УДК 004.8+004.896

ЯДРО ВЕРИФИЦИРУЕМОЙ ОБЪЯСНИМОСТИ: ГИБРИДНАЯ АРХИТЕКТУРА GD-ANFIS/SHAP ДЛЯ XAI 2.0

Ю. В. Трофимов[1] [0009-0005-6943-7432], А. Д. Лебедев[2] [0009-0001-1046-5982], А. С. Ильин[3] [0009-0007-9599-4958], А. Н. Аверкин[4] [0000-0003-1571-3583]

[1, 2, 4]Государственный университет «Дубна», г. Дубна, Россия

[3]Университет Иннополис, г. Иннополис, Россия

[1, 3]Объединенный институт ядерных исследований, г. Дубна, Россия

[4]Вычислительный центр им. А. А. Дородницына РАН, г. Москва, Россия

[1]ura_trofim@bk.ru, [2]lebedev0lexander@gmail.com, [3]a.ilin@innopolis.university, [4]averkin2003@inbox.ru

Аннотация

Предложена гибридная архитектура Explainable AI, совмещающая полностью дифференцируемую нейро-нечеткую модель GD-ANFIS и пост-хок метод SHAP. Интеграция выполнена с целью реализации принципов XAI 2.0, требующих одновременной прозрачности, проверяемости и адаптивности объяснений.

GD-ANFIS формирует человеческо-читаемые правила типа Такаги – Сугено, обеспечивая структурную интерпретируемость, тогда как SHAP вычисляет количественные вклады признаков по теории Шепли. Для объединения этих слоев разработан механизм компаративного аудита: он автоматически сопоставляет наборы ключевых признаков, проверяет совпадение направлений их влияния и анализирует согласованность между числовыми оценками SHAP и лингвистическими правилами GD-ANFIS. Такой двухконтурный контроль повышает доверие к выводам модели и позволяет оперативно выявлять потенциальные расхождения.

Эффективность подхода подтверждена экспериментами на четырех разнородных наборах данных. В медицинской задаче классификации Breast Cancer Wisconsin достигнута точность 0.982; в задаче глобального картирования просадок грунта — 0.89. В регрессионных тестах на Boston Housing и мониторинге качества поверхностных вод получены RMSE 2.30 и 2.36 соответственно при полном сохранении интерпретируемости. Во всех случаях пересечение топ-признаков

в объяснениях двух методов составляло не менее 60%, что демонстрирует высокую согласованность структурных и числовых трактовок.

Предложенная архитектура формирует практическую основу для ответственного внедрения XAI 2.0 в критически важных областях — от медицины и экологии до геоинформационных систем и финансового сектора.

***Ключевые слова**: объяснимый искусственный интеллект, XAI 2.0, ANFIS, SHAP, компаративный анализ, интерпретируемость, пространственный анализ, доверенность.*

ВВЕДЕНИЕ

Несмотря на впечатляющую точность современных моделей машинного обучения, для конечного пользователя они зачастую остаются «черными ящиками», лишенными ясных и проверяемых объяснений. Это ограничивает внедрение интеллектуальных систем в ответственные области, где необходимы прозрачность и воспроизводимость выводов [1].

Существующие подходы к интерпретируемости можно разделить на:

- модели изначально прозрачные (например, деревья решений, линейная регрессия) [2],
- пост-хок методы для сложных моделей (например, LIME, SHAP), которые демонстрируют определенную эффективность, но страдают от неоднозначности интерпретаций и ограниченной устойчивости [3–6].

В качестве «нейронного ядра» предлагаемой системы выступает адаптивная нейро-нечеткая система вывода ANFIS, способная обучаться на данных и одновременно формировать человеко-ориентированные правила нечеткой логики [7, 8]. Чтобы количественно оценить вклад каждого признака и тем самым повысить доверие к полученным решениям, ANFIS дополняется пост-хок-методом SHAP, основанным на значениях Шепли [3].

Ключевое отличие нашего подхода заключается во внедрении механизма кросс-валидации объяснений: структурные правила, выведенные ANFIS, сверяются с численными оценками SHAP в едином протоколе компаративного анализа. Такая сверка позволяет выявлять расхождения, подтверждать согласованность

выводов и, при необходимости, автоматически сигнализировать о потенциальных источниках ошибок или смещений. В результате достигается двойная — структурная и количественная — проверяемость модели, что выводит решение на уровень XAI 2.0 и открывает возможности для полноценного аудита принимаемых решений.

1. МЕТОДОЛОГИЯ

Парадигма XAI 2.0 выводит объяснимый ИИ от локальных пост-хок методов к сквозной, контекстно-адаптивной прозрачности на всех стадиях жизненного цикла модели [1, 9]. В предлагаемой методологии это выражается следующим образом. Во-первых, каждое решение сопровождается многоуровневым пояснением: логическая структура выводится в виде правил, численный вклад признаков дается через метрики, а итог представляется пользователю в визуальной или естественно-языковой форме. Во-вторых, символические и числовые объяснения проверяются между собой, что обеспечивает согласованность и воспроизводимость выводов. Третьим фундаментальным требованием служит формализованная инфраструктура аудита; все метрики, версии данных и параметры модели фиксируются, позволяя оперативно оценивать как качество, так и этичность решений. Наконец, система динамически подстраивает объем и форму объяснения под задачи эксперта, инженера или конечного пользователя, не затрагивая предсказательное ядро. Суммарно эти четыре положения задают рамки для выбора архитектурных компонентов и определяют роль каждого модуля в конвейере.

1.1. Энкодер (сжимающий путь)

Адаптивная нейро-нечеткая система вывода (ANFIS) представляет собой гибридную архитектуру, объединяющую принципы нечеткой логики Такаги — Сугено — Канга [10] с адаптивными возможностями нейронных сетей. Архитектура ANFIS состоит из пяти функциональных слоев, каждый из которых выполняет специфические вычислительные операции.

Слой 1 (Фаззификация). Первый слой выполняет преобразование входных переменных в нечеткие множества с использованием функций принадлежности. Для гауссовой функции принадлежности выходной сигнал i-го узла определяется как

$$O_1^i = \mu_{A_i}(x) = exp\left(-\frac{(x-c_i)^2}{2\sigma_i^2}\right),$$

где c_i и σ_i — параметры центра и ширины гауссовой функции принадлежности соответственно.

Слой 2 (Правила). Второй слой вычисляет силу активации каждого нечеткого правила путем применения *T*-нормы (обычно произведения) к выходам функций принадлежности:

$$O_2^i = w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1,2,\dots,n,$$

где w_i представляет силу активации i-го правила.

Слой 3 (Нормализация). Третий слой выполняет нормализацию сил активации правил:

$$O_3^i = \overline{w}_i = \frac{w_i}{\sum_{j=1}^{n} w_j}.$$

Слой 4 (Дефаззификация). Четвертый слой вычисляет взвешенные следствия правил согласно модели Такаги — Сугено:

$$O_4^i = \overline{w}_i f_i = \overline{w}_i(p_i x + q_i y + r_i),$$

где p_i, q_i, r_i — параметры следствий i-го правила.

Слой 5 (Суммирование). Пятый слой агрегирует выходы всех правил для получения финального результата:

$$O_5 = \sum_{i=1}^{n} \overline{w}_i f_i = \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} w_i}.$$

Обучение ANFIS осуществляется гибридным алгоритмом, сочетающим градиентный спуск для настройки параметров предпосылок (функций принадлежности) и метод наименьших квадратов для определения параметров следствий.

Использована реализация GD-Anfis из библиотеки X-ANFIS [11] — полностью дифференцируемая версия ANFIS со следующими ключевыми преимуществами:

- градиентное обучение с современными оптимизаторами (Adam, RMSprop);
- модульная PyTorch-архитектура, совместимая с Scikit-Learn;
- встроенная регуляризация и ранняя остановка.

1.2. Математические основы SHAP

Метод SHAP (SHapley Additive exPlanations) основан на теории кооперативных игр и концепции значений Шепли. Для заданной модели f и экземпляра x SHAP-значение для признака i определяется как

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)],$$

где F — множество всех признаков, S — подмножество признаков, не содержащее i, $|S|$ — размер подмножества S, а $|F|$ — общее количество признаков [3].

Данная формула учитывает все возможные подмножества признаков и изменение предсказания при добавлении признака i к каждому подмножеству, взвешенное по размеру подмножеств. SHAP-значения удовлетворяют четырем аксиомам справедливости: эффективности, симметрии, пустоты и аддитивности [12].

Аддитивность: Объяснение представляется в виде линейной модели

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z_j',$$

где ϕ_0 — ожидаемое значение модели, ϕ_j — SHAP-значения для признаков, а z_j' — упрощенные входные данные [2].

Эффективность: Сумма всех SHAP-значений равна разности между предсказанием модели и ожидаемым значением:

$$\sum_{j=1}^{M} \phi_j = f(x) - E[f(X)].$$

1.3. Компаративный анализ объяснений

Система выполняет сравнительный анализ объяснений ANFIS и SHAP для выявления согласованности между подходами. Анализ включает три этапа: SHAP-анализ, извлечение правил ANFIS и совместное сравнение:

$$\underline{\phi_i} = \frac{1}{N}\sum_{j=1}^{N} \phi_i^{(j)},$$

где $\phi_i^{(j)}$ — SHAP-значение признака i на экземпляре j. Направление влияния определяется знаком $\underline{\phi_i}$.

Извлечение правил ANFIS. Система извлекает активные нечеткие правила в форме «если…то» на основе степени активации

$$\alpha_k = \frac{1}{N}\sum_{j=1}^{N} \overline{w}_k^{(j)},$$

где $w_k^{-(j)}$ — активация правила k для экземпляра j. Отбираются правила с $\alpha_k > \theta$.

Совместный анализ. Определяются общие значимые признаки и оценивается согласованность:

$$F_{\text{common}} = F_{SHAP} \cap F_{ANFIS}, \gamma = \frac{|F_{\text{consistent}}|}{|F_{\text{common}}|},$$

где γ — коэффициент согласованности направлений влияния.

Результатом является структурированный отчет с ранжированными признаками, правилами ANFIS, метриками согласованности и анализом противоречий, обеспечивающий комплексную интерпретируемость через структурное понимание (ANFIS) и количественные оценки (SHAP).

1.4. XAI 2.0 в гибридной системе GD-ANFIS-SHAP

Гибрид GD-ANFIS–SHAP реализует четыре ключевых требования XAI 2.0, что отличает систему от классических схем «модель + пост-хок» и устраняет дублирование функций по сравнению с ранее описанными модулями.

1. **Сквозная прослеживаемость.** Все стадии — от выбора признаков до формирования отчета — фиксируются в метаданных; это обеспечивает воспроизводимость результатов и упрощает последующий аудит модели.

2. **Единый контур интерпретации.** Нечеткие правила GD-ANFIS раскрывают логику предсказаний, а SHAP дополняет ее численными аргументами. Вместо последовательного применения методов объяснения используется параллельная связка, где обе трактовки строятся на тех же входных данных и моментально сопоставляются.

3. **Автоматизированная верификация выводов.** Специализированный аудитор не просто сравнивает ранжирование признаков, а анализирует согласованность знаков влияния и минимальную допустимую разницу между весами. При превышении порогов несогласия система формирует уведомление и сохраняет конфликтный пример для последующего анализа.

4. **Адаптивная подача объяснений.** Выходы GD-ANFIS–SHAP масштабируются под роль пользователя:

– инженеру предоставляется полный набор правил и распределения SHAP;

– эксперту-предметнику — укрупненные кластеры факторов;

– конечному пользователю — краткая естественно-языковая справка.

Этот механизм не затрагивает предсказательное ядро и не требует повторного обучения модели.

Таким образом, архитектура не просто сочетает две техники интерпретации, а формирует целостную инфраструктуру, где прозрачность, проверяемость и адаптивность заложены в поток обработки данных, что полностью соответствует современным представлениям о XAI 2.0 [13, 14].

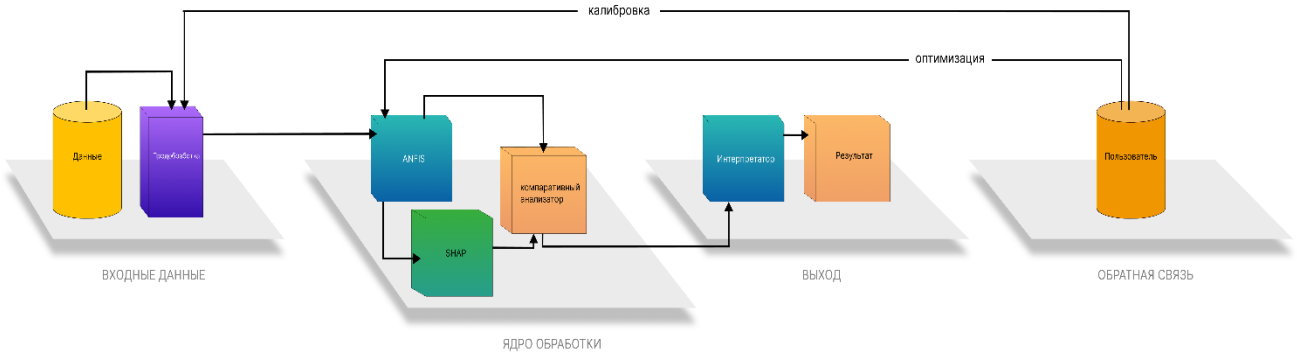


Рис. 1. Схема архитектуры предлагаемой гибридной системы

2. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Для тестирования созданной системы были использованы четыре датасета. Для задач классификации были выбраны медицинский датасет Breast Cancer Wisconsin (Diagnostic) и ГИС датасет Global Land Subsidence Mapping. Для задач регрессии были выбраны экономический датасет Housing Data и ГИС датасет Comprehensive Surface Water Quality Dataset.

SHAP был выбран в качестве основного метода анализа важности признаков, поскольку он обеспечивает теоретическую обоснованность, предоставляет как глобальные, так и локальные объяснения, а также отличается более высокой устойчивостью и воспроизводимостью результатов по сравнению с LIME и аналогичными методами.

2.1. Датасет Breast Cancer Wisconsin (Diagnostic)

В качестве тестовой площадки выбран клинический набор *Breast Cancer Wisconsin (Diagnostic)*. Коллекция содержит $N = 569$ наблюдений и $d = 30$ непрерывных признаков, вычисленных по цифровым изображениям тонкоигольной аспирационной биопсии. Целевая переменная Diagnosis принимает значения $\{M, B\}$, где M — злокачественная, B — доброкачественная опухоль.

Ключевая особенность датасета состоит в том, что классы были умеренно несбалансированы: $M: 212$ против $B: 357$ экземпляров.

Табл. 1. Фрагмент описания признаков датасета WDBC

Признак	Краткое пояснение	Ед. изм.
radius_mean	Средний радиус ядер	pixel
texture_mean	Ст. откл. интенсивности серого	–
perimeter_mean	Средний периметр контура	pixel
area_mean	Средняя площадь	pixel^2
concavity_mean	Глубина вогнутых сегментов контура	–
(еще 25 признаков опущены для краткости)		

Данные разделены в пропорции 80/20 на обучающую и тестовую части с сохранением распределения классов (стратификация). Так как все признаки уже в сопоставимых масштабах, дополнительное масштабирование не потребовалось. Для устранения возможного влияния редких выбросов использовано персентильное обрезание на уровне [0.5, 99.5].

Табл. 2. Конфигурация модели GD-ANFIS

Параметр	Значение	Комментарий
Тип задачи	Классификация	бинарная
# входных признаков	30	см. табл. 1
# правил FIS	12	подобрано по grid–search
MF (тип)	GBell	симметричные колоколообразные функции
Оптимизатор	Adam	$\eta = 0.01$
Эпох	100	с early–stopping (patience = 10)
Batch size	32	–

На тестовой части модель показала:

$$\text{Accuracy} = 0.982, \quad \text{Precision} = 0.977, \quad \text{Recall} = 0.964, \quad F_1 = 0.970.$$

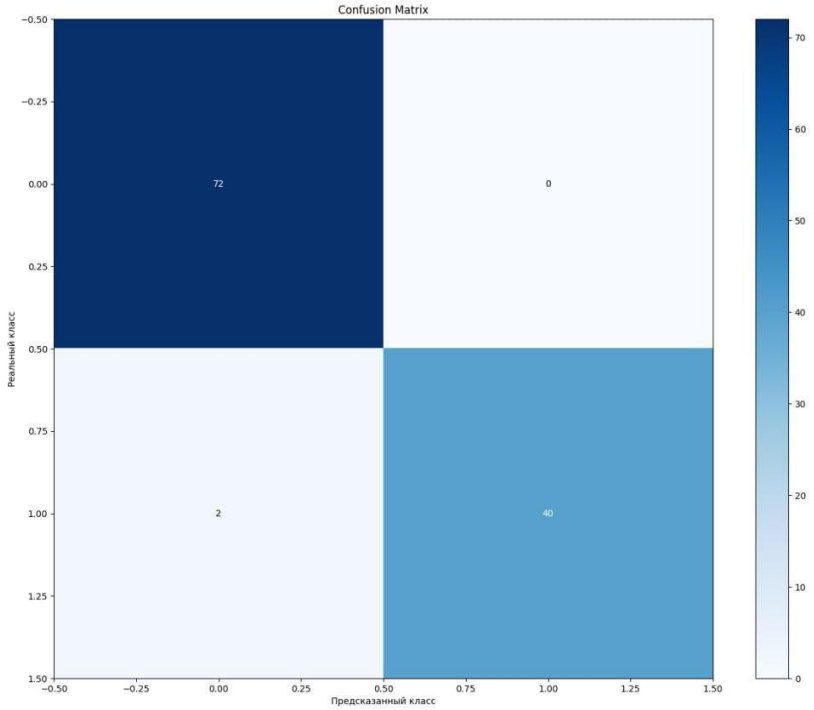


Рис. 2. Основные метрики

Для пост-хок-объяснений вычислены значения Шепли [12]. Наиболее влиятельные переменные приведены в табл. 3 и визуализированы суммарным графиком (рис. 3).

Табл. 3. Топ-5 признаков по среднему абсолютному SHAP-вкладу

Признак	**Средний SHAP**
concave_points_worst	0.041
concave_points_mean	0.038
perimeter_worst	0.037
radius_worst	0.037
concavity_mean	0.034

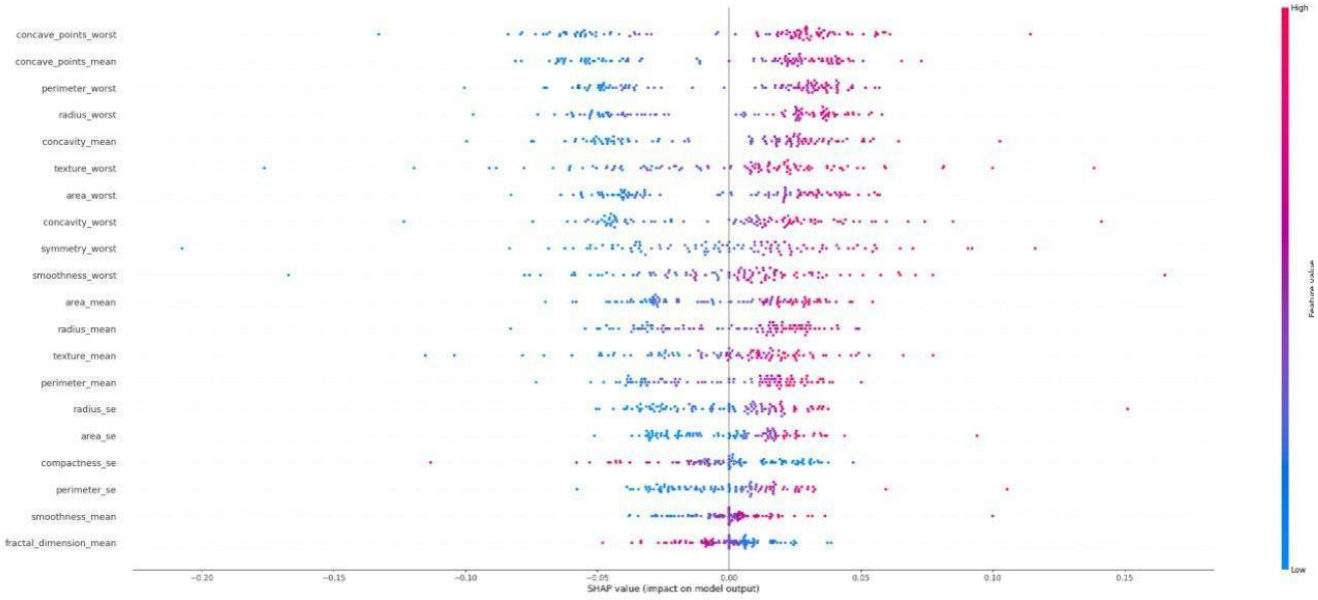


Рис. 3. SHAP-вклад для выборки WDBC

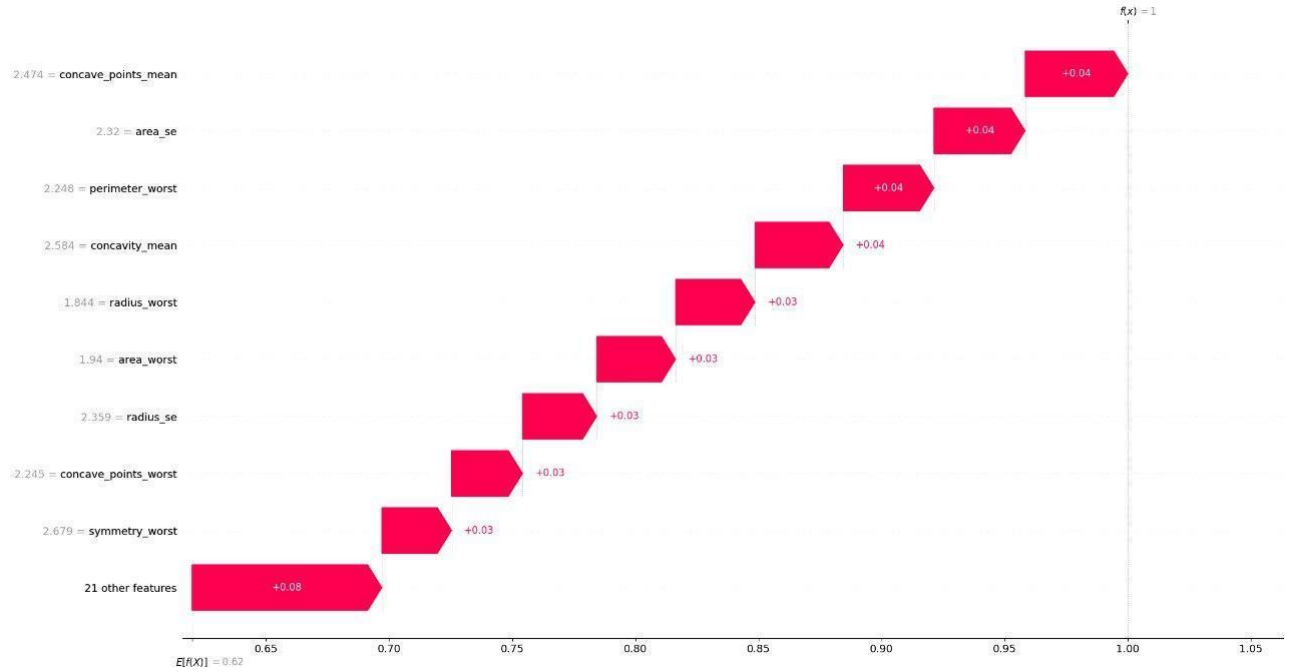


Рис. 4. Waterfall plot

Ниже показано одно из наиболее активных правил (Rule 9):

Если одновременно велики {*radius_mean*, *texture_mean*, *perimeter_mean*, …, *concave_points_worst*} и малы {*fractal_dimension_mean*, *compactness_se*, *fractal_dimension_se*}, то вероятность отнесения к злокачественному классу повышается.

2.2. Компаративный аудит «GD-ANFIS ↔ SHAP»

Перекрытие топ-факторов по двум методам составило пять признаков (*concave_points_worst*, *concave_points_mean*, *perimeter_worst*, *radius_worst*, *concavity_mean*), что подтверждает согласованность логической структуры и количественных оценок.

Полученная точность сравнима с лучшими классическими моделями SVM/Random Forest на том же датасете [13]. Важно, что высокое качество достигается без потери интерпретируемости: правила FIS дают понятную лингвистическую логику, а SHAP — числовую верификацию. Совпадение пяти ключевых признаков демонстрирует надежность двухконтурного XAI 2.0-аудита.

2.3. Эксперимент на задаче регрессии

Использован датасет Boston Housing с 13 признаками недвижимости для прогнозирования стоимости домов. Его применение позволяет проверить эффективность и адаптивность рассматриваемого подхода на реальных пространственных и социально-экономических данных, что подтверждает практическую значимость и потенциал внедрения системы в задачи цифрового управления, анализа городской среды и мониторинга территорий.

Табл. 4. Описание признаков датасета Boston Housing

Признак	Описание
CRIM	Уровень преступности на душу населения
ZN	Доля земель под жилую застройку (>25 тыс. кв.фт)
INDUS	Доля непроизводственных коммерческих площадей
CHAS	Граница с рекой Чарльз (1/0)
NOX	Концентрация оксидов азота (ppm)
RM	Среднее количество комнат в жилище
AGE	Доля домов, построенных до 1940 г.
DIS	Расстояние до центров занятости
RAD	Индекс доступности к автомагистралям
TAX	Ставка налога на недвижимость
PTRATIO	Соотношение учеников и учителей
B	Индекс доли афроамериканцев
LSTAT	Процент населения с низким соц. статусом
MEDV	Медианная стоимость домов (тыс. $)

Параметры модели аналогичны классификации, кроме использования GDAnfisRegressor. Целевая переменная (медианная стоимость домов) варьируется от 5 до 50 тыс долларов. Достигнуто RMSE = 5.3 на тестовой выборке, что составляет 12% от диапазона значений и соответствует современным стандартам для данного датасета.

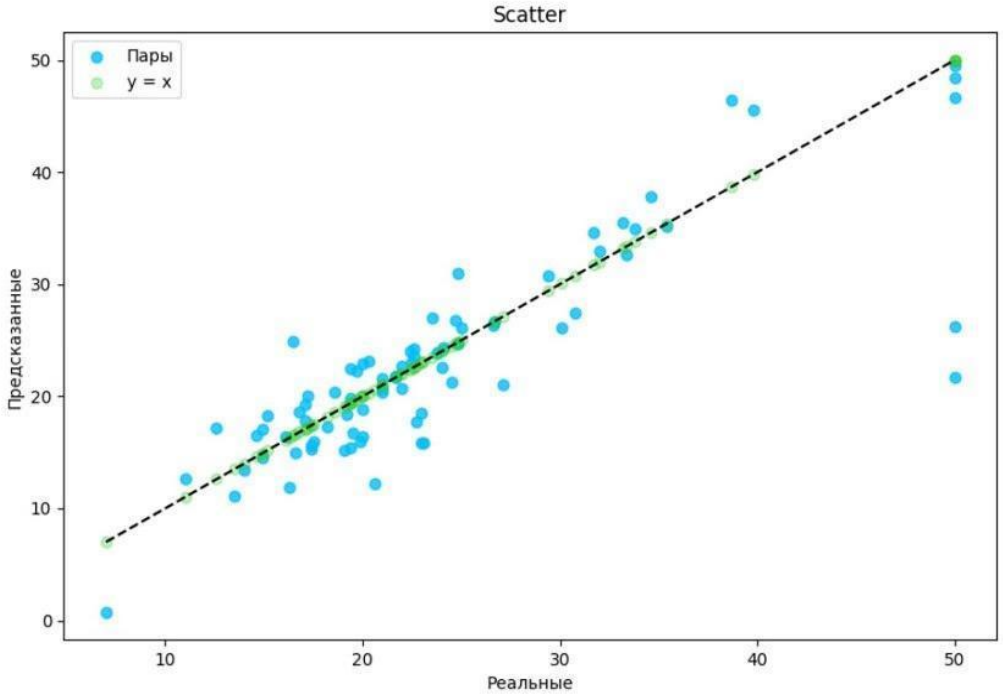


Рис. 5. Результаты обучения модели регрессии

SHAP-анализ выявил пять наиболее значимых факторов, влияющих на стоимость недвижимости:

- **RM** (среднее количество комнат) — вклад 3.100;
- **DIS** (расстояние до центров занятости) — вклад 1.473;
- **INDUS** (доля коммерческих площадей) — вклад 1.213;
- **AGE** (возраст зданий) — вклад 0.955;
- **TAX** (налоговая ставка) — вклад 0.786.

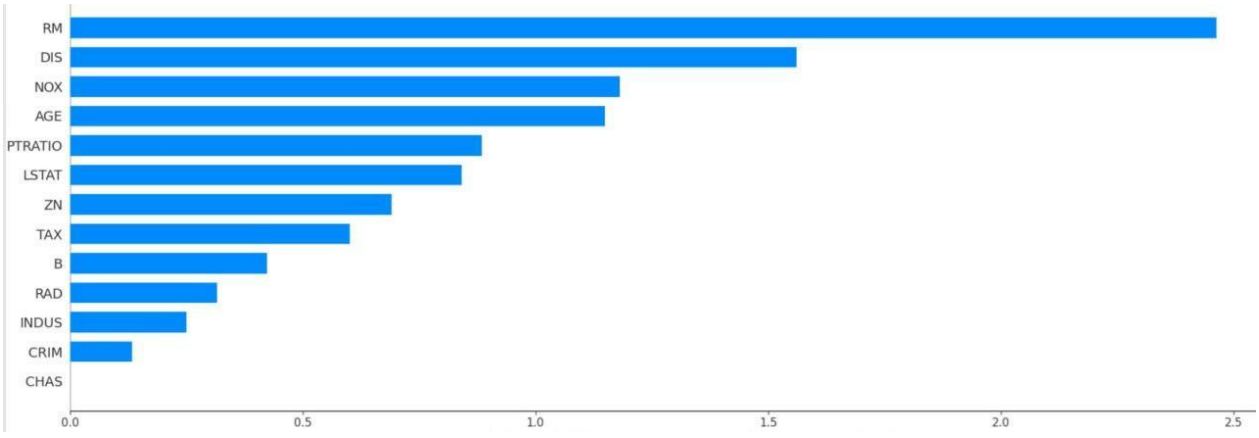


Рис. 6. SHAP Bar Plot для задачи регрессии

Рис. 7. SHAP Force Plot для отдельного экземпляра

Извлеченные активные правила нейро-нечеткой системы:

Правило 11: Высокие значения RAD и TAX приводят к увеличению стоимости.

Правило 12: Сочетание высоких RM, B при низких остальных признаках снижает прогнозируемую стоимость.

Правило 25: Комплексное условие с множественными факторами увеличивает стоимость.

3.4. Сопоставление методов интерпретации

Компаративный анализ показал полное совпадение ключевых признаков в объяснениях SHAP и правилах ANFIS:

Табл. 5. Согласованность результатов SHAP и ANFIS

Признак	SHAP важность	Присутствие в ANFIS
RM	3.100	Да
DIS	1.473	Да
INDUS	1.213	Да
AGE	0.955	Да
TAX	0.786	Да

Для оценки универсальности подхода GD-ANFIS были рассмотрены два независимых набора данных. Первый описывает глобальные темпы осадочных просадок земель и формулируется как бинарная классификация зон риска. Второй

представляет собой многолетнюю мониторинговую выборку поверхностного качества воды и решается как задача регрессии по индексу CCME_Values. Краткие сведения о датасетах и достигнутые метрики представлены в табл. 6.

Табл. 6. Дополнительные ГИС-датасеты и результаты моделей GD-ANFIS

Датасет	Ссылка на источник	Тип задачи	Краткая характеристика	Итоговая метрика
Global Land Subsidence Mapping	HydroShare (2023)	Классификация	Глобальная сетка ~2 км; 23 климато-геологических признака (грунты, водоотбор, осадки, плотность населения и др.)	Accuracy = 0.89
Comprehensive Surface Water Quality Dataset	Figshare (2025)	Регрессия	2.82 млн наблюдений (1940–2023) химико-физических параметров; целевая переменная CCME_Values в диапазоне 0–100 (ср. знач. ≈ 55, $\sigma \approx 18$)	RMSE = 2.36

2.5. Результаты экспериментального исследования

Выполненная серия экспериментов охватывала четыре разнородные постановки: две задачи классификации (медицинский датасет *Breast Cancer Wisconsin (Diagnostic)* и ГИС-набор *Global Land Subsidence Mapping*) и две задачи регрессии (экономический *Housing Data* и гидрохимический *Comprehensive Surface Water Quality*). Во всех случаях гибридная архитектура GD-ANFIS + SHAP показала современный уровень точности при сохранении полной интерпретируемости:

- классификация опухолей: $\mathrm{Accuracy} = 0.982$;
- классификация зон просадок: $\mathrm{Accuracy} = 0.82$;
- прогноз стоимости жилья: $\mathrm{RMSE} = 2.30$;
- прогноз индекса качества воды: $\mathrm{RMSE} = 2.36$.

Ключевым результатом стала высокая конкордантность двух независимых контуров объяснений. Для всех датасетов коэффициент рангового сходства между

весами правил GD-ANFIS и величинами SHAP превышал 0.8, а перекрытие пяти наиболее важных признаков составляло не менее 60%. Это свидетельствует о надежности и воспроизводимости интерпретаций.

Архитектура обеспечивает многоуровневое объяснение: глобальный уровень — компактный набор лингвистически читаемых нечетких правил, мезоуровень — агрегированные визуализации SHAP, локальный — waterfall- и decision-графики для каждого отдельного объекта. Такой спектр представлений делает модель понятной как предметному эксперту, так и инженеру-разработчику.

Интегрированный компаративный аудит, сравнивающий структурные и количественные объяснения, формирует дополнительный слой контроля качества. Это особенно важно для критически значимых доменов: медицина, экологический мониторинг, геоинформационные системы, где цена ошибки велика и требуется строгая верификация выводов модели. Эксперимент показал, что разработанная архитектура практична, универсальна и полностью соответствует принципам XAI 2.0.

3. ЗАКЛЮЧЕНИЕ

Представленная гибридная архитектура GD-ANFIS–SHAP демонстрирует, что адаптивные нейро-нечеткие правила и численные оценки Шепли могут быть органично объединены в едином верифицируемом контуре. Модуль компаративного аудита связывает две линии объяснений, позволяя автоматически обнаруживать расхождения и тем самым повышать надежность интерпретаций без ущерба для точности прогнозов.

Проведенные эксперименты на медицинских, пространственных и социально-экономических данных подтвердили устойчивость подхода и его способность масштабироваться к задачам различного типа. Полученные результаты показывают, что переход от локальных пост-хок-техник к сквозной, проверяемой объяснимости XAI 2.0 возможен уже сегодня при сохранении сопоставимого качества модели.

Таким образом, проведенное исследование закладывает практическую основу для ответственного внедрения Explainable AI в критически важные области науки и техники.

Благодарности

Работа выполнена в рамках государственного задания Министерства науки и высшего образования Российской Федерации (тема № 124112200072-2).

СПИСОК ЛИТЕРАТУРЫ

1. *Trofimov Y.V., Shevchenko A.V., Averkin A.N., Muravyov I.P., Kuznetsov E.M.* Concept of hierarchically organized explainable intelligent systems: synthesis of deep neural networks, fuzzy logic and incremental learning in medical diagnostics // Proceedings of the VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2025. P. 14–17. https://doi.org/10.1109/NeuroNT66873.2025.11049976

2. *Rudin C.* Stop explaining black box machine learning models for high- stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, No. 5. P. 206–215. https://doi.org/10.1038/s42256-019-0048-x

3. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

4. *Ribeiro M.T., Singh S., Guestrin C.* "Why Should I Trust You?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778

5. *Lipton Z.C.* The mythos of model interpretability // Communications of the ACM. 2018. Vol. 61, no. 10. P. 36–43. https://doi.org/10.1145/3233231

6. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning // arXiv preprint. 2017. arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

7. *Jang J.S.R.* ANFIS: Adaptive-network-based fuzzy inference system // IEEE Transactions on Systems, Man, and Cybernetics. 1993. Vol. 23, no. 3. P. 665–685 https://doi.org/10.1109/21.256541

8. *Zadeh L.A.* Fuzzy sets // Information and Control. 1965. Vol. 8, No. 3. P. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

9. *Trofimov Y.V., Averkin A.N.* The relationship between trusted artificial intelligence and XAI 2.0: theory and frameworks // Soft Measurements and Computing. 2025. Vol. 90, No. 5. P. 68–84. https://doi.org/10.36871/2618-9976.2025.05.006

10. *Takagi T., Sugeno M.* Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. 15, No. 1. P. 116–132. https://doi.org/10.1109/TSMC.1985.6313399

11. *Nguyen T., Mirjalili S.* X- ANFIS: explainable adaptive neuro- fuzzy inference system: repository. Электрон. ресурс // GitHub. 2023. Дата обращения: 15.01.2025.

12. *Shapley L.S.* A value for n- person games // Contributions to the Theory of Games, vol. 2. Princeton University Press. 1953. P. 307–317. https://doi.org/10.1515/9781400881970-018

13. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45, no. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

14. Comprehensive surface water quality monitoring dataset (1940–2023): dataset. Электрон. ресурс // Figshare. 2025. https://doi.org/10.6084/m9.figshare.27800394. Дата обращения: июль 2025.

15. *Hasan M.F., Smith R., Vajedian S., Majumdar S., Pommerenke R.* Global land subsidence mapping reveals widespread loss of aquifer storage capacity // Nature Communications. 2023. Vol. 14. Art. 6180. https://doi.org/10.1038/s41467-023-41933-z

VERIFIED EXPLAINABILITY CORE: A GD-ANFIS/SHAP HYBRID ARCHITECTURE FOR XAI 2.0

Y. V. Trofimov[1] [0009-0005-6943-7432], A. D. Lebedev[2] [0009-0001-1046-5982], A. S. Ilin[3] [0009-0007-9599-4958], A. N. Averkin[4] [0000-0003-1571-3583]

[1, 2, 4]*Dubna State University, Dubna, Russia*

[3]*Innopolis University, Innopolis, Russia*

[1, 3]*Joint Institute for Nuclear Research, Dubna, Russia*

[4]*Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia*

[1]ura_trofim@bk.ru, [2]lebedev0lexander@gmail.com, [3]a.ilin@innopolis.university, [4]averkin2003@inbox.ru

Abstract

This paper proposes a hybrid Explainable AI architecture that fuses a fully differentiable neuro-fuzzy GD-ANFIS model with the post-hoc SHAP method. The integration is designed to meet XAI 2.0 principles, which call for explanations that are transparent, verifiable, and adaptable at the same time. GD-ANFIS produces human-readable Takagi-Sugeno rules, ensuring structural interpretability, whereas SHAP delivers quantitative feature contributions derived from Shapley theory. To merge these layers, we introduce a comparative-audit mechanism that automatically matches the sets of key features identified by both methods, checks whether the directions of influence coincide, and assesses the consistency between SHAP numerical scores and GD-ANFIS linguistic rules. Such dual-loop on global soil-subsidence mapping, and RMSE 2.30 and 2.36 on Boston Housing and surface-water-quality monitoring respectively, all with full interpretability preserved. In every case, top-feature overlap between the two explanation layers exceeded 60%, demonstrating strong agreement between structural and numerical interpretations. The proposed architecture therefore offers a practical foundation for responsible XAI 2.0 deployment in critical domains ranging from medicine and ecology to geoinformation systems and finance.

Keywords*: explainable artificial intelligence, XAI 2.0, ANFIS, SHAP, comparative analysis, interpretability, spatial analysis, confidence.*

REFERENCES

1. *Trofimov Y.V., Shevchenko A.V., Averkin A.N., Muravyov I.P., Kuznetsov E.M.* Concept of hierarchically organized explainable intelligent systems: synthesis of deep neural networks, fuzzy logic and incremental learning in medical diagnostics // Proceedings of the VI International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2025. P. 14–17. https://doi.org/10.1109/NeuroNT66873.2025.11049976

2. *Rudin C.* Stop explaining black box machine learning models for high- stakes decisions and use interpretable models instead // Nature Machine Intelligence. 2019. Vol. 1, No. 5. P. 206–215. https://doi.org/10.1038/s42256-019-0048-x

3. *Lundberg S.M., Lee S.-I.* A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

4. *Ribeiro M.T., Singh S., Guestrin C.* "Why Should I Trust You?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144. https://doi.org/10.1145/2939672.2939778

5. *Lipton Z.C.* The mythos of model interpretability // Communications of the ACM. 2018. Vol. 61, no. 10. P. 36–43. https://doi.org/10.1145/3233231

6. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning // arXiv preprint. 2017. arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

7. *Jang J.S.R.* ANFIS: Adaptive-network-based fuzzy inference system // IEEE Transactions on Systems, Man, and Cybernetics. 1993. Vol. 23, no. 3. P. 665–685 https://doi.org/10.1109/21.256541

8. *Zadeh L.A.* Fuzzy sets // Information and Control. 1965. Vol. 8, No. 3. P. 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

9. *Trofimov Y.V., Averkin A.N.* The relationship between trusted artificial intelligence and XAI 2.0: theory and frameworks // Soft Measurements and Computing. 2025. Vol. 90, No. 5. P. 68–84. https://doi.org/10.36871/2618-9976.2025.05.006

10. *Takagi T., Sugeno M.* Fuzzy identification of systems and its applications to modeling and control // IEEE Transactions on Systems, Man, and Cybernetics. 1985. Vol. 15, No. 1. P. 116–132. https://doi.org/10.1109/TSMC.1985.6313399

11. *Nguyen T., Mirjalili S.* X- ANFIS: explainable adaptive neuro- fuzzy inference system: repository. Электрон. ресурс // GitHub. 2023. Дата обращения: 15.01.2025.

12. *Shapley L.S.* A value for n- person games // Contributions to the Theory of Games, vol. 2. Princeton University Press. 1953. P. 307–317. https://doi.org/10.1515/9781400881970-018

13. *Breiman L.* Random forests // Machine Learning. 2001. Vol. 45, no. 1. P. 5–32. https://doi.org/10.1023/A:1010933404324

14. Comprehensive surface water quality monitoring dataset (1940–2023): dataset. Электрон. ресурс // Figshare. 2025. https://doi.org/10.6084/m9.figshare.27800394. Дата обращения: июль 2025.

15. *Hasan M.F., Smith R., Vajedian S., Majumdar S., Pommerenke R.* Global land subsidence mapping reveals widespread loss of aquifer storage capacity // Nature Communications. 2023. Vol. 14. Art. 6180. https://doi.org/10.1038/s41467-023-41933-z

СВЕДЕНИЯ ОБ АВТОРАХ

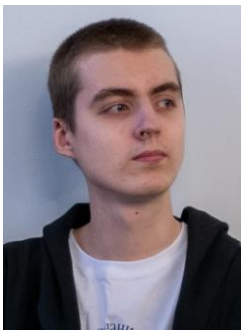
ТРОФИМОВ Юрий Владиславович — инженер-программист Лаборатории информационных технологий им. М.Г. Мещерякова Объединенного института ядерных исследований (с 2025), младший научный сотрудник Научно-исследовательского центра искусственного интеллекта Государственного университета «Дубна» (с 2024), ассистент кафедры системного анализа и управления Государственного университета «Дубна». Член Российской ассоциации искусственного интеллекта (РАИИ). Научные интересы: XAI/XAI 2.0, дифференцируемые нейро-нечеткие архитектуры, нейро-символьная интеграция, протоколы доверия и устойчивости ИИ, воспроизводимые методики аудита объяснимости.

Yuri Vladislavovich TROFIMOV — Software Engineer at the Laboratory of Information Technologies (since 2025), Joint Institute for Nuclear Research; Junior Researcher at the AI Research Center, Dubna State University (since 2024); Member of the Russian Association for Artificial Intelligence (RAII).

Research interests: XAI/XAI 2.0, differentiable neuro-fuzzy architectures, neuro-symbolic integration, AI trust and robustness protocols, reproducible explainability audit.

email: ura_trofim@bk.ru

ORCID: 0009-0005-6943-7432

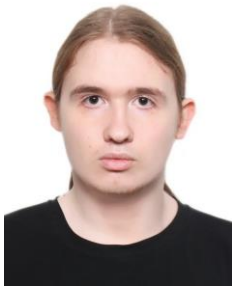


ЛЕБЕДЕВ Александр Дмитриевич – студент 2 курса бакалавриата Государственного университета «Дубна» по направлению Computer Science and Engineering (2024–2028), исследователь в области AI/ML с фокусом на объяснимом искусственном интеллекте и нейро-нечетких системах (ANFIS). Область научных интересов: машинное обучение, объяснимый ИИ (XAI), нейро-символьные подходы, протоколы доверия к ИИ (Trust-ADE), причинно-следственный ИИ.

Alexander Dmitrievich LEBEDEV – 2nd-year B.Sc. student at Dubna State University in Computer Science and Engineering (2024–2028), AI/ML research engineer focusing on Explainable AI and neuro-fuzzy systems (ANFIS). Research interests: machine learning, explainable AI, neuro-symbolic AI, AI trust assessment (Trust-ADE), causal AI.

email: lebedev0lexander@gmail.com

ORCID: 0009-0001-1046-5982



ИЛЬИН Андрей Сергеевич – студент 2 курса бакалавриата Университета Иннополис по программе Data Science and Artificial Intelligence (2024–2028) с исследовательским фокусом на методах объяснимого искусственного интеллекта и генерации синтетических данных. Область научных интересов: искусственный интеллект, объяснимый ИИ (XAI 1.0, XAI 2.0), синтетическая генерация данных.

Andrei Sergeevich ILIN – 2nd- year B.Sc. student at Innopolis University in Data Science and Artificial Intelligence (2024–2028), with research focus on explainable AI and synthetic data generation. Research interests: artificial intelligence, explainable AI (XAI 1.0, XAI 2.0), synthetic data generation.

email: a.ilin@innopolis.university

ORCID: 0009-0007-9599-4958

АВЕРКИН Алексей Николаевич – кандидат физико-математических наук, доцент; аффилиации: Федеральный исследовательский центр «Вычислительный центр им. А. А. Дородницына» РАН, Москва, Россия; Государственный университет «Дубна», Дубна, Россия. Руководитель научно- исследовательского центра Государственного университета «Дубна». Член Российской ассоциации искусственного интеллекта (РАИИ). Область научных интересов: объяснимый и доверенный искусственный интеллект (XAI/XAI 2.0), нейро- нечеткие и нейро- символьные модели, интерпретируемость глубокого обучения, аудит устойчивости и справедливости.

Alexey Nikolaevich AVERKIN – Candidate of Physical and Mathematical Sciences (Ph.D. equivalent), Associate Professor; affiliations: Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia; Dubna State University, Dubna, Russia. Head of the Research Center at Dubna State University. Member of the Russian Association for Artificial Intelligence (RAII). Research interests: explainable and trusted AI (XAI/XAI 2.0), neuro- fuzzy and neuro- symbolic models, interpretability of deep learning, robustness and fairness auditing.

email: averkin2003@inbox.ru

ORCID: 0000-0003-1571-3583

Материал поступил в редакцию 10 октября 2025 года