ISSN 2782-3806
ISSN 2782-3814 (Online)
УДК 61:004.85

МОДЕЛИ ОБЪЯСНЕНИЯ ДИАГНОЗА КАК ЭЛЕМЕНТ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ ДИАГНОСТИКИ В МЕДИЦИНЕ: КРАТКИЙ ОБЗОР

Уткин Л. В., Крылова Ю. И., Константинов А. В.

Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого», Санкт-Петербург, Россия

Контактная информация:
Уткин Лев Владимирович,
ФГАОУ ВО СПбПУ,
ул. Политехническая, д. 29, Санкт-Петербург, Россия, 195251.
E-mail: lev.utkin@gmail.com.

Статья поступила в редакцию 20.10.2022
и принята к печати 03.11.2022.

РЕЗЮМЕ

В работе рассмотрены наиболее важные и эффективные подходы и модели объяснения и интерпретации результатов диагностики, получаемых с использованием интеллектуальных систем диагностики. Необходимость их использования обусловлена тем, что сама интеллектуальная система диагностики является «черным ящиком» и для врача важно не только получить диагноз пациента, но и понять, почему получен такой диагноз, какие элементы информации о пациенте наиболее значимы с точки зрения диагноза. Приведены обзоры основных подходов к объяснению предсказаний моделей машинного обучения в целом и применительно к медицине. Показано, как различная исходная информация о пациенте влияет на выбор моделей объяснения. Рассмотрены модели при наличии визуальной и табличной информации. Также рассмотрены модели объяснения примерами. Цель работы — обзор основных моделей объяснения и их зависимости от вида информации о пациенте.

Ключевые слова: диагноз, интеллектуальная система диагностики, машинное обучение, объяснительный интеллект, персонализированная медицина.

Для цитирования: Уткин Л.В., Крылова Ю.И., Константинов А.В. Модели объяснения диагноза как элемент интеллектуальных систем диагностики в медицине: краткий обзор. Российский журнал персонализированной медицины. 2022;2(6):23-32. DOI: 10.18705/2782-3806-2022-2-6-23-32.

EXPLANATION MODELS AS A COMPONENT OF THE INTELLIGENT COMPUTER-AIDED DIAGNOSIS SYSTEMS IN MEDICINE: A BRIEF REVIEW

Utkin L. V., Krylova Y. I., Konstantinov A. V.

Peter the Great Saint-Petersburg Polytechnic University

Corresponding author:
Utkin Lev V.,
Peter the Great Saint-Petersburg
Polytechnic University,
Politechnicheskaya str., 29, Saint
Petersburg, Russia, 195251.
E-mail: lev.utkin@gmail.com.

Received 20 October 2022; accepted
03 November 2022.

ABSTRACT

The paper considers the most important and effective approaches and models for explaining and interpreting diagnostic results obtained using intelligent computer-aided diagnosis systems. The need to use them is due to the fact that the intelligent computer-aided diagnosis system itself is a "black box" and it is important for the doctor not only to get the patient's diagnosis, but also to understand why such a diagnosis is stated, what elements of the patient information are the most significant from the point of view of the diagnosis. Reviews of the main approaches to explain predictions of machine learning models applied to general areas as well as to medicine are presented. It is shown how different types of the initial patient information impact on the choice of explanation models. Models are considered when visual or tabular information is available. Example-based explanation models are also studied. The purpose of the work is to review the main explanation models and their dependence on types of information about the patient.

Key words: diagnosis, explainable artificial intelligence, intelligent computer-aided diagnosis system, machine learning, personalized medicine.

For citation: Utkin LV, Krylova YI, Konstantinov AV. Explanation models as a component of the intelligent computer-aided diagnosis systems in medicine: a brief review. Russian Journal for Personalized Medicine. 2022;2(6):23-32. (In Russ.) DOI: 10.18705/2782-3806-2022-2-6-23-32.

Список сокращений: ИИ — искусственный интеллект, ИСД — интеллектуальные системы диагностики.

ВВЕДЕНИЕ

Развитие и внедрение интеллектуальных систем диагностики (ИСД) заболеваний в настоящее время является одной из наиболее интенсивно развивающихся областей применения искусственного интеллекта (ИИ), что связано с быстрым ростом объемов обучающих данных по различным заболеваниям, с разработкой новых эффективных моделей машинного обучения (глубоких нейронных сетей, трансформеров и др.), а главное — с пониманием того, насколько ИСД в паре с врачом позволят снизить долю ошибочных диагнозов и решений. В то же время, несмотря на признание значимости ИСД в медицине, существует ряд препятствий для дальнейшего внедрения ИИ в реальные медицинские учреждения, ключевым из которых является тот факт, что ИСД имеют природу «черного ящика», что означает практическую закрытость для врача, использующего ИСД, механизма получения диагноза системой. Отсутствие четкого понимания у врача, почему система поставила тот или иной диагноз определенному пациенту, а также возможные ошибки ИСД, ее уязвимость по отношению к внешним воздействиям не позволяют специалисту доверять в полной мере диагнозу, поставленному системой, что в целом тормозит внедрение ИИ в медицине.

Для преодоления этого препятствия в последние годы интенсивно развивается направление ИИ, связанное с объяснением или интерпретацией решений, получаемых с использованием моделей машинного обучения, которое имеет различные определения и названия: объяснительный интеллект (eXplainable Artificial Intelligence — XAI), интерпретация предсказаний (prediction interpretability), просто модели объяснения (explainable models). Что позволяет, если говорить в медицинских терминах, ответить на следующие вопросы: какие элементы на снимке КТ говорят о конкретном диагнозе, что общего у пациента с другими пациентами с таким же диагнозом, почему выбранное лечение оптимально, каких симптомов не хватает, чтобы поставить другой диагноз, и т. д. Другими словами, цель такой дополнительной подсистемы, которую мы будем называть просто моделью объяснения, заключается в предоставлении врачу или другому пользователю ИСД полного объяснения выданного диагноза. И если в научной и инженерной среде отношение к использованию систем объяснительного интеллекта в различных областях неоднозначно, то все сходятся к единому мнению о необходимости применения этого компонента интеллектуальной поддержки в медицине.

Здесь также необходимо отметить, что ИИ в медицине не ограничивается диагностикой. Более интересные и одновременно более сложные задачи, которые также могут быть решены при помощи ИИ, заключаются в выборе оптимального лечения (оптимальной дозы лекарства, оптимального режима лечения и т. д.) с учетом характеристик пациента. Это как раз и есть реализация задачи персонализированной медицины. Однако создание самих моделей машинного обучения в рамках этого направления находится на начальной стадии, что также ограничивает и развитие соответствующих моделей объяснения. Поэтому ниже будут рассмотрены в основном модели объяснения ИСД, то есть модели объяснения в медицине.

Следует уточнить, что количество публикаций, так или иначе посвященных проблематике моделей объяснения ИСД, в настоящее время превышает все мыслимые пределы, и их полный обзор не представляется возможным. Например, обзоры, полностью посвященные моделям объяснения в самых различных областях, приведены в [1–15]. Обзоры, которые рассматривают только различные аспекты применения моделей объяснения в медицине, приведены в [16–28]. Следующие публикации могут рассматриваться как примеры использования моделей объяснения в радиологии [29], гистопатологии [30], кардиологии [31], онкологии [32–37].

Поэтому цель статьи — не делать детальный обзор всех моделей объяснения в медицине, а рассмотреть основные типы моделей объяснения, основные подходы, которые используются в медицине для объяснения и интерпретации диагноза, рассмотреть само понятие объяснения, что подразумевается под объяснением в различных моделях и при различной информации о пациенте. Кроме того, мы ограничимся рассмотрением *локальных* моделей объяснения, которые основаны на определении значимых признаков или факторов, влияющих на решение ИСД как «черного ящика» только для одного пациента. Этот случай является более интересным, так как в медицинской практике важно иметь объяснение диагноза конкретного пациента с его характеристиками. *Глобальные* модели объяснения, в отличие от локальных, определяют наиболее значимые факторы, влияющие на диагностику всех пациентов в обучающей выборке. Если говорить о медицине, то их область применения ограничена в основном задачами эпидемиологии.

ЛОКАЛЬНЫЕ МОДЕЛИ ОБЪЯСНЕНИЯ ДЛЯ ДАННЫХ ВИЗУАЛИЗАЦИИ

Локальные модели объяснения и методы представления объяснения во многом определяются типом данных, на которых обучена ИСД. Если данные получены в результате использования КТ, МРТ, УЗИ и других методов диагностики, результатом которых является 2D или 3D изображение, то одним из наиболее популярных методов представления является карта значимости (saliency map), которая представляет собой закрашивание определенных участков изображения (пикселей, вокселей), объясняющих поставленный ИСД диагноз, различными цветами, например, красным цветом выделяются самые важные участки с точки зрения объяснения. Пример карты значимости как результата функционирования модели объяснения показан на рисунке 1, где ИСД диагностировала гранулему на рентгеновском снимке, и модель объяснения выдала объяснение в виде карты значимости. Такая визуализация позволяет сразу увидеть, что является причиной диагноза. Однако и такой подход имеет недостаток. Он не позволяет увидеть некоторые скрытые элементы объяснения, которые могут быть более значимыми по сравнению с видимыми элементами. Соответствующими моделями объяснения, позволяющими получить карты значимости, являются Class Activation Mapping (CAM) [38], Gradient-weighted CAM (Grad-CAM) [39], DeepLIFT [40] и ряд других моделей. Создание карт значимости с применением этих моделей основано на использовании линейной комбинации выходных данных последнего слоя сверточной нейронной сети, реализующей ИСД.

Рис. 1. Иллюстрация карты значимости на выходе модели объяснения (МО)

Недостатком этого подхода является то, что «черный ящик» в виде нейронной сети частично «приоткрывается», то есть для реализации подхода необходима информация на последнем слое сети. Поэтому модели не могут быть реализованы как отдельные модели объяснения для готовых ИСД, доступ к программному обеспечению которых закрыт. В то же время такие модели объяснения могут быть встроены в ИСД как элемент. В таких случаях говорят об объясняемой ИСД в целом, а не о модели объяснения как отдельной метамодели.

Несмотря на достаточно широкое распространение моделей объяснения на основе карт значимости, они часто могут давать неоднозначные объяснения, что затрудняет их качественную оценку. Поэтому большой интерес привлекли модели, предоставляющие текстовые объяснения. Именно текстовые объяснения во многих случаях предпочтительнее визуальных объяснений, поскольку они по своей природе понятны людям. А совместные визуальные и текстовые объяснения позволят с большей степенью доверия относиться к решениям ИСД [41]. Такие объяснения основаны на генерации фраз, описывающих диагнозы, поставленные ИСД. В большинстве моделей объяснения генерация осуществляется в два этапа:

1. На основе произвольной модели объяснения выделяются значимые факторы (признаки, пиксели, воксели) или группы значимых факторов.
2. Группам ставится в соответствие объяснение на естественном языке. Это осуществляется с использованием языковых моделей машинного обучения или при помощи специальных словарей, в которых описывается такое соответствие, что зачастую более эффективно, особенно, когда количество вариантов текстового описания диагноза мало.

Это один из наиболее перспективных подходов для объяснения диагноза заболевания в медицине. Однако методы текстовых объяснений являются наиболее сложными с точки зрения реализации и требуют использования других методов объяснения в качестве предварительного анализа, которые выделяют значимые факторы, что позволяет устанавливать соответствие между этими факторами и фразами естественного языка. Однако, несмотря на сложность реализации, модели объяснения с текстовыми объяснениями представляют собой наиболее перспективную группу моделей, что также продиктовано возможностью использования новых языковых моделей на основе трансформеров [42]. Примером реализации модели объяснения с применением трансформеров является представленная в работе [43] модель объяснения, генериру-

ющая диагностический отчет, который содержит информацию по рентгенограмме грудной клетки.

Еще одним интересным подходом в рамках объяснения на естественном языке является использование концептов в объяснении, которые представляют собой ограниченный набор фраз-примитивов, описывающих диагнозы [44, 45]. Например, в работе [45] систематизирован список концептов, описывающих новообразование в легком, следующим образом:

– по типу: очаг, узловое образование, консолидация;
– по форме: сферическая, треугольная, неправильная;
– по структуре: солидная, частично солидная, матовое стекло;
– по включению: кальций, некроз, воздушная полость, жир, воздушная бронхограмма, гомогенная структура;
– по контуру: ровный, неровный, четкий, нечеткий, спикулообразный.

Для каждого диагноза формируется наиболее вероятное подмножество концептов в соответствии с приведенной классификацией.

ЛОКАЛЬНЫЕ МОДЕЛИ ОБЪЯСНЕНИЯ ДЛЯ ТАБЛИЧНЫХ ДАННЫХ

В настоящее время инструменты визуализации, например, рентгеновские снимки, компьютерная томография, ультразвук, являются одной из наиболее активных областей применения объяснительного интеллекта. Однако многие медицинские диагностические исследования также основаны на других типах данных, таких как табличные данные и данные временных рядов, которые могут быть получены из клинической информации, например, в виде электронных медицинских карт. Табличные данные и временные ряды требуют совершенно других методов объяснения. Более того, объяснение диагноза, выделяя в таких данных подмножества значимых атрибутов или факторов, является намного более важной задачей, особенно если количество данных о пациенте достаточно большое (демографические данные, клиническая информация, результаты анкетирования, лабораторные тесты, измерения основных показателей жизнедеятельности), и врачу требуется, соответственно, большее время, чтобы выделить эти подмножества самостоятельно, исходя из своего опыта. Развитие систем мобильного здравоохранения, телемедицины предлагает все больше возможностей для удаленного мониторинга состояния здоровья и накопления информации о состоянии пациента. Интеграция между источниками данных и моделями ИИ может внести фундаментальный вклад в оказание ранней, персонализированной и высококачественной помощи. В этих условиях модели объяснения становятся важнейшей составляющей для получения эффективных объяснений как опытными врачами-экспертами, так и начинающими практикующими врачами.

Основная идея, лежащая в основе моделей объяснения для ситуации табличных данных, заключается в аппроксимации некоторой сложной неявной функции, отображающей, грубо говоря, характеристики пациента в диагноз и реализуемой при помощи ИСД, некоторой простой функцией, которая принадлежит множеству объясняемых функций. Одной из таких функций является линейная функция атрибутов или признаков данных о пациентах. Фактически мы заменяем сложную и неизвестную функцию ИСД линейной функцией в точке в пространстве признаков, определяемой конкретным пациентом, диагноз которого объясняется. Почему именно линейная функция? Потому что значения коэффициентов линейной функции являются как раз численной мерой значимости соответствующих признаков данных о пациенте. Если коэффициент при первом признаке больше коэффициента при втором признаке, то, соответственно, значимость первого признака больше, чем второго. Это выполняется, если все признаки нормированы.

Одним из самых популярных методов локальной интерпретации, основанным на линейной аппроксимации, является метод «Local Interpretable Model-agnostic Explanations» (LIME) [46]. Согласно методу LIME, модель, которая должна быть объяснена, аппроксимируется линейной моделью в локальной области вокруг интерпретируемого примера. Параметры полученной линейной модели используются для определения степени важности признаков анализируемого примера. Для этого, в соответствии с LIME, в окрестности интерпретируемого примера осуществляется генерация случайных примеров. Далее при помощи ИСД определяются метки классов сгенерированных примеров. Полученная обучающая выборка из этих примеров с их метками классов используется для построения линейной разделяющей функции, которая и является той линейной аппроксимацией, по которой можно определить значимость признаков.

Рисунок 2 иллюстрирует, как функционирует метод LIME для примеров с двумя признаками: плотность очага в легком и диаметр очага, по которым определяется наличие Covid-19. Обучающая выборка состоит из заболевших Covid-19 (треугольники) и не заболевших Covid-19 (круги). Раз-

деляющая функция между примерами из различных классов представлена в виде сплошной кривой и реализуется ИСД. Необходимо объяснить, почему пациент, изображенный в виде ромба, не болеет Covid-19, то есть необходимо определить: диаметр очага или его плотность в большей степени повлияли на то, что ИСД поставила такой диагноз этому пациенту. Окрестность вокруг примера на рисунке 2 увеличена, и показаны сгенерированные примеры пациентов. Штриховая прямая (искомая линейная аппроксимация) также показана в увеличенной окрестности. Эта прямая явно соответствует большему коэффициенту для плотности очага, что соответствует его большему влиянию на то, что ИСД поставила для этого пациента диагноз «не Covid-19».

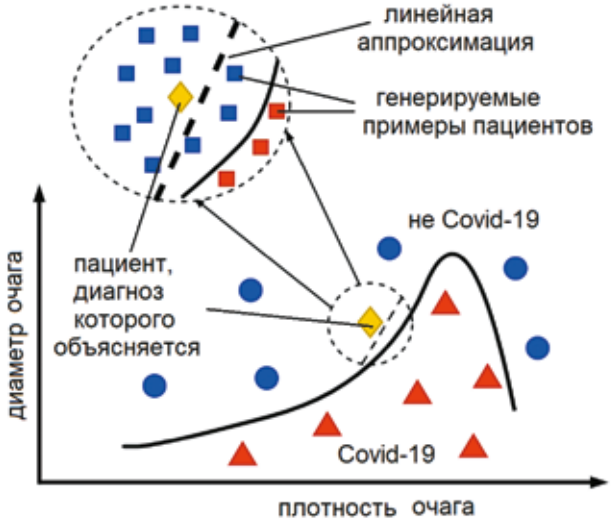

Рис. 2. Объяснение метода LIME

Выбор простой разреженной линейной модели может привести к существенным ошибкам, если функция, реализуемая ИСД, в объясняемых примерах существенно нелинейная. Кроме того, объяснения возникают из-за случайных возмущений исходного входного пространства, которые могут не отражать пример для объяснения. В то же время эффективность LIME привела к созданию целой серии модификаций метода, например, ALIME [47], DLIME [48], Anchor LIME [49], SurvLIME [50]. Одним из интересных обобщений LIME является метод Neural Additive Model (NAM) [51], который основан не на линейной аппроксимации, а на аппроксимации с использованием обобщенной аддитивной модели. В соответствии с этой моделью, аппроксимирующая функция представляется в виде суммы функций отдельных признаков. Основная идея NAM заключается в том, чтобы реализовать эти функции в виде нейронных сетей с одним входом каждая так, что этот вход соответствует одному из признаков. Результатом функционирования NAM является набор функций, каждая из которых показывает, насколько быстро изменяется диагноз пациента при изменении определенного признака. Чем выше скорость изменения функции, тем больше значимость соответствующего признака. На рисунке 3 показана нейронная сеть, обученная на имеющейся обучающей выборке, пациенты в которой имеют признаки (пульс, давление, …, возраст) и диагноз «атеросклероз» или его отсутствие. Функции $g_1, g_2, \ldots, g_m$, реализованные обученными нейронными сетями, показывают, как изменяется вероятность атеросклероза в зависимости от каждого признака. Скорость изменения в точках со значениями пульса, давления, возраста конкретного пациента объясняет поставленный ИСД диагноз.
В работе [52] предложен метод объяснения, аналогичный NAM, но вместо нейронных сетей предлагается использование моделей градиентного бустинга, что позволило значительно повысить точность объяснения для ряда баз данных.

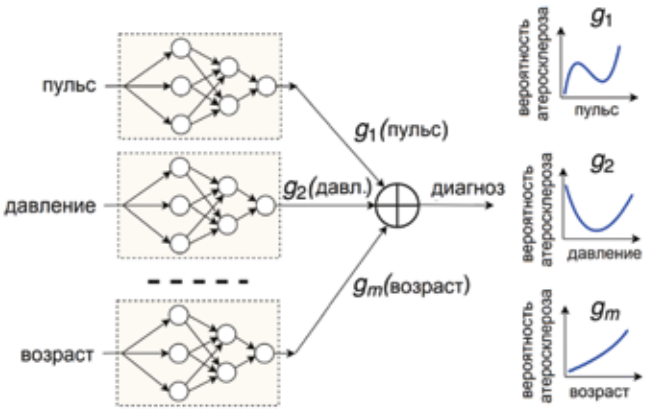

Рис. 3. Иллюстрация метода объяснения NAM

Еще один популярный метод объяснения и анализа значимости признаков объясняемого примера основан на ранжировании признаков с точки зрения их влияния на диагноз, поставленный ИСД, называется метод чисел Шепли или метод SHapley Additive exPlanations (SHAP) [53, 54]. Этот метод строит линейную регрессионную модель, коэффициенты которой определяются с применением чисел Шепли и теоретико-игрового подхода. Метод SHAP требует перебора всех возможных комбинаций признаков. Собственно, он и основан на суммарном сравнении диагнозов, выдаваемых ИСД для пар подмножеств признаков. Поэтому SHAP имеет два существенных недостатка. Во-первых,

полный перебор комбинаций признаков является вычислительно сложной задачей, и при числе признаков больше 20 использование метода не представляется возможным. Во-вторых, ИСД обучена на данных определенной размерности, и рассмотрение подмножеств признаков требует доопределения оставшихся признаков, не входящих в очередное подмножество. Решение этой задачи неоднозначно и может привести к некорректным результатам. Первую проблему частично решают модификации SHAP, например, Kernel SHAP [54] или Random SHAP [55] с сокращенным перебором подмножеств. Однако эти модификации не всегда дают гарантированное корректное объяснение.

Вышеперечисленные МО позволяют объяснять, выделяя наиболее значимые признаки, атрибуты, части изображений. Однако существуют также и другие модели, которые используют аналогичных пациентов с точки зрения их характеристик и диагноза. Такие модели называются моделями объяснения примерами или моделями, основанными на примерах.

МОДЕЛИ ОБЪЯСНЕНИЯ ПРИМЕРАМИ

Стратегия, заложенная в основной идее моделей объяснения примерами, используется в медицине врачами для объяснения причин, почему было принято определенное решение. Один из подходов, используемый в ряде моделей объяснения примерами, заключается в рассуждении на *основе прецедентов* (*case-based reasoning*). При анализе диагноза нового пациента в таких моделях объяснения осуществляется поиск пациентов из базы данных с наиболее близким изображением очага, опухоли, другого региона изображения в соответствии с картой значимости, если диагноз ставится по изображению, или с наиболее близкими табличными характеристиками, если исходные данные о пациенте — табличные. Успешные примеры реализации моделей объяснения на основе прецедентов при объяснении диагнозов различных заболеваний представлены в работах [56–58].

Второй подход к реализации моделей объяснения примерами позволяет получить объяснение, основанное на противопоставлении (counterfactual explanation) [59]. Подход состоит в контролируемом возмущении данных пациента таким образом, чтобы измененные характеристики пациента «перевели» его в другой класс, то есть ИСД поставила бы другой диагноз. Оптимальные возмущения как раз и указывают те признаки, которые позволяют получить объяснение, основанное на противопоставлении. Интуиция подхода заключается в том, что зачастую проще объяснить, чего не хватает пациенту с точки зрения его характеристик, чтобы он имел другой диагноз. Если характеристики пациента представлены в виде изображений (КТ, УЗИ и т. д.), то для реализации моделей объяснения используются порождающие нейронные сети (вариационные автокодеры, GAN), которые фактически и осуществляют контролируемые возмущения [60, 61]. Для случая табличных данных о пациенте обзор многих моделей объяснения, основанных на противопоставлении, представлен в работе [62].

Третий подход в рамках моделей объяснения примерами основан на использовании прототипов — пациентов с определенным диагнозом, структура данных которых является «типичной» для этого диагноза и одновременно близка структуре данных объясняемого пациента. Обоснование этого подхода заключается в том, что во время обучения различные части изображения (КТ, МРТ и др.) выступают в качестве прототипов, представляющих диагнозы. Когда новое изображение необходимо оценить на этапе тестирования, сеть находит прототипы, наиболее похожие на части тестового изображения. Этот подход послужил основой для ряда моделей объяснения [63–65].

ЗАКЛЮЧЕНИЕ

В работе рассмотрена только малая часть большого числа подходов и моделей объяснения и интерпретации результатов диагностики, получаемых с использованием ИСД. Сегодня модели объяснительного интеллекта являются одним из направлений в машинном обучении и ИИ, которое наиболее интенсивно развивается, что обусловило огромный рост числа публикаций, обзоров, программных продуктов, поддерживающих и реализующих соответствующие модели. Поэтому цель работы заключалась в рассмотрении нескольких наиболее важных, эффективных и популярных подходов и моделей, которые в основном перекрывают практически все направления. Было приведено большое количество современных обзоров как по моделям объяснительного интеллекта в целом, так и применительно к медицине. Так как работа ориентирована прежде всего на врача, то математического обоснования всех методов и их особенностей с точки зрения машинного обучения не было рассмотрено.

Необходимо отметить, что целый ряд моделей объяснения не был затронут. Это касается данных временных рядов и моделей выживаемости, когда ИСД анализирует время до некоторого события (рецессии, смерти и т. д.), связанного с пациентом в условиях цензурированности выборки. Важной

особенностью этих моделей является то, что ИСД выдает не диагноз, а вероятностные характеристики времени до события в виде некоторой вероятностной функции времени, например, в виде функции риска или функции выживаемости. В таких задачах объяснение, почему конкретный пациент имеет определенную функцию выживаемости, приобретает новый смысл. Анализ моделей объяснения для таких ситуаций является задачей для дальнейших исследований. Другой важный пласт проблем связан с проблемой оценки эффективности лечения для конкретного пациента. Объяснение, почему для этого пациента необходима определенная доза лекарства, а не какая-то другая, является также важнейшей задачей для дальнейших исследований.

Конфликт интересов / Conflict of interest

Авторы заявляют об отсутствии потенциального конфликта интересов. / The authors declare no conflict of interest.

Финансирование / Funding

Исследование выполнено за счет гранта Российского научного фонда (проект № 21-11-00116). / This work is supported by the Russian Science Foundation under grant 21-11-00116.

СПИСОК ЛИТЕРАТУРЫ / REFERENCES

1. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) // IEEE Access. 2018; 6:52138–52160.

2. Angelov PP, Soares EA, Jiang R, et al. Explainable artificial intelligence: an analytical review // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2021; 11(5):1424.

3. Bodria F, Giannotti F, Guidotti R, et al. Benchmarking and survey of explanation methods for black box models // arXiv:2102.13076. 2021 Feb.

4. Burkart N, Huber, MF. A survey on the explainability of supervised machine learning // Journal of Artificial Intelligence Research. 2021; 70:245–317.

5. Cambria E, Malandri L, Mercorio F, et al. A survey on XAI and natural language explanations // Information Processing & Management. 2023; 60(1): 103111.

6. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: A survey on methods and metrics // Electronics. 2019; 8(8):832.

7. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models // ACM Computing Surveys. 2019; 51(5):1–42.

8. Krenn M, Pollice R, Guo SY, et al. On scientific understanding with artificial intelligence // Nature Reviews Physics. 2022 Oct 11:1–9.

9. Li Z, Zhu Y and Matthijs van Leeuwen. A Survey on Explainable Anomaly Detection // arXiv:2210.06959 (2022).

10. Marcinkevics R and Vogt JE Interpretability and explainability: A machine learning zoo mini-tour // arXiv:2012.01805. Jan 2020.

11. Minh D, Wang HX, Li Y, et al. Explainable artificial intelligence: a comprehensive review // Artificial Intelligence Review. 2021:1–66.

12. Sahakyan M, Aung Z, Rahwan T. Explainable artificial intelligence for tabular data: A survey // IEEE Access. 2021; 9:135392–135422.

13. Schwalbe G, Finzel B. XAI method properties: A (meta-) study // arXiv:2105.07190. 2021 May.

14. Sejr JH, Schneider-Kamp A. Explainable outlier detection: What, for Whom and Why? // Machine Learning with Applications. 2021; 6:100172.

15. Zhang Q, Zhu SC. Visual interpretability for deep learning: a survey // Frontiers of Information Technology & Electronic Engineering. 2018; 19(1):27–39.

16. Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: a survey on tabular and time series data // Artificial Intelligence Review. 2022:1–55.

17. Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine // WIREs Data Mining and Knowledge Discovery. 2019; 9(4): 1–13.

18. Jin D, Sergeeva E, Weng W-H, et al. Explainable deep learning in healthcare: A methodological survey from an attribution view // WIREs Mechanisms of Disease. 2022; Vol.14(3):1–25.

19. Loh HW, Ooi CP, Seoni S, et al. Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022) // Computer Methods and Programs in Biomedicine. 2022 Sep 27:107161.

20. Mohanty A, Mishra S. A Comprehensive Study of Explainable Artificial Intelligence in Healthcare // Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis. Springer, Singapore. 2022: 475–502.

21. Patricio C, Neves JC, Teixeira LF. Explainable Deep Learning Methods in Medical Imaging Diagnosis: A Survey // arXiv:2205.04766, May, 2022.

22. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review // Journal of the American Medical Informatics Association. 2020; 27(7):1173–1185.

23. Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis // Journal of Imaging. 2020 Jun 20; 6(6):52.

24. Slijepcevic D, Horst F, Lapuschkin S, et al. Explaining machine learning models for clinical

gait analysis // ACM Transactions on Computing for Healthcare (HEALTH). 2021; 3(2):1–27.

25. Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI // IEEE Transactions on Neural Networks and Learning Systems. 2020; 32(11): 4793–4813.

26. Tonekaboni S, Joshi S, McCradden MD, et al. What clinicians want: contextualizing explainable machine learning for clinical end use // Machine Learning for Healthcare Conference. PMLR. 2019:359–380.

27. Utkin LV, Meldo AA, Kovalev MS, et al. A Review of Methods for Explaining and Interpreting Decisions of Intelligent Cancer Diagnosis Systems // Scientific and Technical Information Processing. 2021; Vol. 48(5):398–405.

28. Yang CC. Explainable Artificial Intelligence for Predictive Modeling in Healthcare // Journal of Healthcare Informatics Research. 2022; 6(2):228–239.

29. Reyes M, Meier R, Pereira S, et al. On the interpretability of artificial intelligence in radiology: challenges and opportunities // Radiology: Artificial Intelligence. 2020 May 27; 2(3):e190043.

30. Abdelsamea MM, Zidan U, Senousy Z, et al. A survey on artificial intelligence in histopathology image analysis // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2022:e1474.

31. Sakai A, Komatsu M, Komatsu R, et al. Medical professional enhancement using explainable artificial intelligence in fetal cardiac ultrasound screening // Biomedicines. 2022; 10(3):551.

32. Lamy JB, Sekar B, Guezennec G, et al. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach // Artificial intelligence in medicine. 2019; 94:42–53.

33. Rodriguez-Sampaio M, Rincón M, Valladares-Rodríguez S, et al. Explainable Artificial Intelligence to Detect Breast Cancer: A Qualitative Case-Based Visual Interpretability Approach // International Work-Conference on the Interplay Between Natural and Artificial Computation. Springer, Cham. 2022:557–566.

34. Hauser K, Kurz A, Haggenmüller S, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review // European Journal of Cancer. 2022; 167: 54–69.

35. Alsinglawi B, Alshari O, Alorjani M, et al. An explainable machine learning framework for lung cancer hospital length of stay prediction // Scientific Reports. 2022; 12(1):1–10.

36. Kobylińska K, Orłowski T, Adamek M, et al. Explainable Machine Learning for Lung Cancer Screening Models // Applied Sciences. 2022; 12(4):1926.

37. Pintelas E, Liaskos M, Livieris IE, et al. Explainable machine learning framework for image classification problems: case study on glioma cancer prediction // Journal of Imaging. 2020; 6(6):37.

38. Zhou B, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016:2921–2929.

39. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017:618–626.

40. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences // Proceedings of the International Conference on Machine Learning (ICML). 2017; Vol. 70:3145–3153.

41. Gale W, Oakden-Rayner L, Carneiro G, et al. Producing Radiologist-Quality Reports for Interpretable Deep Learning // Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). 2019:1275–1279.

42. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017:5998–6008.

43. Chen Y, Song Z, Chang TH, Wan X. Generating Radiology Reports via Memory-driven Transformer // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020:1439–1449.

44. Graziani M, Andrearczyk V, Marchand-Maillet S, Müller H. Concept attribution: Explaining CNN decisions to physicians // Computers in Biology and Medicine. 2020; 123:103865.

45. Meldo AA, Utkin LV, Kovalev MS, et al. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system // Artificial Intelligence in Medicine. 2020; 108:1–10.

46. Ribeiro MT, Singh S, Guestrin C. Why should I trust you? Explaining the predictions of any classifier // arXiv:1602.04938, Aug 2016.

47. Shankaranarayana SM, Runje D. Alime: Autoencoder based approach for local interpretability // International Conference on Intelligent Data Engineering and Automated Learning. Springer. 2019:454–463.

48. Zafar MR, Khan NM. DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems // arXiv:1906.10263, Jun 2019.

49. Ribeiro MT, Singh S, Guestrin C. Anchors: High-precision model-agnostic explanations // AAAI Conference on Artificial Intelligence. 2018:1527–1535.

50. Kovalev MS, Utkin LV, Kasimov EM. SurvLIME: A method for explaining machine learning survival models // Knowledge-Based Systems. 2020; 203:106164.

51. Agarwal R, Melnick L, Frosst N, et al. Neural additive models: Interpretable machine learning with neural nets // Advances in Neural Information Processing Systems. 2021; 34:4699-4711.

52. Konstantinov AV, Utkin LV. Interpretable machine learning with an ensemble of gradient boosting machines // Knowledge-Based Systems. 2021; 222:1–16.

53. Strumbel E, Kononenko I. An efficient explanation of individual classifications using game theory// Journal of Machine Learning Research. 2010; 11:1–18.

54. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions // Advances in Neural Information Processing Systems. 2017:4765–4774.

55. Utkin LV, Konstantinov AV. Ensembles of Random SHAPs // arXiv:2103.03302, Mar., 2021.

56. Tschandl P, Argenziano G, Razmara M, et al. Diagnostic Accuracy of Content Based Dermatoscopic Image Retrieval with Deep Classification Features // British Journal of Dermatology 181. 2019; 1 (2019):155–165.

57. Barata C and Santiago C. Improving the Explainability of Skin Cancer Diagnosis Using CBIR // Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2021:550–559.

58. Sadeghi M, Chilana PK, Atkins MS. How Users Perceive Content-based Image Retrieval for Identifying Skin Images // Understanding and Interpreting Machine Learning in Medical Image Computing Applications. 2018:141–148.

59. Fong RC, Vedaldi A. Interpretable explanations of black boxes by meaningful perturbation // Proceedings of the IEEE International Conference on Computer Vision, IEEE. 2017:3429–3437.

60. Schutte K, Moindrot O, Hérent P, et al. Using StyleGAN for Visual Interpretability of Deep Learning Models on Medical Images // arXiv:2101.07563, Jan (2021).

61. Kim J, Kim M, Ro YM. Interpretation of Lesional Detection via Counterfactual Generation // Proceedings of the IEEE International Conference on Image Processing (ICIP). 2021:96–100.

62. Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking // Data Mining and Knowledge Discovery. 2022 Apr 28:1–55.

63. Kim S, Seo M, Yoon S. XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021:15719–15728.

64. Ming Y, Xu P, Qu H, et al. Interpretable and steerable sequence learning via prototypes // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019:903–913.

65. Oscar L, Hao L, Chaofan C, et al. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions // Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2018; 32:3530–3537.

Информация об авторах:

Уткин Лев Владимирович, д.т.н., профессор Высшей школы искусственного интеллекта ФГАОУ ВО СПбПУ;

Крылова Юлия Игоревна, аспирант ФГАОУ ВО СПбПУ;

Константинов Андрей Владимирович, аспирант ФГАОУ ВО СПбПУ.

Author information:

Utkin Lev V., Ph.D., Dr.Sci., Professor, the Higher School of Artificial, Peter the Great Saint Petersburg Polytechnic University;

Krylova Julia Y., PhD Student, Peter the Great Saint Petersburg Polytechnic University;

Konstantinov Andrei V., PhD Student, Peter the Great Saint Petersburg Polytechnic University.