ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 1, С. 150–178

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

УДК 510.64–519.816

ОБЪЯСНИТЕЛЬНЫЙ ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В АНАЛИЗЕ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ НЕЙРОННЫХ СЕТЕЙ ГЛУБОКОГО ОБУЧЕНИЯ[1]

© 2024 г. А.Н. Аверкин[a],*, Е.Н. Волков[a],, С.А. Ярушев[a],*****

[a] *РЭУ им. Г.В. Плеханова, Москва, Россия*

**e-mail: averkin2003@inbox.ru,*

***e-mail: envolkoff@gmail.com*

****e-mail: sergey.yarushev@icloud.com*

Поступила в редакцию 06.05.2023 г.
После доработки 31.08.2023 г.
Принята к публикации 02.10.2023 г.

Показаны возможности искусственного интеллекта в анализе цифровых изображений в области медицины с помощью сверхточных нейронных сетей глубокого обучения. Рассматривается новое поколение систем искусственного интеллекта с объяснением пользователю алгоритмов принятия решений — объяснительный искусственный интеллект. Приводится таксономия методов объяснения и описание самих методов. Дано обоснование необходимости применения объяснительного искусственного интеллекта в задачах классификации на примере офтальмологических заболеваний. Проведено исследование используемых в обозреваемых работах составляющих методов глубокого обучения (архитектуры нейронных сетей, точности, характеристики наборов данных) и объяснительного искусственного интеллекта (методы объяснения, критерии точности объяснения). В качестве примера рассматривается задача распознавания двух наиболее часто диагностируемых заболеваний глаза: диабетической ретинопатии и глаукомы искусственными нейронными сетями.

Ключевые слова: объяснимый искусственный интеллект (ОИИ), машинное обучение, глубокое обучение, нейросети, нейро-нечёткие системы, анализ цифровых изображений, точность объяснения, офтальмология

DOI: 10.31857/S0002338824010122, **EDN:** WJCMWV

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN DEEP LEARNING NEURAL NETS-BASED DIGITAL IMAGES ANALYSIS

© 2024 A.N. Averkin[a],*, E.N. Volkov[a],, S.A. Yarushev[a],*****

[a] *Plekhanov Russian University of Economics, Moscow, Russia*

**e-mail: averkin2003@inbox.ru,*

***e-mail: envolkoff@gmail.com*

****e-mail: sergey.yarushev@icloud.com*

This review shows the capabilities of artificial intelligence in the analysis of digital images in the field of medicine using convolutional neural networks of deep learning. A new generation of artificial intelligence systems is described with an explanation of decision-making algorithms to the user — explainable artificial intelligence (XAI). The taxonomy of the methods of explanation and the description of the methods themselves are given. The substantiation of the need to use explainable artificial intelligence in classification tasks is given on the example of ophthalmic diseases. The study of the components of deep learning methods used in the reviewed works (neural network architecture, accuracy, characteristics of data sets) and explainable artificial intelligence (methods of explanation, criteria for the accuracy of explanation). As an example, the problem of recognizing two of the most commonly diagnosed eye diseases: diabetic retinopathy and glaucoma by artificial neural networks is considered.

Keywords: explainable artificial intelligence, machine learning, deep learning, neural networks, neuro-fuzzy systems, digital image analysis, accuracy of explanation, ophthalmology, XAI

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

Введение. Период с 2010 по 2020 г. называют декадой искусственного интеллекта (decade of artificial intelligence), тесно связанной с декадой сознания (decade of mind) по классификации ЮНЕСКО. За последние 10—15 лет количество исследований в области искусственного интеллекта (ИИ) и его приложений возросло в тысячи раз по сравнению с периодом до 2010 г. Использование технологий машинного обучения (ML) и глубокого обучения (DL) перестало быть уделом академических исследований, заняв свою нишу в создании цифровых продуктов. Стремительное развитие искусственных нейронных сетей (ИНС), повышение их точности и оптимизация вычислительных затрат позволили создавать продукты на основе ИИ, востребованные во всех сферах человеческой жизни. По оценке Harvard Business Review [1], в нынешнем десятилетии ИИ принесет более 13 трлн долл. в мировую экономику.

До того как технология ИИ стала мейнстримом, она проделала путь длиной в полвека. Первые два поколения ИИ пришлись на последнюю треть XX и начало XXI в. Первое поколение ИИ (1960—1990 гг.) было завершено переходом от экспертных систем и символьного ИИ к коннекционистким системам ИИ, основанным на глубоком обучении и больших данных, характерных для второго поколения ИИ (2000—2020 гг.). Искусственный интеллект третьего поколения (2020—2030 гг.) предполагает возможности интерпретации и объяснения используемых алгоритмов, что значительно повышает надежность по сравнению с предыдущими поколениями [2].

Подход в оценке развития ИИ с помощью разделения временных промежутков на “волны” или “поколения” предложен Управлением перспективных исследовательских проектов Министерства обороны США DARPA (рис. 1). На сегодняшний день программа DARPA по ИИ является наиболее дорогой правительственной программой по развитию ИИ в истории человечества. К 2022 г. по программе DARPA только в развитие третьего поколения ИИ уже инвестировано более 2 млрд долл. США [3].

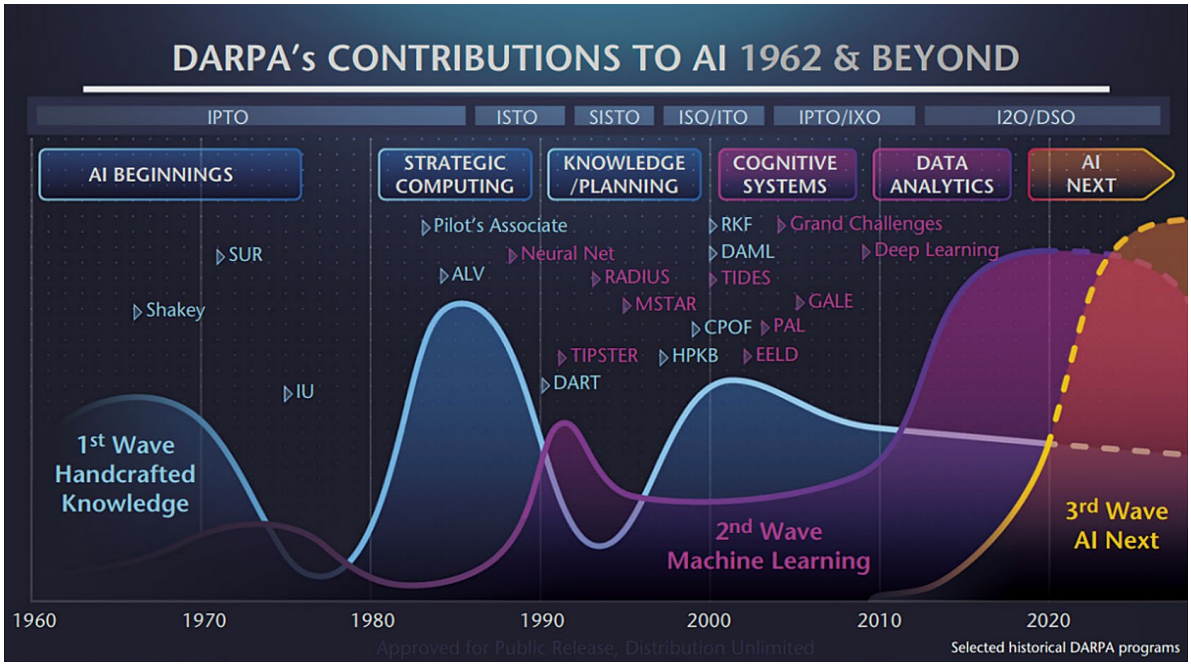

Рис. 1. Вклад DARPA в ИИ [4]

На рис. 1 графически представлен финансовый вклад DARPA в ИИ начиная с 1962 г. [4]. Показано три волны инвестиций DARPA (с несколькими локальными экстремумами), высота каждой волны соответствует величине инвестиций DARPA. На рисунке также показаны избранные исторические программы ИИ, финансируемые DARPA. Первая волна называется “волна знаний ручной работы” (1st wave handcrafted knowledge), что соответствует первому поколению символьного ИИ, основанного на экспертных знаниях. Вторая волна машинного обучения (2rd wave machine learning) соответствует второму поколению ИИ, базирующемуся на нейронных сетях глубокого обучения. Третья волна “следующий искусственный интеллект” (3rd wave AI next) соответствует третьему поколению ИИ, основанному на объяснительном ИИ (ОИИ). Программа AI next предполагает серьезные усилия по устранению ограничений существующего ИИ.

DARPA считает, что текущие инвестиции в исследования и разработки по всему миру слишком сосредоточены на ИИ второго поколения или машинном обучении, которые хорошо подходят для поиска закономерностей в тексте и в изображениях и имеют множество коммерческих приложений, но являются плохо интерпретируемыми для пользователя. Программа AI next сделает ИИ второго поколения волны более надежным для защиты и интерпретации приложений и для обеспечения безопасности, одновременно помогая реализовать третье поколение ИИ (ОИИ и контекстуального мышления). Включение этих технологий в будущие системы ИИ должно облегчить процесс принятия решений; обеспечить совместное понимание большой по объему, неполной и противоречивой информации; расширить возможности автономных робототехнических и беспилотных систем для безопасного выполнения критических миссий.

В настоящее время ОИИ выступает неотъемлемой частью любых интеллектуальных систем (ИС), созданных для работы в ответственных предметных областях. Чаще всего ОИИ используется в медицине как одной их наиболее важных областей для оценки качества полученных ИИ решений. Методы ИИ в медицине применялись с 70-х годов прошлого века, но только с появлением решений на основе машинного обучения и глубокого обучения это направление достигло результатов, сравнимых с возможностями человека-врача, что привело к большому количеству коммерческих применений. Согласно метаобзору [5], уже в 2020 г. в базе данных PubMed было проиндексировано более 11 000 исследований, посвящённых использованию ИИ в медицине. Во многом это стало возможно благодаря накопленным объёмам медицинских данных (медицинские изображения, клинические записи). Применение ИИ в медицине сосредоточено в нескольких направлениях: анализ медицинских изображений [6] и изображений в области микроскопии [7], анализ клинических записей [8], выбор тактики лечения для конкретного заболевания [9], здравоохранение и медицинская помощь [10].

Главной проблемой использования ИИ в ответственных для принятия решений системах всегда была точность и надёжность. С возникновением ИНС прибавилась задача прозрачности процесса классификации и объяснения принятого решения, появилось доверие к технологии. Решением проблемы становится применение ИИ третьего поколения в медицинских системах за счет присутствующей ключевой возможности объяснения результата своей работы. В системах третьего поколения ИИ, как и в системах первого поколения, объясняется внутренняя работа системы. С этой целью системы первого поколения ИИ использовали экспертные знания в правилах и базах знаний и пытались составить языковые описания на базе оценок экспертов. В системах ИИ второго поколения задача оказалась значительно сложнее. Третье поколение ИИ решает эту проблему, хотя недостатки объяснительных систем первого поколения, связанные со слабым уровнем детализации и непонятным языком, могут стать вопросом и для систем третьего поколения. Решение этой проблемы облегчает тот факт, что со времени создания систем объяснения первого поколения, основанных на правилах, компьютерных технологиях и в визуализации данных, анимации, видео и т. п., значительно продвинулись вперед по сравнению с первым поколение, и много новых идей было предложено в качестве потенциальных методов для генерации объяснений [11].

Объяснительный ИИ устроен таким образом, чтобы наблюдающий со стороны пользователь мог понять, почему именно алгоритм принял то или иное решение. Такие методы приходят на смену принципу “черного ящика”, при котором даже сами создатели ИИ не всегда в состоянии объяснить принципы его работы. ОИИ — это важная часть общего тренда, который цикл зрелости технологий Гарднера называют “trust in algorithm” (доверие к алгоритму). Чем больше ИИ заменяет человеческие решения, тем больше он усиливает положительные и отрицательные последствия таких решений. Оставленные без контроля подходы на основе ИИ могут увековечить закрытость и предвзятость, ведущую к проблемам потери производительности и доходов. Двигаясь вперед, организации должны разрабатывать и эксплуатировать системы ИИ на принципах справедливости и прозрачности, заботиться о безопасности и конфиденциальности для общества в целом.

В настоящее время существует постоянно расширяющийся набор методов ОИИ, в англоязычных источниках — XAI (explainable artificial intelligence). Методы ОИИ могут делиться по типам интерпретации результатов. Они содержат модели с внутримодельным объяснением, которые хорошо структурированы и понятны. Эти модели могут давать ответы вместе с объяснениями на базе лингвистического интерпретатора. Другой тип моделей — это модели с постфактумным объяснением. Объяснения в этих моделях получаются с помощью внешних методов анализа моделей ИИ, т.е. методов обратного распространения, отображения активации классов и послойного распространения релевантности.

В моделях ОИИ различаются также модельно-независимые и модельно-зависимые методы, локально-интерпретируемые и глобально-интерпретируемые методы, модели замещения

(суррогатные модели) и модели визуализации, стратегии предварительного моделирования, внутримодельного моделирования и модели корректировки уже разработанной архитектуры нейронной сети. Более детально эти модели будут рассмотрены в следующем разделе.

Массовое появление этих методов связано в основном с начавшейся в 2018 г. программой DARPA по созданию ИИ третьего поколения (ОИИ). Эта программа потребовала не только прогресса в представлении знаний и рассуждений, машинном обучение, языковых технологиях, зрении и робототехники, но также тесной интеграции каждого компонента для реализации надежных и прозрачных ИС, способных работать автономно или в команде, уметь объяснить свои решения. Программа DARPA по созданию систем ОИИ объединяет 11 групп университетов США и стремится создать такие системы ИИ, чьи модели обучения и решения могут быть понятны и должным образом проверены конечными пользователями. Команды разработчиков ОИИ решают эту проблему путем создания и развития технологий объяснимого машинного обучения, исследуя принципы, стратегии и методы взаимодействия человека и компьютера для получения эффективных объяснений.

Система ОИИ обычно состоит из модели принятия решений, которую нужно объяснить, и из объясняющего интерфейса. Подобные системы часто разрабатывается в ситуации, когда у пользователя имеется ИС, которая будет рекомендовать пользователю решения возникающих у пользователя проблем из его предметной области. Для каждого типа ИС строятся свои объяснительные системы, которые в диалоговом режиме с пользователем помогут ему разобраться с возникшими вопросами и получить необходимые объяснения.

Например, команда Исследовательского центра Пало-Альто (PARC) (включающая исследователей из Университета Карнеги—Меллона, Армейского киберинститута, Эдинбургского университета и Университета Мичигана) разрабатывает интерактивную систему объяснений, которая может объяснить возможности и действия смоделированной беспилотной воздушной системы. Объяснения системы XAI должны сообщать, какую информацию она использует для принятия решений и как работает ее модель. Чтобы решить эту проблему, система объяснения и ее пользователи устанавливают общую основу для определения того, какие термины применять в объяснениях и их значения. Это обеспечивается интроспективной моделью дискурса PARC, которая чередует процессы обучения и объяснения [11, 12]. В России аналогичные проблемы ставятся в работе [13] в задачах интеллектуальной поддержки экипажа антропоцентрического объекта в процессе выполнения им задания на сеанс функционирования, которые обеспечиваются бортовыми ИС тактического уровня в форме представления экипажу рекомендуемого способа решения оперативно возникшей тактической задачи.

1. Применение ОИИ в нейросетях глубокого обучения. 1.1. ОИИ для прикладных ИС. В России уделяется большое внимание направлению ОИИ. Один из основных принципов развития и использования технологий ИИ, сформулированных в Национальной стратегии развития ИИ на период с 2018 до 2030 г., является прозрачность, объяснимость работы ИИ и процесса достижения им результатов, недискриминационный доступ пользователей продуктов, которые созданы с помощью технологий ИИ, к информации о применяемых в этих продуктах алгоритмах работы ИИ. Так, Нижегородский государственный университет в 2020 г. стал победителем в конкурсе крупных научных проектов от Минобрнауки России с проектом “Надежный и логически прозрачный искусственный интеллект: технология, верификация и применение при социально-значимых и инфекционных заболеваниях”. Главным результатом проекта стала разработка новых методов и технологий, позволяющих преодолеть два основных барьера систем машинного обучения и ИИ-проблемы ошибок и явного объяснения решений. В 2022 г. РЭУ им. Г.В. Плеханова выиграл проект РНФ “Гибридные модели поддержки принятия решений на основе методов дополненного искусственного интеллекта, когнитивного моделирования и нечеткой логики в задачах персонализированной медицины”, посвященный созданию медицинских систем нового поколения, базирующихся на применении ОИИ для нейросетей глубокого обучения в задачах медицинской диагностики.

Главной составляющей второй и третьей волн ИИ являются технологии глубокого обучения. Множество методов глубокого обучения можно разделить на обучение с учителем и обучение без учителя. В настоящее время эти методы неразрывно связаны с некоторым набором моделей глубоких нейросетей специального вида.

Первая группа методов требует для определения параметров сети наличия размеченных тренировочных наборов данных больших объемов. Как правило, размер наборов для обучения глубоких нейронных сетей на порядки превышает размеры тренировочных баз, используемых для обучения традиционных моделей машинного обучения. К моделям, которые обучаются с учителем, относятся многослойные полностью связанные нейронные сети, сверточные нейронные сети, рекуррентные нейронные сети. Эти модели различаются способами связи

нейронов, находящихся на соседних слоях сети, и количеством нейронов предыдущего слоя, влияющих на состояние отдельных нейронов текущего слоя. При этом для начальной инициализации весов этих моделей — предварительного обучения — могут применятся модели второй группы. Вторая группа моделей не требует наличия размеченных тренировочных наборов данных для настройки параметров сети. К данной группе моделей относятся различные виды автокодировщиков, ограниченные машины Больцмана, разверточные нейронные сети, генеративно-состязательные сети.

Стоит отметить, что около 30% публикаций в Scopus, связанных с использованием глубоких нейросетей в ОИИ, принадлежат авторам, хорошо признанным в области нечеткой логики. В основном это связано с приверженностью нечеткого сообщества созданию объяснимых нечетких систем, поскольку прозрачность и интерпретируемость являются важными составляющими нечетких систем, отмеченными уже в работах создателя аппарата нечетких множеств и мягких вычислений Лотфи Заде. Одним из наиболее интересных направлений в области разработки объяснительных моделей ИИ является извлечение правил с помощью нейронечетких моделей. Системы, основанные на нечетких правилах (СОНП), разработанные с использованием нечеткой логики, стали областью активных исследований с появлением направления мягких вычислений в 1993 г. Эти алгоритмы доказали свои сильные стороны в таких задачах, как управление сложными системами и создание нечетких элементов управления. Взаимосвязь между обоими подходами (ИНС и СОНП) была тщательно изучена, показана их эквивалентность для нейронечетких систем. Это приводит к двум важным выводам. Во-первых, можно применить то, что было обнаружено для одной из моделей, к другой. Во-вторых, можно перевести знания, заложенные в нейронную сеть, на более когнитивно приемлемый язык — нечеткие правила. Другими словами, получаем один из вариантов семантической интерпретации нейронных сетей [11]. В дальнейшем это привело исследовательское сообщество к реализации идей гибридных подходов, таких, как гибридные когнитивные нечеткие модели с функциональным замещением (с сочетанием нечеткой, нейросетевой и эволюционной технологий), композиционные гибридные нечеткие модели с взаимодействием для комплексных задач анализа систем и процессов, которые позднее стали взаимодействовать с нейросетями глубокого обучения [14]. В результате появились новые гибридные системы, классифицируемые как глубокие нейронечеткие системы (ГННС), обладающие высокой интерпретируемостью и сравнимой с классическими системами глубокого обучения точностью. В качестве перспективного типа ГННС можно упомянуть последовательные гибридные сверточные сети, последние полносвязные слои которых заменены нейронечеткой сетью. Исследования, посвященные реализации ГННС, активно применяются в таких областях, как вычислительная техника, здравоохранение, транспорт и финансы [15].

1.2. Методы ОИИ для прикладных ИС в анализе цифровыхизображений. Методы построения систем ОИИ можно подразделить на зависящие от принципов их разработки группы [16]. Рассмотрим основные из них.

1. Модельно-независимые и модельно-зависимые методы. Модельно-независимые методы — это те, которые не учитывают внутренние компоненты модели. Следовательно, их можно применять к любой классической ИНС структуры “черного ящика”. Напротив, модельно-зависимые методы определяются с помощью параметров отдельной модели, таких, как интерпретация весов линейной регрессии, или с использованием выведенных правил из дерева решений, которые будут специфичны для обученной модели. У методов, не зависящих от модели, есть некоторые преимущества, такие, как высокая гибкость для разработчиков в выборе любой модели ИНС 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. Локально интерпретируемые методы используют один результат или конкретные результаты предсказания и классификации модели для создания объяснений. Напротив, глобально интерпретируемые методы используют все возможности модели и общее поведение модели для создания объяснений. В локально интерпретируемых методах существенны только специфические признаки и характеристики. Важность признаков для глобальных методов используется для объяснения общего поведения модели.

3. Модели замещения (суррогатные модели) и модели визуализации. Популярный способ трактовать модель “черного ящика” — применить интерпретируемую приблизительную модель, которая заменяет модель “черного ящика” для объяснения решений. Эта 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. Как и суррогатные модели, модели визуализации предлагают визуальные объяснения и помогают генерировать объяснения в более презентабельной форме, показывают внутреннюю работу многих моделей и не зависят от типа модели (например, какие пиксели помогают отличить кошку от собаки или какие слова помогают решить, является ли документ спамом или нет).

4. Стратегии предварительного моделирования, внутримодельного моделирования и модели корректировки уже разработанной архитектуры нейронной сети ОИИ 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 данных, используемых для построения модели. Общие цели предварительного моделирования заключаются в обобщении данных, описании набора данных, разработке объяснимых признаков и исследовательском анализе данных. Напротив, цель внутримодельного моделирования состоит в том, чтобы разработать объяснимые по своей сути модели, а не создавать модели “черного ящика”. Методологически существуют разные стратегии или способы построения внутримодельных объяснений. Наиболее очевидным является разработка объяснимой по своей сути модели, такой, как линейные модели, деревья решений и наборы правил. Однако необходимы специальные усилия по выбору типов объяснений, встраиваемых в модель.

5. 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, например объединение глубокого скрытого слоя нейронной сети с методом k-ближайших соседей. Кроме того, модель может быть обучена для совместного предоставления прогноза и соответствующего объяснения. 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.

6. Методы постмодельной объяснимости извлекают объяснения, которые по своей сути не могут быть использованы для предварительного описания разработанной модели. Эти популярные апостериорные методы ОИИ 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.

1.3. Классификация алгоритмов ОИИ. Классификация произведена с точки зрения соответствующего типа ОИИ в различных таксономиях по типам алгоритмов ИИ, которые они могут объяснить (табл. 1).

Рассмотрим основные алгоритмы из табл. 1. Среди модельно-независимых стратегий ОИИ — локальные интерпретируемые модель-агностические объяснения LIME (local interpretable model-agnostic explanations). Они являются наиболее популярными алгоритмами, которые объясняют полученный прогноз. LIME реализует линейную интерпретируемую модель, суррогатную (замещающую) модель в качестве локального приближения для объяснения прогноза. Из-за локальной аппроксимации LIME работает со всеми моделями черного ящика и со всеми типами данных (например, с текстовыми данными, табличными данными, изображениями, графиками).

Алгоритм аддитивных объяснений Шепли SHAP (shapley additive explanations) 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 для создания глобальных объяснений. Также существуют алгоритмы оптимизации с точки зрения вычислительной сложности для SHAP, такие, как TreeSHAP (древовидный SHAP) и DeepSHAP (глубокий SHAP), которые не зависят от модели.

Контрафактический алгоритм (counterfactual) — это еще один алгоритм, доступный как для модельно-независимых, так и для модельно-зависимых вариантов. Контрафактический алгоритм 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.

Таблица 1. Классификация алгоритмов ОИИ

Модель XAI	Методы разработки моделей XAI				Алгоритм обучения
	Модельно-независимые и модельно-зависимые методы	Алгоритмы локально интерпретируемые и глобально интерпретируемые	Замещающие методы (суррогатные модели) и модели визуализации	Стратегии предварительного моделирования, внутримодельного моделирования и модели корректировки	
Facets	Модельно-зависимые	Оба метода	Модель визуализации	Предварительного моделирования	Без учителя
LIME	Модельно-зависимые	Локально интерпретируемые	Суррогатная модель	Модели корректировки	С учителем
SHAP	Модельно-зависимые	Оба метода	Суррогатная модель	Модели корректировки	С учителем
Counterfactual	Оба метода	Оба метода	Суррогатная модель	Модели корректировки	С учителем
LRP	Модельно-независимые	Локально интерпретируемые	Модель визуализации	Модели корректировки	Глубокое обучение
PIRL	Модельно-независимые	Глобально интерпретируемые	Суррогатная модель	Внутримодельного моделирования	Обучение с подкреплением
Hierarchical Policies	Модельно-независимые	Локально интерпретируемые	Суррогатная модель	Внутримодельного моделирования	Обучение с подкреплением
Structural Causal Model	Модельно-независимые	Локально интерпретируемые	Оба метода	Модели корректировки	Обучение с подкреплением

Послойное распространение релевантности LRP (layerwise relevance propagation) — это алгоритм, разработанный для объяснения глубоких нейронных сетей с предположением, что классификатор может быть разложен на разные уровни, что делает его модельно-зависимым методом. Алгоритм LPR 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 (explainable reinforcement learning) — многообещающая, но сложная исследовательская ветвь ОИИ, поскольку модель обучения с подкреплением RL (reinforcement learning) часто содержит огромное количество состояний. Тем не менее XRL может ускорить процесс проектирования RL, облегчая разработчикам отладку систем. Существует классификация методов XRL, основанная на типах объяснений (текст или изображения), на уровне объяснения (локальный, только для прогнозов, или глобальный, если объясняется вся модель), а также на типах объясняющих алгоритмов.

Программно-интерпретируемое обучение с подкреплением PIRL (programmatically interpretable reinforcement learning) является примером глобально-интерпретируемого метода. Идея PIRL заключается в использовании языка программирования, гораздо более близкого к человеческому мышлению для имитации поведения модели глубокого обучения с подкреплением DRL (deep reinforcement learning). PIRL применяет структуру под названием нейроуправляемый программный поиск NDPS (neurally directed program search) 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 гораздо более понятными. Модель PIRL успешно применялась в открытом гоночном симуляторе TORCS (the open racing car simulator).

Иерархическое интерпретируемое приобретение навыков в многозадачном обучении с подкреплением 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 Майнкрафт (minecraft), и он реализует иерархическую модель, основанную на двух уровнях, с помощью алгоритма актер-критик, он же использовался для объяснения многозадачной модели RL,

играющей в Майнкрафт. Для объяснения модели RL также применяется модель стохастической временной грамматики, чтобы зафиксировать все отношения между действиями для создания иерархической модели. Люди часто просто просят припарковать автомобиль, вместо того чтобы определять все действия, связанные с рулем, сцеплением, акселератором и тормозом.

Метод иерархических политик (hierarchical policies) является еще одним встроенным в модель методом. Он используется для интерпретации процесса принятия решений в многозадачных сложных системах RL, но локально. Основной принцип этого подхода заключается в разбиении сложной задачи на более мелкие подзадачи, которые будут решаться с помощью уже изученных действий или путем овладения новым навыком.

Структурно-причинные модели SCM (structural causal model) — это очень четкий способ представления причинно-следственных связей событий. В SCM часто используется ориентированный ациклический граф DAG (directed acyclic graph), в котором узлы представляют состояния, а ребра — действия. Проходя по графу, можно наблюдать, какие действия приводят из одного состояния в другое. Процесс состоит из трех основных этапов: создание DAG; использование моделей многомерной регрессии для аппроксимации взаимосвязей с помощью минимального количества переменных; объяснения путем анализа переменных DAG, чтобы ответить на вопросы: “Почему действие A?” и “Почему не действие B?”.

1.4. Особенности применения в медицине. Далее на примере нескольких типов ИС в области медицины, основанных на анализе цифровых изображений нейронными сетями глубокого обучения, рассмотрены результативные объяснительные системы и продемонстрирована их эффективность для пользователя.

Отдельно стоит отметить, что использование объяснений в медицинских системах является традиционным для ИИ. Так, уже первая экспертная система MYCIN имела вопросно-ответный функционал в системе объяснений на базе дерева вывода, что можно считать вариантом ОИИ. Системы, основанные на знаниях, появившиеся в 1970-х годах прошлого века, также имели в своей структуре блок объяснения результата. Примерами таких систем являются NEOMYCIN и GUIDON. Успешное внедрение ИИ и машинного обучения в медицинские системы заставило задуматься о важности объяснения результатов как одной из первоочередных задач создания подобных систем.

Применение методов объяснения результатов в медицинских системах поддержки принятия решений способно вывести их на совершенно новый уровень, устранить семантический пробел, возникший при разработке подобных решений инженерами и дальнейшем их использовании клиницистами. Указанная проблема является общей для всех интерактивных медицинских систем, но особую актуальность приобретает при наличии в основе таких систем технологии нейросетей.

Так, для создателей систем поддержки принятия решений важно понимать, как модель получает те или иные результаты для улучшения ее характеристик и разработки новых алгоритмов. Нередко плохой результата после обучения модели тормозит процесс ее создания, так как нет понимания, почему это произошло. Возникает вопрос о наличии ошибок в реализации модели, приводящих к ухудшению результата. Учитывая многоэтапность создания модели, для поиска ошибок необходимо будет пройти все этапы построения модели и обучения заново. Использование методов ОИИ позволяет избежать необходимости пересмотра этапов создания модели за счет выявления ключевых признаков, сыгравших роль в предсказании результата (рис. 2).

Клиницистам как конечным пользователям продукта важно понимать алгоритм получения моделью результата. Возможна ситуация, когда клиницист убедился в компетентности ИС, которая для ряда возникших случаев дала ему правильные диагнозы и успешно решила возникшие проблемы. Но вот для очередной проблемы ИС рекомендовала решение, с которым пользователь не согласен. Клиницист ставит свой диагноз, отметив при этом, что в базе знаний его ИС произошел “семантический отказ”. Это может быть связано с неполнотой базы знаний, плохими обучающей выборкой или алгоритмом обучения, но может быть правильным решением, в справедливости которого нужно убедить пользователя. Получить обоснование результата можно за счет выделения ключевых признаков, повлиявших на результат предсказания. Эти признаки могут быть показаны клиницисту как в виде визуального объяснения, так и в виде текстового объяснения (рис. 3).

Использование методов ОИИ в медицинских системах соответствует современным требованиям, предъявляемым к таким системам. Например, в Европейском союзе были разработаны основные требования к моделям ИИ в медицине, необходимые для повышения доверия к подобным алгоритмам.

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

Рис. 2. Объяснение для инженеров

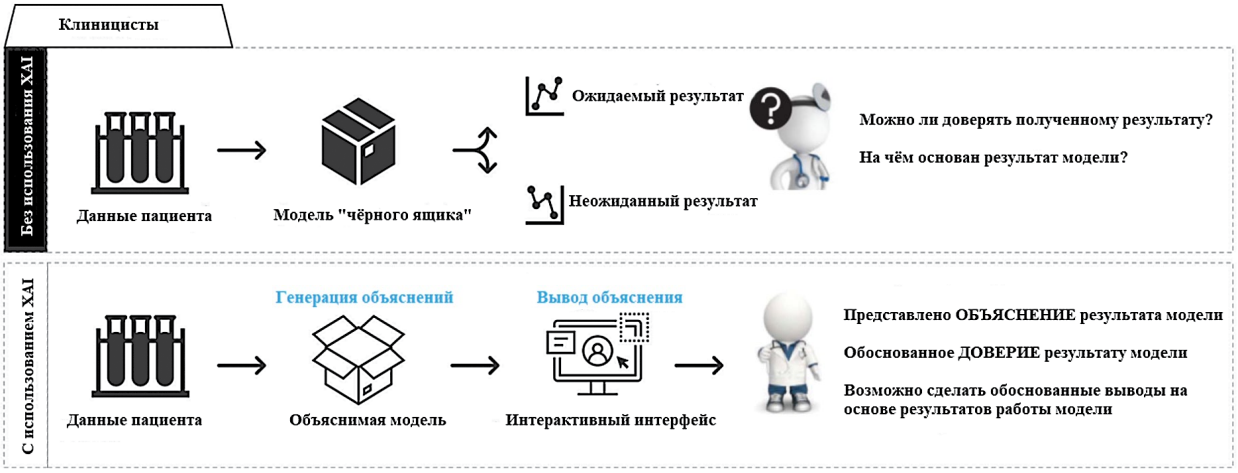

Рис. 3. Объяснение для клиницистов

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 объяснимых 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 маргинализации уязвимых групп населения. Разнообразие подразумевает доступность систем ИИ для всех групп населения, независимо от их социального статуса и состояния здоровья.

6. Социальное и экологическое благополучие. Системы ИИ 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 ИИ. Необходима возможность аудита, оценки алгоритмов и данных, особенно это важно в таких критически значимых отраслях, как здравоохранение.

Применение ОИИ в медицине поможет не только устранить ограничения существующих решений ИИ, но и перейти на качественно новый уровень благодаря новому подходу к медицине — концепции Здравоохранения 5.0 (Healthcare 5.0) — ставящему в основу создание персонализированной медицины, появление и развитие систем медицинского ИИ. В сфере анализа медицинских изображений применение методов ОИИ уже позволило достичь результатов. Предоставление визуального объяснения результата работы ИНС способно увеличить уровень доверия медицинского персонала за счет представления интуитивно понятного объяснения в виде тепловой карты. Согласно обзору [17], на конец 2020 г. в базе данных Scopus суммарно было проиндексировано 223 публикации по исследованию возможностей применения ОИИ в анализе медицинских изображений.

Одной из областей медицины, где использование систем на основе ИИ и с помощью технологии ОИИ позволяет существенно улучшить результат, является офтальмология. Возможности ИИ в анализе изображений глаза с целью диагностики заболеваний исследовались в нескольких обзорах [18—20]. Авторы публикаций отмечают невысокую распространенность применения ИИ в анализе офтальмологических изображений. Системных обзоров применения ОИИ для тех же целей до настоящего времени нами не найдено.

Медицинскими изображениями в офтальмологии являются снимки высокого разрешения, полученные при офтальмоскопии — неинвазивной процедуре, позволяющей оценить состояние сетчатки, хрусталика, диска зрительного нерва, глазного дна, которая проводится с помощью офтальмоскопа. Офтальмоскопия бывает двух типов: прямая (не перевернутое изображение с 15-кратным увеличением) и непрямая (перевернутое изображение с не более чем 5-кратным увеличением). Офтальмоскоп представляет собой прибор (стационарный или ручной), использующий поток направленного света и линзы для увеличения изображения. Современные офтальмоскопы также имеют возможности подключения к персональному компьютеру для вывода изображения [21]. Другим вариантом медицинского изображения в офтальмологии является снимок оптической когерентной томографии (ОКТ) — метода диагностики офтальмологических заболеваний, заключающегося в визуализации структур глаза в сверхвысоком разрешении с помощью низкокогерентной интерферометрии для получения изображений поперечного сечения сетчатки и диска зрительного нерва [22].

Существуют созданные на основе цифровых технологий системы, позволяющие проводить осмотр работающих в опасных условиях. В [23] представлена специализированная система предрейсовых медицинских осмотров, использующая комплексный подход для предполетной диагностики состояния организма пилотов и членов экипажа. Внедрение в подобные системы решений на основе ИИ и ОИИ позволят повысить доверие к результатам работы и снизить нагрузку с оператора системы. Применение методов объяснения результатов в задачах диагностики состояния здоровья в ответственных системах имеет особое значение в контексте полного понимания врачом-диагностом результатов, полученных из системы [24]. Отечественные работы по исследованию возможностей раннего обнаружения патологий зрения методами интеллектуального анализа, отличающимися особым подходом к формулированию проблемы и нестандартными решениями представлены в [25]. В этих публикациях в основном рассматриваются возможности электроретинографии, что более предпочтительно для углубленной диагностики зрения, осуществляемой на базе высокоспециализированных центров, а не массовой диагностики на уровне первичного звена здравоохранения.

2. Методы ОИИ в обработке изображений в офтальмологии на основе нейронных сетей глубокого обучения. 2.1. Методы библиометрического анализа проблемной области. В исследовании авторами был проведён обзор публикаций, за прошедшие 5 лет, с 2018 до 2023 г., размещённых как в онлайн-библиотеках (Scopus, IEEE Digital Library, PubMed, ScienceDirect), так и отобранных с помощью поискового сервиса Google. В последнем случае, публикация должна была индексироваться в научных базах Scopus или Web of Science. Публикации подбирались с использованием поиска по ключевым словам и их комбинациям: “explainable artificial intelligence”, “explainable ai”, “xai”, “eye disease”, “glaucoma”, “diabetic retinopathy”, “cataract”, “retina”, “fundus image”, “eye image”, “ophthalmology”, “ophthalmoscopy”, “oct”. Результатами поиска стали 81 публикация, из которых отобрано 33 работы. Критериями отбора стали упоминание диагностики одного из офтальмологических заболеваний как цели исследования и применение методов ОИИ как визуального объяснения. Распределение исследований по годам публикаций и более подробное деление по исследуемым заболеваниям представлено на диаграммах ниже (рис. 4, 5).

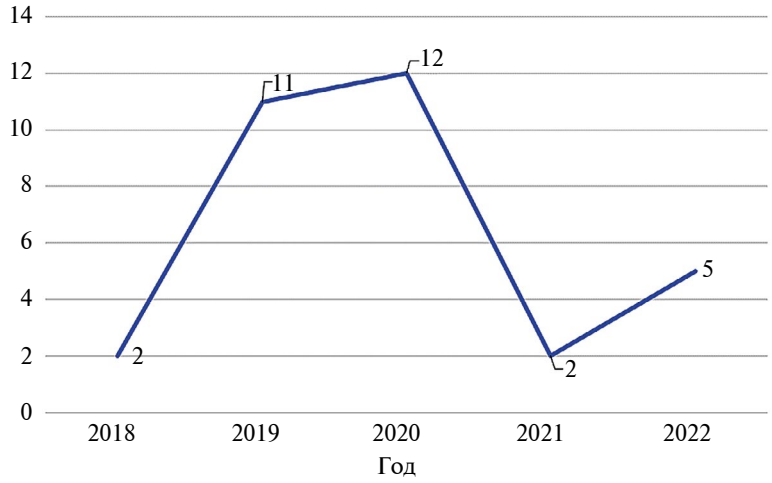

Рис. 4. Количество публикаций по годам

На основании представленных на диаграммах распределений можно сделать вывод о том, что количество исследований с помощью ОИИ в классификации офтальмологических заболеваний невелико по сравнению, например, с использованием ОИИ для анализа КТ/МРТ-снимков головного мозга (n = 42) или грудной клетки (n = 33), проиндексированных в систематическом обзоре [17]. Стоит отметить, что не обнаружено существенной разницы между количеством исследований применения методов объяснения для классификации глаукомы (n = 13) и диабетической ретинопатии (n = 12). Однако нами не было обнаружено отдельных исследований с помощью ОИИ при классификации катаракты. Проиндексированы исследования, содержащий использование методов объяснения в поиске редких глазных заболеваний на основании анализа фото/ОКТ-снимков (n = 4). Отдельно в обзоре упоминаются работы, посвященные улучшению качества исходных снимков при анализе изображений глазного дна и зрительного нерва (n = 3).

2.2. Методы ОИИ. 2.2.1. Определение. Таксономия. Методы объяснения результатов работы ИНС являются основной темой, когда речь идет о применении ОИИ в какой-либо задаче. Рассуждая об этом, мы говорим об использовании одного из многих методов объяснения результата работы ИНС (методов ОИИ).

Для правильного определения ОИИ необходимо найти ключевые термины в этой области — интерпретируемость и объяснимость, разобраться в их соотношении.

Интерпретируемость (interpretability) — способность модели машинного обучения к представлению механизма своей работы понятным для пользователя образом.

В то же время сама по себе интерпретируемость не может решить проблему объяснимости применительно к ИНС, но полезна для улучшения обучаемости [26]. Часто в литературе вместо слова интерпретируемость используется близкое ему по значению — прозрачность.

Прозрачность (transparency) — свойство модели быть понятной априори, без применения методов интерпретации. Выделяют прозрачные (линейная и логическая регрессия, деревья решений, метод k-ближайших соседей, байесовские модели, обобщенные аддитивные модели) и непрозрачные модели (ансамбли деревьев решений, метод опорных векторов, ИНС) машинного обучения [27].

Объяснимость (explainability) — свойство модели, относимое к процессу представления результатов ее работы пользователю в виде понятного интерфейса. Иными словами, краткое описание причины принятия моделью определенного решения, представленное, чаще всего, в интерактивном виде, но не позволяющее понять полный алгоритм его принятия. Таким образом, методы объяснения результата (методы ОИИ) — это внешние (model-agnostic) или использующие внутреннюю структуру объекта объяснения (model-specific) методы, позволяющие получить визуальное представление (visual explanation) о работе исследуемой модели ИНС (объяснение “черного ящика”) путем вывода интерактивного интерфейса. Объяснения могут даваться как в отношении модели в целом (global explanations), так и относительно работы модели на одном конкретном примере (local explanations). Объяснение результатов работы модели является постфактум- объяснением (post-hoc explanation).

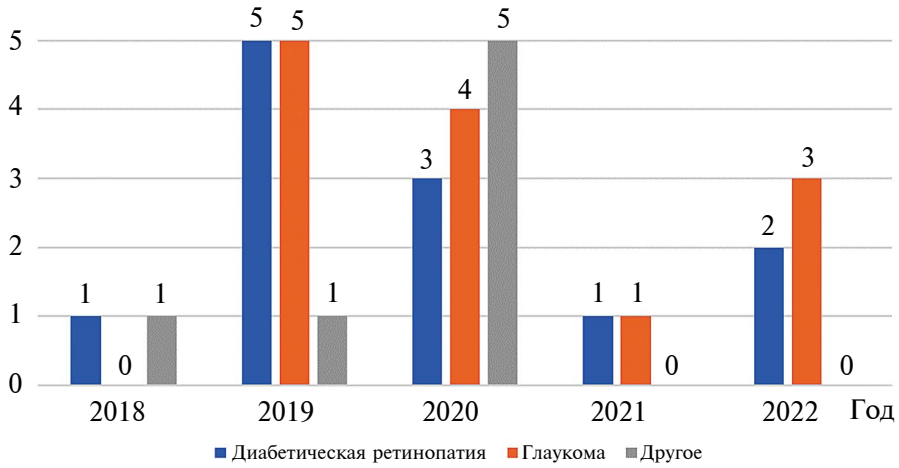

Рис. 5. Количество публикаций по годам с разбиением на глазные заболевания

Существует множество таксономий методов ОИИ [28—30]. В нашем исследовании будем использовать таксономию и определения, введённые в [27, 31].

Среди отобранных нами работ применялись следующие методы объяснения: CAM ($n = 11$), Grad-CAM и его вариации ($n = 8$), GBP ($n = 3$), LIME ($n = 2$), SHAP ($n = 1$), MIL ($n = 2$), IG ($n = 1$), TA ($n = 1$).

Методы LIME, SHAP, CAM, Grad-CAM, GBP, IG, согласно универсальной таксономии, предложенной в [31], относятся к методам, основанным на вычислении важности признаков. Этими признаками для задач работы с изображениями являются пиксели и их группировки на изображении. При этом методы, имеющие в основе своей работы создание суррогатной выборки (LIME) или компоненты теории игр (SHAP), могут применяться к любой модели машинного обучения. В то же время методы (CAM, Grad-CAM, GBP, IG), представляющие объяснения за счет вычисления карт активации классов и обратного распространения градиента, могут давать объяснения лишь для ИНС. Ниже приведено краткое описание принципов работы приведенных методов.

2.2.2. Метод CAM. Метод карты активации класса CAM (class activation mapping) был предложен в 2016 г. Принцип работы метода основан на создания карты локализации искомого класса на исследуемом изображении путем расчета взвешенной суммы карт активации классов последнего свёрточного слоя в объясняемой модели [32].

Метод CAM может применяться практически для всех архитектур ИНС, имеющих в своем составе свёрточные слои. Активно используется при объяснении работы ИНС в задачах классификации и поиска объектов на изображениях, например при объяснении классификации изображений свёрточной нейронной сетью VGG-16. Архитектура VGG-16 представлена пятью блоками свёрточных слоев и следующими за каждым из блоков экстракторами признаков — пятью слоями пулинга, а завершают цепочку три полносвязных слоя перед последним слоем Softmax.

Использование CAM для работы с любой архитектурой ИНС представляет собой имплементацию метода к последнему сверточному слою. Для VGG-16 — это последний свёрточный слой пятого блока (Conv 5-3). Размерность этого слоя свертки составляет $i \times j \times 512$, где i и j — размерность свёртки, а 512 — количество каналов (карт признаков — A_k), сопоставленных с метками классов набора данных для обучения. Так, для VGG-16, обученной на наборе данных ImageNet, количество меток составит 1000.

Карта активации целевого класса M_c показывает с помощью интенсивности цвета на тепловой карте важность пикселя f_k в точке с координатами (x, y) в полученном предсказании, где w_c^k — вес важности предсказанного класса c:

$$S_c = \sum_{x,y}\sum_{k} w_c^k f_k(x,y), \tag{2.1}$$

$$M_c(x,y) = \sum_{k} w_c^k f_k(x,y). \tag{2.2}$$

Обобщенный алгоритм работы метода CAM приведен на рис. 6, а схема получения визуального объяснения работы ИНС — на рис. 7.

Algorithm 1 Class activation mapping.

Require: Image $I^C(H,W)$; Network N
Ensure: Replace FC layer with average pooling layer in Network N
procedure CAM(I, N)
N(I) ▷ Input image into network
$W_k^c \leftarrow (w_1, w_2, w_3, ..., w_k)$ ▷ Get weights from average polling layer
$F_k^c \leftarrow (f_1(x,y), f_2(x,y), f_3(x,y), ..., f_k(x,y))$ ▷ Feature map of the last convolution layer
$M_c(x,y) = \sum_k w_k^c f_k(x,y)$ ▷ Weighted linear summation
$M_c(x,y) = \frac{1}{HW} \sum_i^{H} \sum_j^{W} M_c(x,y)$ ▷ Normalize and up-sample to Network input size
$M_c(x,y) = \text{RELU}(M_c(x,y))$ ▷ Final image heat map
end procedure

Рис. 6. Алгоритм работы метода CAM

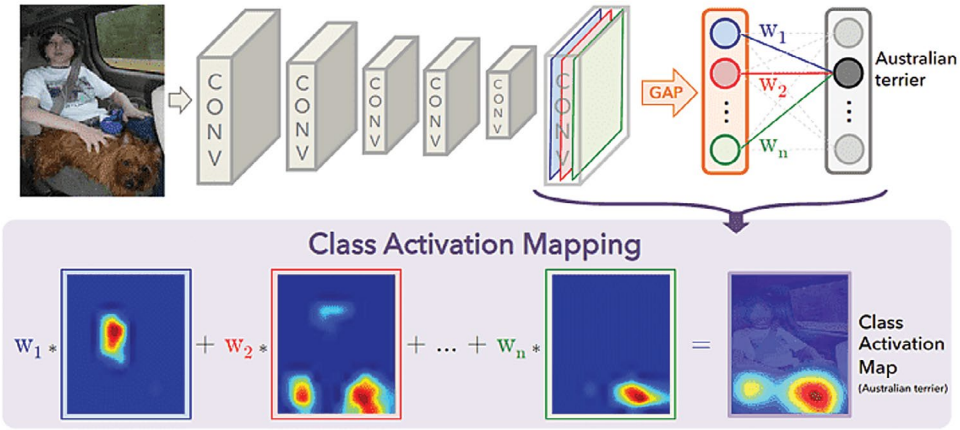

Рис. 7. Получение визуального объяснения с помощью метода CAM

2.2.3. Метод Grad-CAM. Метод градиентной карты активации класса Grad-CAM (gradient-class activation mapping) предложен в 2017 г. [33] и является обобщённой версией CAM, лишенной необходимости замены всех полносвязных слоев свёрточными. Grad-CAM использует вычисление градиента целевого класса для создания карт активации классов.

Алгоритм Grad-CAM (созданный на основе CAM) похож на предшествующий, но имеет несколько принципиальных отличий. Карты признаков A_k объединяются с помощью глобального среднего значения GAP (global average pooling), для каждого класса c ниже рассчитывается итоговое значение Y. Вычисляется градиент оценки класса Y^c относительно карт признаков A_k, который затем усредняется относительно размеров карт активации Z и определяются веса важности нейрона w_k^c, показывающие важность функции k для класса c:

$$Y^c = \sum_k w_k^c \sum_{i,j} A_{i,j}^k, \tag{2.3}$$

$$w_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{i,j}^k}. \tag{2.4}$$

На финальном этапе создается визуальное объяснение в виде тепловых карт с помощью активации прямого распространения ReLU, применяемое к взвешенной комбинации карт активации. ReLU (rectified linear unit) — это функция активации, широко используемая в нейронных сетях. Она определяется следующим образом: $\text{ReLU}(x) = \max(0, x)$. Если значение входа x положительное, то ReLU просто возвращает его. Если x отрицательное или равно нулю, то ReLU равно нулю, что позволяет избежать проблемы исчезающего градиента, присущей

другим популярным функциям активации (сигмоида и гиперболического тангенса). Применение ReLU необходимо для выделения с помощью интенсивности цвета пикселей, входящих в регионы изображения, которые имеют наибольшую важность в предсказании целевого класса через обнуление отрицательных значений. Схема получения объяснения предсказания ИНС с помощью метода Grad-CAM представлена на рис. 8:

$$L^c_{Grad-CAM} = ReLU\left(\sum_k w^c_k A^k\right). \tag{2.5}$$

Основными достоинствами методов CAM и Grad-CAM являются простота имплементации в объясняемую модель, при этом использование Grad-CAM не нуждается в замене слоев. Эти методы совместимы с различными типами выводов свёрточной нейронной сети для решения задач классификации изображений (визуальное объяснение) и создания подписей к изображениям (текстовое объяснение). Методы не требуют больших вычислительных затрат и значительного времени для получения объяснения и для них имеется репозиторий с открытым кодом [34, 35]. К основному недостатку метода Grad-CAM относят визуальное объяснение в виде грубой цифровой карты, которая не может дать представление о точных границах региона значимых признаков. Также при наличии на изображении нескольких объектов одного класса тепловая карта будет захватывать область между ними. Недостатком метода CAM является необходимость замены полносвязных слоев нейросети свёрточными слоями.

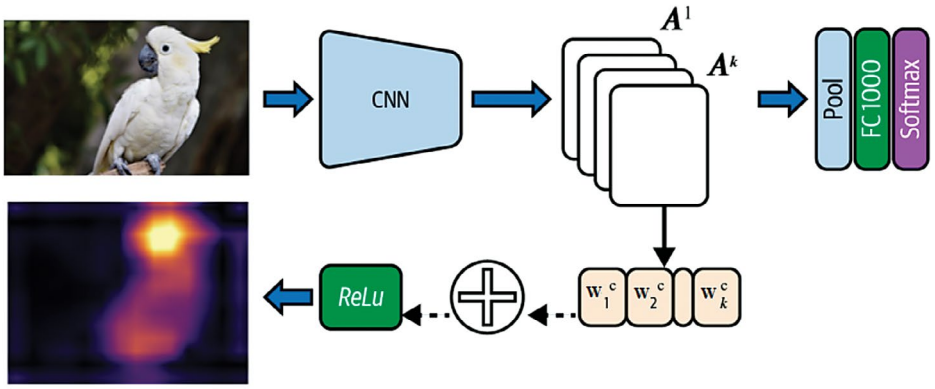

Рис. 8. Получение визуального объяснения с помощью метода Grad-CAM

2.2.4. Метод GBP. Метод управляемого обратного распространения GBP (guided back propagation) был предложен в 2014 г. [36]. Он основывается на идее вычисления обратного распространения градиента и приравнивания отрицательных градиентов к нулю, что оставляет при визуализации объяснения лишь области изображения с наибольшим вкладом в предсказание.

2.2.5. Метод LIME. Метод локальных интерпретируемых модель-агностических объяснений для визуального объяснения результатов работы любой модели машинного обучения, основанный на локальной аппроксимации областей отдельного предсказания путем создания суррогатного набора примеров и признаков, а затем его оптимизации на базе метода Лассо. Визуальный вывод представляется пользователю в виде выделенных цветом областей изображения в зависимости от важности их роли в формировании итогового предсказания [37].

Алгоритм получения объяснения методом Лассо представлен на рис. 9:

$$\varepsilon(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{2.6}$$

где x — признаки, g — модель объяснения из общего семейства моделей G, f — предиктор, π_x — мера близости между экземплярами класса; $L(f, g, \pi_X) = \sum_{z, z' \in Z}^{n} \pi_x(z)$, (2.7)

где g — модель для обучения, z' — экземпляр обучающих суррогатных данных.

Algorithm KernelSHAP algorithm for local explanation

Input: classifier f, input sample x
Output: explainable coefficients from the linear model
1: $z_k \leftarrow$ SampleByRemovingFeature(x)
2: $z_k \leftarrow h_x(z_k)$ ▷ h_x is a feature transformation to reshape to x
3: $y_k \leftarrow f(z_k)$
4: $W_x \leftarrow SHAP(f, z_k, y_k)$
5: `LinearModel(`W_x`).fit()`
6: Return `LinearModel.coefficients()`

Рис. 9. Алгоритм работы метода LIME

Модификацией метода стал GraphLIME [38], предложенный в 2022 г. Данный метод был разработан для применения к графовым нейронным сетям. Его основой послужило использование критерия независимости Гильберта—Шмидта, который представляет собой нелинейный метод выбора признаков. Из-за необходимости генерации суррогатной выборки основным недостатком метода стала скорость получения объяснения, напрямую зависящая от объема выборки [39]. К достоинствам метода LIME относят возможность его применения для объяснения любой модели машинного обучения на любом типе данных, а также открытый доступ к репозиторию с кодом [40].

2.2.6. Метод SHAP. Метод, основанный на теоретико-игровом подходе Шепли, который может быть использован для объяснения любой модели машинного обучения. Он предназначен для локальных объяснений, а его отдельные вариации — для получения глобальных объяснений. Значение Шепли — это принцип оптимальности распределения выигрыша среди игроков коалиции в задачах кооперативных игр [41]. На основе этого подхода С. Лундберг в 2017 г. предложил метод аддитивного объяснения Шепли, заключающийся в объяснении предсказания целевого класса (выигрыша) путем вычисления распределения вклада отдельных признаков (игроков) из общего числа всех признаков (коалиции). В случае объяснения результата задачи классификации изображения, необходимо установить вклад отдельных участков изображения (суперпикселей) в предсказание, определенное ИНС [42].

Получение объяснения методом SHAP происходит путём вычисления значения вклада каждого отдельного признака в общий результата, где φ_i — значение Шепли для признака i, f — модель “черного ящика”, x — входящие данные, z' — это подмножество вектора x' (сохранены только ненулевые элементы) и $|z'|$ равно количеству таких ненулевых элементов вектора z':

$$\phi_i(f,x) = \sum_{z' \in x'} \frac{|z'|!(M - |z'| - 1)!}{M} \left[f_x(z') - f_x\left(\frac{z'}{i}\right) \right]. \tag{2.8}$$

С. Лундберг также предложил несколько вариаций метода для создания объяснения конкретных моделей машинного обучения.

Tree Explainer — метод объяснения для деревьев решений. Идея метода заключается в получении глобального объяснения модели через локальные объяснения. Авторами был создан алгоритм локального объяснения деревьев решений за полиномиальное время на основе точных значений Шепли.

Deep Explainer основан на адаптации метода DeepLIFT и используется для получения объяснений в глубоких нейронных сетях.

Linear Explainer (Linear SHAP) оценивает значения SHAP, используя весовые коэффициенты модели с учетом независимых входных характеристик. Применяется для линейных моделей, если предполагается независимость входных признаков.

Kernel Explainer (Kernel SHAP) выполняет метод атрибуции аддитивных признаков случайным образом, сделав выборку коалиций путем удаления признаков из входных данных и линеаризации влияния модели с помощью ядер SHAP. Универсальность подхода достигается за счет уменьшения количества оценок, необходимых для получения объяснения. Алгоритм работы метода представлен на рис. 10.

Algorithm LIME algorithm for local explanations

Input: classifier f, input sample x, number of superpixels n, number of features to pick m
Output: explainable coefficients from the linear model

$\bar{y} \leftarrow f$.predict(x)
for i in n **do**
  $p_i \leftarrow$ Permute(x) ▷ Randomly pick superpixels
  $obs_i \leftarrow f$.predict(p)
  $dist_i \leftarrow |\bar{y} - obs_i|$
end for
$simscore \leftarrow$ SimilarityScore$(dist)$
$x_{pick} \leftarrow$ Pick$(p, simscore, m)$
$L \leftarrow$ LinearModel.fit$(p, m, simscore)$
return L.weights

Рис. 10. Алгоритм работы метода Kernel SHAP

Достоинством метода является вариативность объектов объяснения и наличие сложившийся библиотеки, находящейся в открытом доступе [43]. Недостатками метода выступают как необходимость больших вычислительных затрат для получения объяснения, так и возможность некорректного объяснения за счет выделения суперпикселей, относящихся к признакам-паттернам (участки фона изображения, повторяющиеся части изображений) набора данных [44]. В [45, 46] отмечено, что значения Шепли как основа метода игнорируют причинно-следственную структуру объясняемой модели. Для решения этой проблемы предлагается использовать ассиметричные значения Шепли, которые за счет применения причинно-следственной информации значительно улучшают точность и полноту объяснения работы исследуемой модели.

2.2.7. Метод MIL. Метод обучения с помощью нескольких примеров MIL (multiple instance learning) как метод визуализации объяснения использует подход на основе патчей для локализации участков изображения, внесших наибольший вклад в предсказание. Для MIL суррогатная выборка состоит из патчей экземпляров целевого класса, при этом патчи маркированы, а экземпляры в них — нет. Патч — это набор $X_i = \left\{x_{i,j} \vee j = \overline{1, N_i}\right\} \subset \mathbb{R}^m$ из N_i экземпляров, где каждый экземпляр является m-мерным вектором признаков [47].

Метод не получил широкого распространения, имеются лишь единичные работы с его использованием в качестве визуального объяснения из-за специфики трактовки принципов его работы и отсутствия открытого репозитория с программной реализацией.

2.2.8. Метод IG. Метод интегрированных градиентов IG (integrated gradients) — метод объяснения, основанный на сопоставлении прогноза модели с ее входными функциями. Он определяет значение важности для значения функции в конкретной точке изображения по формуле путем интегрирования градиента по данному измерению [48]:

$$\phi^{IG}\left(f, x, x'\right) = \left(x_i - x_i^j\right) \int_0^1 \frac{\partial f\left(x' + a\left(x - x'\right)\right)}{\partial x_i} da, \tag{2.9}$$

где $\partial f\left(x\right) / \partial x_i$ — градиент функции $f(x)$ по i-му измерению для входных данных x и базовой точки x'.

Большим недостатком для использования метода интегрированных градиентов для объяснения предсказаний глубоких нейронных сетей является проблема насыщения — градиенты входных признаков могут иметь небольшие значения вокруг выборки, даже если сеть сильно зависит от этих признаков.

2.2.9. Метод TA. Метод обучаемого внимания TA (trainable attention) в отличии от методов, фокусирующихся на определенных участках изображения, выделяет, где и в какой пропорции сеть уделяет внимание входным изображениям для классификации. В дальнейшем эти результаты используются для усиления соответствующих областей и подавления нерелевантных областей [49]. Исходный код TA представлен в открытом репозитории [50].

В табл. 2 приведен сравнительный анализ описанных ранее методов объяснения результатов, принадлежность каждого метода к одной из групп (backpropagation-based, perturbation-based) и типов методов (Model-agnostic, Model-specific) по типу (model-based, post-hoc) и виду (global, local) объяснения.

Таблица 2. Сравнение методов визуального объяснения

Метод XAI	Тип объяснения		Тип метода		Вид объяснения		Группа методов	
	Model -based	Post-hoc	Agnostic	Specific	Local	Global	Backpropagation based	Perturbation based
CAM [32]		+		+	+		+	
Grad-CAM [33]		+		+	+		+	
GBP [36]		+		+	+		+	
LIME [37]		+	+		+			+
SHAP [42]		+	+		+	+		+
MIL [47]		+			+			+
IG [48]		+		+	+		+	
TA [49]	+			+	+		+	

3. Современное состояние в области использования методов ОИИ в офтальмологии. 3.1. Диабетическая ретинопатия. Одним из наиболее частых осложнений у пациентов с сахарным диабетом является диабетическая ретинопатия — хроническое прогрессирующее нейромикрососудистое, последовательно развивающееся заболевание, связанное с поражением сосудов глаза, которое при отсутствии ранней диагностики и должного лечения приводит к частичной или полной слепоте [51]. Заболевание преимущественно поражает больных сахарным диабетом I типа — 27.2% случаев, по сравнению с больными сахарным диабетом II типа, на которых приходится 13% случаев [52]. К основным факторам риска развития диабетической ретинопатии у больных сахарным диабетом относят постоянный повышенный уровень глюкозы в крови (гипергликемия) и повышенное артериальное давление (артериальная гипертензия) [53].

Существует несколько классификаций стадий диабетической ретинопатии. В России общепринятой считается классификация диабетической ретинопатии, выделяющая три стадии (формы) заболевания: непролиферативная (НПДР), преполиферативная (ППДР) и пролиферативная (ПДР) [54]. В отдельных исследованиях по применению ИИ для классификации диабетической ретинопатии можно встретить две основные формы НПДР (NonproliferativeDR (NPDR)), ПДР (ProliferativeDR (PDR)) и три подвида НПДР (Microaneurysms (Mild), Hemorrhages (Moderate), Exudates (Severe)) [55] (рис. 11).

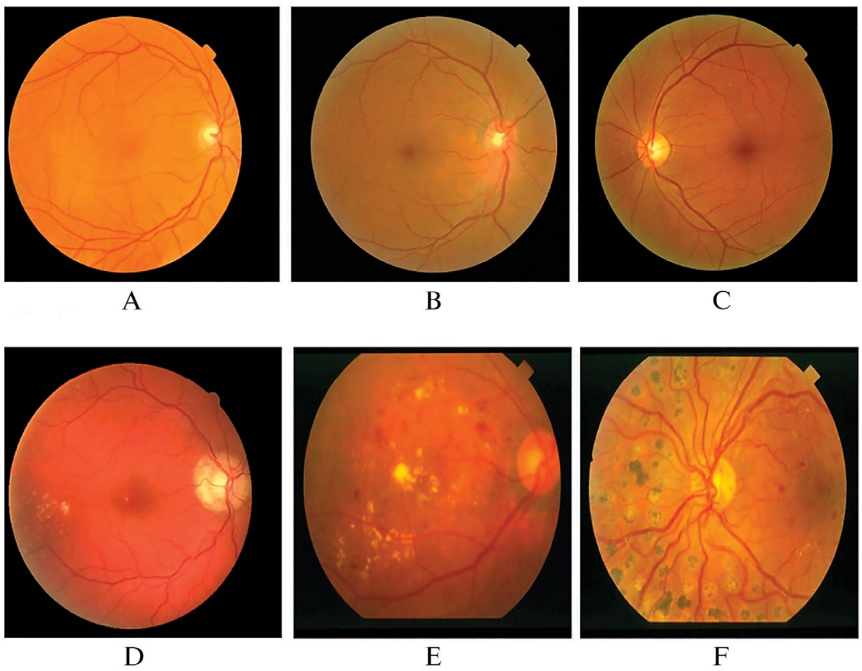

Рис. 11. Стадии диабетической ретинопатии (DR): a) Without DR, b) Early DR, c) Mild DR, d) Moderate NPDR, e) Severe NPDR, f) Proliferative DR [56]

Отобранные для обзора публикации по исследованию применения методов объяснения при анализе медицинских изображений представлены в табл. 3, 4.

Приведенные в табл. 3 DL-составляющие исследований показывают, что меньшинством исследователей ($n = 5$) были выбраны стандартные архитектуры ИНС, представленные в библиотеке моделей Keras [93]. Архитектуры ИНС авторской разработки были выбраны в большинстве случаев ($n = 6$) [73, 80, 82, 84, 88, 89].

В табл. 4 представлены методы объяснения результата и метрики качества объяснения (при наличии), используемые в обозреваемых публикациях. Наиболее часто для объяснения результатов применялся метод CAM ($n = 6$), остальные методы — в единичных случаях: MIL ($n = 2$), Grad-CAM ($n = 1$), IG ($n = 1$). Предпочтение в применении метода CAM во многом связано как с удобством его имплементации, так и возможностями визуализации вывода [94] (рис. 12). Представлена визуализация объяснения классификации диабетической ретинопатии методом CAM.

Анализируя табл. 4, можно увидеть, что в подавляющем большинстве исследований не используются метрики оценки качества объяснения. Это связано с их существующим многообразием и их различием для разных методов ОИИ. Подобный вывод также был сделан в обзоре исследований по применению ОИИ в анализе рентгеновских снимков [95], где 81.56% отобранных публикаций не применяли метрики оценки качества объяснений.

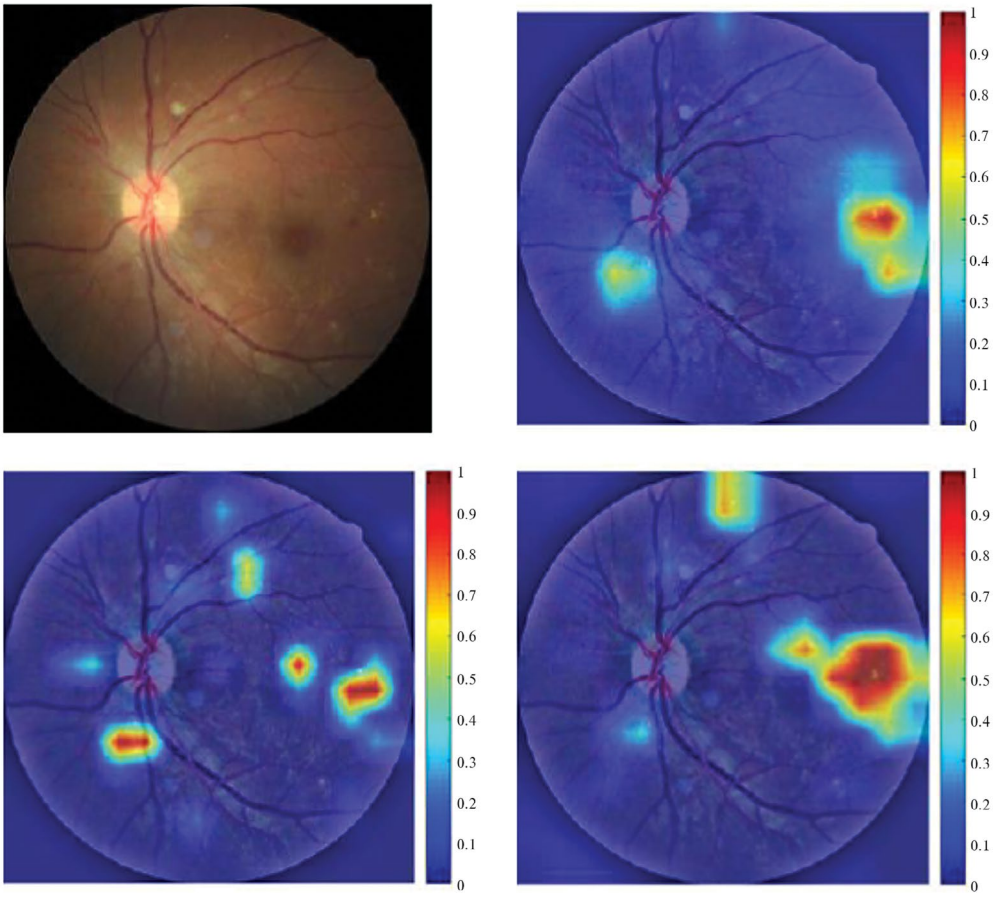

Рис. 12. Результаты использования ОИИ в классификации диабетической ретинопатии [94]

3.2. Глаукома. Глаукомы являются особой группой заболеваний (невропатий) зрительного нерва, характеризующихся постоянно прогрессирующими дегенеративными изменениями ганглиозных клеток сетчатки глаза, обеспечивающих генерацию нервных импульсов. Заболевание сопровождается симптомами повышенного внутриглазного давления и повреждениями структур глаза. Патогенез глаукомы до конца не изучен. Глаукома при несвоевременной диагностике и отсутствии лечения в подавляющем большинстве случаев, приводит к полной потере зрения [96]. Согласно отечественным и зарубежным исследованиям, глаукомы

Таблица 3. Диабетическая ретинопатия

Исследование	Год	DL-блок			
		Характеристика ИНС		Характеристика набора данных	
		Архитектура	Результат	Название, классы, количество изображений	Доступ
[73]	2018	Авторская на основе Inception V4 [74]	Accuracy, % 88.4	Авторский датасет (всего — 2000)	Приватный
[75]	2019	DenseNet121 InceptionResNetV2 InceptionV3 NASNet ResNet50 VGG-16 VGG-19 Xception	Accuracy, % 0.90 0.94 0.93 0.95 0.90 0.94 0.93 0.92	Messidor-2 [76] (всего — 10 274, NPDR — 3564, NRDR — 4630, PDR — 2080)	По запросу
[77]	2019	VGG-16 InceptionV3 Resnet50	AUC, % 90.00 94.53 95.85	Messidor [78]	По запросу
[79]	2019	InceptionV3 Resnet152 Inception-Resnet-V2 Ансамбль моделей	Accuracy, % 87.91 87.20 86.18 88.21	Авторский датасет (всего — 30 244, 16 661 — Negative, 13 583 — Positive)	Приватный
[80]	2019	Авторская (Deep Radiomic CNN Sequencer)	Accuracy, % 73.2	Kaggle diabetic retinopathy dataset [81] (всего — 3662)	Открытый
[82]	2019	Авторская (OCT-NET)	ROC: 93% AUC: 0.99	A2A SD-OCT [83] (всего — 384)	По запросу
[84]	2020	Авторская (DR-GRADUATE)	Quadratic weighted kappa: 0.75 0.71 0.84 0.78 0.74	Kaggle diabetic retinopathy dataset [81] (всего — 3662), Messidor-2 [76] (всего — 10274, NPDR — 3564, NRDR — 4630, PDR — 2080), IDRiD [85] (всего — 516), DMR [86] (всего — 9939), SCREEN-DR (всего — 966)	Открытый По запросу Открытый Открытый Приватный
[87]	2020	AlexNet, VGG-16, ResNet, Inception V3, NASNet, DenseNet, GoogLeNet	Accuracy, % 98.4	Kaggle diabetic retinopathy dataset [81] (всего — 3662)	Открытый
[88]	2020	Авторская (SUNet)	Accuracy, % 0.61	IDRiD [85] (всего — 516)	Открытый
[89]	2021	Авторская (Patho-GAN)	MSE: 0.0093 FID: 81.16 MSE:0.0144 FID: 22.28 MSE: 0.0107 пFID: 20.34	IDRiD [85] (всего — 516), Retinal-Lesions [90], FGADR [91] (всего — 2842)	Открытый Приватный По запросу
[92]	2022	VGG-16 ResNet-18 DenseNet-121	Accuracy, % 48.43 47.86 45.57	APTOS [93] (всего — 35125, Normal — 25809, Mild — 2443, Moderate — 5292, Severe — 873, Proliferative — 708)	Открытый

Таблица 4. Диабетическая ретинопатия (продолжение)

Исследование	XAI-блок		
	Метод объяснения	Оценка объяснимости	
		метрика	значения
[73]	IG	Не использовалось	Не рассчитывалось
[75]	CAM		
[77]	MIL		
[79]	CAM		
[80]			
[82]			
[84]	мMIL	Количественная оценка объяснений предсказанных объектов на основе их пересечений с картами истинности	Oobj-g — 0.506 Oobj — 0.677 Omax — 0.712 Oclass — 0.784 Ogt — 0.526 Oany — 0.729
[87]	CAM	Не использовалось	Не рассчитывалось
[88]			
[89]			

выступают наиболее частой причиной потери зрения среди населения мира вне зависимости от уровня социального благополучия стран проживания. Снимок глазного дна с поражённым глаукомой глазным нервом представлен на рис. 13.

Поскольку единственным доступным неинвазивным методом лечения глаукомы остается снижение уровня внутриглазного давления, то на первый план выходит задача ранней диагностики заболевания. Использование методов компьютерной диагностики за последнее десятилетие возросло в несколько раз.

Учитывая актуальность ранней диагностики глаукомы [97—102], с каждым годом появляется множество исследований по применению методов машинного и глубокого обучения для решения задачи. Благодаря росту использования нейронных сетей в медицине в последние годы установился тренд на преобладание методов, основанных на глубоком обучении в диагностике глаукомы на базе анализа снимков глазного дна [103—107] и ОКТ-снимков [108—110], что показано в представленных систематических обзорах. Точность большинства решений по диагностике глаукомы на основе нейронных сетей составляет более 95% и является приемлемой для практики.

Однако, несмотря на достигнутые успехи, одной из самых актуальных остается проблема отсутствия прозрачности при получении результата, что ведет к снижению доверия к технологии и решениям со стороны врачей-офтальмологов.

Рис. 13. Поражение глаза глаукомой [102]

Представленные в табл. 5 характеристики DL-составляющей исследований говорят о предпочтении собственных архитектур стандартным вариантам из библиотеки Keras.

Таблица 5. Глаукома

Исследование	Год	DL-блок			
		Характеристика ИНС		Характеристика набора данных	
		архитектура	результат	название, классы, количество изображений	доступ
[111]	2019	Авторская (AG-CNN)	Accuracy, % 95.3	LAG-Dataset [112] (всего — 4250, Positive — 1711, Negative — 3143)	По запросу
[113]	2019	VGG-16 Inception-v4 ResNet-152	Accuracy, % 86 91 96	Авторский датасет (всего — 1903, Glaucoma — 1363, Normal — 540)	Приватный
[114]	2019	Авторский (EAMNet)	AUC: 0.88	ORIGA [115] (всего — 650, Glaucoma — 168, Normal — 482)	Открытый
[116]	2019	Conv Layers + Dense Layers, Conv Layers + Random Forest, VGG16 PT + Random Forest, ResNet18 PT + Random Forest, InceptionNet PT + Random Forest	Accuracy, % 95.7 94 95 94.8 94.2	Авторский датасет (всего — 737)	Приватный
[117]	2019	Авторская (Patho-GAN)	Accuracy, % 41.2	LAG-Dataset [112] (всего — 4250, Positive — 1711, Negative — 3143)	По запросу
[118]	2020	Нет доступа	AUC: 0.93	Нет доступа	
[119]	2020	Авторская на основе ResNet	AUC: 0.977 AUC: 0.933	HK Dataset (всего — 975400) Stanford Dataset (всего — 246200)	Приватный
[120]	2020	VGG-16 VGG-19	Accuracy: 0.9463 0.9416	Авторский датасет (всего — 249, Glaucoma — 93, Normal — 156)	Приватный
[121]	2020	Авторская (MCL-Net)	AUC: 0.8698	ORIGA [115] (всего — 650, Glaucoma — 168, Normal — 482), REFUGE [122] (Всего — 400)	Открытый
[123]	2021	Авторская VGG-19	Accuracy, % 94.87 92.77	Авторский датасет (всего — 85497, CNV — 37742, DME — 11840, DRUSEN — 9108, NORMAL — 26807)	Приватный

Окончание таблицы 5

Исследование	Год	DL-блок			
		Характеристика ИНС		Характеристика набора данных	
		архитектура	результат	название, классы, количество изображений	доступ
[124]	2022	DenseNet121 InceptionV3 ResNet50 VGG-16 VGG-19	Accuracy, % 86.81 86.42 94.71 88.63 93.31	LAG-Dataset [112] (всего — 4250, Positive — 1711, Negative — 3143)	По запросу
[125]	2022	Авторская	Accuracy, % 93.5	ORIGA-Light [126] (всего — 650)	По запросу
[127]	2022	Авторская (ANFIS & CNN) AlexNet VGG-16 ResNet	Accuracy, % 96.12 93.61 93.82 96.02	Glaucoma Detection [128] (всего — 650, Positive — 482, Negative — 160)	Открытый

Авторские архитектуры ($n = 7$) рассмотрены в исследованиях [111, 114, 117, 121, 123, 125, 127]. Как следует из табл. 6, метрики оценки объяснимости в данных работах не использовались.

Таблица 6. Глаукома. Сравнительная таблица (продолжение)

Исследование	XAI-блок		
	Метод объяснения	Оценка объяснимости	
		метрика	значение
[111]	Trainable attention	Не использовались	Не рассчитывались
[113]	Grad-CAM		
[114]	CAM		
[116]	Grad-CAM		
[117]	CAM		
[118]	Grad-CAM		
[119]	CAM		
[120]	CAM		
[121]	MCL		
[123]	Grad-CAM		
[124]	LIME		
[125]	CAM		
[127]	LIME		

Применяемые для обучения нейросетей наборы данных также не отличаются разнообразием и в среднем содержат от 600 до 1000 изображений. Наибольшим по количеству примеров, использованных для обучения, стал набор данных LAG, содержащий 4 250 изображений.

3.3. Другие заболевания и методы улучшения результата объяснения. Использование методов объяснения результатов в офтальмологии не ограничивается лишь двумя вышеперечисленными заболеваниями. Существуют также некоторые другие, менее распространенные состояния, к диагностике которых исследователи также применили методы ОИИ. Нельзя забывать о важности наличия качественных снимков для правильной диагностики. В вопросах повышения качества офтальмологических снимков, поисков в них аномалий и устранения артефактов также помогает ОИИ. Речь об этом пойдет в данном разделе.

В [129] предлагается использовать новый Few-Shot Learning фреймворк для обучения ИНС классификации редких заболеваний, таких как передняя ишемическая нейропатия

зрительного нерва. Автор отмечает, что предложенный подход способен устранить проблему необходимости большого количества примеров при традиционном обучении ИНС. Результатами исследования стало достижение показателей AUC в 0.938 при обучении сети на наборе данных, полученном из телемедицинской системы OPHDIAT [130], содержащей 164 660 изображений. Использование метода объяснения GBP позволило минимизировать ошибки детектора и показало возможность классифицировать 37 из 41 заболевания, входящего в OPHDIAT.

Несмотря на большое количество исследований (например, [131—139]), посвящённых объяснению различных медицинских изображений в офтальмологии, обзорные работы по ОИИ в области офтальмологии очень редки. Только в сентябре 2023 г. появился первый обзор по ОИИ [140], гораздо более узкий, чем настоящее исследование. Там также предложено несколько методов ОИИ, которые все чаще применяются в офтальмологических DL-приложениях, преимущественно в задачах анализа медицинских изображений.

Заключение. Рассмотрена эволюция методов ОИИ для распознавания объектов на цифровых изображениях. Рассмотрены классификация методов ОИИ для прикладных интеллектуальных систем, применение ОИИ в нейросетях глубокого обучения, примеры реализованных методов ОИИ в области анализа цифровых изображений на основе нейронных систем глубокого обучения для некоторых офтальмологических заболеваний. Отобранные из различных источников исследования были разделены на несколько групп относительно изучаемых заболеваний и возможностей применений методов ОИИ. Дополнительно для каждой группы исследований составлена проанализированная информация о подходах в области глубокого обучения (архитектура ИНС, точность, набор данных, доступность набора данных) и ОИИ (метод объяснения результата, критерии точности объяснения).

СПИСОК ЛИТЕРАТУРЫ

1. *Fountaine T.*, *McCarthy B.*, *Saleh T.* Building the AI-powered Organization // Harvard Business Review. 2019. V. 97. No. 4. P. 62—73.
2. *Shao Z. et al.* Tracing the evolution of AI in the past decade and forecasting the emerging trends //Expert Systems with Applications. – 2022. – С. 118221. – DOI: 10.1016/j.eswa.2022.118221
3. *Gunning D.*, *Aha D.* DARPA's Explainable Artificial Intelligence (XAI) Program // AI Magazine. 2019. V. 40. No. 2. P. 44—58. DOI: 10.1609/aimag.v40i2.2850.
4. *Fouse S.*, *Cross S.*, *Lapin Z.* DARPA's Impact on Artificial Intelligence // AI Magazine. 2020. V. 41. No. 2. P. 3—8. DOI: 10.1609/aimag.v41i2.5294.
5. *Egger J. Gsaxner C.*, *Pepeet A.* et al. Medical Deep Learning — A Systematic Meta-Review // Computer Methods and Programs in Biomedicine. 2022. V. 221. P. 106874. DOI: 10.1016/j.cmpb.2022.106874.
6. *Shen D.*, *Wu G.*, *Suk H. I.* Deep Learning in Medical Image Analysis // Annual Review of Biomedical Engineering. 2017. V. 19. P. 221—248. DOI: 10.1007/978-3-030-33128-3_1.
7. *Liu Z.*, *Zhichao L.*, *Luhong J.* et al. A Survey on Applications of Deep Learning in Microscopy Image Analysis // omputers in Biology and Medicine. 2021. V. 134. P. 104523. DOI: 10.1016/j.compbiomed.2021.104523.
8. *Xu J.*, *Xi X.*, *Chen J.* et al. A Survey of Deep Learning for Electronic Health Records // Applied Sciences. 2022. V. 12. No. 22. P. 11709. DOI: 10.3390/app122211709.
9. *Feng R.*, *Badgeley M.*, *Mocco J.* et al. Deep Learning Guided Stroke Management: a Review of Clinical Applications // Journal of Neurointerventional Surgery. 2018. V. 10. No. 4. P. 358—362. DOI: 10.1136/neurintsurg-2017-013355.
10. *Abdel-Jaber H.*, *Devassy D.*, *Salam A.* et al. A Review of Deep Learning Algorithms and Their Applications in Healthcare // Algorithms. 2022. V. 15. No. 2. P. 71. DOI: 10.3390/a15020071.
11. *Аверкин А. Н.*, *Ярушев С. А.* Обзор исследований в области разработки методов извлечения правил из искусственных нейронных сетей // Изв. РАН. ТиСУ. 2021. Т. 6. № 6. С. 106—121. DOI 10.31857/S0002338821060044.
12. *Аверкин А. Н.* Объяснимый искусственный интеллект — итоги и перспективы // Авиационные системы в XXI веке: тез. докл. юбилейной Всероссийск. науч.-техн. конф. М.: Государственный научно-исследовательский ин-т авиационных систем, 2022. С. 235—236.
13. *Федунов Б. Е.* Бортовые интеллектуальные системы тактического уровня для антропоцентрических объектов. М.: Де'Либри, 2018. 246 с.
14. *Борисов В. В.* Систематизация нечетких и гибридных нечетких моделей // Мягкие измерения и вычисления. 2020. Т. 29. № 4. С. 98—120.
15. *Talpur N.*, *Abdulkadir S. Alhussian H.* et al. Deep Neuro-Fuzzy System Application Trends, Challenges, and Future Perspectives: A Systematic Survey // Artificial Intelligence Review. 2023. V. 56. No. 2. P. 865—913. DOI: 10.1007/s10462-022-10188-3.

16. *Аверкин А. Н.*, *Ярушев 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.

17. *Van der Velden B. H.M.*, *Kuijf B. H.*, *Gilhuijs H. J.* et al. Explainable Artificial Intelligence (XAI) in Deep Learning-based Medical Image Analysis // Medical Image Analysis. 2022. P. 102470. DOI: 10.1016/j.media.2022.102470.

18. *Anton N.*, *Doroftei B.*, *Curteanu S.* et al. Comprehensive Review on the Use of Artificial Intelligence in Ophthalmology and Future Research Directions // Diagnostics. 2022. V. 13. No. 1. P. 100. DOI: 10.3390/diagnostics13010100.

19. *Srivastava O.*, *Tennant M.*, *Grewal P.* et al. Artificial Intelligence and Machine Learning in Ophthalmology: A Review // Indian J. Ophthalmology. 2023. V. 71. No. 1. P. 11—17. DOI: 10.4103/ijo.ijo_1569_22.

20. *Hogarty D. T.*, *Mackey D. A.*, *Hewitt A. W.* Current State and Future Prospects of Artificial Intelligence in Ophthalmology: A Review // Clinical & Experimental Ophthalmology. 2019. V. 47. No. 1. P. 128—139. DOI: 10.1111/ceo.13381.

21. *Biousse V.*, *Bruce B. B.*, *Newman N. J.* Ophthalmoscopy in the 21st Century: the 2017 H. Houston Merritt Lecture // Neurology. 2018. V. 90. No. 4. P. 167—175. DOI: 10.1212/WNL.0000000000004868.

22. *Minakaran N.*, *de Carvalho E. R.*, *Petzold A.* et al. Optical Coherence Tomography (OCT) in Neuro-ophthalmology // Eye. 2021. V. 35. No. 1. P. 17—32. DOI: 10.1038/s41433-020-01288-x.

23. *Бакуткин В. В.*, *Бакуткин И. В.*, *Зеленов В. А.* Специализированная система предрейсовых медицинских осмотров с применением цифровых технологий // Санитарный врач. 2021. № 5. С. 60—66. DOI: 10.33920/ med-08-2105-07.

24. *Кобринский Б. А.* Автоматизированные регистры медицинского назначения: теория и практика применения. Изд. 2-е, стер. М.; Берлин: Директ-Медиа, 2016. 148 с.

25. *Еремеев А. П.*, *Колосов О. С.*, *Зуева М. В.* и др. Интеграция методов системного анализа и когнитивной графики при ранней диагностике патологий зрения // 20-я Национальная конф. по искусственному интеллекту с международным участием (КИИ-2022). М., 2022. С. 313—329.

26. *Кобринский Б. А.* Интегрированные и гибридные системы искусственного интеллекта: методологические проблемы и вопросы терминологии // Интегрированные модели и мягкие вычисления в искусственном интеллекте ИММВ-2022: Сб. науч. тр. XI Междунар. науч.-практической конф. В 2 т. Коломна: Общероссийская общественная организация “Российская ассоциация искусственного интеллекта”, 2022. С. 37—46.

27. *Arrieta A. B.*, *Díaz-Rodríguez N.*, *Del Ser J.* et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI // Information Fusion. 2020. V. 58. P. 82—115. DOI: 10.1016/j. inffus.2019.12.012.

28. *Schwalbe G.*, *Finzel B.* A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts // Data Mining and Knowledge Discovery. 2023. P. 1—59. DOI: 10.1007/ s10618-022-00867-8.

29. *Speith T.* A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods // ACM Conf. on Fairness, Accountability, and Transparency. Seoul, 2022. P. 2239—2250. DOI: 10.1145/3531146.3534639.

30. *Saeed W.*, *Omlin C.* Explainable AI (XAI): A Systematic Meta-survey of Current Challenges and Future Opportunities // Knowledge-Based Systems. 2023. V. 263. P. 110273. DOI: 10.1016/j.knosys.2023.110273.

31. *Clement T.*, *Kemmerzell N.*, *Abdelaal M.* et al. XAIR: A Systematic Metareview of Explainable AI (XAI) Aligned to the Software Development Process // Machine Learning and Knowledge Extraction. 2023. V. 5. No. 1. P. 78—108. DOI: 10.3390/make5010006.

32. *Zhou B.*, *Khosla A.*, *Lapedriza A.* et al. Learning Deep Features for Discriminative Localization // Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016. P. 2921—2929.

33. *Selvaraju R. R.*, *Cogswell M.*, *Das A.* et al. Grad-cam: Visual Explanations from Deep Networks via Gradient-based Localization // Proc. IEEE Intern. Conf. on Computer Vision (ICCV). Venice, 2017. P. 618—626.

34. Sample Code for the Class Activation Mapping // GitHub URL: https://github.com/zhoubolei/CAM (дата обращения: 13.04.2023).

35. Advanced AI Explainability for PyTorch // GitHub URL: https://github.com/jacobgil/pytorch-grad-cam (дата обращения: 13.04.2023).

36. *Springenberg J. T.* Striving for Simplicity: The all Convolutional Net // arXiv Preprint arXiv:1412.6806.014.

37. *Ribeiro M. T.*, *Singh S.*, *Guestrin C.* Why Should I Trust You? Explaining the Predictions of any Classifier // Proc. 22nd ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. San Francisco, 2016. P. 1135—1144. DOI: 10.1145/2939672.2939778.

38. *Huang Q.*, *Yamada M.*, *Tian Y.* et al. Graphlime: Local Interpretable Model Explanations for Graph Neural Networks // IEEE Transactions on Knowledge and Data Engineering. 2022. V. 35. No. 7. P. 6968—6972. DOI: 10.1109/ TKDE.2022.3187455.

39. *Gramegna A.*, *Giudici P.* SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk // Frontiers in Artificial Intelligence. 2021. V. 4. P. 752558. DOI: 10.3389/frai.2021.752558.

40. Lime // GitHub URL: https://github.com/marcotcr/lime (дата обращения: 13.04.2023).
41. *Shapley L. S.* A Value for N-Person Games. AW Kuhn, HW Tucker, ed., Contributions to the Theory of Games II. Santa-Monica, 1953.
42. *Lundberg S. M.*, *Lee S. I.* A Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems. 2017. V. 30.
43. SHAP // GitHub URL: https://github.com/slundberg/shap (дата обращения: 13.04.2023).
44. *Sundararajan M.*, *Najmi A.* The Many Shapley Values for Model Explanation // Intern. Conf. on Machine Learning. PMLR, Carnegie Mellon, 2020. P. 9269—9278.
45. *Frye C.*, *Rowat C.*, *Feige I.* Asymmetric Shapley Values: Incorporating Causal Knowledge Into Model-agnostic Explainability // Advances in Neural Information Processing Systems. 2020. V. 33. P. 1229—1239.
46. *Janzing D.*, *Minorics L.*, *Blöbaum P.* Feature Relevance Quantification in Explainable AI: A Causal Problem // Intern. Conf. on Artificial Intelligence and Statistics. PMLR, Sydney, 2020. P. 2907—2916.
47. *Cheplygina V.*, *de Bruijne M.*, *Pluim J. P.W.* Not-so-supervised: A Survey of Semi-Supervised, Multi-instance, and Transfer Learning in Medical Image Analysis // Medical Image Analysis. 2019. V. 54. P. 280—296. DOI: 10.1016/j.media.2019.03.009.
48. *Sundararajan M.*, *Taly A.*, *Yan Q.* Axiomatic Attribution for Deep Networks // Intern. Conf. on Machine Learning. PMLR, Sydney, 2017. P. 3319—3328.
49. *Jetley S.*, *Lord N. A.*, *Lee N.* et al. Learn to Pay Attention // arXiv preprint arXiv:1804.02391. 2018.
50. Learn to Pay Attention (ICLR'18) // GitHub URL: https://github.com/saumya-jetley/cd_ICLR18_LearnToPayAttention (дата обращения: 13.04.2023).
51. *Демидова Т. Ю.*, *Кожевников А. А.* Диабетическая ретинопатия: история, современные подходы к ведению, перспективные взгляды на профилактику и лечение // Сахарный диабет. 2020. V. 23(1). P. 95—105. DOI: 10.14341/DM10273.
52. *Дедов И. И.*, *Шестакова М. В.*, *Викулова О. К.* Эпидемиология сахарного диабета в Российской Федерации: клинико-статистический отчет по данным федерального регистра сахарного диабета // Сахарный диабет. 2017. V. 20. No. 1. P. 13—41. DOI: 10.14341/DM8664.
53. *Klein R.*, *Klein B. E.K.* Are Individuals with Diabetes Seeing Better? A Long-term Epidemiological Perspective // Diabetes. 2010. V. 59. No. 8. P. 1853—1860. DOI: 10.2337/db09-1904.
54. *Porta M.*, *Kohner E.* Screening for Diabetic Retinopathy in Europe // Diabetic Medicine. 1991. V. 8. No. 3. P. 197—198. DOI: 10.1111/j.1464-5491.1991.tb01571.x.
55. *Yasin S.*, *Iqbal N.*, *Ali T.*, et al. Severity Grading and Early Retinopathy Lesion Detection Through Hybrid Inception-ResNet Architecture // Sensors. 2021. V. 21. No. 20. P. 6933. DOI: 10.3390/s21206933.
56. *Bidwai P.*, *Gite S.*, *Pahuja K.* et al. A Systematic Literature Review on Diabetic Retinopathy Using an Artificial Intelligence Approach // Big Data and Cognitive Computing. MDPI AG, 2022. V. 6. No. 4. P. 152. DOI: 10.3390/bdcc6040152.
57. *Филиппов В. М.*, *Петрачков Д. В.*, *Будзинская М. В.*, *Сидамонидзе А. Л.* Современные концепции патогенеза диабетической ретинопатии // Вестн. офтальмологии. 2021. V. 137. P. 306—313. DOI: 10.17116/oftalma2021137052306.
58. *Sahiledengle B.*, *Assefa T.*, *Negash W.* et al. Prevalence and Factors Associated with Diabetic Retinopathy among Adult Diabetes Patients in Southeast Ethiopia: A Hospital-Based Cross-Sectional Study // Clinical Ophthalmology. 2022. V. 16. P. 3527—3545. DOI: 10.2147/OPTH.S385806.
59. *Barsegian A.*, *Kotlyar B.*, *Lee J.* et al. Diabetic Retinopathy: Focus on Minority Populations // Intern. J. Clinical Endocrinology and Metabolism. 2017. V. 3. No. 1. P. 034. DOI: 10.17352/ijcem.000027.
60. *Avidor D.*, *Loewenstein A.*, *Waisbourd M.* et al. Cost-effectiveness of Diabetic Retinopathy Screening Programs Using Telemedicine: A Systematic Review // Cost Effectiveness and Resource Allocation. 2020. V. 18. P. 1—9. DOI: 10.1186/s12962-020-00211-1.
61. *Борщук Е. Л.*, *Чупров А. Д.*, *Лосицкий А. О.* и др. Организация скрининга диабетической ретинопатии с применением телемедицинских технологий // Практическая медицина. 2018. Т. 16. № 4. С. 68—70. DOI: 1032000/2072-1757-2018-16-4-68-70.
62. *Russo A.*, *Morescalchi F.*, *Costagliola C.* et al. Comparison of Smartphone Ophthalmoscopy with Slit-lamp Biomicroscopy for Grading Diabetic Retinopathy // American J. Ophthalmology. 2015. V. 159. No. 2. P. 360—364. e1. DOI: 10.1016/j.ajo.2014.11.008.
63. *Rajalakshmi R.*, *Arulmalar S.*, *Usha M.* et al. Validation of Smartphone Based Retinal Photography for Diabetic Retinopathy Screening // PloS One. 2015. V. 10. No. 9. P. e0138285. DOI: 10.1371/journal.pone.0138285.
64. *Shekar S.*, *Satpute N.*, *Gupta A.* Review on Diabetic Retinopathy with Deep Learning Methods // J. of Medical Imaging. 2021. V. 8. No. 6. P. 060901—060901. DOI: 10.1117/1.JMI.8.6.060901.
65. *Nadeem M. W.*, *Goh H. G.*, *Hussain M.* et al. Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions // Sensors. 2022. V. 22. No. 18. P. 6780. DOI: 10.3390/s22186780.

66. *Alyoubi W. L.*, *Shalash W. M.*, *Abulkhair M. F.* Diabetic Retinopathy Detection through Deep Learning Techniques: A Review //Informatics in Medicine Unlocked. 2020. V. 20. P. 100377. DOI: 10.1016/j.imu.2020.100377.

67. *Vij R.*, *Arora S.* A Systematic Review on Diabetic Retinopathy Detection Using Deep Learning Techniques // Archives of Computational Methods in Engineering. 2023. V. 30. No. 3. P. 2211—2256. DOI: 10.1007/s11831-022-09862-0.

68. *Skouta A.*, *Elmoufidi A.*, *Jai-Andaloussi S.* et al. Deep Learning for Diabetic Retinopathy Assessments: A Literature Review // Multimedia Tools and Applications. 2023. P. 1—66. DOI: 10.1007/s11042-023-15110-9.

69. *Tajudin N. M.A.*, *Kipli K.*, *Mahmood M. H.* et al. Deep Learning in the Grading of Diabetic Retinopathy: A Review // IET Computer Vision. 2022. V. 16. No. 8. P. 667—682. DOI: 10.1049/cvi2.12116.

70. *Sowmiya R.*, *Kalpana R.* Survey or Review on the Deep Learning Techniques for Retinal Image Segmentation in Predicting/Diagnosing Diabetic Retinopathy // AI-Enabled Multiple-Criteria Decision-Making Approaches for Healthcare Management. IGI Global. 2022. P. 181—203. DOI: 10.4018/978-1-6684-4405-4.ch010.

71. *Durga N.* A Systematic Review on Diabetic Retinopathy and Common Eye Diseases Detection through Deep Learning Techniques // Journal of Positive School Psychology. 2022. V. 6. No. 4. P. 1905—1919.

72. *Alaeddini Z.* A Review on Machine Learning Methods in Diabetic Retinopathy Detection //J. Ophthalmic and Optometric Sciences. 2021. V. 5. No. 1. DOI: 10.22037/joos.v5i1.39216.

73. *Sayres R.*, *Taly A.*, *Rahimy E.* et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy // Ophthalmology. 2019. V. 126. No. 4. P. 552—564. DOI: 10.1016/j.ophtha.2018.11.016.

74. *Krause J.*, *Gulshan V.*, *Rahimy E.* et al. Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy // Ophthalmology. 2018. V. 125. No. 8. P. 1264—1272. DOI: 10.1016/j.ophtha.2018.01.034.

75. *Ahmad M.*, *Kasukurthi N.*, *Pande H.* Deep Learning for Weak Supervision of Diabetic Retinopathy Abnormalities // IEEE 16th Intern. Sympos. on Biomedical Imaging (ISBI 2019). Venice: IEEE, 2019. P. 573—577. DOI: 10.1109/ISBI.2019.8759417.

76. Messidor-2 // ADCIS URL: https://www.adcis.net/en/third-party/messidor2/ (дата обращения: 13.04.2023).

77. *Costa P.*, *Araújo T.*, *Aresta G.* et al. EyeWes: Weakly Supervised Pre-trained Convolutional Neural Networks for Diabetic Retinopathy Detection // 16th Intern. Conf. on Machine Vision Applications (MVA). Tokyo: IEEE, 2019. P. 1—6. DOI: 10.23919/MVA.2019.8757991.

78. *Decencière E.*, *Zhang X.*, *Cazuguel G.* et al. Feedback on a Publicly Distributed Image Database: the Messidor database // Image Analysis & Stereology. 2014. V. 33. No. 3. P. 231—234. DOI: 10.5566/ias.1155.

79. *Jiang H.* et al. An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification // 41st Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin: IEEE, 2019. P. 2045—2048. DOI: 10.1109/EMBC.2019.8857160.

80. *Kumar D.*, *Taylor G. W.*, *Wong A.* Discovery Radiomics with CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy // IEEE Access. 2019. V. 7. P. 25891—25896. DOI: 10.1109/ACCESS.2019.2893635.

81. Diabetic Retinopathy Detection // Kaggle URL: https://www.kaggle.com/c/diabetic-retinopathy-detection (дата обращения: 13.04.2023).

82. *Perdomo O.*, *Rios H.*, *Rodríguez F. J.* et al. Classification of Diabetes-related Retinal Diseases Using a Deep Learning Approach in Optical Coherence Tomography // Computer Methods and Programs in Biomedicine. 2019. V. 178. P. 181—189. DOI: j.cmpb.2019.06.016.

83. *Farsiu S.*, *Chiu S. J.*, *O'Connell R. V.* et al. Quantitative Classification of Eyes With and Without Intermediate Age-related Macular Degeneration Using Optical Coherence Tomography // Ophthalmology. 2014. V. 121. No. 1. P. 162—172. DOI: 10.1016/j.ophtha.2013.07.013.

84. *Araújo T.*, *Aresta G.*, *Mendonça L.* et al. DR| GRADUATE: Uncertainty-aware Deep Learning-based Diabetic Retinopathy Grading in Eye Fundus Images // Medical Image Analysis. 2020. V. 63. P. 101715. DOI: 10.1016/j.media.2020.101715.

85. *Porwal P.*, *Pachade S.*, *Kamble R.* et al. Indian Diabetic Retinopathy Image Dataset (IDRiD): a Database for Diabetic Retinopathy Screening Research // Data. 2018. V. 3. No. 3. P. 25. DOI: 10.3390/data3030025.

86. *Takahashi H.*, *Tampo H.*, *Arai Y.* et al. Applying Artificial Intelligence to Disease Staging: Deep learning for Improved Staging of Diabetic Retinopathy // PloS One. 2017. V. 12. No. 6. P. e0179790. DOI: 10.1371/journal.pone.0179790.

87. *Narayanan B. N.*, *Hardie R. C.*, *De Silva M. S.* et al. Hybrid Machine Learning Architecture for Automated Detection and Grading of Retinal Images for Diabetic Retinopathy // J. Medical Imaging. 2020. V. 7. No. 3. P. 034501—034501. DOI: 10.1117/1.JMI.7.3.034501.

88. *Tu Z.*, *Gao S.*, *Zhou K.* et al. SUNet: A Lesion Regularized Model for Simultaneous Diabetic Retinopathy and Diabetic Macular Edema Grading // IEEE 17th Inter. Sympos. on Biomedical Imaging (ISBI). Iowa City: IEEE, 2020. P. 1378—1382. DOI: 10.1109/ISBI45749.2020.9098673.

89. *Niu Y.*, *Gu L.*, *Zhao Y.* et al. Explainable Diabetic Retinopathy Detection and Retinal Image Generation // IEEE J. Biomedical and Health Informatics. 2021. V. 26. No. 1. P. 44—55. DOI: 10.1109/JBHI.2021.3110593.

90. *Wei Q.*, *Li X.*, *Yu W.* et al. Learn to Segment Retinal Lesions and Beyond // 25th Intern. Conf. on Pattern Recognition (ICPR). Milano: IEEE, 2021. P. 7403—7410. DOI: 10.1109/ICPR48806.2021.9412088.

91. *Zhou Y.*, *Wang B.*, *Huang L.* et al. A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading, and Transferability // IEEE Transactions on Medical Imaging. 2020. V. 40. No. 3. P. 818—828. DOI: 10.1109/TMI.2020.3037771.

92. *Alghamdi H. S.* Towards Explainable Deep Neural Networks for the Automatic Detection of Diabetic Retinopathy // Applied Sciences. 2022. V. 12. No. 19. P. 9435. DOI: 10.3390/app12199435.

93. APTOS: Eye Preprocessing in Diabetic Retinopathy // Kaggle URL: https://www.kaggle.com/code/ratthachat/aptos-eye-preprocessing-in-diabetic-retinopathy/notebook (дата обращения: 13.04.2023).

94. *Jiang H.*, *Yang K.*, *Gao M.* et al. An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification // 41st Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin: IEEE, 2019. P. 2045—2048.

95. *Miró-Nicolau M.*, *Moyà-Alcover G.*, *Jaume-i-Capó A.* Evaluating Explainable Artificial Intelligence for X-ray Image Analysis // Applied Sciences. 2022. V. 12. No. 9. P. 4459. DOI: 10.3390/app12094459.

96. *Weinreb R. N.*, *Aung T.*, *Medeiros F. A.* The Pathophysiology and Treatment of Glaucoma: A Review // Jama. 2014. V. 311. No. 18. P. 1901—1911. DOI: 10.1001/jama.2014.3192.

97. Клинические рекомендации. Глаукома первичная открытоугольная // Рубрикатор клинических рекомендаций URL: https://cr.minzdrav.gov.ru/recomend/96_1 (дата обращения: 06.04.2023).

98. *Малишевская Т. Н.*, *Косакян С. М.*, *Егоров Д. Б.* и др. Региональный регистр пациентов с глаукомой. Методологические аспекты построения, возможности использования в клинической практике // Российский офтальмологический журнал. 2020. Т. 13. № 4. С. 7—35. DOI: doi.org/10.21516/2072-0076-2020-13-4-supplement-7-35.

99. *Мовсисян А. Б.*, *Куроедов А. В.*, *Архаров М. А.* и др. Эпидемиологический анализ заболеваемости и распространенности первичной открытоугольной глаукомы в Российской Федерации // Клиническая офтальмология. 2022. Т. 22(1). С. 3—10. DOI: 10.32364/2311-7729-2022-22-1-3-10.

100. *Tham Y. C.*, *Li X.*, *Wong T. Y.* et al. Global Prevalence of Glaucoma and Projections of Glaucoma Burden Through 2040: A Systematic Review and Meta-analysis // Ophthalmology. 2014. V. 121. No. 11. P. 2081—2090. DOI: 10.1016/j.ophtha.2014.05.013.

101. *Gallo Afflitto G.*, *Aiello F.*, *Cesareo M.* et al. Primary Open Angle Glaucoma Prevalence in Europe: A Systematic Review and Meta-Analysis // J. Glaucoma. 2022. V. 31. No. 10. P. 783—788. DOI: 10.1111/j.1755-3768.2022.0718.

102. *Mahum R.*, *Rehman S. U.*, *Okon O. D.* et al. A Novel Hybrid Approach Based on Deep CNN to Detect Glaucoma Using Fundus Imaging // Electronics. 2021. V. 11. No. 1. P. 26. DOI: 10.3390/electronics11010026.

103. *Thompson A. C.*, *Jammal A. A.*, *Medeiros F. A.* A Review of Deep Learning for Screening, Diagnosis, and Detection of Glaucoma Progression // Translational Vision Science & Technology. 2020. V. 9. No. 2. P. 42—42. DOI: 10.1167/tvst.9.2.42.

104. *Barros D.*, *Moura J. C.C.*, *Freire C. R.* et al. Machine Learning Applied to Retinal Image Processing for Glaucoma Detection: Review and Perspective // Biomedical Engineering Online. 2020. V. 19. No. 1. P. 1—21. DOI: 10.1186/s12938-020-00767-2.

105. *Jin K.*, *Ye J.* Artificial Intelligence and Deep Learning in Ophthalmology: Current Status and Future Perspectives // Advances in Ophthalmology Practice and Research. 2022. P. 100078. DOI: 10.1016/j.aopr.2022.100078.

106. *Guergueb T.*, *Akhloufi M. A.* A Review of Deep Learning Techniques for Glaucoma Detection // SN Computer Science. 2023. V. 4. No. 3. P. 274. DOI: 10.1007/s42979-023-01734-z.

107. *Alawad M.*, *Aljouie A.*, *Alamri S.* et al. Machine Learning and Deep Learning Techniques for optic Disc and Cup Segmentation — A Review // Clinical Ophthalmology. 2022. P. 747—764. DOI: 10.2147/OPTH.S348479.

108. *Ran A. R.*, *Tham C. C.*, *Chan P. P.* et al. Deep Learning in Glaucoma with Optical Coherence Tomography: A Review // Eye. 2021. V. 35. No. 1. P. 188—201. DOI: 10.1038/s41433-020-01191-5.

109. *Raja H.*, *Akram M. U.*, *Hassan T.* et al. Glaucoma Detection Using Optical Coherence Tomography Images: A Systematic Review of Clinical and Automated Studies // IETE J. Research. 2022. P. 1—21. DOI: 10.1080/03772063.2022.2043783.

110. *Perdomo Charry O. J.*, *González F. A.* A Systematic Review of Deep Learning Methods Applied to Ocular Images // Ciencia e Ingenieria Neogranadina. 2020. V. 30. No. 1. P. 9—26. DOI: 10.18359/rcin.4242.

111. *Li L.*, *Xu M.*, *Liu H.* et al. A Large-scale Database and a CNN Model for Attention-based Glaucoma Detection // IEEE Transactions on Medical Imaging. 2019. V. 39. No. 2. P. 413—424. DOI: 10.1109/TMI.2019.2927226.

112. *Li L.*, *Mai X.*, *Xiaofei W.* et al. Attention Based Glaucoma Detection: a Large-Scale Database and CNN Model // Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019. P. 10571—10580 (LAG Dataset).

113. *Kim M.*, *Han J. C.*, *Hyun S. H.* et al. Medinoid: Computer-aided Diagnosis and Localization of Glaucoma Using Deep Learning // Applied Sciences. 2019. V. 9. No. 15. P. 3064. DOI: 10.3390/app9153064.

114. *Liao W. M.*, *Zou B. J.*, *Zhao R. C.* et al. Clinical Interpretable Deep Learning Model for Glaucoma Diagnosis //IEEE J. Biomedical and Health Informatics. 2019. V. 24. No. 5. P. 1405—1412. DOI: 10.1109/JBHI.2019.2949075.

115. Glaucoma Fundus Imaging Datasets // Kaggle URL: https://www.kaggle.com/datasets/arnavjain1/glaucoma-datasets (дата обращения: 13.04.2023).

116. *Thakoor K. A.*, *Li X.*, *Tsamis E.* et al. Enhancing the Accuracy of Glaucoma Detection from OCT Probability Maps Using Convolutional Neural Networks // 41st Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC). Berlin: IEEE, 2019. P. 2036—2040. DOI: 10.1109/EMBC.2019.8856899.

117. *Wang X.*, *Xu M.*, *Li L.* et al. Pathology-aware Deep Network Visualization and its Application in Glaucoma Image Synthesis // Medical Image Computing and Computer Assisted Intervention (MICCAI 2019): 22nd Intern. Conf., Proceedings. Shenzhen, I 22. China: Springer International Publishing, 2019. P. 423—431. DOI:10.1007/978-3-030-32239-7_47.

118. *Martins J.*, *Cardoso J. S.*, *Soares F.* Offline Computer-aided Diagnosis for Glaucoma Detection Using Fundus Images Targeted at Mobile Devices // Computer Methods and Programs in Biomedicine. 2020. V. 192. P. 105341. DOI: j.cmpb.2020.105341.

119. *Wang X.*, *Chen H.*, *Ran A. R.* et al. Towards Multi-center Glaucoma OCT Image Screening with Semi-Supervised Joint Structure and Function Multi-task Learning // Medical Image Analysis. 2020. V. 63. P. 101695. DOI: j.media.2020.101695.

120. *García G.*, *del Amor R.*, *Colomer A.* et al. Glaucoma Detection from Raw Circumpapillary Oct Images Using Fully Convolutional Neural Networks // IEEE Intern. Conf. on Image Processing (ICIP). Abu Dhabi: IEEE, 2020. P. 2526—2530. DOI: 10.1109/ICIP40778.2020.9190916.

121. *Zhao R.*, *Li S.* Multi-indices Quantification of Optic Nerve Head in Fundus Image via Multitask Collaborative Learning // Medical Image Analysis. 2020. V. 60. P. 101593. DOI: 10.1016/j.media.2019.101593.

122. *Huazhu F.*, *Fei L.*, *Orlando J. I.* et al. Refuge: Retinal Fundus Glaucoma Challenge. IEEE Dataport. 2019. DOI: 10.21227/tz6e-r977.

123. *Apon T. S.*, *Hasan M. M.*, *Islam A.* et al. Demystifying Deep Learning Models for Retinal OCT Disease Classification using Explainable AI // IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE). Brisbane: IEEE, 2021. P. 1—6. DOI: 10.1109/CSDE53843.2021.9718400.

124. *Chayan T I.*, *Islam A.*, *Rahman E.* et al. Explainable AI based Glaucoma Detection using Transfer Learning and LIME // arXiv preprint arXiv:2210.03332. 2022.

125. *Deperlioglu O.*, *Kose U.*, *Gupta D.* et al. Explainable Framework for Glaucoma Diagnosis by Image Processing and Convolutional Neural Network Synergy: Analysis with Doctor Evaluation // Future Generation Computer Systems. 2022. V. 129. P. 152—169. DOI: 10.1016/j.future.2021.11.018.

126. *Zhang Z.*, *Yin F. S.*, *Liu J.* et al. Origa-light: An Online Retinal Fundus Image Database for Glaucoma Analysis and Research // Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology. Buenos Aires: IEEE, 2010. P. 3065—3068. DOI: 10.1109/IEMBS.2010.5626137.

127. *Kamal M. S.*, *Dey N.*, *Chowdhury L.* et al. Explainable AI for Glaucoma Prediction Analysis to Understand Risk Factors in Treatment Planning // IEEE Transactions on Instrumentation and Measurement. 2022. V. 71. P. 1—9. DOI: 10.1109/TIM.2022.3171613.

128. Glaucoma Detection // Kaggle URL: https://www.kaggle.com/datasets/sshikamaru/glaucomadetection?select=Fundus_Train_Val_Data (дата обращения: 13.04.2023).

129. *Quellec G.*, *Lamard M.*, *Conze P. H.* et al. Automatic Detection of Rare Pathologies in Fundus Photographs Using Few-shot Learning // Medical Image Analysis. 2020. V. 61. P. 101660. DOI: 10.1016/j.media.2020.101660.

130. *Massin P.*, *Chabouis A.*, *Erginay A.* et al. OPHDIAT©: A Telemedical Network Screening System for Diabetic Retinopathy in the Île-de-France // Diabetes & Metabolism. 2008. V. 34. No. 3. P. 227—234. DOI: 10.1016/j.diabet.2007.12.006.

131. *Jang Y.*, *Son J.*, *Park K. H.* et al. Laterality Classification of Fundus Images Using Interpretable Deep Neural Network // J. Digital Imaging. 2018. V. 31. P. 923—928. DOI: 10.1007/s10278-018-0099-2.

132. *Shen Y.*, *Sheng B.*, *Fang R.* et al. Domain-invariant Interpretable Fundus Image Quality Assessment // Medical Image Analysis. 2020. V. 61. P. 101654. DOI: 10.1016/j.media.2020.101654.

133. *Wang R.*, *Fan D.*, *Lv B.* et al. OCT Image Quality Evaluation Based on Deep and Shallow Features Fusion Network // IEEE 17th Intern. Sympos. on Biomedical Imaging (ISBI). Iowa City: IEEE, 2020. P. 1561—1564. DOI: 10.1109/ISBI45749.2020.9098635.

134. *Zhou K.*, *Gao S.*, *Cheng J.* et al. Sparse-gan Sparsity-constrained Generative Adversarial Network for Anomaly Detection in Retinal Oct Image // IEEE 17th Intern. Sympos. on Biomedical Imaging (ISBI). Iowa City: IEEE, 2020. P. 1227—1231. DOI: 10.1109/ISBI45749.2020.9098374.

135. *Singh A.*, *Jothi Balaji J.*, *Rasheed M. A.* et al. Evaluation of Explainable Deep Learning Methods for Ophthalmic Diagnosis // Clinical Ophthalmology. 2021. P. 2573—2581. DOI: 10.2147/OPTH.S312236.

136. *Kermany D.*, *Zhang K.*, *Goldbaum M.* et al. Labeled Optical Coherence Tomography (oct) and Chest X-ray Images for Classification // Mendeley Data. 2018. V. 2. No. 2. P. 651. DOI: 10.17632/rscbjbr9sj.

137. *Montavon G.*, *Lapuschkin S.*, *Binder A.* et al. Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition // Pattern Recognition. 2017. V. 65. P. 211—222. DOI: 10.1016/j.patcog.2016.11.008.

138. *Yang H. L.*, *Kim J. J.*, *Kim J. H.* et al. Weakly Supervised Lesion Localization for Age-related Macular Degeneration Detection Using Optical Coherence Tomography Images // PloS One. 2019. V. 14. No. 4. P. e0215076. DOI: 10.1371/journal.pone.0215076.

139. *Meng Q.*, *Hashimoto Y.*, *Satoh S.* How to Extract More Information with Less Burden: Fundus Image Classification and Retinal Disease Localization with Ophthalmologist Intervention // IEEE J. Biomedical and Health Informatics. 2020. V. 24. No. 12. P. 3351—3361. DOI: 10.1109/JBHI.2020.3011805.

140. *Tan T. F.*, *Dai P.*, *Zhang X.* et al. Explainable artificial intelligence in ophthalmology // Current Opinion in Ophthalmology. 2023. V. 34. No. 5. P. 422—430. DOI: 10.1097/ICU.0000000000000983.