ПРАВИТЕЛЬСТВО МОСКВЫ
ДЕПАРТАМЕНТ ЗДРАВООХРАНЕНИЯ ГОРОДА МОСКВЫ

СОГЛАСОВАНО

Главный внештатный специалист
по лучевой и инструментальной
диагностике
Департамента здравоохранения
города Москвы

_____Ю. А. Васильев
«___» _____ 2023 г.

РЕКОМЕНДОВАНО

Экспертным советом по науке
Департамента здравоохранения
города Москвы № 9

«04» АВГУСТА 2023 г.

ОЦЕНКА ЗРЕЛОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЗДРАВООХРАНЕНИЯ

Методические рекомендации № 30

Москва
2023

УДК 004.89+614.2
ББК 32.813
О 93

Серия «Лучшие практики лучевой и инструментальной диагностики»

Серия основана в 2017 году

Организация-разработчик:

Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»

Составители:

Васильев Ю. А. – канд. мед. наук, главный внештатный специалист по лучевой и инструментальной диагностике ДЗМ, директор ГБУЗ «НПКЦ ДиТ ДЗМ»

Владзимирский А. В. – д-р. мед. наук, заместитель директора по научной работе ГБУЗ «НПКЦ ДиТ ДЗМ»

Омелянская О. В. – руководитель по управлению подразделениями дирекции «Наука» ГБУЗ «НПКЦ ДиТ ДЗМ»

Шулькин И. М. – заместитель директора по перспективному развитию ГБУЗ «НПКЦ ДиТ ДЗМ»

Арзамасов К. М. – канд. мед. наук, руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Никитин Н. Ю. – канд. физ.-мат. наук, научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Пестренин Л. Д. – младший научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ «НПКЦ ДиТ ДЗМ»

Шарова Д. Е. – руководитель отдела инновационных технологий ГБУЗ «НПКЦ ДиТ ДЗМ»

О 93 Оценка зрелости технологий искусственного интеллекта для здравоохранения: методические рекомендации / сост. Ю. А. Васильев, А. В. Владзимирский, О. В. Омелянская [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». – Вып. 123. – М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2023. – 28 с.

Рецензенты:

Лебедев Георгий Станиславович – д. техн. наук, доцент, директор Института цифровой медицины, заведующий кафедрой информационных и интернет-технологий Института цифровой медицины ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России (Сеченовский Университет)

Буренчев Дмитрий Владимирович – д-р мед. наук, заведующий отделением рентгенодиагностических и радиоизотопных методов исследования ГБУЗ «ГКБ им. Е. К. Ерамишанцева»

Методические рекомендации предназначены для организаторов здравоохранения, медицинских работников и инженерного персонала, задействованного в тестировании и внедрении программного обеспечения на основе технологий искусственного интеллекта в составе систем поддержки принятия врачебных решений. В издании излагаются практические методы проведения оценки зрелости программного обеспечения с технологиями искусственного интеллекта для внедрения в медицинские организации.

Данные методические рекомендации разработаны в ходе выполнения научно-исследовательской работы «Научные методологии устойчивого развития технологий искусственного интеллекта в медицинской диагностике»

ISBN

СОДЕРЖАНИЕ

НОРМАТИВНЫЕ ССЫЛКИ

В настоящем документе использованы ссылки на следующие нормативные документы (стандарты):

1. Указ Президента Российской Федерации от 10.10.2019 № 490 «О развитии искусственного интеллекта в Российской Федерации».
2. Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья граждан в Российской Федерации».
3. Постановление Правительства Москвы от 21.11.2019 №1543-ПП «О проведении эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы».
4. Приказ Департамента здравоохранения города Москвы от 16.02.2023 № 134 «Об утверждении Порядка и условий проведения эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы».
5. ГОСТ Р 59277-2020. Системы искусственного интеллекта. Классификация систем искусственного интеллекта.
6. ГОСТ 33707-2016 (ISO/IEC 2382-2015). Информационные технологии. Словарь.
7. ГОСТ Р 59921.1-2022. Системы искусственного интеллекта в клинической медицине. Часть 1. Клиническая оценка.
8. ГОСТ Р 59921.2-2021. Системы искусственного интеллекта в клинической медицине. Часть 2. Программа и методика технических испытаний.
9. ГОСТ Р 59921.4-2021. Системы искусственного интеллекта в клинической медицине. Часть 4. Оценка и контроль эксплуатационных параметров.

ОПРЕДЕЛЕНИЯ

В настоящем документе применены следующие термины с соответствующими определениями:

Жизненный цикл – развитие системы, продукции, услуги, проекта или другой создаваемой изготовителем сущности, от замысла до вывода из эксплуатации.

ИИ-сервис – специальное программное обеспечение на основе технологий искусственного интеллекта для решения определенной медико-диагностической задачи.

Искусственный интеллект (ИИ) – комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе то, в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений.

Набор данных – упорядоченная совокупность данных и соответствующих им метаданных, организованных по определенным правилам.

Эксперимент – эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы (mosmed.ai).

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящем документе применены следующие обозначения и сокращения:

БД	–	базы данных
ГБУЗ «НПКЦ ДиТ ДЗМ»	–	государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы»
ДЗМ	–	Департамент здравоохранения города Москвы
ЕМИАС	–	Единая медицинская информационно-аналитическая система
ЕРИС	–	Единый радиологический информационный сервис
ИИ	–	искусственный интеллект
КТ	–	компьютерная томография
МИС	–	медицинские информационные системы
ММГ	–	маммография
ПО	–	программное обеспечение
РГ	–	рентгенография
ТИИ	–	технологии искусственного интеллекта
ЭКГ	–	электрокардиограмма
CAD	–	англ. Computer-assisted Detection (система компьютерной детекции)
DICOM	–	англ. Digital Imaging and Communications in Medicine (цифровые изображения и передача данных в медицине)
PACS	–	англ. Picture Archiving and Communication System (системы передачи и архивации DICOM-изображений)

ВВЕДЕНИЕ

Цифровизация диагностики позволила существенно расширить возможности дополнительной обработки результатов медицинских исследований. В клинической диагностике наблюдается активное внедрение структурированных опросников, позволяющих получить данные анамнеза и клинического осмотра в машиночитаемом виде. В последующем эти данные могут быть обработаны с применением различных алгоритмов (систем поддержки принятия врачебных решений) для получения предварительного диагноза или прогнозирования течения заболевания.

Предоставление результатов исследований в цифровом виде позволило проводить автоматизированный анализ многих инструментальных видов диагностики с предоставлением врачу результатов предварительной обработки. Например, с появлением цифровых электрокардиографов стали активно внедряться в сами аппараты алгоритмы автоматизированной обработки электрокардиограмм (ЭКГ). В настоящее время в связи с централизацией стали применяться алгоритмы обработки ЭКГ, запускаемые с рабочего места врача. Существует большое количество решений для ЭКГ, основанных на информационных технологиях (ИТ). Аналогичная тенденция применения ИТ-решений отмечается и для медицинских изображений, например, в лучевой диагностике.

С появлением первых цифровых изображений стали возможны их машинная обработка и централизованное хранение. Первые радиологические информационные системы начали появляться еще в 1970-х годах, а уже в 1980-х годах стали появляться алгоритмы экспертных систем, позволяющих на основе правил осуществлять постановку диагноза. Первые PACS-системы с возможностью сохранения медицинских исследований в специализированном формате DICOM появились в 1990-х годах, с этого времени начинают активно развиваться модели постановки диагноза на основе атласов.

Следующий этап развития связан с появлением в 2000-х годах систем компьютерного обнаружения и диагностики (CAD), а затем в следствии резкого увеличения вычислительных мощностей в 2010-м появились алгоритмы глубокого обучения для решения задач диагностики.

Отечественная система здравоохранения активно развивается и за последние годы претерпела масштабную цифровизацию и централизацию. Например, в Москве создан и успешно функционирует Единый радиологический информационный сервис (ЕРИС), объединяющий все отделения лучевой, радионуклидной диагностики, отделения

рентгенэндоваскулярных методов диагностики и лечения медицинских учреждений государственного здравоохранения столицы в единую сеть. В 2019–2020 гг. произошла интеграция ЕРИС с Единой медицинской информационно-аналитической системой (ЕМИАС) г. Москвы в качестве единой цифровой платформы для хранения и доступа к результатам лучевых методов исследования для врачей и пациентов учреждений г. Москвы.

Значительный объем исследований диктует необходимость применения ИТ-решений для оптимизации процесса работы врача-рентгенолога. На сегодняшний день на рынок выходит большое количество решений, основанных на информационных технологиях: системы компьютерного обнаружения и диагностики (CAD-системы), а также решения на основе технологий «искусственный интеллект» (ТИИ). Некоторые из этих решений успешно продемонстрировали свою эффективность, в частности, наши предыдущие исследования, показали, что такие алгоритмы для анализа профилактических исследований по показателям диагностической точности сопоставимы со средними для врача-рентгенолога [1, 2] и могут применяться в клинической практике в качестве первого чтения [3].

Тенденцией последних лет является регистрация программного обеспечения (ПО) на основе технологий искусственного интеллекта (ТИИ) как медицинского изделия: это позволяет использовать эти решения не только в лабораторных условиях и экспериментах [4], но также выводить в состав медицинских услуг. Большое количество зарубежных и отечественных работ направлено на оценку качества работы ПО на ограниченном наборе данных, позволяя объективно сравнивать работу таких решений между собой [5, 6]. Если применение ПО на основе ТИИ на локальном компьютере или в качестве облачного помощника врача по запросу не вызывает сложностей, то масштабное внедрении ТИИ в систему здравоохранения в условии централизации хранения и обработки больших объемов данных порождает вопросы.

Принятие решения о внедрении того или иного ПО на основе ТИИ в практическом здравоохранении сопряжено с большими рисками. Несмотря на то, что медицинское изделие на основе ТИИ успешно прошло этап клинических испытаний, при массовом внедрении могут быть обнаружены технологические дефекты, ограничивающие возможность применения данного ПО [3, 7]. Кроме того, показатели диагностической точности, заявляемые разработчиками для разных решений, могут различаться. Более того, показатели диагностической точности одного и того же решения в лабораторных и реальных условиях также могут отличаться [3] Все вышеперечисленное указывает на необходимость разработки методологии оценки зрелости ПО на основе ТИИ для медицинской диагностики, а также методологии объективного сравнения этих решений между собой.

В научной литературе и мировой практике отсутствуют данные об одновременном применении нескольких решений на основе ТИИ в медицинской диагностике и в отделениях лучевой диагностики в частности. Также отсутствует комплексная методология оценки качества и эффективности их применения в условиях централизации хранения и обработки медицинских данных, что подтверждает актуальность данного вопроса.

Цель настоящих методических рекомендаций – описать практический метод оценки зрелости программного обеспечения на базе ТИИ, разработанный на основании эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы согласно данным за период с 2020 по 2022 гг. Представленный материал будет полезен медицинским и техническим специалистам, задействованным в проведении оценки зрелости ПО на основе ТИИ в составе систем поддержки принятия врачебных решений, проводящим его тестирование и внедрение на базе учреждений здравоохранения.

1. ТЕХНОЛОГИЧЕСКИЙ МОНИТОРИНГ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Оценка зрелости программного обеспечения на основе технологий искусственного интеллекта проводится на базе оценки двух основных параметров:

1) удельного веса технологических дефектов, возникающих при работе ПО на основе ТИИ;

2) показателя диагностической точности (производительности классификатора).

Рубрикатор технологических дефектов, возникающих при работе ПО на основе ТИИ, был разработан и актуализирован в ходе проведения эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы в период с 2020 по 2022 гг. (далее – Эксперимент). Затем был рассмотрен актуальный перечень технологических дефектов, который по результатам Эксперимента доказал свою надежность. Несмотря на то, что этот перечень технологических дефектов разработан на основе данных, полученных при обработке лучевых исследований, он может быть адаптирован под любые другие типы исследований.

1.1. Классификация технологических дефектов, возникающих при работе программного обеспечения на основе технологий искусственного интеллекта

При проведении тестирования программного обеспечения на основе технологий искусственного интеллекта была выработана следующая классификация типов технологических дефектов:

1. Дефекты типа «А» – время обработки одного исследования одним ПО на основе ТИИ превышает определенный временной порог (на примере лучевых исследований – более 6,5 минут).
2. Дефекты типа «Б» – отсутствие результатов работы ПО на основе ТИИ в МИС (на примере лучевых исследований – в ЕРИС ЕМИАС).
3. Дефекты типа «В» – некорректная работа заявленного функционала ПО на основе ТИИ, затрудняющая работу врача или делающая ее выполнение невозможным с надлежащим качеством.
4. Дефекты типа «Г» – дефекты, связанные с отображением области изображения. Такие дефекты применимы для медицинских изображений, но могут быть адаптированы, например, для сигнальных данных (ЭКГ и др.).

5. Дефекты типа «Д» – иные нарушения целостности и содержимого файлов с результатами исследований, обуславливающих ограничение их диагностической интерпретации. Эти дефекты применимы для медицинских изображений, но могут быть адаптированы, например, для сигнальных данных (ЭКГ и др.).

При фиксации дефектов типа «А» временем анализа одного исследования считается время, прошедшее с момента публикации сообщения о доступности исследования для скачивания и анализа ПО на основе ТИИ до момента публикации ПО на основе ТИИ сообщения о доступности результатов анализа исследований МИС.

В случае, если при скачивании ПО на основе ТИИ определяется, что продолжительность времени анализа одного исследования будет превышать установленный порог, то ПО на основе ТИИ не приступает к анализу, а отправляет уведомление об ошибке (например, «ОШИБКА А»). Это позволяет не тратить ресурсы на анализ исследований, результатами которых не воспользуется врач. Однако такой показатель (удельный вес необработанных исследований по причине превышения времени) является важным маркером того, требуется ли вносить изменения в ПО на основе ТИИ или увеличивать мощности для его корректного функционирования.

Дефект типа «Б» возникает в случае, если при скачивании исследования ПО на основе ТИИ определяет, что входные данные не соответствуют назначению и/или функционалу данного ПО. Например, на изображении присутствует нецелевая анатомическая область или изображение искажено, или присутствует не полностью. В этом случае рекомендовано не проводить дальнейший анализ данных, а уведомить об ошибке (например, «ОШИБКА Б»). Анализ ошибок такого типа позволяет установить причину дефекта: связан ли он с ошибками маршрутизации или с особенностью работы источника данных – диагностического устройства.

Анализ дефектов типа «А» и «Б» осуществляется на всем количестве проанализированных исследований. Проверка наличия технологических дефектов, возникающих при работе ПО на основе ТИИ, начиная с типа «В» и заканчивая типом «Е», при наличии технической возможности может быть проведена на всем объеме данных, однако это возможно далеко не всегда, следовательно, мы рекомендуем проводить ее на ограниченном объеме выборки.

В ГБУЗ «НПКЦ ДиТ ДЗМ» был проведен ряд работ по определению необходимого и достаточного количества исследований, содержащегося в выборке баланса классов «норма»/«патология», а также объема исследований с долей допустимых дефектных исследований не более 10 % [8].

В настоящее время принято проводить проверку наличия технологических дефектов типа «В», «Г» и «Д» на выборке в размере 80 исследований [8, 9]. Ниже мы детально рассмотрим подклассы технических дефектов на примере лучевых исследований. Для

исследований других типов часть дефектов из этого перечня может быть неприменима.

При анализе лучевого исследования ПО на основе ТИИ создает текстовое описание в формате DICOM SR, а также графическое изображение с локализацией выявленных изменений в формате DICOM SC. В случае возникновения некорректной работы заявленного функционала ПО на основе ТИИ (дефект типа «В») может проявиться один или несколько дефектов следующих типов:

1. Дефект типа «В1» – фиксируется, если отсутствует дополнительная серия изображений (DICOM SC), созданных ПО на основе ТИИ после обработки исследований.
2. Дефект типа «В2» – фиксируется, если по результатам работы ПО на основе ТИИ отсутствует DICOM SR – текстовое описание и заключение.
3. Дефект типа «В3» – фиксируется, если по результатам работы ПО на основе ТИИ возникают два и более текстовых описания (DICOM SR).
4. Дефект типа «В4» – возникает, если по результатам работы ПО на основе ТИИ в выходных данных отсутствует название сервиса.
5. Дефект типа «В5» – возникает, если по результатам работы ПО на основе ТИИ в выходных данных отсутствует номер версии ПО.

Если в результате работы ПО на основе ТИИ исследования содержат нарушения в области отображения изображения «дефект Г», дефекты отображения делятся на пять групп:

1. Дефект типа «Г1» – изображения, полученные по результатам работы ПО на основе ТИИ, имеют меньший размер по одной из осей (обрезаны).
2. Дефект типа «Г2» – изображения, полученные по результатам работы ПО на основе ТИИ, имеют отличную от оригинала яркость или отличный контраст.
3. Дефект типа «Г3» – возникает, если по результатам работы ПО на основе ТИИ проанализированы не все изображения.
4. Дефект типа «Г4» – возникает, если на изображениях, проанализированных ПО на основе ТИИ, отсутствует надпись «Только для использования в исследовательских/научных целях».
5. Дефект типа «Г5» – возникает, если по результатам работы ПО на основе ТИИ происходят изменения в оригинальной серии исследований.

Дефекты типа «Д» фиксируются в случае, если отсутствуют дефекты типа «А»–«Г», и зафиксированы нарушения целостности и содержимого файлов, полученных по результатам работы ПО на основе ТИИ. Выделено два типа нарушений:

1. Дефект типа «Д1» фиксируется, если по результатам работы ПО на основе ТИИ разметка изображения была произведена за пределами целевого органа.

2. Дефект типа «Д2» фиксируется, если по результатам работы ПО на основе ТИИ были проанализированы некорректная анатомическая область, проекция или серия.

Для целей мониторинга возникновения дефектов при работе ПО на основе ТИИ в ГБУЗ «НПКЦ ДиТ ДЗМ» была разработана платформа мониторинга.

На рисунке 1 представлена схема ПО, предназначенного для контроля возникающих в работе сервисов дефектов.

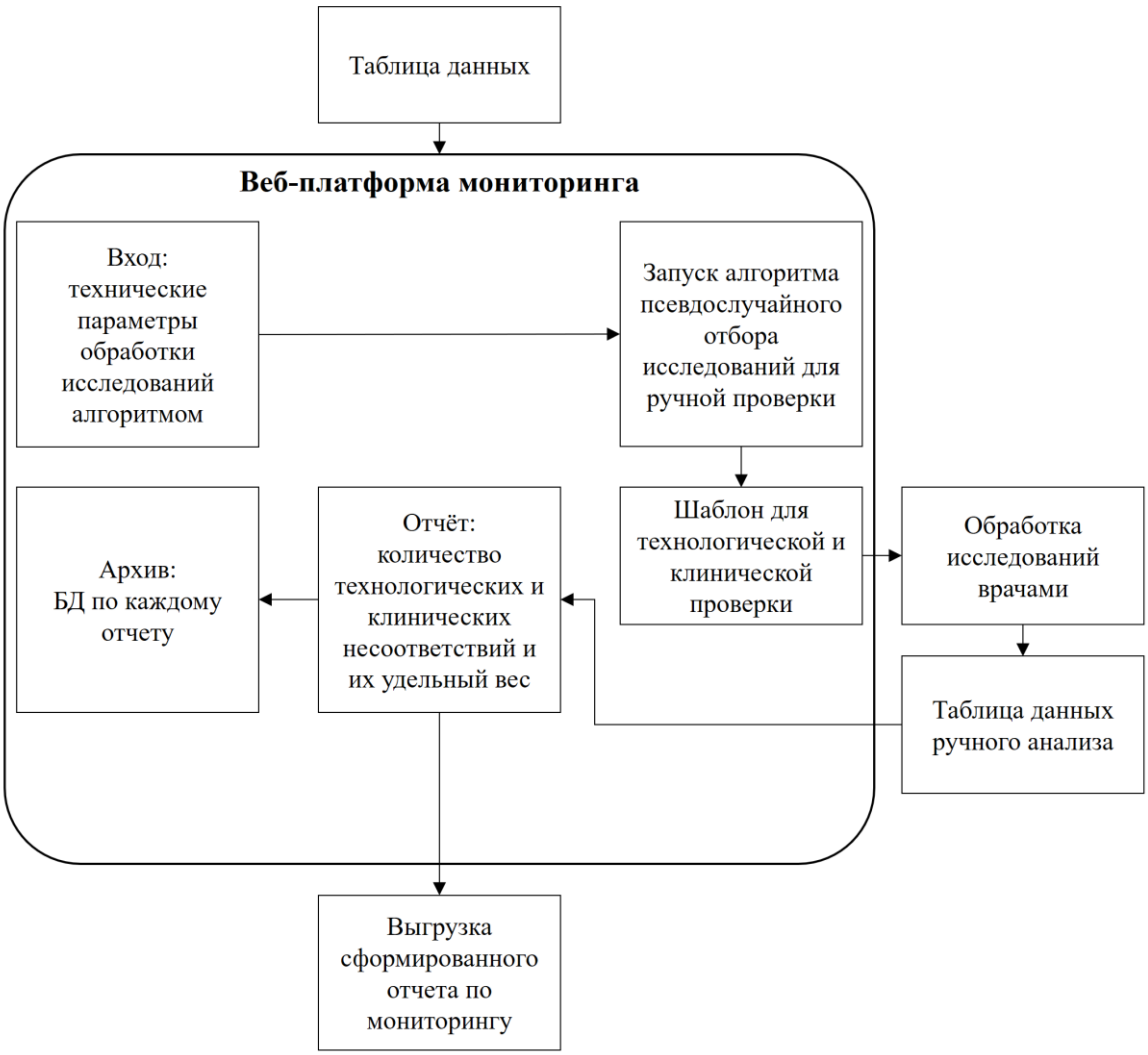


Рисунок 1 – Структурная схема платформы для мониторинга возникновения технологических дефектов при работе ПО на основе ТИИ

На рисунке 2 представлен скриншот Web-интерфейса системы мониторинга возникновения технологических дефектов при работе ПО на основе ТИИ.

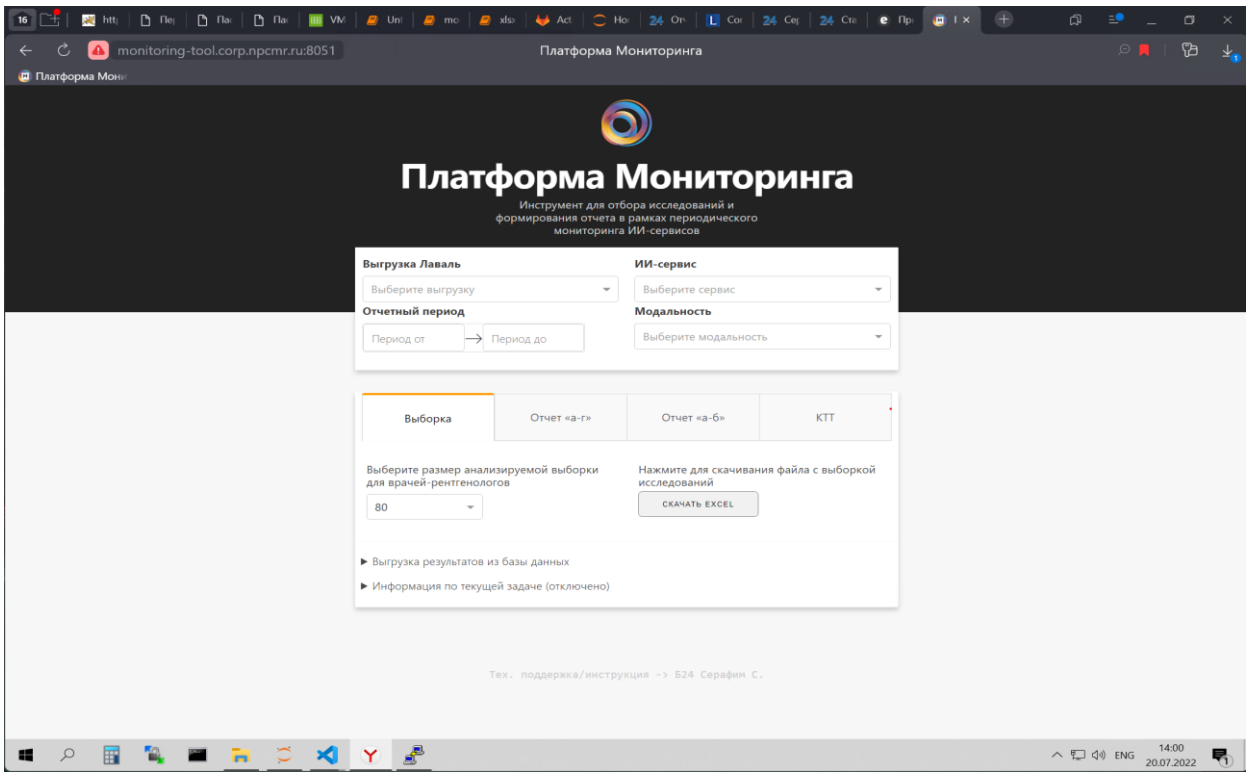


Рисунок 2 – Платформа мониторинга дефектов, возникающих при работе ПО на основе ТИИ

1.2. Диагностическая точность программного обеспечения на основе технологий искусственного интеллекта

При оценке диагностического качества работы ПО на основе ТИИ необходимо рассчитать основные показатели диагностической точности: чувствительность, специфичность и точность. Чувствительность показывает, насколько хорошо алгоритм может выявлять патологию, и отражает количество корректно определенных случаев с патологией. Иными словами, чувствительность показывает вероятность того, что патологический объект будет классифицирован именно как патологический. Специфичность показывает долю отрицательных результатов, которые правильно идентифицированы как таковые, то есть вероятность того, что не патологические объекты будут классифицированы именно как не патологические. Точность – общая вероятность того, что исследование будет правильно классифицировано – определяется как доля всех тестов, которые дают правильный результат.

Описанные выше, а также другие показатели диагностической точности могут быть автоматически рассчитаны при помощи веб-инструмента для выполнения ROC-анализа результатов диагностических тестов (https://roc-analysis.mosmed.ai).

1.3. Расчет показателей диагностической точности (для разработчиков)

Определение диагностической точности классификатора является одной из ключевых проблем при построении моделей классификации. Диагностическая точность модели классификации неразрывно связана с рядом критериев:

1. Наличие или отсутствие подтвержденного результата классификации.
2. Выбор порогового значения при отнесении данных к тому или иному классу.
3. Баланс классов, на котором строилась модель.
4. Объем выборки, на которой была построена модель, и др.

Наличие или отсутствие подтвержденного результата классификации зависит от этапа разработки ПО на основе ТИИ и наличия данных, подтверждающих или опровергающих выставленный класс из других источников (в случае злокачественных образований, например, результатов биопсии). Вопросам баланса классов и объемам выборки посвящено достаточно много работ, в частности исследование С. Ф. Четверикова и соавторов [9].

Для целей оценки зрелости ПО на основе ТИИ основной интерес представляет выбор порогового значения при решении задачи классификации.

Выбор порогового значения может быть осуществлен на основании анализа ROC-кривой с помощью одного из следующих параметров:

1) критерия Неймана-Пирсона [10, 11];
2) площади под ROC-кривой (AUC ROC).

Исходные данные для анализа ROC-кривой как с помощью критерия Неймана-Пирсона, так и с помощью вычисления AUC ROC, размещаются в так называемой «матрице беспорядка» – четырехпольной таблице, содержащей информацию о количестве истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных результатов. Пример такой матрицы представлен в таблице 1.

Таблица 1 – Пример «матрицы беспорядка»

Истинный класс	Прогнозируемый класс		Итого
	Negative	Positive	
Negative	T_n	F_p	C_n
Positive	F_n	T_p	C_p
Итого	R_n	R_p	N

Примечание: Tn и Tp – означает количество истинно отрицательных и истинно положительных результатов классификации соответственно, а Fn и Fp – количество ложноотрицательных и ложноположительных результатов классификации соответственно

Итоговые значения по строкам C_n и C_p – представляют полное количество истинно отрицательных и истинно положительных результатов классификации, а R_n и R_p – полное количество ложноотрицательных и ложноположительных результатов классификации. Общее количество результатов определяется как (1):

$$N = (R_n + R_p) = (C_n + C_p) \quad (1)$$

«Матрица беспорядка» содержит основную необходимую информацию о результатах работы алгоритма классификации, но более информативными являются следующие показатели:

– точность (accuracy) – вычисляется по формуле (2), принимает значения в диапазоне от 0 до 1,0:

$$Accuracy(1 - Error) = \frac{(T_p + T_n)}{(C_p + C_n)} = P(C) \quad (2);$$

– чувствительность (sensitivity) – вычисляется по формуле (3), принимает значения в диапазоне от 0 до 1,0:

$$Sensitivity(1 - \beta) = \frac{T_p}{C_p} = P(T_p) \quad (3);$$

– специфичность (specificity) – вычисляется по формуле (4), принимает значения в диапазоне от 0 до 1,0:

$$Specificity(1 - \alpha) = \frac{T_n}{C_n} = P(T_n) \quad (4);$$

– положительное прогностическое значение (PPV) – вычисляется по формуле (5), принимает значения в диапазоне от 0 до 1,0:

$$PPV = \frac{T_p}{R_p} \quad (5);$$

– отрицательное прогностическое значение (NPV) – вычисляется по формуле (6), принимает значения в диапазоне от 0 до 1,0:

$$NPV = \frac{T_n}{R_n} \quad (6)$$

Показатели, приведенные в уравнениях (2)–(6), являются показателями производительности для одной конкретной точки, причем рабочая точка выбирается таким образом, чтобы свести к минимуму вероятность возникновения ошибки, т.е. *Accuracy (точность)* должна быть максимальной.

Для определения производительности классификатора одним значением рассчитывается показатель площади под ROC-кривой [12]. Разработчиками данного подхода был предложен следующий метод вычисления (7–8):

$$AUC = \sum_i \left\{(1 - \beta_i * \Delta\alpha) + \frac{1}{2}[\Delta(1-\beta) * \Delta\alpha]\right\}, \quad (7)$$

где:

$$\Delta(1-\beta) = (1-\beta_i) - (1-\beta_{i-1}) \quad (8)$$

$$\Delta\alpha = \alpha_i - \alpha_{i-1} \quad (9)$$

Поскольку AUC ROC является частью площади единичного квадрата, ее значение всегда будет находиться в диапазоне от 0 до 1,0. Поскольку случайное угадывание графически выглядит как диагональная линия между точками с координатами (0;0) и (1;1) и имеет площадь под ROC-кривой, равную 0,5, ни один реалистичный классификатор не должен иметь AUC ROC менее 0,5.

Показатель «Площадь под ROC-кривой» обладает важным статистическим свойством: его значение эквивалентно вероятности того, что классификатор оценит случайно выбранный положительный экземпляр выше, чем случайно выбранный отрицательный экземпляр. Это эквивалентно критерию рангов Уилкоксона [13, 14].

Для дальнейшей оценки зрелости ПО на основе ТИИ в качестве показателя диагностической точности применяется показатель площади под ROC-кривой. Для построения этой кривой в качестве истинного значения может применяться как бинарное решение врача (норма или патология), так и любое другое бинарное значение, полученное другим методом (например, с помощью более чувствительного диагностического метода – компьютерной томографии, биопсии и т.д.). Этот показатель может быть применен ко всем диагностическим тестам.

В зависимости от количества исходов может быть использован и другой подход, например, «мультикласс» (например, КТ-степени или категории BI-RADS) или «мультилейбл» (например, сочетанная патология – легочный узел и пневмония). Необходимо отметить, что в этом случае также возможно приведение к бинарной оценке «норма»/«патология» или

попарное сравнение и построение усредненных (macro-/micro-average) ROC-кривых.

Для построения ROC-кривых и их анализа мы рекомендуем использовать ранее разработанную веб-платформу: https://roc-analysis.mosmed.ai [15]. Для работы с предлагаемой веб-платформой необходимо подготовить таблицу, состоящую из столбцов с порядковым номером исследования, вероятностью наличия патологии «result», а также истинным значением «GT», равным «0» или «1».

Для построения матрицы зрелости ПО на основе ТИИ, помимо технологического мониторинга, необходимо провести клиническую оценку сервисов.

2. КЛИНИЧЕСКАЯ ОЦЕНКА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Клиническая оценка работы ПО на основе ТИИ может использоваться в том случае, если, помимо классификатора, данное ПО обладает функционалом подготовки текстового описания и/или графической маркировки выявленных изменений. Клиническая оценка может осуществляться путем ретроспективной проверки исследований, проанализированных ПО на основе ТИИ, в соответствии с данными, полученными из МИС.

Клиническая оценка формируется с помощью ручного просмотра ограниченного числа исследований с целью оценки корректности локализации находок (маркировки) и правильности формирования описания (заключения) с диагностической точки зрения для каждого исследования из выборки.

Клиническая оценка работы сервиса является средним арифметическим оценки маркировки и заключения. Оценка маркировки и заключения – среднее арифметическое всех оценок по выборке исследований, выраженное в процентах по критериям, указанным в таблице критериев и показателей клинической оценки (таблица 2). Необходимо отметить, что в зависимости от цены ошибки ПО на основе ТИИ могут быть изменены баллы за каждый тип ошибки. В текущем виде приведены данные, использующиеся в лучевой диагностике: ложноположительный результат имеет положительную ценность, т.к. лишний раз обратил внимание врача-рентгенолога на исследование. Для других направлений ложноположительный результат может быть критической ошибкой, в этом случае балл может быть снижен вплоть до нуля.

Таблица 2 – Критерии и показатели клинической оценки работы ПО на основе ТИИ по каждому исследованию в отдельности

Критерии оценки	Описание	Баллы	
		Маркировка (M)	Заключение (Z)
Полное соответствие	Отмечены все целевые находки или нормы	1	1
Некорректная оценка (частичный/ избыточный)	Отмечено не менее 1 находки по целевой патологии (или не на всех снимках/проекциях). Неточное оконтуривание находок. Некорректная оценка объема/количества находок	0,5	0,5

Продолжение таблицы 2

Ложно-положительный	Ложная/лишняя находка. Отмечена находка при фактическом полном отсутствии целевой патологии	0,25	0,25
Ложно-отрицательный	Пропуск находки. Не отмечено ни одной находки по целевой патологии при фактическом их наличии	0	0

Клиническая оценка сервиса (K) является средним арифметическим оценки маркировки (M) и заключения (Z). В свою очередь, оценка маркировки (M) и заключения (Z) для выборки объемом n является средним арифметическим оценок по каждому исследованию. При расчете клинической оценки не учитываются исследования с технологическим дефектом, не позволяющим оценить результаты работы ПО на основе ТИИ.

Расчет клинических дефектов также был автоматизирован и возможен при помощи ранее разработанной платформы мониторинга дефектов, возникающих при работе ПО на основе ТИИ (см. рисунок 2) [15].

В зависимости от задачи, решаемой ПО на основе ТИИ, может быть установлено пороговое значение для клинической оценки.

3. МАТРИЦА ЗРЕЛОСТИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Матрица зрелости ПО на основе ТИИ используется для определения наиболее перспективного программного обеспечения. В основе матрицы находятся два параметра успешной эксплуатации ПО на основе ТИИ: качество (совокупность свойств, существенных для использования по назначению) и эффективность (степень соответствия результатов работы, характеризующая приспособленность к достижению цели).

Для качественной составляющей матрицы используется четырехпольная таблица в координатных осях (рисунок 3), где:

– ось 0X – процент технологических дефектов (см. раздел 1.1);

– ось 0Y – единица минус проспективная AUC ROC (см. раздел 1.2);

– граница «0,19» – горизонтальная линия с граничным значением для клинической значимости параметра, равным «1–0,81», в соответствии с методическими рекомендациями [16];

– граница «10» – вертикальная линия на уровне отметки в 10 % технологических дефектов в соответствии с приказом Департамента здравоохранения города Москвы № 134 от 16.02.2023 [8].

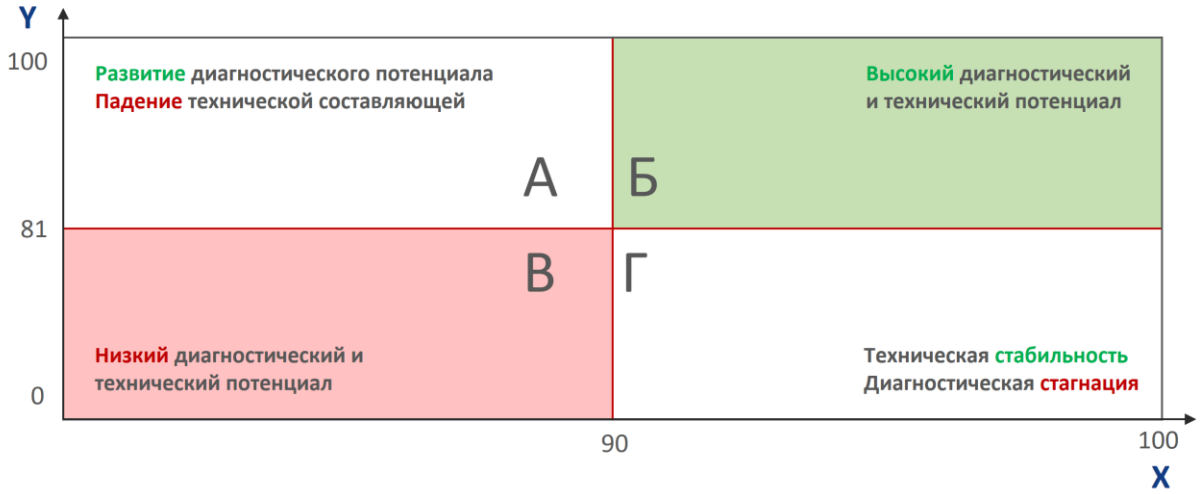


Рисунок 3 – Общий вид матрицы зрелости ПО на основе ТИИ

Физический смысл матрицы зрелости ПО на основе ТИИ заключается в выделении четырех категорий:

1) зона «А», в которой ПО на основе ТИИ развивает и улучшает свою диагностическую составляющую, но теряет свою техническую стабильность;

2) зона «Б», в которой ПО на основе ТИИ обладает на должном уровне техническим и диагностическим свойствами для осуществления качественной работы;

3) зона «В», в которой ПО на основе ТИИ не обладает на должном уровне техническим и диагностическим свойствами для осуществления качественной работы;

4) зона «Г», в которой ПО на основе ТИИ достигает технической стабильности, но не развивает и не улучшает свою диагностическую составляющую.

Качественная составляющая матрицы включает в себя два критерия:

1 Стабильность – свойство ИИ непрерывно сохранять свое качество при заданных воздействиях, характеризуется высоким техническим потенциалом (зоны Б, Г).

2 Долговечность – свойство ИИ сохранять свое качество при заданных воздействиях и условии восстановления свойств, характеризуется высоким диагностическим потенциалом (зоны А, Б).

Для визуализации эффективной составляющей матрицы используется пузырьковая диаграмма (рисунок 4), где:

− диаметр пузырька – клиническая оценка работы ИИ-сервиса;

− орбита пузырька – разброс данных относительно среднего числа по клинической оценке.

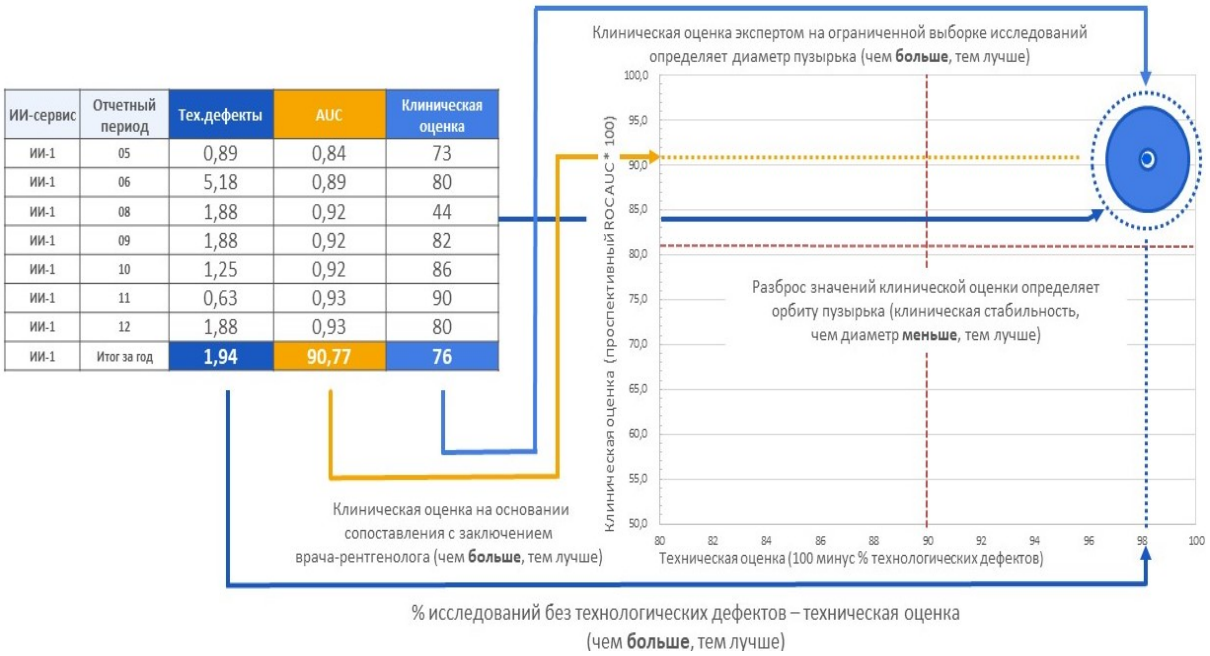


Рисунок 4 – Эффективная составляющая матрицы зрелости ПО на основе ТИИ

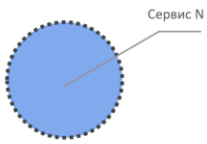
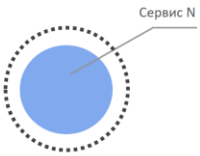
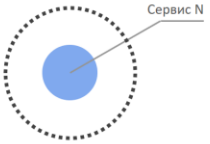
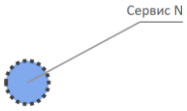
Эффективная составляющая матрицы включает в себя два критерия:

1) пригодность – соответствие ИИ определенным клинико-диагностическим требованиям (диаметр);

2) оптимальность – сбалансированная характеристика клинико-диагностического параметра (орбита).

В зависимости от значений параметров возможны следующие варианты, которые представлены в таблице 3.

Таблица 3 – Варианты эффективной составляющей матрицы

Результат работы ИИ				
Варианты эффективной составляющей	Пригодный и оптимальный результат	Пригодный и неоптимальный результат	Непригодный и неоптимальный результат	Непригодный и оптимальный результат

Предлагаемая матрица зрелости реализована посредством электронной таблицы и может быль легко настроена для работы с другими направлениями медицинской диагностики. Дополнительно настоящая методология представляет возможность оценки динамики развития ПО на основе ТИИ исходя из его «перемещения» в координатах технической и клинической оценки.

Актуальная версия матрицы зрелости ПО на основе ТИИ, участвующих в Эксперименте, опубликована на сайте https://mosmed.ai/ai.

ЗАКЛЮЧЕНИЕ

Представленная методология оценки зрелости ПО на основе ТИИ была апробирована и внедрена в рамках эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы.

Разработанная методология оценки зрелости позволяет не только оценивать решения на основе ТИИ и сравнивать их между собой по величине диагностического и технического потенциала, но и отслеживать изменение этих параметров в динамике и принимать эффективные управленческие решения.

Данная методология легко масштабируется при внедрении ПО на основе ТИИ не только на региональном, но и на федеральном уровне. Она также позволяет выявлять нестабильное ПО на основе ТИИ на ранних этапах и направлять его на доработку, за счет чего достигается экономия ресурсов медицинских организаций и повышается качество работы программного обеспечения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Арзамасов К. М., Семенов С. С., Кокина Д. Ю. [и др.] Критерии применимости компьютерного зрения для профилактических исследований на примере рентгенографии и флюорографии органов грудной клетки // Медицинская физика. 2022. № 4(96). С. 56–63. DOI 10.52775/1810-200X-2022-96-4-56-63 (дата обращения: 12.05.2023).
2. Arzamasov K., Vasilev Y., Vladzymyrskyy A., et al. An International Non-Inferiority Study for the Benchmarking of AI for Routine Radiology Cases: Chest X-ray, Fluorography and Mammography // Healthcare. 2023. № 11. URL: https://www.mdpi.com/2227-9032/11/12/1684 (дата обращения: 12.05.2023).
3. Vasilev Y., Vladzymyrskyy A., Omelyanskaya O., et al. AI-Based CXR First Reading: Current Limitations to Ensure Practical Value // Diagnostics. 2023. Vol. 13, №8. P. 1430. URL: https://doi.org/10.3390/diagnostics13081430 (дата обращения: 12.05.2023).
4. Владзимирский А. В., Васильев Ю. А., Арзамасов К. М. [и др.]. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента. М.: ООО «Издательские решения», 2022. 388 с.
5. Арзамасов К. М., Семенов С. С., Кокина Д. Ю. [и др.]. Критерии применимости компьютерного зрения для профилактических исследований на примере рентгенографии и флюорографии органов грудной клетки // Медицинская физика. 2022. № 4(96). С. 56–63. DOI: 10.52775/1810-200X-2022-96-4-56-63 (дата обращения: 12.05.2023).
6. Арзамасов К. М., Семенов С. С., Кирпичев Ю. С. [и др.]. Оценка диагностической точности ИИ-алгоритмов для выявления злокачественных новообразований по данным маммографии // Медицинская физика. 2022. № 1(93). С. 13–14.
7. Zinchenko V., Chetverikov S., Akhmad E., et al. Changes in software as a medical device based on artificial intelligence technologies // Int J Comput Assist Radiol Surg. 2022. Vol. 17, №10. P. 1969–1977. URL: https://doi.org/10.1007/s11548-022-02669-1 (дата обращения: 12.05.2023).
8. Приказ Департамента здравоохранения города Москвы от 16.02.2023 № 134 «Об утверждении Порядка и условий проведения эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» // Департамент здравоохранения города Москвы: официальный сайт. 2023. 16 февраля. URL: https://mosgorzdrav.ru/ru-RU/document/default/view/2103.html (дата обращения: 10.05.2023).
9. Chetverikov S. F., Arzamasov K. M., Andreichenko A. E., et al. Approaches to Sampling for Quality Control of Artificial Intelligence in

Biomedical Research // Sovremennye tehnologii v medicine. 2023. Vol. 15, № 2. P. 19. DOI: 10.17691/stm2023.15.2.02 (дата обращения: 12.05.2023).

10. Fukunaga K. Introduction to Statistical Pattern Recognition. 2nd Edn. San Diego, California: Academic Press, 1990. 591 p.

11. Therrien C. W. Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics. New York: Wiley, 1989. 251 p.

12. Bradley A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms // Pattern Recogn. 1997. Vol. 30, №7. P. 1145–1159. DOI: 10.1016/S0031-3203(96)00142-2 (дата обращения: 10.05.2023).

13. Hanley J. A., McNeil B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve // Radiology. 1982. Vol. 143, №1. P. 29–36. DOI: 10.1148/radiology.143.1.7063747 (дата обращения: 10.05.2023).

14. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. 2006. Vol. 27, №8. P. 861–874. DOI: 10.1016/j.patrec.2005.10.010 (дата обращения: 10.05.2023).

15. Свидетельство о государственной регистрации программы для ЭВМ № 2022617324 Российская Федерация. Веб-инструмент для выполнения ROC анализа результатов диагностических тестов : № 2022616046 : заявл. 05.04.2022 : опубл. 19.04.2022; заявитель: государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» // EDN ECMPNH.

16. Свидетельство о государственной регистрации программы для ЭВМ № 2023611181 Российская Федерация. Monitoring-ai : № 2022686423 : заявл. 28.12.2022 : опубл. 17.01.2023; заявитель: государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» // EDN IQWUPW.

17. Клинические испытания программного обеспечения на основе интеллектуальных технологий (лучевая диагностика): методические рекомендации / сост. С. П. Морозов, А. В. Владзимирский, В. Г. Кляшторный [и др.] // Серия «Лучшие практики лучевой и инструментальной диагностики». Вып. 57. М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2019. 51 с.

Серия «Лучшие практики лучевой и инструментальной диагностики»

Выпуск 123

Составители:

Васильев Юрий Александрович
Владзимирский Антон Вячеславович
Омелянская Ольга Васильевна
Шулькин Игорь Михайлович
Арзамасов Кирилл Михайлович
Никитин Никита Юрьевич
Пестренин Лев Дмитриевич
Шарова Дарья Евгеньевна

ОЦЕНКА ЗРЕЛОСТИ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ЗДРАВООХРАНЕНИЯ

Методические рекомендации

Технический редактор А. И. Овчарова
Компьютерная верстка Е. Д. Бугаенко

ГБУЗ «НПКЦ ДиТ ДЗМ»
127051, г. Москва, ул. Петровка, д. 24, стр. 1