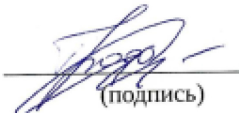
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «МЭИ»

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»

Диссертация на соискание учёной степени
кандидата технических наук

Научный руководитель:
к. ф.-м. н., доцент
Аверкин Алексей Николаевич

Москва – 2026

СОДЕРЖАНИЕ

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ 6

СПИСОК ТЕРМИНОВ 10

ВВЕДЕНИЕ 16

1. СИСТЕМНЫЙ ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ 22

1.1. Место объяснимости в контуре доверенного искусственного интеллекта 25

1.2. Уточнение базовых понятий: прозрачность, интерпретируемость, объяснимость и интерпретация 28

1.3. Нормативно-правовое, этическое и стандартное регулирование систем ИИ и объяснимости 31

1.4. Постмодельные методы объяснения как первый этап развития ОИИ 33

1.5. Ограничения постмодельных подходов 34

1.6. Переход к встроенной и архитектурной объяснимости в ОИИ 2.0 .. 36

1.7. Автоматизированные контуры ОИИ и воспроизводимость объяснений 39

1.8. Агрегирование, ансамблирование, воспроизводимый выбор объяснителя и проблемные зоны современной ОИИ 43

1.9. Выводы по главе и постановка задач диссертационного исследования 48

2. АППАРАТ МНОГОУРОВНЕВОГО ФОРМИРОВАНИЯ, СОГЛАСОВАНИЯ, ОБЪЕДИНЕНИЯ И СОПОСТАВИМОЙ ОЦЕНКИ ОБЪЯСНЕНИЙ В СИСТЕМАХ ИИ 51

2.1. Многоуровневость иерархически организованной интеллектуальной системы 51

2.2. Формализация узлов ориентированного графа объяснимого контура обработки данных 53

2.3. Методы многоуровневого формирования атрибутивных объяснений 57

2.4. Метод межуровневого согласования объяснений 62

2.5. Методы агрегирования, объединения и комплексирования при ансамблировании гетерогенных объяснений 66

2.6. Метод квантификации неопределённости объединённого объяснения 71

2.7. Метрический контур сопоставимой оценки качества объяснений . 74

2.8. Протокол воспроизводимого выбора объяснителя 78

2.9. Метод комплексирования моделей машинного обучения с многоуровневым ансамблированием объяснений 82

2.10. Агентное расширение метода воспроизводимого выбора объяснителя 87

2.11. Выводы по главе 90

3. АРХИТЕКТУРА ГЕТЕРОГЕННОЙ ГИБРИДНОЙ ПЛАТФОРМЫ 92

3.1. Принципы проектирования архитектуры платформы 93

3.2. Концептуальные представления архитектуры платформы 94

3.2.1. Контекстное представление: платформа как среда 94

3.2.2. Компонентное представление: ядро, плагины и реестр XAIRegistry 96

3.2.3. Динамическое представление: жизненный цикл запроса и «гейты» валидации 98

3.3. Архитектура функциональных подсистем 100

3.3.1. Ядро управления 102

3.3.2. Реестр методов и плагинная система 104

3.3.3. Подсистема метрической верификации 107

3.3.4. Контуры контроля SafetyGuard и TrustEvaluator 110

3.3.5. Слой воспроизводимости 114

3.3.6. Локальный интерфейс просмотра отчётов и артефактов 116

3.4. Функциональная организация и маршрутизация 119

3.4.1. Связь AutoML и AutoXAI в маршруте выполнения 122

3.4.2. Формирование объяснений при ансамблировании 122

3.4.3. Воспроизводимый выбор метода объяснения 124

3.4.4. Стратегии маршрутизации и построение плана выполнения. 125

3.5. Архитектура AutoML-ядра..... 126

3.5.1. Многокритериальная постановка и двухэтапная селекция 126

3.5.2. Частные критерии оценки кандидатов и бюджет поиска 128

3.5.3. Нормированные показатели верности объяснения, сложности и вычислительной эффективности..... 128

3.5.4. Нечёткая агрегация критериев выбора 129

3.6. Архитектурные свойства и ограничения..... 129

3.7. Основные выводы по главе 130

4. АПРОБАЦИЯ ГЕТЕРОГЕННОЙ ГИБРИДНОЙ ПЛАТФОРМЫ НА БИОМЕДИЦИНСКИХ ЗАДАЧАХ 131

4.1. Биомедицинская задача: маммология..... 132

4.1.1. Клинико-диагностические предпосылки применения ОИИ в маммологии 132

4.1.2. Клинико-рентгенологические основы маммографической диагностики..... 134

4.1.3. Инфраструктурные ограничения мобильной маммографической диагностики..... 136

4.1.4. Современные подходы к автоматизированному анализу маммографических изображений 137

4.1.5. Постановка задачи..... 139

4.1.6. Практическая реализация..... 141

4.1.7. Основные выводы..... 152

4.2. Биомедицинская задача: ишемическая болезнь сердца 153

4.2.1. Введение 153

4.2.2. Существующие решения 154

4.2.3. Постановка задачи 155

4.2.4. Практическая реализация 155

4.2.5. Основные выводы по результатам 165

4.3. Биомедицинская задача: автоматизированный анализ рентгенограмм грудной клетки при выявлении признаков пневмонии 167

4.3.1. Введение 167

4.3.2. Существующие решения 168

4.3.3. Постановка задачи 169

4.3.4. Введение в предметную область 170

4.3.5. Практическая реализация 171

4.3.6. Основные выводы по результатам 183

ЗАКЛЮЧЕНИЕ 185

СПИСОК ЛИТЕРАТУРЫ 187

ПРИЛОЖЕНИЕ А 210

ПРИЛОЖЕНИЕ Б 214

ПРИЛОЖЕНИЕ В 217

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

ANFIS – *Adaptive Neuro-Fuzzy Inference System*; адаптивная нейро-нечёткая система вывода

AOMO – *AutoML for Model Explanation*; класс решений, в которых построение объяснения включено в контур автоматизированного выбора и проверки

AOPC – *Area Over the Perturbation Curve*; метрика оценки качества объяснения, отражающая изменение качества модели при последовательном возмущении наиболее значимых признаков

AUC – *Area Under the Curve*; площадь под кривой

AutoML – *Automated Machine Learning*; автоматизированное машинное обучение

AutoXAI – *Automated Explainable Artificial Intelligence*; автоматизированный контур выбора, настройки, сопоставления и оценки методов объяснения в составе жизненного цикла модели

BI-RADS – *Breast Imaging-Reporting and Data System*; система стандартизированного описания и категоризации результатов исследований молочной железы

CAD – *Computer-Aided Detection/Diagnosis*; компьютерная поддержка выявления и диагностики

CAM – *Class Activation Mapping*; семейство методов визуализации значимых областей входа, влияющих на решение модели

CLAHE – *Contrast Limited Adaptive Histogram Equalization*; адаптивное выравнивание гистограммы с ограничением контраста

CNN – *Convolutional Neural Network*; свёрточная нейронная сеть

CRF – *Conditional Random Fields*; условные случайные поля

CXR – *Chest X-Ray*; рентгенограмма органов грудной клетки

DARTS – *Differentiable Architecture Search*; дифференцируемый поиск архитектуры нейронной сети

DevOps – *Development and Operations*; практики совместной разработки и эксплуатации программного обеспечения

DICOM – *Digital Imaging and Communications in Medicine*; стандарт хранения и передачи медицинских изображений

ENAS – *Efficient Neural Architecture Search*; эффективный поиск архитектуры нейронной сети

ExplainRunResult – унифицированная структура результата выполнения XAI-конвейера, содержащая полученные объяснения, значения метрических показателей и информацию о параметрах выполнения

ExplainSpec – декларативная спецификация запроса на построение объяснения, определяющая входные данные, целевую задачу, параметры генерации объяснения и уровень детализации результата

FFR – *Fractional Flow Reserve*; фракционный резерв кровотока

Grad-CAM – *Gradient-weighted Class Activation Mapping*; метод визуального объяснения решений нейросетевой модели на основе градиентно-взвешенных карт активации

IEC – *International Electrotechnical Commission*; Международная электротехническая комиссия

IEEE – *Institute of Electrical and Electronics Engineers*; Институт инженеров электротехники и электроники

ISO – *International Organization for Standardization*; Международная организация по стандартизации

KernelSHAP – модельно-независимая реализация метода *SHAP* на основе ядерной аппроксимации

LIME – *Local Interpretable Model-agnostic Explanations*; модельно-независимый метод локального объяснения предсказаний

LLM – *Large Language Model*; большая языковая модель, обученная на больших массивах текста для понимания и генерации естественного языка

MCP – *Model Context Protocol*; протокол, позволяющий языковым моделям подключаться к локальным файлам, базам данных и инструментам разработчика

MLOps – *Machine Learning Operations*; практики эксплуатации и сопровождения моделей машинного обучения

NAS – *Neural Architecture Search*; поиск архитектуры нейронной сети

PACS – *Picture Archiving and Communication System*; система архивирования и передачи медицинских изображений

PGMI – *Perfect, Good, Moderate, Inadequate*; шкала оценки качества укладки при маммографии

RIS – *Radiology Information System*; радиологическая информационная система

RuntimeConfig – конфигурация среды выполнения, фиксирующая вычислительные ресурсы, параметры сохранения артефактов и сведения об окружении, необходимые для обеспечения воспроизводимости эксперимента

self-attention – механизм самовнимания

SHAP – *SHapley Additive exPlanations*; метод объяснения предсказаний на основе значений Шепли

SMAC – *Sequential Model-based Algorithm Configuration*; последовательная модельно-ориентированная настройка алгоритмов

TIMI – *Thrombolysis in Myocardial Infarction*; шкала оценки коронарного кровотока

TreeSHAP – специализированная реализация метода *SHAP* для деревообразных моделей

XAI – *Explainable Artificial Intelligence*; объяснимый искусственный интеллект

XAIRegistry – реестр методов объяснения, обеспечивающий регистрацию, проверку применимости и создание экземпляров объяснителей, используемых в рамках XAI-конвейера

ВАК – Высшая аттестационная комиссия при Министерстве науки и высшего образования Российской Федерации

ГОСТ Р – национальный стандарт Российской Федерации

ГП – городская поликлиника

ИБС – ишемическая болезнь сердца

ИИ – искусственный интеллект

ЛПУ – лечебно-профилактическое учреждение

л/у – лимфатический узел

МО – медицинская организация

НИР – научно-исследовательская работа

ОИИ 1.0 – этап развития объяснимого искусственного интеллекта, основанный преимущественно на постмодельных методах объяснения

ОИИ 2.0 – этап развития объяснимого искусственного интеллекта, ориентированный на встроенную, многоуровневую, воспроизводимую и согласованную объяснимость

ОИЯИ – Объединённый институт ядерных исследований

РМЖ – рак молочной железы

РНФ – Российский научный фонд

СППР – система поддержки принятия решений

ТКО – твёрдые коммунальные отходы

УЗИ – ультразвуковое исследование

ХОБЛ – хроническая обструктивная болезнь лёгких

ЭВМ – электронная вычислительная машина

СПИСОК ТЕРМИНОВ

Агрегирование объяснений – режим совместной обработки однородных объяснений, при котором частные объяснения приводятся к общему пространству представления и единой шкале и сводятся в одно итоговое представление (например, взвешенным усреднением или по правилам коллективного ранжирования).

Адаптер модели – интерфейсный слой между ядром платформы и конкретной моделью машинного обучения, унифицирующий работу с моделями независимо от библиотеки и архитектуры.

Ансамблирование объяснений – совместная обработка нескольких частных объяснений, полученных разными методами, моделями, запусками или для разных модальностей, с целью построения итогового объяснения; включает режимы агрегирования, объединения и комплексирования.

Артефакты эксперимента – обученные модели, карты объяснений, вычисленные метрики, конфигурации запуска и отчёты, полученные при выполнении XAI-конвейера.

Архитектурная объяснимость – форма объяснимости, при которой связь между результатом и механизмом его получения поддерживается на уровне архитектуры системы.

Верность объяснения (faithfulness) – степень соответствия объяснения реальному поведению модели; для карт значимости оценивается по изменению предсказания при последовательном удалении или вставке наиболее значимых признаков.

Внутренняя согласованность объяснения – свойство многоуровневого объяснения, при котором его уровни подтверждают друг друга и образуют единую причинно и содержательно связанную структуру.

Внутриметодная неопределённость – составляющая неопределённости, вызванная разбросом результатов повторных запусков одного и того же метода.

Воспроизводимость объяснения – возможность повторного получения сопоставимого объяснения при фиксированных условиях построения модели, данных и метода.

Воспроизводимый выбор объяснителя – формализованная процедура выбора метода объяснения по единым критериям качества, устойчивости, неопределённости и вычислительной стоимости с учётом контекста задачи, выполняемая в фиксированном протоколе сравнения допустимых кандидатов и потому допускающая повторение.

Встроенная объяснимость – объяснимость, обеспечиваемая архитектурой модели и непосредственно связанная с механизмом её вывода.

Вычислительная стоимость объяснения – время и вычислительные ресурсы, затрачиваемые на построение объяснения.

Гетерогенные объяснения – объяснения, различающиеся по способу построения, форме представления, уровню анализа или модальности данных.

Глобальная объяснимость – характеристика объяснения, раскрывающего общие закономерности поведения модели в целом.

Граф объяснения – ориентированный граф, вершины которого соответствуют уровням объяснения, а рёбра отражают допустимые переходы между ними и их внутреннюю связность.

Доверенный искусственный интеллект – подход к построению и применению систем ИИ, при котором значимыми являются не только результативность, но и проверяемость, воспроизводимость, устойчивость, управляемость рисков и обоснованность выводов.

Интегральная неопределённость объединённого объяснения – характеристика границ доверия к объединённому объяснению, складывающаяся из межметодной, внутриметодной и межмодельной неопределённости; в мультимодальных системах межмодальная неопределённость оценивается отдельно.

Интегральная оценка качества объяснения – обобщённый показатель пригодности объяснения, получаемый свёрткой частных метрик верности, устойчивости, полноты, простоты и вычислительной стоимости.

Интегральный индекс доверия (trust score) – агрегированная оценка пригодности объяснения, построенная по показателям верности, устойчивости и согласованности, результатам sanity-проверки и визуальной пригодности.

Интерпретация – форма, в которой объяснение представлено пользователю, эксперту или исследователю.

Интерпретируемая модель – модель, структуру и логику работы которой специалист может содержательно проанализировать, не прибегая к одним лишь внешним объясняющим процедурам.

Интерпретируемость – свойство модели, показывающее, насколько обозрима её внутренняя структура и можно ли проследить связь между устройством модели и её выводом.

Интерпретируемый уровень объяснения – уровень, на котором внутренние зависимости модели переводятся в правила, концепты, прототипы, нечёткие термы и другие конструкции, соотносимые с предметной областью.

Квантификация неопределённости объяснения – процедура, которая выявляет и количественно выражает неопределённость, нестабильность или вариативность итогового объяснения.

Комплексирование объяснений – режим совместной обработки, при котором итоговое представление сохраняет разнородную структуру входящих объяснений и не сводит их к одному носителю.

Контекст применения – кортеж параметров задачи, учитываемых при выборе объяснителя: тип данных и модальность, класс модели, вычислительный бюджет, прикладные ограничения и требования надёжности, а также цель объяснения.

Локальная объяснимость – объяснимость на уровне отдельного решения модели для конкретного объекта или случая.

Межметодная неопределённость – составляющая неопределённости, вызванная расхождением результатов разных объяснителей.

Межмодальная когерентность – степень согласованности объяснительных сигналов, полученных из разных модальностей данных в мультимодальной системе.

Межмодальная неопределённость – составляющая неопределённости, вызванная конфликтом объяснений, полученных по разным модальностям данных; вычисляется как дополнение межмодальной когерентности до единицы.

Межмодельная неопределённость – составляющая неопределённости, вызванная различием объяснений, построенных для нескольких моделей сопоставимого качества.

Межуровневое расхождение – мера несоответствия между объяснительными объектами двух уровней, которая вычисляется после их приведения к сопоставимой форме и показывает, насколько уровни расходятся по содержанию, структуре или количественным оценкам.

Метрический контур оценки качества объяснений – набор формальных критериев, протоколов проверки и правил сравнения, при котором качество объяснений оценивается сопоставимо и воспроизводимо.

Метрический протокол – процедура количественной оценки качества объяснений. Она выполняется после их генерации, не зависит от объясняющего метода и включает фиксированный набор проверок и метрик.

Многоуровневое объяснение – объяснение, в котором сохранена связь между уровнями представления результата, от признаков и внутренних представлений модели до концептов и итогового вывода.

Многоуровневое представление объяснения – запись объяснения кортежем взаимосвязанных компонентов признакового, модельного, интерпретируемого и результирующего уровней.

Многоуровневость интеллектуальной системы – характеристика организации сложной интеллектуальной системы, при которой путь от входных данных к итоговому выводу проходит через несколько взаимосвязанных уровней представления, обработки и интерпретации.

Модельный уровень объяснения – уровень, на котором описываются внутренние представления модели, активации, латентные переменные, карты внимания и иные объекты, отражающие внутреннюю механику вывода.

Мультимодальное объяснение – объяснение, формируемое на основе нескольких модальностей данных и отражающее их согласованный вклад в итоговый вывод.

Объединение объяснений – режим совместной обработки, при котором итоговое объяснение строится поуровнево из нескольких частных объяснений с сохранением метаданных об их происхождении, весе и роли.

Объяснение – результат процедуры раскрытия оснований вывода модели в форме, пригодной для анализа и проверки.

Объяснимость – свойство интеллектуальной системы представлять основания вывода в понятной, проверяемой и практически значимой форме.

Объяснитель – метод, алгоритм или программный компонент, предназначенный для формирования объяснения вывода модели.

Объяснительный контур – процедуры, средства и критерии, с помощью которых объяснения строятся, представляются, проверяются и используются на протяжении жизненного цикла системы ИИ.

Оператор переноса – отображение, которое переводит объяснительный объект одного уровня в форму, сопоставимую с объектами другого уровня, сохраняя содержательную корректность перехода.

Оператор согласования уровней объяснения – оператор, переводящий исходное многоуровневое объяснение в согласованное. Для этого шкалы уровней нормализуются, области определения унифицируются, атрибуты синхронизируются, а перенос информации между уровнями подчиняется заданным правилам.

Плагин объяснения (*XAI*-плагин) – программный модуль с реализацией конкретного метода объяснения, подключаемый к платформе через стандартизованный интерфейс.

Полнота объяснения – степень, в которой объяснение охватывает значимые факторы, определяющие поведение модели.

Постмодельное объяснение (постмодельная интерпретация) – подход, при котором объяснение строится по уже полученному результату модели и в её внутренний механизм вывода не входит.

Признаковый уровень объяснения – уровень, на котором описывается вклад входных данных в итоговое решение: локальные атрибуции, ранги и карты важности признаков.

Прозрачность – свойство интеллектуальной системы, состоящее в открытости её исходных предпосылок, данных, процедур обучения, обновления и контроля, а также условий применения.

Простота объяснения – степень, в которой итоговая форма объяснения обозрима, компактна и понятна пользователю.

Результирующий уровень объяснения – уровень, на котором согласованное объяснение принимает форму, пригодную для предъявления человеку, экспертной системе или внешнему программному модулю.

Согласование объяснений – установление содержательной и структурной связи между несколькими объяснениями или уровнями объяснения.

Стратегия маршрутизации – механизм выбора метода объяснения или конфигурации конвейера на основе правил, эвристик или процедур оптимизационного поиска.

Устойчивое ядро объяснения – совокупность компонентов объяснения, которые воспроизводятся при использовании разных методов, запусков или моделей и потому могут рассматриваться как наиболее надёжная часть итоговой интерпретации.

Устойчивость объяснения (stability) – степень сохранения структуры и содержания объяснения при малых реалистичных изменениях входа, параметров метода или условий запуска.

Чёрный ящик – модель или система, внутренний механизм вывода которой недоступен для непосредственного анализа либо практически не поддаётся содержательной интерпретации.

ExplainPlan – план выполнения *XAI*-конвейера, представляющий собой упорядоченную последовательность операций обработки запроса, включая выбор методов объяснения и протоколов их проверки.

Provenance – метаданные об условиях выполнения эксперимента, в том числе о версиях библиотек, параметрах среды и конфигурации запуска.

SafetyGuard – подсистема предварительной проверки применимости метода объяснения. До запуска она сверяет выбранный алгоритм с типом модели, структурой данных и условиями выполнения.

Sanity-проверка – проверка чувствительности метода объяснения к рандомизации параметров модели. Она позволяет выявить псевдообъяснения, не зависящие от структуры модели.

TrustEvaluator – модуль, который сводит метрические показатели качества объяснения и сигналы неопределённости модели в интегральный индекс доверия и выдаёт диагностические предупреждения.

XAI-конвейер – формализованная последовательность этапов генерации объяснений, включающая подготовку данных, получение предсказания модели, построение объяснения, метрическую проверку его качества и сохранение артефактов.

ВВЕДЕНИЕ

Актуальность темы. Модели искусственного интеллекта всё шире применяются в задачах, где цена ошибки высока. Высокой точности в таких задачах мало. Нужно проследить, как получено решение, оценить его устойчивость и меру достоверности. Объяснимость становится здесь необходимым условием надёжного, воспроизводимого и практически значимого применения ИИ, и 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» [1] и Указом Президента Российской Федерации от 15.02.2024 № 124 [2]. 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. Ошибка здесь обходится особенно дорого, 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 таких методах растёт по мере того, как технологии ИИ входят в медицинскую деятельность. Растёт она и вместе с цифровой трансформацией 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 [3, 4].

С усложнением интеллектуальных систем одного уровня интерпретации перестаёт хватать. Практически значимое объяснение должно связывать признаки, внутренние представления модели и содержательные выводы в одну согласованную картину. Отсюда 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.

Построение систем поддержки принятия решений, интеллектуальных систем и методов 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. На этих исследованиях выросли современные направления ОИИ, интерпретируемых интеллектуальных систем и анализа данных. Ими опираются и модели, алгоритмы и программные решения для поддержки принятия решений.

Результаты исследования востребованы на практике. Они используются в специализированном программном обеспечении системы поддержки принятия врачебных решений ООО «МАКАО ИТ», в ГБУЗ Московской области «Дубненская больница», в Лаборатории информационных технологий им. М.Г. Мещерякова ОИЯИ, в Научно-исследовательском центре искусственного интеллекта и в учебном процессе Университета «Дубна». Они применялись и при выполнении НИР по государственному заданию Минобрнауки РФ № 124112200072-2 и по грантам РНФ № 22-71-10112 и № 22-17-00045.

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.

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:

1. 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;
2. 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;
3. 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;
4. метод квантификации неопределённости объединённого объяснения, учитывающий расхождения между различными методами объяснения, повторными запусками и моделями, а для мультимодальных систем также согласованность объяснений различных модальностей;
5. 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 заключается в развитии положений объяснимого искусственного интеллекта о многоуровневом иерархическом формировании объ-

яснений, их агрегировании и объединении с учётом неопределённости, а также в построении сопоставимых процедур оценки качества объяснений, учитывающих их устойчивость и воспроизводимость. 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. Результаты работы внедрены при создании специализированного программного обеспечения системы поддержки принятия врачебных решений, входящего в программный продукт ООО «МАКАО ИТ». По результатам исследования получены три свидетельства о государственной регистрации программ для ЭВМ.

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, корректность работы разработанных программных решений.

Реализация результатов работы. По результатам диссертационного исследования разработаны три программы для ЭВМ, на которые получены свидетельства о государственной регистрации (см. приложение В). Разработанные методы, алгоритмы и программные средства внедрены при создании специализированного программного обеспечения системы поддержки принятия врачебных решений в составе программного продукта ООО «МАКАО ИТ», а также использованы в деятельности ГБУЗ Московской области «Дубненская больница», Лаборатории информационных технологий им. М. Г. Мещерякова ОИЯИ, научно-исследовательского центра ИИ и в

учебном процессе Университета «Дубна», что подтверждается актами внедрения, 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), «Научное обоснование безопасного захоронения твёрдых коммунальных отходов в геологической среде» (грант РНФ, проект № 22-17-00045).

Апробация результатов. Основные положения диссертационной работы, разработанные модели, методы, алгоритмы и результаты практического использования, представленные в диссертации, докладывались и обсуждались на научных семинарах в Национальном исследовательском университете «МЭИ» (2026 год) и Государственном университете «Дубна» (2024-2026), а также представлялись и докладывались автором 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), в 10 статьях в изданиях, индексированных в международных базах *Web of Science* и *Scopus*. Получены 3 свидетельства о государственной регистрации программы для ЭВМ (см. приложение В).

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.

1. СИСТЕМНЫЙ ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

Развитие методов ИИ привело к существенному изменению требований к вычислительным системам анализа данных [5, 6]. Если на ранних этапах развития машинного обучения основным критерием эффективности модели выступала точность предсказания, то в современных прикладных областях этого критерия оказывается недостаточно [7, 8]. По мере расширения применения интеллектуальных систем в медицине, промышленности, финансовой аналитике и государственном управлении всё большее значение приобретает вопрос о том, каким образом формируется результат модели и насколько обоснованным является её вывод [9–11]. Иначе стали понимать и качество интеллектуальной системы [12, 13]. Высокая метрика больше не считается исчерпывающим доказательством надёжности, если механизм формирования решения непрозрачен для исследователя и пользователя [14, 15]. Когда от ошибки модели зависит клиническое решение, экономическая оценка или управленческая процедура, скрытый механизм вывода становится источником неопределённости и потенциального риска [16, 17]. Объяснимость стали рассматривать как одно из условий ответственного применения ИИ [12, 18].

Из этих запросов выросло отдельное научное направление — объяснимый искусственный интеллект (ОИИ). Его задача — разрабатывать методы, которые раскрывают внутреннюю логику работы интеллектуальных моделей и представляют её в форме, пригодной для анализа и проверки [16, 19]. Важны здесь прозрачность исходных предпосылок, воспроизводимость процедур, устойчивость модели к изменениям данных и возможность содержательно проверить само объяснение [20, 21]. В современной литературе объяснимость относят к ключевым свойствам доверенного искусственного интеллекта, наряду с безопасностью, устойчивостью, справедливостью и воспроизводимостью вычислительных процедур [18, 22, 23].

Сама задача объяснить результат человеку возникла задолго до ОИИ. В отечественной школе теории принятия решений О.И. Ларичев ввёл термин «система поддержки принятия решений». В рамках вербального анализа решений он рассматривал объяснение выбора лицу, принимающему решение, на привычном для этого лица языке качественных оценок [24, 25].

Для исследования данной области необходим системный подход, позволяющий рассматривать объяснение не как отдельный аналитический приём, а как элемент более широкой архитектуры интеллектуальной системы. В рамках системного анализа сложная информационная система рассматривается как совокупность взаимосвязанных компонентов, взаимодействие которых определяет её функциональные свойства и поведение в процессе эксплуатации [26–28]. Представление системы в виде структурированного набора элементов является частью базовых подходов теории систем [29].

Объяснимую интеллектуальную систему в таком подходе удобно представить структурным кортежем:

$$S_{XAI} = \langle D, C, T, X, Y, M, E, A, V \rangle, \tag{1.1}$$

где D — множество данных, используемых для обучения и функционирования системы;

C — контекст применения системы;

T — оператор преобразования исходных данных и контекста применения во входное представление;

X — пространство входных объектов;

Y — пространство выходных решений;

M — модель машинного обучения;

E — механизм формирования объяснения результата работы модели;

A — механизм агрегирования объяснений;

V — механизм валидации и оценки качества объяснений.

Входное представление системы формируется преобразованием:

$$T: D \times C \to X, \tag{1.2}$$

которое с учётом контекста применения C переводит исходные данные D в пространство входных объектов X. Сюда входят подготовка и структурирование данных и их представление в форме, с которой может работать модель.

Модель машинного обучения задаётся отображением:

$$M : X \to Y, \tag{1.3}$$

где каждому объекту из пространства входных объектов X сопоставляется результат в пространстве решений Y.

Механизм формирования объяснения представлен отображением:

$$E(x) = E(M, x, B), \tag{1.4}$$

где $E(x) \in \mathcal{E}$ — объяснение результата модели для входного объекта $x \in X$, E — пространство объяснений, а $B \subseteq X$ — опорное множество входных объектов, на которое опирается объясняющий метод.

Если объяснений несколько, их можно агрегировать:

$$E^*(x) = A\left(E^{(1)}(x), E^{(2)}(x), \dots, E^{(k)}(x)\right), \tag{1.5}$$

где $E^{(1)}(x), \ldots, E^{(k)}(x)$ — частные объяснения, полученные различными методами или на различных уровнях представления, а $E^{*}(x)$ — итоговое объяснение, полученное их совместной обработкой.

Качество сформированного объяснения оценивается функцией:

$$q = V(E^{*}, X_0), \tag{1.6}$$

где q — совокупность метрик качества объяснения, в том числе устойчивости, согласованности и воспроизводимости, а $X_0 \subseteq X$ — набор входных объектов, на которых выполняется оценка.

Графическая интерпретация данной модели представлена на рисунке 1.1.

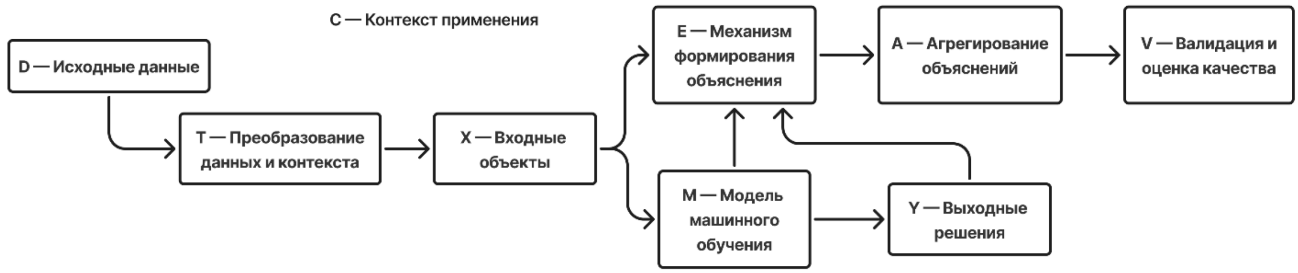


Рисунок 1.1 – Структурная модель объяснимой интеллектуальной системы

Объяснимость в этой модели оказывается свойством, которое возникает из взаимодействия данных, модели, механизмов формирования объяснений и процедур их оценки. Модель задаёт и основу для дальнейшего анализа: существующие подходы к объяснению решений интеллектуальных моделей нужно рассматривать относительно её компонентов, а понятийный аппарат и методологические основания ОИИ — уточнять в тех же терминах.

История ОИИ, развитие понятийного аппарата, нормативные требования к интерпретируемости моделей и эволюция методов объяснения образуют взаимосвязанную линию анализа предметной области [17, 22, 30, 31]. В связи с этим далее последовательно рассматриваются: становление предметной области ОИИ, развитие её базовых понятий и основные направления эволюции методов объяснения результатов моделей машинного обучения. Такой порядок рассмотрения позволяет выявить ограничения существующих решений и подойти к постановке нерешённых задач диссертационного исследования.

1.1. Место объяснимости в контуре доверенного искусственного интеллекта

Обзор предметной области исходит из того, что объяснимость входит в более широкий контур доверенного искусственного интеллекта и к набору методов объяснения не сводится. Так в одну линию анализа связываются развитие моделей машинного обучения, требования к проверяемости результата и необходимость последующего контроля объяснений.

Сама область возникла потому, что изменились технологии ИИ. В экспертных, символических и продукционных системах путь вывода, как правило, оставался обозримым, и объяснение было встроено в структуру рассуждения [32, 33]. Со статистическим обучением по данным эта связь стала ослабевать. Модель начала выявлять зависимости, которые трудно выразить в явном виде. Переход от входа к решению перестал быть содержательно прозрачным и для исследователя, и для пользователя.

В отечественной науке эта проблематика развивалась в работах по кибернетике, системам, основанным на знаниях, и интегрированным интеллектуальным системам. Для настоящего исследования важны результаты по формализации знаний, организации механизмов вывода и программным средствам построения интеллектуальных систем, полученные в работах Л.Т. Кузина, Б.Н. Онукого и Г.В. Рыбиной. Интеллектуальная система рассматривалась в них как программно реализуемое целое, в котором объединены представление знаний, механизм вывода и инструментальные средства разработки.

С ростом предсказательных возможностей моделей изменились и требования к их оценке [9, 10]. Точность на тестовых выборках по-прежнему важна. Но в прикладных сценариях она не говорит, за счёт каких факторов получено решение, насколько оно согласуется с предметной логикой задачи и в каких границах ему можно доверять [9, 11]. Острее всего это ощущается там, где по результату модели ставят диагноз, оценивают риск или выбирают дальнейшие действия. В таких задачах предметно значимую закономерность нужно отличать от статистической регулярности, которая держится только на конкретном корпусе данных.

Так в структуре ОИИ появилось требование содержательной оценки результата [34, 35]. Объяснение стали понимать как способ раскрыть основание вывода в форме, пригодной для анализа, проверки и включения в процедуру принятия решения. Внешним комментарием к готовому ответу его больше не считают. Оценивать приходится уже всю связку «модель – результат – объяснение – процедура проверки».

Объяснимость вошла и в контур доверенного искусственного интеллекта [18, 21, 36]. Модели усложнялись, цена ошибки росла, и объяснимость стали рассматривать вместе с прозрачностью, безопасностью, приватностью, справедливостью и воспроизводимостью вычислительных

процедур. Объяснение берёт на себя и интерпретационную, и контрольную функцию. Через него качество модели связывается с возможностью аудита результата и оценкой границ его применимости.

Место объяснимости среди свойств доверенного ИИ показано на рисунке 1.2 [18, 22, 23, 37]. На схеме она предстаёт одним из компонентов широкой системы требований к надёжному и ответственному использованию интеллектуальных технологий.

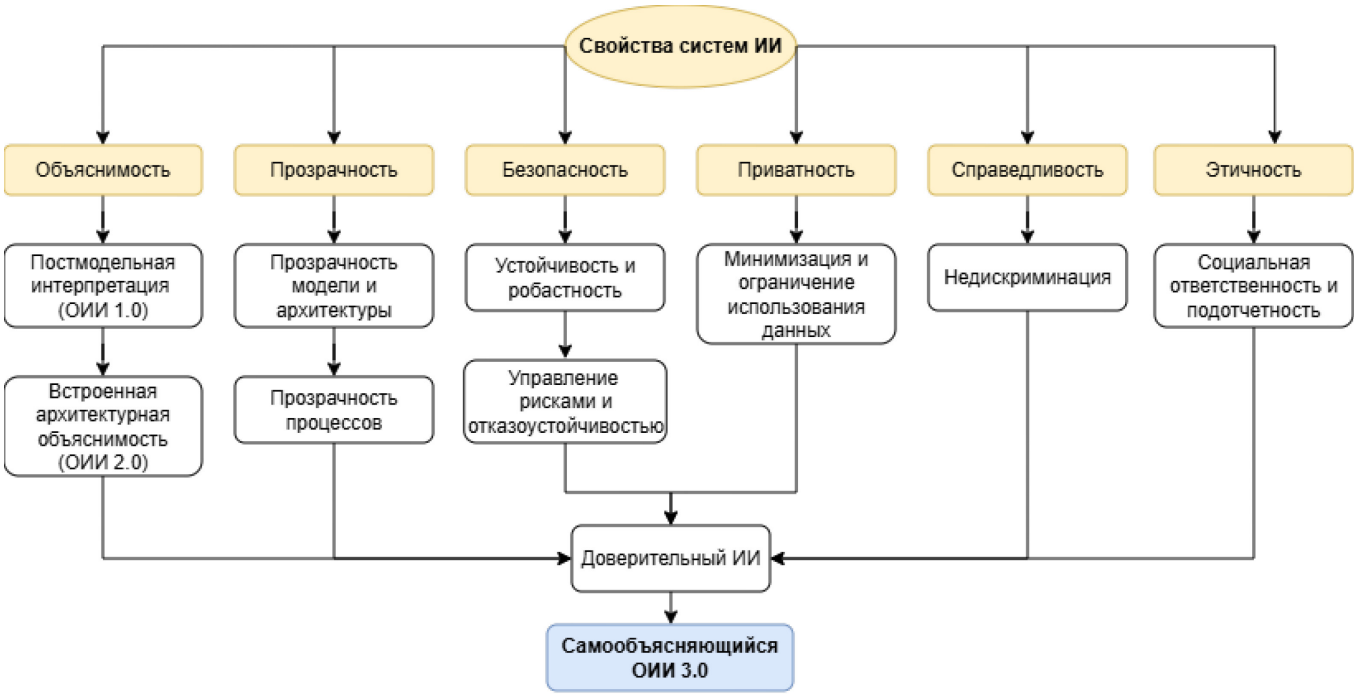


Рисунок 1.2 – Структура свойств доверенного искусственного интеллекта

Развитие ОИИ не было линейным и с самого начала не сводилось к одному классу методов [9, 12]. Одни направления занимались интерпретацией уже обученных моделей, другие — построением интерпретируемых архитектур, третьи уточняли само понятие объяснения и критерии его качества. Постепенно в области выделились четыре линии развития. Это постмодельная интерпретация, интерпретируемые модели, встроенная архитектурная объяснимость и подходы, нацеленные на воспроизводимость, сопоставимость и проверяемость объяснений [13, 19].

Карта этих линий представлена на рисунке 1.3 [14, 19]. По ней видно, как исследовательский фокус смещается от локального объяснения отдельного решения к согласованию, проверке и оценке объяснений на протяжении жизненного цикла модели. Моделирование, построение объяснения, оценка качества, хранение артефактов эксперимента и контроль надёжности всё теснее связываются между собой и становятся компонентами единой интеллектуальной системы.



Рисунок 1.3 – Карта системного анализа предметной области ОИИ

Объяснимость оказывается, таким образом, развивающимся свойством интеллектуальной системы [7, 10–12, 19, 35]. Сравнительная характеристика основных этапов развития ОИИ приведена в таблице 1.1.

Таблица 1.1 – Сравнительная характеристика основных этапов развития ОИИ

Подход / этап развития	**Глубина интерпретации**	**Воспроизводимость**	**Учёт неопределённости**	**Основное ограничение**
Постмодельная интерпретация (ОИИ 1.0)	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 (ОИИ 2.0)	Связывает результат и процесс вывода в едином контуре	Повышенная, так как включена в архитектуру модели	Учитывается в составе объяснения	Трудности согласования различных уровней объяснения

Подход / этап развития	Глубина интерпретации	Воспроизводимость	Учёт неопределённости	Основное ограничение
Измеримый / доверенный ИИ	Раскрывает результат и границы доверия к нему	Высокая, опирается на процедуры контроля	Явный элемент оценки	Требует сопоставимой оценки качества и воспроизводимого выбора метода объяснения

1.2. Уточнение базовых понятий: прозрачность, интерпретируемость, объяснимость и интерпретация

Терминологическая неоднородность в области ОИИ обусловлена тем, что развитие методов длительное время опережало формирование общего понятийного аппарата [38, 39]. В ряде работ прозрачность, интерпретируемость и объяснимость используются как близкие по смыслу характеристики понятности модели, тогда как в других исследованиях между ними проводится принципиальное различие, от которого зависят способ анализа, критерии оценки и ожидаемый результат [8, 40]. В этих условиях уточнение базовых понятий представляет собой необходимое условие последовательного анализа предметной области.

В рамках данного исследования прозрачность рассматривается как характеристика системы в целом и её жизненного цикла [8, 12]. Она относится к степени открытости исходных предпосылок, данных, процедур обучения, модификации и контроля модели [23, 41]. Прозрачность создаёт условия для прослеживаемости, аудита и проверки границ применимости результатов [42, 43]. Следовательно, прозрачность задаёт условия, при которых интеллектуальная система может быть исследована и проверена как технический объект.

С интерпретируемостью связан вопрос об устройстве самой модели [13, 15]. Можно ли проследить, как из структуры модели получается её вывод? Когда внутренние зависимости обозримы, исследователь анализирует механизм решения, а не только его результат [7, 44]. Интерпретируемыми обычно называют модели, ход вывода которых специалист может понять по их структуре.

Объяснимость отвечает на другой вопрос — на чём основан конкретный вывод [9, 16]. Речь идёт о причинах, факторах и зависимостях, которые привели к результату, и о том, можно ли сделать их доступными для анализа и проверки. Внешняя форма, в которой показан результат,

здесь вторична. Объяснимой систему делает способность порождать объяснения, содержательно связанные с её решениями.

Интерпретация замыкает эту цепочку и касается того, как объяснение предъявлено человеку [9, 45]. Форма предъявления подбирается под задачу, контекст применения и подготовку пользователя. Один и тот же объяснительный материал разные аудитории и предметные области прочитают по-разному [43, 46]. Интерпретация не совпадает ни с прозрачностью, ни с интерпретируемостью, ни с объяснимостью. Она описывает предъявление и восприятие объяснения, которое к этому моменту уже сформировано. Соотношение понятий показано на рисунке 1.4 [38, 39].



Рисунок 1.4 – Концептуальная модель понимания поведения ИИ

Чтобы избежать терминологической путаницы, базовые понятия исследования сведены в таблицу 1.2 [12, 38]. Для каждого указаны уровень рассмотрения, основной смысл и роль в дальнейшем анализе.

Таблица 1.2 – Разведение базовых понятий, используемых в исследовании

Понятие	**Уровень рассмотрения**	**Основной смысл**	**Функция в исследовании**
Прозрачность	Система и её жизненный цикл	Открытость предпосылок, данных и процедур функционирования системы	Формирует условия для прослеживаемости, аудита и контроля границ применимости
Интерпретируемость	Архитектура и структура модели	Понятность внутреннего устройства модели и хода вывода	Позволяет анализировать механизм получения результата
Объяснимость	Основание результата	Возможность раскрытия причин и факторов, определивших вывод	Характеризует способность системы порождать объяснение как раскрытие основания решения
Интерпретация	Пользовательское представление	Способ предъявления объяснения человеку в зависимости от контекста	Обеспечивает восприятие и практическое использование объяснения

Разграничение нужно и методологически [19, 35, 47]. Объяснимость бывает локальной, когда раскрываются основания отдельного решения, и глобальной, когда выявляются общие закономерности поведения модели. По месту возникновения она бывает встроенной, заложенной в архитектуру, и постмодельной, которая строится после получения вывода. Одна и та же система может в итоге быть прозрачной лишь отчасти, интерпретируемой лишь в части компонентов и при этом давать объяснения разной глубины и проверяемости.

Отсюда следует ещё одно различие, между объяснением и объяснимостью [19, 35]. Объяснение — это конкретный результат построения интерпретации. Объяснимость означает, что система порождает такие результаты устойчиво и в форме, допускающей анализ и проверку. Оценивать поэтому приходится и содержательность отдельного объяснения, и организацию всего контура, в котором оно формируется. Наконец, объяснение ещё не даёт основания для доверия [41]. Оно раскрывает причины решения, но само по себе не гарантирует, что решение приемлемо в прикладной среде. Доверие появляется, когда к объяснению добавляются прослеживаемость данных, устойчивость модели, контроль качества и понятная процедура проверки. Дальше объяснения рассматриваются в двух качествах. Это отдельные результаты интерпретации и одновременно элементы проверяемого контура, в котором результат формируется, оценивается и предъявляется.

1.3. Нормативно-правовое, этическое и стандартное регулирование систем ИИ и объяснимости

Для объяснимого искусственного интеллекта нормативная и стандартная рамка становится принципиальной тогда, когда интеллектуальные системы покидают исследовательскую среду и начинают работать в регулируемых сценариях [23, 37]. Результат модели перестаёт быть изолированным вычислительным предсказанием. Он входит в процедуру принятия решения, с которой связаны риски, ответственность и последующая проверка [22, 48]. Одного качества предсказания в такой процедуре мало. Учитывать приходится происхождение данных, условия обучения модели и порядок её обновления, устойчивость результатов и возможность восстановить основания принятого решения [23, 37]. Объяснимость получает здесь методологическое и прикладное значение. Она связывает результат интеллектуальной системы с требованиями контролируемости, прослеживаемости и доверия к его использованию [31, 37].

У этой связи есть и другая сторона. Решение, принятое с участием модели, может быть оспорено, и тогда объяснение становится доказательным материалом. Убедительности для этого недостаточно. Объяснение должно сопоставляться с данными, версией модели и условиями запуска. Его нужно сохранить в форме, которую можно передать третьей стороне и проверить спустя время после самого решения.

Российское регулирование искусственного интеллекта складывается в той же логике [1, 2]. Национальная стратегия развития искусственного интеллекта закрепила ориентацию на безопасное и ответственное применение интеллектуальных систем. В число основных принципов развития и использования технологий она включила прозрачность, понимаемую как объяснимость работы ИИ и процесса достижения им результатов. Последующие документы о цифровой трансформации и отраслевых сценариях использования сместили акцент с деклараций на практическую верификацию результата [3, 4]. Принцип, сформулированный на уровне стратегии, на отраслевом уровне превращается в конкретный вопрос. Чем разработчик подтверждает корректность работы системы и кто и как может эту корректность проверить?

Международные стандарты и рамки управления рисками развивают ту же линию в более технологической и процедурной форме [22, 31, 49]. Их внимание сосредоточено на устойчивости, прослеживаемости, управлении рисками, документировании и контроле жизненного цикла модели. Для ОИИ это важно, поскольку фокус смещается с единичного объяснения на процедуру его получения и повторной проверки [37, 50]. Чем выше цена ошибки, тем меньше значит разовое визуально убедительное пояснение без ясного протокола построения.

Большинство таких документов построено на риск-ориентированном подходе. Требования к прозрачности и документированию для разных систем неодинаковы и растут вместе с возможными последствиями ошибки. Конкретных методов объяснения стандарты, как правило, не предписывают. Они задают цели и требуемые свойства результата, а выбор средств оставляют разработчику, который должен этот выбор обосновать. Выбор метода объяснения сам попадает в результате в число решений, которые документируются и предъявляются при внешней оценке.

Требования к объяснимости адресованы при этом разным участникам. Регулятору и аудитору нужно подтверждение того, что система работает в заявленных границах. Оператору нужны сведения, достаточные для контроля конкретного решения. Тому, кого решение касается, нужно понятное изложение его оснований. Одно объяснение редко удовлетворяет всех сразу. Это ещё один довод в пользу многоуровневого представления, в котором формы объяснения различаются, но опираются на общее согласованное основание.

Этические принципы дополняют нормативную и стандартную рамку и переводят объяснимость в плоскость ответственности за применение интеллектуальной системы [51, 52]. Наглядно представить результат модели недостаточно. Важно, позволяет ли объяснение обнаружить скрытую предвзятость, оценить допустимость использования данных и проследить границы применимости вывода. Не менее важно понять, не создаёт ли модель видимость надёжности, когда механизм принятия решения остаётся непрозрачным. Этический аспект определяет область практической значимости объяснимости, и внешним дополнением к ней его считать нельзя.

В российской практике этические принципы закреплены в Кодексе этики в сфере искусственного интеллекта, к которому организации присоединяются добровольно [53]. Мягкий характер таких норм не делает их второстепенными. Они формируют ожидания профессионального сообщества и пользователей раньше, чем эти ожидания становятся обязательными требованиями. Там, где закон пока не даёт прямых указаний, ориентиры задают именно они.

Из нормативной, стандартной и этической рамки вытекает важное методологическое следствие. Объяснимость нужно рассматривать как воспроизводимую и проверяемую процедуру, единичного комментария к ответу модели для этого мало [22, 31]. Допустим, условия получения объяснения нельзя восстановить, устойчивость результата не оценить, а процедуру не повторить на тех же данных и при тех же параметрах. Тогда о пригодности такого объяснения для ответственной практики говорить трудно. Нормативный контур подталкивает ОИИ к более строгой валидации, сопоставимой оценке качества объяснений и выбору метода объяснения по явным критериям.

Воспроизводимость требует, в свою очередь, привязки объяснения к конкретной версии модели и данных. После переобучения или обновления модели прежнее объяснение теряет силу,

даже если внешне поведение системы почти не изменилось. Поддерживать объяснимость приходится на всём жизненном цикле системы, и разовой проверки при вводе в эксплуатацию для этого недостаточно.

Нормативно-правовая, стандартная и этическая среда задаёт ограничения, которые прямо влияют на архитектуру объяснимых интеллектуальных систем. Модель, объяснение, процедура оценки, условия применения и критерии доверия должны быть связаны в единый инженерный контур, где каждое звено документировано и проверяемо. Из этих требований вытекает и следующий шаг анализа — рассмотрение постмодельных подходов, их ограничений и перехода к интерпретируемым и встроенно-объяснимым моделям.

1.4. Постмодельные методы объяснения как первый этап развития ОИИ

Постмодельные методы объяснения стали первым массовым этапом развития ОИИ потому, что позволили работать с уже обученной непрозрачной системой без перестройки её архитектуры [54, 55]. В рассматриваемый период именно такое решение выглядело наиболее реалистичным, поскольку прикладные успехи машинного обучения уже были связаны со сложными моделями, внутренний механизм которых плохо поддавался прямому анализу [5, 34]. Поэтому исследовательское внимание закономерно сместилось к процедурам, способным объяснять готовое предсказание, не вмешиваясь в процесс обучения и устройство самой модели [56, 57].

Изначально область развивалась сразу в нескольких модальностях [34, 44]. В табличных задачах преобладали методы оценки вкладов признаков и локальной аппроксимации [55, 58]. В компьютерном зрении широкое распространение получили карты активации, градиентные методы и методы окклюзии, связывающие ответ модели с определёнными областями изображения [56, 59]. В анализе текста и последовательностей развивались методы возмущений, внимания и контрфактуальные схемы, ориентированные на выделение значимых фрагментов входа [60].

Общий методологический смысл этого этапа состоял в том, что при отсутствии прямой интерпретируемости внутреннего устройства модели исследователь всё же получал возможность анализировать связь между результатом и входными факторами [55, 58]. Если путь вывода остаётся скрытым, необходимо восстановить ту часть информации, которая позволяет понять, за счёт каких признаков, областей входа или фрагментов данных был получен результат [56, 61]. Такая логика сделала постмодельные методы первым рабочим языком взаимодействия с чёрным ящиком [62, 63]. Благодаря этому объяснимость перестала быть исключительно теоретическим требованием и вошла в прикладную практику анализа моделей.

Более существенно, что этот этап впервые превратил проблему непрозрачности из общего теоретического затруднения в самостоятельную инженерную задачу. До этого отсутствие обозримости сложных моделей воспринималось прежде всего, как неизбежное следствие повышения их выразительной способности. Постмодельные методы поставили вопрос иначе: если внутренний механизм вывода скрыт, необходимо разработать процедуры, позволяющие хотя бы частично восстановить основания полученного результата. Благодаря этому объяснимость вошла в инженерную и прикладную практику как реальная процедура анализа уже функционирующей модели, а не как абстрактное требование к будущим системам [10, 11].

Вместе с тем уже на этом этапе обозначилось принципиальное ограничение внешнего объяснения, так как оно формируется после получения результата и потому не входит в сам механизм вывода модели [20, 34, 64, 65]. В дальнейшем именно эта особенность стала одним из оснований критического пересмотра возможностей постмодельных методов [66, 67].

1.5. Ограничения постмодельных подходов

Постмодельных техник со временем стало много, но качество объяснения от этого само по себе не выросло [15, 68]. На первый план вышел другой вопрос. Насколько получаемое объяснение действительно связано с механизмом вывода и не создаёт ли оно лишь внешне убедительную картину результата? Вопрос этот принципиален, потому что постмодельный метод видит модель только снаружи, по её откликам на входные данные. Всё, что он сообщает о внутренней логике решения, восстановлено косвенно, и насколько точно это сделано, тоже приходится проверять.

Серьёзнее всего сказался локальный характер большинства постмодельных процедур [54, 55]. На вопрос о конкретном случае они отвечали хорошо. Восстановить устойчивую структуру поведения модели на множестве ситуаций им удавалось значительно хуже. Локальное объяснение освещает отдельный фрагмент решения, целостного представления о логике модели как системы оно не даёт. Причина заложена в самой конструкции метода. Объяснение строится для окрестности конкретного объекта, границы этой окрестности задаёт исследователь, и от его выбора зависит результат. Свести множество локальных объяснений в общую картину тоже непросто. Объяснения для близких объектов могут расходиться между собой. Взаимодействия признаков, которые определяют поведение модели в целом, в отдельной окрестности часто никак не проявляются.

Глобальные постмодельные процедуры снимают эту трудность лишь отчасти. Они приближают поведение модели более простой интерпретируемой конструкцией или усредняют локальные вклады по выборке. В первом случае объяснение относится уже к суррогатной модели, и между ней и исходной остаётся разрыв, который обычно никак не оценивается. Во втором усреднение сглаживает различия в поведении модели на разных подмножествах данных, и картина выходит проще, чем модель на самом деле.

Характерным выражением локальной логики стали методы локальной аппроксимации, прежде всего *LIME* [54], а также подходы на основе вкладов признаков и визуальных карт значимости [55, 69]. Для анализа частных решений эти инструменты оказались полезны. Именно на них, однако, особенно ясно обозначился вопрос, можно ли считать локальное объяснение устойчивым, переносимым и связанным с внутренней логикой вывода. Результат локальной аппроксимации зависит от того, как порождаются возмущённые объекты в окрестности. При повторном запуске с другой случайной выборкой он может заметно измениться. Возмущённые объекты к тому же нередко выходят за пределы области, на которой обучалась модель. Объяснение начинает тогда описывать её поведение там, где выводы модели ничем не подкреплены. Вклады признаков чувствительны к выбору опорной точки или фонового распределения. Если признаки коррелированы, значимость между ними делится во многом произвольно.

Со временем стало ясно и то, что наглядность ещё не означает объяснимости [44, 70, 71]. Визуально правдоподобная или интуитивно убедительная картина может создавать ощущение контроля там, где связь объяснения с механизмом вывода слаба. Проверки карт значимости, анализ устойчивости к возмущениям и сопоставление разных объяснителей показали, что правдоподобие, воспроизводимость и верность объяснения автоматически не совпадают. Логика таких проверок проста. Если карта почти не меняется после случайной перестройки параметров модели или меток обучающих данных, она отражает скорее структуру входа, чем то, чему научилась модель. Правдоподобие оценивает человек, сопоставляя объяснение со своими ожиданиями. Верность же измеряется вмешательством в работу модели, и эти две оценки могут расходиться сколь угодно сильно.

В ответственных приложениях такое расхождение становится принципиальным [15, 16]. Объяснение, которое создаёт видимость контроля при его фактическом отсутствии, укрепляет ложное доверие к модели. В критических областях это опаснее, чем отсутствие наглядной иллюстрации, ведь эксперт получает неверное представление о реальных основаниях решения. Не имея объяснения, специалист хотя бы сохраняет осторожность и проверяет вывод сам. Убедительное, но неверное объяснение эту осторожность снимает. Сильнее всего это проявляется, когда объяснение совпадает с его собственными ожиданиями и потому принимается без дополнительной проверки.

Отдельная трудность связана с оценкой самих объяснений. Эталонного объяснения для обученной модели, как правило, не существует. О верности приходится судить косвенно, по реакции модели на удаление или замену признаков, которые объяснитель счёл значимыми. Такие процедуры сами опираются на возмущения входа и наследуют часть тех же проблем. Разные метрики могут по-разному упорядочивать одни и те же объяснители. Разные объяснители, в свою очередь, указывают на разные признаки для одного и того же решения. Исследователь оказывается перед выбором, а обоснованного правила для него постмодельная парадигма не даёт.

Первый этап развития ОИИ сыграл двойственную, но принципиально важную роль [10, 11]. Он придал проблеме объяснимости практическую значимость и дал исследователю инструменты для анализа уже построенных моделей. На нём же обнаружились фундаментальные ограничения внешнего объяснения, и у всех у них общий корень. Постмодельный метод строит объяснение вне модели, по наблюдаемым откликам. Результат поэтому зависит от собственных допущений метода не меньше, чем от самой модели. Отсюда и зависимость итога от выбранного объяснителя, и разрыв с внутренней логикой вывода. Для глубокого и устойчивого объяснения сложных интеллектуальных систем этого оказалось недостаточно. Выявленные ограничения закономерно привели к интерпретируемым и встроенно-объяснимым моделям. Связь между устройством системы и её выводом в них короче, содержательно яснее и закреплена в архитектуре [72–76].

1.6. Переход к встроенной и архитектурной объяснимости в ОИИ 2.0

Ограничения постмодельных методов объяснения показали, что внешнего пояснения результата недостаточно в тех случаях, когда требуется не только интерпретировать уже полученный вывод, но и установить его связь с внутренним механизмом формирования решения [15, 77]. В этой связи в современном развитии ОИИ усилился интерес к таким подходам, в которых объяснимость обеспечивается не только на уровне последующей интерпретации, но и на уровне самой архитектуры модели, её промежуточных представлений и организации вычислительного процесса [19]. Особенно это относится к сложным глубоким, гибридным, мультимодальным и иерархически организованным интеллектуальным системам, в которых решение формируется не одномоментно, а последовательно, через несколько взаимосвязанных уровней представления и преобразования данных [16, 19].

При этом интерпретируемые модели как класс не являются новым этапом по отношению к постмодельным методам. Исторически они возникли значительно раньше и изначально были

связаны с такими обозримыми конструкциями, как деревья решений, системы правил и иные модели, допускающие прямое прослеживание хода вывода. Однако в логике ОИИ 2.0 их роль переосмысляется. Речь идёт уже не о возврате к простым моделям как таковым, а о формировании встроенной объяснимости в более сложных интеллектуальных системах, включая глубокие, гибридные, многокомпонентные и иерархически организованные архитектуры [16, 19, 78, 79].

Достоинство интерпретируемых подходов состоит в сокращении разрыва между устройством модели и её результатом [13, 75]. В таких системах анализу доступны итоговый вывод и элементы его формирования: промежуточные признаки, концепты, правила, структурные зависимости и иные содержательно различимые представления. Благодаря этому объяснение опирается не только на внешнюю процедуру интерпретации, но и на саму организацию вывода.

Вместе с тем классические интерпретируемые конструкции не позволяют в полной мере решить задачу объяснимости в современных прикладных сценариях [72, 80]. Это обусловлено тем, что реальные данные характеризуются высокой размерностью, нелинейностью, сложной структурой зависимостей и нередко мультимодальной природой, тогда как практические задачи требуют от модели высокой точности, устойчивости и адаптивности [81, 82]. По этой причине развитие области пошло не по пути отказа от сложных архитектур, а по пути поиска таких способов их организации, при которых объяснение становится встроенным свойством модели. Это существенно и для прикладных задач геоэкологии (мониторинг загрязнения почв, геоэкологическое районирование, выбор площадок размещения полигонов ТКО), где объяснение приходится связывать с конкретными участками территории и факторами риска [83–89].

В этой линии особое место заняли подходы, где основания вывода закреплены в самой архитектуре модели [90, 91]. Это нейро-символьные и концептно-ориентированные модели, многокомпонентные системы с интерпретируемыми промежуточными слоями и гибридные решения. В гибридных решениях статистическое обучение сочетается с явными структурными ограничениями, правилами или предметно различимыми состояниями [74, 78, 92]. Объяснение в таких моделях формируется по ходу их работы и внешняя процедура интерпретации для этого не нужна.

Системная трактовка интеллектуальной системы прослеживается и в работах А.П. Еремеева, В.Н. Вагина, А.А. Башлыкова по интеллектуальным системам принятия решений и управления. В них анализируются модели вывода, а также программно-реализуемый контур функционирования системы, включающий процедуры вывода, сопровождения и контроля. Такая постановка близка настоящему исследованию, поскольку объяснимость здесь рассматривается как свойство архитектурно организованного вычислительного контура.

Важным аргументом в пользу перехода к встроенной объяснимости является вариативность постмодельных объяснений [64, 66]. Даже при фиксированных данных и неизменной модели результат объяснения может зависеть от метода, параметров процедуры, выбора фонового распределения, корреляционной структуры признаков, сдвига распределения данных и чувствительности к возмущениям [65]. Эти источники вариативности представлены на рисунке 1.5. Следовательно, внешнее объяснение не всегда образует устойчивую основу доверия к выводу модели, что усиливает интерес к архитектурным решениям, в которых объяснимость опирается на внутреннюю организацию вычислительного процесса [19, 77].

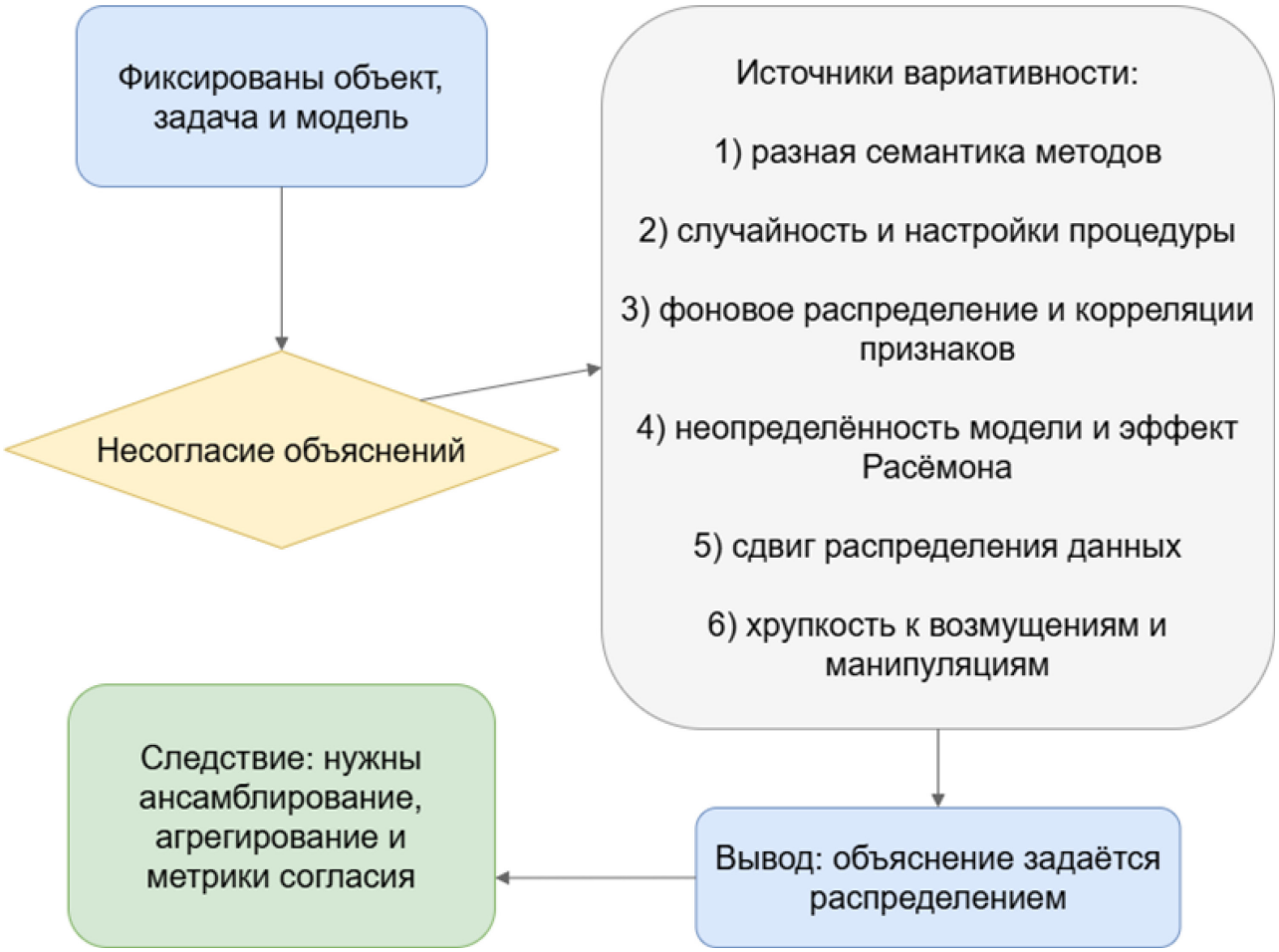


Рисунок 1.5 – Источники вариативности объяснений

По мере перехода к встроенной и архитектурной объяснимости изменяется сама постановка задачи проектирования интеллектуальной системы [13, 19]. Недостаточно дополнить модель внешним поясняющим механизмом или сделать интерпретируемым отдельный компонент. Требуется обеспечить такую организацию вычислительного процесса, при которой различные уровни представления данных, промежуточные состояния модели и итоговый результат сохра-

няют содержательную связь между собой. В этом случае объяснимость начинает рассматриваться не как внешний аналитический слой, а как характеристика архитектуры интеллектуальной системы в целом [16, 19].

В глубоких, многокомпонентных и иерархически организованных архитектурах решение формируется через несколько уровней внутренних представлений, каждый из которых может вносить собственный вклад в итоговый вывод [19, 35]. Поэтому объяснение перестаёт быть однослойным: оно затрагивает признаки, внутренние представления, концептуальные структуры, промежуточные состояния и результирующее решение. В этом состоит принципиальное отличие встроенной объяснимости от первого этапа развития ОИИ. Если постмодельные подходы интерпретировали уже полученный результат, то архитектурная объяснимость ставит вопрос о связи объяснения с самой логикой формирования вывода [16, 19]. Однако по мере усложнения архитектуры возрастает число источников объяснения, различающихся по уровню представления, форме, степени детализации и характеру связи с механизмом вывода. В результате возникает не только задача получения объяснения, но и задача установления его внутренней согласованности в пределах единого объяснительного контура.

Переход к ОИИ 2.0 закономерно подводит к необходимости специального рассмотрения многоуровневого характера объяснений. По мере усложнения интеллектуальных систем становится недостаточным анализировать отдельные объясняющие процедуры изолированно. Требуется учитывать их место в общей архитектуре модели, характер взаимосвязи между уровнями представления, а также наличие разнородных источников объяснения, требующих содержательного сопоставления. Именно это обстоятельство определяет дальнейшее направление исследования, связанное с анализом многоуровневого формирования объяснений и предпосылок к их согласованию в сложных интеллектуальных системах.

1.7. Автоматизированные контуры ОИИ и воспроизводимость объяснений

В развитии ОИИ заметен переход от разрозненного применения отдельных объясняющих методов к автоматизированному контуру, в котором объяснения формируются, оцениваются и используются повторно [19, 93]. Раньше главной задачей было построить локальные и глобальные пояснения для уже обученной модели. Теперь отдельным предметом анализа становится сама процедура их получения. С позиций системного подхода речь идёт о подсистеме объяснимости внутри общей интеллектуальной системы. У этой подсистемы есть свои входы и выходы, ограничения, управляющие параметры и критерии качества.

Воспроизводимой при этом должна быть вся процедура построения объяснения, отдельного численного показателя для этого мало [94, 95]. Чтобы результат можно было проверить по существу, фиксируют состав и версию данных, конфигурацию модели, параметры объясняющего метода, метрики и условия эксперимента [96, 97]. Объяснение становится процессуально определённым объектом. Кроме содержательности и наглядности, у него важны прослеживаемое происхождение и устойчивость к вариациям условий. Важна и возможность получить его повторно в сопоставимой конфигурации.

Острее всего эта проблема стоит при автоматизированном построении моделей [19, 21, 93]. Автоматический выбор модели, гиперпараметров и вычислительного контура давно сложился в отдельное направление. Объяснимость же долго оставалась внешним этапом, который подключали к уже построенной модели. Так возникает архитектурная непрозрачность, при которой скрытыми остаются и сама модель, и процедура её выбора, настройки и включения в вычислительный процесс.

На этом фоне появились работы, где принципы автоматизированного построения моделей и ОИИ соединяются в одном контуре анализа [98–102]. Объяснение в них уже не сводится к постмодельному комментарию. Оно входит в более широкий цикл, который охватывает построение, настройку, анализ и последующую верификацию модели. В ряде работ это направление называют *AutoXAI [98, 101]*.

Одной из ранних попыток такого объединения стала система *AutoXAI* (*Cugny* и др.) [98], в которой поставлена задача автоматического выбора и настройки метода объяснения для заданной пары «данные – обученная модель» с учётом пользовательских требований. Работа показала, что выбор объяснителя поддаётся формализации и может быть отдельной процедурой, не зависящей от эвристических предпочтений исследователя. Объяснение в этой архитектуре остаётся, однако, постмодельным. Рекомендации выдаются после обучения модели, и обратной связи с процедурой оптимизации нет.

Более развитую форму включения объяснимости в автоматизированный конвейер демонстрируют предметно-ориентированные решения, в частности *AutoXAI4Omics* [99] и *AutoML-XAI* для метаболомики [100], где вычисление объяснений автоматически запускается после выбора модели и настройки гиперпараметров. Работа El Shawi и Jamel [101] развивает близкую линию, связывая выбор метода объяснения с метаобучением и метапризнаками данных и модели. Эти подходы смещают акцент с отдельного объяснителя на процедуру его выбора, однако объяснение по-прежнему сохраняет апостериорный характер и не участвует в формировании траектории оптимизационного поиска.

В 2022–2025 гг. появилась ещё одна группа решений, нацеленных на прозрачность самого процесса автоматизированного построения модели [102–105]. *XAutoML* с помощью визуальной

аналитики показывает пространство кандидатов, историю оптимизации и роль признаков в конвейере [102]. *UniAutoML* делает процедуру интерактивной: диалоговый интерфейс поясняет принимаемые решения [103]. Предметно-ориентированная система *xAutoML* встраивает инструменты объяснимости в разные этапы промышленного конвейера [104], а *Knowledge*-Driven *AutoML* опирается на предметные знания, онтологии и экспертные правила [105]. Архитектуры у этих систем разные, а ограничение общее. Они делают прозрачнее уже сложившуюся процедуру или подключают объяснение после оптимизации. Управляемым фактором внутри поиска объяснимость так и не становится.

Из обзора следует, что управлять приходится уже не одной моделью, но и контуром выбора и применения метода объяснения. Растёт роль обратной связи между результатами интерпретации и последующими этапами анализа, и объяснимость всё чаще включают в общую архитектуру управления качеством модели. Автоматизированный контур объяснимости можно понимать как функциональное сопряжение двух подсистем, прежде существовавших раздельно: автоматизированного построения модели и формирования объяснения. Обобщённая схема такого контура представлена на рисунке 1.6. Кроме требований к объяснению и входных ограничений, на ней показаны оптимизационная настройка и контур обратной связи, участие эксперта и контекстно-зависимый выбор метода объяснения. Объяснимость входит в схему как компонент контура управления, связанный с построением, оценкой и корректировкой интеллектуальной системы.



Рисунок 1.6 – Обобщённая схема автоматизированного контура объяснимости и требований к организации оптимизационного процесса

Архитектурная непрозрачность возникает в таком контуре на разных этапах автоматизированного построения модели, от предобработки данных и конструирования признаков до поиска

архитектуры и ансамблирования. На каждом этапе принимаются решения, которые влияют на итоговое предсказание, и восстановить их при последующем анализе удаётся не всегда. Декомпозиция по этапам приведена в таблице 1.3. Из неё видно, что непрозрачность переходит с уровня отдельной модели на уровень всей процедуры её построения [93, 102, 106].

Таблица 1.3 – Декомпозиция компонентов автоматизированного конвейера построения модели, объектов оптимизации и источников архитектурной непрозрачности

Компонент конвейера	**Типичные методы**	**Источник архитектурной непрозрачности**
Предобработка данных	Импутация, нормализация, кодирование, масштабирование	Неочевидность влияния преобразований на итоговую метрику
Конструирование признаков	Полиномиальные признаки, взаимодействия, автоматический отбор	Слабая трассируемость причин включения или исключения признаков
Выбор модели	Линейные модели, деревья решений, бустинг, нейронные сети	Отсутствие формализованного обоснования выбора класса модели
Оптимизация гиперпараметров	*SMAC*, *Bayesian Optimization*, эволюционные алгоритмы, *Random Search*	Непрозрачность траектории поиска, критериев остановки и влияния промежуточных оценок
Поиск архитектур нейронных сетей (*NAS*)	*DARTS, ENAS, AmoebaNet [107, 108]*	Сложность реконструкции причин выбора архитектуры
Ансамблирование	Стекинг, блендинг, голосование	Усложнение интерпретации совокупного вклада моделей в итоговое предсказание

Не менее важны ограничения самих методов объяснения при их включении в автоматизированный контур [43, 68]. Они различаются по вычислительной сложности, универсальности, глубине интерпретации и пригодности для многократного применения в процессе выбора и настройки модели. Поэтому объясняющие методы не являются взаимозаменяемыми, а их выбор должен учитывать тип модели, свойства данных, цель анализа, вычислительные ограничения и требования к устойчивости и воспроизводимости объяснения [40, 109]. Сравнительная характеристика приведена в таблице 1.4.

Таблица 1.4 – Сравнительная характеристика методов ОИИ

Метод	**Вычислительная сложность**	**Методологические ограничения**	**Ограничения для интеграции в AutoML**
SHAP [55]	$O(n * 2\hat{}n)$	Высокая вычислительная стоимость, зависимость от числа признаков	Сложно использовать внутри оптимизационного цикла из-за необходимости многократных вычислений

Метод	Вычислительная сложность	Методологические ограничения	Ограничения для интеграции в AutoML
KernelSHAP	$O(m * C_{model})$	Результат зависит от выбора m и способа сэмплирования	Слишком затратно при больших вычислениях
TreeSHAP	$O(T * D\hat{}2)$	Применим преимущественно к деревообразным моделям	Не универсален для нетривиальных архитектур
LIME [54]	$O(m * C_{model})$	Зависит от генерации выборки и параметров локальной аппроксимации	Плохо переносится между различными архитектурами
Grad-CAM [61]	$O(C_{\text{forward}} + C_{backward})$	Работает в основном для архитектур с картой признаков	Не универсален для произвольных архитектур
Self-attention [110]	$O(L\hat{}2 * d)$	Высокие веса внимания не гарантируют опору модели именно на эти фрагменты	Применим только к моделям с механизмом внимания

Современная ОИИ выходит за пределы построения отдельных объяснений и движется к воспроизводимым, прослеживаемым и сопоставимым контурам объяснимости. В таких контурах данные, модель, метод объяснения и процедуры оценки образуют взаимосвязанную структуру, а доверие к результату определяется возможностью восстановить происхождение объяснения, повторить процедуру его построения и проверить устойчивость результата. Эта линия анализа подводит к задачам обоснованного выбора объяснителя, сопоставления объяснений и оценки их качества в зависимости от контекста задачи.

1.8. Агрегирование, ансамблирование, воспроизводимый выбор объяснителя и проблемные зоны современной ОИИ

Автоматизированный контур объяснимости почти всегда даёт для одной модели или одного решения несколько объяснительных представлений [19, 35]. По форме они бывают разными, от вкладов признаков и карт значимости до правил и контрфактуальных примеров [34, 40, 68]. Анализировать приходится уже всё множество объяснений. Важно, насколько они согласуются, где противоречат друг другу, дублируют одно другое и сохраняются ли при изменении условий [19, 35, 66]. Так в ОИИ складывается отдельная проблема: как сопоставлять и объединять объяснительные представления в одном объяснительном пространстве.

Агрегирование и ансамблирование становятся здесь системообразующими [111–113]. С их помощью набор частных объяснений превращается в более устойчивое и согласованное итоговое представление. Расхождение между объясняющими методами при этом перестаёт быть

только недостатком. По нему можно судить о границах устойчивости результата и о том, насколько модель чувствительна к форме интерпретации [66, 113, 114]. Поиск одного условно «лучшего» объяснителя задачу не решает. Нужны способы сопоставлять и объединять несколько частных объяснений в общей процедуре анализа.

Обобщённая схема перехода от множества частных объяснений к ансамблевому результату показана на рисунке 1.7. Итоговое объяснение получается на ней из обработки, сопоставления и сведения нескольких частных представлений, и ни один метод не даёт его в готовом виде. Ансамблирование стоит поэтому считать отдельной стадией объяснительного анализа.



Рисунок 1.7 – Обобщённая схема ансамблирования объяснений

Несколько объяснений, собранных вместе, ещё не дают осмысленного объединения. Сначала их приводят к сопоставимому виду, нормируя и выравнивая по масштабу и знаку. Затем их сопоставляют по рангам, объединяют вероятностно или применяют правило консенсуса [115–

117]. Агрегирование в современной ОИИ оказывается набором процедур, из которых складывается структурно согласованное объяснение, и к простому усреднению оно не сводится. Классификация этих методов приведена на рисунке 1.8.

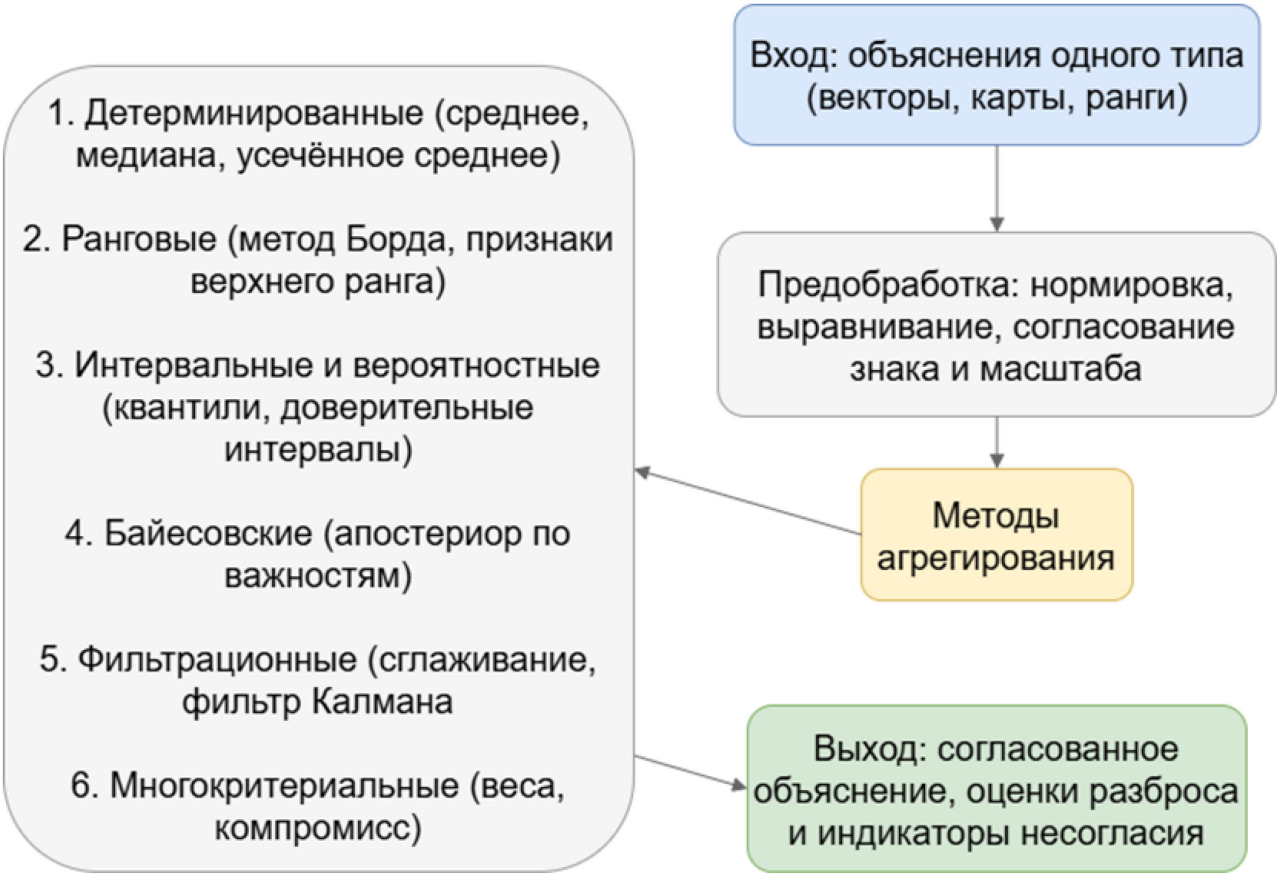


Рисунок 1.8 – Классификация методов агрегирования для построения консенсусного объяснения

Из классификации видно, что агрегирование строится по разным принципам, в том числе на голосовании и вероятностных схемах консенсуса [111, 112, 115, 116]. Универсальной техники здесь нет. Процедуру приходится строить с учётом природы объяснений и того, как они соотносятся друг с другом.

С разнотипными и мультимодальными объяснениями задача заметно усложняется. Объяснения различаются тогда семантикой, формой и уровнем описания, и численного объединения для них мало [19, 77, 118]. Их переводят в сопоставимое представление, согласуя смысловые единицы и разрешая конфликты между модальностями, и только после этого собирают в целостный результат. Такую процедуру мы называем комплексированием объяснений, её архитектура представлена на рисунке 1.9.



Рисунок 1.9 – Архитектура комплексирования мультимодальных объяснений

С комплексированием в ОИИ появляется задача межформатной и межуровневой согласованности объяснений. Агрегирование сводит в основном частные результаты одного класса. Комплексирование объединяет объяснения, относящиеся к разным семантическим и функциональным уровням. Отсюда возникает вопрос, как организовать взаимодействие объяснительных представлений внутри общего контура анализа. Отдельные техники интерпретации оказываются лишь частью этой задачи.

Здесь особое значение получает неопределённость [19, 64, 65, 77]. Пока объяснение остаётся единичным результатом отдельного метода, неопределённость обычно считают ограничением конкретной техники. У агрегированного или комплексированного объяснения она становится свойством самого итогового результата. По ней видно, насколько согласованы частные объяснения и как интерпретация зависит от выбора метода, данных и параметров процедуры. Видно и то, где проходят пределы доверия к выводу [119–122]. Неопределённость должна поэтому входить в итоговое объяснение как его собственная характеристика.

С инженерной точки зрения эти процедуры складываются в последовательность операций (рисунок 1.10). Она начинается с генерации частных объяснений и заканчивается экспертной проверкой и документированием. Объяснение получается здесь в результате организованной процедуры с проверкой, итеративной настройкой и контролем применимости, и разовым аналитическим актом его уже не назовёшь [43, 97, 119, 121].



Рисунок 1.10 – Инженерный конвейер построения и валидации объяснений

Отдельное место в этой процедуре занимает выбор объяснителя [66, 109]. Методы объяснения сильно различаются по области применимости, устойчивости, глубине интерпретации и вычислительной цене [40, 46, 68]. Выбирать метод произвольно или по привычке исследователя поэтому нельзя. Выбор должен опираться на тип данных, архитектуру модели, задачу анализа и критерии качества объяснения, и тогда он становится частью воспроизводимого управления объяснительным контуром. Схема работы агента выбора объяснителя представлена на рисунке 1.11.

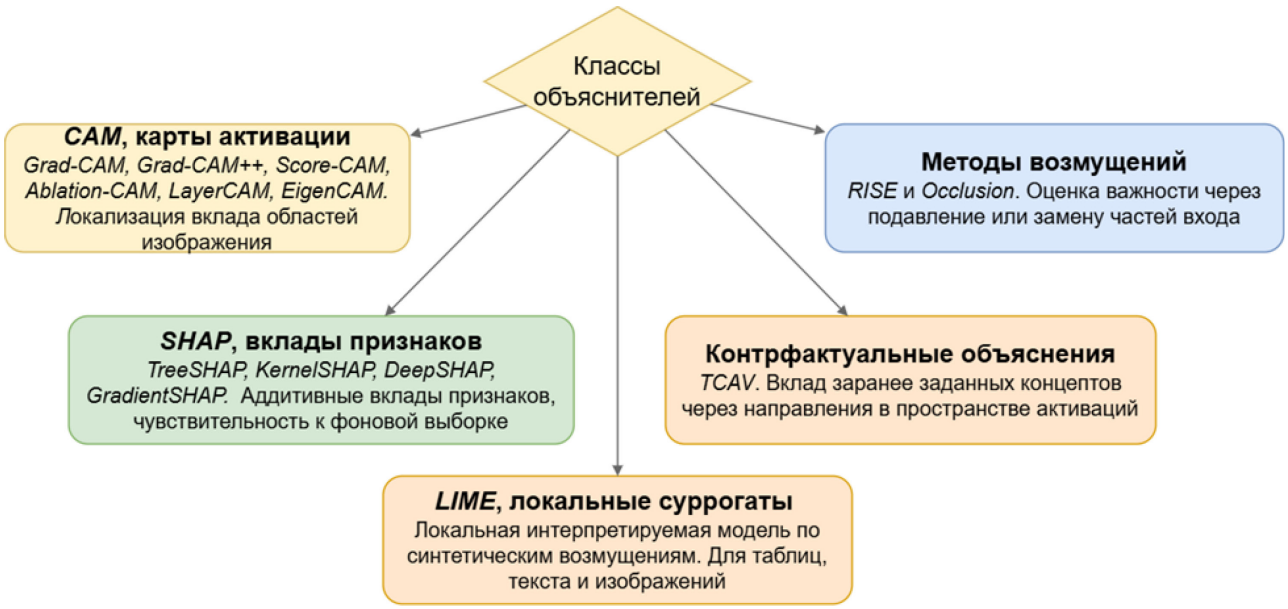


Рисунок 1.11 – Схема работы агента выбора объяснителя

Развитие ОИИ смещает внимание с отдельных объяснений на объяснительный контур в целом. В нём методы объяснения используются совместно, оцениваются сопоставимо и выбираются обоснованно, с учётом типа модели, данных и цели анализа. На этих процедурах и сосредоточен анализ.

1.9. Выводы по главе и постановка задач диссертационного исследования

Проведённый системный анализ предметной области ОИИ показал, что её становление связано с переходом от моделей, основанных на явном представлении знания, к высокоэффективным, но внутренне непрозрачным архитектурам машинного обучения [5, 33]. По мере роста результативности таких систем всё труднее становилось проследить основания их вывода, что превратило объяснимость в самостоятельную научную и прикладную проблему. Объяснение перестало рассматриваться как факультативное дополнение к результату модели и стало условием её содержательной проверяемости и ответственного применения.

Уточнение понятийного аппарата позволило развести прозрачность, интерпретируемость, объяснимость и интерпретацию по их исследовательским функциям [38, 39]. Это разграничение задало понятийную рамку дальнейшего анализа и позволило отделить свойства системы и модели от свойств объяснения и способов его предъявления пользователю.

Анализ нормативной, стандартной и этической рамки показал, что объяснимость имеет не только методологическое, но и практическое значение [22, 37]. В регулируемых и чувствительных областях она связана с требованиями доверия, управления рисками, устойчивости, воспроизводимости и ответственности. Поэтому объяснение должно рассматриваться не как внешний комментарий к полученному выводу, а как элемент проверяемого применения интеллектуальной системы.

Рассмотрение этапов развития ОИИ показало, что постмодельная интерпретация сыграла важную роль в практическом становлении области [54, 55]. Она сделала возможным массовый анализ уже обученных непрозрачных моделей и превратила объяснимость из общего теоретического требования в инженерную процедуру. Вместе с тем её возможности оказались ограничены внешним по отношению к модели характером объяснения, зависимостью результата от выбранного объяснителя, нестабильностью и трудностями глобального обобщения. Это обусловило переход к интерпретируемым, гибридным и встроенно-объяснимым архитектурам, в которых связь между устройством системы и её выводом становится более короткой и содержательно определённой.

Дальнейшее развитие предметной области связано с переходом к ОИИ 2.0, где на первый план выходят воспроизводимость, сопоставимость методов, многокритериальная оценка качества и связь объяснимости с жизненным циклом модели [19, 21]. Современная ОИИ уже не сводится к набору отдельных техник визуализации или локальной интерпретации: в качестве самостоятельного объекта исследования выступает множество объяснительных представлений, требующих сопоставления, согласования и оценки. Это приводит к постановке задач агрегирования, комплексирования, учёта неопределённости и обоснованного выбора объясняющего метода.

Системный анализ позволяет выделить несколько взаимосвязанных проблемных зон современной ОИИ: отсутствие единого контура формирования объяснений, разобщённость модельных и постмодельных представлений, недостаточную разработанность средств многоуровневого согласования, ограниченность формальных процедур агрегирования и комплексирования гетерогенных объяснений, а также недостаточное включение неопределённости и воспроизводимого выбора объяснителя в итоговый объяснительный результат [19, 43, 66, 109, 119]. В совокупности эти ограничения показывают, что дальнейшее развитие ОИИ связано не столько с увеличением числа частных методов объяснения, сколько с построением архитектуры объяснения, в которой формирование, сопоставление, агрегирование, оценка качества и выбор объясняющего метода рассматриваются как взаимосвязанные элементы единого исследовательского контура.

Связь между выявленными ограничениями предметной области и направлением диссертационного исследования представлена на рисунке 1.12. Значение данной схемы состоит в том,

что она позволяет рассматривать задачи диссертации не как произвольный набор частных вопросов, а как ответ на внутреннюю логику развития современной ОИИ.

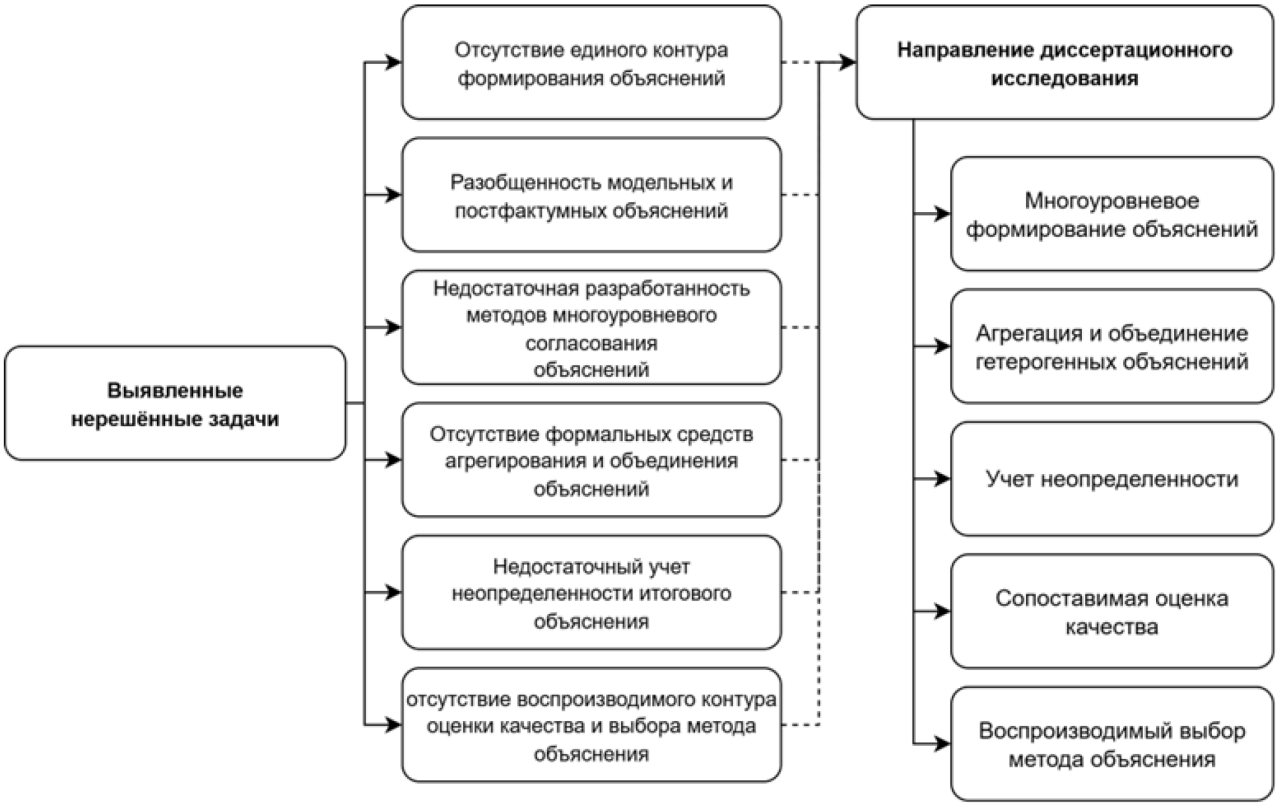


Рисунок 1.12 – Карта выявленных направлений диссертационного исследования

2. АППАРАТ МНОГОУРОВНЕВОГО ФОРМИРОВАНИЯ, СОГЛАСОВАНИЯ, ОБЪЕДИНЕНИЯ И СОПОСТАВИМОЙ ОЦЕНКИ ОБЪЯСНЕНИЙ В СИСТЕМАХ ИИ

Глава посвящена формальному аппарату многоуровневой объяснимости для иерархически организованных и гетерогенных интеллектуальных систем. Объяснение понимается в ней как продукт прослеживаемого вычислительного и объяснительного контура системы, тогда как во многих подходах его строят как внешнее постмодельное описание уже готового результата. Такой взгляд даёт возможность перейти от локальных интерпретаций отдельных компонентов к представлению всего контура обработки данных и уже на нём строить формальные процедуры работы с объяснениями.

Изложение идёт от структуры к оценке. Сначала объяснимость описывается как свойство контура, представимого ориентированным графом, и на узлах этого графа вводятся уровни и объяснительные объекты. Потом объекты разных уровней согласуются между собой, а объяснения из разных источников сводятся в общий результат. Последние разделы отведены неопределённости, метрической оценке качества и выбору объяснителя, который рассматривается сначала для заданной модели, затем совместно с моделью и, наконец, в агентной постановке.

2.1. Многоуровневость иерархически организованной интеллектуальной системы

Многоуровневость здесь понимается прежде всего как свойство организации самой интеллектуальной системы. Большинство исследований ОИИ сосредоточено на локальных и постмодельных объяснениях отдельных решений [7, 19]. В иерархически организованной системе входная информация, однако, преобразуется последовательно, проходя через цепочку взаимосвязанных узлов. В одних выделяются признаки, в других формируются внутренние представления или активируются интерпретируемые структуры, и лишь в конце вырабатывается результирующий вывод.

Такое понимание опирается на классическую теорию иерархических систем. Ещё М. Минский описывал интеллектуальную программу как иерархию подзадач и методов их решения [123]. Г. Саймон показал, что сложные системы, естественные и искусственные, обычно иерархичны и почти разложимы, то есть связи внутри подсистем у них сильнее, чем между подсистемами [124]. М. Месарович, Д. Мако и И. Такахара различили в многоуровневой системе страты,

слои и эшелоны и поставили задачу координации между уровнями [125], а Н.Н. Моисеев рассмотрел информационный обмен в иерархических системах [126]. Из этой традиции заимствована идея стратифицированного описания. Признаковый, модельный, интерпретируемый и результирующий уровни объяснения рассматриваются как страты описания одной и той же системы, а переходы между ними играют роль каналов координации.

Объяснению поэтому приходится охватывать весь путь формирования результата. Итоговая атрибуция сама по себе не говорит, как локальная значимость признаков связана с внутренними состояниями модели и с интерпретируемыми структурами, участвующими в выработке вывода. Если же проследить переходы между узлами, становится видна логика преобразования данных, обнаруживаются границы применимости отдельных модулей и можно установить, откуда взялось итоговое решение.

Для формализации контур обработки данных представляется типизированным ориентированным графом. Его вершины соответствуют функциональным узлам вычислительной и объяснительной обработки, рёбра — допустимым переходам данных, внутренних состояний, объяснительных объектов и оценочных характеристик (см. рисунок 2.1). Объяснимость при этом задаётся системой локальных описаний узлов и переходов, внешняя постобработка результата для неё не нужна. Представлять систему сетью взаимосвязанных функциональных элементов в системном анализе принято давно [127]. Новое здесь — типизация узлов и переходов по их роли в объяснении.

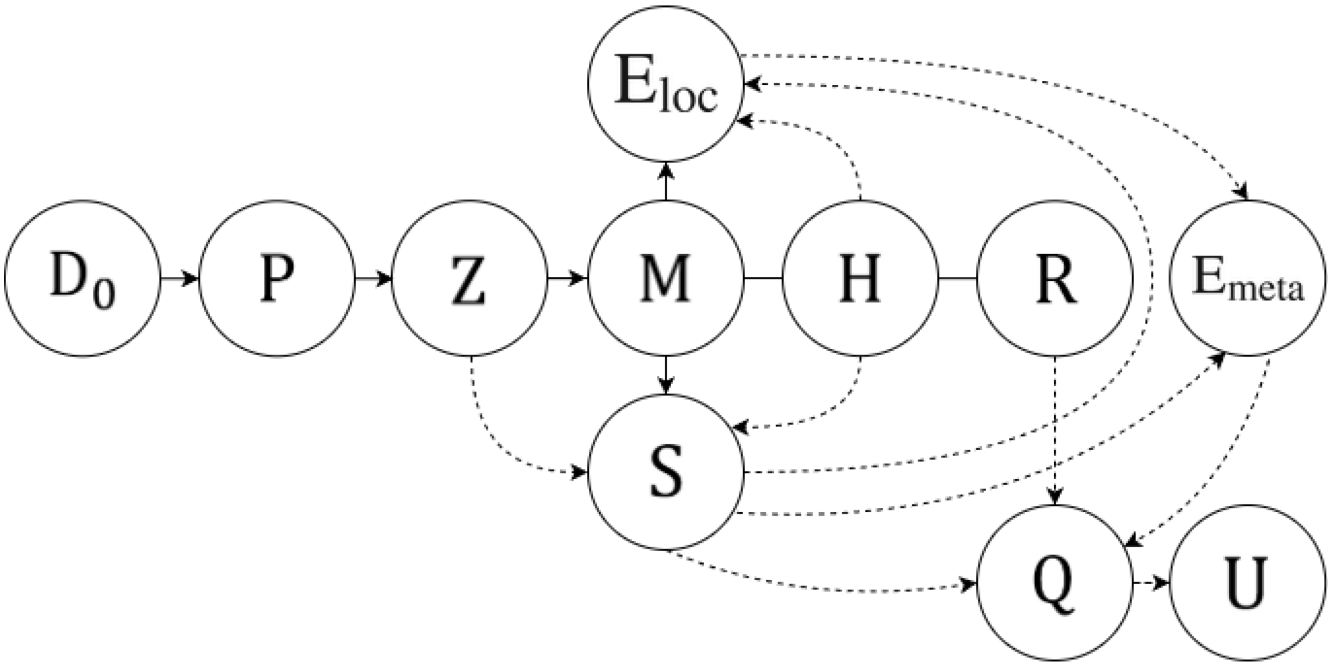


Рисунок 2.1 – Графовое представление объяснимого контура обработки данных иерархически организованной системы ИИ

Вершина D_0 соответствует исходным данным, P — этапу предобработки, Z — пространству признаков, M — базовой модели, H — внутренним представлениям модели, R — результату и журналу переходов, S — интерпретируемым структурам, E_{loc} — локальным объяснениям, E_{meta} — метауровневым объяснениям, Q — оценке качества, трассируемости и неопределённости, U — пользовательскому представлению. В одной схеме оказываются и вычислительный контур получения решения, и объяснительный контур, который отвечает за интерпретацию, согласование и оценку результата.

Объяснимость вычислительного контура понимается как возможность восстановить маршрут преобразования данных по активированным узлам, промежуточным состояниям и результатам переходов. Близкое средство описания интеллектуальных систем — функциональные сети [127, 128], представляющие систему через взаимодействующие функциональные элементы и связи между ними. Они дают структурно-функциональную картину, но для объяснимости её мало, пока к ней не добавлена динамическая трассировка активаций, промежуточных состояний и результатов переходов. Графовое представление можно считать развитием идей функциональных сетей, доведённым до операционной формы под задачу объяснимости.

Онтологическое представление формализует понятия предметной области, отношения между ними и ограничения их совместимости [129, 130]. На нём хорошо определяется семантика признаков и интерпретируемых объектов, но конкретное вычисление по онтологии не восстановить, для этого нужны сведения об активированных операциях, использованных данных и промежуточных результатах. Онтология и граф обработки дополняют друг друга. Первая задаёт семантический слой типизации, второй фиксирует зависимости между объектами конкретного запуска.

2.2. Формализация узлов ориентированного графа объяснимого контура обработки данных

Уровни объяснения, введённые в разделе 2.1, можно теперь описать на языке ориентированного графа объяснимого контура. Каждому узлу графа сопоставляется тип объяснительного объекта (признаковый, модельный, интерпретируемый или результирующий), и многоуровневая объяснимость предстаёт системой узлов и допустимых переходов между ними (см. рисунки 2.2, 2.3) [131–133].

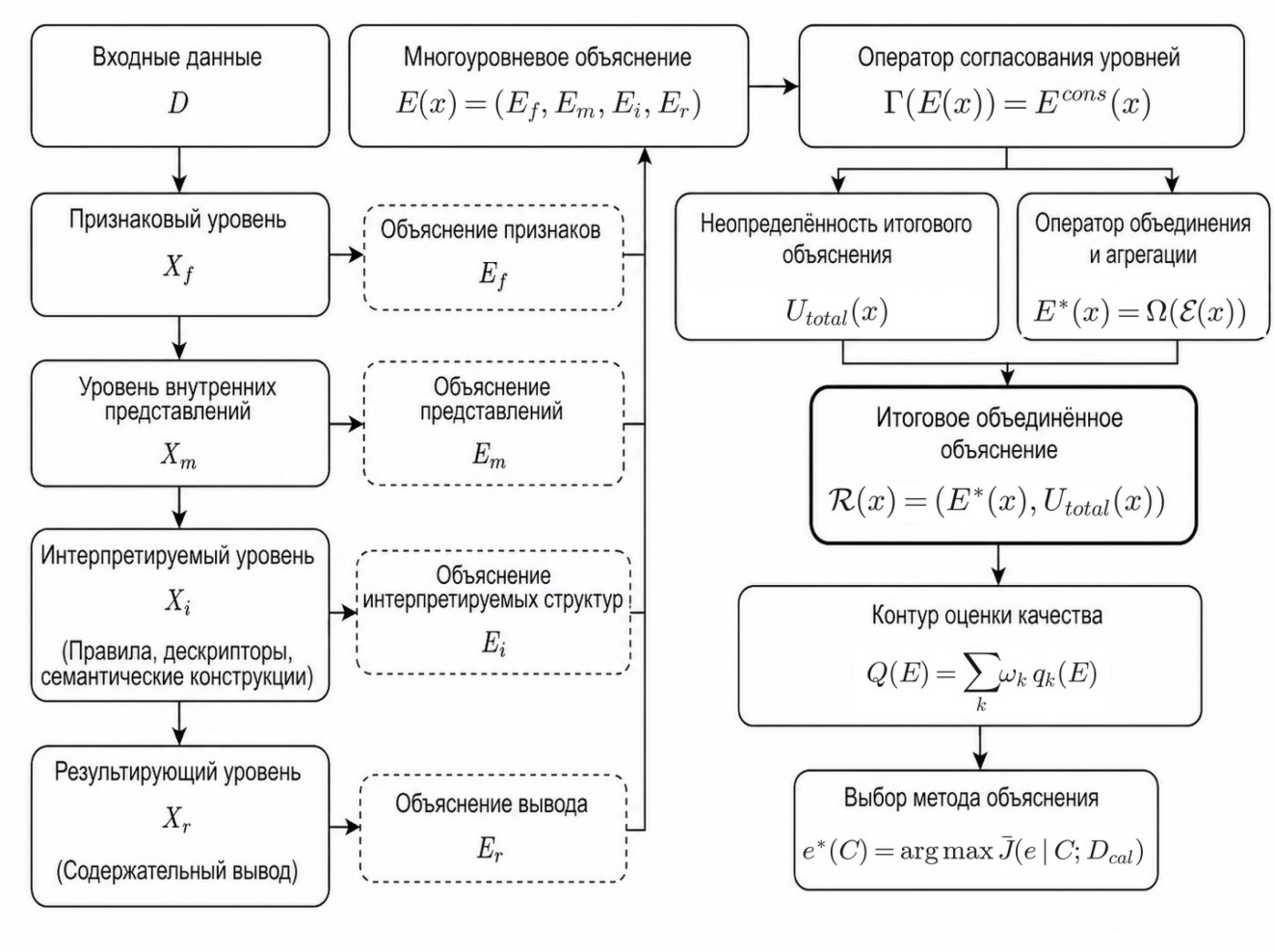


Рисунок 2.2 – Схема соответствия узлов ориентированного графа объяснимого контура обработки данных уровням многоуровневого объяснения



Рисунок 2.3 – Формальное представление узлов графа объяснения, пространств уровней и допустимых межуровневых переходов

У многоуровневости при таком представлении два аспекта. Один касается организации самого контура обработки данных и описывает структуру иерархической системы. Другой относится к тому, как представлены объяснительные объекты, формируемые или извлекаемые на узлах этого контура. Одну и ту же систему можно поэтому описать и как иерархический граф обработки данных, и как граф взаимосвязанных интерпретационных объектов.

Средства ОИИ служат здесь для построения локальных, межузловых и межуровневых представлений объяснения, привязанных к вершинам ориентированного графа.

Чтобы избежать смысловой неоднозначности, введём пространства уровней. Пусть каждому индексу из множества:

$$L = \{f, m, i, r\}, \tag{2.1}$$

соответствует собственное пространство объяснительных объектов, так что:

$$\mathcal{E}_L = \{\mathcal{E}_f, \mathcal{E}_m, \mathcal{E}_i, \mathcal{E}_r\}, \tag{2.2}$$

где $\mathcal{E}_f$ — пространство признаковых объяснений, $\mathcal{E}_m$ — модельных, $\mathcal{E}_i$ — интерпретируемых, $\mathcal{E}_r$ — результирующих [74, 132].

Каждый уровень даёт частичное описание одной и той же цепочки, по которой вычисляется вывод. На признаковом уровне видно, какие элементы входных данных оказались релевантны решению модели. Модельный уровень показывает, как эта релевантность проявляется во

внутренних состояниях вычислительного контура. Интерпретируемый переводит часть внутренней динамики в правила, концепты и другие содержательные дескрипторы, то есть в формы, принятые в моделях представления знаний [129]. Результирующий собирает полученные сведения в форму, пригодную для анализа, проверки и передачи пользователю.

Многоуровневое объяснение представим подграфом объяснительного контура:

$$G_E = (V_E, R_E), \tag{2.3}$$

где $V_E = \mathcal{E}_L$ — множество вершин, образованное пространствами уровней (2.2), а R_E — множество допустимых межуровневых переходов. Вершине сопоставляется объяснительный объект её уровня, поэтому граф G_E описывает и состав уровней, и то, как им разрешено взаимодействовать.

Взаимодействие вершин графа объяснения описывается отображениями между пространствами $\mathcal{E}_u$ и $\mathcal{E}_v$. Каждому допустимому ребру $(u, v) \in R_E$ отвечает оператор межуровневого перехода:

$$T_{u\to v}: \mathcal{E}_u \to \tilde{\mathcal{E}}_v, \tag{2.4}$$

переводящий объяснительный объект из одного уровня представления в другой. Здесь $\tilde{\mathcal{E}}_v$ — пространство представлений, сопоставимых с объектами уровня v. Под взаимодействием вершин понимается допустимое преобразование объяснений между пространствами, и одного соседства разных описаний в общем отчёте для него недостаточно.

Для пары связанных вершин u и v условие межуровневой совместимости записывается в общем виде как:

$$d_v\left(T_{u\to v}\big(E_u(x)\big), E_v(x)\right) \le \tau_{uv}, \tag{2.5}$$

где d_v — мера расхождения в пространстве $\mathcal{E}_v$, а τ_{uv} — допустимый порог межуровневого расхождения. Переход между уровнями допустим, если это неравенство выполнено.

Выбранный граф объяснения должен быть подграфом исходного объяснительного контура:

$$G_E^* \subseteq G_E, \tag{2.6}$$

причём этот подграф покрывает путь от входного описания к результирующему представлению и состоит только из допустимых межуровневых переходов.

2.3. Методы многоуровневого формирования атрибутивных объяснений

Граф объяснимого контура задан, и теперь нужно описать, какие объекты возникают на его узлах и рёбрах и как они попадают в структуру многоуровневого объяснения. Речь пойдёт об атрибутивных объектах, которые формируются или извлекаются на разных этапах работы интеллектуальной системы. Готовым итоговым объяснением такой объект ещё не является, но именно из этих объектов собираются признаковый, модельный, интерпретируемый и результирующий уровни.

Пусть иерархически организованная интеллектуальная система задана ориентированным графом:

$$G = (V, R), \tag{2.7}$$

где V — множество узлов, а $R \subseteq V \times V$ — множество допустимых переходов между ними. Каждому узлу $v \in V$ сопоставим множество его атрибутивных параметров:

$$P_v(x) = \{p_{v,j}(x)\}_{j=1}^{s_v}, \tag{2.8}$$

а каждому ребру $(u, v) \in R$ — множество атрибутивных параметров перехода

$$P_{u\to v}(x) = \{p_{u\to v,j}(x)\}_{j=1}^{t_{uv}}, \tag{2.9}$$

где x — анализируемый объект, s_v — число параметров, фиксируемых в узле v, а t_{uv} — число параметров, характеризующих переход от u к v. В эти множества попадают самые разные характеристики вычислительного процесса, от значений признаков и весов внимания до активаций внутренних слоёв и параметров правил. Объяснительную ценность имеет лишь часть из них, и её выделяют операторы отбора объяснительных атрибутов:

$$\Pi_{\text{expl}}^{(v)}: P_v(x) \to A_v(x), \Pi_{\text{expl}}^{(u\to v)}: P_{u\to v}(x) \to A_{u\to v}(x), \tag{2.10}$$

где $A_v(x)$ — атрибутивный объяснительный объект узла v, а $A_{u\to v}(x)$ — объект перехода. Проекциями в строгом линейно-алгебраическом смысле эти отображения быть не обязаны. От них требуется выделить ту часть локального состояния системы, которую дальше можно интерпретировать, согласовывать и сопоставлять с объектами других уровней.

Объяснительный объект на узле формируется отображением:

$$\Phi_v: \mathcal{X} \times Z_v \times \Theta_v \to \mathcal{A}_v, A_v(x) = \Phi_v(x, z_v, \theta_v), \tag{2.11}$$

где $\mathcal{X}$ — пространство входных объектов, Z_v — пространство внутренних состояний или выходов узла v, Θ_v — параметры модуля, реализующего узел, а $\mathcal{A}_v$ — пространство объяснительных объектов узла [54, 55].

Для ребра $(u, v) \in R$ отображение строится так же:

$$\Psi_{u\to v}: \mathcal{A}_u \times \mathcal{A}_v \times Z_u \times Z_v \times \Theta_{u\to v} \to \mathcal{A}_{u\to v}, \tag{2.12}$$

$$A_{u\to v}(x) = \Psi_{u\to v}(A_u(x), A_v(x), z_u, z_v, \theta_{u\to v}), \tag{2.13}$$

где $\Theta_{u\to v}$ — параметры преобразования между узлами. Отдельное отображение для рёбер нужно потому, что многоуровневое объяснение складывается из компонентов двух видов. Одни описывают локальные атрибуты вершин, другие характеризуют переходы между вершинами, и без вторых связь уровней терялась бы.

Все атрибутивные объекты, сформированные для x, объединяются в множество:

$$\mathcal{A}(x) = \{ A_v(x) \mid v \in V \} \cup \{ A_{u\to v}(x) \mid (u, v) \in R \}, \tag{2.14}$$

из которого затем формируются уровни объяснения. Известные методы ОИИ играют здесь роль частных механизмов, порождающих элементы $\mathcal{A}(x)$. Каждый из них, будь то *SHAP*, *LIME* или карта активации, даёт свой класс атрибутивных объектов, и все эти объекты проходят потом через общий контур межуровневого согласования [134].

Переход от этого множества к кортежу уровней выполняют операторы выделения уровней:

$$E_f(x) = \Pi_f\big(\mathcal{A}(x)\big), E_m(x) = \Pi_m\big(\mathcal{A}(x)\big), \tag{2.15}$$

$$E_i(x) = \Pi_i\big(\mathcal{A}(x)\big), E_r(x) = \Pi_r\big(\mathcal{A}(x)\big), \tag{2.16}$$

где $\Pi_f, \Pi_m, \Pi_i, \Pi_r$ — операторы выделения признакового, модельного, интерпретируемого и результирующего уровней. Реализация операторов зависит от архитектуры системы и типа данных, а назначение у них одно — разложить локально сформированные атрибутивные объекты по компонентам кортежа многоуровневого объяснения.

Атрибутивные объекты получают двумя путями. При встроенном формировании они возникают внутри архитектуры модели из её внутренних состояний, параметров и правил. Постмодельное извлечение выполняется после получения результата и строит приближённое описание поведения уже обученной модели. В первом случае объяснительный объект входит во внутреннюю динамику системы, во втором его восстанавливают по наблюдаемому отклику. С уровнями

объяснения эти пути отождествлять нельзя. Каждый из них может поставлять объекты на разные уровни, и ни один не исчерпывает многоуровневую структуру.

Объекты признакового уровня $\mathcal{E}_f$ чаще всего дают методы локальной атрибуции входных признаков. Базовым среди них остаётся *SHAP* [55], построенный на значениях Шепли из теории кооперативных игр. Для множества признаков $\mathcal{F}$ вклад признака в предсказание для объекта x равен:

$$\phi_j(x) = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!\,(|\mathcal{F}|-|S|-1)!}{|\mathcal{F}|!} \Big(f_{S\cup\{j\}}(x) - f_S(x) \Big), \tag{2.17}$$

здесь $f_S(x)$ — значение модели при доступном подмножестве признаков S, а сумма берётся по всем коалициям, не содержащим j-й признак. Величина $\phi_j(x)$ показывает, насколько в среднем меняется выход модели, когда к уже известным признакам добавляется j-й. Среднее берётся по всем подмножествам остальных признаков с весами из (2.17). Вклады всех признаков в сумме дают разность между предсказанием $f(x)$ и базовым значением ϕ_0, то есть средним выходом модели на фоновой выборке. Значение $\phi_j(x)$ читается как доля этой разности, приходящаяся на j-й признак, а по знаку вклада видно, повышает признак предсказание или понижает. Распределение вкладов по (2.17) аксиоматически корректно, и для объектов признакового уровня этот механизм подходит хорошо. Полного многоуровневого объяснения вектор $\phi(x)$ всё же не образует. Он фиксирует локальную релевантность входных факторов и ничего не говорит о том, как она преобразуется внутри модели и как соотносится с интерпретируемыми структурами более высокого уровня.

Широко используется для этого и *LIME* [54]. Метод подбирает в окрестности объекта x локальную интерпретируемую модель g из класса G:

$$g_x^* = \operatorname*{argmin}_{g \in G} \big(\mathcal{L}(f, g, \pi_x) + \Omega(g) \big), \tag{2.18}$$

где $\mathcal{L}(f, g, \pi_x)$ измеряет ошибку локальной аппроксимации исходной модели f моделью g с учётом весовой функции π_x, задающей близость к объекту x, а $\Omega(g)$ ограничивает сложность интерпретатора. Результат *LIME* тоже относится в первую очередь к признаковому уровню, ведь он описывает локально обозримую структуру значимых факторов. Завершённым объяснением его, как и вектор *SHAP*, считать нельзя, это лишь одна из форм атрибутивного описания входного пространства.

В глубоких моделях признаковую атрибуцию часто строят методом *Integrated Gradients* [62]. Для скалярного выхода модели f_c и базовой точки x' вклад j-го признака равен:

$$IG_j(x) = \left(x_j - x_j'\right) \int_0^1 \frac{\partial f_c\left(x' + \alpha(x - x')\right)}{\partial x_j}\, d\alpha. \tag{2.19}$$

Атрибуция здесь накапливается вдоль траектории от базовой точки к анализируемому объекту. Хотя в вычислении участвуют градиенты модели, результат описывает вклад входных компонент, и место ему на признаковом уровне. Внутреннюю структуру представлений он не характеризует, так что относить *Integrated Gradients* к модельному уровню без оговорок было бы ошибкой.

Модельный уровень $\mathcal{E}_m$ появляется, когда интерпретируются сами внутренние состояния вычислительного контура, его активации, карты внимания или скрытые представления. В свёрточных архитектурах для изображений так работает *Grad-CAM* [61], где карта значимости для класса c определяется формулой:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \tag{2.20}$$

где A^k — k-я карта активаций последнего свёрточного слоя, а коэффициенты вычисляются по формуле:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k}, \tag{2.21}$$

то есть α_k^c получается усреднением градиента выхода для класса c по пространственным координатам карты. Опорой интерпретации служит здесь внутреннее представление модели, входные признаки в ней напрямую не участвуют, и *Grad-CAM* поэтому относится к модельному уровню. До содержательных правил или дескрипторов предметной области такие объекты сами по себе не доходят. Они показывают лишь структуру внутренней активации, релевантной решению.

Интерпретируемый уровень $\mathcal{E}_i$ возникает, если часть вычислительного контура уже допускает прямое содержательное описание. Так обстоит дело в нейро-нечётких и других архитектурах, основанных на правилах [72, 135]. В *ANFIS*-подобной системе k-е правило типа Сугено [83, 84] записывается в виде:

$$R_k: x_1 \in A_1^k, \ldots, x_n \in A_n^k \Rightarrow y_k = p_{k1}x_1 + \cdots + p_{kn}x_n + r_k. \tag{2.22}$$

Степень активации правила и выход системы вычисляются как:

$$w_k = \prod_{j=1}^n \mu_{A_j^k}(x_j), \bar{w}_k = \frac{w_k}{\sum_l w_l}, y = \sum_k \bar{w}_k\, y_k, \tag{2.23}$$

здесь $\mu_{A_j^k}(x_j)$ — функция принадлежности j-го признака терму A_j^k в k-м правиле. Объяснительный объект получается сразу в виде правил и термов, соотносимых с предметной областью, и к постмодельной атрибуции отношения не имеет. Его место на интерпретируемом уровне.

Особый случай представляет объяснительная регуляризация, при которой атрибутивные объекты участвуют в обучении модели, а не только в интерпретации готовой. Функционал обучения в общем виде принимает форму:

$$\mathcal{L}(\theta) = \mathcal{L}_{pred}(\theta) + \lambda\, \mathcal{L}_{expl}(\theta), \tag{2.24}$$

где $\mathcal{L}_{pred}$ отвечает за точность предсказания, $\mathcal{L}_{expl}$ — за выполнение заранее заданного объяснительного ограничения, а $\lambda \geq 0$ — относительный вес ограничения. Вид $\mathcal{L}_{expl}$ зависит от типа модели, природы объяснительного объекта и того, как задана референтная интерпретация. Запись (2.24) описывает целый класс постановок, где объяснимость перенесена внутрь оптимизационного контура [47, 136], и конкретного алгоритма не предписывает.

Если в архитектуре есть отдельный объяснительный модуль, параметры базового предиктора и объяснительного блока могут обновляться совместно по правилу:

$$\theta_{base}^{(t+1)} = \theta_{base}^{(t)} - \eta\, \frac{\partial \mathcal{L}}{\partial \theta_{base}}, \theta_{expl}^{(t+1)} = \theta_{expl}^{(t)} - \eta\, \frac{\partial \mathcal{L}}{\partial \theta_{expl}}, \tag{2.25}$$

где η — шаг оптимизации. Двухблочная запись имеет смысл лишь тогда, когда отдельный объяснительный модуль действительно есть. Если его нет, хватает общей формы $\mathcal{L}(\theta)$, и делить параметры на два подмножества искусственно незачем. Объяснительная регуляризация считается в работе дополнительным режимом формирования атрибутивных объектов, обязательной частью ОИИ-системы она не является.

Механизмы различаются тем, откуда берутся атрибутивные объекты, а порядок построения многоуровневого объяснения у них общий. Всё начинается с графа (2.7), в котором фиксируются узлы обработки и допустимые переходы. Для узлов и переходов формируются множества $P_v(x)$ и $P_{u\to v}(x)$ по (2.8) и (2.9), и операторы (2.10) оставляют в них только параметры, которые можно интерпретировать и сопоставлять с объектами других уровней.

Затем по отображениям (2.11) и (2.12)–(2.13) строятся объекты узлов $A_v(x)$ и переходов $A_{u\to v}(x)$. На этом шаге и работают конкретные методы ОИИ, каждый на своём уровне, а при объяснительной регуляризации (2.24) часть объектов появляется ещё во время обучения. Узловые и межузловые объекты собираются в множество $\mathcal{A}(x)$ по (2.14), операторы (2.15)–(2.16) распределяют его элементы по четырём уровням, и полученный кортеж поступает на межуровневое согласование.

2.4. Метод межуровневого согласования объяснений

Уровни выделены, но из этого ещё не следует, что они складываются в одно объяснение. Атрибуция признаков, внутренние представления модели, правила и итоговый текст могут относиться к одному объекту и при этом расходиться друг с другом. Между уровнями нужна связь, которую можно проверить формально, и её устанавливает оператор согласования уровней. Он переводит исходный кортеж в форму, пригодную для целостной интерпретации, повторной проверки и объединения с другими объяснениями.

Пусть многоуровневое объяснение задано кортежем:

$$E(x) = \big(E_f(x), E_m(x), E_i(x), E_r(x)\big) \in \mathcal{E}_f \times \mathcal{E}_m \times \mathcal{E}_i \times \mathcal{E}_r, \tag{2.26}$$

где $\mathcal{E}_f, \mathcal{E}_m, \mathcal{E}_i, \mathcal{E}_r$ — пространства объектов признакового, модельного, интерпретируемого и результирующего уровней. Оператор согласования отображает произведение этих пространств само в себя:

$$\Gamma\colon \mathcal{E}_f \times \mathcal{E}_m \times \mathcal{E}_i \times \mathcal{E}_r \to \mathcal{E}_f \times \mathcal{E}_m \times \mathcal{E}_i \times \mathcal{E}_r, \tag{2.27}$$

$$E^{cons}(x) = \Gamma\big(E(x)\big), \tag{2.28}$$

где $E^{cons}(x)$ — согласованное многоуровневое объяснение. В нём те же четыре уровня, что и в $E(x)$, но компоненты скорректированы так, чтобы суммарное межуровневое расхождение было минимальным. Согласованным кортеж считается тогда, когда для каждой пары связанных уровней выполнено условие совместимости (2.5) и расхождение не выходит за порог τ_{uv}. На объединение и оценку неопределённости передаётся уже согласованный кортеж (см. рисунок 2.4). Насколько согласование удалось, показывает индекс $Cons(x)$, собранный из локальных индексов для пар уровней.

Одной вычислительной операцией оператор согласования не исчерпывается. За ним стоит класс процедур, куда входят нормализация представлений и приведение к сопоставимым шкалам, построение межуровневых отображений, измерение расхождения и корректировка формы описания [118, 137]. Нужен он из-за неоднородности уровней. Если $E_f(x)$ — вектор важностей признаков, $E_m(x)$ — тензор активаций, $E_i(x)$ — набор нечётких правил, а $E_r(x)$ — текстовый отчёт, сравнить их напрямую нельзя, и сопоставимыми их делает только система межуровневых отображений.

Согласование решений, принимаемых на разных уровнях иерархии, давно изучается в теории систем. Принципы координации многоуровневых систем сформулированы в [125], методы

согласования иерархических решений систематизированы в [138]. Оператор Γ переносит эту задачу с решений на объяснения. Согласовываются интерпретационные описания одного и того же вывода, полученные на разных уровнях, и мерой согласованности служит межуровневое расхождение вместо значения общей целевой функции системы.

Для каждой допустимой пары уровней (u, v), $u, v \in L$, используется оператор переноса $T_{u\to v}$ из (2.4). Обратимым он быть не обязан, его дело — перенести содержание объекта уровня u в форму, сопоставимую с объектом уровня v. Согласование складывается поэтому из множества локальных переходов, и каждый из них проверяется отдельно. Для перехода фиксируются оператор переноса, мера расхождения и порог допустимости (см. рисунок 2.4).

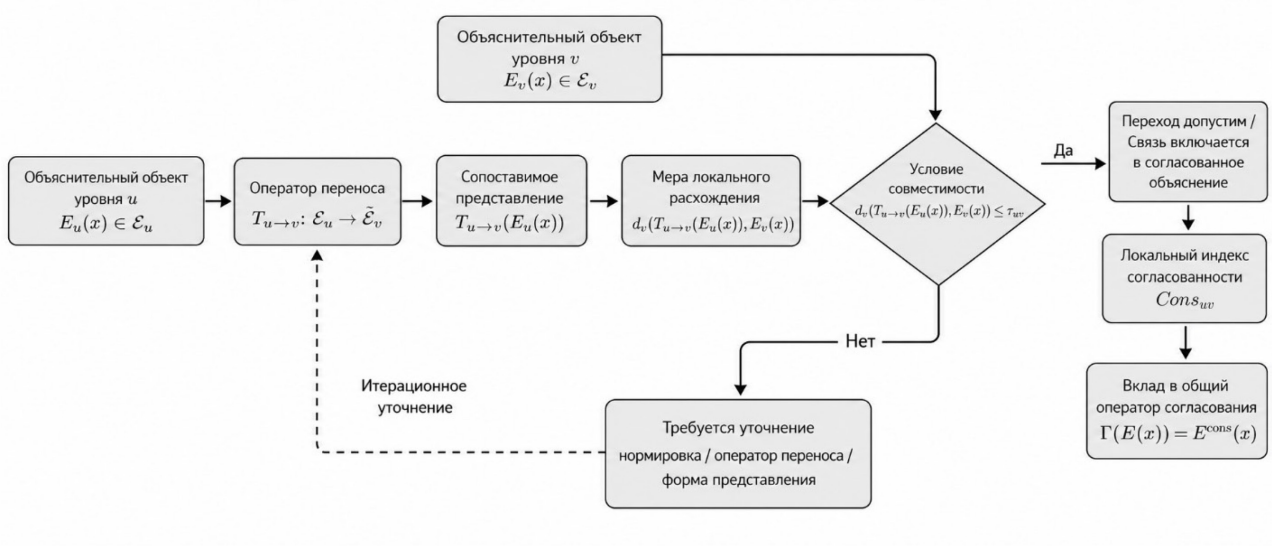


Рисунок 2.4 – Структурная схема локального межуровневого согласования объяснительных объектов

Локальное условие совместимости решает, войдёт ли связь двух уровней в согласованное объяснение. Если порог превышен, уточняют правила нормировки, оператор переноса или форму представления. Все локальные проверки вместе сводятся к задаче минимизации суммарного межуровневого расхождения:

$$\Gamma\big(E(x)\big) = \underset{\hat{E}}{\operatorname{argmin}} \sum_{(u,v)\in R_E} \omega_{uv}\ d_{uv}\left(T_{u\to v}\left(\hat{E}_u(x)\right), \hat{E}_v(x)\right), \tag{2.29}$$

где R_E — множество согласуемых пар уровней, то есть допустимых межуровневых переходов графа (2.3), $\omega_{uv} \geq 0$ — веса связей, d_{uv} — мера расхождения для пары (u, v), а $\hat{E} = \left(\hat{E}_f, \hat{E}_m, \hat{E}_i, \hat{E}_r\right)$ — искомый согласованный кортеж. Согласование сводится здесь к поиску представления, в котором расхождения по фиксированному набору переходов минимальны [113, 114].

Признаки, которые атрибутивный метод счёл значимыми на уровне $E_f(x)$, после переноса на модельный уровень должны попасть в те области внутренних представлений, что действительно поддерживают решение модели. Правило или концепт из $E_i(x)$ обязано согласовываться и с признаковыми атрибуциями, и с внутренними состояниями, на которых оно активируется. Основания решения, о которых результирующее объяснение $E_r(x)$ сообщает пользователю, должны подтверждаться нижними уровнями, а не появляться впервые при вербализации. Набор разрозненных интерпретаций этим требованиям, как правило, не удовлетворяет.

Вид меры d_{uv} определяется природой объектов. Векторные атрибуции сравнивают метрически или по ранговой корреляции, пространственные карты — по пересечению и структурному сходству, правила и концепты — по совпадению условий и активированных признаков. Единой метрики на все случаи у оператора Γ нет. Общим остаётся требование, чтобы согласованность была вычислимой и воспроизводимой, а переходы, меры расхождения и условия применения были явно зафиксированы для конкретного вычислительного контура.

На практике общий оператор дополняют локальными индексами согласованности. Пусть на признаковом уровне выделено множество наиболее значимых факторов:

$$S_f(x) \subseteq \mathcal{F}, \tag{2.30}$$

куда входят n_f признаков с наибольшими $|\phi_j(x)|$ или все признаки, у которых $|\phi_j(x)|$ превышает порог θ_f. Значения n_f и θ_f фиксируются протоколом оценки, а выделение такого подмножества — классическая задача отбора информативных признаков [139]. На интерпретируемом уровне берутся правила базы (2.22), активированные для объекта x со степенью $w_k(x)$ из (2.23) не ниже порога $w_{\min}$:

$$R(x) = \{r_1, \dots, r_K\}. \tag{2.31}$$

Обозначим через $\mathrm{Feat}(R(x)) \subseteq \mathcal{F}$ признаки, входящие в условия активированных правил. Согласованность признакового и интерпретируемого уровней измеряется индексом Жаккара:

$$\mathrm{Cons}_{fi}(x) = \frac{|S_f(x) \cap \mathrm{Feat}(R(x))|}{|S_f(x) \cup \mathrm{Feat}(R(x))|}, \tag{2.32}$$

который принимает значения в $[0,1]$. При $\mathrm{Cons}_{fi}(x) \approx 1$ значимые по атрибуции признаки в основном поддерживаются структурой правил. Малое значение говорит о том, что правило и атрибуция опираются на разные основания решения и единого объяснения не дают.

Так же строятся показатели $\mathrm{Cons}_{fm}(x)$, $\mathrm{Cons}_{mi}(x)$, $\mathrm{Cons}_{ir}(x)$ для тех пар уровней, которые есть в конкретной архитектуре. Общий индекс внутренней согласованности равен их взвешенному среднему:

$$Cons(x) = \sum_{(u,v)\in R_E} \alpha_{uv}\ Cons_{uv}(x), \alpha_{uv} \geq 0, \sum_{(u,v)\in R_E} \alpha_{uv} = 1, \qquad (2.33)$$

который служит сводной мерой целостности многоуровневого объяснения. Через него согласованность учитывается и при оценке качества, и при выборе объяснителя. Оператор Γ приводит уровни к совместимому виду, а индекс количественно показывает, насколько единым получилось объяснение [137].

Реализуется оператор Γ итерационно. Сначала объекты согласуемых уровней нормируются и приводятся к общим шкалам и соглашениям представления, без чего вектор важностей, карту активаций и набор правил не сравнить. Для каждой пары $(u, v) \in R_E$ задаются оператор переноса (2.4), мера расхождения d_{uv} и порог τ_{uv} из (2.5), то есть элементы схемы на рисунке 2.4. Далее по каждой паре вычисляются расхождения, входящие в (2.29), и локальные индексы вида (2.32). Если условие (2.5) для какой-то пары нарушено, уточняют нормировку, оператор переноса или форму представления и пересчитывают расхождения. Цикл заканчивается, когда условия совместимости выполнены для всех пар из R_E. Результатом становится кортеж $E^{cons}(x)$ по (2.28)–(2.29), а достигнутую согласованность выражает индекс $Cons(x)$ по (2.33).

В мультимодальных и гибридных системах без итераций не обойтись, потому что ошибка сопоставления на раннем шаге искажает всю дальнейшую интерпретацию.

У согласования есть и отдельная ценность для воспроизводимости. Пока правила нормировки, переноса и сопоставления не зафиксированы, два исследователя с одной моделью и одним объектом вполне могут получить разные многоуровневые описания, и виновата в этом будет не модель, а протокол интерпретации. С оператором Γ сопоставление уровней становится явной процедурой, которую можно проверить и повторить.

Сильнее всего это заметно в гибридных архитектурах, где нейросетевой контур соседствует с продукционным. Без Γ описание такой системы распадается на два слабо связанных фрагмента. В одном остаются внутренние сигналы глубокой модели, в другом — правила и содержательные дескрипторы. Согласование сшивает их в один объяснительный контур, и как раз оно отделяет многоуровневое описание от многоуровневого объяснения.

2.5. Методы агрегирования, объединения и комплексирования при ансамблировании гетерогенных объяснений

Для одного объекта обычно строят не одно объяснение. Их дают разные методы, модели, запуски и модальности, и каждое уже согласовано внутри себя. Предметом анализа становятся теперь сами согласованные объяснения, которые нужно обрабатывать совместно. По аналогии с ансамблями в машинном обучении, где несколько моделей вместе дают более устойчивое и точное решение, такую обработку будем называть ансамблированием объяснений. Идея близка к множественному описанию объекта в теории принятия решений, где сочетание нескольких методов и систем критериев даёт синергетический эффект и общий результат оказывается богаче суммы частных оценок [140].

Способ совместной обработки зависит от природы объяснительных объектов. Однородные числовые представления можно агрегировать в общем пространстве. Разнородные объяснения лучше хранить как типизированные свидетельства, у каждого из которых указаны происхождение, уровень, мера согласия и неопределённость. Отсюда три разные операции (агрегирование, объединение и комплексирование), которые нужно строго различать.

Пусть для объекта x построено множество согласованных объяснений:

$$\mathcal{E}(x) = \left\{E^{(1)}(x), E^{(2)}(x), \dots, E^{(N_E)}(x)\right\}, \tag{2.34}$$

где каждое $E^{(k)}(x)$ прошло межуровневое согласование, индекс k указывает источник объяснения, а N_E — число частных объяснений. Здесь $E(x)$ обозначает множество объяснений, его не следует путать с кортежем уровней $\mathcal{E}(x)$ из (2.26). Требуется оператор, который переводит это множество в итоговое представление более высокого уровня:

$$\Omega: \mathcal{P}(\mathcal{E}) \to \mathcal{Y}, E^{\Omega}(x) = \Omega\big(\mathcal{E}(x)\big), \tag{2.35}$$

где $\mathcal{P}(\mathcal{E})$ — семейство допустимых конечных наборов согласованных объяснений, $\mathcal{Y}$ — класс допустимых итоговых объектов, $E^{\Omega}(x)$ — результат совместной обработки. Вид результата зависит от режима обработки. В класс $\mathcal{Y}$ входят многоуровневые объяснения вида (2.26), элементы общего пространства, где агрегируются однородные объяснения, и составные объекты, в которых разнородность исходных объяснений сохранена. Обозначение $E^{*}(x)$ закрепим только за результатом режима объединения.

Оператор Ω — общее обозначение, за которым стоят три режима. Агрегирование применяют, если все частные объяснения после приведения укладываются в одно пространство и сравнимы по одной шкале. Объединение нужно, когда итог строится из нескольких объяснений и важно не потерять сведения об их происхождении, весе и роли. К комплексированию прибегают при принципиально разнородных объяснениях, если эту разнородность надо сохранить и не сводить всё к единому носителю.

Агрегирование возможно при наличии общего пространства представления $\mathcal{Z}$ и семейства операторов приведения:

$$\zeta_k: \mathcal{E}^{(k)} \to \mathcal{Z}, k = 1, \dots, N_E. \tag{2.36}$$

С их помощью частные объяснения переводятся в сопоставимую форму:

$$z_k(x) = \zeta_k \left(E^{(k)}(x) \right) \in \mathcal{Z}, \tag{2.37}$$

и агрегированное объяснение получается оператором:

$$E^{agg}(x) = Agg \left(z_1(x), \dots, z_{N_E}(x) \right). \tag{2.38}$$

Для однородных числовых объектов естественная форма агрегирования — взвешенное среднее:

$$E^{\text{agg}}(x) = \sum_{k=1}^{N_E} w_k \ z_k(x), w_k \geq 0, \sum_{k=1}^{N_E} w_k = 1. \tag{2.39}$$

Формула (2.39) корректна, лишь когда все $z_k(x)$ лежат в одном пространстве, имеют одинаковую размерность и приведены к единой шкале. Для произвольных объяснений она не годится.

Если привести объяснения к общему виду нельзя или не нужно, агрегирования мало, а итог не должен стирать различия между формами представления. В таком случае применяют объединение, при котором частные объяснения образуют структурированный ансамбль:

$$E^*(x) = \Omega\big(\mathcal{E}(x), Meta(x)\big), \tag{2.40}$$

где $Meta(x)$ — метаданные происхождения. В них записаны тип метода и уровень объяснения, параметры запуска и модальность, а также показатели качества и надёжности частных компонентов [95, 96].

Объединение естественно вести по уровням. Современные методы ОИИ дают объяснения, различающиеся и по содержательному уровню, и по форме, в которой результат показывают

пользователю [141], и поуровневая процедура эту неоднородность учитывает. Она согласуется и с подходами, где итоговое объяснение собирают из нескольких частных интерпретаций по набору критериев качества [117]. Пусть каждое частное объяснение представлено кортежем:

$$E^{(k)}(x) = \left(E_f^{(k)}(x), E_m^{(k)}(x), E_i^{(k)}(x), E_r^{(k)}(x)\right), \tag{2.41}$$

тогда и объединённое объяснение записывается покомпонентно:

$$E^*(x) = \left(E_f^*(x), E_m^*(x), E_i^*(x), E_r^*(x)\right), \tag{2.42}$$

и для каждого уровня вводится свой оператор объединения:

$$E_f^*(x) = \Omega_f\left(E_f^{(1)}(x), \dots, E_f^{(N_E)}(x)\right), \tag{2.43}$$

$$E_m^*(x) = \Omega_m\left(E_m^{(1)}(x), \dots, E_m^{(N_E)}(x)\right), \tag{2.44}$$

$$E_i^*(x) = \Omega_i\left(E_i^{(1)}(x), \dots, E_i^{(N_E)}(x)\right), \tag{2.45}$$

$$E_r^*(x) = \Omega_r\left(E_r^{(1)}(x), \dots, E_r^{(N_E)}(x)\right). \tag{2.46}$$

Разложение не даёт смешивать объекты разной природы и сохраняет связь с многоуровневой структурой. Оператор Ω_f строится на усреднении, в том числе робастном, или на ранговой агрегации. Оператор Ω_m работает с картами внимания и активационными структурами, оператор Ω_i — с правилами, концептами и контрфактуальными объектами. От Ω_r зависит, в каком виде объединённое объяснение увидит пользователь. Общая схема совместной обработки частных объяснений приведена на рисунке 2.5.

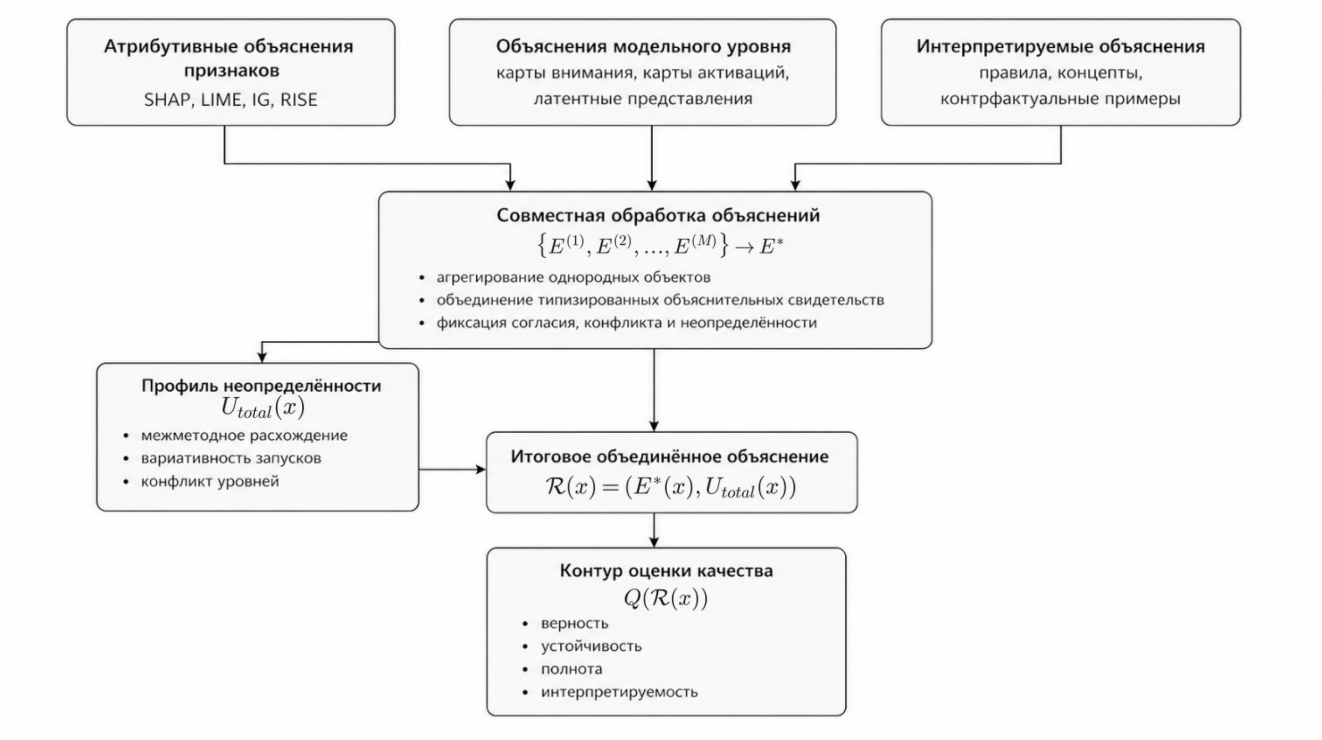


Рисунок 2.5 – Формализованный контур агрегирования, объединения и сопоставимой оценки гетерогенных объяснений

Объяснения различаются качеством и устойчивостью, и это должно отражаться в весах w_k из (2.39). Веса строятся по протокольной оценке качества частного объяснения и мере его неопределённости:

$$w_k = \frac{e^{\alpha Q\left(E^{(k)}(x)\right) - \beta U\left(E^{(k)}(x)\right)}}{\sum_{j=1}^{N_E} e^{\alpha Q\left(E^{(j)}(x)\right) - \beta U\left(E^{(j)}(x)\right)}}, \tag{2.47}$$

где $Q\left(E^{(k)}(x)\right)$ — качество k-го частного объяснения, $U\left(E^{(k)}(x)\right)$ — мера его неопределённости (способ её оценки описан в разделе 2.6), а $\alpha, \beta \geq 0$ задают относительное влияние этих факторов. Нормировка в (2.47) даёт неотрицательные веса, для которых

$$\sum_{k=1}^{N_E} w_k = 1, \tag{2.48}$$

так что их можно прямо подставлять в (2.39).

Структурные и интерпретируемые объяснения объединяются через оптимизацию. Если $\mathcal{H}(x)$ — множество допустимых кандидатов на итоговую интерпретируемую структуру, объединённый интерпретируемый объект находится как:

$$E_i^*(x) = \operatorname{argmax}_{H \in \mathcal{H}(x)} \Big(\operatorname{Supp}\big(H \mid \mathcal{E}(x)\big) + \lambda\, \operatorname{Cons}(H) - \mu\, \operatorname{Comp}(H) \Big), \tag{2.49}$$

где $\mathrm{Supp}\big(H \mid \mathcal{E}(x)\big)$ измеряет поддержку кандидата H со стороны входящих объяснений, $\mathrm{Cons}(H)$ характеризует его внутреннюю согласованность, $\mathrm{Comp}(H)$ штрафует за чрезмерную сложность, а $\lambda, \mu \geq 0$ задают относительный вклад этих факторов.

Ранговые объяснения агрегируются по правилам коллективного ранжирования. Если $\mathrm{rk}_k(a)$ — ранг элемента a в k-м объяснении, суммарный балл по правилу Борда [115, 116] равен:

$$B(a) = \sum_{k=1}^{N_E} \big(N - \mathrm{rk}_k(a)\big), \tag{2.50}$$

где N — число ранжируемых элементов, общее для всех частных объяснений. Итоговое ранжирование идёт по убыванию $B(a)$. Когда источники ранжирования сами образуют иерархию, агрегируют поуровнево, от нижних уровней к верхним [142].

Третий режим, комплексирование, ничего не усредняет и не ищет единственную сводную структуру. Его задача — сохранить содержательно неоднородный результат в форме, пригодной для анализа и аудита. Результатом комплексирования служит составной объяснительный объект:

$$\Xi(x) = \big(Core(x), \mathcal{S}(x), \mathcal{D}(x), Meta(x)\big), \tag{2.51}$$

где $\mathrm{Core}(x)$ — устойчивое ядро объяснения, $\mathcal{S}(x)$ — множество компонентов, поддерживающих ядро, $\mathcal{D}(x)$ — множество компонентов, расходящихся с ним, а $\mathrm{Meta}(x)$ — метаданные происхождения. В такой записи видна и итоговая интерпретация, и то, где объяснения согласны, а где конфликтуют.

В продукционных и нечётких контурах объединение делают мягким. Два правила, расходящиеся лишь в деталях термов, порогов или интервалов, противоречащими не считаются, и для них вводится функция близости:

$$\rho(r_a, r_b) \in [0,1]. \tag{2.52}$$

Если $\rho(r_a, r_b) \geq \tau_\rho$, где τ_ρ — порог близости, правила могут включаться в одно устойчивое ядро с последующим уточнением термов и порогов [72, 81].

Совместная обработка начинается с подготовки частных объяснений. Каждое объяснение из (2.34) нормируется и снабжается метаданными $Meta(x)$ из (2.40), а где возможно, переводится в общее пространство $\mathcal{Z}$ операторами (2.36)–(2.37). Удалось ли это сделать, определяет режим. Объяснения, приведённые к одному пространству, размерности и шкале, агрегируются по (2.38). Если привести их нельзя, но происхождение и роль компонентов терять не хочется, выполняется поуровневое объединение по (2.40) и (2.43)–(2.46). Принципиально разнородные объяснения, разнородность которых нужно сохранить, комплексируются.

Далее вычисляются меры близости и согласованности между объяснениями, для правил это близость (2.52), и по (2.47) назначаются веса. При агрегировании результатом становится $E^{agg}(x)$ или итоговое ранжирование по (2.50), при объединении — кортеж $E^{*}(x)$ вида (2.42), при комплексировании — составной объект $\Xi(x)$ по (2.51).

Пользователю объединённое объяснение $E^{*}(x)$ показывают в виде пары взаимосвязанных компонентов:

$$E^{\text{out}}(x) = \left(E^{\text{core}}(x), E^{\text{view}}(x)\right), \tag{2.53}$$

где $E^{\text{core}}(x)$ — ядро объяснения, прошедшее согласование и метрическую проверку, а $E^{\text{view}}(x)$ — форма, в которой это ядро предъявляется.

2.6. Метод квантификации неопределённости объединённого объяснения

Насколько можно доверять итоговому объяснению, зависит от того, совпадают ли его источники. Когда методы, запуски и модели дают близкие интерпретации, объяснение устойчиво. При заметном расхождении неопределённость приходится явно вписывать в описание результата.

Пусть $E^{*}(x)$ — объединённое объяснение объекта x. Его неопределённость задаётся функционалом от расхождений между частными объяснениями и объединённым представлением:

$$U(x) = \mathcal{U}\left(E^{(1)}(x), \dots, E^{(N_E)}(x);\ E^{*}(x)\right). \tag{2.54}$$

Мерой неопределённости служит наблюдаемый разброс частных объяснений вокруг итогового. Чем ближе они друг к другу, тем меньше $U(x)$, а при расхождении границы доверия расширяются.

Для однородных объектов, приведённых к общему пространству $\mathcal{Z}$, берётся среднее квадратичное расхождение:

$$U_{\text{var}}(x) = \frac{1}{N_E}\sum_{k=1}^{N_E} d^2\left(z_k(x), z^{*}(x)\right), \tag{2.55}$$

где $z_k(x) \in \mathcal{Z}$ — приведённое частное объяснение, $z^{*}(x) \in \mathcal{Z}$ — его объединённое представление, а d — метрика или псевдометрика на $\mathcal{Z}$, вид которой выбирают по типу данных. Формула (2.55) подходит для любых объяснений, допускающих единое числовое представление [143].

Покомпонентно для вектора важностей признаков неопределённость оценивается дисперсией:

$$U_j(x) = \frac{1}{N_E}\sum_{k=1}^{N_E}(\, z_{k,j}(x) - z_j^*(x))^2, j = 1, \dots, p, \tag{2.56}$$

где p — число компонент объяснительного вектора, а интегральная локальная неопределённость равна среднему по компонентам:

$$U_{\text{loc}}(x) = \frac{1}{p}\sum_{j=1}^{p} U_j\,(x). \tag{2.57}$$

Дисперсионная оценка универсальна, но к выбросам чувствительна. Одного аномального объяснителя достаточно, чтобы исказить итог. Робастная альтернатива строится на медианном абсолютном отклонении расстояний до итогового представления:

$$U_{\text{mad}}(x) = \text{median}_k|d(z_k(x), z^*(x)) - \text{median}_l\, d(z_l(x), z^*(x))|. \tag{2.58}$$

От единичных аномальных объяснений эта величина зависит слабо и лучше подходит для ансамблей, где часть методов ведёт себя нестабильно.

Если пользователю показывают точечную оценку вместе с диапазоном устойчивости, неопределённость описывается эмпирическим интервалом. Доверительным интервалом в статистическом смысле он не является, поскольку отражает разброс объяснений внутри ансамбля, а не точность оценки какого-либо параметра. Для компоненты j интервал имеет вид:

$$I_j(x) = \left[Q_{\nu/2}(\{z_{k,j}(x)\}_{k=1}^{N_E}), Q_{1-\nu/2}(\{z_{k,j}(x)\}_{k=1}^{N_E})\right], \tag{2.59}$$

где Q_γ — эмпирическая γ-квантиль, а $\nu \in (0,1)$ — уровень, задающий ширину интервала. Итоговое объяснение можно показывать точечным объектом $E^*(x)$ или набором локальных интервалов устойчивости.

В мультимодальных системах у неопределённости есть ещё один источник. Объяснение может быть стабильным внутри каждой модальности и всё равно конфликтовать между модальностями, так что к внутримодальной вариативности добавляется межмодальная согласованность [118, 144–146].

Если $a_m(x)$ — атрибутивное представление m-й модальности, мера межмодальной когерентности равна:

$$\text{Coh}(x) = \frac{2}{N_{\text{mod}}(N_{\text{mod}}-1)}\sum_{1\le m<l\le N_{\text{mod}}} \text{Sim}\left(a_m(x), a_l(x)\right), \tag{2.60}$$

где $N_{\mathrm{mod}} \geq 2$ — число модальностей, а Sim — мера сходства со значениями в $[0,1]$, выбираемая в соответствии с природой данных.

Для временных сигналов роль Sim играет нормированная взаимная информация или другая мера согласованности временных паттернов [147–149]. Межмодальная неопределённость определяется как:

$$U_{\mathrm{mod}}(x) = 1 - \mathrm{Coh}(x). \tag{2.61}$$

Большое $U_{\mathrm{mod}}(x)$ говорит о том, что каналы данных поддерживают разные объяснительные гипотезы, и в итоговом отчёте это должно быть видно.

Интегральная неопределённость объединённого объяснения складывается по меньшей мере из трёх источников и представляется аддитивной декомпозицией:

$$U_{\mathrm{total}}(x) = \lambda_1 U_{\mathrm{method}}(x) + \lambda_2 U_{\mathrm{run}}(x) + \lambda_3 U_{\mathrm{model}}(x), \lambda_i \geq 0, \sum_{i=1}^{3} \lambda_i = 1, \tag{2.62}$$

где $U_{\mathrm{method}}(x)$ — межметодная неопределённость, отражающая расхождение между различными объяснителями, $U_{\mathrm{run}}(x)$ — внутриметодная неопределённость, то есть вариативность повторных запусков одного метода при фиксированной модели, а $U_{\mathrm{model}}(x)$ — межмодельная неопределённость, то есть расхождение между объяснениями, построенными для разных, но сопоставимо качественных моделей.

Каждое слагаемое вычисляется по (2.55) или (2.58) на своём подмножестве частных объяснений. Межмодальная неопределённость (2.61) в декомпозицию (2.62) не входит. Она описывает конфликт каналов данных, а не источников объяснений, и в отчёте приводится отдельно. Неопределённость частного объяснения $U\left(E^{(k)}(x)\right)$, входящая в веса (2.47), оценивается теми же мерами (2.55) и (2.58) по повторным запускам одного метода.

Разложение подсказывает, где искать причину. Большое $U_{\mathrm{run}}(x)$ требует стабилизировать процедуру построения объяснения и жёстче зафиксировать протокол. Преобладание $U_{\mathrm{method}}(x)$ означает, что объяснители расходятся по интерпретационному смыслу. Когда же основной вклад даёт $U_{\mathrm{model}}(x)$, дело в множественности самих прогнозов, и способ интерпретации здесь вторичен.

На практике итоговое объяснение записывается парой:

$$\mathcal{R}(x) = \left(E^{*}(x), U_{\mathrm{total}}(x)\right), \tag{2.63}$$

где $E^{*}(x)$ — объединённое и согласованное объяснение, а $U_{total}(x)$ — его интегральная неопределённость. Вместе с содержанием объяснение несёт протокольно заданную оценку своей надёжности.

Числовую оценку полезно дополнить качественными режимами. При малом $U_{total}(x)$ на объяснение можно опираться при анализе решения. Промежуточные значения не лишают его информативности, но требуют пометки о неполной устойчивости. При больших значениях система явно фиксирует конфликт между источниками интерпретации, и самостоятельным основанием для вывода такое объяснение служить не может.

Квантификация начинается с группировки. По метаданным $Meta(x)$ из (2.40) частные объяснения из (2.34) делятся на полученные разными методами, полученные повторными запусками одного метода и построенные для разных моделей сопоставимого качества, и каждая группа даёт своё слагаемое в (2.62). Внутри групп расхождение с объединённым представлением вычисляется по (2.55), для оценки по признакам используются (2.56)–(2.57), а при нестабильных объяснителях дисперсионную оценку заменяют робастной (2.58). В мультимодальной системе дополнительно считаются когерентность (2.60) и межмодальная неопределённость (2.61).

Групповые оценки сводятся в $U_{total}(x)$ по (2.62), и по соотношению слагаемых видно, что нужно дорабатывать — протокол построения объяснения, набор объяснителей или сами модели. Итог записывается парой (2.63) и относится к одному из качественных режимов. Если объяснение предъявляется с диапазоном устойчивости, к нему добавляются интервалы (2.59), а в мультимодальном случае U_{mod} выносится в отчёт отдельно.

Пара (2.63) поступает на вход метрического контура оценки качества. Это отвечает современному взгляду на ОИИ, где форма объяснения важна не меньше, чем критерии его оценки, сопоставления и выбора [117, 141].

2.7. Метрический контур сопоставимой оценки качества объяснений

Наглядное и правдоподобное объяснение может плохо соответствовать поведению модели при контролируемых проверках. Чтобы это обнаружить, нужен метрический контур, который переводит результат $\mathcal{R}(x)$ из (2.63) в вектор частных показателей и интегральную оценку качества. Одним числом качество объяснения не описать. К его значимым характеристикам относят верность, устойчивость, простоту, полноту и практическую применимость [141], и в соответствии с этим качество задаётся вектором частных критериев:

$$q(E) = \Big(q_{\mathrm{fid}}(E),\ q_{\mathrm{stab}}(E), q_{\mathrm{comp}}(E), q_{\mathrm{simp}}(E),\ q_{\mathrm{cost}}(E)\Big), \tag{2.64}$$

где E — согласованное и объединённое объяснение, q_{fid} — верность, q_{stab} — устойчивость, q_{comp} — полнота, q_{simp} — простота, q_{cost} — показатель вычислительной стоимости. Все показатели

приведены к [0, 1] и ориентированы одинаково, так что большее значение означает лучшее объяснение, а для стоимости — меньшие затраты.

Верность $q_{\mathrm{fid}}(E)$ показывает, насколько объяснение согласуется с поведением модели при контролируемых вмешательствах и проверках чувствительности. Для признаковых и пространственных атрибуций её оценивают метриками вида *insertion-deletion*, *AOPC* и *infidelity*, которые связывают объяснение с изменением предсказания при заданном изменении входа [119–121]. Причинность в предметной области такой показатель не доказывает. Он измеряет интервенционно проверяемый вклад объяснительных компонент в вывод модели [143, 150–152], и верность понимается здесь операционально, как согласованность с поведением самой модели.

Устойчивость $q_{\mathrm{stab}}(E)$ отражает, как объяснение реагирует на допустимые вариации входа и протокола построения. Для её вычисления фиксируется набор тестовых воздействий, куда входят малые реалистичные возмущения, повторные запуски метода и смена случайной инициализации. В таком операциональном смысле устойчивость относят к ключевым критериям качества объяснений [117, 141]. В общем виде она равна среднему структурному сходству объяснения до и после контрольных тестов:

$$q_{stab}(E) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} Sim\big(E(x), E_t(x)\big), \tag{2.65}$$

где $\mathcal{T}$ — множество тестовых воздействий, $E_t(x)$ — объяснение после теста t, а Sim — мера структурного сходства, вид которой зависит от природы объяснения.

Полнота $q_{\mathrm{comp}}(E)$ характеризует, какую часть факторов, реально влияющих на вывод модели, охватывает объяснение. Для признаковых объяснений это доля релевантности, покрытая выделенным подмножеством признаков, для структурных — охват правил, концептов или контрфактуальных сценариев, на которых держится решение. Верность отвечает на вопрос, правильно ли объяснение, полнота — достаточно ли в нём материала.

Простота $q_{\mathrm{simp}}(E)$ ограничивает интерпретационную сложность. Объяснение с чрезмерным числом правил или признаков может быть формально корректным и при этом неудобным для анализа, так что простоту приходится учитывать как отдельный критерий. При многокритериальном сопоставлении объяснений сложность обычно входит в число критериев [117].

Показатель $q_{\mathrm{cost}}(E)$ учитывает вычислительную стоимость построения объяснения. В интерактивных и ресурсно-ограниченных режимах хорошее по содержанию, но дорогое объяснение может оказаться неприемлемым. Содержательных критериев стоимость не заменяет и играет роль дополнительного ограничения на допустимость метода в конкретном прикладном контексте.

Интегральная оценка качества получается взвешенной свёрткой частных показателей:

$$Q(E) = \sum_{k=1}^{5} \omega_k \; q_k(E), \qquad \omega_k \geq 0, \qquad \sum_{k=1}^{5} \omega_k = 1, \tag{2.66}$$

где $q_k(E)$ — частный показатель, а ω_k — его вес. Через веса метрический контур настраивается на предметную область. Там, где цена ошибки высока, больший вес получают верность и устойчивость, в интерактивных режимах — вычислительная стоимость.

Интегральную оценку нельзя читать в отрыве от неопределённости, и допустимость результата определяется совместным условием:

$$E(x) \text{ допустимо} \Leftrightarrow Q\big(E(x)\big) \geq \tau_Q \;\wedge\; U_{\text{total}}(x) \leq \tau_U, \tag{2.67}$$

где τ_Q и τ_U — пороги качества и неопределённости. Объяснение с приемлемой интегральной оценкой может оставаться ненадёжным из-за большого межметодного или межмодельного разброса, и тогда опираться на него при анализе или принятии решения нельзя.

У аддитивной формы (2.66) есть условия применимости. По Р. Кини и Х. Райфе аддитивная функция ценности корректна, если критерии взаимно независимы по предпочтению. Для измеримой ценности нужна ещё независимость по разности, для полезности в условиях риска — аддитивная независимость по полезности [153]. Показатели (2.64) нормированы к $[0,1]$ и описывают разные свойства объяснения, поэтому независимость по предпочтению выполняется для них приближённо. Предпочтение между двумя уровнями верности при фиксированной стоимости не зависит от того, какова эта стоимость. Где независимость нарушается, компенсировать один критерий другим нельзя, и высокая верность ничего не стоит при неустойчивом объяснении. Эту часть предпочтений задаёт пороговое условие (2.67). Объяснение сначала проходит некомпенсаторный отбор по порогам и лишь потом сравнивается по $Q(E)$. Мультипликативная форма тоже учитывает взаимодействие критериев, но сводит его к одному параметру, который трудно интерпретировать и проверить на предметном материале. В работе принята схема «пороговый отбор, затем аддитивная свёртка», а в задаче совместного выбора модели и метода объяснения (раздел 2.9), где взаимодействие критериев существенно, применяется неаддитивная нечёткая агрегация.

Веса ω_k назначаются экспертно по приоритетам прикладного контекста. Разные процедуры выявления весов, вообще говоря, дают разные результаты [154], и одних значений весов для обоснования выбора мало. Веса варьируются в окрестности заданных значений с сохранением условий $\omega_k \geq 0, \sum_k \omega_k = 1$, и сравнение объяснений считается надёжным, только если его итог при такой вариации не меняется.

Для многоуровневого объяснения вида (2.42) полезна, кроме общей оценки, её раскладка по уровням с частными оценками:

$$Q_f(E_f), Q_m(E_m), Q_i(E_i), Q_r(E_r). \tag{2.68}$$

Интегральная многоуровневая оценка складывается из них с весами:

$$Q_{\text{multi}}(E) = \beta_f Q_f\big(E_f\big) + \beta_m Q_m(E_m) + \beta_i Q_i(E_i) + \beta_r Q_r(E_r), \tag{2.69}$$

где $\beta_l \geq 0, \beta_f + \beta_m + \beta_i + \beta_r = 1$.

В гибридных системах без такого разложения не обойтись. Наглядный итоговый текст бывает слабо связан с модельным уровнем, а качественная модельная интерпретация может дойти до пользователя в перегруженном виде. Оценки по уровням показывают, на каком уровне возникло ограничение.

Метрический контур обязан быть воспроизводимым, и вне фиксированного протокола вычисления метрики смысла не имеют. Пока не зафиксированы данные, схема возмущений, нормировка, меры сходства и критерии допустимости, один и тот же метод будет получать разные оценки в зависимости от случайных деталей эксперимента. Оценка качества поэтому вычисляется в явно заданном протоколе:

$$\mathcal{P}_{eval} = (D,\ M,\ \mathcal{T},\ \varrho,\ \Theta), \tag{2.70}$$

где D — набор данных или калибровочная выборка, M — класс модели, $\mathcal{T}$ — набор стресс-проверок и тестовых сценариев, ϱ — семейство используемых метрик, а Θ — фиксированные параметры протокола. Сопоставимы только оценки $Q(E)$, полученные при одном и том же $\mathcal{P}_{\text{eval}}$, что согласуется с установкой современной методологии ОИИ на контекстно-зависимую, но воспроизводимую оценку объяснений [141].

Оценка начинается с фиксации протокола $\mathcal{P}_{\text{eval}}$ вида (2.70). Затем для объяснения вычисляется вектор (2.64), причём устойчивость считается по (2.65) на множестве тестовых воздействий $\mathcal{T}$. Частные показатели приводятся к общей шкале и сворачиваются в $Q(E)$ по (2.66) с весами, отражающими приоритеты контекста, и сразу проверяется, устойчив ли результат к вариации весов. Вместе с неопределённостью $U_{\text{total}}(x)$ проверяется условие (2.67), и объяснение, не прошедшее хотя бы один порог, основанием для анализа не служит. Для многоуровневого объяснения в конце вычисляются оценки уровней (2.68) и сводная оценка (2.69), по которым находят уровень, ограничивающий качество.

Без метрического контура многоуровневость оставалась бы описанием структуры, согласование — мерой внутренней связности, а объединение — способом свести вместе несколько интерпретаций. Оценка $Q(E)$ даёт им общий язык, на котором можно сравнивать разные объяснители, формы объяснения и протоколы проверки в рамках одной процедуры.

Предметной экспертизы интегральная оценка не заменяет. Она дисциплинирует сравнение, но вопросы доменной применимости, риска ошибки и безопасности решаются отдельно. Числовое превосходство одного объяснения над другим остаётся формально обоснованным ориентиром, не более того.

2.8. Протокол воспроизводимого выбора объяснителя

Последний элемент методического контура — выбор объяснителя. В ранних работах по ОИИ нередко брали метод, удобный в программной библиотеке или привычный исследователю, либо следовали традиции домена. Для первичного анализа этого хватает, но воспроизводимости такой выбор не даёт. При произвольно взятом объяснителе итоговое объяснение зависит от случайного внешнего решения, так что выбор метода нужно включить в общий формализованный контур.

Выбор объяснителя строится на всём введённом аппарате и сводится к поиску допустимого кандидата при заданных задаче, модели и ограничениях применения [98, 101, 109].

Пусть $\mathfrak{E}_{\text{cand}}(C) = \{e_1, \dots, e_N\}$ — множество кандидатов-объяснителей в контексте C. Для каждого кандидата $e \in \mathfrak{E}_{\text{cand}}(C)$ строится объяснение $E_e(x)$ и вычисляются его интегральное качество $Q\big(E_e(x)\big)$, неопределённость $U_e(x)$, устойчивость $S_e(x)$ и вычислительная стоимость $\text{Cost}(e)$. Устойчивость $S_e(x)$ оценивается по (2.65) на сценарии стресс-проверок $\mathcal{T}$, то есть по реакции объяснения на возмущения входа. Неопределённость $U_e(x)$ отражает разброс результатов при повторных запусках и на сопоставимых моделях. Межметодная компонента U_{method} для отдельного объяснителя смысла не имеет, и $U_e(x)$ вычисляется по (2.62) при $\lambda_1 = 0$ и $\lambda_2 + \lambda_3 = 1$. Если сценарий $\mathcal{T}$ содержит повторные запуски, их вклад учитывается только в неопределённости, чтобы один и тот же эффект не попал в оценку дважды. Так же поступают с устойчивостью и стоимостью. Они уже входят в $Q(E)$ через частные показатели (2.64), и при выборе объяснителя их веса в (2.66) полагаются равными нулю, а сами эти свойства учитываются только отдельными слагаемыми функционала.

Базовое правило выбора опирается на функционал:

$$J(e \mid x, C) = Q\big(E_e(x)\big) - \lambda_U U_e(x) + \lambda_S S_e(x) - \lambda_C \,\text{Cost}(e), \tag{2.71}$$

где $\lambda_U, \lambda_S, \lambda_C \geq 0$ — коэффициенты, задающие относительное влияние неопределённости, устойчивости и вычислительной стоимости. Выбирается кандидат с наибольшим значением функционала

$$e^*(x, C) = \underset{e \in \mathfrak{E}_{cand}(C)}{\operatorname{argmax}} J(e \mid x, C). \quad (2.72)$$

В (2.72) заключён главный принцип раздела. Объяснитель выбирается по формально вычислимому критерию, в котором учтены качество объяснения, его устойчивость, неопределённость и ресурсная цена построения, а предпочтения исследователя роли не играют. Без величин $Q(E)$ и $U(x)$ из разделов 2.6 и 2.7 воспроизводимый выбор задать было бы нельзя.

Одного функционала, впрочем, мало, потому что пригодность метода зависит от того, к какой модели и к каким данным его применяют. Правило выбора должно учитывать контекст применения, который зададим кортежем:

$$C = (C_D, C_M, C_B, C_R, C_G), \quad (2.73)$$

где C_D — тип данных и модальность, C_M — класс модели, C_B — вычислительный бюджет, C_R — прикладные ограничения и требования надёжности, а C_G — цель объяснения, то есть конкретный режим его использования. Цель представляется категориальной переменной или вектором требований $C_G = (g_1, \dots, g_k)$, где каждому элементу отвечает свой сценарий, от локальной диагностики решения до аудита устойчивости или подготовки отчёта.

С учётом контекста правило выбора принимает вид:

$$e^*(x, C) = \operatorname{argmax}_{e \in \mathfrak{E}_{\text{cand}}(C)} \Big(\beta_Q Q\big(E_e(x)\big) - \beta_U U_e(x) + \beta_S S_e(x) + \beta_P P_e(C) - -\beta_C \operatorname{Cost}(e) \Big), \quad (2.74)$$

где $P_e(C)$ — пригодность объяснителя e для контекста C, а коэффициенты $\beta_Q, \beta_U, \beta_S, \beta_P, \beta_C \geq 0$ задают приоритеты сценария. Через них в правило входит то, что в интерактивном режиме важнее стоимость, а в высокорисковых задачах — верность, устойчивость и низкая неопределённость.

Выражение в скобках в (2.74) обозначим $J(e \mid x, C)$ и назовём контекстной полезностью. При $\beta_P = 0$ оно совпадает с (2.71) с точностью до обозначения коэффициентов. Входящие в $J(e \mid x, C)$ показатели измеряются в разных единицах, и перед свёрткой их приводят к шкале $[0,1]$ или стандартизуют на калибровочной выборке, иначе коэффициенты λ и β перестают выражать относительные приоритеты. Условия применимости аддитивной свёртки и требования к весам здесь те же, что для (2.66).

Функция $P_e(C)$ измеряет совместимость объяснителя с контекстом. В ней учитывается, поддерживает ли метод данный класс модели и модальность, даёт ли он результат в нужной форме и укладывается ли в ограничения. Несовместимый кандидат отсеивается ещё до ранжирования с помощью индикатора допустимости:

$$I(\,e \mid C\,) \in \{0,1\}, \tag{2.75}$$

где $I(e \mid C) = 0$ означает, что в данном контексте объяснитель e в сравнении не участвует. После отсева остаётся множество допустимых объяснителей:

$$\mathfrak{E}^{adm}(C) = \{e \in \mathfrak{E}_{cand}(C)\colon\ I(e \mid C) = 1\}. \tag{2.76}$$

Контекстная полезность допустимого кандидата усредняется по объектам калибровочной выборки D_{cal}:

$$\bar{J}(e \mid C; D_{\text{cal}}) = \frac{1}{|D_{\text{cal}}|}\sum_{x\in D_{\text{cal}}} J\,(e \mid x, C), \tag{2.77}$$

и итоговое правило выбора объяснителя записывается как:

$$e^{*}(C) = \underset{e\in\mathfrak{E}^{adm}(C)}{\operatorname{argmax}}\ \bar{J}(e \mid C; D_{cal}). \tag{2.78}$$

После усреднения по калибровочной выборке выбор больше не зависит от отдельного объекта и превращается в калиброванную процедуру, которую можно воспроизвести.

Чтобы выбор можно было повторить, протокол сравнения фиксируется явно в виде кортежа:

$$\mathcal{P}_{sel} = (D_{cal},\ \mathfrak{E}^{adm}(C),\ \varrho,\ \mathcal{T},\ \Theta), \tag{2.79}$$

где D_{cal} — калибровочная выборка или набор репрезентативных объектов, ϱ — семейство используемых метрик качества, $\mathcal{T}$ — сценарий стресс-проверок, а Θ — фиксированные параметры вычисления, в том числе способ нормировки показателей и порог близости $\varepsilon > 0$. Кандидаты, у которых средние оценки (2.77) различаются меньше чем на ε, считаются близкими. Значения $Q(E)$, $U(x)$, $S_e(x)$ и ранжирование кандидатов сопоставимы и повторяемы только при заданном $\mathcal{P}_{\text{sel}}$.

Выбор по протоколу $\mathcal{P}_{sel}$ идёт последовательно. После фиксации контекста C вида (2.73) индикатор (2.75) отсеивает объяснители, несовместимые с типом данных, классом модели, целью объяснения или прикладными ограничениями, и остаётся множество $\mathfrak{E}^{adm}(C)$ по (2.76). Каждый допустимый кандидат строит объяснения на калибровочной выборке D_{cal} при одних и тех же параметрах Θ. Для этих объяснений по методикам разделов 2.6 и 2.7 вычисляются качество по семейству метрик ϱ, неопределённость, устойчивость на сценарии стресс-проверок $\mathcal{T}$ и вычислительная стоимость.

Функционал полезности усредняется по D_{cal} согласно (2.77), кандидаты упорядочиваются по средней оценке, и по (2.78) выбирается $e^*(C)$. Если оценки нескольких кандидатов отличаются меньше чем на ε, проводится дополнительная проверка или их контролируемое объединение методами раздела 2.5. Протокол $\mathcal{P}_{\text{sel}}$ и результаты сравнения сохраняются, чтобы выбор можно было повторить или пересмотреть при повторной калибровке.

Схема протокола и метрический контур оценки интерпретации приведены на рисунке 2.6.

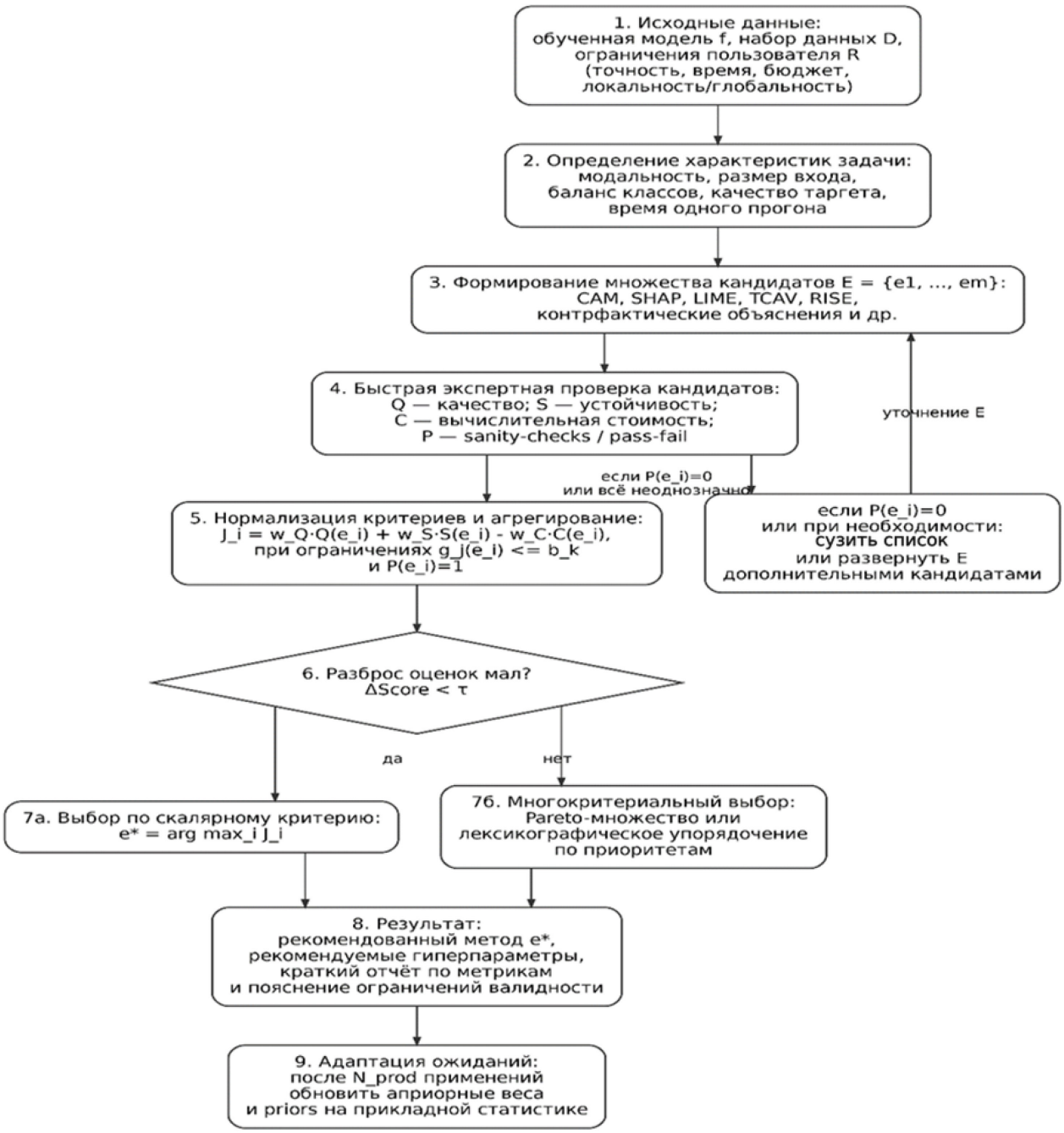


Рисунок 2.6 – Протокол воспроизводимого выбора объяснителя и метрический контур оценки интерпретации

Из протокола следует, что объяснитель нельзя выбрать один раз и навсегда. Когда меняются данные, обновляется модель или ужесточаются требования к надёжности, множество допустимых кандидатов и их ранжирование тоже могут измениться. Воспроизводимый выбор требует поэтому повторной калибровки и хранения артефактов прошлых запусков, и этим формальный контур выбора отличается от единичного вычислительного эксперимента. Вместе с выбранным объяснителем система сохраняет основания выбора, и позже можно восстановить, почему в прежнем контексте решение было именно таким.

Протокол опирается на два допущения. Модель в нём считается заданной, а кандидатов сравнивает один модуль. Первое допущение снимается в разделе 2.9, где модель и метод объяснения выбираются совместно, второе — в разделе 2.10, где выбор распределяется между несколькими агентами.

2.9. Метод комплексирования моделей машинного обучения с многоуровневым ансамблированием объяснений

Один и тот же метод объяснения на архитектурах разной сложности даёт результаты разной верности и устойчивости. Если модель отбирают только по прогностической точности, объяснительные свойства оцениваются, когда выбор уже сделан, и повлиять на него не могут. Здесь модель и метод объяснения выбираются совместно, как конфигурация вычислительного контура, и верность объяснения входит в число критериев отбора наравне с прогностическим качеством.

Селекция идёт в два этапа. Скрининг сужает пространство конфигураций до кандидатов, которых нельзя улучшить сразу и по прогностическому качеству, и по верности объяснения. Эти кандидаты затем оцениваются полностью, и из них выбирается одна конфигурация, которая идёт в запуск. Её артефакты сохраняются, по ней строится объяснительный анализ. Сокращение пространства поиска и окончательное ранжирование разнесены по разным этапам, так что выбор не сводится к одному показателю точности.

Пусть $\mathbb{M}$ — множество моделей-кандидатов, порождаемое пространством поиска $\mathbb{S}$. Каждая модель $M \in \mathbb{M}$ описывается вектором признаков:

$$\chi(M) = \big(a(M), h(M), q(M), s(M), c(M)\big), \tag{2.80}$$

где $a(M)$ — архитектурный тип модели, $h(M)$ — её гиперпараметры, $q(M)$ — показатели прогностического качества, $s(M)$ — характеристика сложности, $c(M) \subseteq \mathfrak{E}_{\text{cand}}(C)$ — методы объяс-

нения, допустимые для модели по индикатору (2.75) при соответствующем ей классе C_M. Допустимые значения каждого признака задаются пространством поиска и ограничениями прикладной задачи.

Ресурсы на обучение и оценку кандидата ограничены бюджетом вычислений:

$$B = \left(N_{\text{ep}}, N_t, N_v\right), \tag{2.81}$$

где N_{ep} — число эпох обучения, N_t и N_v — максимальное число пакетов обучающей и проверочной выборок. Первому этапу отвечает бюджет раннего скрининга B_s, второму — бюджет полной оценки B_f:

$$B_s = \left(N_{\text{ep},s}, N_{t,s}, N_{v,s}\right), B_f = \left(N_{\text{ep},f}, N_{t,f}, N_{v,f}\right), \tag{2.82}$$

причём скрининг обходится дешевле полной оценки:

$$N_{\text{ep},s} < N_{\text{ep},f}, N_{t,s} \leq N_{t,f}, N_{v,s} \leq N_{v,f}. \tag{2.83}$$

Модель считается допустимым кандидатом, если её тип и гиперпараметры лежат в пространстве поиска $\mathbb{S}$, обучение укладывается в бюджет (2.81)–(2.83) и для неё есть хотя бы один допустимый метод объяснения

$$\mathbb{M}^{\text{adm}} = \left\{M \in \mathbb{M}\colon \left(a(M), h(M)\right) \in \mathbb{S},\ c(M) \neq \emptyset\right\}. \tag{2.84}$$

Поиск не выходит за пределы $\mathbb{S}$ и принципиально новых архитектур автоматически не строит. Допустимую пару «модель – метод объяснения» назовём конфигурацией вычислительного контура, а множество конфигураций запишем как

$$\mathcal{K} = \left\{\kappa = (M, e)\colon M \in \mathbb{M}^{\text{adm}},\ e \in c(M)\right\}. \tag{2.85}$$

Когда для модели строится ансамбль объяснений, под e понимается вся процедура его построения и объединения по разделу 2.5, а объяснительные критерии вычисляются для итогового $E^*(x)$. Так комплексирование моделей соединяется с многоуровневым ансамблированием объяснений.

Конфигурация оценивается по четырём частным критериям. Первый из них — прогностическое качество модели, которое вычисляется на валидационной выборке по матрице ошибок:

$$\text{CM} = (c_{ij})_{i,j=1}^{N_c}, \tag{2.86}$$

где N_c — число классов, а элемент c_{ij} равен количеству объектов истинного класса i, отнесённых моделью к классу j, то есть:

$$c_{ij} = \sum_{n=1}^{N} \mathbb{I}\,[y_n = i \wedge \hat{y}_n = j], \tag{2.87}$$

где N — число объектов валидационной выборки, y_n — истинная метка класса, $\hat{y}_n$ — предсказание модели, $\mathbb{I}[\cdot]$ — индикаторная функция. По матрице ошибок для каждого класса вычисляются прецизионность и полнота:

$$\mathrm{Prec}_i = \frac{c_{ii}}{\sum_{j=1}^{N_c} c_{ji} + \varepsilon}, \tag{2.88}$$

$$\mathrm{Rec}_i = \frac{c_{ii}}{\sum_{j=1}^{N_c} c_{ij} + \varepsilon}, \tag{2.89}$$

где $\varepsilon > 0$ — малая константа, предотвращающая деление на нуль. Чтобы оценка меньше зависела от дисбаланса классов, используется макроусреднение:

$$\mathrm{Prec}_{\mathrm{macro}} = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathrm{Prec}_i, \tag{2.90}$$

$$\mathrm{Rec}_{\mathrm{macro}} = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathrm{Rec}_i. \tag{2.91}$$

Итоговая оценка прогностического качества складывается из макропрецизионности и макрополноты с весами:

$$Q_{\mathrm{pred}}(M) = w_{\mathrm{Prec}}\,\mathrm{Prec}_{\mathrm{macro}}(M) + w_{\mathrm{Rec}}\,\mathrm{Rec}_{\mathrm{macro}}(M), w_{\mathrm{Prec}} + w_{\mathrm{Rec}} = 1, \tag{2.92}$$

где $w_{\mathrm{Prec}}, w_{\mathrm{Rec}} \geq 0$ — весовые коэффициенты. В задачах, чувствительных к пропуску значимых классов, вес полноты увеличивается. При выраженной асимметрии стоимости ошибок вместо $\mathrm{Rec}_{\mathrm{macro}}$ или наряду с ним может использоваться полнота целевого класса.

Второй критерий, верность объяснения, оценивается по метрикам вставки и удаления [155]. Метрика вставки отражает рост уверенности модели при поэтапном восстановлении значимых признаков, метрика удаления — её падение при их последовательном удалении. Итоговая величина равна среднему метрики вставки и инвертированной метрики удаления:

$$F(M, e) = \frac{\mathrm{AUC}_{\mathrm{ins}} + (1 - \mathrm{AUC}_{\mathrm{del}})}{2}, \tag{2.93}$$

где $\mathrm{AUC}_{\mathrm{ins}}$ и $\mathrm{AUC}_{\mathrm{del}}$ — площади под кривыми вставки и удаления, $F \in [0,1]$. Чем выше F, тем точнее карта значимости отражает влияние признаков на предсказание модели. Для признаковых и пространственных атрибуций F служит одной из реализаций показателя верности q_{fid} из (2.64).

Третий критерий — параметрическая простота модели. При числе обучаемых параметров N_θ она равна:

$$\mathrm{Simp}(M) = \frac{1}{1+\max\left(0,\log_{10}\left(N_\theta / N_\theta^{\mathrm{ref}}\right)\right)}, \tag{2.94}$$

где N_θ^{ref} — референсный объём параметров. Модели, число параметров которых не превышает N_θ^{ref}, получают $\mathrm{Simp} = 1$, а каждое увеличение числа параметров на порядок сверх референсного снижает оценку.

Четвёртый критерий, вычислительная эффективность, отражает время выполнения этапа конвейера. Если t — фактическое время, то:

$$\mathrm{Eff}(M, e) = \frac{1}{1+t/\tau_{\mathrm{ref}}}, \tag{2.95}$$

где τ_{ref} — опорное время выполнения, задаваемое исходя из вычислительных возможностей системы. Функция убывает с ростом t, принимает значения в $(0,1]$ и при $t = \tau_{\mathrm{ref}}$ равна $0{,}5$. В отличие от показателя q_{cost} из (2.64), относящегося к построению объяснения, Eff характеризует время выполнения этапа конвейера в целом.

Скрининг использует только два критерия, (2.92) и (2.93), вычисленные при бюджете B_s. Для конфигурации $\kappa = (M, e)$ будем писать $Q_{\mathrm{pred}}(\kappa) = Q_{\mathrm{pred}}(M)$ и $F(\kappa) = F(M, e)$. Первичный поиск сводится к двухкритериальной оптимизации:

$$\max_{\kappa \in \mathcal{K}} \left(Q_{\mathrm{pred}}(\kappa), F(\kappa)\right). \tag{2.96}$$

Её решение — фронт Парето, то есть множество конфигураций, для которых нет альтернативы, не худшей по обоим показателям и строго лучшей хотя бы по одному:

$$\mathcal{K}_{\mathrm{par}} = \{\kappa^* \in \mathcal{K} : \nexists\, \kappa' \in \mathcal{K},\ \kappa' \succ \kappa^*\}, \tag{2.97}$$

где $\kappa' \succ \kappa^*$ означает, что κ' доминирует над κ^*, то есть $Q_{\mathrm{pred}}(\kappa') \geq Q_{\mathrm{pred}}(\kappa^*)$, $F(\kappa') \geq F(\kappa^*)$ и хотя бы одно из неравенств строгое.

На втором этапе конфигурации из $\mathcal{K}_{\mathrm{par}}$ оцениваются при бюджете B_f по всем четырём критериям (2.92), (2.93)–(2.95). Оценки сводятся в одну нечётким выводом, выражающим ком-

промиссы между критериями в лингвистической форме. Для фаззификации служат возрастающие и убывающие полутрапециевидные функции принадлежности, описывающие края шкалы, и треугольные, описывающие промежуточные состояния [135]. Переход от чисел к качественным градациям согласуется с вербальным анализом решений, где оценки альтернатив и объяснение выбора формулируются в привычных эксперту словесных шкалах [25]. «Высоким» значениям отвечает возрастающая функция:

$$\mu^{+}(x;a,b)=\begin{cases}0, & x\le a,\\ \frac{x-a}{b-a}, & a<x<b,\\ 1, & x\ge b,\end{cases} \tag{2.98}$$

«низким» — убывающая:

$$\mu^{-}(x;a,b)=\begin{cases}1, & x\le a,\\ \frac{b-x}{b-a}, & a<x<b,\\ 0, & x\ge b,\end{cases} \tag{2.99}$$

промежуточным — треугольная:

$$\mu^{\triangle}(x;a,c,b)=\begin{cases}0, & x\le a,\\ \frac{x-a}{c-a}, & a<x\le c,\\ \frac{b-x}{b-c}, & c<x<b,\\ 0, & x\ge b,\end{cases} \tag{2.100}$$

где a и b задают границы носителя нечёткого множества, а c — точку максимальной принадлежности.

По базе правил вычисляются степени активации уровней предпочтения, а итоговая оценка конфигурации определяется дефаззификацией по методу Сугено нулевого порядка [156]:

$$\mathrm{Score}(M,e)=\frac{\sum_{i\in\Lambda} w_i R_i}{\sum_{i\in\Lambda} R_i}, \tag{2.101}$$

где $\Lambda=\{vh,h,m,l\}$ — множество уровней предпочтения (очень высокий, высокий, средний, низкий), R_i — степень активации правил уровня i, а w_i — числовые синглтон-следствия правил. Значения w_i играют роль априорных полезностей уровней предпочтения и назначаются экспертно. От них зависит итоговая оценка конфигурации, попавшей на тот или иной уровень, поэтому их, как и веса в (2.66), проверяют на устойчивость выбора к вариации.

Правило второго этапа в общем виде записывается как:

$$(M^{*},e^{*})=\mathrm{argmax}_{(M,e)\in\mathcal{K}_{\mathrm{par}}}\Phi\left(Q_{\mathrm{pred}}(M),F(M,e),\mathrm{Simp}(M),\mathrm{Eff}(M,e)\right), \tag{2.102}$$

где Φ — агрегирующая функция многокритериального выбора, реализуемая нечётким выводом (2.98)–(2.101), так что $\Phi = \mathrm{Score}(M, e)$. По структуре (2.102) повторяет правило (2.78), только максимизация идёт по парам «модель – метод объяснения». Аддитивная свёртка вида (2.66) здесь заменена нечётким выводом, потому что критерии второго этапа взаимодействуют и высокая эффективность или простота модели не должна компенсировать низкую верность объяснения.

Метод работает так. Фиксируются пространство $\mathbb{S}$, контекст C и бюджеты (2.81)–(2.83), по (2.84) отбираются допустимые модели и по (2.85) формируется множество конфигураций $\mathcal{K}$. При бюджете B_s для каждой конфигурации вычисляются Q_{pred} по (2.92) и F по (2.93) и строится фронт Парето (2.97). Конфигурации фронта оцениваются заново при бюджете B_f, теперь по всем четырём критериям, включая Simp по (2.94) и Eff по (2.95).

Частные критерии фаззифицируются функциями (2.98)–(2.100), по базе правил вычисляются активации R_i, и каждая конфигурация получает итоговую оценку по (2.101). По (2.102) выбирается конфигурация (M^*, e^*) с наибольшей оценкой. Она сохраняется вместе с параметрами обучения, бюджетами B_s и B_f и параметрами нечёткого вывода, что обеспечивает воспроизводимый запуск и последующий объяснительный анализ.

Программная реализация метода в *AutoML*-ядре платформы и значения параметров, использованные в экспериментах, описаны в разделе 3.5.

2.10. Агентное расширение метода воспроизводимого выбора объяснителя

В базовом протоколе кандидаты сравниваются одним модулем по фиксированному набору метрик, ограничений и контекстных условий. В иерархической системе за разные стороны выбора нередко отвечают разные компоненты, и решение складывается из их согласованной работы. Выбор объяснителя в таком случае естественно рассматривать как результат работы нескольких специализированных агентов [157–160].

Этот взгляд идёт от системно-организационного подхода, где сложная интеллектуальная система понимается как распределённая структура, а решение возникает из координации ролей, ограничений и локальных процедур согласования. У каждого агента своя сторона сравнения кандидатов, будь то допустимость метода, качество объяснения или вычислительная стоимость.

Пусть для объекта x и контекста $C = (C_D, C_M, C_B, C_R, C_G)$ вида (2.73) заданы множество кандидатов-объяснителей $\mathfrak{E}_{\mathrm{cand}}(C) = \{e_1, \dots, e_N\}$ и множество агентов сопоставления $\mathcal{A} = \{a_1, \dots, a_{N_A}\}$. Агент $a_j \in \mathcal{A}$ отвечает за свой аспект выбора и выставляет кандидату $e_k \in \mathfrak{E}_{\mathrm{cand}}(C)$ локальную оценку:

$$s_j(e_k \mid x, C) = \alpha_j Q\left(E_{e_k}(x)\right) - \beta_j U_{e_k}(x) + \gamma_j P_j(e_k \mid C) - \delta_j \operatorname{Cost}(e_k), \qquad (2.103)$$

где $Q\left(E_{e_k}(x)\right)$ — качество объяснения, построенного методом e_k, $U_{e_k}(x)$ — его неопределённость, $P_j(e_k \mid C)$ — пригодность кандидата в контексте задачи с точки зрения агента a_j, $\operatorname{Cost}(e_k)$ — вычислительная стоимость. Коэффициенты $\alpha_j, \beta_j, \gamma_j, \delta_j \geq 0$ задают приоритеты агента, а сами показатели нормируются так же, как в (2.74).

Новых критериев агенты не вводят, они только распределяют по ролям уже имеющиеся, и с базовым протоколом воспроизводимого выбора агентное расширение остаётся совместимым.

Набор независимых оценок $s_j(e_k \mid x, C)$ сам по себе даёт обычное многокритериальное ранжирование, в котором не видно, согласны ли агенты между собой. Между тем от их согласия зависит, насколько устойчив выбор. Для пары агентов a_i и a_j введём попарную согласованность по кандидату e_k:

$$\rho_{ij}(\, e_k \mid x, C\,) = 1 - \frac{|s_i(\, e_k|x, C\,) - s_j(\, e_k|x, C\,)|}{\Delta_{ij}}, \qquad (2.104)$$

где $\Delta_{ij} > 0$ — нормирующий коэффициент, задающий допустимый масштаб расхождения оценок. При корректно выбранном Δ_{ij} согласованность лежит в единичном интервале:

$$\rho_{ij}(\, e_k \mid x, C\,) \in [0,1] \qquad (2.105)$$

и чем больше ρ_{ij}, тем ближе оценки двух агентов по данному кандидату.

Усредняя попарные согласованности по всем парам агентов, получаем параметр порядка:

$$q(e_k \mid x, C) = \frac{2}{N_A(N_A-1)} \sum_{1 \leq i < j \leq N_A} \rho_{ij}\,(e_k \mid x, C). \qquad (2.106)$$

Величина $q(e_k \mid x, C) \in [0,1]$ строится по той же схеме среднего попарного сходства, что и мера когерентности (2.60), и показывает, насколько упорядочена распределённая процедура выбора для кандидата e_k. При $q \approx 1$ агенты оценивают кандидата почти одинаково, и выбор устойчив. Малые значения q говорят о конфликте критериев, ролей или ограничений. Тогда итоговое ранжирование следует считать менее надёжным, даже если отдельные локальные оценки высоки.

С учётом параметра порядка агентная полезность кандидата равна:

$$J_{\text{agent}}(e_k \mid x, C) = \frac{1}{N_A} \sum_{j=1}^{N_A} s_j\,(e_k \mid x, C) + \lambda_q\, q(e_k \mid x, C), \lambda_q \geq 0, \qquad (2.107)$$

где первое слагаемое — средняя локальная полезность кандидата по всем агентам, второе — выигрыш от их согласованности. Под синергией понимается дополнительное преимущество кандидата, по которому агенты приходят к устойчивому согласию, и простое совпадение частных мнений её не создаёт. Выбирается кандидат:

$$e^*_{agent}(x, C) = \underset{e_k \in \mathfrak{C}^{adm}(C)}{\operatorname{argmax}} J_{agent}(e_k \mid x, C), \tag{2.108}$$

где $\mathfrak{C}^{adm}(C)$ — множество кандидатов, допустимых в контексте C.

Базовый протокол раздела 2.8 остаётся частным случаем. При $\lambda_q = 0$ агентная схема сводится к усреднённому многокритериальному выбору, при $\lambda_q > 0$ к нему добавляется требование внутренней согласованности распределённой процедуры. Механизм воспроизводимого выбора агентная постановка надстраивает, ничего в нём не заменяя.

В распределённой системе оценки агентов не остаются неизменными. Агенты уточняют их, учитывая оценки друг друга, и выбор становится итерационной процедурой. Пусть $s_j^{(t)}(e_k)$ — оценка кандидата e_k агентом a_j на шаге t. Межагентное влияние можно описать законом:

$$s_j^{(t+1)}(e_k) = (1 - \eta)\, s_j^{(t)}(e_k) + \eta \sum_{i \in \mathcal{N}_j} w_{ij}\; s_i^{(t)}(e_k), \tag{2.109}$$

где $\eta \in (0,1]$ — шаг согласования, $\mathcal{N}_j$ — агенты, влияющие на a_j, $w_{ij} \geq 0$ — коэффициенты влияния, нормированные условием:

$$\sum_{i \in \mathcal{N}_j} w_{ij} = 1. \tag{2.110}$$

Из-за нормировки (2.110) правая часть (2.109) — выпуклая комбинация оценок, и при итерациях они не выходят из исходной шкалы. На роль универсального закона самоорганизации формула (2.109) не претендует. Это простой шаблон согласования, при котором часть оценки агент сохраняет, а часть уточняет под влиянием связанных с ним агентов. Структура множеств влияния задаёт иерархию связей между агентами, и её выбор — частный случай задачи построения рациональной иерархии управления [161].

Устойчивому режиму соответствует стабилизация последовательностей $s_j^{(t)}(e_k)$ и параметра порядка $q(e_k \mid x, C)$. Выбор объяснителя получается тогда как итог согласованного распределённого процесса, и однократное вычисление скалярной функции его уже не описывает. Для *AutoXAI*-контуров это существенно, поскольку объяснимость в них участвует в подборе модели (раздел 2.9), в проверке устойчивости и в настройке протокола сопоставления.

Чтобы агентная постановка стыковалась с разделом 2.8, её нужно перенести с отдельного объекта на калибровочную выборку. Пусть D_{cal} — фиксированная выборка репрезентативных объектов. По аналогии с (2.77) усреднённая агентная полезность кандидата в контексте C равна:

$$\bar{J}_{agent}(e_k \mid C; D_{cal}) = \frac{1}{|D_{cal}|}\sum_{x \in D_{cal}} J_{agent}(e_k \mid x, C), \tag{2.111}$$

а воспроизводимый агентный выбор, по аналогии с (2.78), задаётся правилом

$$e^*_{agent}(C) = \underset{e_k \in \mathfrak{E}^{adm}(C)}{\operatorname{argmax}} \bar{J}_{agent}(e_k \mid C; D_{cal}). \tag{2.112}$$

В этой форме агентное расширение опирается на калибровочный протокол и не ограничивается ранжированием на одном объекте.

Агентный выбор начинается с фиксации контекста C вида (2.73), допустимых кандидатов $\mathfrak{E}^{\mathrm{adm}}(C)$ по (2.76), агентов $\mathcal{A}$ и коэффициентов их локальных функций (2.103). Каждый агент оценивает по (2.103) всех допустимых кандидатов, после чего оценки уточняются по (2.109) с весами (2.110) до стабилизации последовательностей $s_j^{(t)}(e_k)$. Для каждого кандидата затем считаются попарные согласованности (2.104) и параметр порядка (2.106), и при малом $q(e_k \mid x, C)$ ранжирование помечается как менее надёжное. Агентная полезность (2.107) усредняется по D_{cal} согласно (2.111), и кандидат выбирается по (2.112). Коэффициенты агентов, λ_q, шаг η и число итераций записываются в протокол $\mathcal{P}_{\mathrm{sel}}$.

2.11. Выводы по главе

Во второй главе построен теоретико-методический аппарат многоуровневого формирования, согласования, объединения и сопоставимой оценки объяснений в системах ИИ. На нём основаны положения, выносимые на защиту.

В иерархически организованной системе ИИ объяснение удаётся рассматривать как результат работы прослеживаемого контура обработки данных, а не как внешнее постмодельное описание. Объяснимость системы формализована ориентированным графом узлов и переходов, многоуровневое объяснение задано связанной структурой четырёх уровней, от признакового до результирующего.

Разработаны методы формирования атрибутивных объяснительных объектов на узлах и рёбрах объяснимого контура. Оператор межуровневого согласования формально связывает уровни и превращает набор локальных интерпретаций в целостное многоуровневое объяснение.

Агрегирование, объединение и комплексирование гетерогенных объяснений формально разграничены. Для объединения введены поуровневые операторы и правила взвешивания частных объяснений, а объединённый результат получил форму самостоятельного объекта анализа. Неопределённость объединённого объяснения оценивается количественно, с разложением по источникам разброса.

Метрический контур оценки качества объединяет вектор частных показателей, интегральную оценку и порог допустимой неопределённости. На нём построен протокол воспроизводимого выбора объяснителя, в котором наряду с качеством, устойчивостью и неопределённостью учтены вычислительная стоимость и контекст применения.

Правило выбора распространено на пару «модель – метод объяснения». Селекция конфигураций идёт в два этапа. При бюджете скрининга строится фронт Парето по прогностическому качеству и верности объяснения, затем его конфигурации оцениваются полностью, с учётом параметрической простоты и вычислительной эффективности, и окончательный выбор делается нечёткой агрегацией всех четырёх критериев.

В распределённой постановке протокол расширен до агентной модели. Итоговое решение в ней зависит и от локальных оценок кандидатов, и от того, насколько согласны между собой оценивающие агенты.

Эти результаты служат формальной основой программной реализации платформы, архитектура которой описана в третьей главе.

3. АРХИТЕКТУРА ГЕТЕРОГЕННОЙ ГИБРИДНОЙ ПЛАТФОРМЫ

В третьей главе описана архитектура гетерогенной гибридной платформы для воспроизводимой работы с моделями ИИ, их объяснениями и процедурами проверки. Платформа охватывает весь путь: подготовку данных, обучение или подключение модели, построение объяснений, количественную оценку их качества, контроль доверия и сохранение артефактов эксперимента. Обучение, объяснение и проверка выполняются в одном вычислительном процессе, а не как разрозненные шаги [162].

Архитектура опирается на синергетический подход В. Б. Тарасова. В нём интеллектуальная система рассматривается как организованная среда взаимодействующих компонентов, между которыми распределены функции, локальные решения и ответственность за результат [157, 158]. Платформа устроена так, что за результат отвечает согласованная работа специализированных подсистем и агентов. Синергия возникает там, где результат одной подсистемы влияет на работу другой. Сведения о типе задачи, свойствах модели, допустимых способах интерпретации, итогах метрической проверки и уровне доверия к объяснению определяют следующие шаги. Поэтому итог работы платформы складывается из взаимодействия нескольких её частей, а не из одного запуска объясняющего метода.

Синергия проявляется на двух уровнях. Первый связан с агентной организацией платформы. Отдельные компоненты отвечают за подготовку данных, выбор или подключение модели, построение объяснений, метрическую проверку, контроль доверия и управление артефактами. Их результаты передаются дальше по контуру и учитываются в последующих решениях. Второй уровень относится к взаимодействию двух ключевых контуров: автоматизированного подбора моделей *AutoML* и автоматизированного выбора, настройки, сопоставления и оценки методов объяснения *AutoXAI*. В работах о прозрачности автоматизированного подбора моделей интерпретацию всё чаще рассматривают как отдельную задачу. В ней нужно выбрать метод, настроить его параметры и проверить результат [98, 101, 102, 106].

Для платформы этого мало. Модель должна оцениваться по прогностическому качеству и по тому, насколько она интерпретируема, устойчивы ли её объяснения, можно ли проверить их метриками и включить в доверительный контур. Объяснительные свойства становятся одним из критериев отбора модели наравне с точностью. На практике объяснение часто остаётся отдельной картой значимости или локальным пояснением. Его не воспроизводят, не сравнивают с другими и не связывают с жизненным циклом модели, поэтому оно служит иллюстрацией, а не проверяемым результатом. Между тем в работах по ОИИ всё настойчивее требуют строгой постановки задачи, явных критериев оценки и проверки устойчивости объяснений [7, 19, 67]. Архитектура платформы отвечает на это требование.

Объяснение в платформе является обязательной частью формально заданного эксперимента. В нём фиксируются исходные данные, параметры запуска, методы, показатели оценки и сохраняемые артефакты. Так можно разделить два вопроса, которые часто смешивают: верно ли объяснение модели и согласовано ли оно с данными [134]. При фиксированном протоколе можно также проверить, как меняется объяснение при естественных искажениях входа и при целенаправленных атаках [122, 163]. Архитектурно это обеспечивают формализованный план выполнения и разделение ответственности.

3.1. Принципы проектирования архитектуры платформы

Архитектурное описание платформы выполнено с опорой на положения международного стандарта *ISO*/*IEC*/*IEEE* 42010:2011 (ГОСТ Р 57100–2016) [164]. В логике данного стандарта архитектура описывается через архитектурно значимые запросы заинтересованных сторон. Поэтому проектные решения, принятые при построении платформы, соотносятся с внутренней структурой системы и с требованиями тех групп пользователей, для которых эта система предназначена [165, 166]. Исходя из этого, при проектировании платформы был выполнен анализ заинтересованных сторон. В архитектуре платформы выделены пять ключевых ролей, запросы которых учитывались при выборе основных проектных решений. Эти роли и соответствующие им архитектурные запросы представлены в таблице 3.1.

Таблица 3.1 – Ключевые заинтересованные стороны платформы и их архитектурные запросы

Роль	**Ключевой архитектурный интерес**	**Архитектурное решение**
ML-инженер	Бесшовная интеграция моделей и защита от ошибок	Динамическая валидация применимости методов
Исследователь (*data scientist*)	Подключение новых *XAI*-алгоритмов и их объективное сравнение	Модульная плагинная система и независимые протоколы метрической оценки [162]

Роль	**Ключевой архитектурный интерес**	**Архитектурное решение**
Доменный эксперт (напр., врач)	Предотвращение неверной интерпретации результатов ИИ	Модуль оценки доверия (*trust score)* и генерация предупреждений [41, 167, 168]
Аудитор	Доказуемая воспроизводимость *XAI*-конвейера	Фиксация манифеста окружения (*provenance*) и контроль целостности артефактов [94–96]
DevOps / *MLOps*-инженер	Стабильная работа в разных средах	Разделение ядра-оркестратора и пользовательских интерфейсов [97]

Используются три базовых архитектурных представления, соответствующие точкам зрения, которые приняты в стандарте *ISO*/*IEC*/*IEEE* 42010 и в модели «4+1» [164, 169]. Контекстное представление фиксирует взаимодействие платформы с внешним окружением: источниками данных, внешними сервисами, пользователями и прикладными потребителями результатов. Компонентное представление описывает состав платформы, её основные модули и связи между ними. Динамическое представление раскрывает поведение системы во времени и позволяет проследить исполнение *XAI*-конвейера, включая запуск модели, построение объяснения, метрическую проверку, оценку доверия и сохранение артефактов.

3.2. Концептуальные представления архитектуры платформы

3.2.1. Контекстное представление: платформа как среда

Контекстное представление (см. рисунок 3.1) определяет внешние границы платформы и её место в контуре построения, интерпретации и анализа моделей ИИ. На этом уровне задаются внешние сущности, с которыми платформа связана в процессе эксплуатации: источники данных,

пользователи, внешние модели и сервисы, хранилища, прикладные потребители результатов и регуляторные ограничения.

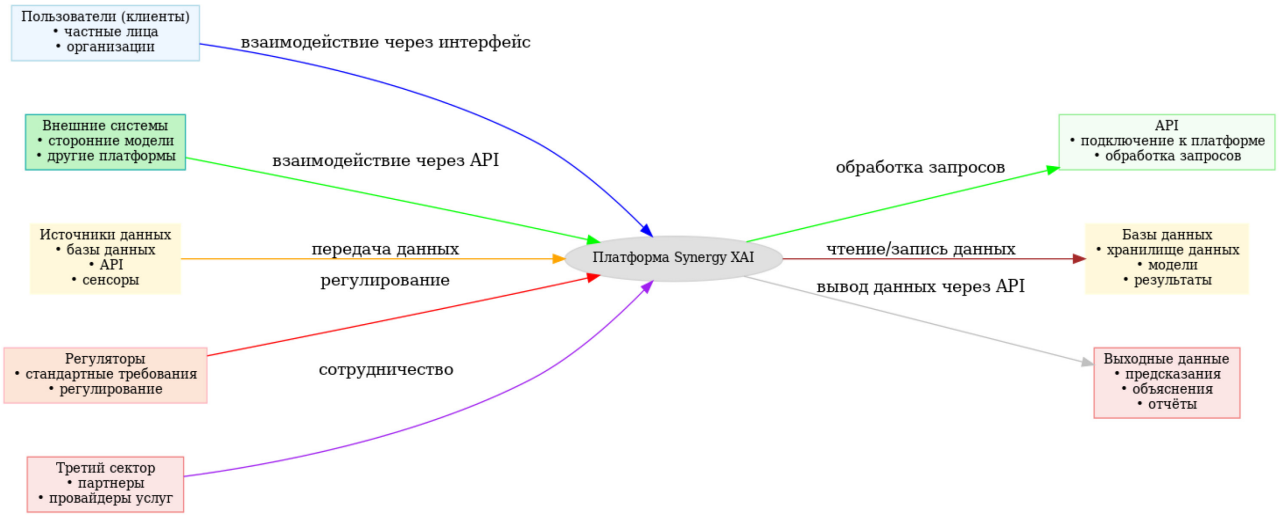


Рисунок 3.1 – Контекстное представление

Во внешнем окружении платформы выделяются три группы взаимодействий:

- к первой группе относятся входные потоки данных и внешние ограничения. Сюда входят поступление исходных данных из баз данных, сенсорных систем и программных интерфейсов, импорт моделей из сторонних систем, взаимодействие с внешними организациями, а также учёт нормативных и регуляторных требований, определяющих допустимые условия применения ИИ.
- Вторая группа охватывает пользовательские и управляющие воздействия. Она включает запросы, поступающие от исследователей, инженеров, доменных специалистов и внешних прикладных систем через графические и программные интерфейсы доступа.
- Третья группа связана с выходными результатами платформы. В ходе работы платформы формируются предсказания, объяснения, метрические оценки, предупреждения и отчёты. Эти артефакты сохраняются в целевых хранилищах, передаются во внешние базы данных и могут предоставляться внешним потребителям через программные интерфейсы.

Тем самым контекстное представление фиксирует состав внешнего окружения платформы, архитектурную границу системы и основные направления обмена данными и результатами. Это задаёт внешнюю рамку для последующего компонентного и динамического описания платформы.

3.2.2. Компонентное представление: ядро, плагины и реестр XAIRegistry

Компонентное представление (см. рисунок 3.2) описывает внутреннюю организацию платформы на уровне основных программных компонентов и связей между ними. Центральное место в этой структуре занимает ядро оркестрации. Оно принимает внешний запрос, формирует внутренний план выполнения, координирует вызов подключённых расширений и собирает итоговый результат. Вокруг ядра организован независимый слой расширений, включающий методы объяснения, адаптеры моделей, метрические протоколы и средства калибровки доверия.

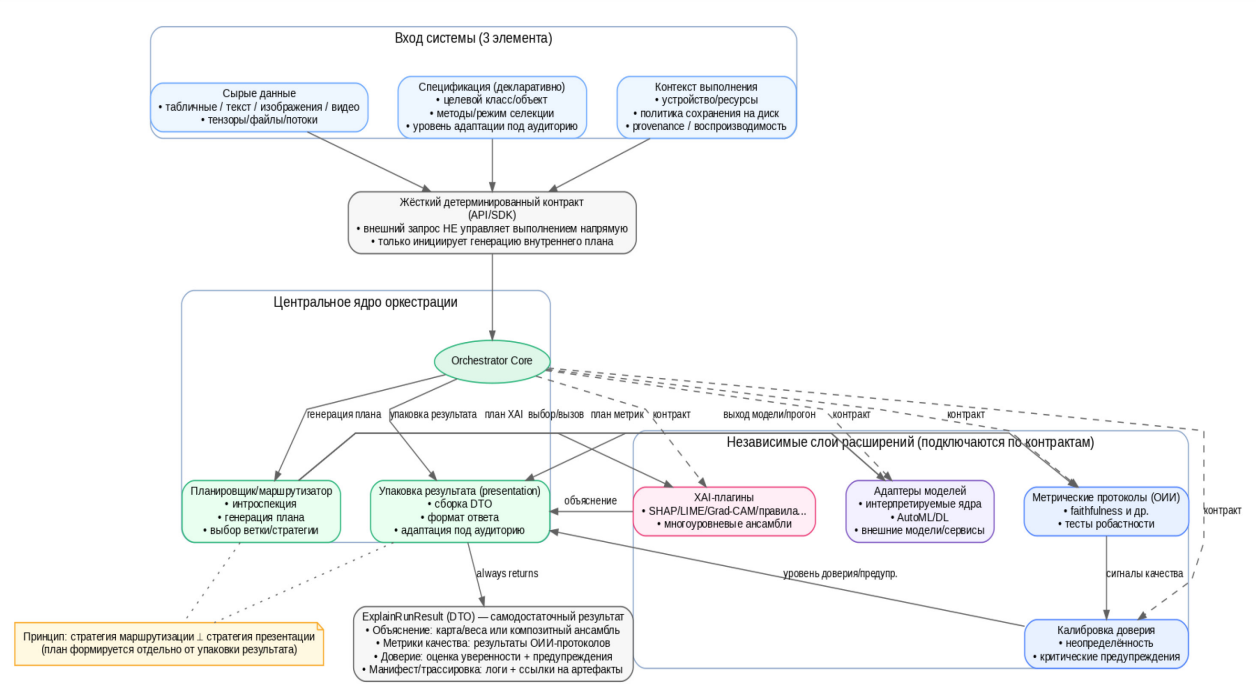


Рисунок 3.2 – Компонентное представление

Вход в платформу задаётся детерминированным контрактом, который включает три группы параметров:

- первая группа описывает исходные данные,
- вторая группа задаёт спецификацию задачи: целевой класс или объект анализа, режим выбора метода, включая ручную настройку или автоматизированный подбор, а также требования к форме представления результата,
- третья группа фиксирует контекст выполнения: вычислительные ресурсы, правила сохранения артефактов и метаданные происхождения эксперимента, необходимые для воспроизводимости и прослеживаемости.

Внешний запрос не управляет маршрутом исполнения напрямую. Его функция состоит в инициации внутреннего плана выполнения, который ядро строит на основе входной спецификации и доступных компонентов платформы. Такая схема исключает зависимость маршрута исполнения от внешнего сценария вызова и делает конвейер детерминированным при фиксированных параметрах задачи и среды.

В составе ядра оркестрации можно выделить два внутренних узла. Первый узел выполняет функции планирования и маршрутизации. Он строит план, интерпретирует правила исполнения и выбирает ветвь обработки. Второй узел отвечает за упаковку результата. Этот узел переводит внутренние согласованные компоненты в выходную форму, задаёт формат представления и учитывает адаптацию к адресату. Логика маршрутизации и логика представления результата не смешиваются, что уменьшает связанность компонентов и позволяет изменять форму выдачи без перестройки основного контура исполнения.

Исполнение плана осуществляется через независимый слой расширений. В его состав входят *XAI*-плагины, реализующие конкретные методы объяснения, адаптеры моделей, обеспечивающие взаимодействие ядра с *AutoML*/*DL* и внешними моделями, а также метрические протоколы, предназначенные для оценки верности и устойчивости объяснений. Отдельный компонент этого слоя связан с калибровкой доверия, оценкой неопределённости и генерацией критических предупреждений. Тем самым объяснение, метрическая проверка и доверительный анализ включаются в единый контур выполнения.

Связующим механизмом между ядром и слоем расширений выступает реестр *XAIRegistry*. В нём фиксируются сведения о доступных расширениях: тип компонента, поддерживаемые входные форматы, ограничения применимости, набор параметров и условия вызова. За счёт этого подключаемые методы и сервисы рассматриваются как зарегистрированные архитектурные элементы с явно заданными интерфейсами. Реестр используется ядром при построении плана выполнения и выборе допустимой последовательности вызовов.

Независимо от того, формируется ли план исполнения вручную или с участием автоматизированного контура подбора, оркестратор возвращает единый результат выполнения. На схеме этот результат представлен объектом *ExplainRunResult*. В его состав входят объяснение, метрические показатели, сигналы доверительного анализа, предупреждения, а также манифест, журналы и ссылки на сохранённые артефакты. Такая форма результата важна тем, что делает выход платформы самодостаточным объектом последующего анализа, проверки и аудита.

3.2.3. Динамическое представление: жизненный цикл запроса и «гейты» валидации

Динамическое представление (см. рисунок 3.3) описывает обработку запроса в платформе как последовательность этапов с условными переходами и встроенными контрольными точками валидации. В этом представлении фиксируются основные стадии выполнения запроса, условия продолжения вычислений и основания для остановки или изменения маршрута обработки.

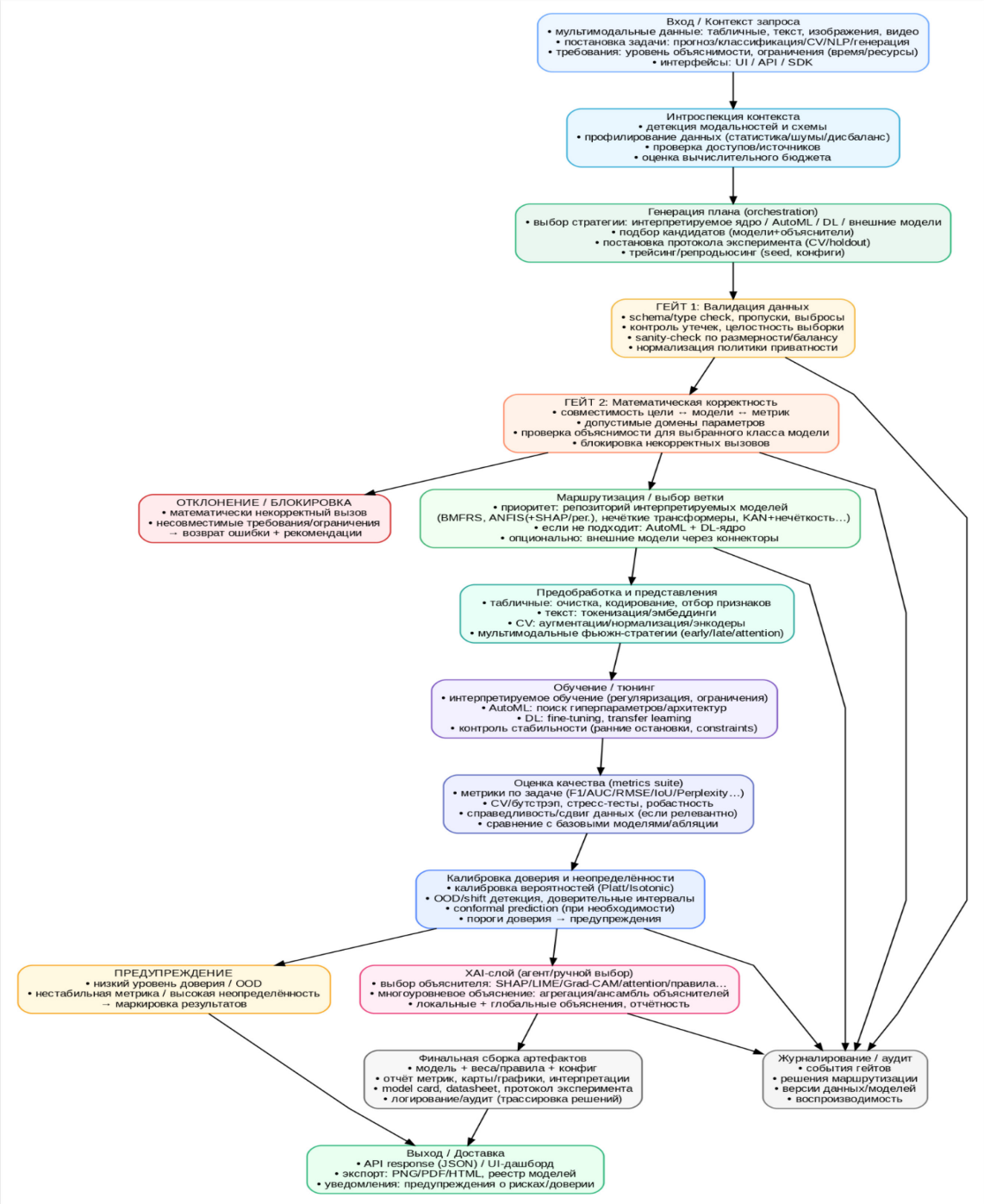


Рисунок 3.3 – Динамическое представление

Жизненный цикл запроса начинается с анализа входного контекста и построения плана выполнения *ExplainPlan*. На этой стадии проводится первичная проверка входных данных, спецификации задачи и параметров запуска. При обнаружении нарушений целостности данных, несовместимости конфигурации или отсутствия обязательных параметров выполнение прекращается на раннем этапе, а запрос отклоняется.

После успешного прохождения входной проверки система переходит к выбору маршрута выполнения. Если сценарий предусматривает использование автоматизированного подбора, в

обработку включается контур *AutoML*, в рамках которого формируются и обучаются модели-кандидаты. В противном случае выполнение продолжается с уже выбранной или подключённой моделью.

Следующая стадия связана с оценкой модели. На этом этапе рассчитываются показатели прогностического качества, выполняется калибровка доверия и анализируется неопределённость результата. Полученные значения используются как основание для дальнейшего продвижения запроса. Если модель не проходит заданные пороги по качеству или доверительности, это учитывается при дальнейшем выполнении запроса. Затем система строит объяснение результата и оценивает его по выбранным метрикам. Одновременно проверяется, применим ли используемый метод объяснения к данной модели, типу данных и режиму выполнения. Если метод неприменим либо объяснение не удовлетворяет установленным критериям, система выбирает альтернативный сценарий обработки или завершает выполнение с указанием причины. На завершающем этапе платформа объединяет предсказание, объяснение, значения метрик, предупреждения и ссылки на сохранённые артефакты. Результаты сохраняются в соответствии с политикой управления артефактами. Воспроизводимость обеспечивается фиксацией параметров запуска, состояния вычислительной среды и ссылок на сформированные артефакты.

3.3. Архитектура функциональных подсистем

Данный раздел раскрывает внутреннюю организацию платформы на уровне функциональных подсистем. Рассматриваются рабочие подсистемы, обеспечивающие выполнение *XAI*-конвейера: ядро управления, реестр методов объяснения, подсистема автоматизированного выбора объяснителей, контур метрической проверки, блок доверительного анализа и механизм сохранения артефактов эксперимента.

Подсистема AutoXAI в составе платформы предназначена для автоматизированного выбора, настройки, запуска, сопоставления и оценки методов объяснения. В этом смысле *AutoXAI* не сводится к прямому вызову отдельного объяснителя, такого как *SHAP*, *LIME* или *Grad-CAM*. Она рассматривает построение объяснения как самостоятельную вычислительную задачу, в которой метод объяснения должен быть согласован с моделью, типом данных, постановкой задачи, требованиями к форме результата и ограничениями предметной области. По своей логике такая подсистема относится к классу решений *AOMO* (*AutoML for Model Explanation*), где объяснение включается в контур автоматизированного выбора и проверки [98, 101, 109].

На вход *AutoXAI* поступают исходные данные или их подготовленное представление, сведения о модели, типе решаемой задачи, целевом объекте объяснения, допустимых методах объяснения, требованиях к форме объяснения, а также доменные и регуляторные ограничения. Эти сведения определяют набор допустимых объяснителей и режим их применения: локальный или глобальный анализ, форму результата, уровень детализации, ограничения по вычислительной стоимости и требования к последующей проверке. На выходе подсистемы формируются объяснения выбранных решений и сопутствующие диагностические сведения: метрики качества объяснения, показатели устойчивости и согласованности, предупреждения о возможной неприменимости метода, а также данные, необходимые для воспроизводимости и отчётности. Результатом работы *AutoXAI* является проверяемый объект анализа, связанный с параметрами запуска, условиями применения метода и результатами метрической оценки.

Внутренняя логика *AutoXAI* определяется несколькими взаимосвязанными механизмами:

- во-первых, выбор метода объяснения выполняется как многокритериальная задача, в которой учитываются свойства модели, тип данных, требования к форме объяснения, ограничения предметной области и критерии качества результата,
- во-вторых, в подсистеме используется обратная связь через объяснения: полученные объяснения и результаты их оценки могут влиять на последующий выбор метода, параметров и маршрута обработки,
- в-третьих, работа *AutoXAI* строится как согласованная процедура, в которой автоматические этапы отбора, метрической проверки и доверительного анализа включаются в единый вычислительный контур.

Один и тот же метод объяснения не является универсально предпочтительным. Его применимость определяется постановкой прикладной задачи, свойствами модели, типом данных, ограничениями предметной области и требованиями к проверяемости результата. В контуре *AutoXAI* допускается экспертное вмешательство, где пользователь может задать ограничения на допустимые методы объяснения, скорректировать критерии их отбора и проанализировать результаты автоматического выбора. Схема подсистемы приведена на рисунке 3.4.



Рисунок 3.4 – Концептуальная схема фреймворка *AutoXAI* класса *AOMO*

3.3.1. Ядро управления

Ядро управления является центральным компонентом платформы, который принимает внешний запрос и преобразует его в формализованный план выполнения. Его задача состоит в координации вычислительного контура: проверке входной спецификации, выборе допустимого маршрута обработки, вызове необходимых подсистем и сборке итогового результата. Поэтому внешний пользователь не управляет конвейером напрямую. Он задаёт постановку задачи и ограничения, а маршрут выполнения формируется ядром на основе спецификации запроса, конфигурации среды и доступных компонентов платформы.

Внутренняя организация ядра опирается на три базовые абстракции, представленные на рисунке 3.5: спецификацию запроса *ExplainSpec*, конфигурацию выполнения *RuntimeConfig* и план выполнения *ExplainPlan*. Так показывается, что должно быть объяснено, в каких условиях выполняется вычисление и каким маршрутом должен пройти запрос.

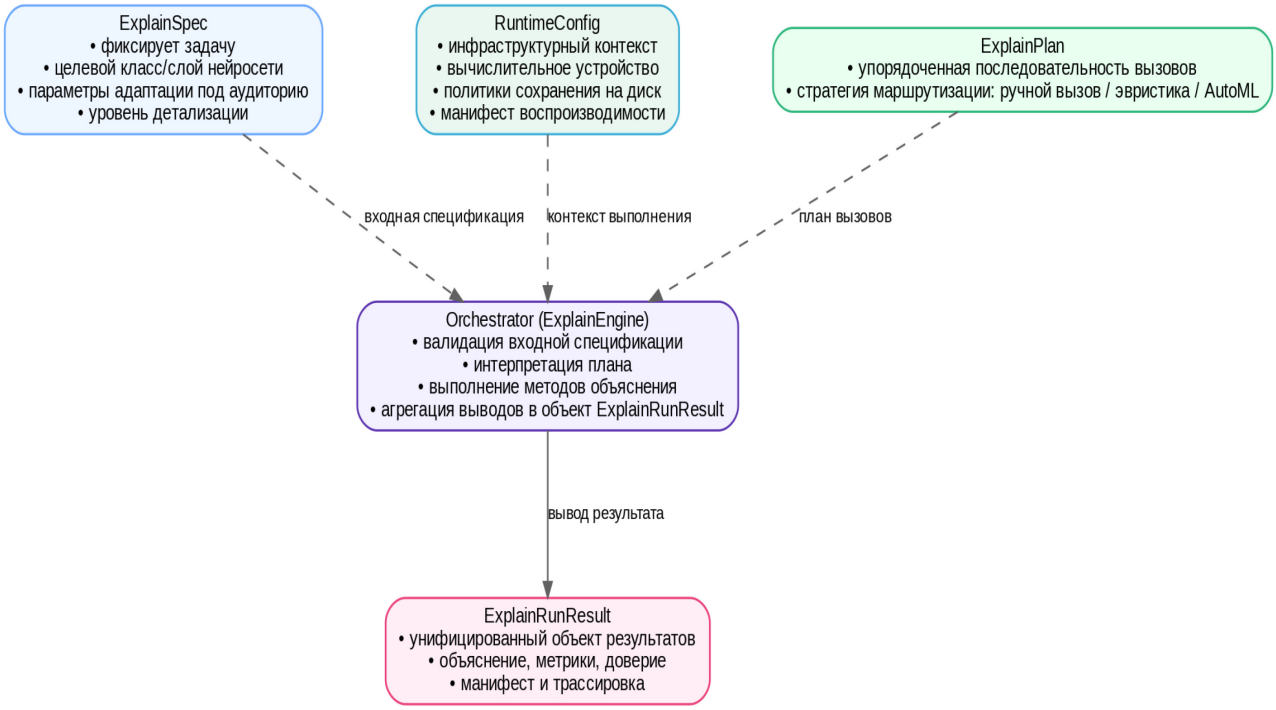


Рисунок 3.5 – Ядро управления платформы

Спецификация запроса задаёт постановку объяснительной задачи. В её состав входят сведения о режиме работы системы, включая использование предварительно обученной модели либо запуск контура обучения, указание целевого класса, целевого выхода или целевого слоя нейронной сети, а также параметры адаптации итогового объяснения к категории пользователя и требуемому уровню детализации. В свою очередь, конфигурация выполнения описывает операционные условия, в которых строится объяснение. Она включает сведения о доступных вычислительных ресурсах, параметрах среды исполнения, правилах сериализации и сохранения артефактов, а также характеристиках окружения, подлежащих фиксации для обеспечения воспроизводимости результатов. Наконец, план выполнения задаёт упорядоченную последовательность этапов вычислительного контура: от подготовки данных и, при необходимости, обучения модели до генерации объяснений, их метрической проверки, доверительного анализа и сохранения результатов. Формирование плана осуществляется подсистемой маршрутизации и может опираться на декларативное описание, набор правил или процедуры автоматизированного поиска. Так *ExplainPlan* связывает постановку задачи и условия запуска с конкретным маршрутом обработки.

Оркестратор интерпретирует сформированный план и обеспечивает согласованное выполнение его этапов. Он проверяет корректность входной спецификации с учётом ограничений используемой модели и выбранных методов объяснения, инициирует вызовы модулей в требуемой последовательности, контролирует корректность вычислительных процедур и передаёт результаты в подсистемы оценки, доверительного анализа и сохранения артефактов. За счёт этого

ядро обеспечивает выполнение *XAI*-конвейера и прослеживаемость процесса построения объяснения.

Методы объяснения подключаются к ядру через унифицированный механизм регистрации и вызова, который рассматривается далее в подразделе 3.3.2.

3.3.2. Реестр методов и плагинная система

Реестр методов объяснения *XAIRegistry*, представленный на рисунке 3.6, является связующим слоем между ядром управления и подключаемыми компонентами платформы. Его назначение состоит в том, чтобы методы объяснения, адаптеры моделей, метрические протоколы и связанные с ними ограничения применимости регистрировались как элементы с явно заданными интерфейсами, параметрами и условиями вызова.

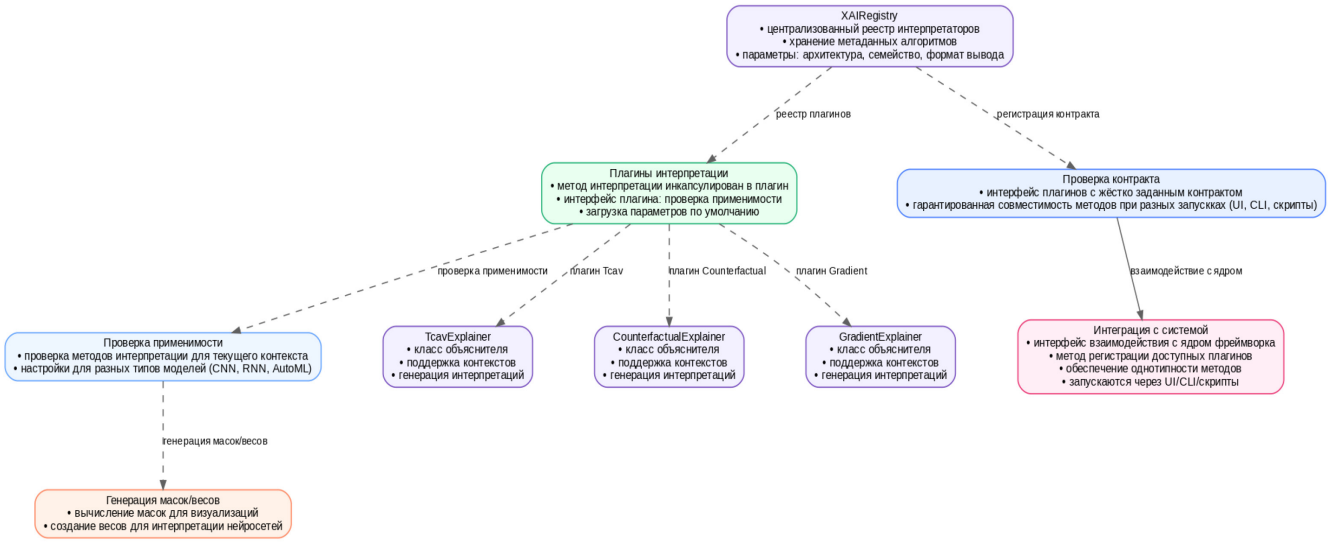


Рисунок 3.6 – Модуль *XAIRegistry*

В реестре фиксируются сведения о доступных объяснителях и вспомогательных компонентах: тип метода, поддерживаемые модели и форматы данных, требуемые внутренние представления, допустимые параметры, ограничения на входные данные и формат возвращаемого результата. При построении плана выполнения ядро использует эти сведения для проверки применимости метода к текущему контексту задачи. Каждый метод объяснения включается в архитектуру как подключаемый модуль, реализующий фиксированный интерфейс. После проверки применимости реестр обеспечивает инициализацию объяснителя с заданными или принятыми по умолчанию параметрами и передачу результата в стандартизованном представлении.

XAIRegistry поддерживает несколько семейств апостериорных методов объяснения, применяемых к моделям типа «чёрный ящик»: градиентные методы атрибуции, локальные суррогатные модели, методы на основе значений Шепли, пертурбационные методы, методы семейства *CAM*, а также концептные и контрфактуальные объяснения [54–56, 60–62, 155, 170–174]. Несмотря на различие математических оснований, эти методы включаются в единый архитектурный контур по общей схеме: выбор целевого объекта объяснения, настройка параметров, построение локального или глобального объяснения и передача результата в модуль оценки качества и доверия.

К числу базовых градиентных методов относится построение карты чувствительности, в которой важность признака определяется через производную выхода модели по соответствующей входной координате [20, 69]. Для входа $x \in \mathbb{R}^{|\mathcal{F}|}$ и целевого класса c классическая карта чувствительности определяется как:

$$S^{(c)}(x) = | \nabla_x f_c(x) |, \tag{3.1}$$

где $f_c(x)$ – выход модели для класса c, а $\nabla_x f_c(x)$ – градиент этого выхода по входу. Данный подход лежит в основе *saliency*-методов и позволяет быстро получить локальную карту вклада признаков.

При насыщении нелинейностей градиент в (3.1) может обращаться в нуль и для тех признаков, которые на самом деле влияют на решение. Этот недостаток устраняет метод Integrated Gradients [62], определяемый формулой (2.19), где вклад j-го признака накапливается вдоль траектории от опорной точки x' к анализируемому объекту x. На практике интеграл в (2.19) заменяется дискретной суммой по m шагам:

$$IG_j(x) \approx \left(x_j - x_j'\right) \frac{1}{m} \sum_{t=1}^{m} \frac{\partial f_c\left(x' + \frac{t}{m}(x - x')\right)}{\partial x_j}. \tag{3.2}$$

Для сглаживания локальных градиентных карт в платформе могут использоваться методы семейства *SmoothGrad* [170]. Если $\xi_k \sim \mathcal{N}(0, \sigma^2 I)$ – случайный шум, то сглаженная оценка важности задаётся выражением:

$$SG^{(c)}(x) = \frac{1}{K} \sum_{k=1}^{K} \nabla_x f_c(x + \xi_k), \tag{3.3}$$

такое усреднение ослабляет влияние локальных флуктуаций и повышает интерпретируемость получаемых карт.

Отдельное семейство образуют локальные суррогатные методы, прежде всего *LIME* [54]. В этом случае исходная модель f аппроксимируется интерпретируемой локальной моделью $g \in$

G, выбранной из класса простых моделей, например линейных. Соответствующая оптимизационная задача задаётся формулой (2.18), а её решение даёт локальное, понятное человеку объяснение в окрестности конкретного объекта.

Для более строгой декомпозиции предсказания по признакам используются методы, основанные на значениях Шепли [55]. Вклад j-го признака определяется формулой (2.17) и интерпретируется как средний маржинальный вклад признака по всем возможным коалициям. На этой основе в реестре поддерживаются вычислительные варианты *SHAP*, включая *kernel*, *permutation* и *partition*. При сохранении общей аддитивной интерпретации вклада признаков они различаются способом аппроксимации величин ϕ_j:

$$f(x) \approx \phi_0 + \sum_{j \in \mathcal{F}} \phi_j\,(x). \tag{3.4}$$

Пертурбационные методы строят объяснение на основе контролируемых изменений входа и анализа реакции модели [56]. В методе окклюзии важность области R может быть выражена через падение отклика модели после маскирования этой области:

$$Occ_R(x) = f_c(x) - f_c\big(x \odot (1 - \mathbb{M}_R)\big), \tag{3.5}$$

где $\mathbb{M}_R$ — бинарная маска области R, а $\odot$ обозначает поэлементное умножение. Чем сильнее снижается отклик модели, тем выше считается значимость скрытой области.

Более общим градиент-независимым методом является *RISE*, в котором используется набор случайных масок $\mathbb{M}_k$, $k = 1, \dots, N$ [155]. Карта значимости оценивается как:

$$E_{RISE}(i,j) = \frac{1}{N\,\mathbb{E}[\mathbb{M}(i,j)]} \sum_{k=1}^{N} f_c\,(x \odot \mathbb{M}_k)\mathbb{M}_k(i,j), \tag{3.6}$$

где $\mathbb{M}_k(i,j) \in \{0,1\}$ – значение случайной маски в позиции (i,j). Такой подход позволяет оценивать вклад областей изображения по статистике отклика модели на множество частично замаскированных входов.

Для свёрточных архитектур важную роль играют методы семейства *CAM*, в частности *Grad-CAM* [61]. Для k-й карты активаций A^k выбранного свёрточного слоя вес α_k^c для класса c вычисляется по формуле (2.21), где y^c – логит класса c, а Z – число пространственных позиций в карте активаций. Итоговая карта локализации определяется формулой (2.20).

На этой базовой конструкции строятся модификации семейства, используемые в платформе: *Grad-CAM++, Score-CAM, HiResCAM*, *LayerCAM*, *AblationCAM*, *XGradCAM* и *EigenCAM* [171–173]. Они различаются способом вычисления весов $\alpha_k^{(c)}$, процедурой агрегации карт активации и степенью учёта пространственной структуры признаков, но сохраняют общую задачу – локализацию областей входа, определяющих решение модели.

Для объяснения через концепты высокого уровня в реестре могут использоваться методы семейства *TCAV* [174]. Пусть v_C – вектор концепта, построенный в пространстве внутренних представлений модели, а $h_l(x)$ – представление объекта на слое l. Тогда локальное влияние концепта на целевой класс c выражается через направленную производную:

$$S_{C,c,l}(x) = \nabla_{h_l} f_c\big(h_l(x)\big) \cdot v_C, \tag{3.7}$$

если эта величина положительна, увеличение активации концепта C повышает отклик модели для класса c. На множестве объектов итоговая концептная значимость может быть выражена долей примеров, для которых выполняется условие $S_{C,c,l}(x) > 0$.

Наконец, контрфактуальные объяснения описываются как задача поиска минимального изменения входа, переводящего модель к альтернативному решению [60]. В общем виде такая задача записывается как:

$$x^{\mathrm{cf}} = \mathrm{argmin}_{z \in X}\big(\lambda\, \ell(f(z), y_{\mathrm{tar}}) + d(z, x)\big), \tag{3.8}$$

где x^{cf} – контрфактуальный объект, y_{tar} – целевой альтернативный класс, ℓ – функция потерь, поощряющая достижение нужного прогноза, $d(z, x)$ – мера отклонения от исходного объекта, а $\lambda > 0$ – коэффициент компромисса.

3.3.3. Подсистема метрической верификации

Метрическая верификация в платформе вынесена в отдельную подсистему независимых метрических протоколов (см. рисунок 3.7) и предназначена для того, чтобы объяснение не рассматривалось как завершённый результат только на основании факта его построения. Без проверки карта значимости, таблица вкладов признаков или другое локальное объяснение остаются вспомогательным материалом, поэтому в платформе объяснение включается в контур оценки, для него рассчитываются показатели качества, устойчивости, согласованности и применимости, а результаты проверки передаются далее в подсистемы доверительного анализа, сопоставления методов и сохранения артефактов.

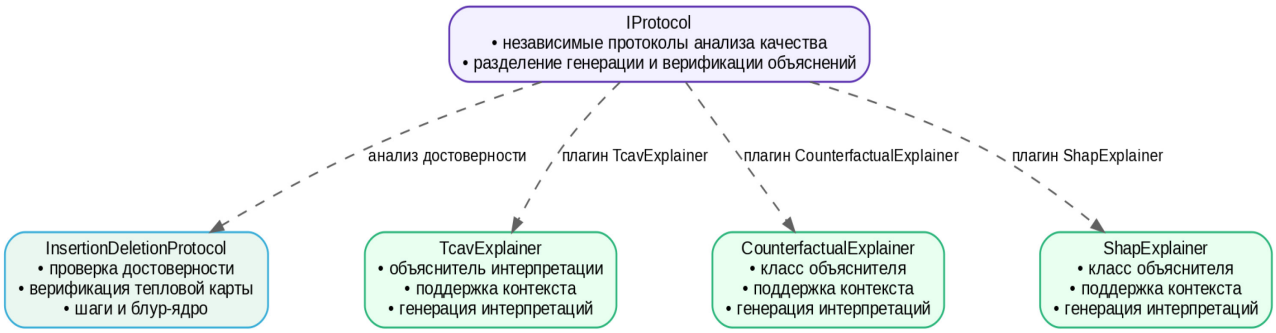


Рисунок 3.7 – Подсистема независимых метрических протоколов

В составе метрического контура выделяются несколько групп проверок:

- первая группа связана с верностью объяснения модели и оценивает, насколько выделенные признаки действительно влияют на предсказание,
- вторая группа связана с устойчивостью и проверяет, сохраняются ли предсказание и карта объяснения при малых возмущениях входных данных,
- третья группа оценивает согласованность объяснений, полученных разными методами или при разных условиях запуска,
- четвёртая группа включает контрольные проверки, позволяющие выявить объяснения, которые слабо зависят от параметров модели, целевого класса или других значимых условий вычисления.

Результаты метрической верификации используются как основание для дальнейшего выбора маршрута обработки. Если метод объяснения неприменим к заданной модели, типу данных или режиму выполнения, либо если объяснение не проходит заданные пороги качества, система может выбрать альтернативный объяснитель, изменить параметры запуска или завершить выполнение с указанием причины. В базовом режиме проверка воспроизводимости объяснений обеспечивается фиксацией параметров запуска, состояния вычислительной среды и ссылок на сформированные артефакты. Дополнительно может применяться проверка на серии повторных запусков, при которой один и тот же метод объяснения оценивается при нескольких независимых выполнениях вычислительной процедуры. При явном задании минимального допустимого уровня воспроизводимости этот показатель может использоваться как дополнительное условие отбора метода.

Ниже для иллюстрации работы независимого метрического протокола рассматривается один из базовых механизмов оценки *Insertion–Deletion* [155], применяемый для карт признаковой важности. Данный протокол анализирует изменение предсказательной способности модели при последовательном возмущении входного объекта в соответствии с картой значимости признаков. Пусть $f(x)$ – функция предсказания модели, возвращающая вероятность целевого класса для

входного объекта x, тогда на каждом шаге процедуры формируется модифицированный объект x_t, для которого вычисляется вероятность предсказания:

$$p_t = f(x_t), \tag{3.9}$$

где $t = 1, \dots, N_{st}$— номер шага процедуры возмущения.

Для количественной оценки изменения уверенности модели используется площадь под кривой изменения вероятности (AUC):

$$AUC = \frac{1}{N_{st}} \sum_{t=1}^{N_{st}} p_t, \tag{3.10}$$

где N_{st}— общее число шагов процедуры.

В сценарии *Insertion* пиксели ранжируются по убыванию значений карты значимости. Процедура начинается с размытой копии изображения, и на каждом шаге в неё возвращается очередная порция наиболее важных пикселей оригинала, так что к последнему шагу изображение восстанавливается полностью [155]. Если карта значимости корректно выделяет признаки, существенные для модели, вероятность целевого класса должна быстро расти. В предельном случае $AUC_{ins} \to 1$.

В сценарии *Deletion* пиксели исходного изображения последовательно удаляются или заменяются фоном в том же порядке. Если карта действительно отражает признаки, на которые опирается модель, вероятность целевого класса должна быстро снижаться. В предельном случае $AUC_{del} \to 0$.

Для перехода от оценки качества объяснения отдельного объекта к оценке объяснимости модели на валидационной выборке вводится усреднённая метрика верности:

$$\bar{F}(M) = \frac{1}{n} \sum_{i=1}^{n} F_i\,(M), \tag{3.11}$$

здесь $F_i(M)$ — значение метрики верности объяснения для i-го объекта выборки, а n— число объектов валидационной выборки.

Дополнительно оценивается разброс значений этой метрики по выборке:

$$D_F(M) = \frac{1}{n} \sum_{i=1}^{n} (\,F_i(M) - \bar{F}(M))^2, \tag{3.12}$$

чем меньше $D_F(M)$, тем более однородно качество объяснений на различных объектах.

3.3.4. Контуры контроля SafetyGuard и TrustEvaluator

В архитектуре платформы предусмотрены два контура контроля (см. рисунок 3.8): *SafetyGuard* и *TrustEvaluator*.

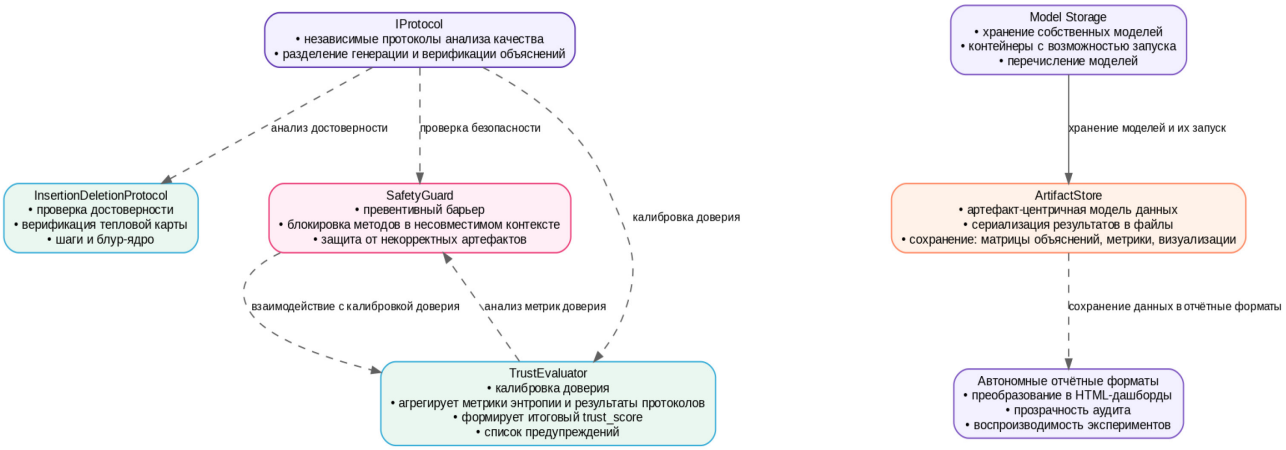


Рисунок 3.8 – Архитектура системы оценки доверия и безопасности

Подсистема *SafetyGuard* используется на этапе подготовки выполнения. Она контролирует соответствие объяснителя типу модели, наличие требуемых внутренних представлений, допустимость входных данных, ограничения вычислительного окружения и требования прикладного контекста. При выявлении несовместимости вызов метода блокируется до начала исполнения. Это предотвращает появление артефактов, которые формально построены, но методологически некорректны.

Подсистема *TrustEvaluator* выполняет агрегированную оценку пригодности результата. Такая оценка не должна трактоваться как доказательство истинности объяснения, она является инженерным индикатором того, насколько результат можно использовать в дальнейшем анализе. *TrustEvaluator* объединяет сигналы качества модели, метрики качества объяснения, показатели устойчивости и согласованности, сведения о неопределённости, результаты контрольных проверок и предупреждения [167, 168]. Поэтому в платформе важен не только итоговый числовой индекс доверия, но и список предупреждений, указывающий на ограничения применимости объяснения.

Неопределённость в этом контуре рассматривается не только как характеристика предсказания модели, но и как фактор надёжности объяснительного результата. Если модель демонстрирует высокую неопределённость, а объяснение при этом нестабильно, плохо согласуется с другими методами или нарушает предметные ограничения, такая ситуация должна снижать доверие

к результату независимо от визуальной убедительности карты значимости. Поэтому предупреждения и диагностические признаки в *TrustEvaluator* имеют самостоятельное значение и не должны скрываться за одной агрегированной оценкой.

Для формализации работы доверительного контура далее вводится интегральный индекс доверия, который определяется как взвешенная агрегация частных показателей качества объяснения и результатов независимых проверок.

$$I_{\text{trust}} = \lambda_F F + \lambda_S S_e + \lambda_{\text{Cons}} \text{Cons}_{\text{pair}} + \lambda_{\text{San}} \text{San} + \lambda_V V, \tag{3.13}$$

где F — метрика верности, S_e — метрика устойчивости, $\text{Cons}_{\text{pair}}$ — метрика согласованности, San — результат проверки чувствительности к разрушению параметров модели, V — метрика визуальной пригодности, λ_i — весовые коэффициенты при условиях:

$$\lambda_F, \lambda_S, \lambda_{Cons}, \lambda_{San}, \lambda_V \geq 0, \lambda_F + \lambda_S + \lambda_{Cons} + \lambda_{San} + \lambda_V = 1. \tag{3.14}$$

Перед агрегированием все частные показатели нормируются к диапазону [0,1]. Чем выше значение I_{trust}, тем выше степень доверия к объяснению. Условия применимости аддитивной формы (3.13) и требования к выбору коэффициентов λ те же, что и для интегральной оценки качества (2.66).

Для конкретизации частных компонентов интегрального индекса доверия далее вводятся показатели согласованности, устойчивости и проверки чувствительности к разрушению параметров модели. Согласованность объяснений, полученных различными методами, оценивается нормированной косинусной мерой:

$$\text{Cons}_{\text{pair}}\left(E^{(1)}, E^{(2)}\right) = \frac{1}{2}\left(1 + \frac{\langle E^{(1)}, E^{(2)} \rangle}{\|E^{(1)}\|_2 \, \|E^{(2)}\|_2}\right), \tag{3.15}$$

здесь $E^{(1)}$ и $E^{(2)}$ — векторы объяснений, полученные различными методами для одного и того же объекта. Значение $\text{Cons}_{\text{pair}}\left(E^{(1)}, E^{(2)}\right)$ тем выше, чем ближе карты значимости по направлению и пространственному распределению значимых признаков.

Устойчивость в доверительном контуре рассматривается на двух уровнях: на уровне предсказания модели и на уровне карты объяснения. На уровне предсказания метрика S_p показывает, насколько меняется вероятность исходно предсказанного класса, если на вход подать слегка возмущённый объект $x + \delta_k$. Для каждого из K возмущённых вариантов вычисляется модуль разности $|p_k - p_0|$, затем эти значения усредняются и вычитаются из единицы. Если вероятность не меняется ни при одном возмущении, $S_p = 1$, чем сильнее колеблется уверенность модели, тем меньше значение метрики. Формально:

$\hat{y}_0 = \text{argmax} f(x)$, $p_k = P(\hat{y}_0 \mid x + \delta_k)$, $p_0 = P(\hat{y}_0 \mid x)$,

$$S_p(x) = 1 - \frac{1}{K}\sum_{k=1}^{K} \mid p_k - p_0 \mid \tag{3.16}$$

Показатель R_{cls} задаёт долю возмущённых примеров, для которых модель сохраняет исходный предсказанный класс:

$$R_{\text{cls}}(x) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\,[\text{argmax}_c\, f_c(x + \delta_k) = \hat{y}_0]. \tag{3.17}$$

Устойчивость карты объяснения оценивается по степени изменения карты значимости при малых возмущениях входных данных:

$$S_e(x) = 1 - \frac{1}{K}\sum_{k=1}^{K} \frac{\|E(x) - E(x+\delta_k)\|_1}{\|E(x)\|_1 + \|E(x+\delta_k)\|_1 + \varepsilon}, \tag{3.18}$$

здесь $E(x)$— объяснение исходного объекта, $E(x + \delta_k)$— объяснение после малого возмущения, $\varepsilon > 0$— малая константа, предотвращающая деление на нуль. Чем ближе S_e к единице, тем устойчивее объяснение.

Метрика San показывает, насколько объяснение зависит от параметров модели. Для этого карта объяснения сравнивается с картой, полученной после рандомизации весов [70]. Если после такой рандомизации объяснение заметно меняется, *sanity*-проверка считается пройденной:

$$\tilde{E} = \frac{E - \min(E)}{\max(E) - \min(E) + \varepsilon},$$

$$San(x) = 1 - \frac{\langle \tilde{E}_{base}, \tilde{E}_{rand} \rangle}{\|\tilde{E}_{base}\|_2 \, \|\tilde{E}_{rand}\|_2 + \varepsilon}. \tag{3.19}$$

Помимо этих протоколов, в платформе могут использоваться и прикладные эвристики визуальной пригодности объяснения. Они учитывают ограничения предметной области и позволяют скорректировать итоговую оценку доверия. Например, оценка может снижаться, если карта внимания концентрируется на фоне или выходит за пределы допустимой области интереса.

Пусть $E \in [0,1]^{m \times n}$ — нормированная карта значимости, построенная методом объяснения для рассматриваемого объекта, где m и n — размеры сетки признаков. Предполагается, что большие значения $E(i,j)$ соответствуют более высокой важности соответствующей области для принятого моделью решения.

Для количественной оценки нежелательной концентрации внимания на фоне вводится штраф за фон. Пусть $\mathbb{M}_{bg} \in \{0,1\}^{m \times n}$ — бинарная маска фоновой области, полученная методами

сегментации, пороговой фильтрации или иными доменно-специфическими процедурами. Тогда доля значимости вычисляется как отношение суммарной значимости, приходящейся на фоновые пиксели, к общей значимости карты E:

$$Pen_{bg} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} E(i,j)\,\mathbb{M}_{bg}(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} E(i,j)+\varepsilon}, \tag{3.20}$$

где $\varepsilon > 0$ — малая константа, предотвращающая деление на нуль. Чем больше значение Pen_{bg}, тем большая часть внимания модели сосредоточена на областях, не содержащих содержательно значимой информации.

Для учёта предметно допустимой области внимания вводится штраф за выход за пределы области интереса. Пусть $\mathbb{M}_{ROI} \in \{0,1\}^{m\times n}$ — бинарная маска допустимой области интереса, определяемая анатомическими, структурными или иными семантическими ограничениями предметной области. Тогда доля активации вне допустимой области задаётся выражением:

$$Pen_{ROI} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} E(i,j)\left(1-\mathbb{M}_{ROI}(i,j)\right)}{\sum_{i=1}^{m}\sum_{j=1}^{n} E(i,j)+\varepsilon}, \tag{3.21}$$

чем выше значение Pen_{ROI}, тем сильнее карта объяснения смещена за пределы области, в которой с точки зрения предметной интерпретации ожидается наличие релевантных признаков.

В общем случае может использоваться не один, а несколько доменных штрафов Pen_k, где $k = 1, \dots, N_{pen}$, отражающих различные типы пространственных или семантических ограничений. Тогда компонент визуальной пригодности может быть определён как скорректированная оценка верности объяснения:

$$V = \max\left(0, F_{faith} - \sum_{k=1}^{N_{pen}} \lambda_k\, Pen_k\right), \tag{3.22}$$

где $F_{\text{faith}} \in [0,1]$ — метрика верности объяснения, $Pen_k \in [0,1]$— доменные штрафы, а $\lambda_k \geq 0$ — коэффициенты строгости, задаваемые экспертом или определяемые на этапе настройки системы. Для сохранения нормировки удобно выбирать коэффициенты так, чтобы выполнялось условие:

$$\sum_{k=1}^{N_{pen}} \lambda_k \leq 1. \tag{3.23}$$

В этом случае величина V характеризует визуальную пригодность: чем больше её значение, тем лучше объяснение удовлетворяет критериям верности и ограничениям предметной области. Если доменные ограничения не заданы, соответствующие штрафы принимаются равными нулю, и компонент V совпадает с базовой метрикой верности F_{faith}.

Подсистема *TrustEvaluator* объединяет универсальные протоколы оценки качества объяснений и прикладные эвристики. За счёт этого оценка учитывает согласованность объяснения с поведением модели и его интерпретируемость в контексте конкретной предметной области.

3.3.5. Слой воспроизводимости

Репозиторий моделей и хранилище артефактов образуют инфраструктурный слой воспроизводимости платформы (см. рисунок 3.9). Задача слоя состоит в том, чтобы каждый запуск платформы оставлял проверяемый след. Благодаря этому результат работы *XAI*-конвейера может быть просмотрен и проверен, сопоставлен с другими запусками и воспроизведён при сохранении исходных условий.

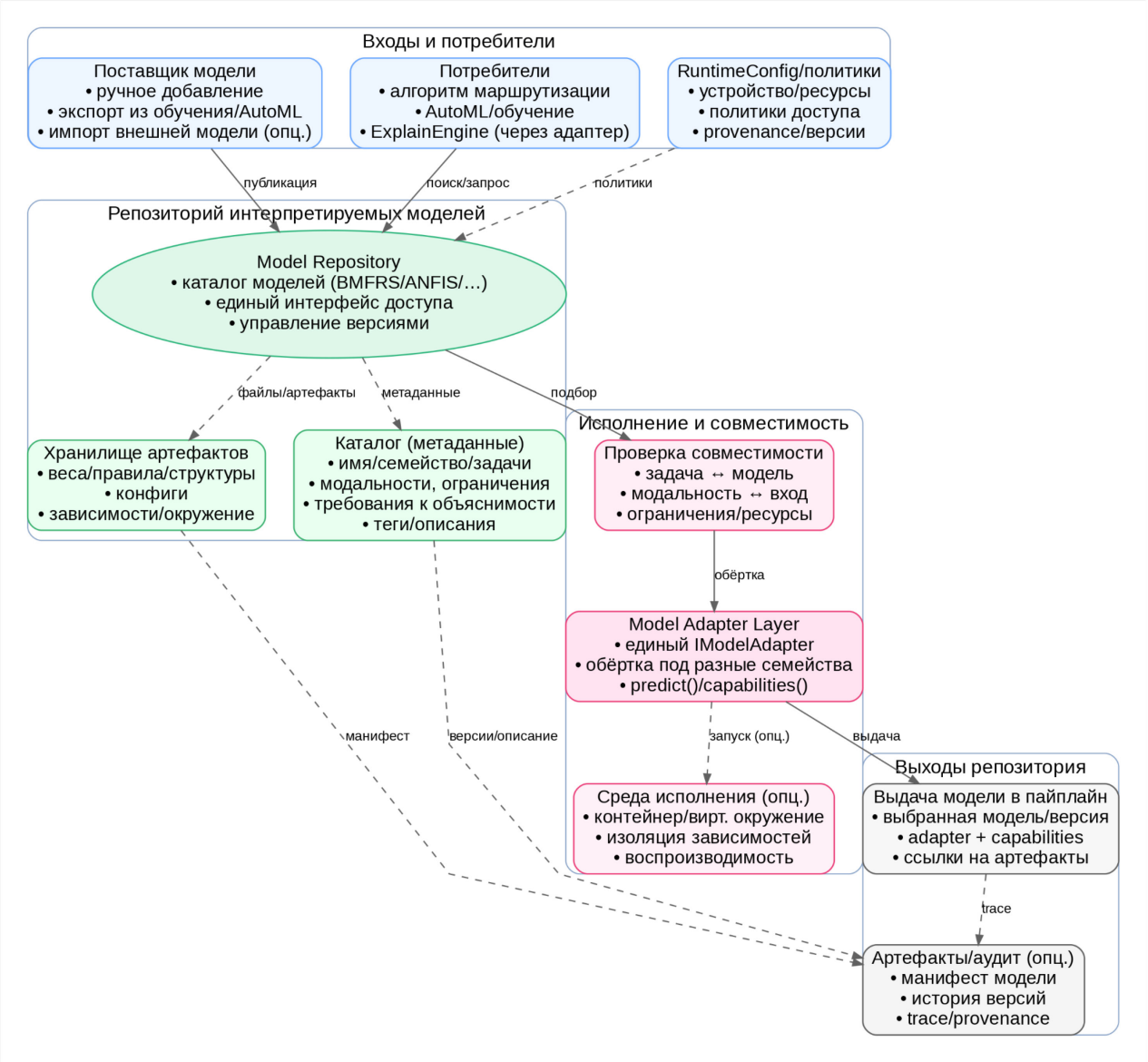


Рисунок 3.9 – Инфраструктурный слой воспроизводимости платформы

На этапе передачи модели в вычислительный контур репозиторий возвращает выбранную версию модели вместе с её метаданными, ссылками на связанные артефакты и адаптером доступа к модели. Так модель включается в платформу как формализованный объект, снабжённый описанием происхождения, версии, параметров и интерфейса подключения. Это принципиально важно для воспроизводимости: одна и та же архитектура при разных версиях весов, настройках обучения и наборах данных не должна рассматриваться как один и тот же объект анализа.

Каталог моделей выполняет функцию хранения уже имеющихся архитектур. В нём фиксируются сведения о типе модели, дате формирования версии и связях с соответствующими вычислительными артефактами. Хранилище артефактов предназначено для сохранения промежуточных и итоговых результатов выполнения. К числу таких объектов относятся обученные мо-

дели, конфигурации запуска, журналы исполнения, карты объяснений, метрические отчёты, результаты сравнительного анализа объяснителей и иные данные, возникающие в ходе работы конвейера. Централизованное хранение этих объектов позволяет восстановить итоговый результат и условия его получения.

Слой проверки обеспечивает корректность подключения модели к остальным компонентам системы. На этом уровне проверяются соответствие версии модели ожидаемому формату данных, наличие необходимых метаданных, корректность ссылок на артефакты и возможность использования модели в конкретном вычислительном сценарии. Благодаря этому исключается передача в конвейер неполных, несогласованных или устаревших объектов.

Репозиторий моделей и хранилище артефактов образуют инфраструктурную основу воспроизводимого контура платформы. Их функция состоит в том, чтобы связать вычислительное исполнение, версионность моделей, метаданные эксперимента, результаты объяснения и метрической проверки в единую проверяемую запись запуска. Такой записью выступает итоговый объект выполнения *ExplainRunResult*, включающий объяснение, значения метрик, сигналы доверительного анализа, предупреждения, манифест запуска, журналы и ссылки на сохранённые артефакты.

3.3.6. Локальный интерфейс просмотра отчётов и артефактов

Интерфейс предоставляет сводное представление результатов *AutoML* и *AutoXAI*. На рисунках 3.10–3.11 показаны экран итоговой сводки и экран анализа результатов отбора.

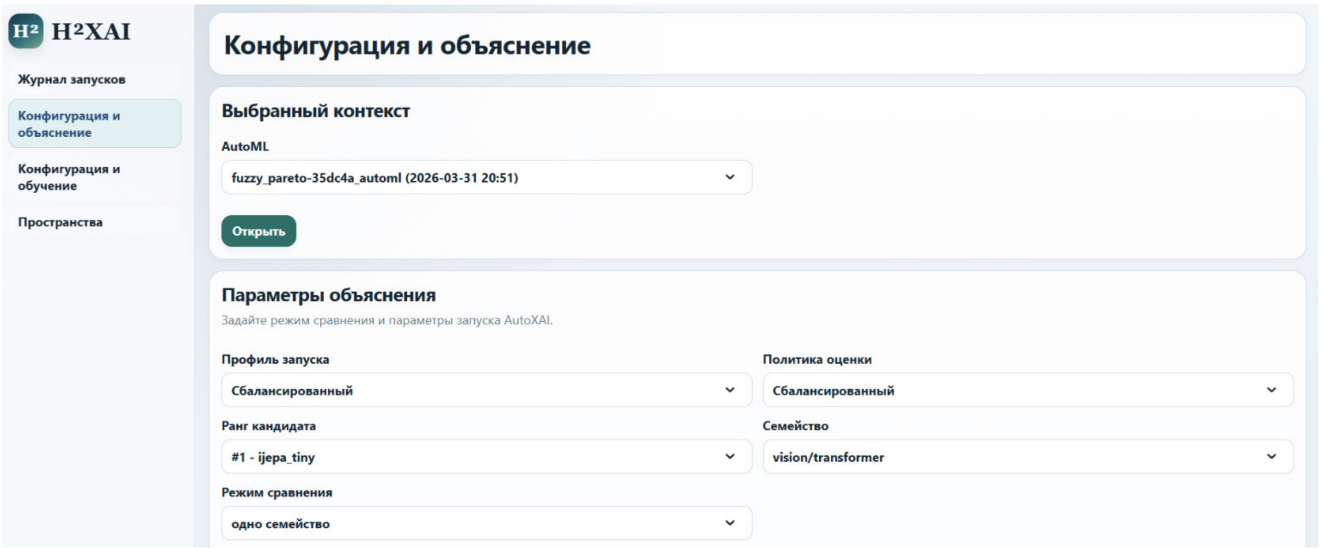


Рисунок 3.10 – Локальный интерфейс сводки по выбору модели и метода объяснения

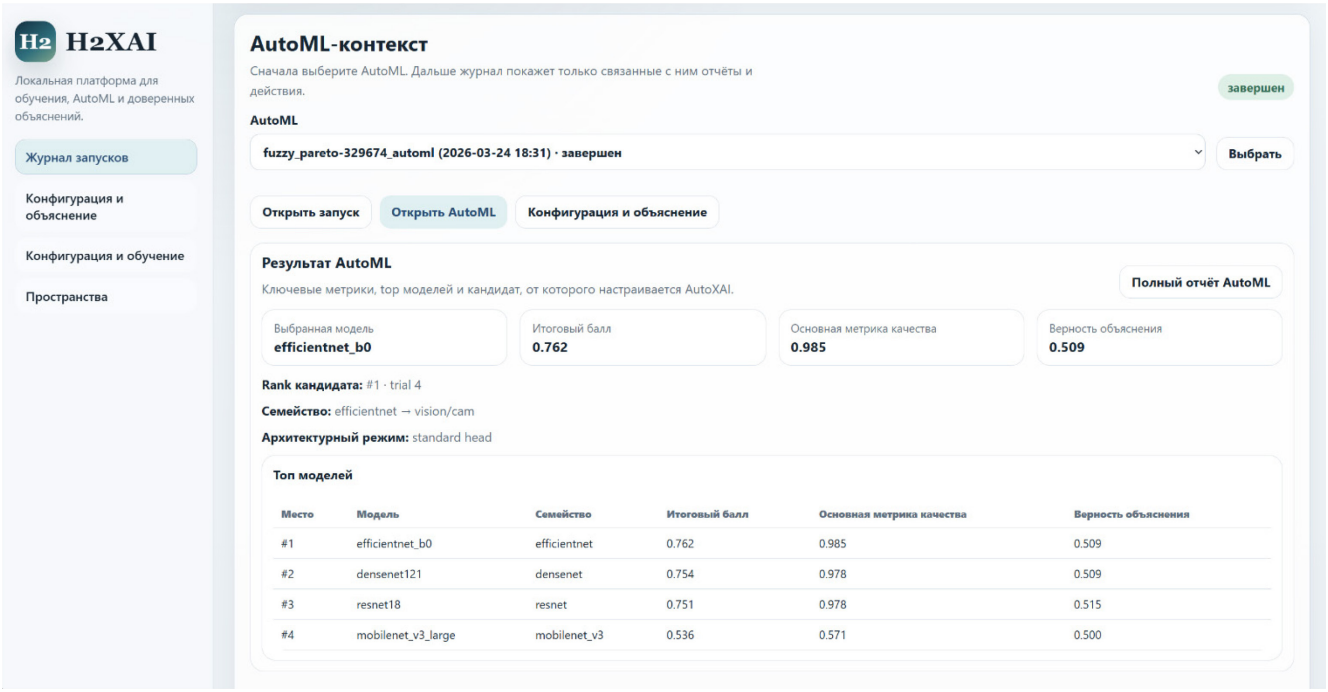


Рисунок 3.11 – Экран анализа результатов *AutoML* и *AutoXAI*

На рисунке 3.12 приведён экран, на котором выполняется сравнение методов объяснения для выбранной модели-кандидата. В таблице собраны значения основных показателей: верности объяснения, визуальной пригодности, чувствительности к разрушению параметров модели, устойчивости предсказания, устойчивости карты и согласованности методов. Такой формат позволяет сравнивать объяснители не по одному показателю, а по совокупности критериев.

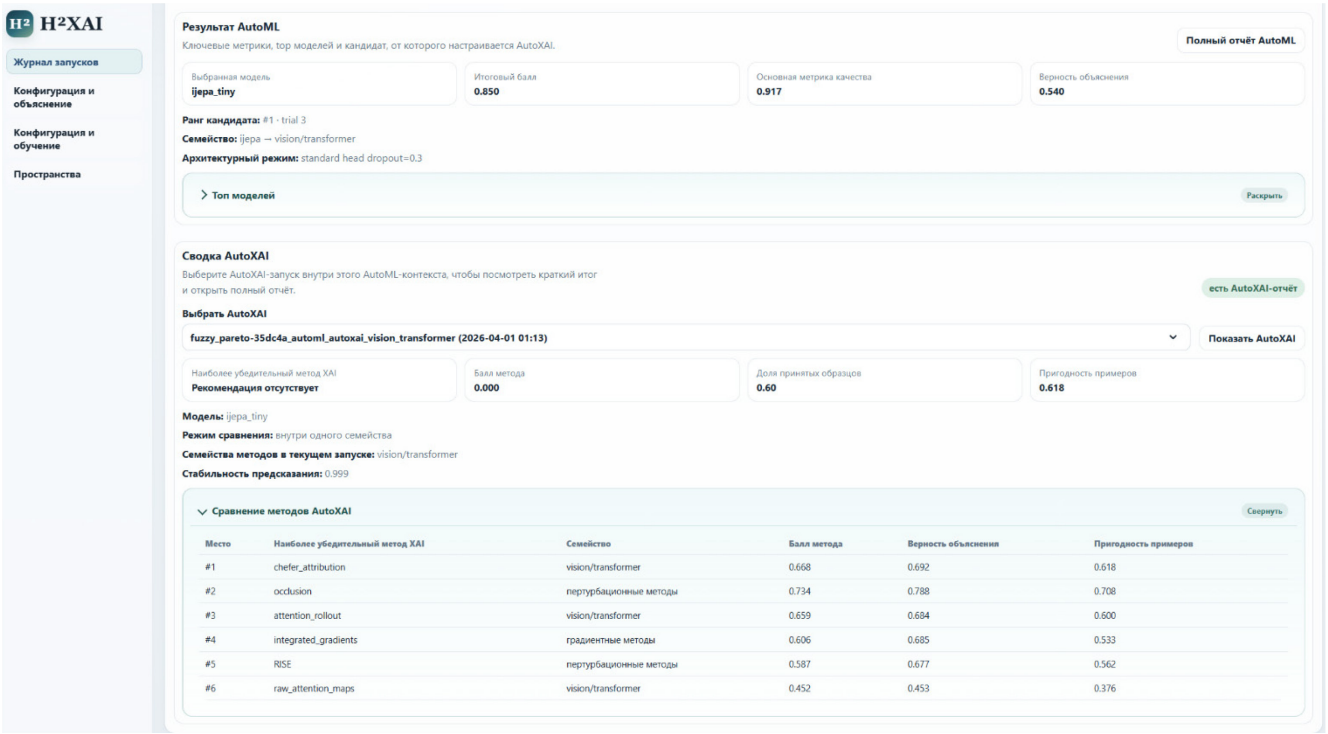


Рисунок 3.12 – Экран сравнений анализа *AutoXAI*

На рисунке 3.13 представлена предварительная настройка эксперимента перед запуском контура *AutoML*. Через этот интерфейс задаются исходные параметры обучения: набор данных, конфигурация и вычислительная среда. Экран используется как этап предварительной проверки, на котором пользователь может проверить настройки, доступность данных и готовность вычислительных ресурсов.

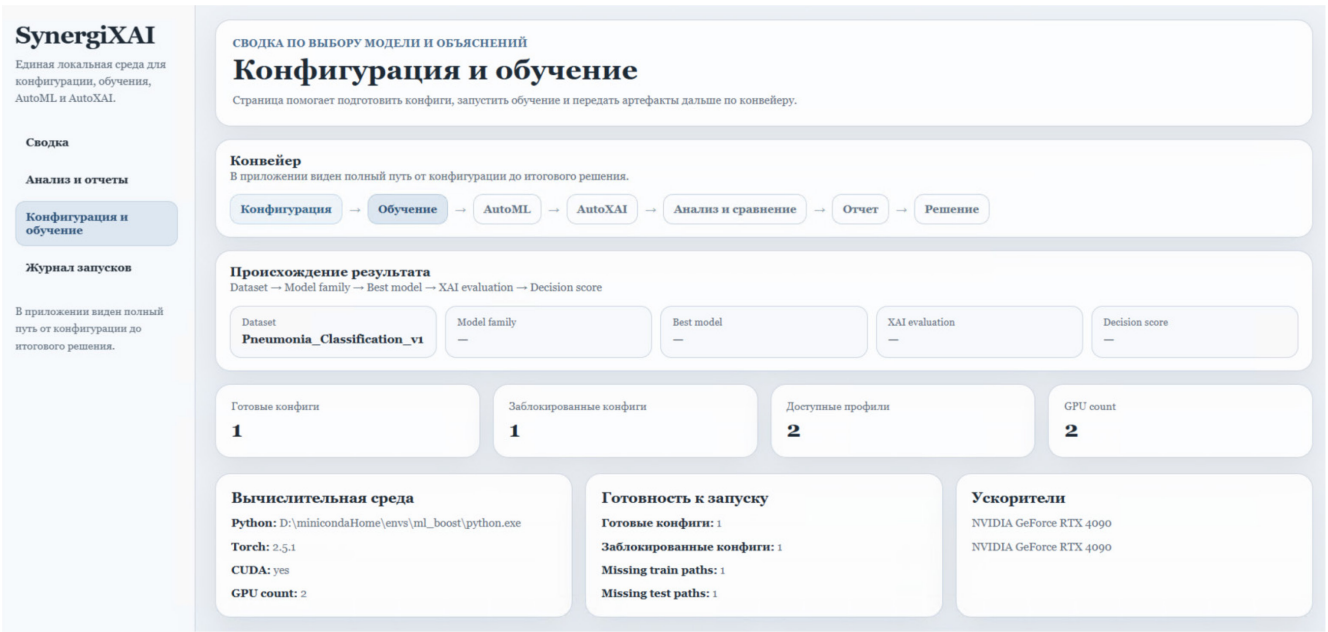


Рисунок 3.13 – Экран конфигурации обучения

На рисунке 3.14 представлен запуск обучения, предназначенный для формирования и выполнения конфигурации, где пользователь выбирает профиль обучения, указывает конфигурационный файл и среду выполнения. На экране также доступен каталог уже имеющихся профилей, которые задают разные сценарии обучения, включая базовый режим и режим с ограничениями. Дополнительно отображаются конфигурационные файлы с их параметрами и состоянием готовности, а также история последних запусков со статусами и сохранёнными контрольными точками. За счёт этого пользователь может прослеживать весь ход эксперимента и контролировать обучение модели.

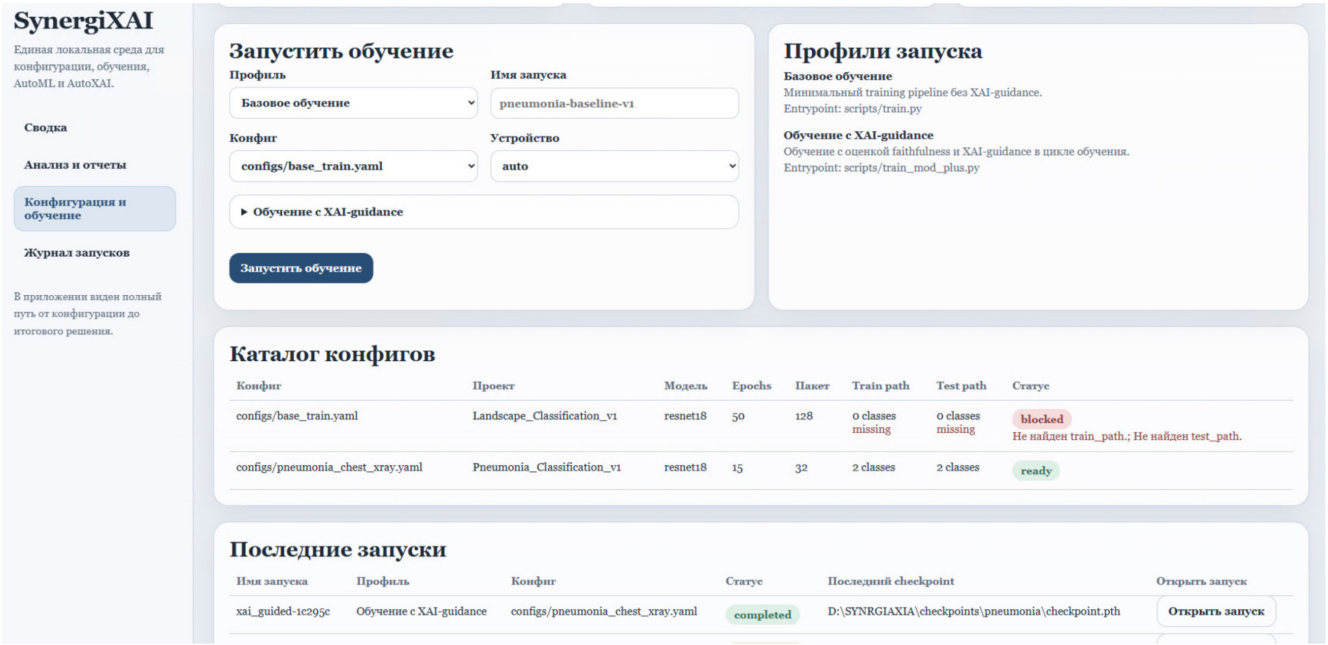


Рисунок 3.14 – Экран запуска обучения

3.4. Функциональная организация и маршрутизация

Раздел 3.4 объединяет подсистемы платформы в общий маршрут обработки запроса. Если в разделе 3.3 рассматривались функции платформы по отдельности, то далее они представлены как единый маршрут выполнения: от поступления запроса и построения плана до выбора модели, формирования объяснений, их проверки, агрегирования и сохранения итоговых результатов.

Организация маршрутизации запроса в платформе представлена на рисунке 3.15. Обработка начинается с поступления запроса в ядро управления, где на основе поставленной задачи или конфигурации выполнения формируется план. Далее через реестр *XAIRegistry* выбираются модели, объяснители, адаптеры и метрические протоколы. После этого выполняется модельный

контур, где используется уже подключённая модель либо запускается контур *AutoML*. Затем управление передаётся в *AutoXAI*-контур, где выбираются и применяются методы объяснения. Полученные объяснения проходят метрическую верификацию, доверительный анализ, формирование предупреждений и сохраняются вместе с параметрами запуска, метриками и ссылками на артефакты.

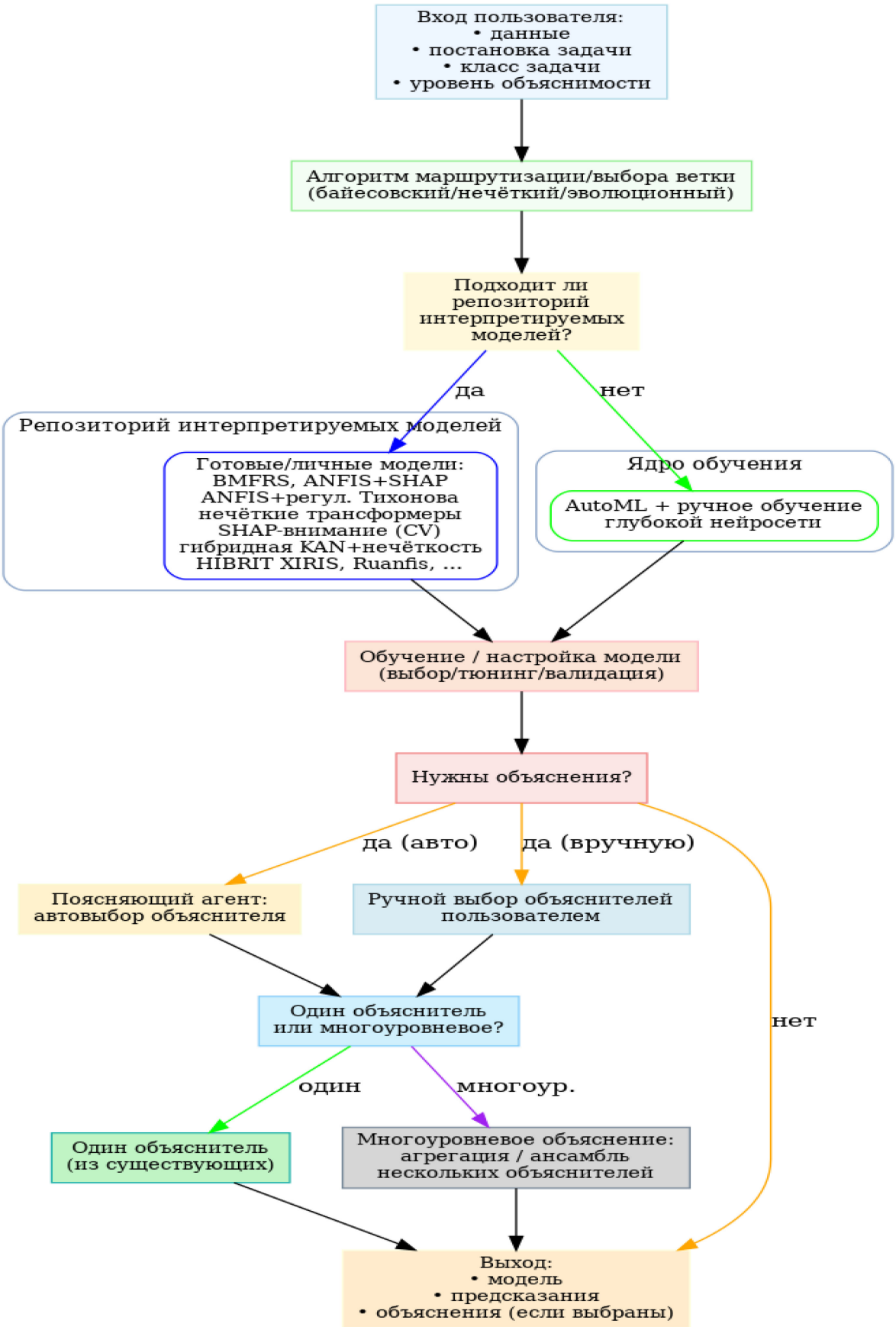


Рисунок 3.15 – Контур маршрутизации запроса в платформе

3.4.1. Связь AutoML и AutoXAI в маршруте выполнения

В маршруте выполнения платформы модельный и объяснительный контуры рассматриваются как связанные этапы одного вычислительного процесса. *AutoML*-контур отвечает за выбор, обучение или подключение модели-кандидата, тогда как *AutoXAI*-контур выполняет выбор, настройку и оценку метода объяснения для полученной модели. Поэтому переход к блоку интерпретации осуществляется только после того, как модель включена в конвейер как проверенный объект с известными метаданными, параметрами и результатами оценки.

Если постановка задачи допускает использование моделей с внутренней интерпретируемостью, запрос может быть направлен в репозиторий интерпретируемых моделей, включая нейро-нечёткие и иерархические архитектуры [89, 131, 175–180]. Если подходящая конфигурация отсутствует или требуется автоматизированный подбор модели, активируется контур *AutoML*. После выбора конфигурации или модели управление передаётся в *AutoXAI*, где определяется набор возможных методов объяснения, строятся объяснения и проводится их метрическая проверка.

Модель в платформе не оценивается только по прогностической точности. В маршрут выполнения включаются также пригодность формируемых объяснений, устойчивость результатов, доверительная оценка, предупреждения о возможной неприменимости метода и вычислительные ограничения. Поэтому итоговая конфигурация рассматривается как связка «модель — метод объяснения — условия выполнения», а не как отдельно выбранная модель с последующим произвольным объяснителем.

Формальная постановка совместного выбора модели и метода объяснения дана в разделе 2.9, а её программная реализация рассматривается в разделе 3.5 при описании *AutoML*-ядра.

3.4.2. Формирование объяснений при ансамблировании

При использовании нескольких методов объяснения для формирования объяснений результат работы должен представлять собой набор связанных карт значимости, таблиц вкладов или текстовых пояснений. В платформе такие результаты объединяются в композитный профиль объяснения, который сохраняет сведения о происхождении отдельных объяснений, их типе, параметрах построения, метрических оценках, согласованности и неопределённости объединённого результата.

Композитный профиль не следует трактовать как универсальную алгебру, сводящую любые виды объяснений к единому формату. Разные типы объяснений имеют различную природу: карты значимости описывают пространственное или признаковое распределение вклада, объяснения на уровне отдельных признаков задают вклад конкретных признаков, концептные объяснения работают с содержательными признаками более высокого уровня, а контрфактуальные объяснения описывают минимальные изменения входа. Поэтому в платформе используется типизированное объединение объяснительных свидетельств: однородные объяснения могут агрегироваться напрямую, а разнородные сохраняются и сопоставляются через общий профиль результата.

Для карт значимости допускается ансамблирование после приведения к сопоставимому пространственному и числовому представлению. Поскольку карты, полученные разными методами, могут иметь различное разрешение, масштаб и диапазон значений, перед агрегированием выполняется их пространственная и амплитудная унификация. Пусть $E^{(k)}$, $k = 1, \dots, N$, обозначает карту значимости, построенную k-м методом объяснения. Для приведения всех карт к единому пространственному разрешению используется оператор ресемплирования Res, реализуемый, например, билинейной интерполяцией. После этого выполняется робастная перцентильная нормализация, отображающая значения каждой карты в единый диапазон $[0,1]$. Нормированная карта $E'^{(k)}$ определяется выражением:

$$E'^{(k)}(i,j) = \frac{\mathrm{clip}\left(\mathrm{Res}\left(E^{(k)}\right)(i,j), p_{low}^{(k)}, p_{high}^{(k)}\right) - p_{low}^{(k)}}{p_{high}^{(k)} - p_{low}^{(k)} + \varepsilon}, \tag{3.24}$$

где $p_{low}^{(k)}$ и $p_{high}^{(k)}$ — нижний и верхний перцентили распределения значений k-й карты, $\mathrm{clip}(\cdot)$ — операция ограничения значений заданным интервалом, а $\varepsilon > 0$— малая константа, предотвращающая деление на нуль. Такая процедура позволяет ослабить влияние выбросов и привести карты, полученные разными *XAI*-методами, к сопоставимому метрическому пространству.

На основе множества подготовленных карт $E'^{(1)}, E'^{(2)}, \dots, E'^{(N)}$,платформа формирует несколько композитных артефактов, обеспечивающих многомерный анализ интерпретации.

Во-первых, строится консенсусная карта, отражающая усреднённое распределение значимости по всем методам:

$$E_{mean}(i,j) = \frac{1}{N}\sum_{k=1}^{N} E'^{(k)}(i,j). \tag{3.25}$$

Формула (3.25) представляет собой частный случай взвешенного усреднения при равных весах. Данная карта позволяет выявить области, которые в среднем признаются значимыми большинством объяснителей и тем самым служит базовой агрегированной формой ансамблевого объяснения.

Во-вторых, формируется карта строгого согласия, выделяющая области, одновременно поддержанные всеми методами объяснения. Для этого используется операция нечёткого пересечения в форме T-нормы Заде:

$$E_{int}(i,j) = \min_{k=1,\dots,N} E'^{(k)}(i,j). \tag{3.26}$$

Карта E_{int} носит более консервативный характер по сравнению с усреднением и акцентирует только те зоны, в отношении которых между объяснителями достигается наибольшая степень согласованности.

В-третьих, вычисляется карта пространственной неопределённости, характеризующая разброс оценок значимости между методами:

$$E_{var}(i,j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(E'^{(k)}(i,j) - E_{mean}(i,j)\right)^2}. \tag{3.27}$$

Высокие значения E_{var} указывают на области, в которых различные методы объяснения расходятся в оценке важности признаков. Такие зоны могут рассматриваться как области пониженной интерпретационной устойчивости и учитываться подсистемой *TrustEvaluator* при формировании итогового индекса доверия. По построению (3.27) задаёт поэлементное стандартное отклонение, то есть квадратный корень из дисперсионной оценки вида (2.56), вычисленной по пикселям карты.

3.4.3. Воспроизводимый выбор метода объяснения

Воспроизводимый выбор метода объяснения в платформе рассматривается как отдельная процедура сопоставления объяснителей по единому набору критериев. Метод объяснения не выбирается только по факту доступности в реестре или визуальной наглядности результата. Для каждого допустимого объяснителя оцениваются качество объяснения, устойчивость, согласованность с другими методами, применимость к данной модели и типу данных, а также возможность воспроизведения результата при повторном выполнении вычислительной процедуры.

На первом этапе через реестр *XAIRegistry* формируется множество методов, применимых к текущей постановке задачи. Из рассмотрения исключаются объяснители, несовместимые с типом модели, форматом входных данных, требуемым уровнем доступа к внутренним представлениям или ограничениями предметной области. Тем самым выбор выполняется не среди всех известных *XAI*-методов, а среди тех, которые корректны для данного вычислительного контекста.

На втором этапе для каждого допустимого метода рассчитываются показатели метрической верификации: верность объяснения, устойчивость карты или признаковых вкладов, согласованность с другими объяснителями и результаты контрольных проверок. Эти значения используются для сопоставления методов в единой шкале и позволяют отличать объяснитель, дающий устойчивый и проверяемый результат, от метода, который формирует визуально убедительное, но нестабильное или слабо согласованное объяснение.

Отдельно учитывается воспроизводимость объяснения. Она может оцениваться на серии повторных запусков одной и той же вычислительной процедуры, при которых анализируется стабильность получаемых карт, вкладов признаков или рангового положения метода среди других объяснителей. В базовом режиме этот показатель сохраняется как диагностическая характеристика. При явном задании минимально допустимого уровня воспроизводимости он может использоваться как дополнительное условие отбора метода.

3.4.4. Стратегии маршрутизации и построение плана выполнения

Для формирования последовательности операций обработки запроса используются стратегии маршрутизации. Их задача состоит в построении плана выполнения, то есть определении набора методов и порядка их применения. Независимо от выбранного способа построения плана его дальнейшее исполнение осуществляется единым управляющим модулем. Этот модуль не изменяет структуру плана, а последовательно реализует заданные в нём действия. Параметры стратегии и сформированный план фиксируются в документации запуска, что обеспечивает возможность повторения вычислений при одинаковых исходных условиях.

Платформа поддерживает три уровня маршрутизации:

1. Прямая маршрутизация — ручное задание метода и его параметров. Такой режим применяется при проведении сравнительных экспериментов и в регламентированных вычислительных процессах, где структура обработки заранее определена и не изменяется в ходе выполнения.

2. Маршрутизация на основе правил — выбор метода определяется системой формализованных условий. При этом учитываются свойства модели (например, её дифференцируемость

или архитектурный тип), характер решаемой задачи и ограничения по вычислительным ресурсам.

3. Оптимизационная маршрутизация (многокритериальный поиск) используется в тех случаях, когда план выполнения должен быть выбран по нескольким количественным критериям одновременно. Каждая допустимая конфигурация конвейера, включающая архитектуру модели, гиперпараметры обучения и выбор семейства методов объяснения, рассматривается как кандидат с двумя целевыми показателями — прогностическим качеством и верностью объяснения, а задача первичного поиска ставится как двухкритериальная оптимизация (2.96) над множеством конфигураций (2.85) по аналогии с задачами поиска архитектур нейронных сетей [107, 108, 181].

Для обеспечения воспроизводимости в артефактах запуска сохраняются: пространство поиска, состав целевых функций, ограничения, параметры оптимизатора, правила ранней остановки и сформированный план выполнения. Тем самым на этапе маршрутизации определяется множество допустимых кандидатов, а окончательная селекция осуществляется в *AutoML*-ядре.

3.5. Архитектура AutoML-ядра

3.5.1. Многокритериальная постановка и двухэтапная селекция

AutoML-ядро платформы выполняет не простой перебор моделей с выбором максимальной прогностической точности, а многокритериальную селекцию конфигураций вычислительного контура (см. рисунок 3.16). В качестве кандидата рассматривается архитектура модели и связанная с ней конфигурация объяснения: параметры обучения, способ подключения модели, допустимые методы объяснения, показатели качества объяснения и вычислительные ограничения.

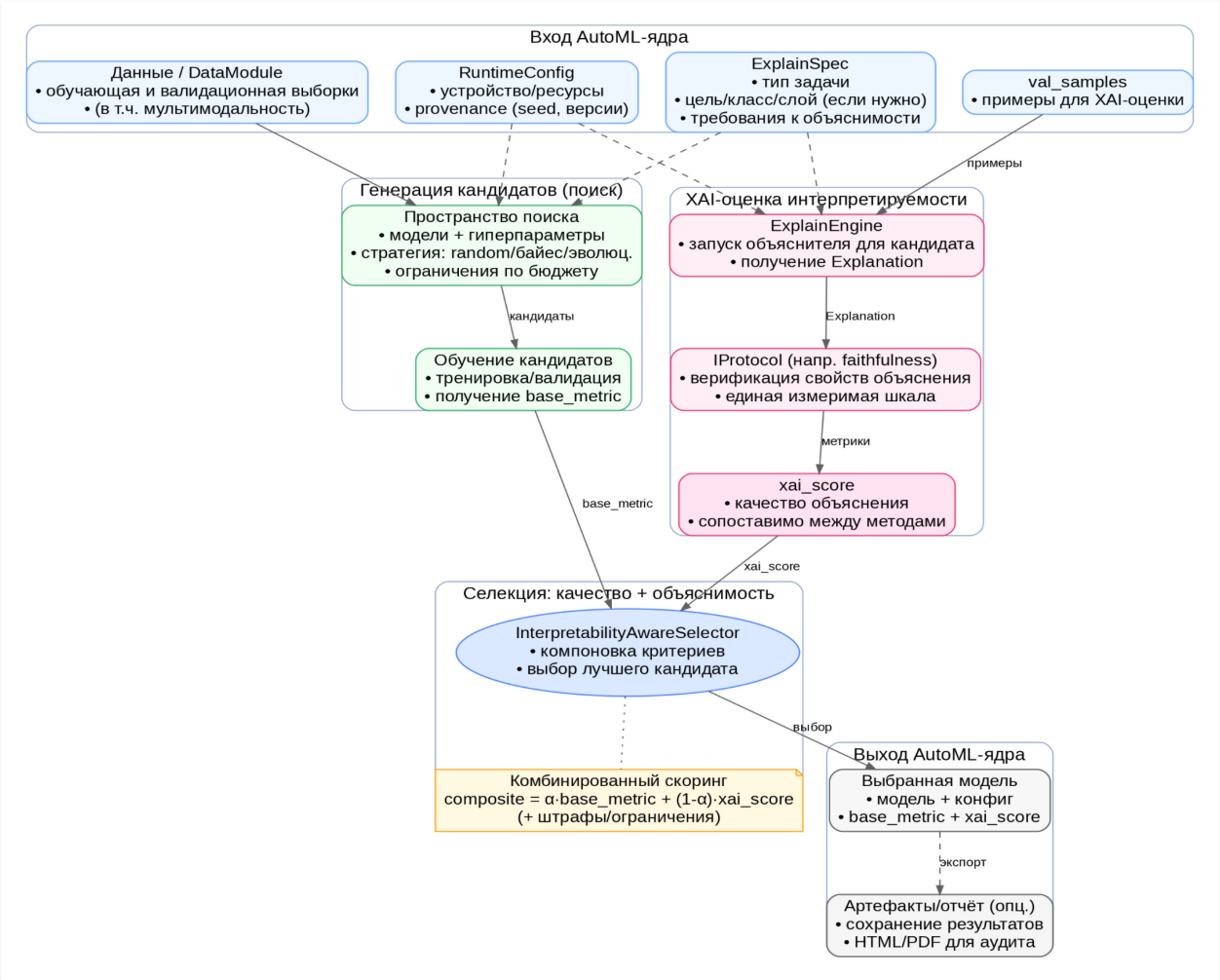


Рисунок 3.16 – Архитектура *AutoML*-ядра с многокритериальной селекцией

Селекция реализует двухэтапную процедуру, формально описанную в разделе 2.9. На первом этапе по оценкам, полученным при бюджете раннего скрининга, строится фронт Парето (2.97) по прогностическому качеству и верности объяснения. На втором этапе модуль *InterpretabilityAwareSelector* выбирает из фронта Парето конфигурацию по правилу (2.102), после чего она передаётся на исполнение и сохраняется вместе с артефактами.

Финальная селекция в платформе (см. рисунок 3.16) основана не на линейной свёртке критериев, а на последовательном вычислении частных показателей и их нечёткой агрегации. Модуль *InterpretabilityAwareSelector* вычисляет для каждого кандидата значения (2.92) и (2.93)–(2.95), фаззифицирует их и определяет итоговую оценку по (2.101). Такой подход позволяет учитывать компромиссы между прогностическим качеством модели, верностью объяснения, параметрической сложностью и вычислительной эффективностью в интерпретируемой лингвистической форме.

Многокритериальный поиск и финальная нечёткая селекция выполняются последовательно: сначала формируется множество Парето-эффективных кандидатов, затем из него выбирается конфигурация, используемая для исполнения и сохранения артефактов.

3.5.2. Частные критерии оценки кандидатов и бюджет поиска

Частные критерии селекции — прогностическое качество (2.92), верность объяснения (2.93), параметрическая простота (2.94) и вычислительная эффективность (2.95) — вычисляются в платформе по результатам обучения и проверки каждого кандидата.

Прогностическое качество рассчитывается на валидационной выборке по матрице ошибок (2.86)–(2.87) с макроусреднением (2.90)–(2.91). Весовые коэффициенты w_{Prec} и w_{Rec} в (2.92) задаются декларативно — параметрами контура *AutoML* или конфигурацией селектора, поэтому критерий настраивается под предметную область без изменения программного кода.

AutoML-поиск организован как многоступенчатая процедура. Бюджеты раннего скрининга B_s и полной оценки B_f вида (2.81)–(2.83) задаются в конфигурации запуска числом эпох и максимальным числом пакетов обучающей и проверочной выборок и сохраняются в артефактах эксперимента. Такая организация поиска сокращает вычислительные затраты на раннем этапе без отказа от детальной оценки наиболее перспективных кандидатов.

3.5.3. Нормированные показатели верности объяснения, сложности и вычислительной эффективности

Верность объяснения F вычисляется по (2.93) подсистемой метрической верификации (см. раздел 3.3.3) на основе метрик вставки и удаления.

Для нормировки параметрической сложности по (2.94) в проведённых экспериментах использовалось референсное значение $N_\theta^{\mathrm{ref}} = 10^6$ параметров. Опорное время в оценке вычислительной эффективности (2.95) задавалось эмпирически по вычислительным возможностям системы и составляло $\tau_{\mathrm{ref}} = 180$ с. При таком значении конфигурация, этап которой выполняется за три минуты, получает $\mathrm{Eff} = 0{,}5$.

3.5.4. Нечёткая агрегация критериев выбора

В модуле *InterpretabilityAwareSelector* частные критерии фаззифицируются функциями принадлежности (2.98)–(2.100): возрастающие функции описывают термы «высокое» и «очень высокое», убывающие — «низкое» и «нежелательно высокое», треугольные — промежуточные значения. Итоговая оценка кандидата вычисляется дефаззификацией Сугено нулевого порядка по (2.101).

Синглтон-следствия правил w_i интерпретируются как априорные полезности уровней предпочтения и задаются на основе экспертной оценки значимости соответствующих уровней качества модели.

В проведённых экспериментах использовались значения:

$$w_{vh} = 0{,}92, \qquad w_h = 0{,}75, \qquad w_m = 0{,}52, \qquad w_l = 0{,}25.$$

Такая схема агрегирования соответствует классическому нечёткому выводу типа Сугено и обеспечивает непрерывную итоговую оценку альтернатив [156].

3.6. Архитектурные свойства и ограничения

Архитектура платформы обладает свойством масштабируемости. Новые методы и компоненты могут подключаться без изменения базового контура исполнения. Это обеспечивается модульной организацией платформы и использованием контрактных интерфейсов, рассмотренных в предыдущих подразделах.

Внешние механизмы оркестрации на базе *MCP*, включая подключаемые *LLM*-компоненты, в рамках настоящей работы не рассматриваются как штатная часть базового воспроизводимого контура. Они образуют перспективный внешний слой расширения, способный повысить гибкость маршрутизации, автоматизировать выбор шагов вычислительного контура и адаптировать конфигурацию выполнения к контексту задачи. Вместе с тем подобные механизмы вносят дополнительную вариативность в процесс построения плана и потому требуют самостоятельной архитектурной и экспериментальной проверки с точки зрения трассируемости, детерминированности и воспроизводимости получаемых решений.

По этой причине в текущей реализации платформы формирование плана, выбор методов и оценка результатов опираются прежде всего на формализованные процедуры, явные правила,

независимые метрические протоколы и фиксируемые артефакты выполнения. Внешние интеллектуальные оркестраторы рассматриваются как допустимое направление дальнейшего развития платформы, но не как стабилизированный эксплуатационный слой базовой архитектуры.

3.7. Основные выводы по главе

В главе разработана архитектура платформы, предназначенной для воспроизводимой организации полного цикла работы с моделями искусственного интеллекта, их объяснениями и процедурами проверки качества. Архитектура описана в контекстном, компонентном и динамическом представлениях, что позволило определить состав платформы, функции её основных подсистем и логику их взаимодействия в едином вычислительном контуре.

В составе платформы выделены ядро управления, реестр методов объяснения *XAIRegistry*, подсистема *AutoXAI*, независимые метрические протоколы, контуры контроля допустимости и доверия, репозиторий моделей и хранилище артефактов. Показано, что объяснение в платформе включается в процедуру метрической проверки, сопоставления, доверительного анализа и сохранения вместе с параметрами запуска, предупреждениями и ссылками на артефакты.

Предложенная архитектура обеспечивает построение объяснений и их проверку, сравнение, типизированное объединение и воспроизводимое сохранение. Для случаев использования нескольких объяснителей введён композитный профиль результата, позволяющий сохранять однородные и разнородные объяснительные свидетельства без искусственного сведения всех типов объяснений к единой форме. Отдельно рассмотрена процедура воспроизводимого выбора метода объяснения, в которой учитываются качество, устойчивость, согласованность, применимость и воспроизводимость результата.

Дополнительно реализован предложенный в разделе 2.9 контур многокритериального выбора моделей и методов объяснения, учитывающий прогностическое качество, верность объяснения, структурную простоту и вычислительную эффективность. Тем самым создана программно-архитектурная основа платформы, обеспечивающая системную интеграцию *AutoML*, *AutoXAI*, метрической верификации, доверительного анализа и управления артефактами в единый воспроизводимый контур анализа данных.

4. АПРОБАЦИЯ ГЕТЕРОГЕННОЙ ГИБРИДНОЙ ПЛАТФОРМЫ НА БИОМЕДИЦИНСКИХ ЗАДАЧАХ

Настоящая глава посвящена апробации гетерогенной гибридной платформы при решении прикладных медицинских задач. Рассматриваемые сценарии выбраны с учётом высоких требований к объяснимости, надёжности и воспроизводимости результатов интеллектуального анализа данных. Платформа представляет собой программно-алгоритмический комплекс, в котором средства автоматизированного подбора моделей, процедуры интерпретации, оценка качества, калибровка доверия и сохранение результатов вычислений объединены в рамках единого контура обработки данных. Такой подход обеспечивает формирование согласованной конфигурации системы с учётом совокупности диагностически значимых критериев.

Выбор медицинской области обусловлен особенностями клинической практики, в которой результат работы модели должен быть понятен специалисту и пригоден для использования при принятии диагностического решения. Существенное значение имеют признаки, области изображения и вычислительные основания, определившие вывод системы, а также устойчивость полученного результата и границы его применимости. В этих условиях объяснимость и оценка доверия к результату приобретают значение обязательных характеристик интеллектуальной системы [17, 82, 182, 183].

В главе рассматриваются три предметные области: маммология, диагностика коронарных стенозов при ишемической болезни сердца и анализ рентгенологических изображений органов грудной клетки для выявления пневмонии и хронической обструктивной болезни лёгких. Указанные задачи различаются по типу данных, диагностическому содержанию и требованиям к моделям, однако объединяются общей постановкой, связанной с необходимостью получения интерпретируемого и практически применимого результата. В этом контексте платформа рассматривается как архитектурная основа построения доверенных медицинских интеллектуальных систем, в которых интерпретируемость учитывается при выборе и применении итогового решения.

4.1. Биомедицинская задача: маммология

4.1.1. Клинико-диагностические предпосылки применения ОИИ в маммологии

Рак молочной железы остается одной из наиболее значимых медико-социальных проблем современной медицины. Маммография является одним из основных методов его раннего выявления как в клинической практике, так и в рамках массовых профилактических обследований. Вместе с тем, несмотря на широкое применение и подтвержденную клиническую эффективность, маммографическое исследование имеет ряд объективных ограничений, влияющих на качество диагностики. По данным исследований, чувствительность маммографии при выявлении рака молочной железы в среднем составляет 75–90%, однако данный показатель существенно зависит от плотности ткани молочной железы и опыта специалиста [184]. Вследствие этого часть опухолевых образований может оставаться не выявленной либо получать неоднозначную интерпретацию.

Существенной особенностью маммографической диагностики является вариативность трактовки заключений даже при использовании стандартизированных систем описания, в частности, при применении шкалы *BI-RADS (Breast Imaging-Reporting and Data System)* (см. приложение А). Для оценки согласованности интерпретации в клинических исследованиях используется коэффициент внутриклассовой корреляции. Согласованность между врачами, выполнявшими первичную оценку, и врачами, проводившими первый пересмотр, находилась в пределах 0,836–0,875 при оценке по полной шкале *BI-RADS*. Аналогичная согласованность между врачами, выполнявшими первичную оценку, и экспертами составляла 0,838–0,879 [185]. Это свидетельствует о возможности риска диагностических расхождений, способных приводить к пропуску патологических изменений.

Дополнительные трудности возникают при высокой плотности ткани молочной железы. Несмотря на то, что маммография считается основным методом ранней диагностики, исследования показывают, что доля ложноотрицательных результатов составляет от 10% до 30% [186]. Сложность интерпретации также связана со слабо выраженными признаками патологии, включая микрокальцинаты и архитектурные искажения, малозаметные на фоне нормальных анатомических структур. Точность интерпретации в таких условиях определяется качеством визуализации, профессиональным опытом врача, а также факторами рабочей нагрузки и утомления [187].

Значимость данной проблемы возрастает в условиях массовых профилактических обследований. По имеющимся данным, до 97% результатов профилактической маммографии не содержат отклонений от нормы, тогда как патологические состояния выявляются в среднем в 6,9%

случаев, а злокачественные новообразования молочной железы — примерно в 1% исследований [187]. Врач-рентгенолог вынужден анализировать значительный поток изображений без патологических изменений, что увеличивает нагрузку и повышает вероятность пропуска малозаметных аномалий. Одним из способов повышения надежности диагностики является чтение маммографических исследований двумя независимыми специалистами [187]. Однако такой подход сопровождается увеличением трудозатрат, времени анализа и нагрузки на медицинский персонал.

В этих условиях возрастает значение средств автоматизации анализа маммографических изображений. По данным клинических исследований, средняя продолжительность двойного чтения при использовании алгоритмов ИИ снижается на 66,4% по сравнению с традиционным двойным чтением врачами [188]. При этом в 72,9% случаев врачи-рентгенологи согласились с результатами анализа алгоритма, а уровень согласованности между экспертами и алгоритмом может достигать 84,1% [188]. Эти данные указывают на перспективность интеллектуальных систем как средства поддержки диагностических решений.

Вместе с тем существенным ограничением существующих решений остается недостаточная объяснимость получаемых результатов. Во многих случаях алгоритм формирует итоговый результат без раскрытия признаков изображения, повлиявших на вывод, и без указания степени устойчивости результата. Для медицинской практики это обстоятельство имеет принципиальное значение, поскольку автоматизированный анализ должен сопровождаться интерпретацией, понятной врачу и соотносимой с клинически значимыми признаками.

Потребность в объяснимых интеллектуальных системах в маммологии определяется как клиническими, так и организационно-техническими факторами. Практически значимым является получение результата, который может быть использован врачом в процессе принятия решения и сопоставлен с визуальными признаками на изображении. С инженерной точки зрения необходимы решения, в которых процедуры интерпретации не приводят к чрезмерному увеличению вычислительных затрат и сохраняют применимость в реальных условиях эксплуатации, включая сценарии с ограниченными вычислительными ресурсами.

Особое значение данное требование имеет в условиях обеспечения доступности диагностики в удаленных регионах. В этих случаях востребованы решения, пригодные для применения в передвижных диагностических комплексах и иных инфраструктурно ограниченных условиях, где существенны надежность алгоритмов, вычислительная экономичность и интерпретируемость результата. Это обусловливает актуальность разработки систем, сочетающих высокое качество автоматизированного анализа маммографических изображений с интерпретируемостью, вычислительной эффективностью и возможностью применения в реальных условиях медицинской практики.

4.1.2. Клинико-рентгенологические основы маммографической диагностики

Маммография представляет собой специализированный метод рентгеновской визуализации молочных желез, основанный на различиях в коэффициентах ослабления рентгеновского излучения биологическими тканями различного морфологического состава [189]. За счет неодинакового поглощения и прохождения рентгеновских лучей через нормальные и патологически измененные структуры молочной железы на изображении формируется дифференцированная рентгенологическая картина, позволяющая выявлять признаки доброкачественных и злокачественных образований [189, 190].

Исследование выполняется с использованием маммографа - специализированной рентгенодиагностической системы, предназначенной для получения высококонтрастных изображений ткани молочной железы [189]. В процессе обследования молочная железа размещается на опорной платформе и подвергается дозированной компрессии, что необходимо для уменьшения толщины исследуемого объекта, снижения рассеяния излучения, повышения пространственного разрешения и улучшения качества визуализации [191]. После этого рентгеновская трубка генерирует излучение, проходящее через ткани молочной железы и регистрируемое либо на пленке в аналоговых системах, либо цифровыми детекторами в современных цифровых маммографических комплексах [189, 190].

Полученные изображения используются для последующего диагностического анализа. В цифровых системах маммограммы передаются на рабочую станцию для обработки, визуализации и интерпретации, в аналоговых — фиксируются на плёночном носителе [190]. Итоговое заключение формируется врачом-рентгенологом на основании анализа визуализируемых структурных изменений [190].

Представленная в приложении А таблица А.1 описывает стандартизацию описания результатов маммографического исследования по системе BI-RADS, обеспечивающей унифицированную классификацию выявленных изменений и формулировку диагностического заключения [192]. Использование данной системы позволяет соотносить рентгенологические признаки с категориями клинического риска и соответствующей тактикой дальнейшего ведения пациентки. Категории *BI-RADS* варьируют от 0 до 6 и отражают степень определенности заключения, вероятность злокачественного процесса и необходимость дополнительных диагностических мероприятий.

По характеру рентгенологической плотности ткани выделяют четыре типа молочных желез согласно классификации ACR (A, B, C и D) [192]; их характеристика приведена в таблице 4.1. Указанная классификация имеет существенное диагностическое значение, поскольку плотность

ткани непосредственно влияет на чувствительность маммографии и вероятность визуализации патологических изменений. При типах *A* и *B* чувствительность метода остается высокой, тогда как при типах *C* и *D* она снижается, что затрудняет выявление малых образований и требует применения дополнительных методов визуализации, прежде всего ультразвукового исследования [192].

Таблица 4.1 – Типы плотности ткани молочных желез

Класс	**Описание**	**Чувствительность**
A	Молочная железа практически полностью представлена жировой тканью (низкая плотность – термин «плотность» здесь и далее в данном контексте характеризует степень ослабления рентгеновского излучения при прохождении через ткань молочной железы).	Маммография обладает высокой чувствительностью.
B	На фоне жировой ткани определяются отдельные участки фиброзножелезистой (низкая плотность).	Маммография обладает высокой чувствительностью.
C	Молочная железа представлена неоднородной железисто-фиброзной тканью (гетерогенная плотность), либо имеются отдельные достаточно плотные участки в молочной железе, способные скрывать небольшие образования (повышенная плотность).	Маммография обладает низкой чувствительностью, требуется проведение УЗИ.
D	Молочная железа представлена железистой тканью (высокая плотность).	Маммография обладает низкой чувствительностью, требуется проведение УЗИ.
N.B	Наличие эндопротезов молочных желез указывается дополнительно!	-

Одной из основных причин диагностических затруднений при анализе маммограмм является высокая плотность ткани молочной железы. При преобладании железистой и фиброзной ткани рентгенологическое изображение становится менее однородным, а визуальная дифференциация отдельных структур усложняется. Поскольку плотная ткань снижает чувствительность маммографии, это осложняет интерпретацию изображений и в ряде случаев требует привлечения дополнительных диагностических методов.

Предметная область маммографической диагностики характеризуется рядом клинико-рентгенологических особенностей, непосредственно влияющих на процесс интерпретации изображений. Вышеперечисленные особенности определяют требования к системам автоматизиро-

ванного анализа маммограмм, включая чувствительность к слабо выраженным признакам патологии, устойчивость к вариациям плотности ткани и возможность интерпретации получаемых результатов в клинически значимой форме.

4.1.3. Инфраструктурные ограничения мобильной маммографической диагностики

Для территорий, где наблюдается ограниченная доступность диагностических обследований, удалённость населённых пунктов, неравномерное распределение медицинских учреждений и потребность в регулярном выездном обслуживании населения, мобильные формы оказания медицинской помощи приобретают особую значимость [193, 194]. Суть подхода заключается в применении передвижных медицинских комплексов, которые дают возможность проводить необходимые обследования непосредственно по месту проживания пациентов, без необходимости обращаться в крупные стационарные медицинские центры. Наибольшую эффективность мобильный формат демонстрирует при организации профилактических и скрининговых исследований, в том числе маммографии. Его практическая роль особенно велика в регионах, где создание стационарной медицинской инфраструктуры связано со значительными финансовыми и организационными затратами [194, 195].

Сельские и труднодоступные территории характеризуются сочетанием организационных и инфраструктурных ограничений. К числу наиболее значимых факторов относятся недостаточный уровень цифровизации, нестабильный сетевой доступ, зависимость транспортной логистики от погодных условий и ограниченность вычислительных ресурсов. При такой организации работы передвижной медицинский комплекс служит не только средством доставки оборудования, но и автономной рабочей средой, в которой выполняются обследования, сохраняются диагностические данные и обеспечивается их последующая передача в медицинские информационные системы [194].

В работе используются передвижные комплексы «ЛУИДОР», рассчитанные на эксплуатацию в условиях ограниченной инфраструктуры и применяемые мобильными медицинскими бригадами [196]. Их конструктивные и инженерные решения обеспечивают автономный или полуавтономный режим работы вне стационарных учреждений. По этой причине такой комплекс следует рассматривать как технологическую среду, в пределах которой должны работать диагностические, информационные и аналитические модули, включая систему автоматизированного анализа маммографических изображений.

Особенности мобильной маммографической диагностики определяют требования к системе поддержки принятия врачебных решений. Для рассматриваемого проекта принципиален

отказ от сложной внешней вычислительной архитектуры, постоянного высокоскоростного сетевого подключения и облачных вычислительных сервисов. Рациональным решением является включение модуля анализа в существующий контур работы с медицинскими изображениями с использованием стандарта *DICOM* и взаимодействием с системой *PACS* как средством хранения, передачи и дополнительной обработки данных.

Мобильная диагностика предполагает экономное использование вычислительных, сетевых и организационных ресурсов. Поэтому разрабатываемая система ориентирована на локальную работу в инфраструктуре медицинской организации и использование данных, формируемых в ходе стандартных диагностических процедур. Это позволяет применять систему в передвижных маммографических комплексах без существенного изменения сложившегося диагностического процесса.

Следовательно, инфраструктурные ограничения мобильной маммографической диагностики требуют разработки решений, сочетающих вычислительную экономичность, локальный характер работы, совместимость с существующими форматами медицинских данных и практическую пригодность в условиях ограниченных ресурсов. Учет этих требований необходим при создании системы автоматизированного анализа маммографических изображений, предназначенной для реального контура эксплуатации.

4.1.4. Современные подходы к автоматизированному анализу маммографических изображений

После получения маммографических изображений их интерпретация выполняется врачом-рентгенологом на рабочей станции, связанной с системой *PACS*. В ходе анализа оцениваются структура ткани молочной железы, наличие асимметрий, микрокальцинатов, архитектурных искажений и объемных образований. По совокупности выявленных признаков формируется диагностическое заключение с присвоением категории по системе *BI-RADS*, которое затем используется при определении тактики ведения пациентки и выборе дополнительных методов обследования [192].

Обычно результаты исследования фиксируются в виде текстового заключения, содержащего описание структуры молочной железы, категорию маммографической плотности, характеристику выявленных образований, категорию *BI-RADS* и рекомендации по дальнейшему обследованию [192]. При этом хранящееся в системе *PACS* изображение, как правило, не сопровождается детальной визуальной разметкой подозрительных участков. Это затрудняет повторную оценку исследования, сопоставление маммограмм в динамике и использование изображения как

наглядной основы для обсуждения результата.

Для повышения воспроизводимости диагностики и снижения нагрузки на врача в маммографии применяются системы компьютерной поддержки принятия решений. Их основное назначение состоит в автоматизированном выявлении патологически значимых изменений, локализации подозрительных участков и формировании дополнительных количественных характеристик, учитываемых при интерпретации исследования.

Одним из наиболее известных решений является *ProFound AI* — продукт компании *iCAD* (США), представляющий собой проприетарное коммерческое решение. Система использует методы глубокого обучения для обнаружения мягкотканных образований и микрокальцинатов, а также формирует количественные оценки отдельных очагов и исследования в целом. Дополнительно предусмотрены автоматизированная оценка плотности молочной железы и краткосрочная оценка индивидуального риска развития рака молочной железы [197].

Система *Genius AI* разработана компанией *Hologic* (США) и также относится к проприетарным коммерческим решениям. Она ориентирована преимущественно на анализ изображений цифрового томосинтеза молочной железы. Алгоритмы используются для выявления опухолевых образований, микрокальцинатов и других клинически значимых изменений, а результаты отображаются непосредственно в контуре врачебного чтения. В расширенных конфигурациях поддерживаются анализ предшествующих обследований, автоматизированная оценка плотности ткани и контроль качества изображений [198].

Алгоритм *Lunit INSIGHT MMG* создан компанией *Lunit* (Республика Корея) и представляет собой проприетарное коммерческое решение. Он предназначен для автоматического обнаружения подозрительных областей на маммограммах и расчета количественной оценки вероятности патологии. Система выявляет опухолевые образования, микрокальцинаты, асимметрию тканей и архитектурные искажения. Результаты анализа представляются в виде локализации подозрительных участков и числового показателя *abnormality score* [199].

Отечественным примером является платформа *SberMedAI*, разработанная компанией СберМедИИ (Россия) и относящаяся к проприетарным коммерческим решениям. В направлении маммографии она предусматривает анализ изображений молочных желез для выявления признаков патологических изменений, а также средства технического контроля качества исследования по критериям *PGMI*. Тем самым автоматизированный анализ используется как для поиска подозрительных очагов, так и для оценки качества исходных диагностических данных [200].

Отдельное направление исследований связано с применением визуальных трансформеров к анализу медицинских изображений. Механизм внимания позволяет учитывать пространственно удалённые зависимости и глобальную структуру входных данных [110]. В маммографии примером такого подхода служит *MammoFormer*, где исследованы архитектуры визуальных

трансформеров и их сочетание со средствами объяснимого искусственного интеллекта [201]. Механизм внимания не отменяет потребности в других методах объяснения.

Ещё одно направление составляет совместная обработка медицинских изображений и текстовых описаний визуально-языковыми моделями. В *Mammo-CLIP* общее представление строится по парам маммограмм и соответствующих им рентгенологических описаний. Это позволяет классифицировать и локализовать маммографические признаки и связывать области изображения с текстовыми понятиями [202]. Такие решения относятся к мультимодальным системам, в которых требуется согласовывать информацию различных модальностей и оценивать межмодальное расхождение [118, 146]. Применение визуально-языковых моделей в рассматриваемом мобильном диагностическом сценарии требует учёта вычислительных ограничений, локальной обработки медицинских данных и отдельной проверки достоверности формируемых текстов.

Проведенный обзор показывает, что современные подходы к автоматизированному анализу маммографических изображений сосредоточены прежде всего на выявлении подозрительных изменений, визуальном выделении областей интереса и расчете вероятности патологии. В отдельных системах дополнительно реализованы оценка плотности ткани, контроль качества исследования и анализ предшествующих обследований. Однако интерпретация результата в большинстве случаев сводится к указанию подозрительной области и числовому показателю риска. Средства интерпретируемого обоснования вывода алгоритма, пригодного для практической работы врача, развиты ограниченно. Кроме того, все рассмотренные решения относятся к закрытым промышленным системам, что ограничивает возможности независимой модификации алгоритмов и их адаптации к специализированным условиям эксплуатации. В меньшей степени проработан и вопрос локального функционирования таких систем в составе мобильной и ресурсно-ограниченной диагностической инфраструктуры.

Следовательно, сохраняется задача разработки системы автоматизированного анализа маммографических изображений, сочетающей выявление патологических признаков с интерпретируемым сопровождением результата, совместимостью с существующим контуром работы с медицинскими изображениями и возможностью практического применения в передвижных медицинских комплексах.

4.1.5. Постановка задачи

В настоящем исследовании решается задача построения прикладного решения для автоматизированного анализа маммографических изображений с использованием гетерогенной ги-

бридной платформы. Платформа используется как средство выбора модели, организации обучения, подбора механизма интерпретации и формирования воспроизводимого контура обработки данных для конкретной медицинской задачи. На её основе должно быть получено прикладное решение, предназначенное для поддержки маммографической диагностики в условиях мобильной медицинской инфраструктуры.

Рассматриваемое решение ориентировано на анализ маммографических изображений, получаемых в передвижном медицинском комплексе «ЛУИДОР» [196], с целью выявления патологических изменений в тканях молочной железы. Предполагаемая область применения — поддержка работы врача-рентгенолога при интерпретации исследований в условиях выездной медицинской практики.

Функционирование решения предполагается в контуре действующей инфраструктуры хранения и передачи медицинских изображений *PACS*. Маммографические исследования, выполняемые в мобильном медицинском кабинете, передаются в формате *DICOM* в *PACS*, после чего поступают на автоматизированную обработку. Такая схема обеспечивает встраивание решения в существующий диагностический процесс без дублирования оборудования, создания отдельного архива изображений и изменения стандартной процедуры обследования пациентов.

В исследовании используются анонимизированные медицинские данные, предоставленные в рамках сотрудничества с ГБУЗ Московской области «Дубненская больница». На их основе формируется набор маммографических изображений, включающий исследования без патологических изменений и исследования с признаками аномалий. Подготовленные данные используются для обучения и последующей оценки моделей, отбираемых средствами платформы.

Построенное решение должно обеспечивать автоматическую локализацию патологических областей на маммографических изображениях и формирование объяснения полученного результата. Наличие механизма интерпретации требуется для соотнесения вывода модели с визуально наблюдаемыми признаками на изображении и использования результата в качестве вспомогательного средства при врачебной оценке исследования.

К разрабатываемому решению предъявляются следующие требования: совместимость с контуром *PACS*/*DICOM*, возможность локального функционирования в условиях ограниченных вычислительных ресурсов, автоматическая локализация подозрительных областей на маммограммах и наличие интерпретируемого сопровождения результата. Следовательно, задача включает не только достижение приемлемого качества детекции, но и построение рабочего процесса, пригодного для практического применения в составе мобильного диагностического комплекса. Для решения поставленной задачи необходимо:

1. подготовить и структурировать набор маммографических изображений для решения задачи автоматизированного анализа;

2. выполнить разметку маммографических изображений с выделением областей патологических изменений;
3. сформировать набор данных, пригодный для обучения и оценки моделей;
4. с использованием платформы подобрать и обучить модель детекции патологических областей на маммографических изображениях;
5. определить и реализовать механизм объяснения результатов работы модели;
6. построить прикладное решение, совместимое с контуром *PACS*/*DICOM* и пригодное для эксплуатации в передвижном медицинском комплексе.

4.1.6. Практическая реализация

Для построения прикладного решения использовался набор маммографических исследований, предоставленный медицинской организацией в формате *PNG*. До передачи данных медицинское учреждение выполнило анонимизацию, в результате которой были удалены персональные данные пациентов и иные идентифицирующие сведения. Последующая работа проводилась исключительно с обезличенными медицинскими изображениями.

Исходные данные представляли собой цифровые маммографические изображения молочной железы, полученные в стандартных диагностических проекциях. В набор вошли маммографические исследования как без патологических изменений, так и с признаками аномалий. Для клинической интерпретации выявленных изменений использовались категории *BI-RADS* 0–5. На этапе построения обучающего набора целевыми областями считались участки изображения, содержащие рентгенологические признаки, требующие внимания врача-рентгенолога.

Разметка изображений выполнялась под руководством врачей-рентгенологов с использованием локально развернутой открытой платформы *Label Studio* [203]. Локальное размещение системы позволило организовать процесс аннотирования без передачи медицинских данных во внешние облачные сервисы. Разметка проводилась в веб-интерфейсе платформы с помощью ограничивающих прямоугольников (*bounding boxes*), внутри которых выделялись области с признаками патологических изменений. Подготовка датасета выполнялась совместно с врачами-экспертами, что обеспечило соответствие аннотаций клиническим диагностическим критериям. Сложность разметки заключалась в нахождении всех патологически аномальных областей с последующей проверкой врачами-рентгенологами и нахождением общего времени для просмотра всех размеченных данных. Платформа использовалась как средство организации разметки, распределения задач и централизованного хранения результатов аннотирования (см. рисунок 4.1).

Сформированные аннотации затем экспортировались в формат, совместимый с задачами детекции объектов, включая *COCO*.

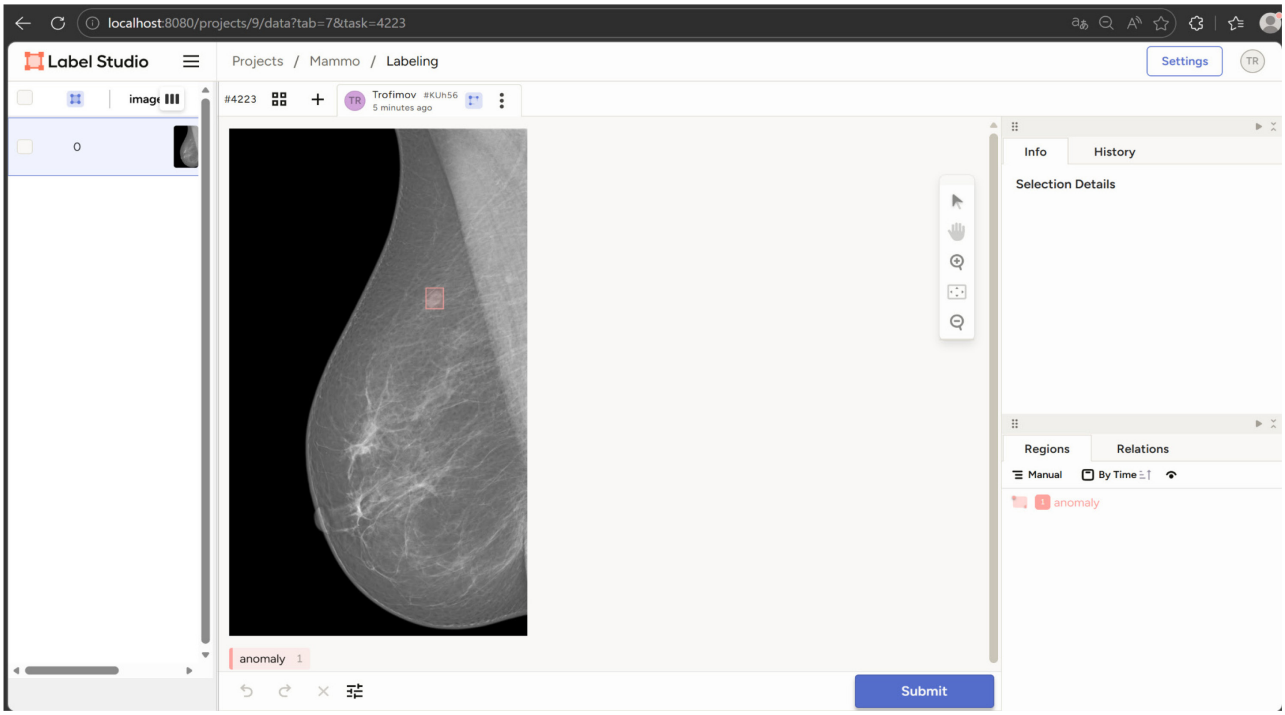


Рисунок 4.1 – Пример разметки

В обучающий набор были включены все 1150 исследований в различных проекциях: 1612 изображений с патологическими изменениями и 2823 изображения без них. Корректность разметки проверялась врачом-рентгенологом. Такая верификация позволила подтвердить релевантность выделенных областей и исключить неоднозначные либо технически неудовлетворительные исследования.

Подготовленный набор данных был загружен на платформу, используемую в качестве технологической среды для построения диагностического конвейера. С помощью модуля *AutoML* проведён отбор модели по совокупности метрик качества и ресурсоемкости. Особое внимание уделялось требованиям вычислительной экономичности, поскольку создаваемое решение ориентировано на применение в составе передвижного медицинского комплекса. По результатам отбора наиболее приемлемой признана архитектура *YOLOv8* [204] (см. рисунок 4.2).

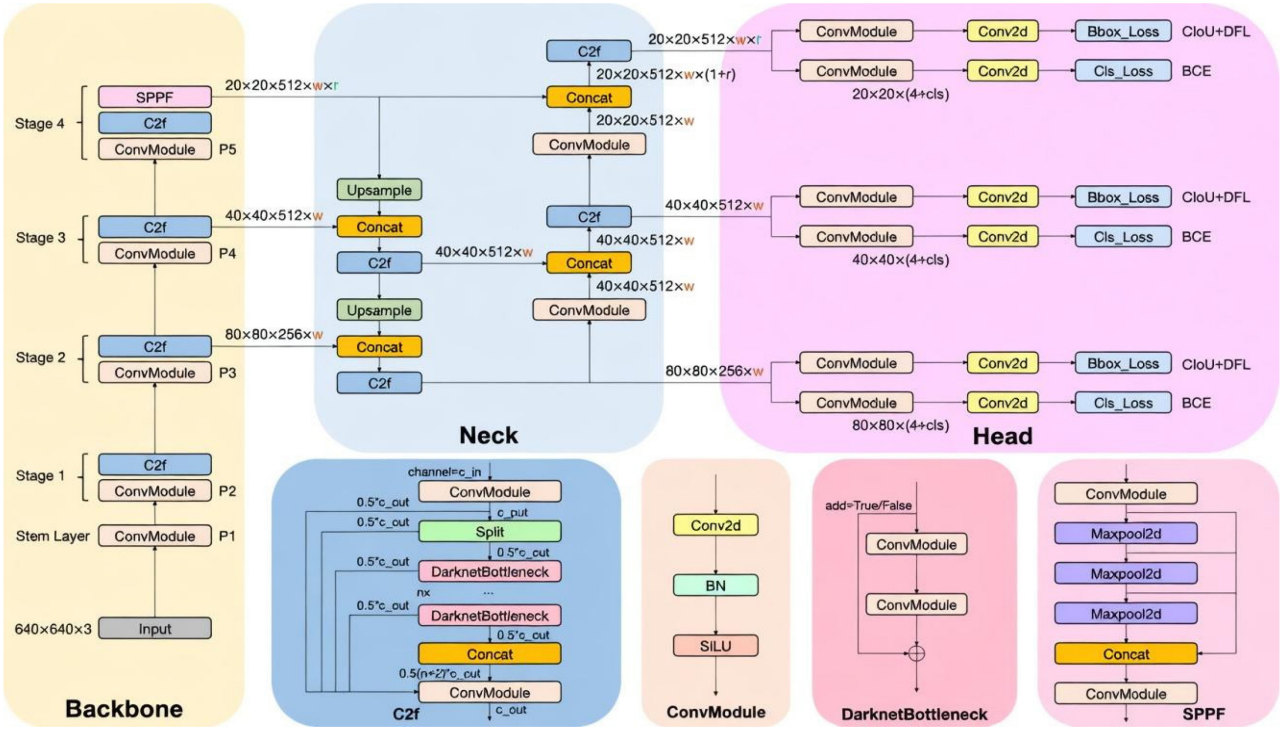


Рисунок 4.2 – Архитектура *YOLOv8*

Применение детекционной модели позволяет ускорить анализ маммографических изображений, однако само по себе не дает врачу достаточного представления о том, какие именно признаки послужили основанием для срабатывания. Альтернативой являлось обучение отдельной сегментационной сети. Такой подход потребовал бы подготовки нового корпуса данных с пиксельной разметкой, получение которой существенно более трудоемко и экспертозависимо, чем разметка прямоугольных областей. В работе был выбран иной подход, не требующий обучения отдельной сегментационной модели.

Построенный конвейер платформы использовался для выбора и обучения детекционной модели и для формирования объяснимого контура анализа на основе локализованных карт *GradientSHAP* [55, 62, 136, 205] (см. рисунок 4.3).

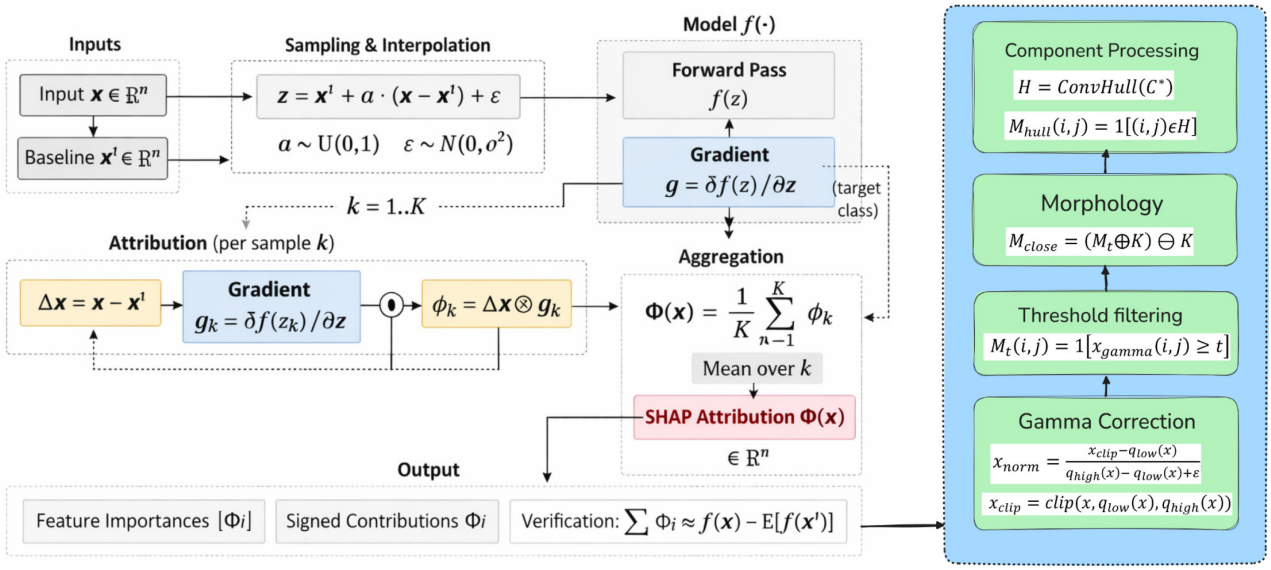


Рисунок 4.3 – Локализация *GradientSHAP*

Это позволило сформировать единый воспроизводимый конвейер, в котором к результату детекции добавляется модуль построения области интереса, отражающей предполагаемые границы патологического очага. Такое решение позволяет перейти от грубой прямоугольной локализации к более содержательному представлению подозрительной области без повторной разметки изображений на пиксельном уровне.

В маммографическом сценарии объяснитель выбирается по агентной схеме раздела 2.10. Кандидаты — GradientSHAP, Integrated Gradients и окклюзия, применённые к выходу YOLOv8 внутри найденной рамки. Их оценивают четыре агента. Агент верности смотрит, как падает уверенность детектора при удалении значимых пикселей рамки по 5 % за шаг. Агент устойчивости сравнивает карты пяти повторных запусков. Клинический агент проверяет совпадение области с контуром образования, микрокальцинатами и спикулами. Ресурсный агент учитывает время расчёта на оборудовании передвижного комплекса. В локальной функции (2.103) собственный критерий агента получает вес 0,6, остальные три: 0,2, 0,1 и 0,1. Показатели нормированы к [0, 1], поэтому нормирующий коэффициент в (2.104) равен 1. Согласование по (2.109) идёт с шагом 0,3 и равными весами влияния 1/3. Итерации прекращаются, когда изменение оценок становится меньше 0,001, но их не больше 50. Вес согласованности в (2.107) равен 0,2. Если параметр порядка (2.106) опускается ниже 0,7, объяснение детекции выводится врачу с предупреждением. Количественная оценка отдельных компонентов интерпретационного контура приведена в таблице 4.2.

Таблица 4.2 – Количественная оценка отдельных компонентов интерпретационного контура маммографического анализа

Что оценивается	**Метрика**	**Результат**
Выявление микрокальцинатов	*AUC*, 5-кратная перекрёстная проверка	**0,88**
Выявление спикуляции по совокупности геометрических и яркостных признаков	*AUC*, 5-кратная перекрёстная проверка	**0,728**
Выявление архитектурных нарушений	*AUC*, 5-кратная перекрёстная проверка	**0,71**
Оценка злокачественности по полному вектору признаков	*AUC*, 5-кратная перекрёстная проверка	**0,70**
Воспроизводимость карт *GradientSHAP* между повторными запусками	согласованность ранжирования	**≈ 0,99**
Вариативность объяснения между повторными запусками	U_{run}	**≈ 0,0001**

Полученные результаты подтверждают работоспособность отдельных компонентов интерпретационного контура. Наиболее высокое значение *AUC* получено при выявлении микрокальцинатов – 0,88. Для оценки спикуляции по совокупности геометрических и яркостных признаков *AUC* составила 0,728, для выявления архитектурных нарушений – 0,71. Повторные вычисления карт *GradientSHAP* продемонстрировали высокую воспроизводимость: согласованность ранжирования областей составила около 0,99 при межзапусковой вариативности порядка 10^{-4}. Эти показатели характеризуют отдельные компоненты интерпретационного контура и не являются оценкой клинической диагностической эффективности системы в целом.

Построение подозрительной области осуществлялось на основе вычисления *SHAP*-маски и состояло из четырёх последовательных этапов: предобработки изображения, пороговой фильтрации, морфологической обработки и выделения связной компоненты с последующим формированием сегментированной области.

На этапе предобработки исходное изображение x обрезается по квантилям и подвергается гамма-коррекции. Обрезка по квантилям задавалась выражением:

$$x_{clip} = clip(x, q_{low}(x), q_{high}(x)), \tag{4.1}$$

где $q_{low}(x)$ и $q_{high}(x)$ — нижний и верхний квантили соответственно.

Далее выполнялась нормализация:

$$x_{norm} = \frac{x_{clip} - q_{low}(x)}{max(q_{high}(x) - q_{low}(x), \varepsilon)}, \quad (4.2)$$

где $\varepsilon > 0$ – малое число, предотвращающее деление на нуль.

Для усиления контраста применялась гамма-коррекция:

$$x_\gamma = x_{norm}^\gamma, \quad (4.3)$$

где $\gamma > 0$ – параметр яркостного преобразования.

После этого выполнялась пороговая фильтрация:

$$\mathbb{M}_t(i,j) = \mathbb{I}\left[x_\gamma(i,j) \geq t\right], \quad (4.4)$$

где t – порог выделения значимых областей, $\mathbb{I}[\cdot]$ — индикаторная функция.

Для устранения разрывов и сглаживания контура к бинарной маске применялась морфологическая операция закрытия:

$$\mathbb{M}_{close} = (\mathbb{M}_t \oplus SE) \ominus SE, \quad (4.5)$$

где SE — структурирующий элемент, $\oplus$ обозначает дилатацию, а $\ominus$ — эрозию.

На следующем этапе в маске $\mathbb{M}_{close}$ выделялись связные компоненты:

$$Lab = \text{label}(\mathbb{M}_{close}), \quad (4.6)$$

где $Lab(i,j)$ – метка компоненты в пикселе (i,j). Фоновая область имела метку 0, а компоненты нумеровались с 1.

Для каждой компоненты вычислялась площадь:

$$A_k = \sum_{i,j} \mathbb{I}\,[Lab(i,j) = k]. \quad (4.7)$$

В качестве основной выбиралась крупнейшая компонента:

$$k^* = \arg\max_k A_k. \quad (4.8)$$

Правило (4.8) оставляет одну связную структуру, поэтому при нескольких детекциях на снимке его следует применять к окрестности каждой ограничивающей рамки отдельно: тогда выбор одной компоненты не отбрасывает другие очаги, найденные детектором, а их число учитывается при интерпретационной оценке. Внутри окрестности одной детекции мелкие компоненты могут соответствовать шуму атрибуции. При фрагментированном очаге, например при группе

микрокальцинатов, часть фрагментов может быть исключена. Естественным расширением правила (4.8), в текущей реализации не применявшимся, является объединение компонент, площадь которых превышает заданную долю площади A_{k^*}.

Для выбранной компоненты строилась выпуклая оболочка:

$$H = ConvHull(k^*), \tag{4.9}$$

где $ConvHull(x)$ — выпуклая оболочка, вычисляемая алгоритмом Грэхама.

После этого формировалась соответствующая бинарная маска:

$$\mathbb{M}_{hull}(i,j) = \mathbb{I}[(i,j) \in H]. \tag{4.10}$$

На заключительном шаге выполнялось расширение полученной области:

$$\mathbb{M}_{final} = \mathbb{M}_{hull} \oplus SE_r, \tag{4.11}$$

где SE_r — структурирующий элемент финальной дилатации.

Закрытие (4.5), построение выпуклой оболочки (4.9) и финальная дилатация (4.11) только расширяют область. По построению результат каждой из этих операций содержит исходное множество (для закрытия и дилатации, при структурирующих элементах, содержащих начало координат). Поэтому итоговая маска M_{final} не отсекает пиксели выбранной компоненты, а её отклонение от этой компоненты одностороннее и выражается в возможном завышении площади. Отклонение от истинной границы очага зависит также от порога t в (4.4) и от самой карты атрибуции и может иметь любой знак. Для вспомогательной визуализации такое консервативное свойство предпочтительно. Количественная оценка этого отклонения требует пиксельной экспертной разметки, которая в данном наборе не выполнялась, поскольку разметка велась ограничивающими прямоугольниками и пиксельные границы очагов внутри прямоугольников не размечались.

На основе итоговой бинарной маски $\mathbb{M}_{final}$ вычислялись морфометрические признаки образования. Площадь определялась как:

$$A = \sum_{i,j} \mathbb{M}_{final}\,(i,j). \tag{4.12}$$

Для оценки формы использовался коэффициент округлости, связывающий площадь A из (4.12) с периметром контура Per:

$$Circ = \frac{4\pi A}{Per^2}. \tag{4.13}$$

Шероховатость границы определялась сравнением реального контура с его выпуклой оболочкой:

$$R_b = \frac{Per}{Per_H} - 1, \tag{4.14}$$

где Per_H — периметр выпуклой оболочки H.

Для оценки спикуляции использовалась радиальная функция контура относительно центроида объекта $r(\theta)$. Индекс спикуляции, построенный по аналогии с [206], вычислялся по формуле:

$$SI = \frac{1}{Per}\sum_{k=1}^{N_{pk}} \frac{h_k}{w_k}, \tag{4.15}$$

где h_k — высота k-го радиального пика, а w_k — его угловая ширина, N_{pk} — число радиальных пиков. Увеличение значения SI соответствует более выраженной спикуляции и ассоциируется с большей вероятностью злокачественного характера образования.

После вычисления морфометрических признаков выполнялась эвристическая интерпретация обнаруженного образования в терминах категорий *BI-RADS*:

$$BI\text{-}RADS = \begin{cases} 1, & N_{det} = 0 \\ 2, & Circ \geq \theta_C^{(high)} \wedge R_b < \theta_R^{(low)} \wedge SI < \theta_{SI}^{(low)} \\ 3, & \theta_C^{(mid)} \leq Circ < \theta_C^{(high)} \wedge \theta_R^{(low)} \leq R_b < \theta_R^{(mid)} \wedge SI < \theta_{SI}^{(mid)} \\ 4, & Circ < \theta_C^{(mid)} \wedge \left(R_b \geq \theta_R^{(mid)} \vee SI \geq \theta_{SI}^{(mid)}\right) \\ 5, & Circ < \theta_C^{(low)} \wedge R_b \geq \theta_R^{(high)} \wedge SI \geq \theta_{SI}^{(high)} \end{cases}, \tag{4.16}$$

здесь N_{det} — число обнаруженных областей, а θ_C, θ_R и θ_{SI} — эмпирически задаваемые пороговые значения, используемые в интерпретационной схеме. Результат выражения (4.16) используется для приоритизации обнаруженных областей при последующем анализе врачом-рентгенологом и не представляет собой автоматически присвоенную клиническую категорию *BI-RADS*. Выражение (4.16) не является формализацией полного стандарта *BI-RADS* и отражает только морфометрическую составляющую оценки степени подозрительности. Разделение категории 4 на подкатегории *4A*, *4B* и *4C* в данной схеме не выполняется по требованию врачей-рентгенологов, чтобы не исключать из процесса диагностики лицо, принимающее решение. Окончательная категория *BI-RADS* определяется врачом-рентгенологом с учётом полного состава клинико-рентгенологических признаков. Численные значения параметров предобработки и морфологической обработки (q_{low}, q_{high}, γ, t, размеры SE и SE_r) и пороги θ фиксируются так же, как параметры вычисления Θ протокола (2.70), и сохраняются в конфигурации запуска платформы вместе с версией модели.

Итоговая архитектура прикладного решения, сформированная средствами платформы,

представлена на рисунке 4.4.

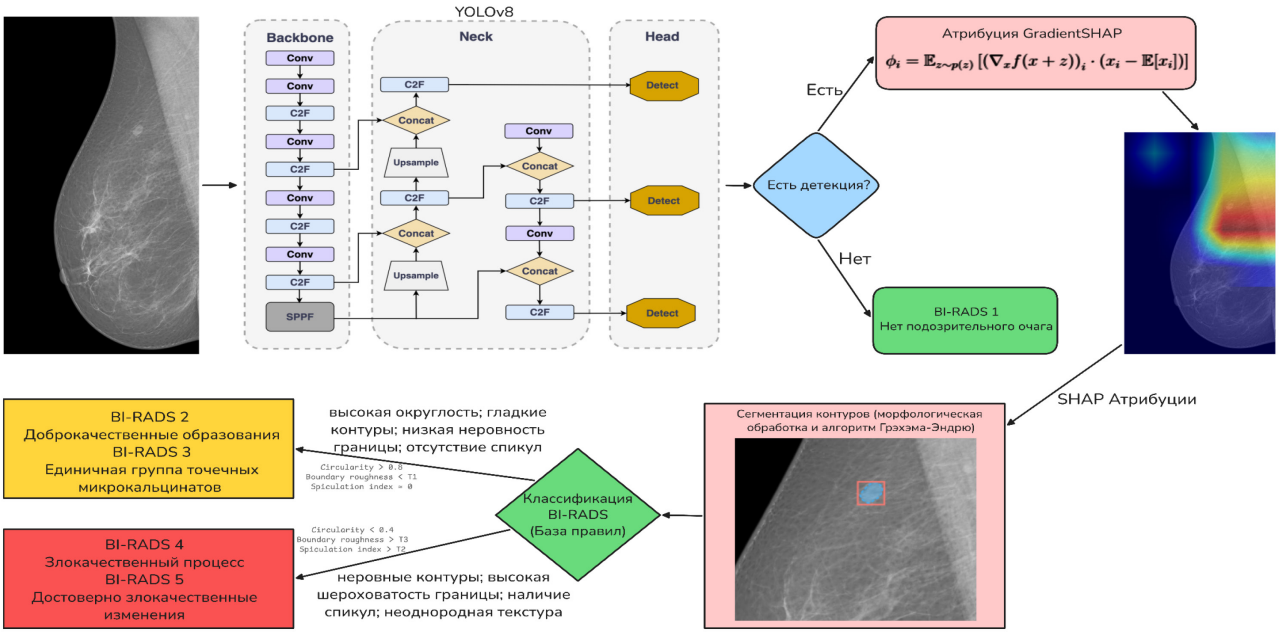


Рисунок 4.4 – Архитектура системы ИИ

Первичная реализация *XAI*-контура в платформе основывалась на построении градиентных карт значимости для изображения в целом. Практическая апробация показала, что такие глобальные карты не всегда обеспечивают достаточно точное объяснение отдельных срабатываний детектора, поскольку их пространственная структура может быть размытой и включать вклады, не относящиеся непосредственно к анализируемой области.

Предлагаемый подход основан на двух положениях. Во-первых, объяснение должно строиться для конкретной обнаруженной области, а не для усреднённой глобальной активности сети. Во-вторых, карта атрибуции рассматривается как исходный сигнал значимости, требующий последующей обработки для получения устойчивой и геометрически ограниченной области интереса.

Построение локализованных объяснений для каждой отдельной детекции связано с существенными вычислительными затратами. Если глобальная карта атрибуции вычисляется однократно для всего изображения, то в локализованной постановке объяснение необходимо формировать отдельно для каждого ограничивающего прямоугольника, полученного на этапе детекции. Для методов семейства *SHAP* такая схема особенно ресурсоёмка, поскольку требует многократного вычисления градиентов на множестве базовых и интерполяционных состояний входа. В наивной реализации это приводит к близкому к линейному росту времени обработки при увеличении числа обнаруженных объектов и делает этап построения объяснений существенно более затратным, чем сама детекция.

Для снижения вычислительной сложности в платформе была реализована схема локально ограниченной множественной атрибуции области объяснений *SHAP*, представленная на рисунке 4.5 [205]. В этой схеме прямой проход детекционной сети выполняется по полному изображению, что позволяет сохранить глобальный контекст сцены и корректные активации модели. При этом генерация *SHAP* атрибуций ограничивается локальными областями, соответствующими найденным ограничивающим прямоугольникам.



Рисунок 4.5 – Распараллеливание вычислений атрибуций

В отличие от построения атрибуций для всего изображения, в данной постановке пространство атрибуции ограничивается окрестностями сопоставленных детекций, тогда как вне этих областей входное изображение остается неизменным. Это сокращает число варьируемых признаков и уменьшает объём вычислений при сохранении связи объяснения с исходным решением детектора. В результате сохраняется интерпретационная ценность локальной атрибуции области объяснений, а их вычисление становится пригодным для практического использования. Поскольку атрибуция вычисляется в окрестности рамки детектора, ошибка локализации *YOLOv8* переносится на карту объяснения смещённая или слишком узкая рамка ограничивает область поиска значимости. Поэтому локализованная карта трактуется как объяснение конкретной детекции, а не как самостоятельная сегментация очага, и не может исправить пропуск или ошибку детектора. Выбросы градиента на локальных перепадах яркости ослабляются тем, что *GradientSHAP* усредняет градиенты по опорным изображениям и зашумлённым точкам интерполяции [55, 136]. Межзапусковая воспроизводимость карт (таблица 4.2) характеризует устойчивость к случайному выбору опорных точек.

Общая организация потока данных в системе приведена на рисунке 4.6.

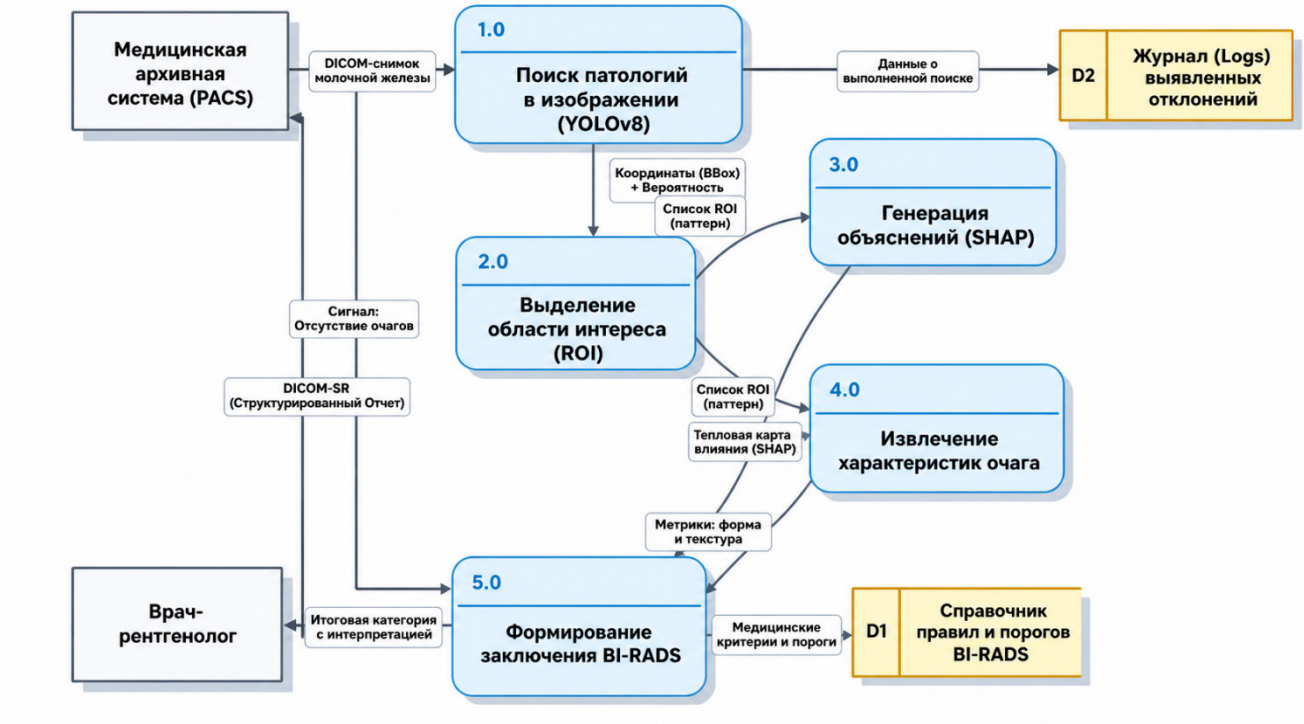


Рисунок 4.6 – Диаграмма потока данных

Практическая реализация предложенного подхода выполнена в виде программной системы поддержки принятия решений при анализе маммографических изображений. Архитектура, сформированная средствами платформы, была развернута в веб-интерфейсе. Система включает модуль загрузки и просмотра исследований, модуль автоматической детекции патологически значимых областей, модуль локализованного вычисления *GradientSHAP*-атрибуций и модуль визуального представления результатов анализа. Интерфейс системы приведен на рисунке 4.7.

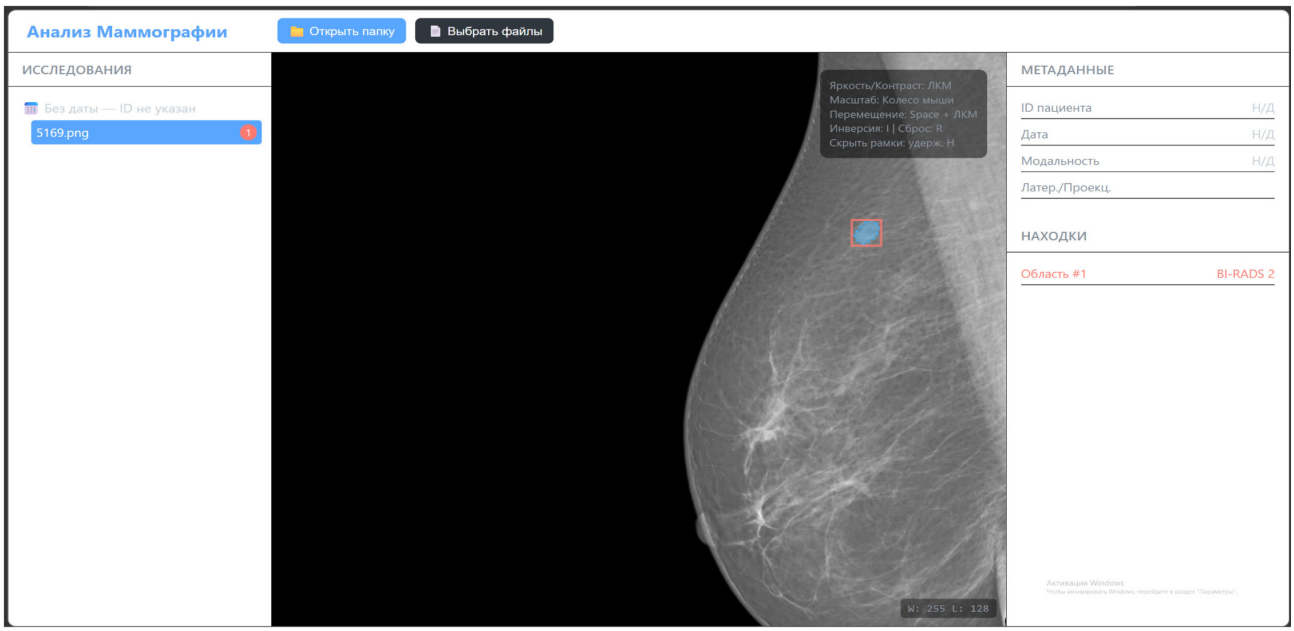


Рисунок 4.7 – Интерфейс веб-платформы

Реализованная система ориентирована на конечный сценарий эксплуатации и обеспечивает представление результатов в едином интерфейсе, где исходное изображение, выделенные области интереса, сопутствующие метаданные и результаты, вспомогательной интерпретации рассматриваются как элементы единого диагностического контура. Тем самым предложенный метод реализован в форме прикладного программного решения, пригодного для использования в составе медицинской информационной инфраструктуры.

4.1.7. Основные выводы

В рамках маммологического кейса была проведена апробация платформы при построении прикладного решения для анализа маммографических изображений. В ходе работы был сформирован и верифицирован набор аннотированных маммографических исследований, пригодный для обучения и оценки моделей детекции патологических областей.

С использованием средств платформы, включая механизмы *AutoML*, была подобрана и обучена модель автоматической детекции патологических областей на маммографических изображениях. Полученные результаты показали, что детекция в формате ограничивающих прямоугольников представляет практический интерес для анализа маммограмм, однако не в полной мере отвечает требованиям клинической интерпретации, поскольку не отражает предполагаемые границы и форму патологического очага.

В рамках апробации платформы реализован объяснимый контур на локализованных картах *GradientSHAP* без обучения отдельной сегментационной модели и повторной пиксельной разметки, что позволило перейти от прямоугольной локализации к интерпретируемому представлению подозрительного участка и дополнить детекционный контур локализованным объяснением конкретного срабатывания вместо неоднозначных глобальных карт значимости.

Результаты по созданию маммологической системы показали, что платформа может использоваться как инструментальная основа для построения прикладных решений в задачах маммографической диагностики, объединяющих автоматическую детекцию патологических областей и интерпретируемое сопровождение результата.

4.2. Биомедицинская задача: ишемическая болезнь сердца

4.2.1. Введение

Коронарные стенозы являются патологическим сужением просвета коронарных артерий, возникающим вследствие формирования атеросклеротических поражений сосудистой стенки. Такие изменения приводят к уменьшению кровотока в миокарде и являются ключевым патофизиологическим механизмом ишемической болезни сердца (ИБС) [207, 208]. В современной клинической практике инвазивная коронарная ангиография рассматривается как основной метод выявления и характеристики коронарных стенозов [209, 210]. Высокая пространственная детализация ангиографических изображений позволяет визуализировать коронарное русло и выявлять локализацию, протяжённость и выраженность сосудистых поражений.

Несмотря на высокую диагностическую ценность ангиографии, интерпретация коронарных стенозов не является полностью однозначной. Ограничения связаны со сложной геометрией сосудистого русла, вариативностью анатомического строения коронарных артерий, неоднородностью контрастирования, наличием шумов и артефактов изображения, а также высокой зависимостью визуальной оценки от опыта специалиста [211–214]. Дополнительной проблемой является значительный объём данных. Исследования могут включать тысячи кадров, что увеличивает нагрузку на специалиста.

Снизить нагрузку на специалиста возможно с помощью систем автоматического анализа исследования. Но разработка систем ограничена в данной области из-за дефицита наборов данных с качественной разметкой коронарных поражений. В опубликованных работах по набору данных *CADICA* это обозначается как одна из ключевых причин, затрудняющих разработку и сравнение алгоритмов автоматического анализа [215].

В последние годы методы глубокого обучения нейросетей стали одним из основных инструментов анализа изображений. Они успешно применяются для задач классификации, сегментации и детекции объектов [216]. Благодаря способности извлекать косвенные признаки из изображений, глубокие нейронные сети позволяют строить сложные системы анализа данных. Но при применении нейросетей в данной задаче возникают затруднения, связанные с объяснением работы алгоритма.

В этих условиях актуальной становится задача построения интеллектуальных систем анализа ангиографических изображений, способных не только автоматически выделять сосудистые структуры и оценивать степень стеноза, но и сопровождать результат интерпретируемым объяснением. В этой связи особый интерес представляет переход к архитектурам ОИИ 2.0, в которых

нейросетевые методы объединяются с механизмами представления экспертного знания и интерпретации результата [19, 217, 218].

В настоящем разделе задача диагностики коронарных стенозов рассматривается как апробационный сценарий платформы. Платформа в данном случае выступает не как система подготовки данных, а как инструментальная среда, в которой в ручном режиме строятся и настраиваются глубокие модели, объяснительный контур и гибридный нейро-нечёткий конвейер интерпретации. Подготовка и верификация данных выполняются отдельно, вне платформы. Результатом же работы платформы являются обученные модели, подобранные объяснения, согласованный конвейер анализа и рекомендации по построению иерархической системы. Затем именно на основе этих артефактов создаётся прикладное программное обеспечение для конкретной медицинской задачи. Такая логика непосредственно соотносится с защищаемыми положениями диссертации о многоуровневости, иерархичности и объяснимости платформенного контура.

4.2.2. Существующие решения

В настоящее время разработан ряд программных комплексов, позволяющих выполнять обработку ангиографических изображений, проводить количественный анализ сосудов и визуализировать патологические изменения.

Среди наиболее распространённых решений можно выделить следующие системы:

- ***HeartFlow – программная*** платформа для анализа коронарных артерий на основе данных компьютерной томографии. Система позволяет моделировать кровоток и рассчитывать функциональные показатели стеноза [219].
- ***Siemens syngo.via – медицинская*** система визуализации, применяемая для анализа кардиологических изображений, включая ангиографические исследования [220].
- ***Philips IntelliSpace Cardiovascular – программный*** комплекс для обработки кардиологических данных и анализа сосудистых структур [221].
- ***GE AW Cardiac – система*** обработки медицинских изображений, используемая для анализа сердечно-сосудистой системы и диагностики патологий коронарных артерий [222].

Большинство указанных решений являются проприетарными и ориентированы преимущественно на визуализацию и количественный анализ изображений, не обеспечивая достаточный уровень объяснимости решения. Прецедентный подход здесь лишь дополняет предлагаемый: база продукционных правил (таблица 4.3) кодирует обобщённые клинические критерии, а не отдельные случаи, поэтому не требует репозитория размеченных прецедентов и допускает проверку каждого правила по отдельности.

4.2.3. Постановка задачи

Целью работы является апробация платформы на задаче анализа ангиографических изображений при диагностике коронарных стенозов и построение на её основе прикладного программного обеспечения для поддержки врача.

Для достижения поставленной цели необходимо решить следующие задачи:

- Выполнить отбор источника данных, подготовку и верификацию исходных медицинских данных вне платформы, экспертную ревизию и формирование рабочей выборки.
- Построить на платформе в ручном режиме модели ИИ, обеспечивающие автоматический отбор диагностически значимых кадров, сегментацию сосудистых структур и формирование объяснительного контура.
- Разработать и встроить в платформенный контур гибридный модуль интерпретации, объединяющий нейросетевые результаты с нейро-нечёткой логикой и базой продукционных правил.
- Создать прикладное программное обеспечение, ориентированное на использование в контуре поддержки принятия решений.

4.2.4. Практическая реализация

На первом этапе был проведён анализ открытых наборов данных; в качестве исходного источника был выбран открытый набор медицинских изображений CADICA v2 [215]. Набор данных включает изображения 42 пациентов и более 6000 кадров с разрешением 512×512 пикселей. Данный набор данных был выбран как один из немногих открытых ресурсов, ориентированных именно на задачу анализа коронарных поражений по ангиографическим изображениям.

На втором этапе были собраны требования врачей-кардиологов к системе. В частности, системе необходимо было поддерживать процентную оценку степени сужения, категориальное описание выраженности стеноза, морфологическую интерпретацию поражения, а также формирование рекомендаций, связанных с риском гемодинамически значимого поражения, оценкой фракционного резерва кровотока и шкалами коронарного кровотока. Эти требования определили структуру последующего гибридного конвейера.

На третьем этапе выполнялось дополнительное аннотирование набора данных *CADICA* v2. Был проведён автоматизированный отбор изображений, в ходе которого были удалены кадры с выраженными артефактами и недостаточным качеством визуализации сосудистых структур.

После первичной фильтрации осталось 3271 изображение. Далее была проведена врачебная ревизия данных, в результате которой сформирован рабочий набор 1412 диагностически значимых изображений, использованных для обучения и тестирования моделей сегментации. Для повышения качества анализа изображений был реализован многоэтапный конвейер предобработки, включающий подавление спекл-шума, локальное повышение контрастности методом *CLAHE*, морфологические операции и выделение границ с использованием оператора Собеля. Он использовался для получения предварительных контуров сосудов, применявшихся на этапе аннотирования. Примеры коронарных сосудов, содержащихся в датасете, приведены на рисунке 4.8, а результаты работы оператора Собеля представлены на рисунке 4.9.

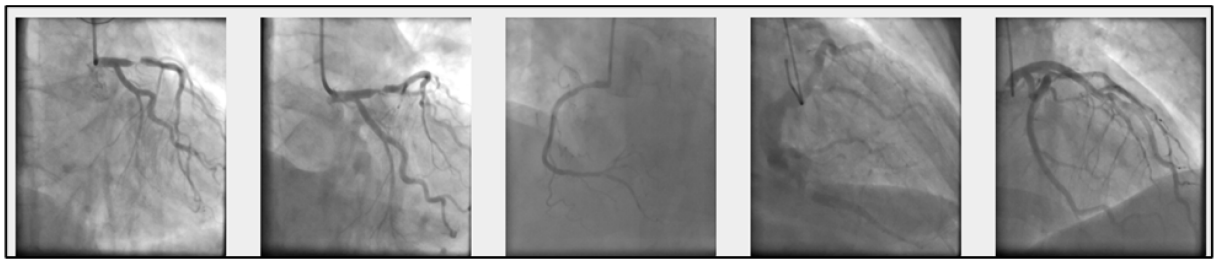


Рисунок 4.8 – Пример ангиографического изображения коронарных сосудов

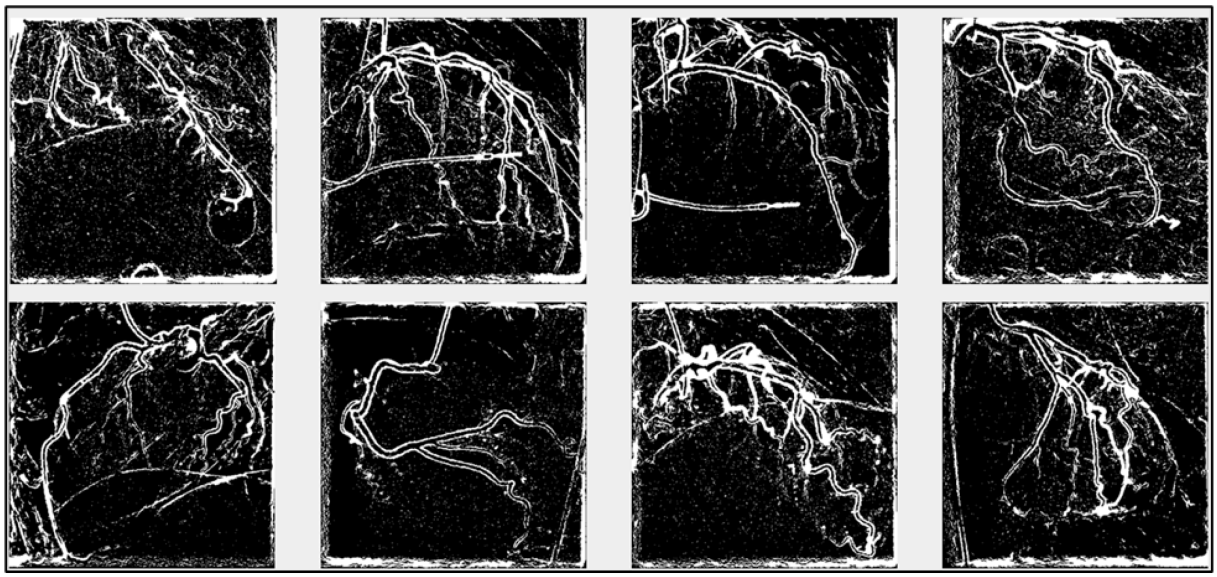


Рисунок 4.9 – Пример маски Собела

С целью ускорения подготовки обучающих данных был разработан инструмент полуавтоматической аннотации *ConAnnotator*. В качестве основы использовались предварительно построенные маски Собела, по которым формировались контуры сосудов. Далее эксперт вручную уточнял границы, корректируя форму сосудистых сегментов и стенозов. Использование такого под-

хода позволило заметно сократить трудоёмкость создания обучающих масок и повысить согласованность аннотаций. Этот этап также относится к подготовительному контуру и выполнялся до платформенной апробации. Общая схема подхода продемонстрирована на рисунке 4.10.

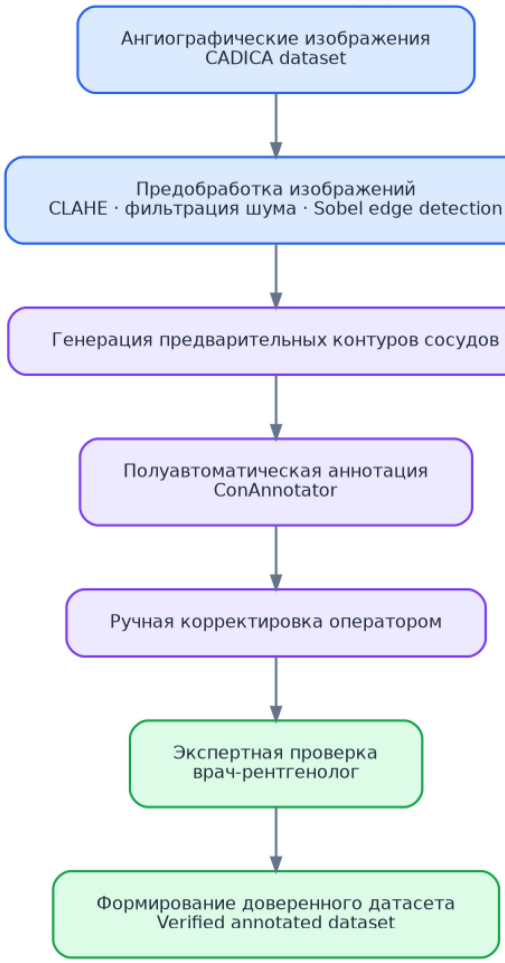


Рисунок 4.10 – Схема процесса аннотации изображений в системе *ConAnnotator*

На четвертом этапе происходила подготовка моделей на платформе в ручном режиме, из-за повышенных требований к объяснимости и необходимости построения специфического гибридного конвейера. В данном сценарии платформа использовалась не как автоматический конструктор готового решения, а как инструментальная среда для ручной сборки и настройки иерархического контура анализа, включающего модуль отбора кадров, сегментационный модуль, *XAI*-контур и нечётко-логическую интерпретацию.

Для изображений был создан модуль классификации качества кадров на основе архитектуры ResNet-34 [223] (пример архитектуры представлен на рисунке 4.11). Модель обучалась различать диагностически значимые изображения и кадры с недостаточной информативностью. В ходе экспериментов на платформе были достигнуты следующие показатели качества модели: *Accuracy* 0,965; *Recall* 0,972; *AUC* 0,987. Использование данного фильтра позволило автоматически отбирать изображения, пригодные для дальнейшего анализа и сегментации, и тем самым сократить влияние шумовых и малоинформативных кадров на последующие этапы конвейера.

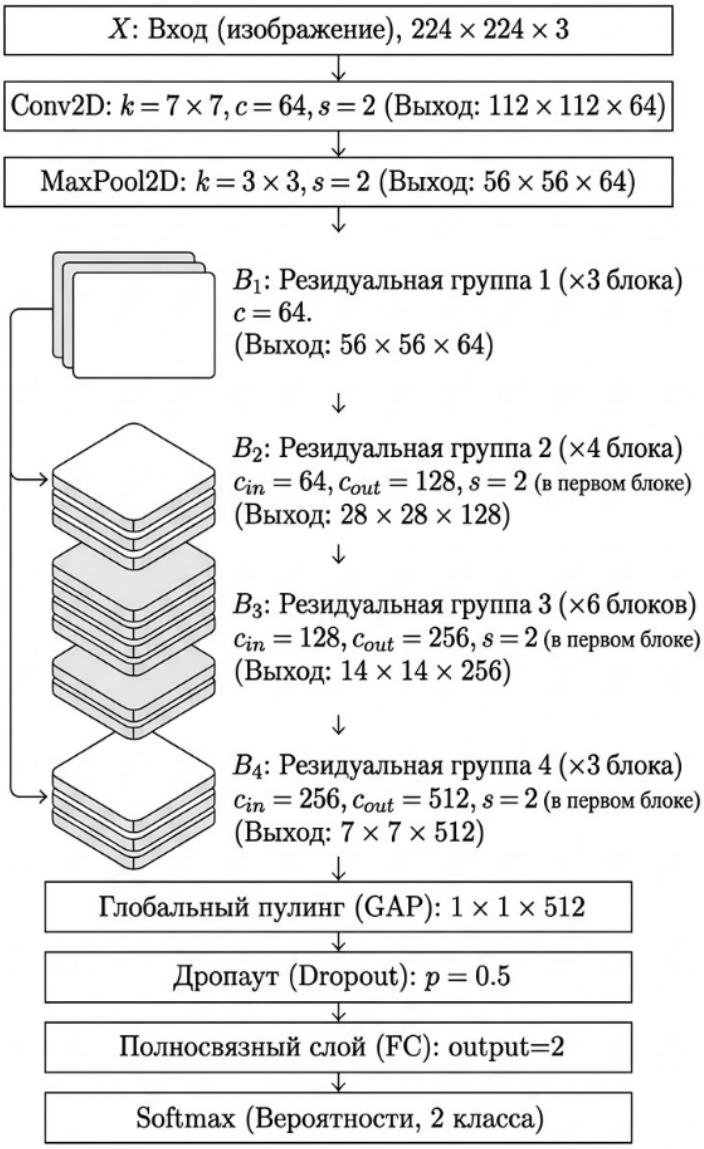


Рисунок 4.11 – Архитектура *ResNet-34*

Следующим этапом обработки являлась сегментация сосудистых структур. В ходе исследования были протестированы архитектуры U-Net [224], SegNet и DeepLabV3 [225]. По результатам экспериментальной оценки наилучшие показатели продемонстрировала архитектура *U-Net* с постобработкой на основе *Conditional Random Fields (CRF) [226]*. Архитектура U-Net представлена на рисунке 4.12. Использование *CRF*-постобработки позволило учитывать пространственные зависимости между пикселями и уточнять границы сегментируемых сосудов. Результат сегментации приведен на рисунке 4.13. Полученные метрики сегментации составили *Dice* 0,84 и *IoU* 0,78. В логике платформенной апробации это означает, что платформа позволила в ручном режиме сформировать устойчивый сегментационный модуль, пригодный для последующего включения в гибридный объяснимый контур.

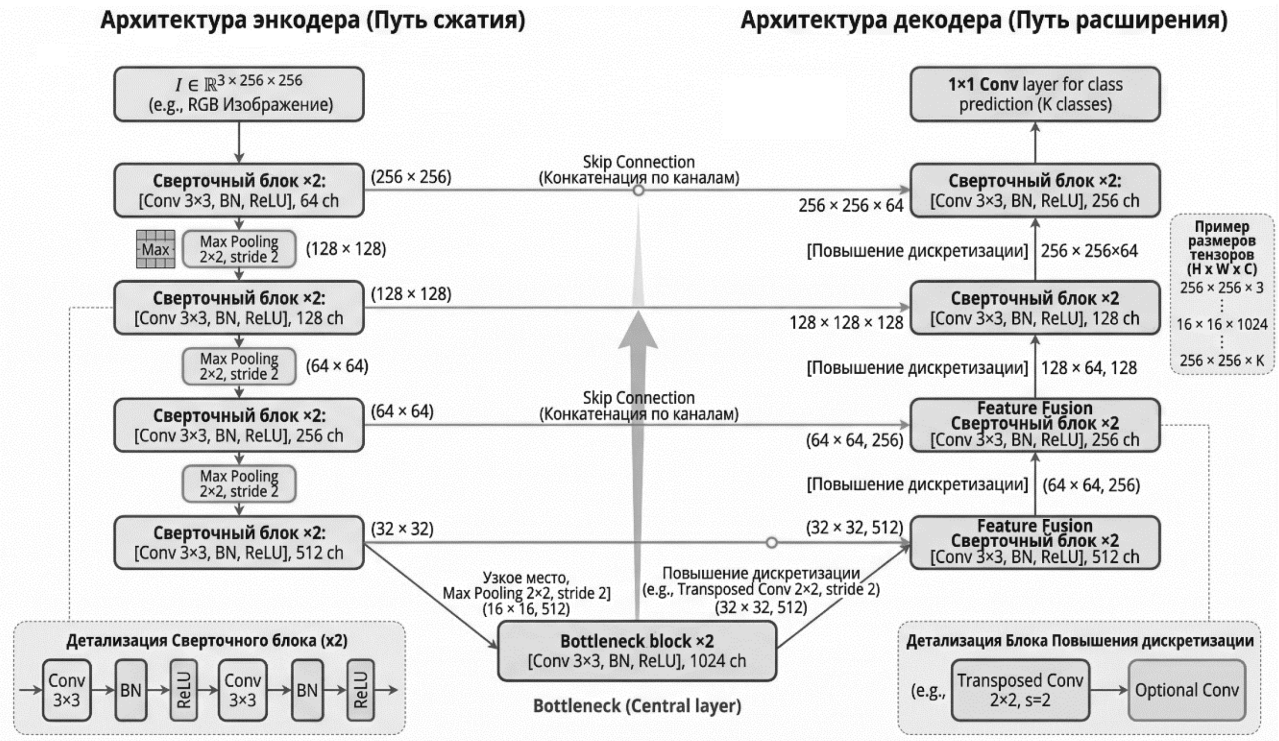


Рисунок 4.12 – Архитектура *U-Net*

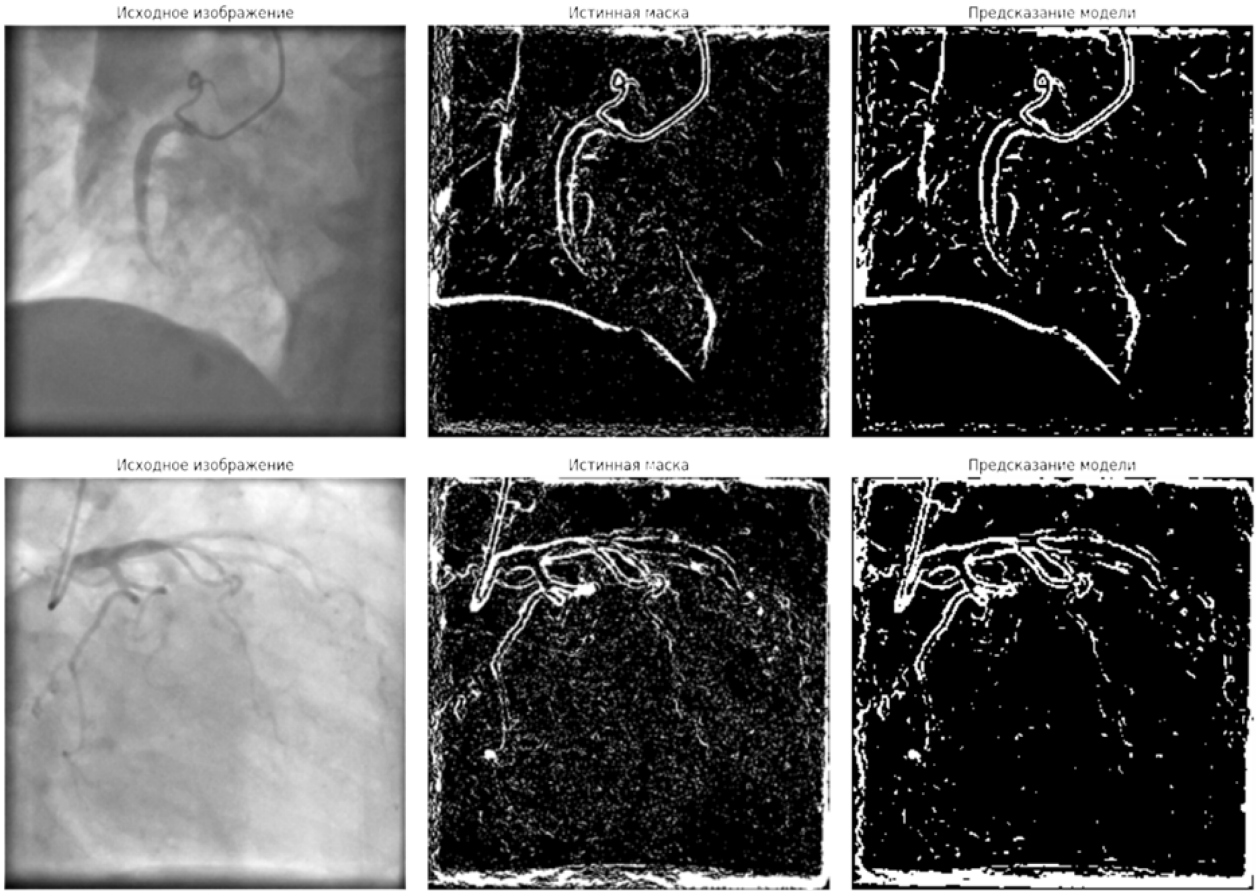


Рисунок 4.13 – Результат сегментации коронарного сосуда с использованием *U-Net*

Для обеспечения интерпретируемости решений модели применялись методы ОИИ. В частности, использовался метод Score-CAM [172], позволяющий визуализировать области изображения, оказавшие наибольшее влияние на результат сегментации. Формируемая карта внимания показывает вклад различных участков изображения в результат работы модели и даёт специалисту возможность визуально оценить корректность локализации предполагаемого стеноза (см. рисунок 4.14). В терминах главы 2 данный компонент выступает как частный механизм формирования атрибутивных объяснительных объектов модельного уровня.

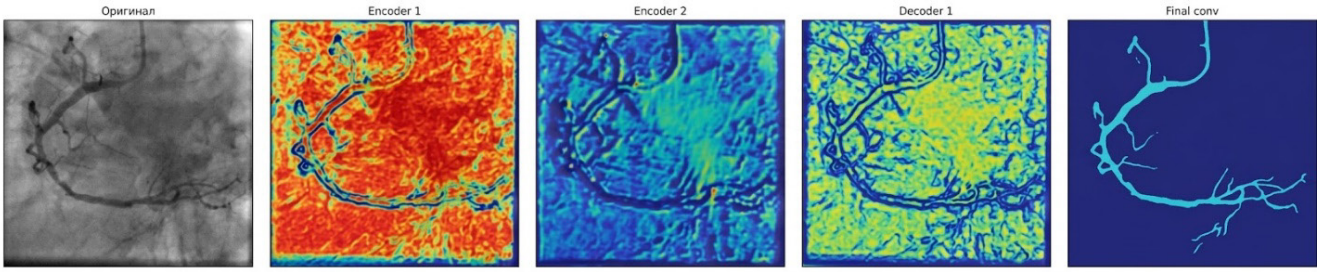


Рисунок 4.14 – Визуализация области внимания модели методом *Score-CAM* при выявлении коронарного стеноза

Для оценки согласованности полученных результатов анализа использовался интегральный показатель значимости, рассчитываемый на основе нескольких параметров, включая метрики сегментации, показатели карт внимания и геометрические характеристики выделенных сосудистых структур. Метрики сегментации *Dice* и *IoU* требуют экспертной маски, поэтому они вычислялись только на выборке разработки, где маски получены с помощью *ConAnnotator*, и служили для настройки порогов показателя. Полученное значение используется для определения степени достоверности результата и выбора дальнейшего этапа обработки (см. рисунок 4.15). Тем самым в практической реализации была апробирована идея согласования и совместного использования разнородных объяснительных и вычислительных компонент, сформулированная во второй главе.

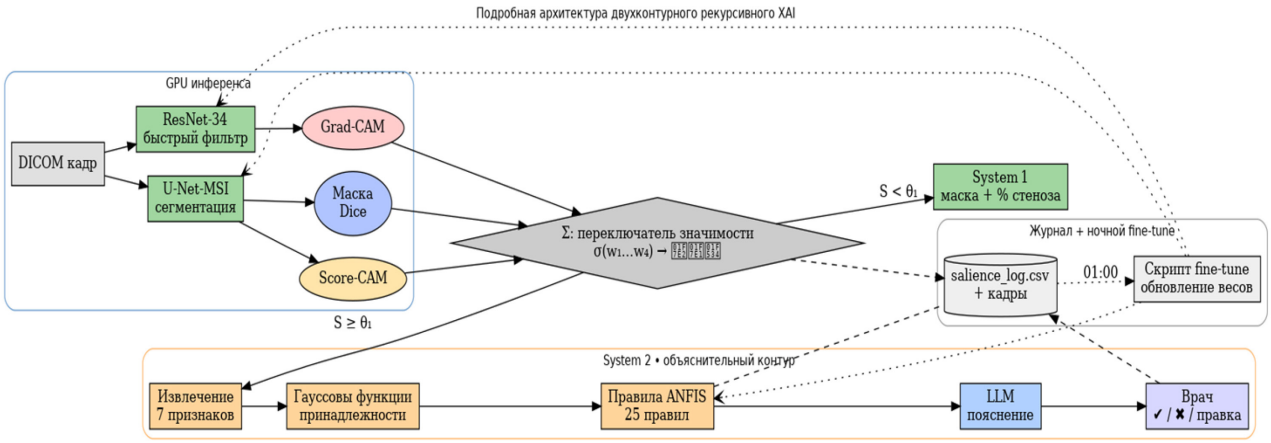


Рисунок 4.15 – Архитектура интеллектуальной системы анализа

Следующим этапом обработки данных является интерпретация результатов нейросетевого анализа. Для этой цели был разработан модуль нечётко-логического вывода, использующий продукционные правила вида «ЕСЛИ – ТО». В дополнительной программной реализации входы правил разделены по происхождению. Степень стеноза принимается из внешней количественной ангиографической оценки или экспертной разметки. Фракционный резерв кровотока (*FFR*) и степень коронарного кровотока *TIMI* передаются как внешние клинические параметры при наличии соответствующих записей. Сведения о длительности поражения также берутся из клинической записи. Эти показатели не вычисляются сетью *U-Net* и не выводятся из карт внимания. Для каждого входа сохраняются значение, единица измерения, идентификатор случая и ссылка на источник. Если хотя бы один обязательный вход отсутствует или не проходит проверку соответствия источнику, правило помечается как неприменимое и в выводе не участвует.

Модуль нечётко-логического вывода опирается на базу знаний из 25 продукционных правил, связывающих признаки сосудистого поражения с лингвистическими описаниями [72, 227, 228]. Опорные условия и заключения правил сформулированы совместно со специалистом-кардиологом; пример структуры правил приведён в таблице 4.3. В дополнительной программной реализации условия по непрерывным входам заданы кусочно-линейными функциями принадлежности, а условия по дискретным признакам — категориальными. Анатомические, функциональные и контекстные заключения хранятся раздельно. Степень поддержки заключения показывает, насколько входы соответствуют условиям базы, и не является вероятностью диагноза. В терминах главы 2 база правил формирует интерпретируемый уровень объяснительного контура.

Для пары взаимодополняющих условий по *FFR* с опорной границей 0,80 использованы линейные функции принадлежности на интервале [0,78; 0,82]:

$$\mu_+(x) = \min\left(1, \max\left(0, \frac{x-0{,}78}{0{,}04}\right)\right), \ \mu_-(x) = 1 - \mu_+(x), \qquad (4.17)$$

где x — значение FFR. Для опорного условия «степень стеноза менее 50 %» переход задан на

интервале [45 %; 55 %], для условия «50–69 %» использована трапециевидная функция с параметрами (45; 50; 69; 74) %. Полный набор термов и параметров хранится в конфигурации программы. При наличии всех обязательных входов степень активации правила r равна

$$\alpha_r = \min_i \mu_{ri}(x_i), \tag{4.18}$$

где μ_{ri} — функция принадлежности i-го условия правила r, x_i — значение соответствующего входа. Поддержка заключения равна максимуму α_r по правилам с этим заключением в соответствующей группе. Параметры функций принадлежности заданы явно для воспроизводимой инженерной проверки.

Таблица 4.3 – Пример структуры продукционных правил

Условие	**Вывод**
Степень стеноза < 50%	Слабый стеноз
Степень стеноза < 50% И фракционный резерв кровотока > 0,80	Слабый стеноз
Степень стеноза < 30% И степень коронарного кровотока TIMI = 3	Слабый стеноз
Степень стеноза < 50% И тромб отсутствует	Слабый стеноз
Степень стеноза < 50% И поражение локальное	Слабый стеноз
50% ≤ степень стеноза ≤ 69%	Умеренный стеноз
50–69% И фракционный резерв кровотока > 0,80	Умеренный гемодинамически незначимый
50–69% И фракционный резерв кровотока ≤ 0,80	Умеренный гемодинамически значимый
50–69% И степень коронарного кровотока TIMI = 2–3	Умеренный стеноз
50–69% И ангиографическая картина неоднозначна	Требуется дополнительная функциональная оценка
70% ≤ степень стеноза ≤ 90%	Выраженный стеноз
≥ 70% И фракционный резерв кровотока ≤ 0,80	Гемодинамически значимый стеноз
≥ 70% И степень коронарного кровотока TIMI = 2	Выраженный стеноз
≥ 70% И поражение нескольких сосудов	Выраженный стеноз
≥ 70% И проксимальная локализация	Высокий риск

Условие	Вывод
90% ≤ степень стеноза ≤ 99%	Критический стеноз
≥ 90% И степень коронарного кровотока TIMI ≤ 2	Критический стеноз
≥ 90% И фракционный резерв кровотока ≤ 0,75	Критический стеноз
≥ 90% И наличие тромба	Критический стеноз
Степень стеноза 95–99%	Субтотальный (критический) стеноз
Степень стеноза = 100%	Окклюзия
Степень коронарного кровотока TIMI = 0	Окклюзия
Отсутствие антеградного кровотока	Окклюзия
Наличие коллатерального кровотока при отсутствии основного	Хроническая окклюзия
Степень стеноза = 100% И длительность > 3 месяцев	Хроническая тотальная окклюзия

Исполнимость базы правил проверена на синтетических случаях: для каждого из 25 правил сформирован пример, активирующий его опорное условие, и выполнено 42 проверки с удалением обязательного входа. Во всех проверках неполное правило исключалось из вывода. При степени стеноза 60 % и *FFR* = 0,79 поддержка правил об умеренном гемодинамически незначимом и значимом стенозе составила соответственно 0,25 и 0,75, что согласуется с (4.17) и (4.18). При отсутствии *FFR* оба правила становились неприменимыми, а анатомическое описание умеренного стеноза сохранялось. Независимый пересчёт подтвердил сохранённые степени активации и соответствие входов источникам.

Итоговым этапом платформенного контура являлось формирование вспомогательного интерпретируемого диагностического заключения. Заключение строилось на основе результатов сегментации сосудов, карт внимания и вывода нечётко-логического контура. Результат платформы включал обученные модели, а также согласованный гибридный конвейер объяснений, пригодный для переноса в прикладное программное средство.

После завершения платформенной апробации были получены ключевые артефакты: обученный модуль отбора кадров, сегментационная модель, объяснительный контур, гибридный нейро-нечёткий конвейер и согласованное интерпретируемое представление результата. На основе этих компонентов был разработан прототип веб-приложения, реализующего описанный функционал: пользователь загружает изображения, после чего приложение выполняет автомати-

ческую фильтрацию кадров, сегментацию сосудов, оценку степени стеноза и формирование отчёта, показанного на рисунке 4.16. В этой логике прикладное ПО рассматривается не как сама платформа, а как надстройка над моделями и конвейером, полученными в результате её апробации.

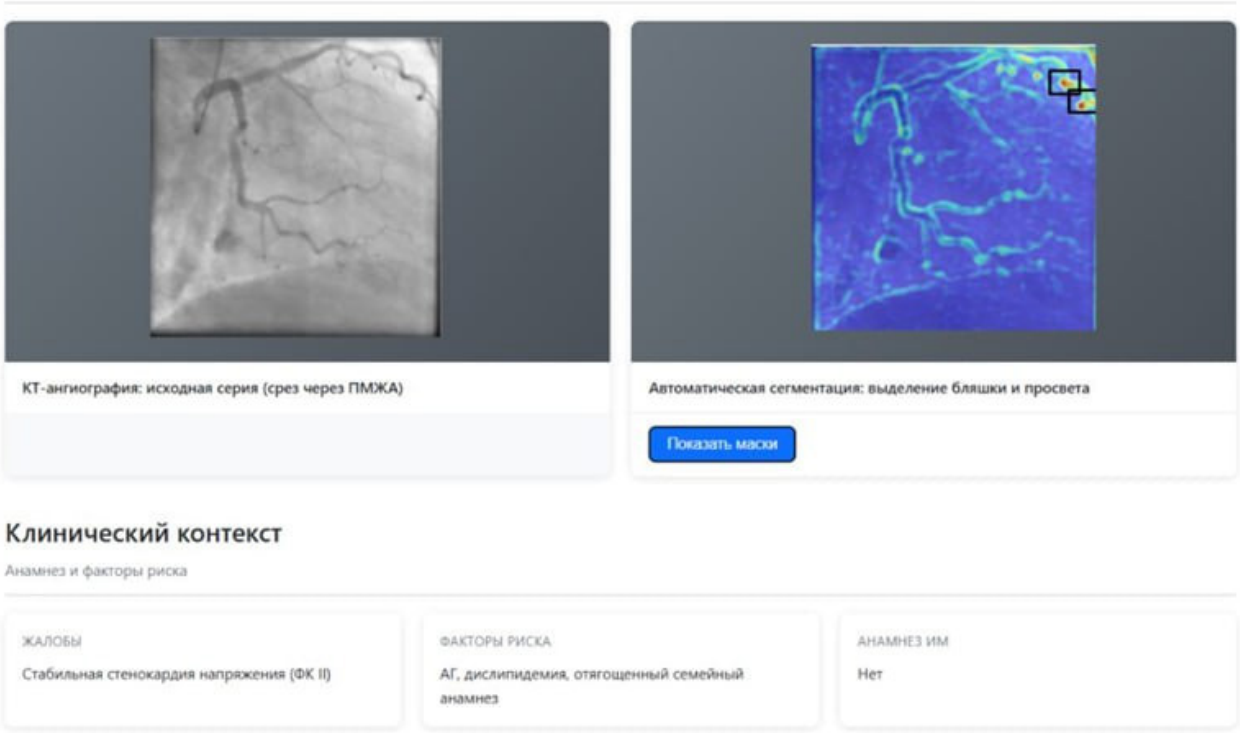


Рисунок 4.16 – Пример интерфейса программы

Дополнительно в коронарном эксперименте проверялась чувствительность объяснений к параметрам модели и обучающей разметке. Проверка построена по схеме sanity-проверок карт значимости [70] и выполнена на выборке разработки из 62 пациентов. В первом тесте параметры модели последовательно заменялись случайными значениями, начиная с выходного классифицирующего слоя и далее к начальным слоям. Во втором тесте модель заново обучалась на данных со случайно перемешанными метками классов. Карты объяснений каждого варианта сравнивались с картами исходной обученной модели по абсолютному значению коэффициента ранговой корреляции Спирмена. Доверительные интервалы рассчитывались бутстрепом по пациентам (5000 повторных выборок с возвращением). Проверка считалась пройденной, если верхняя граница 95%-го доверительного интервала средней абсолютной корреляции была меньше 0,8, а доля пригодных для сравнения пар составляла не менее 90 %. Результаты приведены в таблице 4.4.

Таблица 4.4 – Проверка чувствительности объяснений к параметрам модели и обучающей разметке

Проверка	**Средняя абсолютная корреляция карт ↓**	**95%-й доверительный интервал**	**Доля пригодных пар**	**Условия выполнены**
Замена параметров классифицирующего слоя случайными значениями	0,287	[0,279; 0,294]	98,88 %	Да
Дополнительная замена параметров четвёртого блока	0,245	[0,238; 0,252]	99,53 %	Да
Дополнительная замена параметров третьего блока	0,213	[0,207; 0,219]	99,60 %	Да
Дополнительная замена параметров второго блока	0,227	[0,221; 0,234]	99,55 %	Да
Дополнительная замена параметров первого блока	0,252	[0,243; 0,260]	99,57 %	Да
Дополнительная замена параметров начальных слоёв	0,256	[0,247; 0,265]	98,20 %	Да
Повторное обучение на перемешанных метках классов	**0,244**	**[0,238; 0,251]**	**95,33 %**	**Да**

Все варианты удовлетворили условиям проверки. Замена параметров модели случайными значениями заметно меняла карты по сравнению с картами исходной обученной модели; сходный результат дало повторное обучение на данных с перемешанными метками. Следовательно, объяснения зависят и от параметров модели, и от разметки, использованной при обучении. Проверка выполнена на выборке разработки и не устанавливает клинической правильности локализации стеноза.

4.2.5. Основные выводы по результатам

В разделе 4.2 выполнена апробация платформы на задаче анализа ангиографических изображений при диагностике коронарных стенозов. Полученные результаты показали, что для задач

данного класса принципиальное значение имеют методы сегментации и классификации, перевода результата автоматического анализа в объяснимую форму, согласованную с клинической логикой предметной области.

Вне платформы был сформирован рабочий набор данных на основе *CADICA v2*, включающий изображения, прошедшие автоматическую фильтрацию, экспертную ревизию и дополнительное аннотирование. Реализован конвейер предобработки, включающий подавление спекл-шума, повышение контрастности методом *CLAHE* и выделение границ сосудов с использованием оператора Собеля. Разработанный инструмент *ConAnnotator* позволил ускорить подготовку обучающих масок за счёт использования предварительных контуров сосудов и последующей экспертной корректировки. Тем самым сформирован внешний подготовительный контур данных, который выступает входом для последующей платформенной апробации.

На платформе в ручном режиме были построены и настроены основные компоненты интеллектуального контура анализа. Разработан модуль автоматической фильтрации изображений на основе архитектуры *ResNet*-34 для отбора диагностически информативных кадров. Полученные показатели качества составили *Accuracy* 0,965, *Recall* 0,972, *AUC* 0,987. Реализована модель сегментации сосудистых структур на основе архитектуры *U-Net* с *CRF*-постобработкой, позволившая достичь показателей *Dice* 0,84 и *IoU* 0,78. В систему интегрирован объяснительный контур на основе *Score-CAM*, обеспечивающий визуализацию областей изображения, оказавших наибольшее влияние на решение модели.

Существенным результатом является построение специфического гибридного конвейера, объединяющего нейросетевые методы обработки изображений, механизмы ОИИ и нечётко-логическую интерпретацию результатов. Сформирована база знаний из 25 продукционных правил, связывающих геометрические и визуальные характеристики сосудистого поражения со степенью выраженности стеноза. Тем самым в прикладной реализации платформы подтверждена работоспособность защищаемых положений о многоуровневом, иерархически организованном и объяснимом контуре анализа медицинских изображений.

В результате работы платформы получены обученные модели, подобранные объяснения и согласованный гибридный конвейер интерпретации. На основе этих артефактов было разработано прикладное программное обеспечение в виде прототипа веб-приложения, реализующего автоматическую фильтрацию кадров, сегментацию сосудов, оценку степени стеноза и формирование интерпретируемого диагностического заключения.

4.3. Биомедицинская задача: автоматизированный анализ рентгенограмм грудной клетки при выявлении признаков пневмонии

4.3.1. Введение

Автоматизированный анализ рентгенограмм грудной клетки является одной из наиболее востребованных областей применения ИИ в медицинской визуализации [229]. В настоящей работе в качестве одного из сценариев апробации платформы рассматривается задача выявления признаков пневмонии по рентгеновским изображениям органов грудной клетки. Выбор данного сценария обусловлен клинической значимостью заболевания, широким использованием рентгенографии в рутинной практике, доступностью открытых данных и возможностью воспроизводимой проверки методов классификации, интерпретации и пользовательского представления результатов [230, 231].

Рентгенография органов грудной клетки остаётся одним из основных методов лучевой диагностики при подозрении на пневмонию благодаря своей доступности, оперативности и диагностической информативности [230, 231]. Вместе с тем интерпретация рентгенограмм представляет собой сложную экспертную задачу: диагностическое заключение зависит от выраженности патологических изменений, качества снимка, сопутствующей патологии и клинического опыта специалиста. В литературе показана межнаблюдательная вариабельность при интерпретации рентгенограмм грудной клетки, что подтверждает ограниченную однозначность визуальной оценки и обосновывает применение интеллектуальных средств поддержки принятия врачебных решений [232, 233].

В логике диссертационного исследования данный раздел выполняет функцию прикладной апробации теоретико-методического аппарата, разработанного во второй главе. Платформа рассматривается как готовое прикладное программное средство и инструментальная среда, в которой на основе заранее подготовленных данных обучаются модели, подбираются объяснительные механизмы и формируется многоуровневый контур интерпретации. Общая схема апробации платформы в рассматриваемом сценарии представлена на рисунке 4.17. Полученные модельные и объяснительные артефакты далее используются при создании прикладного программного обеспечения для конкретной медицинской задачи.

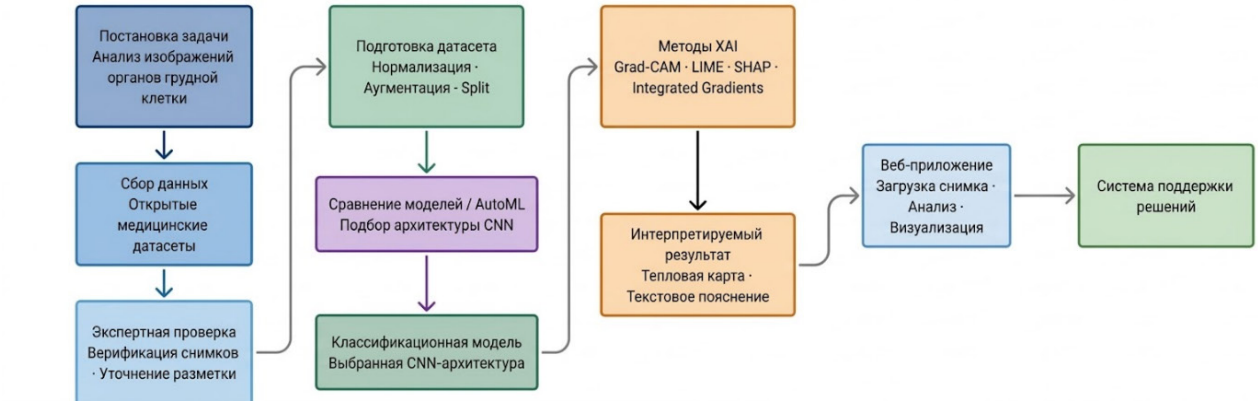


Рисунок 4.17 – Общая схема апробации платформы на задаче автоматизированного анализа рентгенограмм грудной клетки при выявлении признаков пневмонии

4.3.2. Существующие решения

В анализе рентгенограмм грудной клетки (*CXR*, *Chest X-Ray*) в настоящее время активно применяются методы глубокого обучения и компьютерного зрения. На практике такие решения используются либо как системы компьютерной поддержки диагностики, либо как инструменты поддержки принятия врачебных решений. Для настоящего исследования наибольший интерес представляют системы, которые не только выявляют патологические изменения, но и позволяют локализовать подозрительные участки изображения и представить результат в интерпретируемой форме.

К числу наиболее известных решений относится *Lunit INSIGHT CXR*, предназначенная для анализа рентгенограмм грудной клетки и поддержки выявления ряда торакальных патологий, включая консолидацию, плевральный выпот, пневмоторакс и другие изменения [234]. В контексте пневмонии подобная система может использоваться как инструмент поддержки врача при обнаружении участков консолидации и приоритизации подозрительных исследований. Другим примером является *qXR* (*Qure.ai*) — система автоматизированного анализа рентгенограмм грудной клетки, выполняющая предварительный просмотр исследований, выделение патологических находок и поддержку формирования заключения [235]. Решение применяется как инструмент *CAD*/СППР для анализа широкого спектра изменений на *CXR* и, в частности, может использоваться при поиске признаков пневмонии в составе общего анализа лёгочной патологии. К комплексным решениям относится *Annalise Enterprise CXR* (Harrison.ai), предназначенная для ана-

лиза рентгенографии грудной клетки и поддержки выявления большого числа рентгенологических находок. В числе её практических достоинств выделяются наличие функций ОИИ, локализации изменений и интеграции в клинический контур *PACS*/*RIS* [236].

Наряду с рентгенографическими решениями в смежной области представлены системы анализа заболеваний органов грудной клетки по данным компьютерной томографии. К ним относятся *AI-Rad Companion Chest CT* и *AVIEW COPD*, используемые для автоматизированной постобработки, количественного анализа лёгочной ткани и поддержки диагностики хронических заболеваний лёгких [237, 238]. В контексте настоящего исследования такие решения рассматриваются не как прямые аналоги задачи выявления пневмонии на *CXR*, а как примеры переноса интеллектуальных и интерпретируемых контуров на смежные клинические сценарии.

Существующие решения подтверждают высокий потенциал методов глубокого обучения для автоматизированного анализа рентгенограмм грудной клетки, включая выявление воспалительных изменений, локализацию подозрительных областей и поддержку формирования заключения. Вместе с тем задачи многоуровневой интерпретации, сопоставимой оценки различных методов объяснимости, объединения диагностического и объяснительного контуров в единую платформу, а также воспроизводимого выбора объяснителя остаются недостаточно проработанными. Это определяет актуальность апробации платформы как интегрального решения, объединяющего классификацию, интерпретацию, пользовательское представление и оценку качества объяснений.

4.3.3. Постановка задачи

В исследовании рассматривается задача автоматизированного анализа рентгенограмм грудной клетки на платформе. В качестве базового прикладного сценария выбрано выявление признаков пневмонии. Такой выбор обусловлен клинической значимостью задачи, доступностью открытых данных и возможностью воспроизводимой проверки методов классификации и интерпретации в медицинском контуре.

Разрабатываемая система предназначена для поддержки врача-рентгенолога при анализе рентгенограмм грудной клетки. На вход системы подаётся рентгеновское изображение x, а на выходе формируются: классификационное решение $\hat{y}(x)$ о наличии или отсутствии признаков пневмонии, согласованное объяснительное представление $E^*(x)$, включающее визуальные и атрибутивные компоненты, оценка неопределённости $U_{\text{total}}(x)$ и пользовательское представление

результата, пригодное для последующей экспертной интерпретации. В соответствии с архитектурой, разработанной во второй главе, в прикладном сценарии должны быть реализованы следующие компоненты: подготовка и структурирование набора рентгеновских изображений грудной клетки, предобработка медицинских изображений и формирование обучающей, валидационной и тестовой выборок, обучение и сравнительный анализ современных архитектур свёрточных нейронных сетей, построение автоматизированного контура выявления признаков пневмонии и интеграция методов ОИИ, обеспечивающих визуализацию значимых областей и формирование интерпретируемого вывода, а также оценка практической применимости разработанной системы в качестве инструмента поддержки принятия врачебных решений.

В рамках поставленной задачи необходимо различать внешний контур подготовки данных и платформенный контур апробации. Подготовка и верификация рентгенографических данных выполняются отдельно, вне платформы. На платформе в автоматизированном режиме *AutoML* строятся и настраиваются классификационные модели, подбираются объяснительные механизмы и формируется иерархически организованный контур интерпретации результата. Результатом работы платформы в данном сценарии являются обученная модель ИИ, согласованный набор объяснений модели и рекомендации по построению прикладной иерархической системы анализа изображений. Уже на основе указанных артефактов реализуется прикладное программное обеспечение, ориентированное на использование в контуре поддержки принятия врачебных решений.

4.3.4. Введение в предметную область

Пневмония представляет собой инфекционно-воспалительное заболевание лёгких, при котором патологический процесс затрагивает респираторные отделы лёгочной ткани и сопровождается воспалительной экссудацией [230]. Заболевание может протекать с различной степенью тяжести, от относительно лёгких форм до состояний, угрожающих жизни пациента. Клиническая значимость пневмонии определяется риском нарушения газообмена, ухудшения дыхательной функции и развития осложнений, особенно у пациентов пожилого возраста, детей и лиц с сопутствующей хронической патологией [230].

Клинические проявления пневмонии включают повышение температуры тела, кашель, слабость, боль в грудной клетке, одышку и признаки дыхательной недостаточности. Однако клиническая картина заболевания не всегда бывает однозначной, особенно у ослабленных пациентов и лиц старших возрастных групп. В связи с этим важную роль в диагностике пневмонии играют

методы лучевой визуализации, позволяющие подтвердить наличие воспалительных изменений в лёгочной ткани и оценить их распространённость [230, 231].

Одним из основных методов первичной инструментальной диагностики пневмонии является рентгенография органов грудной клетки [230, 231]. Воспалительные изменения на рентгенограммах обычно проявляются в виде участков инфильтрации и уплотнения, визуализируемых как затемнения различной формы, интенсивности и локализации. По этой причине рентгенография широко применяется для подтверждения диагноза и оценки распространённости воспалительного процесса.

Интерпретация рентгенограмм грудной клетки при пневмонии представляет собой сложную диагностическую задачу. Рентгенологическая картина зависит от стадии заболевания, выраженности инфильтрации, возраста пациента и наличия сопутствующей патологии. На ранних этапах пневмонии воспалительные изменения на рентгенограммах могут быть малозаметны, что повышает вероятность пропуска патологии. Диагностику также затрудняет то, что признаки пневмонии нередко сходны с проявлениями других состояний, включая ателектазы, застойные изменения, опухолевые процессы и туберкулёзные очаги [230, 231].

Дополнительные трудности связаны с тем, что разные специалисты могут по-разному интерпретировать один и тот же снимок, особенно при слабовыраженных или пограничных изменениях. Исследования показывают, что согласованность экспертных оценок рентгенологических признаков внебольничной пневмонии остаётся невысокой [232, 233]. Это указывает на объективную сложность визуальной диагностики, ограничивает воспроизводимость результатов и делает особенно востребованными объяснимые системы поддержки принятия решений.

Следовательно, интерпретация рентгенограмм грудной клетки при пневмонии является не только клинически важной, но и методически сложной задачей. Для её решения недостаточно добиться высокой точности классификации, необходим и объяснимый вывод, позволяющий врачу понять, какие признаки изображения повлияли на решение модели. Поэтому сценарий выявления признаков пневмонии выбран в настоящей работе как базовая прикладная задача для апробации платформы.

4.3.5. Практическая реализация

Практическая реализация включала два взаимосвязанных контура: внешний контур подготовки данных и платформенный контур апробации. Подготовка и верификация медицинских данных выполнялись отдельно, вне платформы. На данном этапе использовались открытые дата-

сеты рентгенограмм грудной клетки, включающие нормальные исследования и снимки с признаками пневмонии. Перед формированием итоговой выборки изображения проходили экспертную проверку с участием врача-рентгенолога: некорректные, неоднозначные и низкокачественные случаи исключались. После этого выполнялись разметка классов, предобработка данных и разделение выборки на обучающую, валидационную и тестовую части. Количественная характеристика исходных и итоговых выборок приведена в таблице 4.5.

В качестве основного источника данных использовался набор *Kaggle Chest X-Ray Images* [239], дополненный вспомогательной выборкой изображений для ручной проверки, верификации разметки и отладки методов объяснимости. Для формирования итоговой выборки выполнялись исключение дубликатов, удаление некорректно размеченных или дефектных изображений, нормализация структуры классов и разделение набора на обучающую, валидационную и тестовую подвыборки. Принятая схема подготовки данных обеспечила воспроизводимую основу для последующей апробации платформенного классификационного и объяснительного контуров. Схема формирования итогового датасета представлена на рисунке 4.18.

Таблица 4.5 — Характеристика исходных и итоговых выборок рентгенограмм грудной клетки

Источник данных	**Исходное число изображений**	**Число исключённых изображений**	**Критерии исключения**	**Итоговое число изображений**	**Классы**	**Распределение train/val/test**
Kaggle Chest X-Ray Images (Pneumonia) / Kermany et al. [239]	5 856	624	дубликаты, низкое качество снимка, некорректная или неоднозначная разметка	5 232	*Normal / Pneumonia*	4 186 / 523 / 523
NIH ChestX-ray8 / ChestX-ray14 [240]	108 948	101 746	нерелевантные классы, многометочные случаи вне целевой постановки, низкое качество, дубликаты	7 202	*Normal / Pneumonia*	5 761 / 720 / 721
CheXpert-small [241]	224 316	213 548	нерелевантные наблюдения, *uncertain-label* случаи, низкое качество, несоответствие бинарной постановке	10 768	*Normal / Pneumonia*	8 614 / 1 077 / 1 077

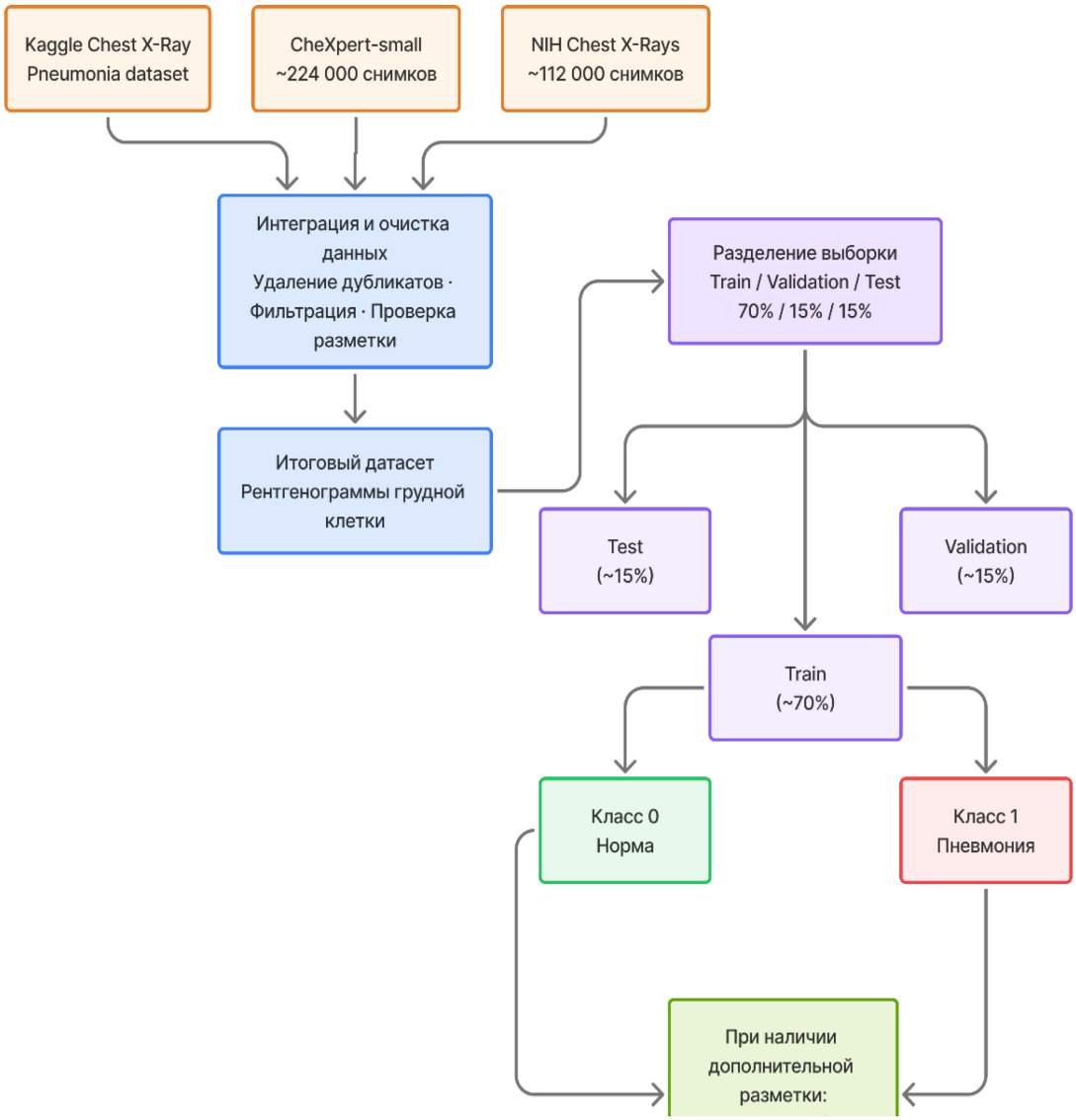


Рисунок 4.18 – Формирование итогового набора данных рентгенограмм грудной клетки

Рисунок 4.18 отражает не только структуру подготовленного набора данных, но и границу между внешним этапом формирования данных и последующей платформенной апробацией. На внешнем этапе формируется и верифицируется набор изображений, после чего на платформе строятся классификационная модель, объяснительный контур и рекомендации по организации иерархической системы анализа.

Для повышения устойчивости модели и снижения риска переобучения в процессе обучения применялась медицински допустимая аугментация рентгеновских изображений грудной клетки. Использовались небольшие геометрические преобразования, не нарушающие анатомическую правдоподобность снимка, включая малые повороты, ограниченные сдвиги и масштаби-

рование, а также умеренные фотометрические преобразования, моделирующие вариации условий съёмки и цифровой обработки, прежде всего изменение яркости и контрастности. Аугментации, способные исказить анатомически корректную пространственную ориентацию рентгенограммы, в контур обучения не включались.

На следующем этапе осуществлялась собственно платформенная апробация. В данной задаче использовался автоматизированный режим *AutoML*, в рамках которого платформа применялась как среда выбора, настройки и оценки моделей глубокого обучения. При этом платформа использовалась не для подготовки исходных медицинских данных, а для построения и оценки модельного и объяснительного контуров на заранее сформированной выборке.

На этапе построения классификационной модели обучение осуществлялось на платформе в автоматизированном режиме *AutoML*, предусматривающем подбор архитектуры, настройку гиперпараметров и сравнительную оценку кандидатов на валидационной выборке. В процедуре отбора рассматривались современные архитектуры свёрточных нейронных сетей, включая *ResNet-34*, *DenseNet121* и *EfficientNetB0*; исходным уровнем служила базовая свёрточная сеть *Custom CNN*. По результатам автоматизированного подбора в качестве основной архитектуры была выбрана модифицированная сеть *ResNet-34*, продемонстрировавшая наилучшее соотношение точности, устойчивости и вычислительной эффективности для задачи анализа рентгенограмм грудной клетки. Модель принимает на вход изображения размером 224×224 пикселей и формирует вероятностный вывод о наличии признаков пневмонии. Сравнительные результаты по архитектурам-кандидатам приведены в таблице 4.6. После завершения обучения выполнялась итоговая оценка качества на тестовой выборке. Для количественной оценки эффекта платформенного отбора результаты выбранной архитектуры сопоставлены с базовой моделью *Custom CNN*, принятой за исходный уровень. При переходе от базовой модели к выбранной средствами *AutoML* сети *ResNet-34* значение *Accuracy* выросло с 0,83 до 0,93, *AUC* с 0,88 до 0,97, *F1-score* с 0,85 до 0,95, то есть на 0,10, 0,09 и 0,10 в абсолютном выражении. Этот эффект характеризует выбор более подходящей архитектуры в фиксированной экспериментальной постановке, а не изменение качества одной и той же модели после её переноса на платформу.

Таблица 4.6 – Сравнение архитектур-кандидатов, рассмотренных в контуре *AutoML* при выявлении признаков пневмонии

Архитектура	**Accuracy**	**AUC**	**F1-score**	**Precision**	**Recall**	**Итог**
ResNet-34 (модель-победитель)	**0,93**	**0,97**	**0,95**	**0,91**	**0,99**	**Выбрана как основная модель**

Архитектура	Accuracy	AUC	F1-score	Precision	Recall	Итог
DenseNet121	0,91	0,95	0,93	0,89	0,97	уступает победителю
EfficientNetB0	0,90	0,95	0,92	0,88	0,96	уступает победителю
ResNet-18	0,89	0,94	0,91	0,87	0,95	уступает победителю
MobileNetV3-Large	0,88	0,93	0,90	0,86	0,94	уступает победителю
ConvNeXt-Tiny	0,89	0,94	0,91	0,88	0,93	уступает победителю
InceptionV3	0,87	0,92	0,89	0,85	0,93	уступает победителю
VGG16	0,85	0,90	0,87	0,83	0,91	уступает победителю
Xception	0,88	0,93	0,90	0,87	0,92	уступает победителю
ViT-B/16	0,86	0,91	0,88	0,84	0,92	уступает победителю
Custom CNN baseline	0,83	0,88	0,85	0,81	0,89	базовый уровень

По результатам автоматизированного отбора архитектур в контуре *AutoML* наилучшие значения основных метрик продемонстрировала модель *ResNet-34*, обеспечившая *Accuracy* = 0,93, *AUC* = 0,97, *F1-score* = 0,95, *Precision* = 0,91 и *Recall = 0,99*. С учётом достигнутого качества классификации, высокой полноты выявления патологических случаев и устойчивости валидационных результатов данная архитектура была выбрана в качестве основной модели для последующего построения объяснительного контура платформы.

Таким образом, на данном этапе платформа используется как среда получения воспроизводимых модельных артефактов, которые затем могут быть перенесены в прикладной программный контур.

Для интерпретации итогового решения в системе был сформирован отдельный объяснительный контур, связанный с классификационной моделью. В реализации использовались методы *Grad-CAM* [61], *LIME* [54], *SHAP* [55] и *Integrated Gradients* [62]. Указанные методы обес-

печивали получение взаимодополняющих визуальных и атрибутивных объяснений, которые сопоставлялись и агрегировались в рамках единого платформенного контура. Их состав, роль и ограничения систематизированы в таблице 4.7 (расширенный перечень методов приведён в таблице Б.1 приложения Б), а архитектура объяснительного контура представлена на рисунке 4.19. В логике главы 2 они выступают как частные механизмы формирования объяснительных объектов на различных уровнях интерпретации.

Карты четырёх методов относятся к одному изображению и имеют общую пространственную природу, поэтому их совместная обработка выполняется в режиме агрегирования (раздел 2.5). Каждая карта приводится к сетке входного изображения 224×224 и нормируется по (3.24), после чего строятся консенсусная карта (3.25), карта строгого согласия (3.26) и карта межметодного разброса (3.27). Консенсусная карта является частным случаем (2.39) с равными весами, что соответствует весам (2.47) при $\alpha = \beta = 0$; при ненулевых α и β вклад метода определяется его качеством Q и неопределённостью U; способ получения таких оценок показан в калибровочном эксперименте (таблица 4.8). Среднее по пикселям значение квадрата карты (3.27) соответствует дисперсионной оценке (2.56)–(2.57) и может служить оценкой межметодной составляющей U_{method} в (2.62). Структурированное представление, передаваемое *LLM*-модулю, содержит результаты этих этапов: классификационное решение $y(x)$ и его вероятность, описание значимых областей агрегированных карт и оценку межметодной неопределённости $U_{\text{method}}(x)$.

В данном сценарии *LLM*-модуль вербализовал агрегированное объяснительное представление. В экспериментальном контуре использовался облачный доступ к языковым моделям. Для клинического развёртывания та же схема рассчитана на локально развёрнутую модель. Конкретная *LLM* не закреплена в архитектуре, модуль получает только структурированный вход, поэтому замена модели не меняет классификационного вывода, карт объяснения и метрик. Из этой роли следуют требования к шаблону инструкции: *LLM* должна назвать класс и его вероятность, описать расположение значимых областей и степень согласия методов, указать оценку неопределённости и не добавлять признаков, диагнозов и рекомендаций, отсутствующих во входном объекте. Чтобы возможные неподтверждённые утверждения не влияли на результаты анализа, текст не возвращается в вычислительный контур и не используется при расчёте метрик качества, неопределённости и выборе объяснителя. Автоматическая фактологическая проверка сгенерированного текста в этой реализации не выполнялась.

Таблица 4.7 – Методы ОИИ, использованные в платформенном контуре при анализе рентгенограмм грудной клетки

Метод	Тип объяснения	Уровень объяснительного контура	Роль в интерпретации	Основные преимущества	Ограничения
Grad-CAM	визуальное, активационное	модельный	локализация областей изображения, повлиявших на решение классификационной модели	нагляден, хорошо применим к *CNN*, удобен для анализа медицинских изображений	ограниченная пространственная точность, зависит от архитектуры сети
LIME	локальное, суррогатное	признаковый	построение локального интерпретируемого приближения решения модели	модельно-независим, даёт локально понятное объяснение	чувствителен к параметрам локальной аппроксимации, возможна нестабильность результата
SHAP	атрибутивное, на основе значений Шепли	признаковый	оценка вклада признаков или областей изображения в итоговое предсказание	строгая теоретическая основа, сопоставимость вкладов	высокая вычислительная стоимость, чувствительность к выбору фонового распределения
Integrated Gradients	атрибутивное, градиентное	признаковый	оценка вклада входных компонент относительно базовой точки	аксиоматически корректен, хорошо применим к глубоким нейронным сетям	зависит от выбора *baseline*, чувствителен к способу задания опорного состояния

Представленные в таблице 4.7 методы образуют фактически использованный объяснительный контур апробации платформы на задаче выявления признаков пневмонии по рентгенограммам грудной клетки. Их совместное применение позволило дополнить классификационный

вывод визуальными и атрибутивными объяснениями, а последующее агрегирование результатов и передача их в LLM-модуль обеспечили переход к связному текстовому пояснению, более удобному для восприятия специалистом.



Рисунок 4.19 – Архитектура объяснительного контура платформы при анализе рентгенограмм грудной клетки

Для количественной проверки протокола воспроизводимого выбора объяснителя из раздела 2.8 выполнен отдельный калибровочный эксперимент. Использовались отдельно обученная модель *ResNet-34* и калибровочная выборка D_{cal} из 80 рентгенограмм, по 40 изображений каждого класса. Для методов *Grad-CAM*, *LIME*, *SHAP* и *Integrated Gradients* фиксировались одни и те же изображения, модель, целевой класс «пневмония» и параметры вычисления. Верность объяснения и межметодное расхождение оценивались на 80 изображениях, устойчивость и повторяемость — на 16, sanity-проверка чувствительности к рандомизации параметров модели — на 8. По этим показателям кандидаты сопоставлялись по качеству, устойчивости, неопределённости и вычислительной стоимости, как предусмотрено протоколом раздела 2.8. Результаты приведены в таблице 4.8.

В таблице 4.8 F — показатель верности объяснения, S — его устойчивость, $U_{\text{внутр}}$ — вариативность результатов повторных запусков метода при фиксированной модели (оценка составляющей U_{run} из (2.62)), $D_{\text{меж}}$ — межметодное расхождение (оценка составляющей U_{method}), *Sanity* — чувствительность объяснения к рандомизации параметров модели, Q — интегральная оценка качества, J — итоговый критерий выбора объяснителя по протоколу раздела 2.8. Межмодельная составляющая U_{model} не оценивалась, поскольку для всех методов использовалась одна модель *ResNet-34*. Пороговые значения, веса критериев и параметры вычисления зафиксированы до сопоставления методов. Полнота q_{comp} из (2.64) в этом эксперименте отдельно не вычислялась и в интегральную оценку Q не входила.

Таблица 4.8 – Сопоставление методов объяснения в калибровочном эксперименте

Метод	**F ↑**	**S ↑**	**Uвнутр ↓**	**Dмеж ↓**	**Sanity ↑**	**Q ↑**	**J ↑**
Grad-CAM	0,532	0,819	0,000	0,182	0,532	0,674	0,642
LIME	0,476	0,542	0,228	0,191	0,530	0,537	0,488
SHAP	0,414	0,869	0,000	0,166	0,559	0,615	0,579
Integrated Gradients	0,467	0,724	0,000	0,151	0,414	0,611	0,586

Всем обязательным пороговым условиям удовлетворил только *Grad-CAM*; итоговый критерий выбора для него составил $J = 0{,}642$. *LIME* не прошёл пороги по F и S, *SHAP* — по F, *Integrated Gradients* — по F, S и *sanity*-проверке. Тем самым основания допуска или исключения каждого кандидата зафиксированы явно. На всей калибровочной выборке верность *Grad-CAM* составила 0,532, на подвыборке из 40 снимков с признаками пневмонии — 0,502. Точечная оценка *sanity*-проверки для *Grad-CAM* выше порога, но её доверительный интервал пересекает границу допуска, поэтому этот результат следует считать пограничным. Эксперимент подтверждает работоспособность процедуры сопоставимого и воспроизводимого выбора объяснителя, но сам по себе не характеризует анатомическую корректность карт и их пригодность для специалиста. Согласие между методами также не подтверждает правильность локализации патологических изменений.

Поэтому результаты дополнительно рассмотрел врач-рентгенолог в рамках качественной экспертной оценки. Оценивались соответствие локализованных областей рентгенологически значимым участкам изображения, понятность визуального представления и возможность использовать объяснение как вспомогательный элемент анализа. По результатам просмотра отмечено содержательное соответствие формируемых областей объяснения рентгенологически значимым участкам исследованных изображений. Экспертная оценка не рассматривалась как самостоятельная клиническая валидация. Замечания эксперта были использованы при формировании пользовательского контура системы.

Результат калибровочного эксперимента показывает и условие, при котором предпочтителен один объяснитель. Если порогам допуска удовлетворяет единственный кандидат, включение остальных методов в агрегирование (2.39) перенесло бы в итоговое объяснение их низкую верность или неустойчивость. Основной контур рисунка 4.19 и калибровочный эксперимент выполнены раздельно: в основном контуре четыре карты агрегируются с равными весами (3.25), а калибровка показывает, что при переносе её порогов в этот контур основой пользовательского представления должна стать карта *Grad-CAM*, а карты LIME, SHAP и Integrated Gradients — сопоставительными свидетельствами, не входящими в агрегат. Совместная обработка нескольких

объяснителей оправдана, когда порогам удовлетворяют несколько кандидатов с близкими значениями J, а межметодное расхождение $D_{\text{меж}}$ невелико: согласованные области усиливаются в консенсусной карте (3.25), а расходящиеся выделяются картой разброса (3.27) и увеличивают U_{method}. Тем самым ошибки отдельных объяснителей не маскируются усреднением, а отражаются в оценке неопределённости.

На завершающем этапе был реализован пользовательский интерфейс системы в виде веб-приложения, обеспечивающего представление классификационного результата, визуального объяснения и текстового сопровождения в единой рабочей форме. Пользовательский слой платформы реализует переход от внутреннего модельного и объяснительного контура к форме представления, пригодной для последующей экспертной интерпретации и использования в контуре поддержки принятия решений. Пример работы реализованного конвейера и пример пользовательского интерфейса приведены соответственно на рисунках 4.20 и 4.21.

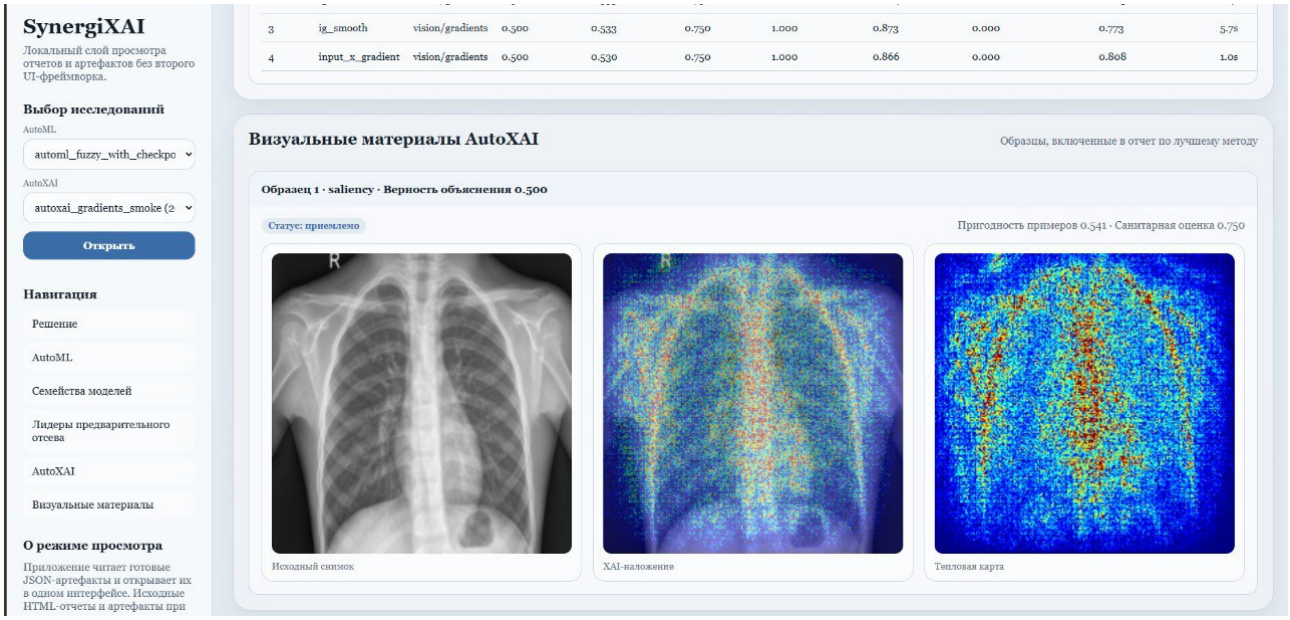


Рисунок 4.20 – Работа реализованного конвейера на базе платформы для анализа рентгенограмм грудной клетки

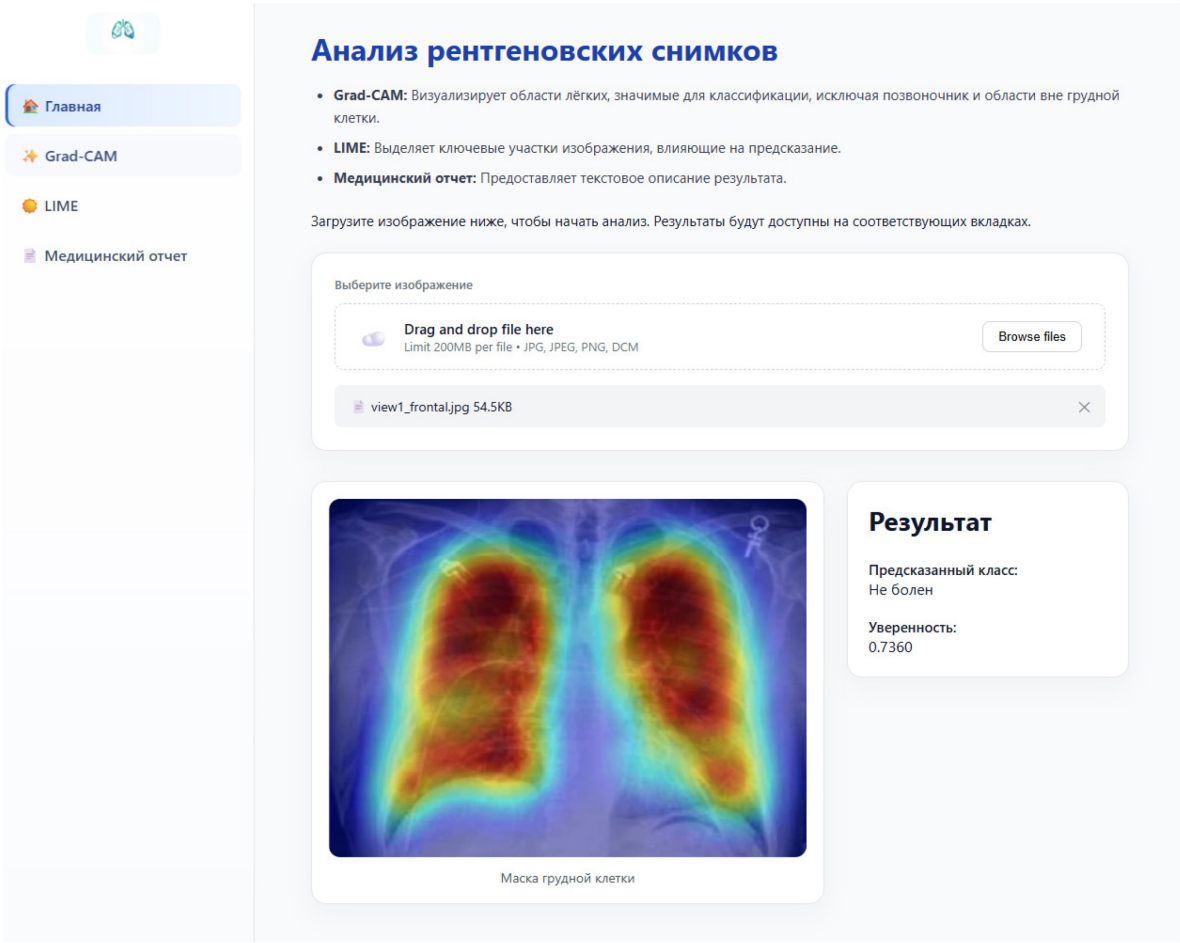


Рисунок 4.21 – Пример пользовательского интерфейса веб-приложения

В рамках работы были разработаны и зарегистрированы программы для ЭВМ, отражающие как базовый сценарий интерпретируемой классификации пневмонии, так и последующее развитие предложенного решения. Первоначально платформенный и объяснительный контур был апробирован на задаче выявления признаков пневмонии по рентгеновским изображениям. В дальнейшем на основе сформированной архитектуры, пользовательского интерфейса и механизмов интерпретации было разработано отдельное веб-приложение для мультиуровневого и интерпретируемого анализа ХОБЛ. В настоящем разделе задача пневмонии рассматривается как исходный прикладной сценарий, а последующее решение для ХОБЛ — как подтверждение переносимости разработанного платформенного подхода на смежные клинические задачи.

После завершения платформенной апробации были получены ключевые артефакты: обученная классификационная модель, подобранный набор объяснительных методов, согласованный контур визуальной и атрибутивной интерпретации, а также рекомендации по организации иерархической системы анализа изображений. Уже на основе этих артефактов был реализован пользовательский интерфейс в виде веб-приложения, обеспечивающего представление классификационного результата, визуального объяснения и текстового сопровождения в единой рабочей форме. В этой логике прикладное программное обеспечение рассматривается не как сама

платформа, а как надстройка над моделями и объяснительным конвейером, полученными в результате её апробации.

4.3.6. Основные выводы по результатам

В разделе 4.3 выполнена прикладная апробация платформы на задаче автоматизированного анализа рентгенограмм грудной клетки при выявлении признаков пневмонии. Полученные результаты показали, что для задач данного типа существенное значение имеют как показатели классификационного качества, так и возможность представить результат анализа в объяснимой форме для клинического анализа.

На внешнем подготовительном этапе выполнены отбор, экспертная верификация и структурирование открытых наборов данных рентгенограмм грудной клетки, включающих исследования с признаками пневмонии и нормальные снимки. В качестве основного источника данных использовался набор *Kaggle Chest X-Ray Images*, дополненный вспомогательной выборкой изображений для ручной проверки, верификации разметки и отладки методов объяснимости. Это обеспечило формирование выборки, пригодной для воспроизводимой апробации платформенного классификационного и объяснительного контуров.

На платформе в автоматизированном режиме *AutoML* выполнены подбор архитектуры и сравнительный отбор моделей-кандидатов. По результатам данного этапа в качестве основной классификационной архитектуры выбрана модифицированная *ResNet-34*, обеспечившая наилучшее сочетание качества классификации, устойчивости и вычислительной эффективности. Одновременно в платформенном контуре были подобраны и согласованы методы объяснения.

Существенным результатом стала интеграция классификационного и объяснительного контуров в единую платформенную схему. Использование нескольких методов ОИИ позволило дополнить классификационный вывод визуальными и атрибутивными объяснениями, локализующими области изображения, повлиявшие на решение модели. Дополнительное формирование текстового представления результата обеспечило переход от внутреннего объяснительного контура к пользовательскому уровню предъявления [242, 243]. Тем самым в прикладном медицинском сценарии апробированы основные элементы теоретико-методического аппарата главы 2: формирование объяснительных объектов, их совместное использование, переход к согласованному объяснению и пользовательскому представлению.

В результате работы платформы получены обученная модель ИИ, подобранные объяснения модели и согласованный иерархический конвейер анализа, пригодный для дальнейшего пе-

реноса в прикладное программное обеспечение. На основе указанных артефактов было реализовано веб-приложение, объединяющее автоматическую классификацию, визуальное объяснение и текстовое сопровождение диагностического вывода. Следовательно, в разделе 4.3 демонстрируется не только частное решение задачи анализа *CXR* при пневмонии, но и прикладное подтверждение выносимого на защиту положения об апробации платформы на задачах биомедицинского анализа изображений.

Одновременно показано, что разработанный в базовом сценарии платформенный контур обладает переносимостью на смежные задачи анализа заболеваний органов грудной клетки. Это подтверждено последующей реализацией веб-приложения для мультиуровневого и интерпретируемого анализа ХОБЛ, построенного на тех же архитектурных и объяснительных механизмах, которые были апробированы на задаче классификации пневмонии. Платформа может рассматриваться как основа для построения интерпретируемых медицинских приложений более общего класса.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена научная задача, заключающаяся в разработке методов, моделей, алгоритмов и программных решений для многоуровневого формирования, согласования, объединения и сопоставимой оценки объяснений в системах ИИ с учётом неопределённости, а также в создании и программной реализации платформы, обеспечивающей воспроизводимый контур их применения в прикладных задачах анализа данных и поддержки принятия решений.

Основные результаты диссертационного исследования состоят в следующем.

1. Выполнен системный обзор современного состояния методов ОИИ, 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.

Полученные результаты в совокупности образуют целостный научно-методический и программно-алгоритмический аппарат ОИИ, ориентированный на многоуровневое формирование и согласование объяснений, их объединение с учётом неопределённости, а также на обеспечение воспроизводимости и сопоставимости результатов интерпретации. Предложенные решения обеспечивают переход от использования отдельных объясняющих методов к их системной интеграции в составе единого вычислительного контура, что расширяет возможности применения ОИИ в задачах с высокой ценой ошибки и повышенными требованиями к обоснованности, проверяемости и практической значимости результата.

СПИСОК ЛИТЕРАТУРЫ

1. О развитии искусственного интеллекта в Российской Федерации : Указ Президента Российской Федерации от 10 октября 2019 г. № 490. – Текст : электронный // Официальный интернет-портал правовой информации. – URL: http://publication.pravo.gov.ru (дата обращения: 01.10.2025).
2. О внесении изменений в Указ Президента Российской Федерации от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» и в Национальную стратегию, утвержденную этим Указом : Указ Президента Российской Федерации от 15 февраля 2024 г. № 124. – Текст : электронный // Официальный интернет-портал правовой информации. – URL: http://publication.pravo.gov.ru (дата обращения: 01.10.2025).
3. Об утверждении стратегического направления в области цифровой трансформации здравоохранения : распоряжение Правительства Российской Федерации от 17 апреля 2024 г. № 959-р. – Текст : электронный // Официальный интернет-портал правовой информации. – URL: http://publication.pravo.gov.ru (дата обращения: 01.10.2025).
4. Об установлении экспериментального правового режима в сфере цифровых инноваций и утверждении Программы экспериментального правового режима в сфере цифровых инноваций по направлению медицинской деятельности, в том числе с применением телемедицинских технологий и технологий сбора и обработки сведений о состоянии здоровья и диагнозах граждан : постановление Правительства Российской Федерации от 18 июля 2023 г. № 1164. – Текст : электронный // Официальный интернет-портал правовой информации. – URL: http://publication.pravo.gov.ru (дата обращения: 01.10.2025).
5. LeCun, Y. Deep learning / Y. LeCun, Y. Bengio, G. Hinton // Nature. – 2015. – Vol. 521, no. 7553. – P. 436–444. – DOI 10.1038/nature14539.
6. Schmidhuber, J. Deep learning in neural networks: An overview / J. Schmidhuber // Neural Networks. – 2015. – Vol. 61. – P. 85–117. – DOI 10.1016/j.neunet.2014.09.003.
7. Doshi-Velez, F. Towards a rigorous science of interpretable machine learning / F. Doshi-Velez, B. Kim. – Текст : электронный // arXiv. – 2017. – arXiv:1702.08608. – URL: https://arxiv.org/abs/1702.08608 (дата обращения: 01.10.2026).
8. Lipton, Z. C. The mythos of model interpretability / Z. C. Lipton // Queue. – 2018. – Vol. 16, no. 3. – P. 31–57. – DOI 10.1145/3236386.3241340.
9. Miller, T. Explanation in artificial intelligence: Insights from the social sciences / T. Miller // Artificial Intelligence. – 2019. – Vol. 267. – P. 1–38. – DOI 10.1016/j.artint.2018.07.007.

10. Adadi, A. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI) / A. Adadi, M. Berrada // IEEE Access. – 2018. – Vol. 6. – P. 52138–52160. – DOI 10.1109/ACCESS.2018.2870052.
11. A survey of methods for explaining black box models / R. Guidotti, A. Monreale, S. Ruggieri [et al.] // ACM Computing Surveys. – 2018. – Vol. 51, no. 5. – Art. 93. – DOI 10.1145/3236009.
12. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI / A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser [et al.] // Information Fusion. – 2020. – Vol. 58. – P. 82–115. – DOI 10.1016/j.inffus.2019.12.012.
13. Interpretable machine learning: Fundamental principles and 10 grand challenges / C. Rudin, C. Chen, Z. Chen [et al.] // Statistics Surveys. – 2022. – Vol. 16. – P. 1–85. – DOI 10.1214/21-SS133.
14. Saeed, W. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities / W. Saeed, C. Omlin // Knowledge-Based Systems. – 2023. – Vol. 263. – Art. 110273. – DOI 10.1016/j.knosys.2023.110273.
15. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead / C. Rudin // Nature Machine Intelligence. – 2019. – Vol. 1, no. 5. – P. 206–215. – DOI 10.1038/s42256-019-0048-x.
16. Causability and explainability of artificial intelligence in medicine / A. Holzinger, G. Langs, H. Denk [et al.] // WIREs Data Mining and Knowledge Discovery. – 2019. – Vol. 9, no. 4. – Art. e1312. – DOI 10.1002/widm.1312.
17. Tjoa, E. A survey on explainable artificial intelligence (XAI): Toward medical XAI / E. Tjoa, C. Guan // IEEE Transactions on Neural Networks and Learning Systems. – 2021. – Vol. 32, no. 11. – P. 4793–4813. – DOI 10.1109/TNNLS.2020.3027314.
18. Trustworthy AI: From principles to practices / B. Li, P. Qi, B. Liu [et al.] // ACM Computing Surveys. – 2023. – Vol. 55, no. 9. – Art. 177. – DOI 10.1145/3555803.
19. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions / L. Longo, M. Brcic, F. Cabitza [et al.] // Information Fusion. – 2024. – Vol. 106. – Art. 102301. – DOI 10.1016/j.inffus.2024.102301.
20. Montavon, G. Methods for interpreting and understanding deep neural networks / G. Montavon, W. Samek, K.-R. Müller // Digital Signal Processing. – 2018. – Vol. 73. – P. 1–15. – DOI 10.1016/j.dsp.2017.10.011.
21. 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.

22. Artificial Intelligence Risk Management Framework (AI RMF 1.0) : NIST AI 100-1. – Gaithersburg : National Institute of Standards and Technology, 2023. – DOI 10.6028/NIST.AI.100-1.
23. ISO/IEC TR 24028:2020. Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence. – Geneva : International Organization for Standardization, 2020.
24. Ларичев, О. И. Проблемы взаимодействия человек-ЭВМ в системах поддержки принятия решений / О. И. Ларичев // Процедуры оценивания многокритериальных объектов : сборник трудов / под ред. О. И. Ларичева. – Москва : ВНИИСИ, 1984. – Вып. 9. – С. 20–28.
25. Ларичев, О. И. Вербальный анализ решений / О. И. Ларичев. – Москва : Наука, 2006. – 181 с.
26. Берталанфи, Л. фон. Общая теория систем: критический обзор / Л. фон Берталанфи // Исследования по общей теории систем : сборник переводов. – Москва : Прогресс, 1969. – С. 23–82.
27. Клир, Дж. Системология. Автоматизация решения системных задач / Дж. Клир ; пер. с англ. – Москва : Радио и связь, 1990. – 544 с.
28. Волкова, В. Н. Теория систем и системный анализ : учебник для вузов / В. Н. Волкова, А. А. Денисов. – Москва : Юрайт, 2019. – 462 с.
29. Моисеев, Н. Н. Математические задачи системного анализа / Н. Н. Моисеев. – Москва : Наука, 1981. – 488 с.
30. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis / B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, M. A. Viergever // Medical Image Analysis. – 2022. – Vol. 79. – Art. 102470. – DOI 10.1016/j.media.2022.102470.
31. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. – Geneva : International Organization for Standardization, 2023.
32. Shortliffe, E. H. Computer-Based Medical Consultations: MYCIN / E. H. Shortliffe. – New York : Elsevier, 1976.
33. Buchanan, B. G. Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project / B. G. Buchanan, E. H. Shortliffe. – Reading, MA : Addison-Wesley, 1984.
34. Explaining deep neural networks and beyond: A review of methods and applications / W. Samek, G. Montavon, S. Lapuschkin [et al.] // Proceedings of the IEEE. – 2021. – Vol. 109, no. 3. – P. 247–278. – DOI 10.1109/JPROC.2021.3060483.

35. Vilone, G. Notions of explainability and evaluation approaches for explainable artificial intelligence / G. Vilone, L. Longo // Information Fusion. – 2021. – Vol. 76. – P. 89–106. – DOI 10.1016/j.inffus.2021.05.009.
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 т. (Москва, 27–28 марта 2025 г.). – Москва : Московский университет им. С. Ю. Витте, 2025. – С. 134–151. – EDN AMOTXS.
37. ГОСТ Р 59276–2020. Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения : национальный стандарт Российской Федерации : издание официальное. – Москва : Стандартинформ, 2020.
38. Explaining explanations: An overview of interpretability of machine learning / L. H. Gilpin, D. Bau, B. Z. Yuan [et al.] // 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). – Turin, 2018. – P. 80–89. – DOI 10.1109/DSAA.2018.00018.
39. Doran, D. What does explainable AI really mean? A new conceptualization of perspectives / D. Doran, S. Schulz, T. R. Besold. – Текст : электронный // arXiv. – 2017. – arXiv:1710.00794. – URL: https://arxiv.org/abs/1710.00794 (дата обращения: 02.01.2026).
40. Carvalho, D. V. Machine learning interpretability: A survey on methods and metrics / D. V. Carvalho, E. M. Pereira, J. S. Cardoso // Electronics. – 2019. – Vol. 8, no. 8. – Art. 832. – DOI 10.3390/electronics8080832.
41. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI / A. Jacovi, A. Marasović, T. Miller, Y. Goldberg // Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21). – New York : ACM, 2021. – P. 624–635. – DOI 10.1145/3442188.3445923.
42. Sokol, K. Explainability fact sheets: A framework for systematic assessment of explainable approaches / K. Sokol, P. Flach // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). – New York : ACM, 2020. – P. 56–67. – DOI 10.1145/3351095.3372870.
43. Explainable machine learning in deployment / U. Bhatt, A. Xiang, S. Sharma [et al.] // Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20). – New York : ACM, 2020. – P. 648–657. – DOI 10.1145/3351095.3375624.
44. Definitions, methods, and applications in interpretable machine learning / W. J. Murdoch, C. Singh, K. Kumbier [et al.] // Proceedings of the National Academy of Sciences. – 2019. – Vol. 116, no. 44. – P. 22071–22080. – DOI 10.1073/pnas.1900654116.

45. Vassiliades, A. Argumentation and explainable artificial intelligence: a survey / A. Vassiliades, N. Bassiliades, T. Patkos // The Knowledge Engineering Review. – 2021. – Vol. 36. – Art. e5. – DOI 10.1017/S0269888921000011.
46. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems / R. Tomsett, D. Braines, D. Harborne [et al.] // ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). – Stockholm, 2018. – arXiv:1806.07552.
47. 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.
48. ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. – Geneva : International Organization for Standardization, 2023.
49. NISTIR 8269 (Draft). A Taxonomy and Terminology of Adversarial Machine Learning. – Gaithersburg : National Institute of Standards and Technology, 2019. – DOI 10.6028/NIST.IR.8269-draft.
50. ISO/IEC TR 24029-1:2021. Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview. – Geneva : International Organization for Standardization, 2021.
51. OECD Principles on Artificial Intelligence (Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449). – Paris : Organisation for Economic Co-operation and Development, 2019.
52. Recommendation on the Ethics of Artificial Intelligence. – Paris : United Nations Educational, Scientific and Cultural Organization (UNESCO), 2021.
53. Кодекс этики в сфере искусственного интеллекта. – Москва : Альянс в сфере искусственного интеллекта, 2021. – Текст : электронный. – URL: https://ethics.a-ai.ru/ (дата обращения: 01.10.2026).
54. Ribeiro, M. T. “Why should I trust you?”: Explaining the predictions of any classifier / M. T. Ribeiro, S. Singh, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). – New York : ACM, 2016. – P. 1135–1144. – DOI 10.1145/2939672.2939778.
55. Lundberg, S. M. A unified approach to interpreting model predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 4765–4774.
56. Zeiler, M. D. Visualizing and understanding convolutional networks / M. D. Zeiler, R. Fergus // Computer Vision – ECCV 2014. – Cham : Springer, 2014. – P. 818–833. – (Lecture Notes in Computer Science ; vol. 8689). – DOI 10.1007/978-3-319-10590-1_53.

57. Is object localization for free? Weakly-supervised learning with convolutional neural networks / M. Oquab, L. Bottou, I. Laptev, J. Sivic // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). – 2015. – P. 685–694. – DOI 10.1109/CVPR.2015.7298668.
58. Robnik-Šikonja, M. Explaining classifications for individual instances / M. Robnik-Šikonja, I. Kononenko // IEEE Transactions on Knowledge and Data Engineering. – 2008. – Vol. 20, no. 5. – P. 589–600. – DOI 10.1109/TKDE.2007.190734.
59. Learning deep features for discriminative localization / B. Zhou, A. Khosla, A. Lapedriza [et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). – 2016. – P. 2921–2929. – DOI 10.1109/CVPR.2016.319.
60. Wachter, S. Counterfactual explanations without opening the black box: Automated decisions and the GDPR / S. Wachter, B. Mittelstadt, C. Russell // Harvard Journal of Law & Technology. – 2018. – Vol. 31, no. 2. – P. 841–887.
61. Grad-CAM: Visual explanations from deep networks via gradient-based localization / R. R. Selvaraju, M. Cogswell, A. Das [et al.] // Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017). – 2017. – P. 618–626. – DOI 10.1109/ICCV.2017.74.
62. Sundararajan, M. Axiomatic attribution for deep networks / M. Sundararajan, A. Taly, Q. Yan // Proceedings of the 34th International Conference on Machine Learning (ICML 2017). – 2017. – P. 3319–3328. – (Proceedings of Machine Learning Research ; vol. 70).
63. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation / S. Bach, A. Binder, G. Montavon [et al.] // PLoS ONE. – 2015. – Vol. 10, no. 7. – Art. e0130140. – DOI 10.1371/journal.pone.0130140.
64. Alvarez-Melis, D. On the robustness of interpretability methods / D. Alvarez-Melis, T. S. Jaakkola // ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). – Stockholm, 2018. – arXiv:1806.08049.
65. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods / D. Slack, S. Hilgard, E. Jia [et al.] // Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES ’20). – New York : ACM, 2020. – P. 180–186. – DOI 10.1145/3375627.3375830.
66. The disagreement problem in explainable machine learning: A practitioner’s perspective / S. Krishna, T. Han, A. Gu [et al.] – Текст : электронный // Transactions on Machine Learning Research. – 2024. – URL: https://arxiv.org/abs/2202.01602 (дата обращения: 03.11.2025).
67. Hase, P. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? / P. Hase, M. Bansal // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). – 2020. – P. 5540–5552. – DOI 10.18653/v1/2020.acl-main.491.

68. General pitfalls of model-agnostic interpretation methods for machine learning models / C. Molnar, G. König, J. Herbinger [et al.] // xxAI – Beyond Explainable AI : International Workshop, Held in Conjunction with ICML 2020. – Cham : Springer, 2022. – P. 39–68. – (Lecture Notes in Computer Science ; vol. 13200). – DOI 10.1007/978-3-031-04083-2_4.
69. Simonyan, K. Deep inside convolutional networks: Visualising image classification models and saliency maps / K. Simonyan, A. Vedaldi, A. Zisserman. – Текст : электронный // arXiv. – 2013. – arXiv:1312.6034. – URL: https://arxiv.org/abs/1312.6034 (дата обращения: 01.10.2026).
70. Sanity checks for saliency maps / J. Adebayo, J. Gilmer, M. Muelly [et al.] // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 9525–9536.
71. Explanations can be manipulated and geometry is to blame / A.-K. Dombrowski, M. Alber, C. J. Anders [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.
72. Rajab, S. Handling interpretability issues in ANFIS using rule base simplification and constrained learning / S. Rajab // Fuzzy Sets and Systems. – 2019. – Vol. 368. – P. 36–58. – DOI 10.1016/j.fss.2018.11.010.
73. 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.
74. Concept of Hierarchically Organized Explainable Intelligent Systems: Synthesis of Deep Neural Networks, Fuzzy Logic and Incremental Learning in Medical Diagnostics / Y. V. Trofimov, A. V. Shevchenko, A. N. Averkin [et al.] // 2025 VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). – Saint Petersburg, 2025. – P. 14–17. – DOI 10.1109/NeuroNT66873.2025.11049976.
75. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model / B. Letham, C. Rudin, T. H. McCormick, D. Madigan // The Annals of Applied Statistics. – 2015. – Vol. 9, no. 3. – P. 1350–1371. – DOI 10.1214/15-AOAS848.
76. Learning certifiably optimal rule lists for categorical data / E. Angelino, N. Larus-Stone, D. Alabi [et al.] // Journal of Machine Learning Research. – 2018. – Vol. 18, no. 234. – P. 1–78.
77. Moving beyond post hoc explainable artificial intelligence: a perspective paper on lessons learned from dynamical climate modeling / R. J. O'Loughlin, D. Li, R. Neale, T. A. O'Brien // Geoscientific Model Development. – 2025. – Vol. 18. – P. 787–802. – DOI 10.5194/gmd-18-787-2025.

78. 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 г.). – Санкт-Петербург : СПб ФИЦ РАН, 2025. – С. 97–107. – DOI 10.15622/rcai.2025.040.

79. 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.

80. Shihabudheen, K. V. Regularized extreme learning adaptive neuro-fuzzy algorithm for regression and classification / K. V. Shihabudheen, G. N. Pillai // Knowledge-Based Systems. – 2017. – Vol. 127. – P. 100–113. – DOI 10.1016/j.knosys.2017.04.007.

81. Constrained neuro fuzzy inference methodology for explainable personalised modelling with applications on gene expression data / B. Singh, M. Doborjeh, Z. Doborjeh [et al.] // Scientific Reports. – 2023. – Vol. 13. – Art. 456. – DOI 10.1038/s41598-022-27132-8.

82. Interpretability of machine learning-based prediction models in healthcare / G. Stiglic, P. Kocbek, N. Fijacko [et al.] // WIREs Data Mining and Knowledge Discovery. – 2020. – Vol. 10, no. 5. – Art. e1379. – DOI 10.1002/widm.1379.

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

84. 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.

85. 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.

86. 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.

87. 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 т. (Москва, 25–27 апреля 2025 г.). – Москва : РУДН, 2025. – С. 390–394. – EDN KUYLYY.
88. 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, January 29, 2025). – Delhi : Infinity, 2025. – P. 159–167. – DOI 10.34660/INF.2025.79.43.089. – EDN SOHBOR.
89. RuFLEX: A Hybrid Deep Neuro-Fuzzy Intelligence Platform for Geospatial Data with Logical Rule Traceability and Counterfactual Explainability / Y. V. Trofimov, A. N. Averkin, M. D. Lebedev [et al.] // 2026 VII International Conference on Neural Networks and Neurotechnologies (NeuroNT). – Saint Petersburg, 2026. – P. 11–14. – DOI 10.1109/NeuroNT71829.2026.11651423.
90. This looks like that: Deep learning for interpretable image recognition / C. Chen, O. Li, D. Tao [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 8930–8941.
91. Alvarez-Melis, D. Towards robust interpretability with self-explaining neural networks / D. Alvarez-Melis, T. S. Jaakkola // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 7775–7784.
92. 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 г.). – Санкт-Петербург : СПб ФИЦ РАН, 2025. – С. 108–119. – DOI 10.15622/rcai.2025.041.
93. 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. – № 1-2. – С. 41–60. – DOI 10.36871/26189976.2026.01-2.005.
94. Reproducibility in machine-learning-based research: Overview, barriers, and drivers / H. Semmelrock, T. Ross-Hellauer, S. Kopeinik [et al.] // AI Magazine. – 2025. – Vol. 46, no. 2. – Art. e70002. – DOI 10.1002/aaai.70002.
95. Gundersen, O. E. Do machine learning platforms provide out-of-the-box reproducibility? / O. E. Gundersen, S. Shamsaliei, R. J. Isdahl // Future Generation Computer Systems. – 2022. – Vol. 126. – P. 34–47. – DOI 10.1016/j.future.2021.06.014.

96. Schlegel, M. Capturing end-to-end provenance for machine learning pipelines / M. Schlegel, K.-U. Sattler // Information Systems. – 2025. – Vol. 132. – Art. 102495. – DOI 10.1016/j.is.2024.102495.
97. Kreuzberger, D. Machine learning operations (MLOps): Overview, definition, and architecture / D. Kreuzberger, N. Kühl, S. Hirschl // IEEE Access. – 2023. – Vol. 11. – P. 31866–31879. – DOI 10.1109/ACCESS.2023.3262138.
98. Why should I choose you? AutoXAI: a framework for selecting and tuning eXplainable AI solutions / R. Cugny, J. Aligon, M. Chevalier [et al.] // Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22). – New York : ACM, 2022. – P. 3429–3433.
99. AutoXAI4Omics: an automated explainable AI tool for omics and tabular data / J. Strudwick, L. Gardiner, K. Denning-James [et al.] // Briefings in Bioinformatics. – 2025. – Vol. 26, no. 1. – Art. bbae593. – DOI 10.1093/bib/bbae593.
100. Bifarin, O. O. Automated machine learning and explainable AI (AutoML-XAI) for metabolomics: improving cancer diagnostics / O. O. Bifarin, F. M. Fernández // Journal of the American Society for Mass Spectrometry. – 2024. – Vol. 35, no. 6. – P. 1089–1100.
101. El Shawi, R. AutoXAI: a meta-learning approach for recommendation of explanation techniques / R. El Shawi, L. Jamel // Scientific Reports. – 2025. – Vol. 15. – Art. 41912. – DOI 10.1038/s41598-025-25872-x.
102. XAutoML: A Visual Analytics Tool for Understanding and Validating Automated Machine Learning / M.-A. Zöller, W. Titov, T. Schlegel, M. F. Huber // ACM Transactions on Interactive Intelligent Systems. – 2023. – Vol. 13, no. 4. – DOI 10.1145/3625240.
103. UniAutoML: A Human-Centered Framework for Unified Discriminative and Generative AutoML with Large Language Models / J. Guo, Z. Chen, Y. Ji [et al.]. – Текст : электронный // arXiv. – 2024. – arXiv:2410.12841. – URL: https://arxiv.org/abs/2410.12841 (дата обращения: 12.02.2026).
104. Explainable AutoML (xAutoML) with adaptive modeling for yield enhancement in semiconductor smart manufacturing / W. Zhai, X. Shi, Y. D. Wong [et al.]. – Текст : электронный // arXiv. – 2024. – arXiv:2403.12381. – URL: https://arxiv.org/abs/2403.12381 (дата обращения: 12.02.2026).
105. Cofaru, C. A knowledge-driven AutoML architecture / C. Cofaru, J. Loeckx. – Текст : электронный // arXiv. – 2023. – arXiv:2311.17124. – URL: https://arxiv.org/abs/2311.17124 (дата обращения: 18.02.2026).

106. Pereira, A. C. G. F. Integration of an Explainable Dashboard to Enhance AutoML Transparency / A. C. G. F. Pereira, A. M. A. Maciel // Lecture Notes in Networks and Systems. – Cham : Springer, 2026. – Vol. 1717. – P. 725–736. – DOI 10.1007/978-3-032-10721-3_62.
107. Regularized evolution for image classifier architecture search / E. Real, A. Aggarwal, Y. Huang, Q. V. Le // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, no. 1. – P. 4780–4789. – DOI 10.1609/aaai.v33i01.33014780.
108. Elsken, T. Neural architecture search: A survey / T. Elsken, J. H. Metzen, F. Hutter // Journal of Machine Learning Research. – 2019. – Vol. 20, no. 55. – P. 1–21.
109. Explainable AI evaluation: A top-down approach for selecting optimal explanations for black box models / S. Mirzaei, H. Mao, R. R. O. Al-Nima, W. L. Woo // Information. – 2024. – Vol. 15, no. 1. – Art. 4. – DOI 10.3390/info15010004.
110. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.
111. Dietterich, T. G. Ensemble methods in machine learning / T. G. Dietterich // Multiple Classifier Systems. MCS 2000. – Berlin : Springer, 2000. – P. 1–15. – (Lecture Notes in Computer Science ; vol. 1857). – DOI 10.1007/3-540-45014-9_1.
112. A survey on ensemble learning / X. Dong, Z. Yu, W. Cao [et al.] // Frontiers of Computer Science. – 2020. – Vol. 14, no. 2. – P. 241–258. – DOI 10.1007/s11704-019-8208-z.
113. Reingold, O. Dissenting explanations: Leveraging disagreement to reduce model overreliance / O. Reingold, J. H. Shen, A. Talati. – Текст : электронный // arXiv. – 2023. – arXiv:2307.07636. – URL: https://arxiv.org/abs/2307.07636 (дата обращения: 01.10.2026).
114. Agree to disagree: Exploring consensus of XAI methods for ML-based NIDS / K. Dietz, M. Hajizadeh, J. Schleicher [et al.] // 2024 20th International Conference on Network and Service Management (CNSM). – 2024. – P. 1–7.
115. Boehmer, N. Rank aggregation using scoring rules / N. Boehmer, R. Bredereck, D. Peters // Proceedings of the AAAI Conference on Artificial Intelligence. – 2023. – Vol. 37, no. 5. – P. 5515–5523. – DOI 10.1609/aaai.v37i5.25685.
116. Conitzer, V. Common voting rules as maximum likelihood estimators / V. Conitzer, T. Sandholm // Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI 2005). – Arlington : AUAI Press, 2005. – P. 145–152.
117. Chatterjee, S. Multi-criteria rank-based aggregation for explainable AI / S. Chatterjee, E. R. Colombo, M. M. Raimundo. – Текст : электронный // arXiv. – 2025. – arXiv:2505.24612. – URL: https://arxiv.org/abs/2505.24612 (дата обращения: 01.10.2026).

118. Joshi, G. A review on explainability in multimodal deep neural nets / G. Joshi, R. Walambe, K. Kotecha // IEEE Access. – 2021. – Vol. 9. – P. 59800–59821. – DOI 10.1109/ACCESS.2021.3070212.
119. Miró-Nicolau, M. A comprehensive study on fidelity metrics for XAI / M. Miró-Nicolau, A. Jaume-i-Capó, G. Moyà-Alcover // Information Processing & Management. – 2025. – Vol. 62, no. 1. – Art. 103900. – DOI 10.1016/j.ipm.2024.103900.
120. A benchmark for interpretability methods in deep neural networks / S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 9737–9748.
121. On the (in)fidelity and sensitivity of explanations / C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala [et al.] // Advances in Neural Information Processing Systems. – 2019. – Vol. 32. – P. 10967–10978.
122. Evaluating the robustness of explainable AI in medical image recognition under natural and adversarial data corruption / S. Repetto, I. Maljkovic, M. Lotto [et al.] // Machine Learning. – 2026. – Vol. 115, no. 1. – Art. 4. – DOI 10.1007/s10994-025-06919-6.
123. Minsky, M. Steps toward artificial intelligence / M. Minsky // Computers and Thought / ed. by E. A. Feigenbaum, J. Feldman. – New York : McGraw-Hill, 1963. – P. 406–450.
124. Саймон, Г. Науки об искусственном / Г. Саймон ; пер. с англ. – 2-е изд. – Москва : Едиториал УРСС, 2004. – 144 с.
125. Месарович, М. Теория иерархических многоуровневых систем / М. Месарович, Д. Мако, И. Такахара ; пер. с англ. – Москва : Мир, 1973. – 344 с.
126. Моисеев, Н. Н. Информационная теория иерархических систем / Н. Н. Моисеев // Избранные труды : в 2 т. / под ред. А. А. Петрова. – Москва : Тайдекс Ко, 2003. – Т. 1. – С. 214–266.
127. Белов, П. Г. Системный анализ и моделирование опасных процессов в техносфере : учебное пособие для студентов высших учебных заведений / П. Г. Белов. – Москва : Академия, 2003. – 512 с.
128. Castillo, E. Functional networks / E. Castillo // Neural Processing Letters. – 1998. – Vol. 7, no. 3. – P. 151–159. – DOI 10.1023/A:1009656525752.
129. Гаврилова, Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский. – Санкт-Петербург : Питер, 2000. – 384 с.
130. Gruber, T. R. A translation approach to portable ontology specifications / T. R. Gruber // Knowledge Acquisition. – 1993. – Vol. 5, no. 2. – P. 199–220. – DOI 10.1006/knac.1993.1008.

131. Hybrid and Hierarchical Explainable AI Based on Kolmogorov-Arnold Networks / A. N. Averkin, Y. V. Trofimov, A. D. Lebedev, A. S. Ilin // Proceedings of the Ninth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’25). Vol. 2. – Cham : Springer, 2026. – P. 36–43. – (Lecture Notes in Networks and Systems ; vol. 1763). – DOI 10.1007/978-3-032-13612-1_4.
132. 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 г.). – Санкт-Петербург : СПбГЭТУ «ЛЭТИ», 2025. – С. 3–6. – EDN CJPUMF.
133. 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, January 15, 2025). – Beijing : Infinity, 2025. – P. 130–135. – DOI 10.34660/INF.2025.87.97.124. – EDN CDNNDC.
134. True to the model or true to the data? / H. Chen, J. D. Janizek, S. Lundberg, S.-I. Lee. – Текст : электронный // arXiv. – 2020. – arXiv:2006.16234. – URL: https://arxiv.org/abs/2006.16234 (дата обращения: 01.10.2026).
135. Zadeh, L. A. Fuzzy sets / L. A. Zadeh // Information and Control. – 1965. – Vol. 8, no. 3. – P. 338–353. – DOI 10.1016/S0019-9958(65)90241-X.
136. Improving performance of deep learning models with axiomatic attribution priors and expected gradients / G. Erion, J. D. Janizek, P. Sturmfels [et al.] // Nature Machine Intelligence. – 2021. – Vol. 3. – P. 620–631. – DOI 10.1038/s42256-021-00343-w.
137. 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. – № 2-2. – DOI 10.36871/26189976.2026.02-2.008.
138. Методы и модели согласования иерархических решений / под ред. А. А. Макарова. – Новосибирск : Наука, 1979. – 240 с.
139. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск : Изд-во Ин-та математики, 1999. – 270 с.
140. Ройзензон, Г. В. Синергетический эффект в принятии решений / Г. В. Ройзензон // Системные исследования. Методологические проблемы : ежегодник / под ред. Ю. С. Попкова, В. Н. Садовского, В. И. Тищенко. – Москва : URSS, 2012. – Вып. 36. – С. 248–272.

141. Schwalbe, G. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts / G. Schwalbe, B. Finzel // Data Mining and Knowledge Discovery. – 2024. – Vol. 38, no. 5. – P. 3043–3101. – DOI 10.1007/s10618-022-00867-8.
142. Макеев, С. П. Упорядочение объектов в иерархических системах / С. П. Макеев, И. Ф. Шахнов // Известия АН СССР. Техническая кибернетика. – 1991. – № 3. – С. 29–46.
143. Временная атрибуция в стохастических диффузионных процессах: математическая формализация, объяснимость и каузальная валидация / Ю. В. Трофимов, А. Н. Аверкин, М. А. Лопатин [и др.] // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2026. – № 1. – С. 89–137. – DOI 10.26456/vtpmk773.
144. Fusion-driven multimodal learning for biomedical time series in surgical care / J. Che, M. Sun, Y. Wang, Z. Xu // Frontiers in Physiology. – 2025. – Vol. 16. – Art. 1605406. – DOI 10.3389/fphys.2025.1605406.
145. Physiological signals fusion oriented to diagnosis – a review / Y. F. Uribe, K. C. Alvarez-Uribe, D. H. Peluffo-Ordoñez, M. A. Becerra // Advances in Computing. CCC 2018. – Cham : Springer, 2018. – P. 1–15. – (Communications in Computer and Information Science ; vol. 885). – DOI 10.1007/978-3-319-98998-3_1.
146. Мультимодальная объяснимость для ОРИТ-сигналов: метрические и асимптотические результаты / Ю. В. Трофимов, А. Н. Аверкин, Е. М. Кузнецов [и др.] // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2025. – № 4. – С. 1–28.
147. A wide and deep transformer neural network for 12-lead ECG classification / A. Natarajan, Y. Chang, S. Mariani [et al.] // 2020 Computing in Cardiology (CinC). – Rimini, 2020. – P. 1–4. – DOI 10.22489/CinC.2020.107.
148. A transformer-based framework for multivariate time series representation learning / G. Zerveas, S. Jayaraman, D. Patel [et al.] // Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD ’21). – New York : ACM, 2021. – P. 2114–2124. – DOI 10.1145/3447548.3467401.
149. Physio-JEPA for ICU Signals: Temporal Multimodal Explainability and the Temporal Range Coherence Metric for Ventricular Tachycardia Alarm Validation / E. Kuznetsov, A. Averkin, Y. Trofimov [et al.] // 2026 VII International Conference on Neural Networks and Neurotechnologies (NeuroNT). – Saint Petersburg, 2026. – P. 28–31. – DOI 10.1109/NeuroNT71829.2026.11650778.
150. Pearl, J. Causality: Models, Reasoning, and Inference / J. Pearl. – 2nd ed. – Cambridge : Cambridge University Press, 2009.

151. Carloni, G. The role of causality in explainable artificial intelligence / G. Carloni, A. Berti, S. Colantonio // WIREs Data Mining and Knowledge Discovery. – 2025. – Vol. 15, no. 2. – Art. e70015. – DOI 10.1002/widm.70015.
152. Trofimov, Y. Causality-Guided Admission Control for Diffusion-Generated Synthetic Medical Image / Y. Trofimov, A. Shevchenko, I. Trusov // 2026 International Russian Smart Industry Conference (SmartIndustryCon). – Sochi, 2026. – P. 1093–1098. – DOI 10.1109/SmartIndustryCon68821.2026.11492854.
153. Кини, Р. Л. Принятие решений при многих критериях: предпочтения и замещения / Р. Л. Кини, Х. Райфа ; пер. с англ. – Москва : Радио и связь, 1981. – 560 с.
154. Методы определения коэффициентов важности критериев / А. М. Анохин, В. А. Глотов, В. В. Павельев, А. М. Черкашин // Автоматика и телемеханика. – 1997. – № 8. – С. 3–35.
155. Petsiuk, V. RISE: Randomized input sampling for explanation of black-box models / V. Petsiuk, A. Das, K. Saenko // Proceedings of the British Machine Vision Conference (BMVC 2018). – Newcastle, 2018. – arXiv:1806.07421.
156. Sugeno, M. Theory of fuzzy integrals and its applications : PhD thesis / M. Sugeno. – Tokyo : Tokyo Institute of Technology, 1974.
157. Тарасов, В. Б. Системно-организационный подход в искусственном интеллекте / В. Б. Тарасов // Программные продукты и системы. – 1997. – № 3. – С. 6–13.
158. Тарасов, В. Б. Агенты, многоагентные системы, виртуальные сообщества: стратегическое направление в информатике и искусственном интеллекте / В. Б. Тарасов // Новости искусственного интеллекта. – 1998. – № 2. – С. 5–63.
159. Kirikov, I. A. Measuring protocol consistency in cohesive hybrid intelligent multi-agent systems / I. A. Kirikov, S. V. Listopad // Open Semantic Technologies for Intelligent Systems. OSTIS 2021. – Cham : Springer, 2022. – P. 39–58. – (Communications in Computer and Information Science ; vol. 1625). – DOI 10.1007/978-3-031-15882-7_3.
160. 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. – № 3. – С. 56–78. – DOI 10.36871/2618-9976.2026.03.005.
161. Мишин, С. П. Оптимальные иерархии управления в экономических системах / С. П. Мишин. – Москва : ПМСОФТ, 2004. – 190 с.
162. 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. – № 3-2. – С. 101–119. – DOI 10.36871/26189976.2026.03-2.011.

163. Исследование сигнальных методов выявления состязательных атак на нейронные модели через объяснительную визуализацию / Д. А. Чеколаев, А. В. Шевченко, А. К. Алексеев [и др.] // Вопросы кибербезопасности. – 2025. – № 6 (70). – С. 140–146. – DOI 10.21681/2311-3456-2025-6-140-146. – EDN JSORTT.
164. ГОСТ Р 57100–2016/ISO/IEC/IEEE 42010:2011. Системная и программная инженерия. Описание архитектуры : национальный стандарт Российской Федерации : издание официальное : дата введения 2017-09-01. – Москва : Стандартинформ, 2019. – 36 с.
165. van Heesch, U. A documentation framework for architecture decisions / U. van Heesch, P. Avgeriou, R. Hilliard // Journal of Systems and Software. – 2012. – Vol. 85, no. 4. – P. 795–820. – DOI 10.1016/j.jss.2011.10.017.
166. Architecture definition in complex system design using model theory / C. E. Dickerson, M. K. Wilkinson, E. Hunsicker [et al.] // IEEE Systems Journal. – 2021. – Vol. 15, no. 2. – P. 1847–1860. – DOI 10.1109/JSYST.2020.2975073.
167. Dimitriadis, T. Stable reliability diagrams for probabilistic classifiers / T. Dimitriadis, T. Gneiting, A. I. Jordan // Proceedings of the National Academy of Sciences of the United States of America. – 2021. – Vol. 118, no. 8. – Art. e2016191118. – DOI 10.1073/pnas.2016191118.
168. Balanya, S. A. Adaptive temperature scaling for robust calibration of deep neural networks / S. A. Balanya, J. Maroñas, D. Ramos // Neural Computing and Applications. – 2024. – Vol. 36, no. 14. – P. 8073–8095. – DOI 10.1007/s00521-024-09505-4.
169. Kruchten, P. Architectural blueprints – the “4+1” view model of software architecture / P. Kruchten // IEEE Software. – 1995. – Vol. 12, no. 6. – P. 42–50. – DOI 10.1109/52.469759.
170. SmoothGrad: removing noise by adding noise / D. Smilkov, N. Thorat, B. Kim [et al.] // ICML Workshop on Visualization for Deep Learning. – Sydney, 2017. – arXiv:1706.03825.
171. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks / A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian // 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). – Lake Tahoe, 2018. – P. 839–847. – DOI 10.1109/WACV.2018.00097.
172. Score-CAM: Score-weighted visual explanations for convolutional neural networks / H. Wang, Z. Wang, M. Du [et al.] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – 2020. – P. 111–119. – DOI 10.1109/CVPRW50498.2020.00020.
173. LayerCAM: Exploring hierarchical class activation maps for localization / P.-T. Jiang, C.-B. Zhang, Q. Hou [et al.] // IEEE Transactions on Image Processing. – 2021. – Vol. 30. – P. 5875–5888. – DOI 10.1109/TIP.2021.3089943.

174. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV) / B. Kim, M. Wattenberg, J. Gilmer [et al.] // Proceedings of the 35th International Conference on Machine Learning (ICML 2018). – 2018. – P. 2668–2677. – (Proceedings of Machine Learning Research ; vol. 80).
175. Trofimov, Yu. V. HYBRID-XIRIS: Neuro-fuzzy architecture for explanatory biometric identification by iris / Yu. V. Trofimov, I. A. Matveev, A. N. Averkin // Soft Measurements and Computing. – 2025. – Vol. 12, no. 2. – Art. 013.
176. 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.
177. Trofimov, Yu. V. A fuzzy transformer for multimodal AI: Differential fuzzy attention and adaptive explanations / Yu. V. Trofimov, A. N. Averkin, A. S. Ilin // Soft Measurements and Computing. – 2025. – Vol. 11. – Art. 003.
178. 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.
179. 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.
180. Image segmentation algorithm based on anisotropic fuzzy local convolution / Y. Trofimov, A. Averkin, M. Lebedev [et al.] // 2026 IEEE 27th International Conference of Young Professionals in Electron Devices and Materials (EDM). – Novosibirsk, 2026. – P. 1–5. – DOI 10.1109/EDM69524.2026.11632120.
181. NSGA-Net: neural architecture search using multi-objective genetic algorithm / Z. Lu, I. Whalen, V. Boddeti [et al.] // Proceedings of the Genetic and Evolutionary Computation Conference (GECCO ’19). – New York : ACM, 2019. – P. 419–427. – DOI 10.1145/3321707.3321729.
182. A method for generating synthetic cellular data using rules, variational modeling, and explainable artificial intelligence / T. Kim, A. Averkin, A. Alexeev [et al.] // 2026 IEEE 27th International Conference of Young Professionals in Electron Devices and Materials (EDM). – Novosibirsk, 2026. – P. 1–4. – DOI 10.1109/EDM69524.2026.11632177.

183. 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.
184. Joseph, J. Comparison of diagnostic accuracy of digital mammography, digital breast tomosynthesis and contrast enhanced mammography in the preoperative assessment of BI-RADS III and above lesions in dense breasts: A cross-sectional study / J. Joseph, K. G. Ramakrishnan, R. Narayanan // Journal of Clinical and Diagnostic Research. – 2025. – Vol. 19, no. 3. – P. TC19–TC24. – DOI 10.7860/JCDR/2025/75298.20802.
185. Согласованность между рентгенологами города Москвы при интерпретации маммографических исследований по шкале BI-RADS / А. С. Азарян, Л. Д. Пестренин, Ю. А. Васильев [и др.] // Альманах клинической медицины. – 2024. – Т. 52, № 7. – С. 377–384. – DOI 10.18786/2072-0505-2024-52-035.
186. Comparetto, C. Cutting-edge imaging breakthroughs for early breast cancer detection / C. Comparetto, F. Borruto // Cancer Screening and Prevention. – 2025. – Vol. 4, no. 1. – P. 21–39. – DOI 10.14218/CSP.2024.00032.
187. Результативность профилактической маммографии в Российской Федерации: сравнение итогов первого этапа диспансеризации в 2019 и 2022 гг. / Ю. А. Васильев, И. М. Сон, Л. Д. Пестренин [и др.] // Менеджер здравоохранения. – 2024. – № 11. – С. 63–76. – DOI 10.21045/1811-0185-2024-11-63-76.
188. Влияние технологий искусственного интеллекта на длительность проведения двойного чтения маммографических исследований / Н. Д. Кудрявцев, Д. Д. Кожихина, И. В. Гончарова [и др.] // Профилактическая медицина. – 2024. – Т. 27, № 5. – С. 32–37. – DOI 10.17116/profmed20242705132.
189. Prakash, D. Mammography: Physics, principle of operation, quality control, and safety / D. Prakash, R. P. Kotian // Fundamentals of X-ray Imaging. – Singapore : Springer, 2025. – P. 239–277. – DOI 10.1007/978-981-96-7328-5_4.
190. Digital Mammography / ed. by E. D. Pisano, M. J. Yaffe, C. M. Kuzmiak. – Philadelphia : Lippincott Williams & Wilkins, 2004. – 231 p.
191. Poulos, A. The application of breast compression in mammography: a new perspective / A. Poulos, D. McLean // Radiography. – 2004. – Vol. 10, no. 2. – P. 131–137. – DOI 10.1016/j.radi.2004.02.012.
192. Методические рекомендации по использованию международной системы BI-RADS при маммографическом обследовании / О. О. Мануйлова, Т. В. Павлова, В. В. Диденко [и

др.]. – Москва : ГБУЗ «НПКЦ ДиТ ДЗМ», 2025. – 32 с. – (Лучшие практики лучевой и инструментальной диагностики ; вып. 152).

193. Калининская, А. А. Передвижные медицинские комплексы в условиях села / А. А. Калининская, Н. А. Баянова, Ф. А. Сулькина // Современные проблемы здравоохранения и медицинской статистики. – 2019. – № 1. – DOI 10.24411/2312-2935-2019-10009.
194. Огрызко, Е. В. Динамика оснащенности некоторыми передвижными подразделениями по Российской Федерации в целом и федеральным округам за 2019–2023 годы / Е. В. Огрызко, А. В. Поликарпов, С. В. Моравская // Современные проблемы здравоохранения и медицинской статистики. – 2025. – № 2. – С. 811–825. – DOI 10.24412/2312-2935-2025-2-811-825.
195. Корженкова, Г. П. Опыт работы мобильных маммографических комплексов в России / Г. П. Корженкова, Н. Н. Блинов – Текст : электронный // Опухоли женской репродуктивной системы. – 2005. – № 1. – URL: https://cyberleninka.ru/article/n/opyt-raboty-mobilnyh-mammograficheskih-kompleksov-v-rossii (дата обращения: 14.01.2026).
196. Передвижной медицинский комплекс на базе ПАЗ 4234. – Текст : электронный // Луидор : официальный сайт. – URL: https://luidorbus.ru/product/meditsinskie-avtomobili/peredvizhnoy-meditsinskiy-kompleks-paz-4234/ (дата обращения: 10.12.2025).
197. ProFound AI Breast Cancer Health Suite. – Текст : электронный // iCAD : официальный сайт. – URL: https://www.icadmed.com/breast-health/ (дата обращения: 26.12.2025).
198. Genius AI Detection: Mammography AI Software. – Текст : электронный // Hologic : официальный сайт. – URL: https://www.hologic.com/hologic-products/breast-health-solutions/genius-ai-detection-technology (дата обращения: 26.12.2025).
199. AI Mammography Software for Early Detection (Lunit INSIGHT MMG). – Текст : электронный // Lunit : официальный сайт. – URL: https://www.lunit.io/en/ai-radiology-software/insight-mmg/ (дата обращения: 26.12.2025).
200. СберМедИИ — искусственный интеллект для здравоохранения. – Текст : электронный // SberMedAI : официальный сайт. – URL: https://sbermed.ai/ (дата обращения: 26.12.2025).
201. Transformer-based explainable deep learning for breast cancer detection in mammography: The MammoFormer framework / O. O. Ejiga Peter, D. Emakporuena, B. D. Tunde [et al.] // American Journal of Computer Science and Technology. – 2025. – Vol. 8, no. 2. – P. 121–137. – DOI 10.11648/j.ajcst.20250802.16.
202. Mammo-CLIP: A vision language foundation model to enhance data efficiency and robustness in mammography / S. Ghosh, C. B. Poynton, S. Visweswaran, K. Batmanghelich // Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. – Cham : Springer,

2024. – P. 632–642. – (Lecture Notes in Computer Science ; vol. 15012). – DOI 10.1007/978-3-031-72390-2_59.

203. Label Studio: Open Source Data Labeling. – Текст : электронный // Label Studio : официальный сайт. – URL: https://labelstud.io/ (дата обращения: 14.02.2026).
204. Varghese, R. YOLOv8: A novel object detection algorithm with enhanced performance and robustness / R. Varghese, Sambath M. // 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). – Chennai, 2024. – P. 1–6. – DOI 10.1109/ADICS58448.2024.10533619.
205. Localized SHAP Explainability for Object Detection Based on Parallel Attribution Analysis / Y. V. Trofimov, A. N. Averkin, A. V. Shevchenko [et al.] // 2026 XXIX International Conference on Soft Computing and Measurements (SCM). – Saint Petersburg, 2026. – P. 1–3. – DOI 10.1109/SCM71573.2026.11625907.
206. Measures of acutance and shape for classification of breast tumors / R. M. Rangayyan, N. M. El-Faramawy, J. E. L. Desautels, O. A. Alim // IEEE Transactions on Medical Imaging. – 1997. – Vol. 16, no. 6. – P. 799–810. – DOI 10.1109/42.650876.
207. Assessment of risk factors for coronary artery disease and severity by coronary computed tomography angiography imaging / H. Wu, G. Yang, S. Zhang [et al.] // International Journal of General Medicine. – 2024. – Vol. 17. – P. 4607–4615. – DOI 10.2147/IJGM.S475008.
208. Herrmann, J. Coronary artery disease: development and progression / J. Herrmann, A. Lerman // Coronary Artery Disease / ed. by G. W. Barsness, D. R. Holmes. – London : Springer, 2012. – DOI 10.1007/978-1-84628-712-1_3.
209. Real-time coronary artery stenosis detection based on modern neural networks / V. V. Danilov, K. Y. Klyshnikov, O. M. Gerget [et al.] // Scientific Reports. – 2021. – Vol. 11. – Art. 7582. – DOI 10.1038/s41598-021-87174-2.
210. Contemporary management of stable coronary artery disease / D. T. Bertolone, E. Gallinoro, G. Esposito [et al.] // High Blood Pressure & Cardiovascular Prevention. – 2022. – Vol. 29. – P. 207–219. – DOI 10.1007/s40292-021-00497-z.
211. Herrmann, S. C. Current concepts in coronary physiology for the interventionalist / S. C. Herrmann, A. El-Shafei, M. J. Kern // International Journal of Cardiovascular Interventions. – 2003. – Vol. 5, no. 3. – P. 109–131. – DOI 10.1080/14628840310017357.
212. Sun, Z. Coronary CT angiography: current status and continuing challenges / Z. Sun, G. H. Choo, K. H. Ng // British Journal of Radiology. – 2012. – Vol. 85, no. 1013. – P. 495–510. – DOI 10.1259/bjr/15296170.
213. Deep learning applications in medical image analysis / J. Ker, L. Wang, J. Rao, T. Lim // IEEE Access. – 2018. – Vol. 6. – P. 9375–9389. – DOI 10.1109/ACCESS.2017.2788044.

214. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis / X. Liu, L. Faes, A. U. Kale [et al.] // The Lancet Digital Health. – 2019. – Vol. 1, no. 6. – P. e271–e297. – DOI 10.1016/S2589-7500(19)30123-2.
215. CADICA: a new dataset for coronary artery disease detection by using invasive coronary angiography / A. Jiménez-Partinen, M. A. Molina-Cabello, K. Thurnhofer-Hemsi [et al.] // Expert Systems. – 2024. – Vol. 41, no. 12. – Art. e13708. – DOI 10.1111/exsy.13708.
216. A survey on deep learning in medical image analysis / G. Litjens, T. Kooi, B. E. Bejnordi [et al.] // Medical Image Analysis. – 2017. – Vol. 42. – P. 60–88. – DOI 10.1016/j.media.2017.07.005.
217. Averkin, A. Hybrid neural networks for time series forecasting / A. Averkin, S. Yarushev // Artificial Intelligence. RCAI 2018 / ed. by S. O. Kuznetsov, G. S. Osipov, V. L. Stefanuk. – Cham : Springer, 2018. – P. 230–239. – (Communications in Computer and Information Science ; vol. 934). – DOI 10.1007/978-3-030-00617-4_21.
218. Hulsen, T. Explainable artificial intelligence (XAI): concepts and challenges in healthcare / T. Hulsen // AI. – 2023. – Vol. 4, no. 3. – P. 652–666. – DOI 10.3390/ai4030034.
219. HeartFlow FFRCT Analysis. – Текст : электронный // HeartFlow : официальный сайт. – URL: https://www.heartflow.com/ (дата обращения: 01.10.2026).
220. syngo.via – advanced visualization platform for medical imaging. – Текст : электронный // Siemens Healthineers : официальный сайт. – URL: https://www.siemens-healthineers.com (дата обращения: 14.03.2026).
221. IntelliSpace Cardiovascular – integrated cardiovascular imaging and information platform. – Текст : электронный // Philips Healthcare : официальный сайт. – URL: https://www.philips.com (дата обращения: 14.03.2026).
222. Advantage Workstation (AW) Cardiac – advanced cardiac imaging analysis platform. – Текст : электронный // GE HealthCare : официальный сайт. – URL: https://www.gehealthcare.com (дата обращения: 14.03.2026).
223. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). – 2016. – P. 770–778. – DOI 10.1109/CVPR.2016.90.
224. Ronneberger, O. U-Net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. – Cham : Springer, 2015. – P. 234–241. – (Lecture Notes in Computer Science ; vol. 9351). – DOI 10.1007/978-3-319-24574-4_28.

225. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs / L.-C. Chen, G. Papandreou, I. Kokkinos [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2018. – Vol. 40, no. 4. – P. 834–848. – DOI 10.1109/TPAMI.2017.2699184.
226. Krähenbühl, P. Efficient inference in fully connected CRFs with Gaussian edge potentials / P. Krähenbühl, V. Koltun // Advances in Neural Information Processing Systems. – 2011. – Vol. 24. – P. 109–117.
227. Аверкин, А. Н. Извлечение правил из классических, глубоких нейронных сетей и нейро-нечетких сетей / А. Н. Аверкин – Текст : электронный // Речевые технологии. – 2020. – № 3–4. – URL: https://cyberleninka.ru/article/n/izvlechenie-pravil-iz-klassicheskih-glubokih-neyronnyh-setey-i-neyro-nechetkih-setey (дата обращения: 14.03.2026).
228. 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 г.). – Санкт-Петербург : Университет ИТМО, 2025. – С. 938–939. – EDN TLYRHK.
229. 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.
230. Внебольничная пневмония у взрослых : клинические рекомендации / Российское респираторное общество [и др.]. – 2024. – Текст : электронный // Рубрикатор клинических рекомендаций Министерства здравоохранения Российской Федерации. – URL: https://cr.minzdrav.gov.ru (дата обращения: 01.10.2026).
231. Ellis, S. M. The WHO manual of diagnostic imaging: radiographic anatomy and interpretation of the chest and the pulmonary system / S. M. Ellis, C. D. R. Flower. – Geneva : World Health Organization, 2006. – ISBN 978-92-4-154677-5.
232. Interobserver reliability of the chest radiograph in community-acquired pneumonia / M. N. Albaum, L. C. Hill, M. Murphy [et al.] // Chest. – 1996. – Vol. 110, no. 2. – P. 343–350. – DOI 10.1378/chest.110.2.343.
233. Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections / R. M. Hopstaken, T. Witbraad, J. M. A. van Engelshoven, G.-J. Dinant // Clinical Radiology. – 2004. – Vol. 59, no. 8. – P. 743–752. – DOI 10.1016/j.crad.2004.01.011.

234. Lunit INSIGHT CXR: one AI solution for all your chest X-ray needs. – Текст : электронный // Lunit : официальный сайт. – URL: https://www.lunit.io/en/ai-radiology-software/insight-cxr/ (дата обращения: 16.03.2026).
235. qXR: AI enabled comprehensive chest X-ray reporting tool. – Текст : электронный // Qure.ai : официальный сайт. – URL: https://www.qure.ai/product/qxr (дата обращения: 16.03.2026).
236. Harrison.ai CXR — AI chest X-ray solution. – Текст : электронный // Harrison.ai : официальный сайт. – URL: https://harrison.ai/cxr/ (дата обращения: 16.03.2026).
237. AI-Rad Companion Chest CT. – Текст : электронный // Siemens Healthineers : официальный сайт. – URL: https://www.siemens-healthineers.com/en-th/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion/chest-ct (дата обращения: 16.03.2026).
238. AVIEW COPD: AI for quantitative COPD analysis. – Текст : электронный // Coreline Soft : официальный сайт. – URL: https://www.corelinesoft.com/en/solutions/copd (дата обращения: 16.03.2026).
239. Identifying medical diagnoses and treatable diseases by image-based deep learning / D. S. Kermany, M. Goldbaum, W. Cai [et al.] // Cell. – 2018. – Vol. 172, no. 5. – P. 1122–1131.e9. – DOI 10.1016/j.cell.2018.02.010.
240. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases / X. Wang, Y. Peng, L. Lu [et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). – 2017. – P. 2097–2106. – DOI 10.1109/CVPR.2017.369.
241. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison / J. Irvin, P. Rajpurkar, M. Ko [et al.] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, no. 1. – P. 590–597. – DOI 10.1609/aaai.v33i01.3301590.
242. 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.
243. 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 г.). – Санкт-Петербург : Университет ИТМО, 2025. – С. 815–817. – EDN LPKXXE.

ПРИЛОЖЕНИЕ А

Таблица А.1 – Категории оценки по системе BI-RADS

Категория BI-RADS	**Тип исследований**	**Характеристики**	**Рекомендации**	**Вероятность РМЖ**
BI-RADS 0	Профилактическое обследование	Невозможно прийти к однозначному выводу по результатам визуализации (маммография плохого качества, неправильная укладка, недостаточно проекций). Необходимы дополнительные изображения или данные предыдущего обследования.	При невозможности интерпретации выполненного исследования необходимо проведение повторного исследования в ЛПУ по месту прикрепления или выполнение дополнительных методов визуализации.	Не определена
BI-RADS 1	Профилактическое обследование	Изменений в молочных железах не выявлено (возрастная норма). Допускаются: неизменные аксиллярные лимфатические узлы	Плановая маммография согласно плану обследования в ЛПУ по месту прикрепления. При выявлении «плотного фона» (тип плотности по ACR – C, D (см. таблицу 4.1)) обязательно выполнение УЗИ в плановом порядке в ЛПУ по месту прикрепления.	0 %
BI-RADS 2	Профилактическое обследование	– Доброкачественные образования (простые кисты, липомы/фибролипомы, фиброаденомы без отрицательных изменений в динамике); – доброкачественные обызвествления (рас-	Плановая маммография согласно плану обследования в ЛПУ по месту прикрепления. При выявлении «плотного фона» (тип плотности по	0 %

Категория BI-RADS	Тип исследований	Характеристики	Рекомендации	Вероятность РМЖ
		пределение диффузное, региональное либо единичное); – интрамаммарные лимфатические узлы; – галактоцеле; – гамартомы; имплантаты молочных желез; – нарушения архитектоники (рубцовые изменения); – отек после лучевой терапии.	ACR – C, D) обязательно выполнение УЗИ в плановом порядке в ГП по месту прикрепления.	
BI-RADS 3	Диагностическое обследование	– Впервые выявленные вероятно доброкачественные изменения после проведенного УЗИ: фиброаденома, киста, или нет возможности сравнить с данными предыдущих исследований; – очаговая асимметрия ткани; – единичная группа точечных микрокальцинатов; – мастит; – втяжение соска без видимого объемного образования.	Контроль через 6 месяцев. Не исключается проведение биопсии и дальнейшего динамического контроля.	> 0, но < 2 %
BI-RADS 4	Диагностическое обследование. Выставляется по результатам комплексного клинико-лучевого обследования.	– Изменения, подозрительные на злокачественный процесс; – односторонняя подозрительная лимфаденопатия. 4а – низкая степень вероятности наличия злокачественного процесса: – объемное образование с частично размы-	Трепан-биопсия с последующей морфологической верификацией.	4a: > 2 % ≤ 10 % 4b: >10 % ≤ 50 % 4c: > 50 % < 95 %

Категория BI-RADS	Тип исследований	Характеристики	Рекомендации	Вероятность РМЖ
		тым контуром (атипичная фиброаденома); – пальпируемые одиночные и сложные кисты; – локальная асимметрия; – нарушение архитектоники; – абсцесс. 4b – промежуточная степень вероятности наличия злокачественного процесса: – сгруппированные аморфные или мелкие плеоморфные микрокальцинаты; – образование с нечёткими краями; – увеличение размеров ранее выявленного образования; – дефекты наполнения при дуктографии; – л/у с подозрением на метастатическое поражение. 4c – высокая степень вероятности наличия злокачественного процесса: – скопление плеоморфных микрокальцинатов, увеличение их количества в динамике; – узловое образование, имеющее признаки рака.		
BI-RADS 5	Диагностическое обследование. Выставляется по результатам комплексного кли-	Достоверно злокачественные изменения: – объемное образование высокой плотности с неровными лучистыми контурами;	Трепан-биопсия с последующей морфологической верификацией.	≥ 95%

Категория BI-RADS	**Тип исследований**	**Характеристики**	**Рекомендации**	**Вероятность РМЖ**
	нико-лучевого обследования.	– сегментарное или линейное распределение мелких линейных и плеоморфных кальцинатов; объемное образование с неровным лучистым контуром в сочетании с плеоморфными кальцинатами.		
BI-RADS 6		Верифицированный неоперированный рак молочной железы.	Лечение в МО 3-го уровня.	Доказана

ПРИЛОЖЕНИЕ Б

Таблица Б.1 – Расширенный набор методов ОИИ

Метод	Тип объяснения	Уровень объяснительного контура	Роль в интерпретации	Основные преимущества	Ограничения
Grad-CAM	визуальное, активационное	модельный	локализация областей изображения, повлиявших на решение CNN	нагляден, хорошо применим к CNN, удобен для медицинских изображений	ограниченная пространственная точность, зависит от архитектуры
Grad-CAM++	визуальное, активационное	модельный	уточнённая локализация значимых областей при множественных очагах	лучше работает при нескольких активных зонах	вычислительно сложнее базового Grad-CAM
Score-CAM	визуальное, активационное	модельный	построение карт значимости без прямого использования градиентов	меньше шумит, устойчивее к нестабильным градиентам	более высокая вычислительная стоимость
LayerCAM	визуальное, многослойное	модельный	выделение значимых областей на разных уровнях глубины сети	учитывает информацию нескольких слоёв	сложнее интерпретировать и настраивать
EigenCAM	визуальное, активационное	модельный	получение компактной карты наиболее информативных зон	не требует расчёта градиентов, быстро работает	слабее связан с целевым классом
HiResCAM	визуальное, активационное	модельный	более точная пространственная локализация релевантных областей	повышенная детализация тепловой карты	зависит от структуры модели
Occlusion Sensitivity	визуальное, perturbation-based	признаковый / модельный	оценка вклада участков изображения путём	интуитивно понятен, хорошо согласуется с	очень медленный на больших изображениях

Метод	Тип объяснения	Уровень объяснительного контура	Роль в интерпретации	Основные преимущества	Ограничения
			маскирования фрагментов	идеей причинного вклада	
RISE	визуальное, perturbation-based	признаковый / модельный	построение карт значимости на основе случайных масок	модельно-независим, применим к чёрным ящикам	высокая вычислительная стоимость, стохастичность
LIME	локальное, суррогатное	признаковый	построение локального интерпретируемого приближения решения модели	модельно-независим, даёт локально понятное объяснение	чувствителен к параметрам локальной аппроксимации
SHAP	атрибутивное, на основе значений Шепли	признаковый	оценка вклада признаков/суперпикселей в итоговое решение	строгая теоретическая основа, сопоставимость вкладов	ресурсоёмок, требует аккуратного выбора фонового распределения
KernelSHAP	атрибутивное, модельно-независимое	признаковый	универсальная оценка вкладов признаков для произвольных моделей	применим к широкому классу моделей	медленный, чувствителен к выбору выборки фона
TreeSHAP	атрибутивное, специализированное	признаковый	объяснение деревообразных моделей и ансамблей	точный и быстрый для tree-based моделей	неприменим к CNN напрямую
Integrated Gradients	атрибутивное, градиентное	признаковый	оценка вклада входных компонент относительно базовой точки	аксиоматически корректен, хорошо работает для глубоких сетей	зависит от выбора baseline
DeepLIFT	атрибутивное, градиентоподобное	признаковый / модельный	оценка вклада входов по сравнению с референсным состоянием	часто стабильнее обычных градиентов	зависит от выбора опорного состояния

Метод	Тип объяснения	Уровень объяснительного контура	Роль в интерпретации	Основные преимущества	Ограничения
SmoothGrad	атрибутивное, градиентное	признаковый	подавление шума в градиентных картах путём усреднения по зашумлённым входам	улучшает визуальное качество карты	не является самостоятельным объяснителем, а усиливает базовый метод
Guided Backpropagation	визуально-градиентное	модельный	выделение тонких локальных деталей, влияющих на активацию	высокая детализация визуального результата	низкая устойчивость как самостоятельного объяснения
CAM	визуальное, активационное	модельный	базовая локализация значимых областей для совместимых CNN-архитектур	простота и историческая интерпретируемость	требует архитектурных ограничений
Counterfactual Explanations	контрфактуальное	результирующий / интерпретируемый	формирование альтернативного сценария: что нужно изменить для смены решения	полезно для экспертного анализа и СППР	сложно строить для изображений без артефактов
Prototype-based Explanations	прототипное	интерпретируемый	сопоставление случая с похожими примерами из обучающей выборки	хорошо воспринимается врачом как аналогия	требует качественного репозитория прототипов
Concept Activation Vectors (TCAV)	концептуальное	интерпретируемый	оценка вклада семантических концептов в решение модели	переход от пикселей к содержательным понятиям	требует заранее заданных концептов

ПРИЛОЖЕНИЕ В

Свидетельства о государственной регистрации программ для ЭВМ



РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025690674

Интерактивная демонстрационная платформа для изучения принципов объяснимого ИИ через моделирование состязательных атак

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Университет "Дубна" (RU)***

Авторы: ***Кузнецов Егор Михайлович (RU), Муравьев Иван Павлович (RU), Чеколаев Дмитрий Алексеевич (RU), Трофимов Юрий Владиславович (RU), Шевченко Алексей Валерьевич (RU), Аверкин Алексей Николаевич (RU)***

Заявка № **2025689137**
Дата поступления **25 октября 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***10 ноября 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 00a570e4f7add8d531b4b8818e75f29506
Владелец **Зубов Юрий Сергеевич**
Действителен с 04.09.2025 по 28.11.2026

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025690301

Веб-приложение для интерпретируемой классификации пневмонии лёгких (версия 1.0)

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Университет "Дубна" (RU)***

Авторы: ***Кузнецов Егор Михайлович (RU), Муравьев Иван Павлович (RU), Трофимов Юрий Владиславович (RU), Шевченко Алексей Валерьевич (RU), Аверкин Алексей Николаевич (RU)***

Заявка № **2025689088**
Дата поступления **25 октября 2025 г.**
Дата государственной регистрации
в Реестре программ для ЭВМ ***06 ноября 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 00a570e4f7add8d531b4b8818e75f29506
Владелец **Зубов Юрий Сергеевич**
Действителен с 04.09.2025 по 28.11.2026

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2025690503

Веб-приложение для мультиуровневого и интерпретируемого анализа ХОБЛ (версия 1.0)

Правообладатель: ***Федеральное государственное бюджетное образовательное учреждение высшего образования "Университет "Дубна" (RU)***

Авторы: ***Кузнецов Егор Михайлович (RU), Муравьев Иван Павлович (RU), Трофимов Юрий Владиславович (RU), Шевченко Алексей Валерьевич (RU), Аверкин Алексей Николаевич (RU)***

Заявка № **2025689067**

Дата поступления **25 октября 2025 г.**

Дата государственной регистрации в Реестре программ для ЭВМ ***07 ноября 2025 г.***

Руководитель Федеральной службы по интеллектуальной собственности

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ
Сертификат 00a570e4f7add8d531b4b8818e75f29506
Владелец **Зубов Юрий Сергеевич**
Действителен с 04.09.2025 по 28.11.2026

Ю.С. Зубов