Компьютерные инструменты в образовании, 2023
№ 2: 41–61
УДК: 004.89:614.2
http://cte.eltech.ru
doi:10.32603/2071-2340-2023-2-41-61

ОБЪЯСНИТЕЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В МОДЕЛЯХ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ ЗДРАВООХРАНЕНИЯ 5.0*

Аверкин А. Н.[1], канд. физ.-мат. наук, доцент, ✉ averkin2003@inbox.ru,
orcid.org/0000-0003-1571-3583
Ярушев С. А.[1], канд. техн. наук, доцент, sergey.yarushev@icloud.com,
orcid.org/0000-0003-1352-9301

[1]Российский экономический университет имени Г. В. Плеханова,
Стремянный пер., 36, 117997, Москва, Россия

Аннотация

В основу пятой промышленной революции легла персонализация — персонализированные сервисы, умные устройства, роботы-помощники, а теперь и персонализированная медицина — направление, развиваемое в рамках философии Здравоохранения 5.0. В данной работе рассматриваются технологические аспекты применения моделей искусственного интеллекта нового поколения в задачах персонализированной медицины для Здравоохранения 5.0. Проанализированы возможности применения моделей объяснительного искусственного интеллекта в задачах здравоохранения. Проведена классификация методов объяснительного искусственного интеллекта (XAI), а также рассмотрены наиболее популярные алгоритмы XAI. Представлен обзор применения алгоритмов XAI в медицине, в котором рассмотрены задачи, конкретные алгоритмы и архитектуры искусственных нейронных сетей.

Ключевые слова: *объяснительный искусственный интеллект, XAI, искусственный интеллект, глубокое обучение, Здравоохранение 5.0, персонализированная медицина.*

Цитирование: Аверкина А. Н., Ярушева С. А. Объяснительный искусственный интеллект в моделях поддержки принятия решений для Здравоохранения 5.0 // Компьютерные инструменты в образовании. 2023. № 2. С. 41–61. doi:10.32603/2071-2340-2023-2-41-61

1. ВВЕДЕНИЕ

В настоящее время в совершенно разных областях жизнедеятельности человека все больше внимания уделяется персонализации продуктов, которые соответствуют его уникальным и особым требованиям. Пятая промышленная революция, известная также как Индустрия 5.0, открыла новые возможности для удовлетворения персонализированных

* Исследование выполнено за счет гранта Российского научного фонда № 22-71-10112, https://rscf.ru/project/22-71-10112/.

потребностей потребителей. Ранее Индустрия 4.0 предлагала массовую настройку, но это было недостаточно. Теперь продукты предоставляются потребителям в соответствии с их конкретными требованиями. Эта промышленная революция связана с взаимодействием между людьми и машинами, чтобы сделать работу лучше и быстрее. Качество продукции значительно повышается, обеспечивая безопасность и сокращая количество отходов.

Индустрия 5.0 также коснулась и отрасли здравоохранения. Здравоохранение 5.0 можно считать пятой промышленной революцией в области здравоохранения, которая обеспечивает «массовую персонализацию». Персонализированная медицина развивается большими темпами, разрабатываются персонализированные устройства, которые могут измерять различные параметры здоровья человека, такие как уровень сахара в крови, артериальное давление, пульс и др. Подобные персонализированные технологии позволяют докторам получать информацию о здоровье пациентов в режиме реального времени. Встраиваемый искусственный интеллект полностью меняет жизнь человека, позволяет изучать, как реагирует организм на те или иные условия.

В Здравоохранении 5.0 предполагается массовое применение технологий интернета вещей, связывающих миллионы различных датчиков, которые работают в сетях пятого поколения. Главной целью ставится обеспечение цифрового благополучия, внедрения интеллектуальных технологий на основе искусственного интеллекта и создание интеллектуального здравоохранения.

Технологии сетей пятого поколения и интернета вещей в сочетании с искусственным интеллектом формируют возможности для интеграции умных персональных устройств, которые позволят в режиме реального времени удаленно следить за здоровьем пациентов и оперативно оказывать помощь. Усовершенствованные устройства интернета вещей, подключенные к пациентам, собирают жизненно важные медицинские данные, отслеживают прогресс и диагностируют состояние здоровья [1, 2] без необходимости взаимодействия с пациентом. Алгоритмы искусственного интеллекта, такие как глубокие нейронные сети (DNN), позволяют обрабатывать огромные наборы данных, решать задачи распознавания изображений и текста и обеспечивают точное прогнозирование и обнаружение заболеваний [3].

Ограничения, которые существуют в традиционных моделях искусственного интеллекта, такие как отсутствие интерпретируемости и объяснимости полученных результатов, привели к мощному всплеску исследований в области объяснительного искусственного интеллекта. На рисунке 1 показан прогноз роста рынка объяснительного ИИ, сделанный агентством Next Move Strategy Consulting [4].

В области здравоохранения XAI применяется в различных моделях поддержки принятия клинических решений, для анализа медицинских данных, в задачах клинической диагностики, уменьшения систематической ошибки медицинских датчиков и классификации заболеваний [5].

В данной работе будут рассмотрены возможности применения объяснительного искусственного интеллекта в задачах здравоохранения пятого поколения.

2. РОЛЬ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СТАНОВЛЕНИИ ПЯТОЙ ПРОМЫШЛЕННОЙ РЕВОЛЮЦИИ

Первая промышленная революция началась в 1780 году с выработки механической энергии, за которой последовало применение электроэнергии на сборочных конвейерах. Позднее уже информационные технологии начали применяться для автоматизации де-

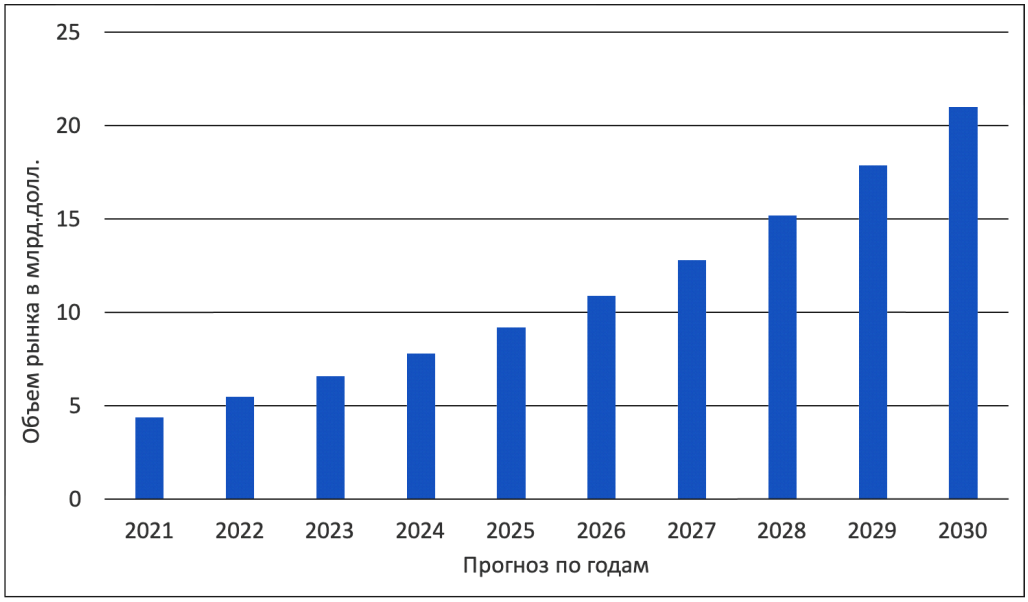

Рис. 1. Прогноз объема рынка объяснительного искусственного интеллекта до 2030 года

ятельности в производственной отрасли. Например, четвертая промышленная революция или Индустрия 4.0 привнесла Интернет вещей и облачные технологии для соединения виртуального и физического пространства, позже получившего название киберфизических систем [6, 7]. Несмотря на то, что стандарт 4.0 изменил обрабатывающую промышленность, оптимизация процессов ставила на второй план человеческие ресурсы, что привело к повышению безработицы. Поэтому пионеры отрасли с нетерпением ждут следующей революции, когда человеческий интеллект и машины будут интегрированы для совместного решения важных задач. Четвертая промышленная революция была направлена на преобразование производственных агентов в кибер-физические системы (КФС) из полноценных физических систем за счет эффективной интеграции бизнес-процессов и производства. Это включает в себя интеграцию всех субъектов в цепочке поставок обрабатывающей промышленности, от поставщиков до производственных линий и конечных пользователей с использованием интернета вещей (IoT) [8]. Индустрия 4.0 использует КФС для связи со всеми объектами через сеть IoT. В результате этого значительный объем накопленных данных сохраняется в облачной среде для их дальнейшей эффективной обработки. Индустрия 4.0 использует такие технологические концепции, как КФС, IoT, AI (искусственный интеллект), робототехника, облачные вычисления, аналитика больших данных, виртуальная реальность и кибербезопасность для достижения своей главной цели — умное производство (Smart Manufacturing) [9, 10]. Индустрия 4.0 позволила снизить затраты на производство, логистику и управление качеством за счет увеличения массового производства. И, тем не менее, Индустрия 4.0 повысила себестоимость производства, она проигнорировала человеческие затраты за счет оптимизации процессов. Это непреднамеренно приводит к обратному оттоку занятости и вызывает повышение безработицы, что препятствует более глубокому внедрению Индустрии 4.0 [11]. Таким образом, Индустрия 5.0 призвана решить эту проблему за счет более активного участия людей. Промышленные революции способствовали ускорению загрязнения окружающей среды, и вплоть до четвертой промышленной революции не разрабатывалось никаких решений для защиты окружающей среды. Таким образом, потребность в техноло-

гическом решении для обеспечения экологически чистых производственных процессов привела к следующей промышленной революции [12]. Индустрия 5.0, или пятая промышленная революция, обеспечивает устойчивость цивилизации за счет сокращения образования отходов благодаря внедрению биоэкономики, что приводит к поддержанию экологически чистой окружающей среды. Пятая промышленная революция сосредоточится на интеллектуальном производстве, вернув человеческий интеллект на производство, позволив роботам не заменять людей, а наоборот сотрудничать и помогать им. Индустрия 5.0 будет привлекать людей к совместной работе с роботами на заводах, тем самым используя человеческий интеллект и креативность, так необходимую для интеллектуальных процессов. Так как пятая промышленная революция только начинается и для нее еще нет единого общепризнанного определения, на сегодняшний день существует большое количество интерпретаций и определений пятой промышленной революции.

Рассмотрим несколько наиболее интересных описаний Индустрии 5.0:

1. Индустрия 5.0 — это первая промышленная эволюция, в центре которой находится человек. Она основывается на принципах 6R — Recognize, Reconsider, Realize, Reduce, Reuse and Recycle (Распознавать, Пересматривать, Реализовывать, Сокращать, Повторно использовать и Перерабатывать), а также на принципе цикличности, технологиях сокращения отходов, эффективном проектировании логистических систем для повышения уровня жизни, инновационных разработках и производстве высококачественных продуктов [13]. Данное определение дал Майкл Рада, один из основоположников Индустрия 5.0.

2. Пятая промышленная революция возвращает человеческую рабочую силу на завод, где человек и машина объединяются для повышения эффективности процесса за счет использования человеческого интеллекта и креативности и за счет интеграции рабочих процессов с интеллектуальными системами. [14].

3. Европейский экономический и социальный комитет заявляет, что новая промышленная революция, Индустрия 5.0, объединяет сильные стороны киберфизических производственных систем и человеческого интеллекта для создания синергетических фабрик [15]. Кроме того, для решения проблемы сокращения рабочей силы в результате Индустрии 4.0 в Индустрии 5.0 предлагается инновационный, этичный и ориентированный на человека подход.

4. Фридман и Хендри [16] предполагают, что Индустрия 5.0 вынуждает различных отраслевых практиков, специалистов по информационным технологиям и философов сосредоточиться на рассмотрении человеческого фактора при использовании технологий в промышленных системах.

5. Индустрия 5.0 — это ориентированное на человека решение, в котором идеальный человек-компаньон и коллаборативные роботы (коботы) сотрудничают с персоналом, чтобы обеспечить персонализируемое автономное производство через корпоративные социальные сети. Это, в свою очередь, позволяет человеку и машине работать рука об руку. Коботы не являются программируемыми машинами, но они могут чувствовать и понимать присутствие человека. В этом контексте коботы будут использоваться для повторяющихся задач и трудоемкой работы, в то время как человек позаботится о креативном и критическом мышлении (нестандартном мышлении).

В представлении производственных систем, ориентированных на человека, ключевая миссия цифровых технологий, описанная в рамках пятой промышленной революции, заключается в том, чтобы объяснить причины решений, принимаемых встроенными моделями искусственного интеллекта в системах промышленной автоматизации для

того, чтобы обеспечить взаимодействие человека и технологий в производственном цикле. Объяснительный искусственный интеллект (XAI) — один из подходов к решению этой проблемы. Цель XAI — предоставить алгоритмам искусственного интеллекта (ИИ) описательную функциональность — способность сообщить человеку об основных шагах, предпринятых для достижения решения. Еще один пример, проект Blackbox AI — автоматизированная система принятия решений, использующая машинное обучение для анализа больших данных, которая предоставляет функционал для прогнозирования поведенческих черт человека без раскрытия причины анализа. Проблемы, связанные с Blackbox AI, заключаются в отсутствии прозрачности работы самого ИИ. Иногда более быстрый производственный процесс может привести к перепроизводству и потерям товаров. Поэтому следует учитывать прозрачность реализации технологий. Необходимо внедрять объяснительный ИИ, чтобы повысить доверие к системам, в основе которых лежит искусственный интеллект с помощью объяснительных моделей искусственного интеллекта. Рисунок 2 демонстрирует применимость XAI для Индустрии 5.0, где цели XAI имеют различное назначение: XAI помогает конечным пользователям доверять решению ИИ, в то же время XAI позволяет инженерам или ученым полностью понять процесс работы системы ИИ.

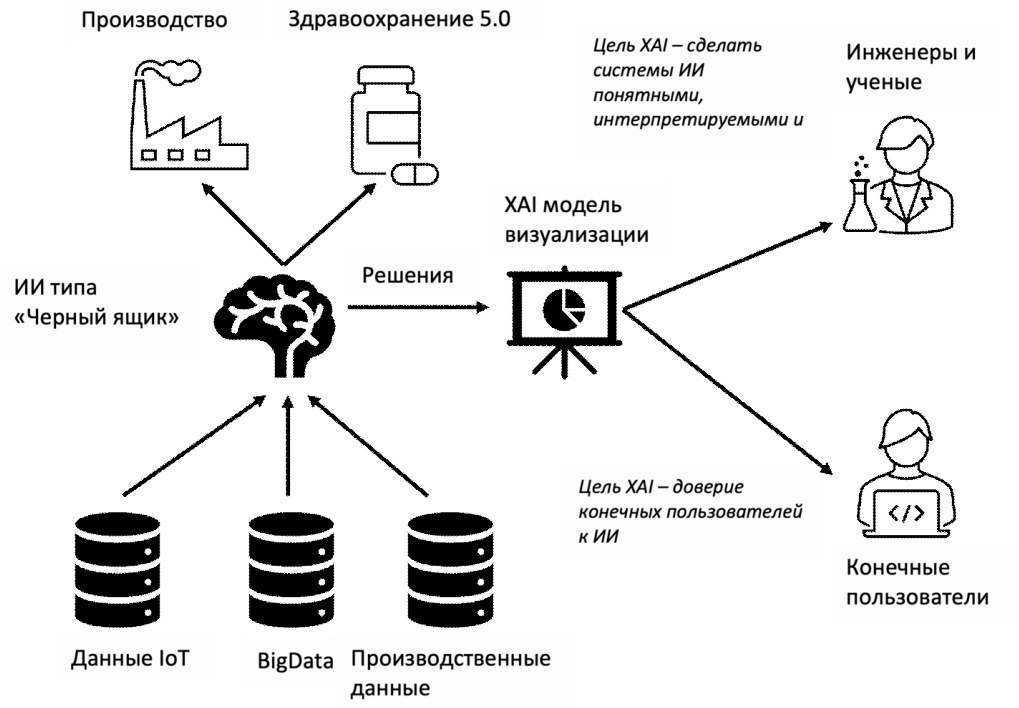

Рис. 2. Применимость XAI для Индустрии 5.0

3. ОСОБЕННОСТИ ОБЪЯСНИТЕЛЬНЫХ АЛГОРИТМОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В МЕДИЦИНЕ

Во всем мире активно ведутся работы не только в направлении разработки новых алгоритмов искусственного интеллекта, но и в направлении разработки требований и правовых основ для разрабатываемых алгоритмов искусственного интеллекта, в том числе

и для алгоритмов XAI в медицине. В Европейском Союзе были разработаны основные требования к моделям искусственного интеллекта в медицине [17], необходимые для повышения доверия к подобным алгоритмам:

1. **Надежный контроль и свобода для человеческой деятельности.** Подразумевается, что системы искусственного интеллекта должны позволять людям самостоятельно принимать обоснованные решения и укреплять их собственные права, а не ущемлять их. Искусственный интеллект должен расширять возможности людей, но не заменять их. Но также необходимо обеспечить надежный контроль за качеством выполняемых работ.

2. **Помехоустойчивость и безопасность.** Системы ИИ должны обладать высоким уровнем помехоустойчивости и безопасности. Системы, в том числе и медицинские, должны обеспечивать возможность разработки запасных планов действий, если основной план не сработал, они должны быть точными и легко воспроизводимыми вновь в случае сбоя.

3. **Конфиденциальность и управление данными.** Системы ИИ должны обеспечивать высокий уровень конфиденциальности и защиты данных, а также обладать всеми принципами эффективного управления данными.

4. **Прозрачность моделей ИИ.** Системы ИИ должны быть прозрачными, что является одним из основных принципов объяснительных моделей ИИ. Системы ИИ должны иметь такие способности объяснения, которые позволят адаптировать их для всех сторон — врачей и пациентов, профессионалов и простых пользователей. Системы должны информировать пользователей обо всех своих возможностях и ограничениях.

5. **Разнообразие, недискриминация и справедливость.** Системы ИИ должны избегать предвзятости, дискриминации и маргинации уязвимых групп населения. Разнообразие подразумевает доступность систем искусственного интеллекта для всех групп населения, независимо от их социального статуса и состояния здоровья.

6. **Социальное и экологическое благополучие.** Системы искусственного интеллекта должны приносить пользу всем людям, в том числе и будущим поколениям. Они должны способствовать улучшению экологии, взаимодействовать и оценивать окружающую среду и социальное воздействие.

7. **Подотчетность.** Необходимо внедрять механизмы для обеспечения полной подотчетности систем искусственного интеллекта. Необходима возможность аудита, оценки алгоритмов и данных, особенно это важно в таких критически значимых отраслях, как здравоохранение.

3.1. Категории методов XAI для применения в задачах здравоохранения

Методы XAI могут делиться по типам интерпретации результатов:

- **Модели с внутримодельным объяснением.** Модели данного типа структурированы и понятны. Модели могут давать ответы вместе с объяснениями на базе лингвистического интерпретатора [18].
- **Модели с пост-фактумным объяснением.** Объяснения получаются при помощи внешних методов анализа моделей искусственного интеллекта: методы обратного распространения, методы отображения активации классов и послойное распространение релевантности [19, 20].

4. ОСНОВНЫЕ МЕТОДЫ РАЗРАБОТКИ СИСТЕМ ОБЪЯСНИТЕЛЬНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Методы XAI можно разделить на несколько различных методов [21–24]. Рассмотрим основные из них:

1. **Модельно-независимые и модельно-зависимые методы.** Модельно-независимые методы — это те методы, которые не учитывают внутренние компоненты модели (то есть вес модели и структурные параметры). Следовательно, их можно применять к любой классической искусственной нейронной сети структуры «черного ящика». Напротив, модельно-зависимые методы определяются с использованием параметров отдельной модели, таких как интерпретация весов линейной регрессии или использование выведенных правил из дерева решений, которые будут специфичны для обученной модели [25]. У методов, не зависящих от модели, есть некоторые преимущества [26], такие как высокая гибкость для разработчиков в выборе любой модели искусственной нейронной сети для генерации и интерпретации, которая отличается от фактической модели «черного ящика», генерирующей решения.

2. **Локально-интерпретируемые и глобально-интерпретируемые методы.** В зависимости от объема пояснений предоставленные методы можно разделить на два класса: локальные и глобальные методы. Локальные интерпретируемые методы используют один результат или конкретные результаты предсказания и классификации модели для создания объяснений. Напротив, глобально интерпретируемые методы используют всю способность модели к выводу или общее поведение модели [27] для создания объяснений. В локально-интерпретируемых методах существенны только специфические признаки и характеристики. Для глобальных методов важность признаков можно использовать для объяснения общего поведения модели.

3. **Модели замещения (суррогатные модели) и модели визуализации.** Популярный способ объяснить модель черного ящика — применить интерпретируемую приблизительную модель, которая заменяет модель черного ящика для объяснения решений. Эта интерпретируемая приблизительная модель называется суррогатной моделью или моделью-заменителем, которая обучена аппроксимировать прогнозы черного ящика и позже используется для получения объяснений, интерпретирующих решения из модели черного ящика. Примером модели черного ящика может быть глубокая нейронная сеть, тогда как любая интерпретируемая модель может быть суррогатной моделью, такой как деревья решений или линейные модели. Помимо суррогатных моделей, модели визуализации предлагают визуальные объяснения и помогают генерировать объяснения в более презентабельной форме, показывая внутреннюю работу многих моделей, не зависящих от типа модели (например, какие пиксели или область пикселя помогают отличить кошку от собаки, или какие слова помогают решить, является ли документ спамом или нет).

4. **Стратегии предварительного моделирования, внутри-модельного моделирования и модели корректировки уже разработанной архитектуры нейронной сети.** XAI можно применять на протяжении всего процесса разработки модели нейронной сети. Цель предварительного моделирования объяснимости состоит в том, чтобы описать набор данных, с целью получения более полного представления о наборе данных, используемом для построения модели. Общие цели предварительного моделирования заключаются в обобщении данных, описании набора данных, разработке объяснимых признаков

и исследовательском анализе данных. Google Facets2 — это пример объяснений предварительного моделирования, которые позволяют изучать шаблоны из больших объемов данных. Напротив, цель внутри-модельного моделирования состоит в том, чтобы разработать объяснимые по своей сути модели, а не создавать модели черного ящика. Методологически существуют разные стратегии или способы построения внутри-модельных объяснений. Наиболее очевидным является разработка объяснимой по своей сути модели, такой как линейные модели, деревья решений и наборы правил. Однако необходимы специальные усилия для создания объяснений с использованием этих методов, таких как выбор важных функций. Предлагаются и другие подходы, выходящие за рамки объяснимых по своей сути моделей, такие как гибридные модели, совместное прогнозирование и объяснение, а также объяснимость посредством архитектурных корректировок. При гибридном подходе модели черного ящика сочетаются с объяснимыми моделями для разработки высокопроизводительной и объяснимой модели, такой как объединение глубоко скрытого слоя нейронной сети с методом k-ближайших соседей [28]. Кроме того, модель может быть обучена для совместного предоставления прогноза и соответствующего объяснения [29]. Идея здесь состоит в том, чтобы создать обучающий набор данных, в котором решение дополняется обоснованием. Наконец, объяснения посредством корректировки архитектуры сосредоточены на архитектуре глубокой сети для повышения объяснимости, например, использование фильтров более высокого уровня для представления части объекта, а не смеси шаблонов [30]. Данные внутри-модельные подходы XAI имеют два основных недостатка: во-первых, они предполагают наличие объяснений в обучающем наборе данных, что часто не так. Во-вторых, объяснения, генерируемые этими методами, не обязательно отражают то, как были сделаны прогнозы модели, а скорее то, что люди хотели бы видеть в качестве объяснения.Метод постмодельной объяснимости извлекает объяснения, которые по своей сути не объяснимы для описания предварительно разработанной модели. Эти популярные апостериорные методы XAI обычно работают с четырьмя ключевыми характеристиками: цель — что должно быть объяснено в отношении модели; драйвер — почему решение должно быть объяснено; семейство объяснений — как объяснение будет представлено пользователю; оценщик — вычислительный процесс, порождающий объяснение.

4.1. Классификация алгоритмов объяснительного искусственного интеллекта

Классификация алгоритмов объяснительного искусственного интеллекта представлена в таблице 1. Классификация произведена с точки зрения соответствующего типа XAI в различных таксономиях с типами алгоритмов ИИ, которые они могут объяснить.

Рассмотрим основные алгоритмы подробнее. Среди модельно-независимых стратегий XAI LIME является одним из наиболее популярных алгоритмов, которые объясняют полученный прогноз. LIME реализует линейную интерпретируемую модель, суррогатную (замещающую) модель в локальной области в качестве локального приближения для объяснения прогноза. Благодаря локальной аппроксимации LIME работает со всеми моделями черного ящика и типами данных (например, текстовые данные, табличные данные, изображения, графики).

Алгоритм SHAP [31] принципиально отличается от LIME с точки зрения того, каким образом получаются оценки важности, и в целом работает лучше, чем LIME. В SHAP вклад функции в решение оценивается значениями Шепли — классическим методом оценки предельного вклада. Однако агрегированные локальные прогнозы могут использоваться

Таблица 1. Классификация алгоритмов объяснительного искусственного интеллекта

Алгоритм XAI	**Методы разработки моделей XAI**				**Алгоритм обучения**
	Модельно-независимые и модельно-зависимые методы	**Локально-интерпретируемые и глобально-интерпретируемые**	**Замещающие методы (суррогатные модели) и модели визуализации**	**Стратегии предварительного моделирования, внутримодельного моделирования и модели корректировки**	
Facets	модельно-зависимы	оба метода	модель визуализации	предварительного моделирования	без учителя
LIME	модельно-зависимы	локально-интерпретируемые	суррогатная модель	модели корректировки	с учителем
SHAP	модельно-зависимы	оба метода	суррогатная модель	модели корректировки	с учителем
Counterfactual	оба метода	оба метода	суррогатная модель	модели корректировки	с учителем
LRP	модельно-независимы	локально-интерпретируемые	модель визуализации	модели корректировки	глубокое обучение
PIRL	модельно-независимы	глобально-интерпретируемые	суррогатная модель	внутримодельного моделирования	обучение с подкреплением
Hierarchical Policies	модельно-независимы	локально-интерпретируемые	суррогатная модель	внутримодельного моделирования	обучение с подкреплением
Structural Causal Model	модельно-независимы	локально-интерпретируемые	оба метода	модели корректировки	обучение с подкреплением

для создания глобальных объяснений. Также существуют алгоритмы оптимизации с точки зрения вычислительной сложности для SHAP, такие как TreeSHAP [32] и Deep SHAP, но они не зависят от модели.

Counterfactual — еще один алгоритм, доступный как для модельно-независимых [33], так и для модельно-зависимых [34] вариантов. Алгоритм Counterfactual основан на объяснении прогноза алгоритма предиктора путем нахождения малейшего изменения значений входных признаков, вызывающего изменение исходного прогноза. Например, если изменение индекса массы тела человека привело к изменению первоначального прогноза с болезни на выздоровление, то использование значения индекса массы тела является

ориентировочным объяснением для корреляции с исходным прогнозом, таким образом подразумевая подходящие для человека объяснения. Объяснение получается простое, но с несколькими возможными объяснениями.

Layerwise Relevance Propagation (LRP) [35] — это алгоритм, разработанный для объяснения глубоких нейронных сетей с предположением, что классификатор может быть разложен на разные уровни, что делает его модельно-зависимым методом. LRP алгоритм разработан с учетом того фактора, что определенные уровни входных данных имеют отношение к прогнозу. Кроме того, чтобы получить представление о том, какие нейроны являются значимыми, оценки активации каждого нейрона учитываются посредством обратного прохода и в конечном итоге дают информацию о входных данных. Алгоритм наиболее часто применяется к данным изображения, чтобы выделить значимые пиксели, которые позволяют сделать определенный прогноз.

Объяснимое обучение с подкреплением (XRL) — многообещающая, но сложная исследовательская ветвь XAI [36], поскольку модель обучения с подкреплением (RL) часто содержит огромное количество состояний. Тем не менее, XRL может ускорить процесс проектирования RL, облегчая разработчикам отладку систем. A. Heuillet и др. [37] представили строгую классификацию методов XRL, основанную на типах объяснений (текст или изображения), на уровне объяснения (локальном, если это было для прогнозов, или глобальном, если объяснялась вся модель), а также какой алгоритм является объясняющим. В частности, можно выделить три следующих алгоритма XRL: программно-интерпретируемое обучение с подкреплением (PIRL) является примером глобально-интерпретируемого метода [38]. Идея PIRL заключается в использовании языка программирования, гораздо более близкого к человеческому мышлению для имитации поведения модели глубокого обучения с подкреплением (DRL). PIRL использует структуру под названием Neurally Directed Program Search (NDPS) для изучения поведения модели DRL путем имитации обучения. Таким образом, существует два шага: построение DRL и извлечение знаний о модели DRL для создания последовательности действий. Прогнозы, получаемые алгоритмом PIRL, не такие точные, как у нейронной сети, но они могут быть довольно близкими по точности и гораздо более понятными. Модель PIRL успешно применялась в симуляторе Open Racing Car Simulator (TORCS) [39].

Иерархическое и интерпретируемое приобретение навыков в многозадачном обучении с подкреплением (RL) [40] является примером локально-интерпретируемой модели. Этот подход состоит в представлении модели с высокоуровневыми действиями в виде последовательности более простых действий, поскольку она более знакома людям. Этот подход использовался для игры в Minecraft, и он реализует иерархическую модель, основанную на двух уровнях, с использованием алгоритма актер-критик, он же использовался для объяснения многозадачной модели RL, играющей в Minecraft. Модель также использует модель стохастической временной грамматики, чтобы зафиксировать отношения между действиями для создания иерархической модели. Люди просто говорят, например, припарковать автомобиль, вместо того чтобы определять все действия, связанные с рулем, сцеплением, акселератором и тормозом. Точно так же, прежде чем переместить объект, его нужно найти. Этот метод представляет инструкции высокого уровня, такие как «Синий стек», где «Стек синий» состоит из «Найти / Получить / Положить / Синий», а в то же время «Найти синий» состоит из множества менее низкоуровневых «Идти налево, Двигаться вперед, Повернуть направо...».

Структурно-причинные модели (SCM) — это очень четкий способ представления причинно-следственных связей событий. P. Madumal et al. [41] предложили структуру, которая попадает в категорию модели корректировок, основанную на SCM для объяснения

поведения немодальных агентов RL. В данном случае используется ориентированный ациклический граф (DAG), в котором узлы представляют состояния, а ребра — действия. Проходя по графу, можно наблюдать, какие действия переходят из одного состояния в другое. Процесс состоит из трех основных этапов: создание DAG; использование моделей многомерной регрессии для аппроксимации взаимосвязей с использованием минимального количества переменных; создание объяснений путем анализа переменных DAG, чтобы ответить на вопросы: «Почему действие A?» и «Почему не действие B?». В своем исследовании они создают модель на основе случайных структур для оценки шести доменов с использованием шести различных алгоритмов RL для игры в Starcraft II. Ученые также провели исследование, в котором группа из 120 человек оценила качественные объяснения, и было обнаружено, что люди, участвовавшие в исследовании, предпочли свои объяснения, основанные на случайных моделях, другим базовым показателям.

5. ОБЗОР ИССЛЕДОВАНИЙ ПО ПРИМЕНЕНИЮ АЛГОРИТМОВ XAI В МЕДИЦИНЕ

В таблице 2 представлен обзор исследований по применению рассмотренных алгоритмов объяснительного искусственного интеллекта в задачах здравоохранения. Исследования рассматривались в зависимости от используемого алгоритма объяснения, объект исследования в данном случае — изображения каких органов человека обрабатывались, каким методом были получены данные, к примеру, рентгеновские снимки, ультразвук или компьютерная томография. Также дано краткое описание исследования.

Таблица 2

Обследуемый орган	**Архитектура нейронной сети**	**Алгоритм XAI**	**Метод получения данных**	**Краткое описание исследования**
Кость	Сверточная нейронная сеть	CAM	Рентген	В работе разработана модель прогнозирования степени повреждения коленного сустава и уровня степени боли с помощью рентгеновского изображения [42].
Легкое	VGG-16 и VGG-CAM	CAM	Ультразвуковое исследование (УЗИ), рентген	В качестве обучающих наборов данных исследователи использовали три вида ультразвуковых изображений легких и две нейронные сети, VGG-16 и VGG-CAM для классификации пневмонии на три типа [43].
Легкое	MobileNet и ResNet	CAM	Рентген, компьютерная томография (КТ)	В данной работе исследователи использовали две модели нейронных сетей, первая основана на алгоритме MobileNet для классификации изображений рентгенограммы грудной клетки COVID-19, другая нейронная сеть — ResNet для классификации изображений КТ [44].

Обследуемый орган	Архитектура нейронной сети	Алгоритм XAI	Метод получения данных	Краткое описание исследования
Легкое	DRE-Net	CAM	КТ	В исследовании используются данные, полученные при помощи КТ о здоровых пациентах и о пациентах с COVID-19 для обучения модели нейронной сети DRE-Net [45].
Легкое		Grad-CAM	КТ	В работе рассматривается метод глубокого слияния признаков на изображениях КТ. Метод обеспечивает лучшую производительность, чем классические сверточные нейронные сети [46].
Легкое	Discrimination-DL и the Localization-DL в связке с ResNet	Grad-CAM	Рентген	Предлагаемая в работе система автоматизированного обнаружения состоит из двух этапов с DL (deep learning): Discrimination-DL и Localization-DL. Первый алгоритм DL был разработан для извлечения признаков легких из рентгенограмм грудной клетки для распознавания COVID-19 и обучен с использованием 3548 рентгенограмм грудной клетки. Вторая нейросеть DL была обучена с 406-пиксельными участками и применена к распознанным рентгенограммам, чтобы локализовать и отнести их к левому легкому, правому легкому или бипульмональному типу. В качестве объяснительного алгоритма используется Grad-CAM для обнаружения аномальных областей на изображении [47].
Толстый кишечник	DenseNet121	Grad-CAM	Колоноскопия	Исследователи используют нейронную сеть DenseNet121 для прогнозирования наличия у пациента язвенного колита [48].
Легкое	Res2Net	Grad-CAM	КТ	В работе представлена модель классификации на основе нейронной сети Res2Net. В исследовании используется метод Grad-CAM для повышения интерпретируемости общей системы совместной классификации и сегментации [49].

Обследуемый орган	Архитектура нейронной сети	Алгоритм XAI	Метод получения данных	Краткое описание исследования
Грудь	NAGNN	Grad-CAM	КТ	В данном исследовании предлагается нейронная сеть с учетом соседнего графа (neighboring aware graph neural network NAGNN) для обнаружения очагов COVID-19 на основе изображений компьютерной томографии грудной клетки [50].
Легкое	COVID-CXNet (на основе нейронной сети CheXNet)	Grad-CAM, LIME	Рентгеновский	В данной работе представлен набор рентгеновских данных COVID-19 и предлагается разработанная нейронная сеть под названием COVID-CXNet, основанная на нейронной сети CheXNet с использованием трансферного обучения [51].
Легкое	NASNetLarge	Grad-CAM, LIME	Рентген, КТ	В исследовании проводится сравнительный анализ пяти моделей глубокого обучения и используется метод визуализации для объяснения результатов работы нейронной сети NASNetLarge [52].
Грудь	A3 Net	Attention	Рентгеновский	В данной работе разработана модель обучения тройному вниманию A3 Net для диагностики 14 заболеваний грудной клетки [53].
Легкое, кожа	CNN	Saliency	КТ, рентген	В работе представлены показатели количественной оценки значимости объяснительных алгоритмов ИИ [54].
—	CNN	SHAP	Электронная медицинская карта (ЭМК)	В исследовании представлена объяснимая медицинская система поддержки принятия решений, которая помогает врачам на ранних стадиях обнаруживать у женщин риск развития сахарного диабета [55].
—		SHAP	Радиомика	В исследовании представлена система для анализа медицинских изображений с помощью радиомики [56].
Легкое		SHAP	КТ	В данной статье представлена модель прогнозирования мутаций у пациентов с немелкоклеточным раком легкого [57].

Обследуемый орган	Архитектура нейронной сети	Алгоритм XAI	Метод получения данных	Краткое описание исследования
Грудь		LIME, SHAP	Рентген	В данной работе исследователи предлагают фреймворк для повышения качества объяснений работы сверхточной нейронной сети. В работе использовались сразу два метода объяснения — LIME и SHAP [58].
Легкое		SHAP, LIME, Scoped Rules	ЭМК	Исследователи проводят сравнение трех методов XAI на наборе данных электронных медицинских карт. В результате исследований был сделан вывод о невозможности полной замены людей данными методами [59].
Грудь		Image caption	КТ	В работе предложена модель Medical Visual Language BERT (Medical-VLBERT) для создания отчетов о компьютерной томографии пациентов с COVID-19 [60].

В основном, исследователи применяют объяснительные алгоритмы для решения задач распознавания медицинских снимков, полученных разными методами, такими как рентген, компьютерная томография или УЗИ. Но некоторые исследователи применяют XAI алгоритмы для разработки моделей поддержки принятия решений. В подобных задачах алгоритмы анализируют электронные медицинские карты пациентов, оценивают ранее сдаваемые медицинские анализы и выявляют риски развития того или иного заболевания.

6. ЗАКЛЮЧЕНИЕ

В данной работе рассмотрены модели объяснительного искусственного интеллекта, применяемые в задачах персонализированной медицины в рамках Здравоохранения 5.0. Представлен прогноз объема рынка объяснительного искусственного интеллекта, как показывает прогноз, к 2030 году объемы рынка XAI превысят 20 миллиардов долларов. В работе также были проанализированы возможности применения моделей объяснительного искусственного интеллекта в задачах здравоохранения. Классифицированы методы объяснительного искусственного интеллекта (XAI), а также рассмотрены наиболее популярные алгоритмы XAI, такие как Facets, LIME, SHAP, Counterfactual, LRP, PIRL, HierarchicalPolicies, StructuralCausal Model, CAM, Grad-CAM. Также представлен обзор применения алгоритмов XAI в медицине, в котором рассмотрены задачи, конкретные алгоритмы и архитектуры искусственных нейронных сетей.

Список литературы

1. *Pasluosta C. F., Gassner H., Winkler J., Klucken J., Eskofier B. M.* An emerging era in the management of parkinson's disease: Wearable technologies and the internet of things // IEEE Journal of Biomedical and Health Informatics. 2015. Vol. 19, № 6. P. 1873–1881. doi:10.1109/JBHI.2015.2461555

2. *Laplante P. A., Laplante N.* The internet of things in healthcare: Potential applications and challenges // IT Professional. 2016. Vol. 18, № 3. P. 2–4. doi:10.1109/MITP.2016.42
3. *Mohanta B., Das P., Patnaik S.* Healthcare 5.0: A paradigm shift in digital healthcare system using artificial intelligence, iot and 5g communication // 2019 Int. Conf. on Applied Machine Learning (ICAML), Bhubaneswar, India. 2019. P. 191–196. doi:10.1109/ICAML48257.2019.00044
4. Next Move Strategy Consulting. Explainable AI (XAI) Market (2021 to 2030). 2022. URL: https://www.nextmsc.com/report/explainable-ai-Market (date: 22.06.2023).
5. *Shaban-Nejad A., Michalowski M., Buckeridge D. L.* Explainable AI in Healthcare and Medicine / Studies in Computational Intelligence. Springer Cham, 2021. doi:10.1007/978-3-030-53352-6
6. *Gao J., Liu N., Lawley M., Hu X.* An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums // Journal of Healthcare Engineering. 2017. Vol. 2017. P. 1–12. doi:10.1155/2017/2460174
7. *Yang S. C.-H., Shafto P.* Explainable Artificial Intelligence via Bayesian Teaching // Proc. of the 31st Conf. on Neural Information Processing Systems Workshop on Teaching Machines, Robots and Humans. 2017. P. 127–137.
8. *Sharma P. K., Kumar N., Park J. H.* Blockchain-Based Distributed Framework for Automotive Industry in a Smart City // IEEE Transactions on Industrial Informatics. 2019. Vol. 15, № 7. P. 4197–4205. doi:10.1109/tii.2018.2887101
9. *He D., Ma M., Zeadally S., Kumar N., Liang K.* Certificateless Public Key Authenticated Encryption With Keyword Search for Industrial Internet of Things // IEEE Transactions on Industrial Informatics. 2018. Vol. 14, № 8. P. 3618–3627. doi:10.1109/tii.2017.2771382
10. *Khan I. H., Javaid M.* Role of Internet of Things (IoT) in Adoption of Industry 4.0 // Journal of Industrial Integration and Management. 2021. Vol. 07, № 04. P. 515–533. doi:10.1142/s2424862221500068
11. *Kim J. H.* A Review of Cyber-Physical System Research Relevant to the Emerging IT Trends: Industry 4.0, IoT, Big Data, and Cloud Computing // Journal of Industrial Integration and Management. 2017. Vol. 02, № 03, p. 1750011.
12. *Chen H.* Theoretical Foundations for Cyber-Physical Systems: A Literature Review // Journal of Industrial Integration and Management. 2017. Vol. 02, № 03. P. 1750013. doi:10.1142/s2424862217500130
13. *Demir K. A., Döven G., Sezen B.* Industry 5.0 and Human-Robot Co-working // Procedia Computer Science. 2019. Vol. 158. P. 688–695. doi:10.1016/j.procs.2019.09.104
14. *Xu L. D., Xu E. L., Li L.* Industry 4.0: state of the art and future trends // International Journal of Production Research. 2018. Vol. 56, № 8. P. 2941–2962. doi:10.1080/00207543.2018.1444806
15. *Rada M.* Industry 5.0 definition. URL: https://michael-rada.medium.com/industry-5-0-definition-6a2f9922dc48 (date: 22.06.2023).
16. *Nahavandi S.* Industry 5.0—A Human-Centric Solution // Sustainability. 2019. Vol. 11, № 16. P. 4371. doi:10.3390/su11164371
17. *Smuha N. A.* The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence // Computer Law Review International. 2019. Vol. 20, № 4. P. 97–106. doi:10.9785/cri-2019-200402
18. *Wu J., Mooney R. J.* Faithful Multimodal Explanation for Visual Question Answering // arXiv:1809.02805, 2018.
19. *Simonyan K., Vedaldi A., Zisserman A.* Deep inside convolutional networks: Visualising image classification models and saliency maps // arXiv:1312.6034, 2013.
20. *Bach S., Binder A., Montavon G., Klauschen F., Müller K.-R., Samek W.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation // PLOS ONE. 2015. Vol. 10, № 7. P. e0130140. doi:10.1371/journal.pone.0130140
21. Многозначные логики и их применения. Т. 2. Логики в системах искусственного интеллекта / Под ред. В. К. Финна. М.: Изд-во ЛКИ, 2008.
22. *Stiglic G., Kocbek P., Fijacko N., Zitnik M., Verbert K., Cilar L.* Interpretability of machine learning-based prediction models in healthcare // WIREs Data Mining and Knowledge Discovery. 2020. Vol. 10, № 5. doi:10.1002/widm.1379
23. *Arya V. et al.* One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. Vol. 2 // arXiv:1909.03012 [Preprint], 2019.

24. *Vilone G., Longo L.* Explainable artificial intelligence: a systematic review. Vol. 4 // arXiv:2006.00093 [Preprint], 2020.
25. *Singh A., Sengupta S., Lakshminarayanan V.* Explainable Deep Learning Models in Medical Image Analysis // Journal of Imaging. 2020. Vol. 6, № 6. P. 52. doi:10.3390/jimaging6060052
26. *Arras L., Osman A., Müller K.-R., Samek W.* Evaluating recurrent neural network explanations // arXiv: 1904.11829, 2019.
27. *Ribeiro M. T., Singh S., Guestrin C.* Model-agnostic interpretability of machine learning // arXiv:1606. 05386, 2016.
28. *Samek W., Müller K.-R.* Towards explainable artificial intelligence // Explainable AI: interpreting, explaining and visualizing deep learning. Springer Cham, 2019. P. 5–22.
29. *Papernot N., McDaniel P.* Deep *k*-nearest neighbors: Towards confident, interpretable and robust deep learning // arXiv:1803.04765, 2018.
30. *Hind M. et al.* Ted: Teaching ai to explain its decisions // Proc. of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. 2019. P. 123–129. doi:10.1145/3306618.3314273
31. *Zhang Q., Wu Y. N., Zhu S.-C.* Interpretable convolutional neural networks // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. P. 8827–8836. doi:10.1109/cvpr.2018.00920
32. *Lundberg S. M., Lee S.-I.* A unified approach to interpreting model predictions // Proc. of the 31st international conference on neural information processing systems. 2017. P. 4768–4777.
33. *Lundberg S. M., Erion G. G., Lee S.-I.* Consistent individualized feature attribution for tree ensembles // arXiv:1802.03888, 2018.
34. *Sharma S., Henderson J., Ghosh J.* Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models // arXiv:1905.07857, 2019.
35. *Wachter S., Mittelstadt B., Russell C.* Counterfactual explanations without opening the black box: Automated decisions and the gdpr // Harv. JL & Tech. 2017. Vol. 31. P. 841.
36. *Montavon G., Binder A., Lapuschkin S., Samek W., Muller K.-R.* Layer-wise relevance propagation: an overview // Explainable AI: interpreting, explaining and visualizing deep learning. 2019. P. 193–209.
37. *Heuillet A., Couthouis F., Díaz-Rodríguez N.* Explainability in deep reinforcement learning // Knowledge-Based Systems. 2021. Vol. 214, P. 106685. doi:10.1016/j.knosys.2020.106685
38. *Verma A., Murali V., Singh R., Kohli P., Chaudhuri S.* Programmatically interpretable reinforcement learning // Proc. of. International Conference on Machine Learning. PMLR, 2018. P. 5045–5054.
39. *Wymann B., Espié E., Guionneau C., Dimitrakakis C., Coulom R., Sumner A.* Torcs, the open racing car simulator // http://torcs.sourceforge.net, 2000. Vol. 4, № 6. P. 2.
40. *Shu T., Xiong C., Socher R.* Hierarchical and interpretable skill acquisition in multi-task reinforcement learning // arXiv:1712.07294 [Preprint], 2017.
41. *Madumal P., Miller T., Sonenberg L., Vetere F.* Explainable reinforcement learning through a causal lens // Proc. of the AAAI Conference on Artificial Intelligence. 2020. Vol. 34, № 03. P. 2493–2500.
42. *Pierson E., Cutler D. M., Leskovec J., Mullainathan S., Obermeyer Z.* An algorithmic approach to reducing unexplained pain disparities in underserved populations // Nature Medicine. 2021. Vol. 27, № 1. P. 136–140. doi:10.1038/s41591-020-01192-7
43. *Born J. et al.* Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis // Applied Sciences. 2021. Vol. 11, № 2, p. 672. doi:10.3390/app11020672
44. *Jia G., Lam H.-K., Xu Y.* Classification of COVID-19 chest X-Ray and CT images using a type of dynamic CNN modification method // Computers in Biology and Medicine. 2021. Vol. 134. P. 104425. doi:10.1016/j.compbiomed.2021.104425
45. *Song Y. et al.* Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images // IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2021. Vol. 18, № 6. P. 2775–2780. doi:10.1109/tcbb.2021.3065361
46. *Wang S.-H., Govindaraj V. V., Górriz J. M., Zhang X., Zhang Y.-D.* Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network // Information Fusion. 2021. Vol. 67. P. 208–229. doi:10.1016/j.inffus.2020.10.004
47. *Wang Z. et al.* Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays // Pattern Recognition. 2021. Vol. 110. P. 107613. doi:10.1016/j.patcog.

2020.107613
48. *Sutton R. T., Zaïane O. R., Goebel R., Baumgart D. C.* Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images // Scientific Reports. 2022. Vol. 12, № 1. doi:10.1038/s41598-022-06726-2
49. *Wu Y.-H. et al.* JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation // IEEE Transactions on Image Processing. 2021. Vol. 30. P. 3113–3126. doi:10.1109/tip.2021.3058783
50. *Lu S., Zhu Z., Gorriz J. M., Wang S., Zhang Y.* NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network // Int. J. Intell. Syst. 2022. Vol. 37, № 2. P. 1572–1598. doi:10.1002/int.22686
51. *Haghanifar A., Majdabadi M. M., Choi Y., Deivalakshmi S., Ko S.* COVID-CXNet: Detecting COVID-19 in frontal chest X-ray images using deep learning // Multimedia Tools and Applications. 2022. Vol. 81, № 21. P. 30615–30645. doi:10.1007/s11042-022-12156-z
52. *Punn N. S., Agarwal S.* Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks // Applied Intelligence. 2020. Vol. 51, № 5. P. 2689–2702. doi:10.1007/s10489-020-01900-3
53. *Wang H., Wang S., Qin Z., Zhang Y., Li R., Xia Y.* Triple attention learning for classification of 14 thoracic diseases using chest radiography // Medical Image Analysis. 2021. Vol. 67. P. 101846. doi:10.1016/j.media.2020.101846
54. *Hu B., Vasu B., Hoogs A.* X-MIR: EXplainable Medical Image Retrieval // Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022. 2022. P. 440–450.
55. *Du Y., Rafferty A. R., McAuliffe F. M., Wei L., Mooney C.* An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus, // Scientific Reports. 2022. Vol. 12, № 1. P. 1170. doi:10.1038/s41598-022-05112-2
56. *Severn C., Suresh K., Görg C., Choi Y. S., Jain R., Ghosh D.* A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features // Sensors. 2022. Vol. 22, № 14. P. 5205. doi:10.3390/s22145205
57. *Le N. Q. K., Kha Q. H., Nguyen V. H., Chen Y.-C., Cheng S.-J., Chen C.-Y.* Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer // International Journal of Molecular Sciences. 2021. Vol. 22, № 17. P. 9254. doi:10.3390/ijms22179254
58. *Abeyagunasekera S. H. P., Perera Y., Chamara K., Kaushalya U., Sumathipala P., Senaweera O.* LISA: Enhance the explainability of medical images unifying current XAI techniques // Proc. of 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Apr. 2022. 2022. doi:10.1109/i2ct54291.2022.9824840
59. *Duell J., Fan X., Burnett B., Aarts G., Zhou S.-M.* A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records // Proc. of 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Jul. 2021. 2021. doi:10.1109/bhi50953.2021.9508618
60. *Liu G. et al.* Medical-VLBERT: Medical Visual Language BERT for COVID-19 CT Report Generation With Alternate Learning // IEEE Transactions on Neural Networks and Learning Systems. 2021. Vol. 32, № 9. P. 3786–3797. doi:10.1109/tnnls.2021.3099165

Поступила в редакцию 10.06.2023, окончательный вариант — 22.06.2023.

Аверкин Алексей Николаевич, кандидат физико-математических наук, доцент, ведущий научный сотрудник научного центра перспективных исследований в искусственном интеллекте РЭУ им. Г. В. Плеханова, Москва, ✉ averkin2003@inbox.ru

Ярушев Сергей Александрович, кандидат технических наук, доцент, директор научного центра перспективных исследований в искусственном интеллекте РЭУ им. Г. В. Плеханова, Москва, sergey.yarushev@icloud.com

Computer tools in education, 2023
№ 2: 41–61
http://cte.eltech.ru
doi:10.32603/2071-2340-2023-2-41-61

Explainable Artificial Intelligence in Decision Support Models for Healthcare 5.0

Averkin A. N.[1], Cand. Sc., Associate Professor, ✉ averkin2003@inbox.ru, orcid.org/0000-0003-1571-3583
Yarushev S. A.[1], Cand. Sc., Associate Professor, sergey.yarushev@icloud.com, orcid.org/0000-0003-1352-9301

[1]Russian University of Economics. G. V. Plekhanov, 36 Stremyanny lane, 117997, Moscow, Russia

Abstract

Industry 5.0 was based on personalization — personalized services, smart devices, assistant robots, and now personalized medicine, a direction developed within the framework of the Healthcare 5.0 philosophy. This paper discusses the technological aspects of the application of new generation artificial intelligence models in the tasks of personalized medicine for Healthcare 5.0. The possibilities of using explanatory artificial intelligence models in healthcare tasks are analyzed. The classification of explainable artificial intelligence (XAI) methods is carried out, and the most popular XAI algorithms are considered. It also provides an overview of the application of XAI algorithms in medicine, which considers tasks, specific algorithms and architectures of artificial neural networks.

Keywords: *explainable artificial intelligence, XAI, artificial intelligence, deep learning, Healthcare 5.0, personalized medicine.*

Citation: A. N. Averkin and S. A. Yarushev, “Explainable Artificial Intelligence in Decision Support Models for Healthcare 5.0,” *Computer tools in education*, no. 2, pp. 41–61, 2023 (in Russian); doi:10.32603/2071-2340-2023-2-41-61

References

1. C. F. Pasluosta, H. Gassner, J. Winkler, J. Klucken, and B. M. Eskofier, “An emerging era in the management of parkinson’s disease: Wearable technologies and the internet of things,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1873–1881, 2015; doi:10.1109/JBHI.2015.2461555
2. P. A. Laplante and N. Laplante, “The internet of things in healthcare: Potential applications and challenges,” *IT Professional*, vol. 18, no. 3, pp. 2–4, 2016; do:10.1109/MITP.2016.42
3. B. Mohanta, P. Das, and S. Patnaik, “Healthcare 5.0: A paradigm shift in digital healthcare system using artificial intelligence, iot and 5g communication,” in *2019 Int. Conf. on Applied Machine Learning (ICAML), Bhubaneswar, India*, pp. 191–196, 2019; doi:10.1109/ICAML48257.2019.00044
4. Next Move Strategy Consulting, “Explainable AI (XAI) Market (2021 to 2030),” in *www.nextmsc.com*, 2022. [Online]. Available: https://www.nextmsc.com/report/explainable-ai-Market
5. A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, eds., “Explainable AI in Healthcare and Medicine,” in *Studies in Computational Intelligence*, Springer Cham, 2021; doi:10.1007/978-3-030-53352-6
6. J. Gao, N. Liu, M. Lawley, and X. Hu, “An Interpretable Classification Framework for Information Extraction from Online Healthcare Forums,” *Journal of Healthcare Engineering*, vol. 2017, pp. 1–12, 2017; doi:10.1155/2017/2460174
7. S. C.-H. Yang and P. Shafto, “Explainable Artificial Intelligence via Bayesian Teaching,” in *Proc. of the 31st Conf. on Neural Information Processing Systems Workshop on Teaching Machines, Robots and Humans*, pp. 127–137, 2017.

8. P. K. Sharma, N. Kumar, and J. H. Park, "Blockchain-Based Distributed Framework for Automotive Industry in a Smart City," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 4197–4205, 2019; doi:10.1109/tii.2018.2887101
9. D. He, M. Ma, S. Zeadally, N. Kumar, and K. Liang, "Certificateless Public Key Authenticated Encryption With Keyword Search for Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 8, pp. 3618–3627, 2018; doi:10.1109/tii.2017.2771382
10. I. H. Khan and M. Javaid, "Role of Internet of Things (IoT) in Adoption of Industry 4.0," *Journal of Industrial Integration and Management*, vol. 07, no. 04, pp. 515–533; 2021; doi:10.1142/s2424862221500068
11. J. H. Kim, "A Review of Cyber-Physical System Research Relevant to the Emerging IT Trends: Industry 4.0, IoT, Big Data, and Cloud Computing," *Journal of Industrial Integration and Management*, vol. 02, no. 03, p. 1750011, 2017; doi:10.1142/s2424862217500117
12. H. Chen, "Theoretical Foundations for Cyber-Physical Systems: A Literature Review," *Journal of Industrial Integration and Management*, vol. 02, no. 03, p. 1750013, 2017; doi:10.1142/s2424862217500130
13. K. A. Demir, G. Döven, and B. Sezen, "Industry 5.0 and Human-Robot Co-working," *Procedia Computer Science*, vol. 158, pp. 688–695, 2019; doi:10.1016/j.procs.2019.09.104
14. L. D. Xu, E. L. Xu, and L. Li, "Industry 4.0: state of the art and future trends," *International Journal of Production Research*, vol. 56, no. 8, pp. 2941–2962, 2018; doi:10.1080/00207543.2018.1444806
15. M. Rada, "Industry 5.0 definition," in *https://michael-rada.medium.com 2020*. [Online]. Available: https://michael-rada.medium.com/industry-5-0-definition-6a2f9922dc48
16. S. Nahavandi, "Industry 5.0 — A Human-Centric Solution," *Sustainability*, vol. 11, no. 16, p. 4371, 2019; doi:10.3390/su11164371
17. N. A. Smuha, "The EU Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," *Computer Law Review International*, vol. 20, no. 4, pp. 97–106, 2019; doi:10.9785/cri-2019-200402
18. J. Wu and R. J. Mooney, "Faithful Multimodal Explanation for Visual Question Answering," in *arXiv:1809.02805*, 2018.
19. K. Simonyan and A.Vedaldi, "Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv:1312.6034*, 2013.
20. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, 2015; doi:10.1371/journal.pone.0130140
21. V. K. Finn et al., eds., "Logic in artificial intelligence systems," *Multivalued logics and their applications*, vol. 2, Moscow: Publishing house LKI, 2008 (in Russian).
22. G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 5, 2020; doi:10.1002/widm.1379
23. V. Arya et al., "One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques," v. 2, in *arXiv:1909.03012* [Preprint], 2019.
24. G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," v. 4, in *arXiv:2006.00093* [Preprint], 2020.
25. A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020; doi:10.3390/jimaging6060052
26. L. Arras, A. Osman, K.-R. Müller, and W. Samek, "Evaluating recurrent neural network explanations," in *arXiv:1904.11829*, 2019.
27. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," in *arXiv:1606.05386*, 2016.
28. W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer Cham, pp. 5–22, 2019.
29. N. Papernot and P. McDaniel, "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning," in *arXiv:1803.04765*, 2018.
30. M. Hind et al, "Ted: Teaching ai to explain its decisions," in *Proc. of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 123–129, 2019. doi:10.1145/3306618.3314273
31. Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018. doi:10.1109/cvpr.2018.00920
32. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.
33. S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," in *arXiv:1802.03888*, 2018.
34. S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency,

interpretability, and fairness of artificial intelligence models," in *arXiv:1905.07857*, 2019.
35. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.
36. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Muller, "Layer-wise relevance propagation: an overview," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 193–209, 2019.
37. A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowledge-Based Systems*, vol. 214, p. 106685, 2021; doi:10.1016/j.knosys.2020.106685
38. A. Verma, V. Murali, R. Singh, P. Kohli, and S. Chaudhuri, "Programmatically interpretable reinforcement learning," in *Proc. of. International Conference on Machine Learning. PMLR, 2018*, pp. 5045–5054, 2018.
39. B. Wymann, E. Espié, C. Guionneau, C. Dimitrakakis, R. Coulom, and A. Sumner, "Torcs, the open racing car simulator," in http://torcs.sourceforge.net, vol. 4, no. 6, p. 2, 2000.
40. T. Shu, C. Xiong, and R. Socher, "Hierarchical and interpretable skill acquisition in multi-task reinforcement learning," in arXiv:1712.07294 [Preprint], 2017.
41. P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 03, pp. 2493–2500, 2020.
42. E. Pierson, D. M. Cutler, J. Leskovec, S. Mullainathan, and Z. Obermeyer, "An algorithmic approach to reducing unexplained pain disparities in underserved populations," *Nature Medicine*, vol. 27, no. 1, pp. 136–140, 2021; doi:10.1038/s41591-020-01192-7
43. J. Born et al., "Accelerating Detection of Lung Pathologies with Explainable Ultrasound Image Analysis," *Applied Sciences*, vol. 11, no. 2, p. 672, 2021; doi:10.3390/app11020672
44. G. Jia, H.-K. Lam, and Y. Xu, "Classification of COVID-19 chest X-Ray and CT images using a type of dynamic CNN modification method," *Computers in Biology and Medicine*, vol. 134, p. 104425, 2021; doi:10.1016/j.compbiomed.2021.104425
45. Y. Song et al., "Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 6, pp. 2775–2780, 2021; doi:10.1109/tcbb.2021.3065361
46. S.-H. Wang, V. V. Govindaraj, J. M. Górriz, X. Zhang, and Y.-D. Zhang, "Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network," *Information Fusion*, vol. 67, pp. 208–229, 2021; doi:10.1016/j.inffus.2020.10.004
47. Z. Wang et al., "Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays," *Pattern Recognition*, vol. 110, p. 107613, 2021; doi:10.1016/j.patcog.2020.107613
48. R. T. Sutton, O. R. Zaïane, R. Goebel, and D. C. Baumgart, "Artificial intelligence enabled automated diagnosis and grading of ulcerative colitis endoscopy images," *Scientific Reports*, vol. 12, no. 1, 2022; doi:10.1038/s41598-022-06726-2
49. Y.-H. Wu et al., "JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3113–3126, 2021; doi:10.1109/tip.2021.3058783
50. S. Lu, Z. Zhu, J. M. Gorriz, S. Wang, and Y. Zhang, "NAGNN: Classification of COVID-19 based on neighboring aware representation from deep graph neural network," *Int. J. Intell. Syst.*, vol. 37, no. 2, pp. 1572–1598, 2022; doi:10.1002/int.22686
51. A. Haghanifar, M. M. Majdabadi, Y. Choi, S. Deivalakshmi, and S. Ko, "COVID-CXNet: Detecting COVID-19 in frontal chest X-ray images using deep learning," *Multimedia Tools and Applications*, vol. 81, no. 21, pp. 30615–30645, 2022; doi:10.1007/s11042-022-12156-z
52. N. S. Punn and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," *Applied Intelligence*, vol. 51, no. 5, pp. 2689–2702, 2020; doi:10.1007/s10489-020-01900-3
53. H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, and Y. Xia, "Triple attention learning for classification of 14 thoracic diseases using chest radiography," Medical Image Analysis, vol. 67, p. 101846, 2021; doi:10.1016/j.media.2020.101846
54. B. Hu, B. Vasu, and A. Hoogs, "X-MIR: EXplainable Medical Image Retrieval," in *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2022*, pp. 440–450, 2022.
55. Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," *Scientific Reports*, vol. 12, no. 1, p. 1170, 2022; doi:10.1038/s41598-022-05112-2
56. C. Severn, K. Suresh, C. Görg, Y. S. Choi, R. Jain, and D. Ghosh, "A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features," *Sensors*, vol. 22, no. 14, p. 5205, 2022; doi:10.3390/s22145205
57. N. Q. K. Le, Q. H. Kha, V. H. Nguyen, Y.-C. Chen, S.-J. Cheng, and C.-Y. Chen, "Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer," *International Journal of*

Molecular Sciences, vol. 22, no. 17, p. 9254, 2021; doi:10.3390/ijms22179254
58. S. H. P. Abeyagunasekera, Y. Perera, K. Chamara, U. Kaushalya, P. Sumathipala, and O. Senaweera, "LISA : Enhance the explainability of medical images unifying current XAI techniques," in *Proc. of 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Apr. 2022*, 2022; doi:10.1109/i2ct54291.2022.9824840
59. J. Duell, X. Fan, B. Burnett, G. Aarts, and S.-M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records," in *Proc. of 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Jul. 2021*, 2021; doi:10.1109/bhi50953.2021.9508618
60. G. Liu et al., "Medical-VLBERT: Medical Visual Language BERT for COVID-19 CT Report Generation With Alternate Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3786–3797, 2021; doi:10.1109/tnnls.2021.3099165

Received 10-06-2023, the final version — 22-06-2023.

Alexey Averkin, Candidate of Sciences (Phys.-Math.), Associate Professor, Leading Researcher of the Research Center for Advanced Studies in Artificial Intelligence, REU them. G. V. Plekhanov, Moscow, ✉ **averkin2003@inbox.ru**

Sergey Yarushev, Candidate of Sciences (Tech.), Associate Professor, Director of the Research Center for Advanced Research in Artificial Intelligence, REU them. G. V. Plekhanov, Moscow, sergey.yarushev@icloud.com