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. П. Муравьев, Е. М. Кузнецов

Государственный университет «Дубна»

E-mail ura_trofim@bk.ru

Аннотация. **Развитие объяснимого искусственного интеллекта (XAI) выходит за рамки традиционных постхок-методов интерпретации, формируя новое поколение когнитивно-ориентированных интеллектуальных систем, основанных на глубокой интеграции нейросетевых методов, нечеткой логики и инкрементного обучения. В докладе представлена концепция иерархически организованных XAI-систем, обеспечивающих адаптивную объяснимость и непрерывное обновление знаний в процессе их практического применения. Реализация опирается на динамическое формирование объяснений, включающее каскадное соединение локальных и глобальных интерпретаторов, продукционных моделей и адаптивных механизмов генерации медицинских заключений. Представленный подход формирует парадигму XAI 2.0, обеспечивающую устойчивость к неопределённости медицинских данных, когнитивную прозрачность решений и возможность многоуровневой интерпретации результатов. В работе обсуждаются теоретические и практические принципы построения таких систем, а также обосновывается необходимость интеграции глубоких нейросетей и нечеткой логики для создания интеллектуальных медицинских систем нового поколения.**

Ключевые слова: искусственный интеллект, интеллектуальные медицинские системы, интерпретируемый искусственный интеллект, объяснимый искусственный интеллект, когнитивно-ориентированные системы, нечеткая логика, инкрементное обучение, адаптивная объяснимость

I. Введение

Современное развитие технологий искусственного интеллекта (ИИ) в медицине сталкивается с проблемой недостаточной объяснимости принимаемых решений. Традиционные методы построения глубокой нейронной сети (Deep Neural Network, DNN) сфокусированы на достижении высокой точности, однако их внутренняя структура часто представляется «чёрным ящиком», что препятствует полноценному доверию врачей и экспертов при формировании клинических заключений. Чтобы преодолеть эту проблему, интенсивно развиваются подходы XAI, нацеленные на создание прозрачных и интерпретируемых систем, сохраняющих при этом конкурентное качество распознавания и анализа [1]. Вместе с тем, классические пост-хок методы (например, Grad-CAM, SHAP или LIME) нередко рассматриваются лишь как дополнительные слои интерпретации, накладываемые на уже обученную сеть. Подобная стратегия улучшает локальную объяснимость, но не всегда учитывает возникающую необходимость глобально объяснять свои выводы в формах, воспринимаемых человеком.

Совсем недавно появилось направление XAI 2.0, призванное расширить классические методы за счёт междисциплинарных подходов и устранить оставшиеся проблемы трактовки сложных моделей [2]. Идеология XAI 2.0 предполагает глубокую интеграцию разнородных технологий – нейросетевых моделей, методов нечёткой логики, продукционных правил и языковых моделей – в единую иерархическую архитектуру объяснимой интеллектуальной системы. Такой подход нацелен на обеспечение когнитивной интерпретируемости (т. е. соответствия объяснений способу рассуждений врача), адаптивной объяснимости (динамического подстройки объяснений под контекст) и устойчивости к неопределённости данных (надёжной работы с шумами и неполнотой информации).

Цель статьи – предложить концептуальную архитектуру иерархически организованной объяснимой интеллектуальной системы, в которой (1) глубокие нейронные сети (включая сверточные и рекуррентные модули) дополняются слоями нечеткой логики, (2) пост-хок интерпретаторы работают параллельно с системой формирования «мягких дескрипторов», (3) все метрики, объяснения и локальные правила агрегируют в специальную промежуточную многоуровневую модель, и (4) итоговая информация о принятом решении и диагностических факторах передается в языковую модель, генерирующую медицинское заключение. Инкрементный характер обучения позволяет периодически переобучать систему с учётом новых данных, что формирует динамично обновляемые и одновременно объяснимые медицинские решения.

Структура работы следующая. В разделе 2 кратко излагаются теоретические предпосылки создания гибридных систем на основе глубоких сетей и нечеткой логики, а также основные идеи пост-хок методов объяснения и их ограничения. Кроме того,

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

затрагиваются теоретические основы инкрементного обучения, обеспечивающие непрерывное обновление модели. В разделе 3 описывается многослойная архитектура XAI, включающая нечеткие слои, агрегаторы метрик и языковую модель. В разделе 4 приводятся принципы оценки качества, включающие когнитивную интерпретируемость, устойчивость к шуму и эффективность инкрементного апдейта. В заключении (раздел 5) суммируются основные результаты работы и очерчиваются перспективы развития XAI 2.0 для медицинской диагностики».

II. Теоретические основы: нечеткая логика и пост-хок объяснения и инкрементное обучение

A. *XAI в медицинской диагностике*

Согласно определению Doshi-Velez и Kim, интерпретируемость – это способность объяснить или представить работу модели понятными для человека терминами [3]. В медицинском контексте это означает, что алгоритм должен обоснованно указывать, какие входные данные (симптомы, биомаркеры, признаки на изображении) привели к определенному заключению.

За последнее десятилетие опубликовано множество исследований, посвященных XAI в медицине [4–5]. Наиболее распространенные методы – это пост-хок объяснения для уже обученных моделей глубокого обучения. Эти методы успешно повышают прозрачность отдельных предсказаний. Однако они не меняют саму модель и не всегда удовлетворяют потребности глобальной объяснимости – т. е. общего понимания того, как модель принимает решения во всех случаях, и соответствия этой логики медицинским знаниям.

Альтернативный подход – разработка интерпретируемых моделей с нуля, закладывая объяснимость в их структуру. Примеры таких моделей включают решающие деревья, байесовские и прототипные сети, различные нечеткие экспертные системы. В медицине исторически распространены экспертные системы на базе нечеткой логики, где знание врача задается в виде набора правил вида «ЕСЛИ X и Y, ТО диагноз Z» с нечеткими лингвистическими термами. Они легко интерпретируются человеком и допускают формальный вывод с учетом неопределенности входных данных. Такие системы успешно применялись для диагностики широкого круга заболеваний, от пневмонии до эндокринных нарушений, но вручную составленные правила ограничены охватом и гибкостью. Современные исследования стремятся объединить преимущества обучения на данных с интерпретируемостью правил, что привело к возрождению интереса к нейро-нечетким гибридным моделям [6].

В литературе описаны различные архитектурные решения для нейро-нечетких сетей. Одни из них внедряют слой нечёткого вывода внутрь нейросети, другие – обучают нейросеть и последовательно аппроксимируют ее решения набором правил (педагогический метод извлечения правил) [7–9]. Варианты реализованы как с правилами Мамдани, более интерпретируемыми для человека, так и с правилами Сугено, удобными для гибридного обучения. Отдельно развивается направление эволюционирующих нечетких нейронных сетей – систем, которые в процессе обучения сами добавляют новые правила или нейроны по мере необходимости, эффективно реализуя инкрементное наращивание знаний. В частности, существуют методы, где при поступлении новых данных, плохо объясняемых текущей моделью, генерируется новый нечеткий кластер и соответствующее правило, тем самым расширяя модель без переписывания старых правил. Такой подход особенно перспективен для долгосрочного использования в медицине, где система может эволюционировать по мере накопления клинического опыта

B. *Инкрементное обучение и устойчивость к изменениям данных*

Инкрементное обучение – это способность модели обучаться последовательно на новых порциях данных или новых задачах, не забывая ранее изученного. Инкрементное обновление предпочтительнее, чем полное переобучение, когда имеется большой уже накопленный опыт – дообучение экономит вычислительные ресурсы и позволяет непрерывно внедрять улучшения.

Основная сложность при дообучении нейросетей – избежать катастрофического забывания, когда адаптация на новых данных приводит к деградации качества на старых данных. В последние годы разработан ряд стратегий для борьбы с этим: регуляризационные методы (например, Elastic Weight Consolidation – добавление штрафа за отклонение важных весов от старых значений), репетиционные методы (rehearsal, опыт с воспроизведением части старых данных при обучении на новых), архитектурные методы (динамическое увеличение архитектуры сети для новых знаний) и др. [10–13]. В контексте медицинских моделей эти подходы только начинают применяться.

Интеграция инкрементного обучения с XAI представляет особый интерес. С одной стороны, объяснимость может помочь контролировать процесс дообучения – эксперт видит, какие новые правила добавила модель, и может оценить их соответствие знаниям. С другой стороны, сам механизм нечеткой логики благоприятствует инкрементальности: новые знания могут добавляться в систему в виде новых правил, не разрушая старую базу (если конфликтуют – возможно присвоение им меньшего веса либо разрешение конфликтов по специфичности условия). Нечеткая система, будучи модульной, позволяет «локально» обновлять ту часть правил, которая связана с новым типом данных, не влияя на остальные. Это дает потенциальное преимущество перед монолитными нейросетями, где любое изменение требует глобального перераспределения весов. В связи с этим синергия нечеткой логики и инкрементного обучения выглядит естественной: первая обеспечивает структуру знаний, а второе – механизм ее эволюции.

Таким образом, обзоры существующих исследований показывают: (1) объяснимость является критически важной характеристикой для медицинских ИИ-систем, и наиболее перспективны подходы, встроенные в саму модель (не пост-хок), (2) нейро-нечеткие гибриды предоставляют интепретируемый каркас с выводами правил и уже демонстрируют успехи в ряде медицинских задач, (3) инкрементное обучение – востребованный функционал, позволяющий интеллектуальным системам «учиться в процессе использования», и его сочетание с интерпретируемыми

моделями открывает новые возможности для создания доверенных систем ИИ в медицине.

III. Многоуровневая XAI-архитектура: глубокая сеть, нечеткая логика и агрегатор объяснений

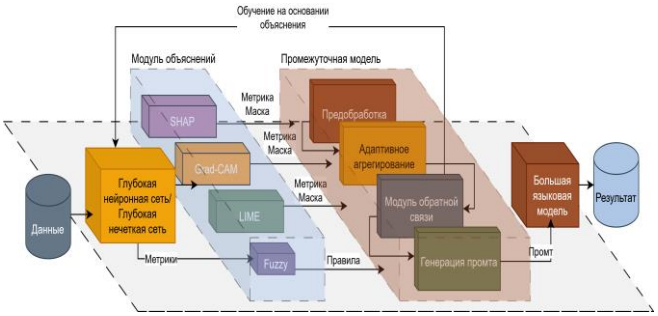

Рис. 1. Многоуровневая XAI-архитектура

A. *Общая схема*

Архитектура имеет несколько уровней, организованных каскадно (рис. 1). Разные уровни объяснений формируются и консолидируются в рамках единого конвейера, рис. 1:

- Исходные данные поступают на вход глубокой сети с нечеткими слоями (DNN + Fuzzy).
- Выходы глубокой сети параллельно подаются на пост-хок методы, которые обеспечивают локальное объяснение.
- Нечетко-логический блок (вшитый в архитектуру) порождает набор мягких дескрипторов, которые можно рассматривать как продукционные правила или как параметры принадлежности к нечетким термам.
- Специальный модуль-коллектор собирает метрики (точность, recall, precision), выходы пост-хок объяснений, результаты нечетких правил и прочую дополнительную информацию о значимости признаков.
- Многоуровневая промежуточная модель (рис. 2) агрегирует все вышеперечисленные данные, проводит объединение и формирует единое представление о работе системы. Идея состоит в том, что каждый метод объяснения имеет свои сильные стороны и ограничения, поэтому их композитное использование может дать более надежное и понятное объяснение.
- Контур обучения «с объяснениями» (рис. 3). Именно на этом уровне становится возможным обучение на основе объяснений: модель оптимизирует не только функцию потерь на входных данных, но и учитывает согласованность, «понятность» и стабильность локальных объяснений.
- Модуль генерации промта (рис. 4) создаёт итоговые обобщённые выводы, передает в языковую модель, которая генерирует прозрачное для врача-эксперта медицинское заключение.

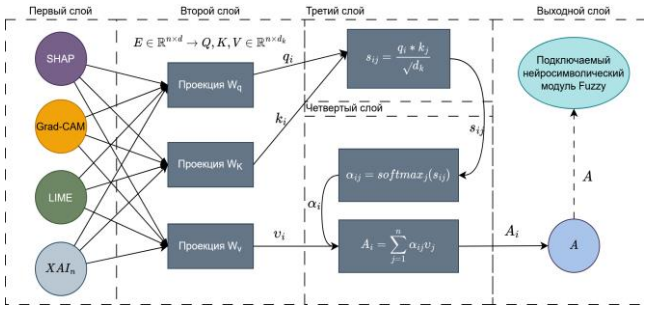

Рис. 2. Промежуточная модель (адаптивное агрегирование)

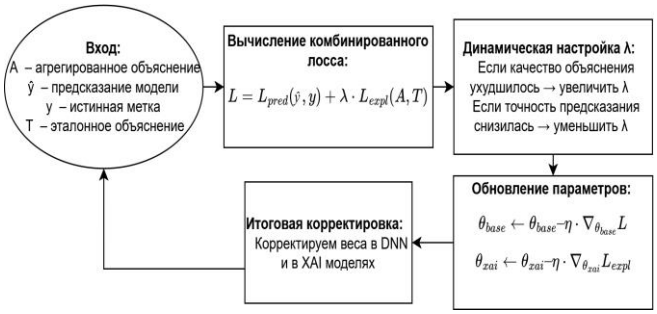

Рис. 3. Модуль обратной связи

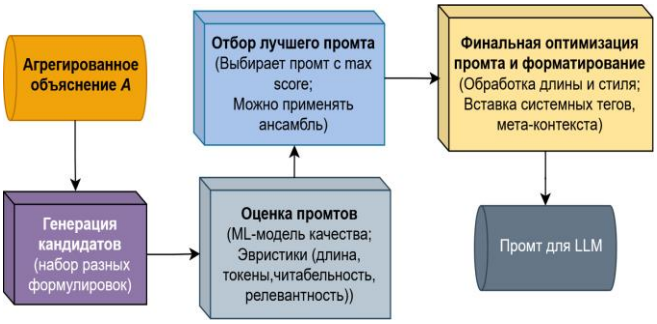

Рис. 4. Модуль генерации промта

IV. Принципы оценки качества, включающие когнитивную интерпретируемость

Еще один момент – оценка качества объяснений. В настоящее время нет общепринятой метрики «интерпретируемости». Мы можем измерить точность предсказаний, но как формально измерить, насколько хорошо система объясняет свои решения? Здесь могут помочь пользовательские исследования (опросы врачей, сравнение скорости и правильности принятия решений с помощью системы и без нее). Но для широкого признания верификация XAI-архитектур должна опираться на трансдисциплинарную методологию когнитивных измерений, интегрирующую психометрику, эргономику и информационную метрику. Предлагаемый каркас включает семь операциональных индикаторов: (1) понятностная прозрачность объяснения, (2) субъективная удовлетворённость и (3) доверие пользователя, (4) когнитивная вовлечённость, (5) сформированная ментальная модель, (6) прирост продуктовой эффективности (скорость / точность решения клинической задачи) и (7) управляемость человек-машина диалога. Количественная операционализация реализуется через объединённый индекс простоты (число правил, сложность графа зависимостей), коэффициент конкордантности с экспертными эвристиками и перцептивное время интерпретации [14].

V. Заключение

Предложенная XAI 2.0-архитектура демонстрирует, что объединение нейросетевых и символических методов позволяет достигнуть баланса между точностью и объяснимостью, необходимого в медицине. Когнитивная интерпретируемость системы обусловлена тем, что объяснения формируются на нескольких уровнях: от низкоуровневых до высокоуровневых правил и выводов на естественном языке. Это многоуровневое объяснение соответствует человеческому стилю рассуждений: сначала внимание на деталях, затем синтез в общую картину и вывод. Кроме того, использование терминов и правил, понятных врачам, делает объяснения когнитивно приемлемыми – специалисты легче воспринимают и доверяют таким решениям ИИ, поскольку они говорят на «их языке». Адаптивная объяснимость достигается за счёт способности системы подстраивать механизм объяснения под новые данные и ситуации. Инкрементное обучение позволяет обновлять как саму модель, так и интерпретирующие компоненты по мере накопления случаев, где предыдущие объяснения были недостаточны. Кроме того, система может менять детализацию объяснения в зависимости от аудитории: например, для врача предоставить подробный разбор по признакам и правилам, а для пациента – более простое объяснение через языковую модель. Такой гибкий подход к объяснениям обеспечивает устойчивость системы в реальной клинической практике, где каждый кейс уникален.

Архитектура успешно апробирована на четырёх клинических задачах: (1) автоматизированная диагностика пневмонии по рентгенограммам грудной клетки; (2) многофакторная стратификация риска хронической обструктивной болезни лёгких (ХОБЛ); (3) неинвазивная детекция стенозов коронарных артерий и атеросклеротических поражений; (4) морфологический анализ и классификация клеточных элементов крови. Во всех сценариях достигнута высокая точность и воспроизводимая когнитивная объяснимость решений, подтверждённая клинической апробацией [5, 13].

Список литературы

[1] Аверкин А.Н. Объяснительный искусственный интеллект // Государственный университет "Дубна". 30 лет в науке : Сборник научных трудов. Дубна : Университет Дубна, 2024. С. 314-321.

[2] L. Longo et al., "Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions", Information Fusion, 2024.

[3] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. 2017; Available from: http://arxiv.org/abs/1702.08608

[4] Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. Sensors (Basel). 2023 Jan 5;23(2):634. doi: 10.3390/s23020634. PMID: 36679430; PMCID: PMC9862413.

[5] Трофимов Ю.В., Семашко В.С. (науч. рук. Аверкин А.Н.) Нечёткие продукционные правила и нейросети глубокого обучения: объяснимый искусственный интеллект 2.0 для диагностики коронарных стенозов // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. СПб: Университет ИТМО, [2025]. URL: https://kmu.itmo.ru/digests/article/15055

[6] Talpur, N., Abdulkadir, S.J., Alhussian, H. et al. Deep Neuro-Fuzzy System application trends, challenges, and future perspectives: a systematic survey. Artif Intell Rev 56, 865–913 (2023). https://doi.org/10.1007/s10462-022-10188-3

[7] Wang L. X., Mendel J. M. Generating fuzzy rules by learning from examples //IEEE Transactions on systems, man, and cybernetics. 1992. Т. 22. №. 6. С. 1414-1427.

[8] Guillaume S. Designing fuzzy inference systems from data: An interpretability-oriented review //IEEE Transactions on fuzzy systems. 2001. Т. 9. №. 3. С. 426-443.

[9] Duţu L.C., Mauris G., Bolon P. A fast and accurate rule-base generation method for Mamdani fuzzy systems //IEEE Transactions on Fuzzy Systems. 2017. Т. 26. №. 2. С. 715-733.

[10] Chaudhry A. et al. Efficient lifelong learning with a-gem //arXiv preprint arXiv:1812.00420. 2018.

[11] Chaudhry A. et al. On tiny episodic memories in continual learning //arXiv preprint arXiv:1902.10486. 2019.

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

[13] Кузнецов Е.М., Трофимов Ю.В., Муравьев И.П. (науч. рук. Аверкин А.Н.) Объяснимый инкрементный подход в диагностике пневмонии: от свёрточных нейросетей до генеративных текстовых заключений // Сборник тезисов докладов конгресса молодых ученых. Электронное издание. СПб: Университет ИТМО, [2025]. URL: https://kmu.itmo.ru/digests/article/15266

[14] Прокопчина С.В. Новое направление в искусственном интеллекте: измерительный искусственный интеллект // Международная конференция по мягким вычислениям и измерениям. 2024. Т. 1. С. 3-6. EDN RTHBVL.