т.д.) генерируются и подстраиваются на основе поведения и прогресса конкретного пользователя. Такая интеграция должна повысить внутреннюю мотивацию и удержание пользователей, компенсируя недостатки каждого подхода [11].

В статье представлены обзор литературы, обсуждение теоретических основ, описание архитектуры предложенной модели и её ключевых компонентов, а также практические рекомендации и прогнозы эффективности.

Теоретические основы геймификации

Геймификация – применение игровых приёмов и элементов (баллы, значки, лидерборды, сюжеты) в нерыночных контекстах – направлена на повышение вовлечённости пользователей [4]. Игры создают необходимые избыточные препятствия, которые люди охотно преодолевают ради удовольствия и чувства достижений [5].

В результате при правильно спроектированной геймификации участники обретают ощущение мастерства, компетентности и увлечённости процессом (flow) [5].

Например, лабораторные исследования показывают, что в дистанционном обучении при использовании элементов геймификации (очки, значки, таблицы лидеров) существенно увеличивается запоминание материала и улучшение результатов работы: геймификация значимо повышает удержание знаний и **учебные результаты пользователей** [11].

Авторы эксперимента с European Training Sessions сообщают, что применённая геймификация прямо усиливала вовлечённость участников обучения, а это, в свою очередь, улучшало способность запоминать и применять знания на практике [11].

С точки зрения теорий мотивации, геймификация подкрепляет фундаментальные психологические потребности человека. Так, SDT (Теория самоопределения Деси и Райана) выделяет три базовые потребности: автономию (самостоятельность), компетентность (умение) и сопричастность (социальные связи) [6].

Грамотно выстроенная геймификация позволяет удовлетворять эти потребности: например, система очков и уровней укрепляет ощущение компетентности, задания с выбором ветки обучения – автономию, а совместные квесты – социальную сопричастность. Кроме того, модель типов игроков Бартла (1996) выделяет четыре категории мотивации: достигатели, исследователи, социализаторы и «килеры» [7]. В **дистанционном обучении** обучении это означает, что разным **обучающимся** требуются разные стимулы: достигателям важны видимые индикаторы прогресса, исследователям – новые области знаний и скрытые сюжеты, социализаторам – командные соревнования, а соревнующимся («килерам») – рейтинги и вызовы [7]. Игровая система должна учитывать эти различия, иначе унифицированная геймификация рискует потерять эффективность или даже демотивировать некоторых пользователей [8].

Однако исследования отмечают и ограничения геймификации.

Во-первых, эффект часто краткосрочен, если игровые элементы не адаптированы к контексту и личности пользователя [12].

Во-вторых, перенасыщение внешними мотиваторами (баллами, ранжированием) может отвлекать от учебных целей, если не учитывать реальные потребности обучающихся [4]. Следовательно, сегодня во многих работах подчёркивается необходимость *адаптивной* или *персонализированной* геймификации, когда игра подстраивается под индивидуальные характеристики обучающегося [8].

Эта идея лежит в основе предлагаемой концепции.

Теоретические основы персонализации

Персонализация обучения – это адаптация учебного процесса под уникальные характеристики каждого ученика (или пользователя).

В цифровых системах это обычно реализуется через анализ данных о пользователях (результаты тестов, предпочтения, скорость выполнения заданий) и динамический подбор контента и траекторий. Адаптивные технологии обучения способны в реальном времени изменять содержание, уровень и темп материала в зависимости от текущих потребностей ученика.

Например, система может автоматически усложнять задачи, когда пользователь быстро справляется с учебным материалом, или, наоборот, предлагать дополнительные пояснения и упрощённые материалы при затруднениях. Такие технологии позволяют обеспечить обучение, «подстроенное» под индивидуальные темпы и стили, что, по данным многих исследований, повышает эффективность усвоения.

Обзор 69 исследований показывает, что персонализированное адаптивное обучение приводило к улучшению академической успеваемости в 59 % случаев, а в 36 % наблюдалось повышение вовлечённости обучающихся [9].

Иными словами, в большинстве экспериментов внедрение адаптивных механизмов дало статистически значимый прирост результатов по сравнению со стандартным обучением. **В прикладных исследованиях цифрового обучения также отмечается**, что внедрение рекомендательных механизмов повышает завершение курсов и активность пользователей: в случае **онлайн-платформы** Tele2 после внедрения рекомендательной системы по выбору курсов 91 % сотрудников прошли тренинги, и те, кто получал персональные рекомендации, стали активнее учиться и чаще доводить курсы до конца, участники без рекомендаций [10]. Персонализация усиливает чувство личного прогресса и осмысленности обучения – **пользователи** видят, что компания инвестирует в их развитие, что, в свою очередь, снижает текучесть кадров и повышает лояльность.

Тем не менее, персонализация также имеет ограничения.

Во-первых, эффективная адаптация требует сбора и обработки больших объёмов данных о пользователях, качественных алгоритмов искусственного интеллекта и аналитической платформы. Это увеличивает технологические и финансовые затраты на внедрение таких систем [9].

Во-вторых, возможны риски неправильных рекомендаций или «ловушек» узких специализаций, когда система слишком уж ориентируется на выявленные слабости пользователя. Наконец, интеграция персонализации и геймификации усложняет архитектуру приложения и может вызывать вопросы безопасности данных. Эти ограничения важно учитывать при проектировании системы и учитывать их влияние на конечные результаты.

Интеграция геймификации и персонализации

Современные исследования отмечают тренд синергетического объединения геймификации и персонализации – когда игровые механики адаптируются к индивидуальному прогрессу и потребностям обучающегося.

Такая адаптивная геймификация обеспечивает более сильный эффект, чем раздельное применение методов.

Во-первых, играя на сильных мотивационных сторонах пользователя, система стимулирует его вовлечённость и желание действовать (например, зарабатывать очки и прохождение уровней).

Во-вторых, в момент ослабления интереса персонализация подстраивает сложность и формат материала, не допуская фрустрации.

Комбинация этих подходов даёт эффект «непрерывного прогресса»: по мере обучения пользователь видит, что достижение игровых целей и образовательных результатов идут «рука об руку», что поддерживает его мотивацию [11].

На практике это означает, что система геймификации принимает решения на основе данных персонализации. Например, на начальных этапах обучения новичкам могут выдавать повышенные «приветственные» очки для быстрого достижения первого уровня (стимул продолжать), а продвинутые пользователи получат более сложные вызовы. Матрица игровых контекстов строится с учётом профиля пользователя: для «социализаторов» особенно важны групповые квесты и командные соревнования, для «исследователей» – побочные миссии и скрытые достижения. Таким образом, предлагаемый подход учитывает психологию разных **обучающихся** и **образовательные цели** одновременно.

По предварительным оценкам, интеграция геймификации и персонализации должна привести к значительному росту ключевых показателей: повышению вовлечённости, запоминания материала и удовлетворённости обучением.

Например, в описанном **эксперименте** из JIK (2024) геймификация уже усилила сохранение знаний и результативность обучения [1].

С учётом добавления персонализации можно ожидать дальнейшего усиления таких эффектов, поскольку игра станет *осмысленной частью* индивидуального учебного плана каждого пользователя.

Архитектура модели и компоненты системы

Предлагаемая архитектура интегрированной модели включает три ключевых компонента: модуль персонализации, модуль геймификации и интеграционный слой. Ниже приведена их функция в системе.

Модуль персонализации (User Profiling): собирает данные о поведении и прогрессе каждого **пользователя** – результаты тестов, время на задачах, ответы и т.д.

На основе этой информации строится динамический пользовательский профиль. Алгоритмы машинного обучения анализируют пробелы в знаниях и цели обучения и формируют индивидуальную программу. Система адаптивно выбирает темы и задания нужной сложности, предлагает дополнительные пояснения в слабых областях, а также рекомендует курсы по интересам пользователя. Эти механизмы реализуют контекстную адаптацию содержимого и гибкие учебные траектории [9].

Модуль геймификации (Game Design Engine): отвечает за построение игровых стимулов и их динамическое изменение. На этом уровне определяется состав игровых элементов (очки опыта, значки, уровни, лидерборды, сюжеты) и правил их начисления.

При этом алгоритм учитывает профиль пользователя: какие игровые мотиваторы для него наиболее важны. Модуль генерирует задания и сюжеты (квесты) в соответствии с учебными вехами, создавая видимый прогресс (например, герой проходит уровни, когда **пользователь** осваивает темы).

При реализации используются кластеры игровых элементов: направление деятельности (достижения, очки за выполнение задач), социальные элементы (общие задачи, рейтинги, совместные действия), персональные мотивации (личные цели, аватары, разблокируемый контент) [8].

Интеграционный слой (Integration Layer): обеспечивает связку между LMS (обучающей платформой) и игровыми модулями.

Он синхронизирует события обучения и геймификации: например, при успешном завершении урока система передачи данных уведомляет игровую логику о новых очках или достижениях. Этот слой контролирует общий прогресс, следит за выполнением геймифицированных условий и при необходимости корректирует пользовательский профиль.

Важная задача – обеспечить консистентность метрик (например, количество изученных уроков соответствует уровню в игре). Интеграционный интерфейс также управляет безопасностью и конфиденциальностью данных пользователей, которые используются для персонализации.

Таблица 4.1 Компоненты системы и их функции

Компонент системы	Функции
Модуль персонализации	Анализ результатов тестов, построение профиля, подбор контента по индивидуальным потребностям, формирование адаптивных траекторий обучения.
Модуль геймификации	Создание и управление игровыми элементами (очки, уровни, значки), генерация квестов и челленджей в соответствии с прогрессом, мотивирование через сюжеты и соревнования.
Интеграционный слой	Обмен данными между LMS и игровой платформой, синхронизация прогресса обучения и игровых достижений, объединение метрик, обеспечение безопасности.

Архитектура позволяет в реальном времени соединять обучение и игру. По мере прохождения дистанционного курса происходит событие-лог (например, тест сдан, пройдён модуль), которое поступает на интеграционный слой. Тот обновляет персональный профиль и передаёт информацию в модуль геймификации: формируется соответствующее игровое событие (например, повышение уровня).

Одновременно модуль персонализации решает, какие следующие учебные задания предложить пользователю. В результате обучение становится «игрой» с адаптивной подстройкой.

Ожидаемые эффекты и ограничения

Ожидается, что интеграция геймификации и персонализации приведёт к значительному росту мотивации и удержанию **пользователей** в обучающих приложениях.

По аналогии с исследованиями в образовании, можно прогнозировать улучшение усвоения знаний, повышение успешности прохождения курсов и большей концентрации внимания [11][9]. Внутренние эксперименты показывают: геймифицированные курсы усиливают вовлечённость и положительно влияют на отношение к обучению [11][12].

А адаптивные механизмы дополнительно увеличивают образовательную эффективность, поскольку поддерживают внимание и помогают ученикам продвигаться в своём темпе [9][10]. Вместе эти эффекты должны обеспечить более высокий процент завершённых программ и лучшую интеграцию знаний в практическую деятельность.

При этом необходимо учитывать ограничения.

Во-первых, технически интегрированная система сложна и может требовать значительных ресурсов на разработку и внедрение [9]. Затраты связаны с созданием качественного контента, аналитической инфраструктуры и поддержкой алгоритмов ИИ.

Во-вторых, эффективность решения напрямую зависит от качества исходных данных: неполные или неточные пользовательские профили могут приводить к ошибочным персональным рекомендациям. Также возможна «усталость» от игрового контента, если механика будет слишком навязчивой или несоответствующей цели обучения [4].

Наконец, существует риск увеличения разрыва между технически подкованными и менее опытными пользователями, если система не адаптируется под разные группы пользователей.

Успешность внедрения во многом зависит от грамотного проектирования и оценки результатов.

Практические рекомендации

Для эффективной реализации интегрированной модели рекомендуется:

1. **Чёткое выравнивание с бизнес-целями.** Перед внедрением нужно определить ключевые показатели (KPI) обучения: повышение производительности, снижение ошибок,

удержание персонала и т.п. Игровые механики и адаптация контента должны служить этим целям. Например, если задача – улучшить освоение новых продуктов, геймификация может поощрять изучение соответствующих модулей достижениями и бонусами.

2. **Сегментация аудитории и поэтапное тестирование.** Целесообразно начать с пилотных групп разных профилей (новичков и опытных, разных возрастных групп и уровней подготовки) и отрегулировать систему под их потребности. Разные группы могут предъявлять разные требования к интерфейсу и темпам обучения. Постепенное расширение системы позволит выявить «узкие места» и скорректировать механику без больших потерь.

3. **Гибкая настройка игровых элементов.** Не стоит добавлять много элементов сразу – это может перегрузить пользователя. Лучше динамически «включать» новые стимулы (например, новые уровни или командные миссии) по мере роста навыков пользователей. Управление такими настройками должно быть централизованным и доступным администраторам платформы.

4. **Использование аналитики и обратной связи.** Важно регулярно анализировать метрики вовлечённости и успеваемости: например, сравнивать группу с геймификацией и контрольную без неё. Обратная связь от **пользователей** также поможет улучшать модель. При снижении активности нужно оперативно вмешиваться – скорректировать сложность задач, сменить формат подачи.

5. **Обеспечение безопасности и прозрачности.** Поскольку система опирается на персональные данные, необходимо соблюдать законы о защите информации. Пользователи должны знать, какие данные собираются и как они влияют на обучение. Прозрачность увеличивает доверие к системе и принятию персонализации.

Заключение. В статье рассмотрен подход к повышению удержания пользователей в дистанционных обучающих приложениях через интеграцию геймификации и персонализации. Проанализированы теоретические основы обоих методов (игровые мотиваторы, психологические потребности, адаптивное обучение) и показано, что их комбинация – адаптивная геймификация – потенциально преодолевает ограничения каждого подхода в отдельности [11].

Предложена концептуальная архитектура системы, включающая модули персонализации и геймификации, интегрированные на уровне данных и бизнес-логики.

Ожидается, что применение такой модели в цифровой образовательной среде повысит мотивацию сотрудников, улучшит результаты обучения и снизит текучесть кадров. При этом требуются тщательная настройка и учёт особенностей целевой аудитории.

Список литературы

[1] Gallup. State of the Global Workplace 2025 [Электронный ресурс]. – 2025. – Режим доступа: https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx (дата обращения: 19.02.2026).

[2] Лебедев П. Стали известны лидеры российского рынка LMS для корпоративного обучения [Электронный ресурс] // Skillbox Media. – 2025. – Режим доступа: https://skillbox.ru/media/corptrain/stali-izvestny-lidery-rossiyskogo-rynka-lms-dlya-korporativnogo-obucheniya/ (дата обращения: 19.02.2026).

[3] Чигинцева А.А. Актуальные проблемы дистанционного обучения // Скиф. 2018. № 3 (19). С. 10–13.

[4] Gamification (GF) – What is Gamification? [Электронный ресурс] // Interaction Design Foundation. – Режим доступа: https://www.interaction-design.org/literature/topics/gamification (дата обращения: 19.02.2026).

[5] Hamari J., Koivisto J., Sarsa H. Does gamification work? A literature review of empirical studies on gamification // Proc. 47th Hawaii Int. Conf. on System Sciences. 2014. С. 3025–3034.

[6] Ryan R.M., Deci E.L. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being // American Psychologist. 2000. Vol. 55, № 1. С. 68–78. – Режим доступа: https://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf (дата обращения: 19.02.2026).

[7] Bartle taxonomy of player types [Электронный ресурс] // Википедия. – Режим доступа: https://en.wikipedia.org/wiki/Bartle_taxonomy_of_player_types (дата обращения: 19.02.2026).

[8] Hong Y., Saab N., Admiraal W. Approaches and game elements used to tailor digital gamification for learning: a systematic literature review [Электронный ресурс] // Computers & Education. 2024. – Режим доступа: https://www.sciencedirect.com/science/article/pii/S0360131524000149 (дата обращения: 19.02.2026).

[9] du Plooy E., Casteleijn D., Franzsen D. Personalized adaptive learning in higher education: a scoping review of key characteristics and impact on academic performance and engagement [Электронный ресурс] // Heliyon. 2024. Vol. 10, № 21. e39630. – Режим доступа: https://www.sciencedirect.com/science/article/pii/S2405844024156617 (дата обращения: 19.02.2026).

[10] Исследования показали, как корпоративное обучение влияет на вовлечённость персонала и бизнес-показатели [Электронный ресурс] // Комсомольская правда – Челябинск. – 2023. – 20 апр. – Режим доступа: https://www.chel.kp.ru/daily/27498/4758224/ (дата обращения: 19.02.2026).

[11] Capatina A., Juarez-Varon D., Micu A., Micu A.E. Leveling up in corporate training: Unveiling the power of gamification to enhance knowledge retention, knowledge sharing and job performance [Электронный ресурс] // Journal of Innovation & Knowledge. 2024. Т. 9. № 3. Статья 100530. – Режим доступа: https://www.sciencedirect.com/science/article/pii/S2444569X24000696 (дата обращения: 19.02.2026).

[12] Khoshnoodifar M., Ashouri A., Taheri M. Effectiveness of Gamification in Enhancing Learning and Attitudes: A Study of Statistics Education for Health School Students [Электронный ресурс] // J Adv Med Educ Prof. 2023. Т. 11. № 4. С. 230–239. – Режим доступа: https://pmc.ncbi.nlm.nih.gov/articles/PMC10611935/ (дата обращения: 19.02.2026).

Авторский вклад

Мосендз Александра Викторовна – постановка цели и задач исследования, анализ теоретических основ геймификации и персонализации, разработка концептуальной модели интеграции игровых механик с адаптивным обучением, описание архитектуры и ключевых компонентов системы, подготовка текста статьи, формулирование выводов и практических рекомендаций.

Хацкевич Геннадий Алексеевич – критический анализ предлагаемой модели, проверка корректности научного аппарата, обсуждение ограничений исследования, доработка итогового текста.

INFLUENCE OF GAMIFICATION AND PERSONALIZATION INTEGRATION ON USER RETENTION IN DIGITAL EDUCATIONAL APPLICATIONS

A.V. Mosenz
Master's student, Belarusian State University of Informatics and Radioelectronics

G.A. Khatskevich
Professor, Department of Economic Informatics, BSUIR, Doctor of Economic Sciences, Professor

Abstract. In the context of the growing need to increase motivation and user retention in digital learning, combined approaches are considered. The article analyzes the theoretical foundations of gamification and learning personalization, identifies their strengths and limitations when applied separately. A conceptual model for integrating game mechanics with adaptive learning is proposed, in which dynamic game incentives are correlated with individual user progress. The architecture and key components of the model are described, including personalization mechanisms, contextual design of game incentives, and integration mechanisms. It is noted that the synergy of gamification and personalization should enhance engagement and retention, compensating for the shortcomings of each approach. Expected effects, practical recommendations, and limitations of the proposed model are discussed.

Keywords: gamification, learning personalization, adaptive learning, user retention, engagement, digital educational applications, adaptive gamification, game mechanics, motivation, integration model.