На правах рукописи



ТРОФИМОВ ЮРИЙ ВЛАДИСЛАВОВИЧ

МЕТОДЫ И ГЕТЕРОГЕННАЯ ГИБРИДНАЯ ПЛАТФОРМА МНОГОУРОВНЕВОГО АНСАМБЛИРОВАНИЯ ОБЪЯСНЕНИЙ ПРИ КОМПЛЕКСИРОВАНИИ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Специальность: 1.2.1 –

«Искусственный интеллект и машинное обучение»

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2026

Работа выполнена на кафедре прикладной математики и искусственного интеллекта федерального государственного бюджетного образовательного учреждения высшего образования «Национальный исследовательский университет «МЭИ».

Научный руководитель:	**Аверкин Алексей Николаевич** Кандидат физико-математических наук, доцент, ведущий научный сотрудник отдела № 14, ФИЦ ИУ РАН
Официальные оппоненты:	**Палюх Борис Васильевич** Доктор технических наук, профессор, заведующий кафедрой информационных систем Тверского государственного технического университета (ТвГТУ) **Харламов Александр Александрович** Доктор технических наук, профессор департамента программной инженерии факультета компьютерных наук НИУ «Высшая школа экономики»
Ведущая организация:	ФГБОУ ВО «Тверской государственный университет» (ТвГУ)

Защита диссертации состоится «__» ____ 2026 г. в __час. __мин. на заседании диссертационного совета 24.1.224.03 на базе ФИЦ ИУ РАН по адресу: 119333, г. Москва, ул. Вавилова, д. 42.

С диссертацией можно ознакомиться в библиотеке ФИЦ ИУ РАН и на официальном сайте http://www.frccsc.ru.

Автореферат разослан «__» ____ 2026 г.

Учёный секретарь
диссертационного совета 24.1.224.03,
к.т.н. И.А. Рейер

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Модели искусственного интеллекта (ИИ) находят всё более широкое применение в областях, где цена ошибки велика. В таких задачах высокая точность сама по себе ещё не даёт оснований доверять результату. Нужно проследить, как получено решение, оценить его устойчивость и степень достоверности. Объяснимость становится здесь необходимым условием надёжного и практически значимого применения ИИ, а также воспроизводимости его результатов. Считать её факультативным свойством модели уже нельзя. Государственная значимость развития ИИ закреплена Указом Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации» и Указом Президента Российской Федерации от 15.02.2024 № 124. Последним внесены изменения в указанный акт и в Национальную стратегию развития искусственного интеллекта на период до 2030 года.

Острее всего проблема объяснимости стоит в здравоохранении. Ошибочное решение здесь напрямую влияет на жизнь и здоровье пациента, медицинские данные часто неполны и неоднозначны, а результат модели должен согласовываться с профессиональной логикой врача. Поэтому к интерпретируемости и обоснованности вычислительного вывода предъявляются повышенные требования. Потребность в методах объяснимого ИИ растёт по мере внедрения технологий ИИ в медицинскую практику и цифровой трансформации отрасли, которая предусматривает единый цифровой контур здравоохранения, структурированные электронные медицинские документы и развитие систем поддержки принятия врачебных решений. Эти направления отражены в распоряжении Правительства Российской Федерации от 17.04.2024 № 959-р и в постановлении Правительства Российской Федерации от 18.07.2023 № 1164.

С усложнением интеллектуальных систем одного уровня интерпретации перестаёт хватать. Практически значимое объяснение должно связывать признаки, внутренние представления модели и содержательные выводы в единую картину. Отсюда следует необходимость формировать объяснения на нескольких уровнях, объединять их и явно учитывать неопределённость объединённого результата. Переход от отдельных объяснений к их согласованной системе и составляет одну из ключевых научных задач исследования.

Этим определяется актуальность темы диссертационного исследования, посвящённого разработке методов, моделей и алгоритмов многоуровневого формирования, согласования и объединения объяснений в системах ИИ с учётом неопределённости.

Объект и предмет исследования. Объектом исследования являются процессы формирования, интерпретации и использования объяснений в системах ИИ в условиях неопределённости.

Предметом исследования являются методы, модели и алгоритмы многоуровневого формирования, согласования, агрегирования и оценки объяснений в гетерогенных системах объяснимого искусственного интеллекта.

Степень разработанности проблемы. Построение систем поддержки принятия решений, интеллектуальных систем и методов объяснимого искусственного интеллекта (ОИИ) изучалось как в СССР и России, так и за рубежом. Из отечественных учёных вклад в эти направления внесли А.Н. Аверкин, В.В. Борисов, В.Б. Тарасов, О.И. Ларичев, Н.Н. Моисеев, В.П. Кутепов, А.П. Еремеев, Г.В. Рыбина, И.И. Зурабов, В.Л. Арлазаров, Е.В. Морозов, Ю.Н. Бабанский, А.Г. Ивахненко, Б.Г. Соколов, О.В. Бондаренко, Е.А. Ляпунов, А.А. Смирнов и Д.В. Садовников. На современные подходы существенно повлияли также работы Г. Саймона, М. Минского, М. Месаровича, Л. Заде, Р. Алиева, Б. Коско, Я. Лекуна, Й. Бенжио, Д. Хинтона, Ю. Шмидхубера, Т. Диттериха, Т. Митчелла, Дж. Шоу-Тейлора, Т. Хасти, Р. Тибширани, К. Бишопа, В. Вапника и Дж. Пёрла. Нерешёнными при этом остаются согласование модельных и постмодельных объяснений разных уровней, формальные процедуры агрегирования и комплексирования гетерогенных объяснений, учёт неопределённости итогового объяснения и воспроизводимый выбор объяснителя.

Целью исследования является повышение качества и воспроизводимости объяснений, формируемых для решений моделей машинного обучения, на основе разработки методов их многоуровневого формирования, согласования и объединения с учётом неопределённости, а также автоматизированного выбора метода объяснения.

Для достижения поставленной цели были решены следующие **задачи:**

1. провести системный анализ методов, моделей и подходов объяснимого искусственного интеллекта с целью выявления ограничений существующих решений, нерешённых задач многоуровневого формирования объяснений, их объединения и оценки качества;
2. разработать методы многоуровневого формирования объяснений и их согласования в иерархически организованных объяснимых системах ИИ;
3. разработать механизм агрегирования и объединения объяснений, включая операции над многоуровневыми модельными и постмодельными объяснениями и оценку неопределённости объединённого результата;
4. сформировать протокол воспроизводимого выбора объяснителя и метрический контур оценки качества объяснений, связывающий качество с устойчивостью и воспроизводимостью;

5. разработать архитектуру платформы и программных компонентов, обеспечивающих оркестрацию выбора модели, формирования объяснения и процедуру проверки качества;
6. апробировать платформенный подход на комплексе биомедицинских задач, демонстрирующих применимость многоуровневых объяснений, объединения и учёта неопределённости в практическом контуре.

Научная новизна диссертационного исследования заключается в следующем:

- метод многоуровневого формирования объяснений в иерархически организованных системах ИИ, отличающийся представлением объяснимого контура в виде ориентированного графа узлов и переходов и формированием объяснительных объектов на его узлах и рёбрах с последующим выделением признакового, модельного, интерпретируемого и результирующего уровней;
- метод согласования уровней объяснения, отличающийся использованием операторов межуровневого переноса, мер расхождения и показателей согласованности для установления формально проверяемой связи между признаковым, модельным, интерпретируемым и результирующим уровнями;
- формальный аппарат агрегирования, объединения и комплексирования гетерогенных объяснений, отличающийся разделением способов совместной обработки однородных и разнородных объяснительных объектов и сохранением их типа и происхождения в итоговом объяснении;
- метод квантификации неопределённости объединённого объяснения, учитывающий расхождения между различными методами объяснения, повторными запусками и моделями, а для мультимодальных систем также согласованность объяснений различных модальностей;
- протокол воспроизводимого выбора метода объяснения и метрический контур его оценки, отличающиеся совместным учётом качества, устойчивости, неопределённости, вычислительной стоимости и контекста применения в фиксированной процедуре сопоставления кандидатов.

Методы исследования. Задачи решаются методами дискретной математики, теории алгоритмов и теории вероятностей. Применяются также методы интеллектуального анализа данных и машинного обучения, обработки и анализа изображений, системного анализа, а также проектирования, реализации и верификации программных систем и вычислительных контуров.

Теоретическая и практическая значимость работы. Теоретическая значимость заключается в развитии положений объяснимого искусственного интеллекта о многоуровневом

иерархическом формировании объяснений, их агрегировании и объединении с учётом неопределённости, а также в построении сопоставимых процедур оценки качества объяснений, учитывающих их устойчивость и воспроизводимость. Практическая значимость состоит в том, что на этой основе разработаны архитектура и программные компоненты платформы программно-методического комплекса для обучения и применения моделей, автоматизированного подбора моделей и методов объяснения и построения конвейеров обработки и анализа данных. Платформа формирует и оценивает объяснения с учётом неопределённости и повышает воспроизводимость вычислительных экспериментов.

Положения, выносимые на защиту:

1. Методы многоуровневого иерархического формирования и согласования объяснений в гетерогенных системах искусственного интеллекта, реализуемые в составе вычислительного контура интеллектуальной системы.
2. Методы агрегирования, объединения и комплексирования при ансамблировании гетерогенных объяснений, обеспечивающие совместное использование различных типов объяснений с учётом неопределённости итогового результата.
3. Протокол воспроизводимого выбора метода объяснения на основе сопоставимой оценки качества, устойчивости, неопределённости, вычислительной стоимости и воспроизводимости объяснений, обеспечивающий автоматизацию выбора.
4. Архитектурные, алгоритмические и программные решения гетерогенной гибридной платформы, обеспечивающие оркестрацию вычислений, реализацию разработанных методов формирования и оценки объяснений и их применение в задачах анализа данных и поддержки принятия решений.

Достоверность научных результатов подтверждается корректным использованием методов, обоснованием постановки задач, а также экспериментальными исследованиями на открытых данных и данных индустриальных партнёров, демонстрирующими эффективность предложенных методов и алгоритмов и корректность работы разработанных программных решений.

Реализация результатов работы. По результатам диссертационного исследования разработаны три программы для ЭВМ, на которые получены свидетельства о государственной регистрации. Разработанные методы, алгоритмы и программные средства внедрены при создании специализированного программного обеспечения системы поддержки принятия врачебных решений в составе программного продукта ООО «МАКАО ИТ» и использованы в совместных с ГБУЗ Московской области «Дубненская больница» научно-практических исследованиях в области ИИ для медицинской диагностики, в деятельности Лаборатории

информационных технологий им. М. Г. Мещерякова Объединённого института ядерных исследований, научно-исследовательского центра ИИ и в учебном процессе Университета «Дубна», что подтверждено актами внедрения и использования и письмами поддержки.

Результаты исследований также использованы при выполнении научно-исследовательских работ «Применение объяснительного искусственного интеллекта для интерпретации алгоритмов машинного обучения» (государственное задание Министерства науки и высшего образования Российской Федерации, тема № 124112200072-2), «Гибридные модели поддержки принятия решений на основе методов дополненного искусственного интеллекта, когнитивного моделирования и нечёткой логики в задачах персонализированной медицины» (грант РНФ, проект № 22-71-10112) и «Научное обоснование безопасного захоронения твёрдых коммунальных отходов в геологической среде» (грант РНФ, проект № 22-17-00045).

Апробация результатов. Основные положения диссертационной работы докладывались и обсуждались на научных семинарах в Национальном исследовательском университете «МЭИ» (2026) и Государственном университете «Дубна» (2024–2026), а также на 26 международных и всероссийских научных конференциях: AI-ZAMAN 2025 Interdisciplinary Conference on Artificial Intelligence: from theory to practice (Казань, 2025), IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (Новосибирск, 2024), VI и VII Международной конференции по нейронным сетям и нейротехнологиям (NeuroNT'2025, 2026, Санкт-Петербург, ЛЭТИ), 11th International Conference “Distributed Computing and Grid Technologies in Science and Education” (GRID'2025, Дубна), XXII Национальной конференции по искусственному интеллекту с международным участием (КИИ-2025, Санкт-Петербург), IX Международной научной конференции «Интеллектуальные информационные технологии для индустрии» (Федеральная территория «Сириус», 2025), XXXIII Международной конференции «Математика. Компьютер. Образование» (Дубна, 2026), XVII SRPSKI Simpozijum O Hidrogeologiji sa međunarodnim učešćem (Пирот, Сербия, 2024), International Science Conference “Science. Education. Practice” (Дели, Индия, 2025), Scientific Research of the SCO Countries: Synergy and Integration (Пекин, Китай, 2025), XXIX Международная конференция по мягким вычислениям и измерениям (SCM'2026 Санкт-Петербург, ЛЭТИ), 27th International Conference of Young Professionals in Electron Devices and Materials (EDM) (Республика Алтай, 2026), International Conference on Industrial Engineering (Сочи, 2026), International Russian Smart Industry Conference (Сочи, 2026), XIII Международная научно-практическая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (ИММВ-2026, Коломна), Донецкий международный круглый стол искусственный интеллект: теоретические аспекты и практическое применение (ДМКС-ИИ-

2026), а также на XXIX, XXX и XXXI Всероссийских научно-практических конференциях студентов, аспирантов и молодых специалистов (Дубна, 2024, 2025, 2026).

Публикации. Результаты диссертации опубликованы в 40 работах, в том числе в 15 статьях в журналах из перечня ВАК РФ (3 статьи по специальности 1.2.1) и в 10 статьях в изданиях, индексированных в международных базах Web of Science и Scopus. Получены 3 свидетельства о государственной регистрации программ для ЭВМ.

Личный вклад автора. Основные научные результаты, представленные в диссертации, получены автором лично под научным руководством к.ф.-м.н., доцента Аверкина А.Н. Постановка научной задачи, формирование концептуальных подходов и определение направлений исследования осуществлялись совместно с научным руководителем. Автором лично разработаны методы, алгоритмы и модели, выполнено проектирование архитектуры и реализация программного обеспечения, проведены вычислительные эксперименты, анализ и интерпретация полученных результатов, а также подготовка научных публикаций по теме исследования.

Соответствие паспорту специальности.

Исследование выполнено в соответствии со следующими четырьмя научными направлениями паспорта специальности 1.2.1 – «Искусственный интеллект и машинное обучение»:

п. 2. Исследования в области оценки качества и эффективности алгоритмических и программных решений для систем искусственного интеллекта и машинного обучения. Методики сравнения и выбора алгоритмических и программных решений при многих критериях.

п. 4. Разработка методов, алгоритмов и создание систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке, для изображений, речи, биомедицины и других специальных видов данных.

п. 7. Разработка специализированного математического, алгоритмического и программного обеспечения систем искусственного интеллекта и машинного обучения. Методы и средства взаимодействия систем искусственного интеллекта с другими системами и человеком-оператором.

п. 12. Исследования в области «доверенных» систем класса ИИ, включая проблемы формирования тестовых выборок прецедентов, надежности, устойчивости, переобучения и т.д.

Структура и объём диссертации. Диссертация состоит из введения, четырёх глав, заключения, списка литературы и трёх приложений. Полный объём диссертации составляет 219 страниц, включая 55 рисунков и 15 таблиц. Список литературы содержит 243 наименования.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность темы, сформулированы цель и задачи, определены объект и предмет исследования, раскрыты научная новизна, теоретическая и практическая значимость, приведены положения, выносимые на защиту.

В первой главе выполнен системный обзор предметной области ОИИ (рисунки 1, 2). Рост точности нейросетевых моделей сопровождался снижением их прозрачности. В медицинских приложениях эта непрозрачность стала самостоятельной научной проблемой. Объяснимость рассматривается как одно из свойств доверенного ИИ наряду с безопасностью, устойчивостью, справедливостью и воспроизводимостью вычислений.



Рисунок 1 – Карта системного анализа предметной области объяснимого искусственного интеллекта

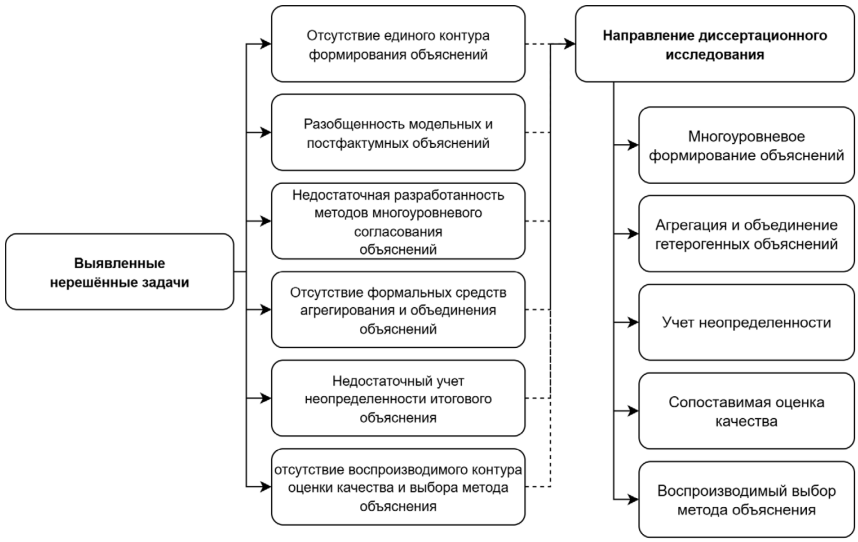


Рисунок 2 – Карта выявленных направлений диссертационного исследования

В работе разграничены понятия, которые в литературе нередко смешиваются. Прозрачность характеризует систему и её жизненный цикл, интерпретируемость — устройство модели, объяснимость — возможность раскрыть основания конкретного вывода, интерпретация — форму, в которой объяснение предъявляется человеку. Анализ стандартов ГОСТ Р 59276–2020 и ISO/IEC 42001:2023 показал, что от системы требуется процедура объяснения, воспроизводимая на тех же данных и при тех же параметрах.

Исторически первыми получили распространение постмодельные методы, позволяющие объяснять уже обученную модель без её изменения. Их объяснения локальны и зависят от выбранного метода и его параметров, а наглядная карта значимости может слабо отражать действительный механизм вывода. Эти ограничения обусловили переход к интерпретируемым моделям, а затем к встроенной архитектурной объяснимости, при которой объяснение формируется в ходе вычислительного процесса (таблица 1).

Таблица 1 – Сравнительная характеристика основных этапов развития объяснимого искусственного интеллекта

Этап развития	**Глубина интерпретации**	**Воспроизводимость**	**Учёт неопределённости**	**Основное ограничение**
Постмодельная интерпретация (первое поколение)	Объясняет отдельные решения, не раскрывая внутренний механизм вывода	Ограниченная, зависит от внешнего метода	Выражен слабо	Локальность и зависимость от метода
Интерпретируемые модели	Позволяют проследить структуру модели и ход решения	Выше, так как связана с устройством модели	Представлен ограниченно	Ограниченная применимость к сложным моделям и данным
Встроенная архитектурная объяснимость (второе поколение)	Связывает результат и процесс вывода в едином контуре	Повышенная, так как включена в архитектуру модели	Учитывается в составе объяснения	Трудности согласования различных уровней объяснения
Измеримый / доверенный искусственный интеллект	Раскрывает результат и границы доверия к нему	Высокая, опирается на процедуры контроля	Явный элемент оценки	Требует сопоставимой оценки качества и воспроизводимого выбора метода объяснения

В решениях, совмещающих автоматизированное машинное обучение с подбором объяснителя (AutoXAI, AutoXAI4Omics, XAutoML), объяснение подключается после оптимизации модели и на ход поиска не влияет. По итогам обзора в число задач диссертации включены согласование уровней объяснения, объединение разнородных объяснений с учётом неопределённости и воспроизводимый выбор метода объяснения.

Во второй главе разработан формальный аппарат многоуровневой объяснимости. Контур обработки данных иерархически организованной системы представлен типизированным ориентированным графом, вершины которого соответствуют функциональным узлам, а рёбра — допустимым переходам данных, внутренних состояний и объяснительных объектов. Операторы отбора извлекают из локального состояния узлов и переходов (значений признаков, весов внимания, активаций слоёв, параметров правил) атрибутивные объекты, образующие признаковый, модельный, интерпретируемый и результирующий уровни (таблица 2). Вслед за Г. Саймоном и М. Месаровичем уровни трактуются как страты описания одной системы, а переходы между ними — как каналы координации. Известные методы SHAP, LIME, Grad-CAM и нечёткие правила используются как механизмы получения объектов отдельных уровней.

Таблица 2 – Уровни объяснения и их функции

Уровень	**Обозначение**	**Что интерпретируется**	**Форма объяснения**	**Роль в итоговом выводе**
Признаки и их вклады	E_f	Значимые входные факторы	Локальная интерпретация	Выделение значимых факторов
Внутренние представления	E_m	Промежуточные зависимости модели	Структурное объяснение	Переход к внутренней логике вывода
Интерпретируемые правила и дескрипторы	E_i	Промежуточные содержательные описания	Содержательное объяснение	Формирование логики вывода
Итоговый вывод	E_r	Конечное решение системы	Сводное объяснение	Обоснование результата

Методы согласования и объединения объяснений сведены в структурно-функциональную модель (рисунок 3).

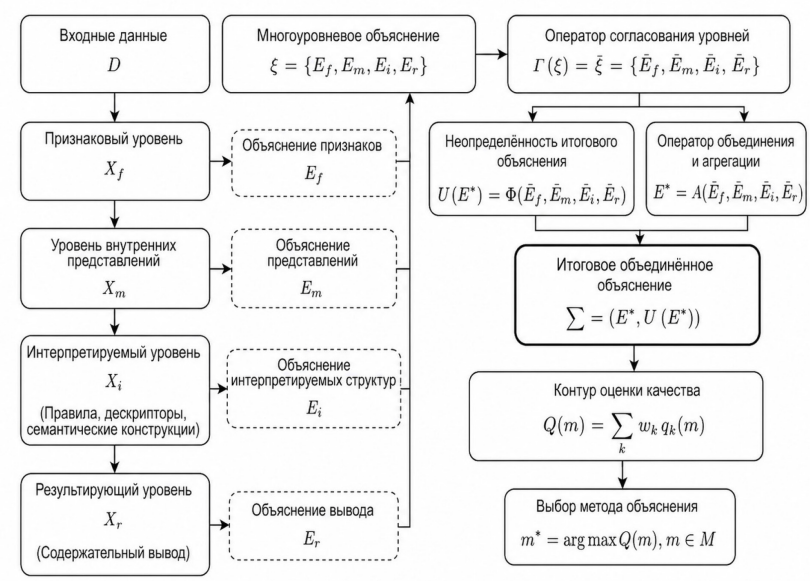


Рисунок 3 – Структурно-функциональная модель согласования и объединения многоуровневых объяснений

Согласование уровней выполняет оператор Γ, переводящий кортеж $E(x) = (E_f, E_m, E_i, E_r)$ в согласованное объяснение $E^{\text{cons}}(x)$. Для каждой допустимой пары уровней задаются оператор переноса $T_{u\to v}$, мера расхождения и порог. Связь включается в объяснение при выполнении условия

$$d_v\left(T_{u\to v}\left(E_u(x)\right), E_v(x)\right) \le \tau_{uv}.$$

Оператор Γ минимизирует взвешенное межуровневое расхождение, а достигнутая согласованность характеризуется индексом $\text{Cons}(x)$. Тем самым известная в теории иерархических систем задача согласования решений перенесена на объяснения.

Для нескольких объяснений одного объекта введены три режима совместной обработки. При агрегировании однородные объяснения усредняются в общем пространстве с весами, зависящими от их качества и неопределённости. Объединение выполняется по уровням с сохранением происхождения компонентов. Комплексирование формирует составной объект из устойчивого ядра, поддерживающих и расходящихся компонентов и метаданных, благодаря чему структура согласия и конфликта сохраняется (рисунок 4).

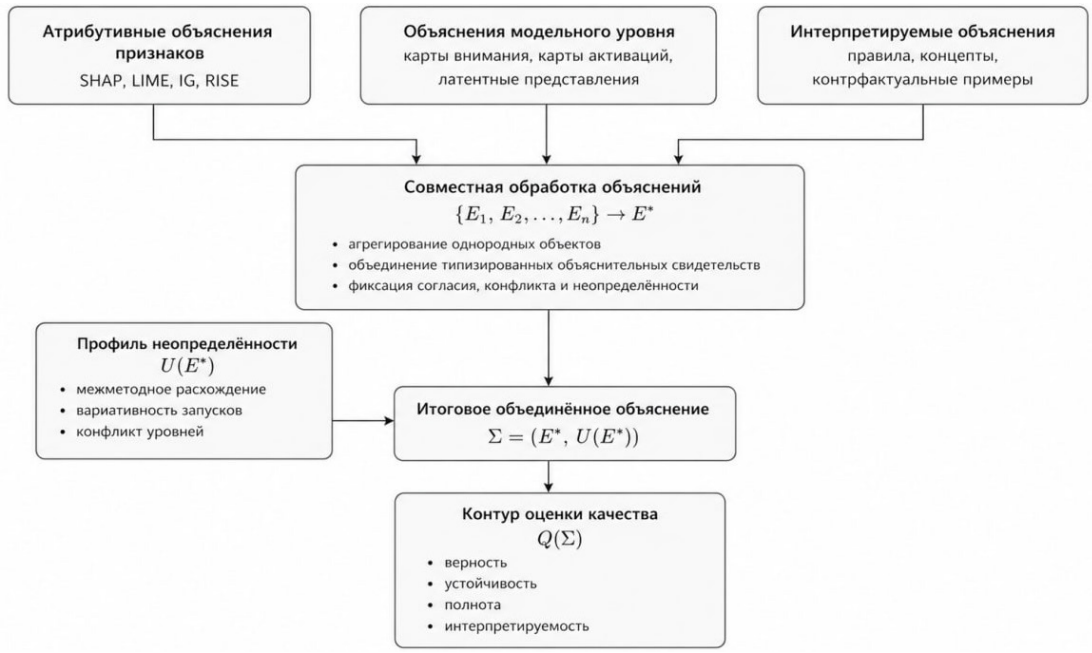


Рисунок 4 – Формализованный контур агрегации, объединения и сопоставимой оценки гетерогенных объяснений

Неопределённость объединённого объяснения оценивается по разбросу частных объяснений и раскладывается на межметодную, внутриметодную и межмодельную составляющие:

$$U_{\text{total}}(x) = \lambda_1 U_{\text{method}}(x) + \lambda_2 U_{\text{run}}(x) + \lambda_3 U_{\text{model}}(x),$$

межмодальная неопределённость приводится отдельно. Результат представляется парой $\mathcal{R}(x) = \big(E^*(x), U_{\text{total}}(x)\big)$.

Метрический контур сворачивает частные показатели качества (верность, устойчивость, полноту, простоту, стоимость) в оценку $Q(E)$. К сравнению допускаются объяснения, для которых $Q(E) \geq \tau_Q$ и $U_{\text{total}} \leq \tau_U$. Аддитивная свёртка применяется только после такого отбора. По Р. Кини и Х. Райфе она корректна при взаимной независимости критериев по предпочтению, а для нормированных показателей это условие выполняется лишь приближённо. Результат сравнения признаётся надёжным, если он сохраняется при варьировании экспертных весов. Классы объяснителей показаны на рисунке 5. Протокол выбора фиксирует контекст задачи, набор кандидатов, калибровочную выборку, метрики и параметры (рисунок 6). Выбирается кандидат с наибольшей средней по выборке контекстной полезностью, к которой в агентной постановке добавляется выигрыш от согласованности оценок агентов.

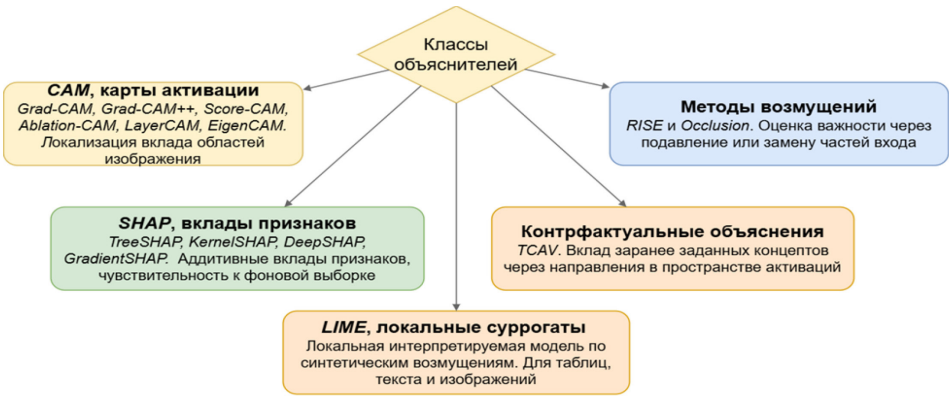


Рисунок 5 – Основные классы объяснителей в задаче выбора метода объяснения

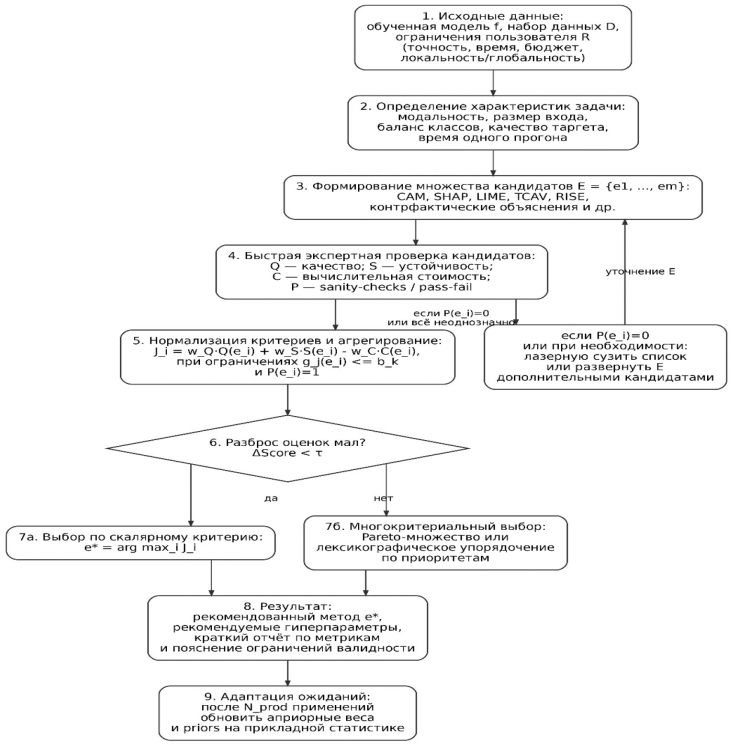


Рисунок 6 – Протокол выбора объяснителя и метрический контур оценки интерпретации

Правило выбора распространено на совместный выбор модели и метода объяснения по четырём критериям. Это прогностическое качество Q_{pred}, достоверность объяснения F (по метрикам вставки и удаления), простота модели и вычислительная эффективность. При заданном бюджете скрининга строится фронт Парето $\mathcal{K}_{\text{par}}$ по Q_{pred} и F, из которого окончательный вариант выбирается нечётким выводом Сугено по всем критериям:

$$(M^*, e^*) = \operatorname{argmax}_{(M,e)\in\mathcal{K}_{\text{par}}} \frac{\sum_{i\in\Lambda} w_i R_i}{\sum_{i\in\Lambda} R_i},$$

где R_i — активация правил уровня предпочтения i, w_i — полезность уровня (а не вес критерия). Нечёткий вывод применяется вместо аддитивной свёртки, чтобы высокая вычислительная эффективность не компенсировала низкую достоверность объяснения.

<u>В третьей главе</u> описана архитектура гетерогенной гибридной платформы, объединяющей автоматизированное машинное обучение и ОИИ в едином воспроизводимом контуре. Описание выполнено по ГОСТ Р 57100–2016 (ISO/IEC/IEEE 42010:2011) в контекстном, компонентном и динамическом представлениях. Организационной основой служит синергетический подход В. Б. Тарасова. Общая схема приведена на рисунке 7.

Методы второй главы реализованы в подсистемах платформы. Ядро управления формирует план выполнения по спецификации запроса. Реестр XAIRegistry хранит допустимые модели, объяснители и метрические протоколы. Подсистема AutoXAI выбирает, настраивает, запускает и сопоставляет методы объяснения. Подсистема метрической верификации оценивает верность, устойчивость и согласованность объяснений, а контуры SafetyGuard и TrustEvaluator проверяют допустимость запуска и рассчитывают индекс доверия с доменными штрафами. Репозиторий моделей и хранилище артефактов фиксируют манифест окружения, что обеспечивает повторяемость эксперимента. Результаты нескольких объяснителей сохраняются в композитном профиле с указанием происхождения, типа, параметров и оценок каждого объяснения, без приведения разнородных объяснений к единой форме. Для ансамбля карт строятся консенсусная карта, карта строгого согласия и карта разброса методов.

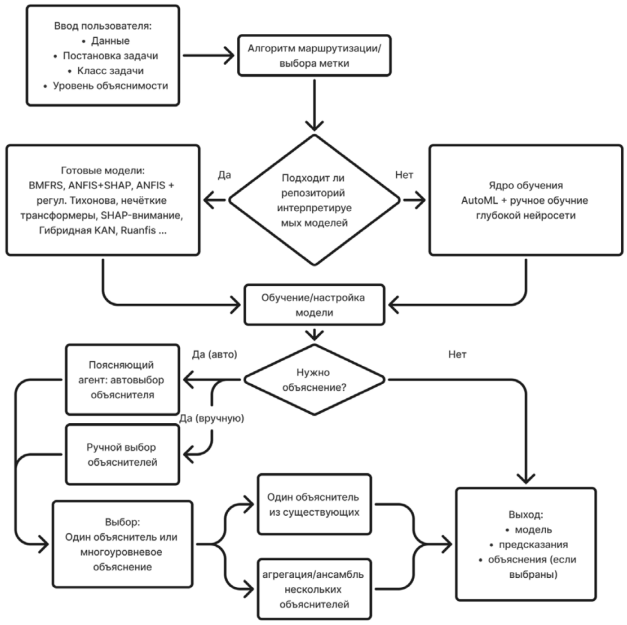


Рисунок 7 – Архитектура и контур формирования объяснений платформы

Маршрутизация запроса (рисунок 8) допускает прямое задание метода, выбор по формализованным правилам и многокритериальный поиск. Многокритериальный поиск выполняет AutoML-ядро. По оценкам, полученным при бюджете раннего скрининга, строится фронт Парето, из которого модуль InterpretabilityAwareSelector выбирает конфигурацию нечётким выводом Сугено. Объяснители назначаются автоматически или вручную, возможно и их ансамблирование.

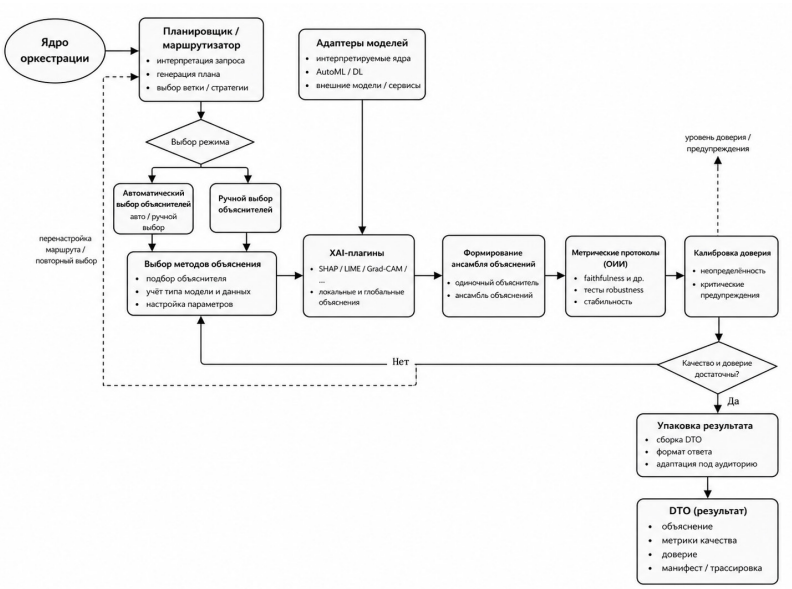


Рисунок 8 – Контур формирования и оценки объяснений в архитектуре платформы

Внешние оркестраторы на основе больших языковых моделей в базовый контур не включены, поскольку они вносят вариативность в построение плана и требуют отдельной проверки на воспроизводимость.

В четвёртой главе выполнена апробация платформы на трёх биомедицинских задачах. В каждой из них платформа формировала обученные модели, подобранные объяснения и конвейер анализа, на основе которых затем разрабатывалось прикладное программное обеспечение.

Первая задача — анализ маммограмм. Набор из 1150 обезличенных исследований (1612 изображений с патологическими изменениями и 2823 без них) предоставлен ГБУЗ Московской области «Дубненская больница». Корректность разметки проверялась врачом-рентгенологом. Модулем автоматизированного машинного обучения отобран детектор YOLOv8, поскольку для передвижного комплекса вычислительная экономичность была обязательным требованием.

Чтобы обойтись без попиксельной разметки, предложена локализованная атрибуция. Карта GradientSHAP вычисляется в окрестности каждой найденной рамки, тогда как прямой проход сети выполняется по полному снимку. В результате фон ткани молочной железы не влияет на объяснение, а пороговая и морфологическая обработка карты даёт область интереса. По её площади, округлости, шероховатости границы и индексу спикуляции формируется ориентировочная оценка по шкале BI-RADS (Breast Imaging Reporting and Data System), которая используется для приоритизации областей. Окончательную категорию определяет врач-рентгенолог. При пятикратной перекрёстной проверке площадь под ROC-кривой (AUC) составила 0,88 для микрокальцинатов, 0,728 для спикуляции и 0,71 для архитектурных нарушений, согласованность карт между повторными запусками — около 0,99. Эти значения

характеризуют отдельные компоненты контура и не являются оценкой клинической эффективности системы. Схема конвейера приведена на рисунке 9.

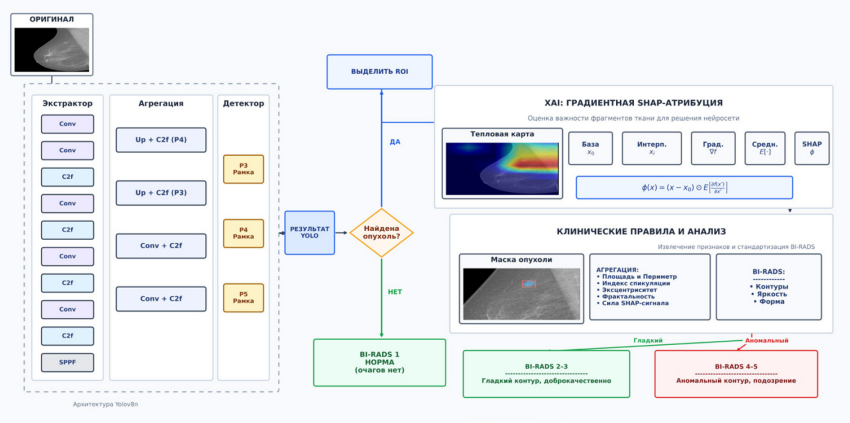


Рисунок 9 – Конвейер локализованной атрибуции SHAP и классификации по шкале BI-RADS

На этой основе разработана веб-система поддержки принятия решений для передвижного маммографического комплекса «ЛУИДОР».

Вторая задача — диагностика коронарных стенозов по ангиографическим изображениям. Использован открытый набор CADICA, дополнительно размеченный в разработанном инструменте ConAnnotator. Модуль отбора информативных кадров на основе ResNet-34 показал долю верных ответов 0,965 и AUC 0,987. Сеть U-Net с постобработкой условными случайными полями сегментирует сосуды с коэффициентом Дайса 0,84. Области, повлиявшие на решение, визуализируются методом Score-CAM (Score-weighted Class Activation Mapping). Степень стеноза, фракционный резерв кровотока и степень кровотока TIMI поступают из количественной ангиографической оценки или экспертной разметки и интерпретируются нечётко-логическим модулем с базой из 25 продукционных правил. При рандомизации параметров модели и обучении на перемешанных метках (62 пациента) корреляция карт с исходными не превысила 0,29, следовательно, объяснения зависят от параметров модели и обучающей разметки. Интерфейс приложения показан на рисунке 10.

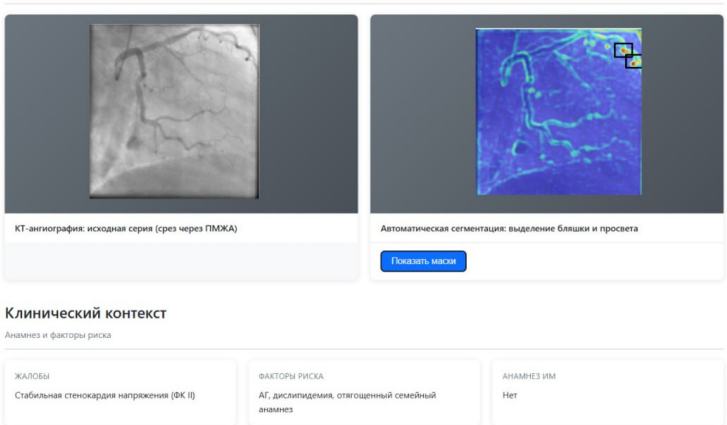


Рисунок 10 – Интерфейс приложения для анализа коронарографических исследований

Третья задача — выявление пневмонии на рентгенограммах грудной клетки. Основным источником данных служил набор Kaggle Chest X-Ray Images. Среди десяти архитектур-кандидатов, сравнённых в режиме автоматизированного машинного обучения, лучший результат показала ResNet-34 с долей верных ответов 0,93, AUC 0,97 и полнотой 0,99 (у базовой Custom CNN доля верных ответов и AUC составили 0,83 и 0,88). Объяснительный контур включает Grad-CAM, LIME, SHAP и метод интегрированных градиентов. Протокол выбора объяснителя проверен в калибровочном эксперименте на 80 рентгенограммах. Все пороговые условия выполнил только Grad-CAM (итоговый критерий 0,642), причём результат его проверки на рандомизацию параметров оказался пограничным. LIME не прошёл пороги по верности и устойчивости, SHAP — по верности, интегрированные градиенты — по верности, устойчивости и проверке на рандомизацию. Текстовое пояснение для врача формирует вербализатор на основе большой языковой модели. Сгенерированный текст в вычислительный контур не возвращается и не влияет ни на диагноз, ни на выбор объяснителя. Архитектура системы приведена на рисунке 11.

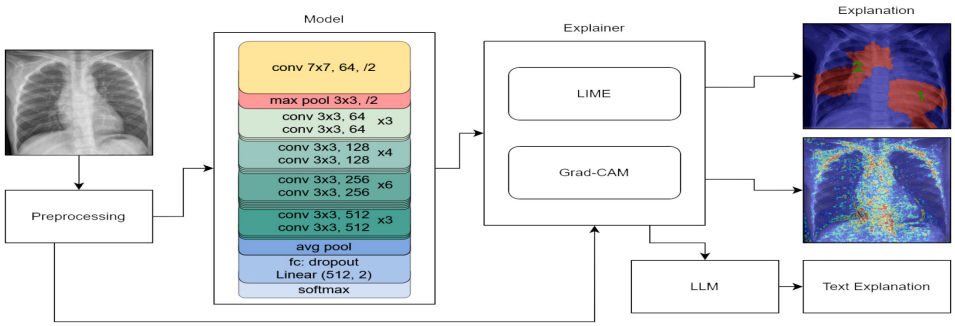


Рисунок 11 – Архитектура системы интерпретируемой диагностики пневмонии

Решение реализовано в виде веб-приложения и перенесено на анализ хронической обструктивной болезни лёгких (программы для ЭВМ № 2025690301 и № 2025690503).

В заключении сформулированы основные результаты диссертационного исследования. Решена научная задача многоуровневого формирования, согласования, объединения и сопоставимой оценки объяснений в системах ИИ с учётом неопределённости. Создана платформа, обеспечивающая воспроизводимый контур их применения в прикладных задачах анализа данных и поддержки принятия решений.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИИ

1. Выполнен системный обзор современного состояния методов объяснимого искусственного интеллекта, позволивший выявить ограничения существующих подходов и обосновать необходимость многоуровневого формирования, согласования и объединения объяснений в гетерогенных интеллектуальных системах.

2. Разработаны методы многоуровневого формирования и согласования объяснений, обеспечивающие взаимосвязь признакового, модельного, интерпретируемого и результирующего уровней представления.

3. Разработаны методы агрегирования, объединения и комплексирования при ансамблировании гетерогенных объяснений, а также формальный аппарат операций над многоуровневыми модельными и постмодельными объяснениями.

4. Разработаны методы учёта и квантификации неопределённости объединённого объяснения, позволяющие рассматривать итоговое объяснение как самостоятельный объект анализа.

5. Разработан протокол автоматизированного воспроизводимого выбора метода объяснения и метрический контур сопоставимой оценки качества объяснений, учитывающий согласованность, устойчивость, воспроизводимость и пригодность для практического применения.

6. Создана и программно реализована платформа, обеспечивающая оркестрацию вычислений, выбор моделей и методов объяснения, формирование объяснений и сопоставимую оценку их качества в воспроизводимом контуре анализа данных. Для её реализации разработаны соответствующие архитектурные, алгоритмические и программные решения.

7. Проведена апробация разработанных методов и платформенных решений на биомедицинских задачах анализа маммографических изображений, диагностики стенозов коронарных артерий при коронарографии и диагностики заболеваний органов грудной клетки.

8. Подтверждена практическая значимость результатов. Разработаны три программы для ЭВМ, результаты внедрены в специализированное программное обеспечение системы поддержки принятия врачебных решений, использованы в деятельности медицинских и научных организаций, при выполнении научно-исследовательских работ и в учебном процессе.

ПУБЛИКАЦИИ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

В изданиях, входящих в международную базу цитирования Scopus:

1. **Trofimov Y.**, Averkin A. Enhancing Results with ANFIS and XAI for Soil Contamination Monitoring and Public Health Risk Assessment // 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). Novosibirsk, Russian Federation, 2024. P. 1120–1123. DOI: 10.1109/PIERE62470.2024.10805065. (Scopus)

2. **Trofimov Y.V.**, Shevchenko A.V., Averkin A.N., Muravyov I.P., Kuznetsov E.M. Concept of Hierarchically Organized Explainable Intelligent Systems: Synthesis of Deep Neural Networks, Fuzzy Logic and Incremental Learning in Medical Diagnostics // 2025 VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). Saint Petersburg, Russian Federation, 2025. P. 14–17. DOI: 10.1109/NeuroNT66873.2025.11049976. (Scopus)

3. Verified Explainability Core: A GD–ANFIS/SHAP Hybrid Architecture for XAI 2.0 / **Y. Trofimov**, A. Lebedev, A. Ilin, A. Averkin // Automatic Documentation and Mathematical Linguistics. – 2025. – Vol. 59, Suppl. 5. – P. S469–S478. – DOI 10.3103/S0005105525701420. (Scopus)

4. Averkin A.N., **Trofimov Y.V.**, Lebedev A.D., Ilin A.S. Hybrid and Hierarchical Explainable AI Based on Kolmogorov-Arnold Networks // Proceedings of the Ninth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’25), Volume 2 / edited by S. Kovalev, I. Kotenko, A. Sukhanov. – Cham : Springer, 2026. – (Lecture Notes in Networks and Systems, vol. 1763). – DOI: 10.1007/978-3-032-13612-1_4. (Scopus)

5. **Trofimov Y.**, Shevchenko A., Trusov I. Causality-Guided Admission Control for Diffusion-Generated Synthetic Medical Image // Proceedings of the 2026 International Russian Smart Industry Conference (SmartIndustryCon). Sochi, Russian Federation, 2026. P. 1093–1098. DOI: 10.1109/SmartIndustryCon68821.2026.11492854. (Scopus)

6. **Trofimov Y.**, Averkin A., Lebedev M., Lebedev A., Ilyin A., Alekseev A. RuFLEX: A Hybrid Deep Neuro-Fuzzy Intelligence Platform for Geospatial Data with Logical Rule Traceability and Counterfactual Explainability // Proceedings of the 2026 VII International Conference on Neural Networks and Neurotechnologies (NeuroNT). Saint Petersburg, Russian Federation, 2026. P. 11–14. DOI: 10.1109/NeuroNT71829.2026.11651423. (Scopus)

7. Kuznetsov E., Averkin A., **Trofimov Y.**, Lopatin M., Kondrashova E. Physio-JEPA for ICU Signals: Temporal Multimodal Explainability and the Temporal Range Coherence Metric for Ventricular Tachycardia Alarm Validation // Proceedings of the 2026 VII International Conference on Neural Networks and Neurotechnologies (NeuroNT). Saint Petersburg, Russian Federation, 2026. P. 28–31. DOI: 10.1109/NeuroNT71829.2026.11650778. (Scopus)

8. Kim T., Averkin A., Alexeev A., Trusov I., **Trofimov Y.** A Method for Generating Synthetic Cellular Data Using Rules, Variational Modeling, and Explainable Artificial Intelligence // 2026 IEEE 27th International Conference of Young Professionals in Electron Devices and Materials (EDM). Novosibirsk, Russian Federation, 2026. P. 1–4. DOI: 10.1109/EDM69524.2026.11632177. (Scopus)

9. **Trofimov Y.**, Averkin A., Lebedev M., Lebedev A., Ilin A. Image Segmentation Algorithm Based on Anisotropic Fuzzy Local Convolution // 2026 IEEE 27th International Conference of Young Professionals in Electron Devices and Materials (EDM). Novosibirsk, Russian Federation, 2026. P. 1–5. DOI: 10.1109/EDM69524.2026.11632120. (Scopus)

10. **Trofimov Y.**, Averkin A., Shevchenko A., Trusov I., Alekseev A., Kondrashova E. Localized SHAP Explainability for Object Detection Based on Parallel Attribution Analysis // Proceedings of the 2026 XXIX International Conference on Soft Computing and Measurements (SCM). Saint Petersburg, Russian Federation, 2026. P. 1–3. DOI: 10.1109/SCM71573.2026.11625907. (Scopus)

В изданиях, рекомендованных ВАК РФ по специальности 1.2.1:

11. Мультимодальная объяснимость для ОРИТ-сигналов (VTaC): метрические и асимптотические результаты / **Ю. В. Трофимов**, А. Н. Аверкин, Е. М. Кузнецов, А. П. Еремеев, А. В. Нечаевский // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2025. – № 4. – С. 43–80. – DOI 10.26456/vtpmk757. (ВАК К2)

12. Временная атрибуция в стохастических диффузионных процессах: математическая формализация, объяснимость и каузальная валидация / **Ю. В. Трофимов**, А. Н. Аверкин, М. А. Лопатин, И. А. Трусов, И. П. Муравьев, А. С. Ильин, А. В. Шевченко // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2026. – № 1. – С. 89–137. – DOI 10.26456/vtpmk773. (ВАК К2)

13. Согласование глобальных и локальных объяснений моделей искусственного интеллекта с помощью регуляризации чувствительности признаков / **Ю. В. Трофимов**, А. В. Шевченко, А. Н. Аверкин, И. П. Муравьев // Программные продукты и системы. – 2026. – Т. 39, № 3. – С. 359–367. – DOI 10.15827/0236-235X.155. (ВАК К1)

В других, рекомендованных ВАК РФ:

14. Исследование сигнальных методов выявления состязательных атак на нейронные модели через объяснительную визуализацию / Д. А. Чеколаев, А. В. Шевченко, А. К. Алексеев, **Ю. В. Трофимов**, А. Н. Аверкин // Вопросы кибербезопасности. – 2025. – № 6(70). – С. 140-146. – DOI 10.21681/2311-3456-2025-6-140-146. – EDN JSORTT. (ВАК К-1)

15. Алгебраические операции над многоуровневыми объяснениями и квантификация их неопределённости / **Ю. В. Трофимов**, А. Н. Аверкин, А. Д. Лебедев [и др.] // Мягкие измерения и вычисления. – 2026. Т. 2–2. – DOI: 10.36871/26189976.2026.02-2.008. (ВАК К-3)

16. Трофимов, Ю. В. Обзор и анализ методов объяснительного искусственного интеллекта для решения задач геоэкологического районирования и медицинской профилактики населения / **Ю. В. Трофимов**, А. Н. Аверкин, Е. Н. Черемисина // Геоинформатика. – 2024. – № 4. – С. 93-118. – DOI 10.47148/1609-364X-2024-4-93-118. – EDN ZXOVIB. (ВАК К-2) (RCSI У3)

17. Аверкин, А. Н. Применение адаптивных нейро-нечетких систем вывода и объяснительного искусственного интеллекта для мониторинга загрязнения почв и оценки рисков для здоровья населения в городе Дубне / А. Н. Аверкин, **Ю. В. Трофимов** // Мягкие измерения и вычисления. – 2024. – Т. 84, № 11-1. – С. 70-90. – DOI 10.36871/2618-9976.2024.11.006. – EDN CSJTNK. (ВАК К-3)

18. **Трофимов Ю. В**. Связь доверенного искусственного интеллекта и XAI 2.0: теория и фреймворки / Ю. В. Трофимов, А. Н. Аверкин // Мягкие измерения и вычисления. – 2025. – Т. 90, № 5. – С. 68-84. – DOI 10.36871/2618-9976.2025.05.006. – EDN XUYADQ. (ВАК К-3)

19. Генерация синтетических данных для обучения модели предсказания степени выживаемости раковых клеток с последующим XAI-анализом / Т. В. Ким, И. А. Булякова, **Ю. В.**

Трофимов, А. Н. Аверкин // Мягкие измерения и вычисления. – 2025. – Т. 93, № 8. – С. 28-48. – DOI 10.36871/2618-9976.2025.08.004. – EDN JGDIUZ. (ВАК К-3)

20. Нечёткий трансформер для мультимодального ИИ: дифференцируемое нечёткое внимание и адаптивные объяснения / **Ю. В. Трофимов**, А. Н. Аверкин, А. С. Ильин [и др.] // Мягкие измерения и вычисления. – 2025. – Т. 96, № 11. – С. 34-55. – DOI 10.36871/2618-9976.2025.11.003. – EDN NBBCRM. (ВАК К-3)

21. Многоуровневая ГИС-архитектура и объяснимый ИИ для территориального районирования при размещении полигонов ТКО: обзор и концепция интеграции / **Ю. В. Трофимов**, И. З. Каманина, А. Н. Аверкин [и др.] // Мягкие измерения и вычисления. – 2025. – Т. 97, № 12. – С. 23-40. – DOI 10.36871/2618-9976.2025.12.003. – EDN ESPEKG. (ВАК К-3)

22. HYBRID-XIRIS: нейро-нечеткая архитектура для объяснительной биометрической идентификации по радужной оболочке глаза / **Ю. В. Трофимов**, И. А. Матвеев, А. Н. Аверкин [и др.] // Мягкие измерения и вычисления. – 2025. – Т. 97, № 12-2. – С. 119-132. – DOI 10.36871/2618-9976.2025.12-2.013. – EDN QBGESK. (ВАК К-3)

23. SCR-анализ сквозной объяснимости в AutoML и AutoXAI и синергетическое обоснование архитектурной интеграции / **Ю. В. Трофимов**, Е. С. Кондрашова, А. Н. Аверкин [и др.] // Мягкие измерения и вычисления. – 2026. – Т. 2. – С. 41-60. – DOI: 10.36871/26189976.2026.01-2.005. (ВАК К-3)

24. Агентная модель сопоставимого выбора объяснителя для моделей машинного обучения / **Ю. В. Трофимов**, С. О. Косарев, А. Н. Аверкин [и др.] // Мягкие измерения и вычисления. – 2026. – Т. 1. – С. 56-78. – DOI: 10.36871/2618-9976.2026.03.005. (ВАК К-3)

25. Архитектурная модель платформы SynergiXAI для воспроизводимого управления жизненным циклом моделей искусственного интеллекта / **Ю. В. Трофимов**, Е. М. Кузнецов, А. Н. Аверкин [и др.] // Мягкие измерения и вычисления. – 2026. – Т. 2. – С. 101-119. – DOI: 10.36871/26189976.2026.03-2.011. (ВАК К-3)

В других изданиях:

26. Автоматизация анализа рентгеновских снимков грудной клетки с использованием методов глубокого обучения и объяснимого искусственного интеллекта / М. И. Беляев, А. Н. Аверкин, **Ю. В. Трофимов** [и др.] // Системный анализ в науке и образовании. – 2025. – № 2. – С. 32-41. – EDN VHJCOS. (РИНЦ)

27. Нечёткие продукционные правила и нейросети глубокого обучения: объяснимый искусственный интеллект 2.0 для диагностики коронарных стенозов / **Ю. В. Трофимов**, В. С. Семашко, И. П. Муравьев [и др.] // Системный анализ в науке и образовании. – 2025. – № 2. – С. 73-82. – EDN ONRAKG. (РИНЦ)

28. Новый плитогенно-нечеткий многослойный перцептрон (PN-MLP) на основе нечёткой, интуиционистской, нейтрософической и плитогенной логик с объяснительным искусственным интеллектом второго поколения (XAI 2.0) для платформы клинического анализа молочной железы / **Ю. В. Трофимов**, И. П. Муравьев, А. Н. Аверкин [и др.] // Нечеткие системы и мягкие вычисления. – 2025. – Т. 20, № 1. – С. 5-35. – DOI 10.26456/fssc132. – EDN CAKQYD. (РИНЦ)

29. Объяснительная интервальная ANFIS Типа 2 с интегралом Шоке и самонастраиваемыми параметрическими T-нормами для прогнозирования и стратификации риска гестационного

сахарного диабета / **Ю. В. Трофимов**, А. В. Шевченко, Е. М. Кузнецов [и др.] // Нечеткие системы и мягкие вычисления. – 2025. – Т. 20, № 2. – С. 43-67. – DOI 10.26456/fssc134. – EDN IVOAWY. (РИНЦ)

30. Концепция иерархически организованных объяснимых интеллектуальных систем: синтез глубоких нейросетей, нечеткой логики и инкрементного обучения в медицинской диагностике / **Ю. В. Трофимов**, А. В. Шевченко, А. Н. Аверкин [и др.] // VI Международная конференция по нейронным сетям и нейротехнологиям (NeuroNT'2025): Материалы конференции, Санкт-Петербург, 04–05 июня 2025 года. – Санкт-Петербург: Санкт-Петербургский государственный электротехнический университет "ЛЭТИ" им. В.И. Ульянова (Ленина), 2025. – С. 3-6. – EDN CJPUMF. (РИНЦ)

31. Использование нечетких множеств для построения гибридного объяснительного искусственного интеллекта нового поколения / **Ю. В. Трофимов**, А. Н. Аверкин, А. К. Алексеев, А. В. Шевченко // XXII национальная конференция по искусственному интеллекту с международным участием (КИИ-2025): Труды конференции, Санкт-Петербург, 06–10 октября 2025 года. – Санкт-Петербург: Санкт-Петербургский Федеральный исследовательский центр РАН, 2025. – С. 97-107. – DOI 10.15622/rcai.2025.040

32. Создание прототипа рекурсивного объяснимого искусственного интеллекта на основе поведенческо-логических моделей / **Ю. В. Трофимов**, А. Н. Аверкин, И. П. Муравьев [и др.] // XXII национальная конференция по искусственному интеллекту с международным участием (КИИ-2025) : Труды конференции, Санкт-Петербург, 06–10 октября 2025 года. – Санкт-Петербург: Санкт-Петербургский Федеральный исследовательский центр РАН, 2025. – С. 108-119. – DOI 10.15622/rcai.2025.041

33. **Трофимов Ю. В.** Мягкие измерения нейронечётких составляющих XAI 2.0: мультиуровневая гибридная архитектура (CNN + ANFIS) / Ю. В. Трофимов, Е. М. Кузнецов, И. П. Муравьев // Наука настоящего и будущего. – 2025. – Т. 1. – С. 44-47. – EDN WTOSIF.

34. **Трофимов Ю. В**. Многоуровневый объяснительный ИИ для диагностики пневмонии в рентген-снимках: от Grad-CAM и LIME до генеративного автоописания с MLOPS-интеграцией / Ю. В. Трофимов, И. П. Муравьев, Е. М. Кузнецов // Наука настоящего и будущего. – 2025. – Т. 1. – С. 60-63. – EDN ORUBXV.

35. Интеграция объяснимого искусственного интеллекта и спутниковых данных в задаче районирования территории / **Ю. В. Трофимов**, И. З. Каманина, А. Н. Аверкин [и др.] // Актуальные проблемы экологии и природопользования: сборник научных трудов XXVI Международной научно-практической конференции. В 3 томах, Москва, 25–27 апреля 2025 года. – Москва: Российский университет дружбы народов (РУДН), 2025. – С. 390-394. – EDN KUYLYY.

36. **Трофимов Ю. В.** Объяснительный интеллект как основа доверия и адаптации знаний для устойчивого развития / Ю. В. Трофимов // Искусственный и естественный интеллект: алгоритмы, мышление и образовательные технологии: материалы XXI международного конгресса с элементами научной школы для молодых ученых: в 2-х томах, Москва, 27–28 марта 2025 года. – Москва: Московский университет им. С.Ю. Витте, 2025. – С. 134-151. – EDN AMOTXS. (РИНЦ)

37. **Trofimov Yu. V**. Hierarchical approach to multimodal Explainable Artificial Intelligence: Incremental integration of explanatory / Yu. V. Trofimov, A. N. Averkin // Scientific research of the SCO countries: synergy and integration: Proceedings of the International Conference, Beijing, 15 января 2025 года. – Beijing: Infiniti, 2025. – P. 130-135. – DOI 10.34660/INF.2025.87.97.124. – EDN CDNNDC.

38. **Trofimov Yu. V**. Zoning based on XAI and knowledge bases: a theoretical and applied approach to managed spatial systems / Yu. V. Trofimov, A. N. Averkin // Science. Education. practice : Proceedings of the International Science Conference, Delhi, 29 января 2025 года. – Delhi: OOO "Инфинити", 2025. – P. 159-167. – DOI 10.34660/INF.2025.79.43.089. – EDN SOHBOR.

39. Кузнецов, Е. М. Объяснимый инкрементный подход в диагностике пневмонии: от свёрточных нейросетей до генеративных текстовых заключений / Е. М. Кузнецов, И. П. Муравьев, **Ю. В. Трофимов** // XIV Конгресс молодых ученых : Тезисы научных работ, Санкт-Петербург, 07–11 апреля 2025 года. – Санкт-Петербург: Национальный исследовательский университет ИТМО, 2025. – С. 815-817. – EDN LPKXXE. (РИНЦ)

40. **Трофимов Ю. В**. Нечёткие продукционные правила и нейросети глубокого обучения: объяснимый искусственный интеллект 2.0 для диагностики коронарных стенозов / Ю. В. Трофимов, В. С. Семашко // XIV Конгресс молодых ученых : Тезисы научных работ, Санкт-Петербург, 07–11 апреля 2025 года. – Санкт-Петербург: Национальный исследовательский университет ИТМО, 2025. – С. 938-939. – EDN TLYRHK. (РИНЦ)

Свидетельства о государственной регистрации программ для ЭВМ

41. **Ю. В. Трофимов** [и др.] свидетельство о государственной регистрации программы для ЭВМ № 2025690301 Российская Федерация. Веб-приложение для интерпретируемой классификации пневмонии лёгких (версия 1.0): заявл. 25.10.2025: опубл. 06.11.2025

42. **Ю. В. Трофимов** [и др.] свидетельство о государственной регистрации программы для ЭВМ № 2025690503 Российская Федерация. Веб-приложение для мультиуровневого и интерпретируемого анализа ХОБЛ (версия 1.0): заявл. 25.10.2025: опубл. 07.11.2025

43. Д. А. Чеколаев, **Ю. В. Трофимов** [и др.] свидетельство о государственной регистрации программы для ЭВМ № 2025690674 Российская Федерация. Интерактивная демонстрационная платформа для изучения принципов объяснимого ИИ через моделирование состязательных атак: заявл. 25.10.2025: опубл. 10.11.2025