ВАСИЛЬЕВ Ю.А.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: npcmr@zdrav.mos.ru

АРЗАМАСОВ К.М.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: ArzamasovKM@zdrav.mos.ru

КОЛСАНОВ А.В.,
профессор РАН, д.м.н., профессор, ФГБОУ ВО СамГМУ Минздрава России, Самара, Россия,
e-mail: info@samsmu.ru

ВЛАДЗИМИРСКИЙ А.В.,
д.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: npcmr@zdrav.mos.ru

ОМЕЛЯНСКАЯ О.В.,
ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: npcmr@zdrav.mos.ru

ПЕСТРЕНИН Л.Д.,
ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: PestreninLD@zdrav.mos.ru

НЕЧАЕВ Н.Б.,
к.м.н., ГБУЗ «НПКЦ ДиТ ДЗМ», Москва, Россия, e-mail: NechaevNB@zdrav.mos.ru

ОПЫТ ПРИМЕНЕНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ДАННЫХ 800 ТЫСЯЧ ФЛЮОРОГРАФИЧЕСКИХ ИССЛЕДОВАНИЙ

DOI: 10.25881/18110193_2023_4_54

Аннотация. *Цель: Оценить опыт применения программного обеспечения на основе технологий искусственного интеллекта в рамках Московского эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений.*

Материал и методы: проведено ретроспективное исследование. В работу включены заключения 3 ИИ-сервисов по 822 тысячам флюорографических исследований за период с 05.01.2022 по 29.12.2022. В 28341 исследовании присутствовала патология (3,4%). Оценка проводилась с помощью метрик качества бинарных классификаторов и статистических методов. Произведена оценка метрик в зависимости от порога срабатывания ИИ-сервиса.

Результаты: Отмечается выраженный дисбаланс исследований с нормой и патологией. Получены высокие значения дисбаланс-чувствительных метрик и низкие значения дисбаланс-нечувствительных метрик, что связано с высокой долей ложноположительных и ложноотрицательных результатов. При изменении порога срабатывания можно добиться снижения количества ложноотрицательных результатов. Так, например, один из ИИ-сервисов при пороге 0,05 правильно выявил 46,8% исследований с нормой при отсутствии ложноотрицательных результатов.

Выводы: Количество ложноотрицательных заключений для рассмотренных версий ИИ-сервисов является препятствием для автономного их внедрения в рутинную практику, что требует их доработки. Оптимизацией порога срабатывания сервиса можно добиться безошибочного определения 46,8% исследований с нормой, но ввиду закрытости ИИ-сервисов этот метод ограничен. Дальнейшие варианты оптимизации сервисов требуют дополнительного изучения.

Ключевые слова: *флюорография; рентгенологические исследования; нейронные сети*

Для цитирования: *Васильев Ю.А., Арзамасов К.М., Колсанов А.В., Владзимирский А.В., Омелянская О.В., Пестренин Л.Д., Нечаев Н.Б. Опыт применения программного обеспечения на основе технологий искусственного интеллекта на данных 800 тысяч флюорографических исследований. Врач и информационные технологии. 2023; 4: 54-65. doi: 10.25881/18110193_2023_4_54.*

VASILEV Y.A.,
PhD, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: npcmr@zdrav.mos.ru

ARZAMASOV K.M.,
PhD, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: ArzamasovKM@zdrav.mos.ru

KOLSANOV A.V.,
Prof. of RAS, DSc, Prof., Samara State Medical University, Samara, Russia, e-mail: info@samsmu.ru

VLADZYMYRSKYY A.V.,
DSc, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: npcmr@zdrav.mos.ru

OMELYANSKAYA O.V.,
State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: npcmr@zdrav.mos.ru

PESTRENIN L.D.,
State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: PestreninLD@zdrav.mos.ru

NECHAEV N.B.,
PhD, State Budget-Funded Health Care Institution of the City of Moscow "Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department", Moscow, Russia, e-mail: NechaevNB@zdrav.mos.ru

EXPERIENCE OF APPLICATION ARTIFICIAL INTELLIGENCE SOFTWARE ON 800 THOUSAND FLUOROGRAPHIC STUDIES

DOI: 10.25881/18110193_2023_4_54

Abstract.*Aim: To evaluate the experience of using software based on artificial intelligence technologies as part of the Moscow experiment on the use of innovative technologies in the field of computer vision for the analysis of medical images.*

Material and methods: A retrospective study was conducted. The work includes the conclusion outputs of 3 AI services on 822 thousand fluorographic studies for the period from 05.01.2022 to 29.12.2022. Pathology was present in 28,341 studies (3.4%). The assessment was carried out using quality metrics of binary classifiers and statistical methods. The metrics were assessed depending on the AI services threshold.

Results: There was a pronounced imbalance between studies with norm and pathology. High values of imbalance-sensitive metrics and low values of imbalance-insensitive metrics were obtained, which was associated with a high rate of false positive and false negative results. By changing the threshold, it was possible to reduce the number of false negative results. For example, one of the AI services, with a threshold of 0.05, correctly identified 46.8% of studies with the norm, and with no false negative results.

Conclusions: The number of false negative results for the studied versions of AI services is an obstacle to their autonomous implementation into routine practice, which requires their improvement. By optimizing the service threshold, it is possible to achieve error-free identification of 46.8% of studies with the norm, but due to the closed nature of AI services, this method is limited. Further options for optimizing services require additional study.

Keywords: *fluorography; X-ray examinations; neural networks*

For citation: *Vasilev Y.A., Arzamasov K.M., Kolsanov A.V., Vladzymyrskyy A.V., Omelyanskaya O.V., Pestrenin L.D., Nechaev N.B. Experience of application artificial intelligence software on 800 thousand fluorographic studies. Medical doctor and information technology. 2023; 4: 54-65. doi: 10.25881/18110193_2023_4_54.*

ВВЕДЕНИЕ

Заболевания легких, такие как хроническая обструктивная болезнь легких и инфекции нижних дыхательных путей, занимают 3 и 4 место среди причин глобальной смертности населения по данным Всемирной организации здравоохранения (ВОЗ) [1], а злокачественные новообразования являются ведущей причиной смертности среди пациентов с онкологическими заболеваниями [2]. В Российской Федерации болезни органов дыхания занимают 2 место среди причин смертности от неонкологических заболеваний [3]. Летальность в первый год после установления диагноза рака легкого составляет 47,2% [4].

Рентгенологические исследования грудной клетки оказались одним из самых эффективных методов диагностики как с точки зрения диагностической ценности, так и с точки зрения экономической эффективности и доступности [5-9]. В условиях большого потока пациентов все более широкое применение находят алгоритмы машинного обучения, в первую очередь основанные на анализе визуальных данных, помогающие улучшить диагностику тех или иных патологических состояний, которые могли быть пропущены врачом ввиду низкого качества изображения, небольшого размера изменений (менее 1 см) или человеческого фактора [10–12].

Уровень точности алгоритмов глубокого машинного обучения, которые применяются при анализе рентгенологических исследований, достигает 98% [10, 11], что сопоставимо с точностью врачей лучевой диагностики или превышает ее [13, 14]. В совокупности с экономической составляющей [15] это открывает перспективы улучшения диагностики заболеваний легких при использовании данных алгоритмов, особенно в регионах, не обеспеченных достаточным количеством врачей-рентгенологов [16].

С 19.02.2020 в медицинских организациях города Москвы начат эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы (Эксперимент) [17]. Одним из направлений этого Эксперимента является обработка флюорографических (ФЛГ) исследований сервисами с программным обеспечением (ПО) на основе технологий искусственного интеллекта (именуемыми в Эксперименте и далее по тексту ИИ-сервисы) [18].

Цель исследования: проанализировать промежуточные результаты обработки ФЛГ ИИ-сервисами, которые применялись в рамках Эксперимента, и оценить эволюцию показателей их диагностической точности.

МАТЕРИАЛЫ И МЕТОДЫ

Было проведено ретроспективное исследование работы ИИ-сервисов, участвующих в Эксперименте. Для оценки результатов Эксперимента сравнивались заключения, предоставленные ИИ-сервисом, с врачебными заключениями того же исследования. Для работы были включены 955138 исследований за период с 05.01.2022 по 29.12.2022.

Для сравнения результатов работы ИИ-сервиса с заключениями врачей-рентгенологов данные переведены в форму бинарной классификации: «1» — при наличии описания целевой патологии, «0» — при отсутствии. Целевыми патологиями для ИИ-сервисов были: пневмоторакс, гидроторакс, инфильтрация, очаговое образование, диссеминация, эмфизема легких, наличие полости, кальцинат, патология костей, расширение средостения и кардиомегалия. Так как врачебные заключения были написаны в свободной форме, для их анализа было разработано программное обеспечение (интерпретатор) с использованием технологий обработки естественного языка, результатом работы которого было выявление в заключении рентгенолога описания хотя бы одной из вышеуказанных патологий и приведение заключения к бинарной классификации. На основе работы интерпретатора создана выборка врачебных оценок, принятая за истину.

ИИ-сервисы в рамках Эксперимента предоставляли изображение с маркировкой патологических областей, а также текстовое описание исследования. В рамках настоящей работы использовался только показатель «вероятность наличия патологии в исследовании в целом», который принимал значения в диапазоне от 0 до 100%. Установление факта наличия патологии зависело от порога срабатывания. Если вероятность наличия патологии, выданная ИИ-сервисом, была выше или равна установленному порогу срабатывания, то результат расценивался как положительный, т.е. присваивалось значение «1», в противном случае — «0». Таким образом, формировались

таблицы соответствия оценок, полученных от ИИ-сервисов и по результатам работы интерпретатора текстов протоколов. Важно отметить, что архитектура алгоритмов глубокого обучения, лежащих в основе ИИ-сервисов, была неизвестна.

Для интерпретации ФЛГ применялось ПО от следующих производителей: ООО «ФтизисБиоМед», Россия; ООО «Платформа Третье Мнение», Россия и ООО «Медицинские скрининг системы», Россия. За выбранный период сервисы неоднократно меняли свои версии, поэтому для детального анализа из вышеуказанных продуктов были взяты по три версии ПО: последняя версия, на конец анализируемого периода, и две версии с наибольшим количеством обработанных исследований. В рамках исследования ИИ-сервисы были анонимизированы. В заключительной выборке из 955138 было оставлено 822100 исследований, содержащих результаты работы ИИ-сервисов, а также описание и заключение врача-рентгенолога.

Возраст исследуемых пациентов составил от 18 до 102 лет (средний возраст составил 49,9±17,7 лет). Распределение по полу было следующим: 531160 (64,6%) мужчин и 290833 (35,4%) женщин.

Описание исследований выполнено 571 врачом-рентгенологом. Дизайн исследования предполагал сравнение результатов работы ИИ-сервисов и врачей-рентгенологов, поэтому качество описания исследований врачом-рентгенологом имело большое значение. В связи с этим были отобраны заключения, описанные врачами-рентгенологами, получившими высокие оценки на очередном врачебном аудите (10 врачей). Эти врачи наиболее полно предоставляют описание исследования, данные исследования были выделены в выборку экспертной группы.

Для оценки достоверности работы интерпретатора из выборки врачебных оценок были созданы 5 независимых выборок, содержащих 1000 случайных исследований в каждой, проверенных вручную врачом с 10-летним стажем. На данных заключениях была проведена кросс-валидации интерпретатора. Также из исследований, описанных экспертной группой врачей для последних версий сервисов, создана «выборка оценки интерпретатора». В нее были включены все исследования, в которых интерпретатор выявил описание патологии, а также случайные исследования, в которых интерпретатор и ИИ-сервис не выявили патологию. Общее количество исследований в выборке составило 10% исследований для данной версии сервиса. Данная выборка была проверена вручную врачом с 10-летним стажем на предмет наличия одной из оцениваемых патологий в заключении врача-рентгенолога. Далее результаты ручной оценки сравнивались с результатами интерпретатора статистическими методами.

В работе оценивались следующие метрики диагностической точности: точность, чувствительность, специфичность, Точность отрицательного прогноза, F1-мера, коэффициент Каппа-Коэна, коэффициент корреляции Метьюса, ROC-AUC (AUC) — площадь под кривой ошибок, а также доля ложноотрицательных ответов сервиса [19].

Для статистической обработки результатов работы ИИ-сервисов и оценки работы интерпретаторов использовались тесты Колмогорова-Смирнова и Манна-Уитни. Для сравнения ИИ-сервисов между собой использовался тест МакНемара.

РЕЗУЛЬТАТЫ

На первом этапе была произведена оценка качества работы интерпретатора с целью определения доверительного интервала точности при чтении заключений врача-рентгенолога, которое было проведено методом кросс-валидации на 5 независимых выборках. Полученные результаты представлены в таблице 1. Также установлено, что статистически значимая разница между «выборкой оценки интерпретатора», проведенной вручную, и интерпретатором отсутствовала (Сервис 1 версия 3 p = 0,990; Сервис 2 версия 3 p = 1,000; Сервис 3 версия 3 p = 0,990). Это свидетельствует о том, что метрические показатели работы интерпретатора для заключений врачей экспертной группы близки к 100%.

Произведено сравнение наличия патологии, полученной интерпретатором в выборке, описанной врачами экспертной группы, и выборке, не содержащей заключения врачей экспертной группы, с помощью критерия Хи-квадрат. Статистически значимой разницы между выборками получено не было (p = 0,652).

Описание патологических находок было определено в 28341 исследовании, что составляет 3,4% от всех анализируемых ФЛГ. Диагностические метрики, полученные в процессе Эксперимента, представлены в таблице 2.

Таблица 1 — Оценка работы интерпретатора заключений врача-рентгенолога

Метрика	Патология
Точность	0,990 (95% ДИ от 0,960 до 1,000)
Специфичность	0,990 (95% ДИ от 0,970 до 1,000)
Чувствительность	0,970 (95% ДИ от 0,920 до 1,000)
F1-мера	0,970 (95% ДИ от 0,900 до 1,000)
Точность отрицательного прогноза	0,990 (95% ДИ от 0,980 до 1,000)
Коэффициент корреляции Метьюса	0,960 (95% ДИ от 0,880 до 1,000)
Коэффициент Каппа-Коэна	0,960 (95% ДИ от 0,880 до 1,000)

Таблица 2 — Параметры диагностической точности ИИ-сервисов, полученные в ходе исследования

Сервис	Характер выборки	Всего исследований	Из них с патологией	AUC	F1-мера	Точность	Чувствительность	Специфичность	Точность Отрицательного прогноза	Доля ложноотрицательных ответов сервиса	Коэффициент Каппа-Коэна	Коэффициент Метьюса
Сервис 1 вер. 1	Все исследования	40847	1265 (3,10%)	0,697 [0,682-0,711]	0,137 [0,129-0,144]	0,772 [0,768-0,776]	0,587 [0,560-0,614]	0,778 [0,774-0,782]	0,969 [0,967-0,971]	29,21%	0,088 [0,087-0,088]	0,149 [0,137-0,160]
	Исследования, описанные врачами-экспертами	1361	56 (4,11%)	0,761 [0,693-0,830]	0,156 [0,144-0,221]	0,765 [0,742-0,787]	0,732 [0,616-0,848]	0,766 [0,743-0,789]	0,959 [0,946-0,971]	21,13%	0,112 [0,109-0,113]	0,195 [0,139-0,252]
Сервис 1 вер. 2	Все исследования	119325	4922 (4,12%)	0,648 [0,641-0,655]	0,177 [0,174-0,185]	0,837 [0,835-0,839]	0,433 [0,419-0,446]	0,855 [0,853-0,857]	0,960 [0,957-0,960]	36,20%	0,122 [0,122-0,123]	0,156 [0,149-0,164]
	Исследования, описанные врачами-экспертами	4353	125 (2,87%)	0,713 [0,668-0,758]	0,162 [0,128-0,176]	0,822 [0,811-0,833]	0,592 [0,506-0,678]	0,829 [0,817-0,840]	0,968 [0,966-0,977]	28,57%	0,106 [0,105-0,107]	0,171 [0,130-0,211]
Сервис 1 вер. 3	Все исследования	88612	2433 (2,75%)	0,715 [0,704-0,725]	0,126 [0,125-0,134]	0,766 [0,763-0,769]	0,638 [0,619-0,657]	0,770 [0,767-0,772]	0,973 [0,971-0,974]	26,58%	0,085 [0,084-0,085]	0,155 [0,147-0,163]
	Исследования, описанные врачами-экспертами	2616	83 (3,17%)	0,741 [0,687-0,796]	0,158 [0,134-0,184]	0,773 [0,757-0,789]	0,687 [0,587-0,787]	0,776 [0,760-0,792]	0,967 [0,960-0,976]	23,85%	0,105 [0,103-0,106]	0,183 [0,137-0,226]
Сервис 2 вер. 1	Все исследования	161989	6189 (3,82%)	0,669 [0,663-0,675]	0,230 [0,232-0,245]	0,899 [0,897-0,900]	0,413 [0,401-0,425]	0,918 [0,916-0,919]	0,961 [0,960-0,963]	36,99%	0,193 [0,193-0,194]	0,216 [0,208-0,224]
	Исследования, описанные врачами-экспертами	10491	357 (3,40%)	0,738 [0,711-0,765]	0,272 [0,226-0,271]	0,883 [0,877-0,889]	0,566 [0,514-0,617]	0,894 [0,888-0,900]	0,967 [0,962-0,970]	30,27%	0,205 [0,204-0,206]	0,255 [0,224-0,285]

Таблица 2 — Параметры диагностической точности ИИ-сервисов, полученные в ходе исследования (продолжение)

Сервис	Характер выборки	Всего исследований	Из них с патологией	AUC	F1-мера	Точность	Чувствительность	Специфичность	Точность Отрицательного прогноза	Доля ложноотрицательных ответов сервиса	Коэффициент Каппа-Коэна	Коэффициент Метьюса
Сервис 2 вер. 2	Все исследования	162919	5110 (3,14%)	0,708 [0,701-0,715]	0,263 [0,255-0,269]	0,915 [0,914-0,917]	0,479 [0,465-0,492]	0,930 [0,928-0,931]	0,970 [0,968-0,969]	34,28%	0,227 [0,226-0,227]	0,257 [0,248-0,266]
	Исследования, описанные врачами-экспертами	2917	67 (2,30%)	0,777 [0,715-0,839]	0,308 [0,216-0,314]	0,923 [0,913-0,933]	0,612 [0,495-0,729]	0,930 [0,921-0,940]	0,979 [0,971-0,983]	27,96%	0,239 [0,234-0,239]	0,295 [0,228-0,357]
Сервис 2 вер. 3	Все исследования	91429	2698 (2,95%)	0,667 [0,658-0,677]	0,214 [0,207-0,226]	0,914 [0,912-0,916]	0,394 [0,376-0,413]	0,930 [0,928-0,932]	0,969 [0,969-0,972]	37,70%	0,178 [0,177-0,178]	0,202 [0,191-0,210]
	Исследования, описанные врачами-экспертами	3361	121 (3,60%)	0,776 [0,729-0,821]	0,341 [0,280-0,373]	0,919 [0,910-0,929]	0,603 [0,516-0,690]	0,931 [0,922-0,940]	0,968 [0,957-0,970]	27,98%	0,292 [0,290-0,295]	0,334 [0,277-0,391]
Сервис 3 вер. 1	Все исследования	83391	2971 (3,56%)	0,689 [0,679-0,699]	0,172 [0,165-0,177]	0,823 [0,821-0,826]	0,513 [0,495-0,531]	0,835 [0,832-0,837]	0,964 [0,963-0,966]	32,74%	0,119 [0,118-0,119]	0,168 [0,158-0,177]
	Исследования, описанные врачами-экспертами	1918	111 (5,79%)	0,767 [0,719-0,813]	0,229 [0,216-0,285]	0,749 [0,730-0,769]	0,739 [0,657-0,820]	0,750 [0,730-0,770]	0,950 [0,929-0,954]	20,71%	0,171 [0,169-0,173]	0,251 [0,203-0,300]
Сервис 3 вер. 2	Все исследования	170902	5818 (3,40%)	0,660 [0,654-0,667]	0,219 [0,217-0,230]	0,910 [0,908-0,911]	0,381 [0,369-0,394]	0,928 [0,927-0,930]	0,965 [0,965-0,967]	38,22%	0,184 [0,184-0,185]	0,204 [0,196-0,212]
	Исследования, описанные врачами-экспертами	7813	260 (3,33%)	0,775 [0,744-0,805]	0,248 [0,228-0,282]	0,878 [0,871-0,885]	0,635 [0,576-0,693]	0,887 [0,879-0,894]	0,966 [0,963-0,971]	26,76%	0,216 [0,214-0,217]	0,277 [0,246-0,311]
Сервис 3 вер. 3	Все исследования	34119	1000 (2,93%)	0,745 [0,728-0,761]	0,211 [0,188-0,209]	0,861 [0,857-0,864]	0,589 [0,559-0,620]	0,869 [0,865-0,872]	0,970 [0,968-0,972]	29,13%	0,158 [0,157-0,159]	0,219 [0,203-0,235]
	Исследования, описанные врачами-экспертами	1073	31 (2,89%)	0,903 [0,835-0,969]	0,328 [0,212-0,328]	0,883 [0,863-0,902]	0,871 [0,753-0,989]	0,883 [0,863-0,902]	0,971 [0,959-0,981]	11,43%	0,234 [0,228-0,236]	0,339 [0,261-0,410]

Примечание: значения параметров диагностической точности приводятся в таблице в формате: значение (95% доверительный интервал).

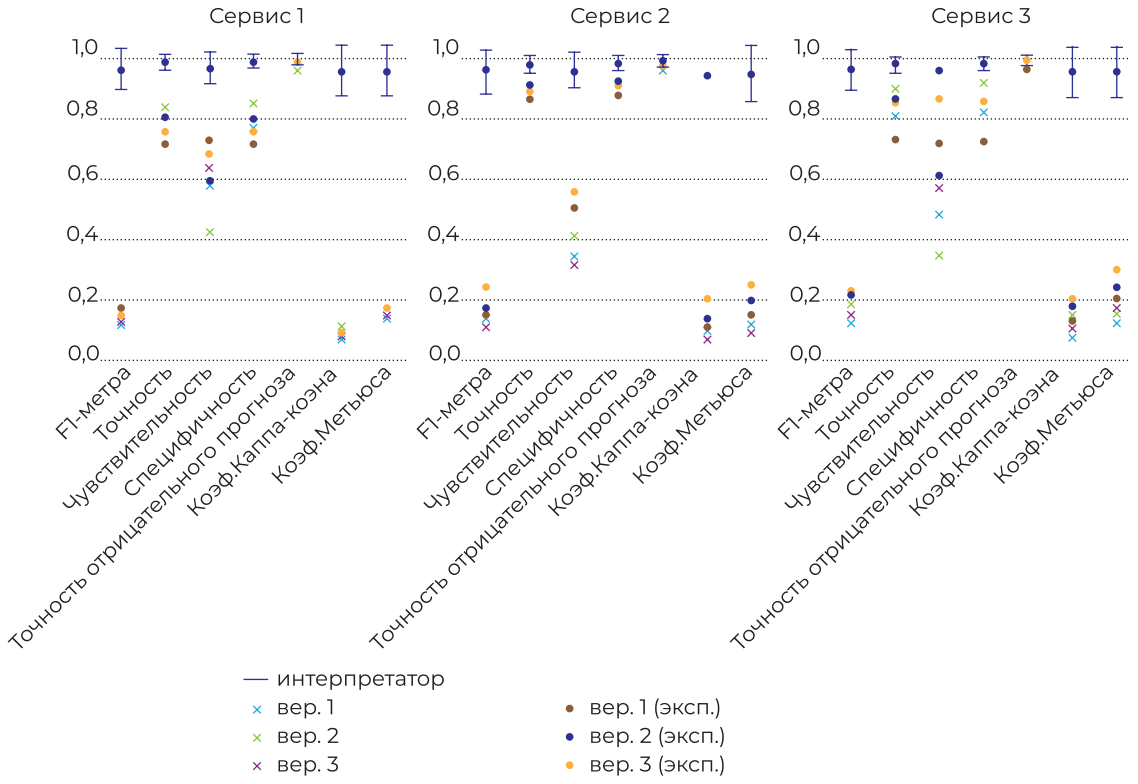

Рисунок 1 — Значения метрик ИИ-сервисов, их отношение к доверительному интервалу интерпретатора врачебных заключений. Версии (эксп.) — значения метрик ИИ-сервисов, полученные на выборках, описанных врачами экспертной группы.

В ошибочных заключениях ИИ-сервисов доля ложноотрицательных результатов (33,5±4,3%) значимо превышала долю ложноположительных (11,4±4,9%). Однако с учетом дисбаланса классов абсолютное количество ложноположительных значений было статистически значимо выше количества ложноотрицательных (11653±4847 против 1953±1259, $p<0,05$).

По сравнению с общей выборкой врачебных оценок отмечается повышение ряда метрик в экспертной группе, что связано с более полными описаниями и заключениями, и, в частности, описанием минимальных изменений на ФЛГ, которые также определяет ИИ-сервис.

На рис. 1 представлено соотношение полученных диагностических метрик и доверительных интервалов интерпретатора врачебных заключений. Значения всех метрик, за исключением точности отрицательного прогноза, находятся вне доверительных интервалов интерпретатора, что свидетельствует о достоверности полученных интерпретатором данных. Учитывая вышеописанную работу интерпретатора, также можно говорить и о достоверности полученной точности отрицательного прогноза.

На выборках, описанных врачами экспертной группы, произведена оценка дисбаланс-независимых метрик и доли ложноотрицательных результатов для различных порогов срабатывания с шагом 0,05 (результаты представлены на рис. 2). У большинства сервисов F1-мера и коэффициент Метьюса практически не изменялись при изменении порога срабатывания, за исключением сервиса 3 версии 3. Для него отмечалось небольшое плато в виде нулевого количества ложноотрицательных значений при максимальном значении порога срабатывания, равном 0,05. При этом же пороге доля истинно отрицательных значений составила 46,8% от всех исследований с нормой в обработанной сервисом выборке (рис. 3).

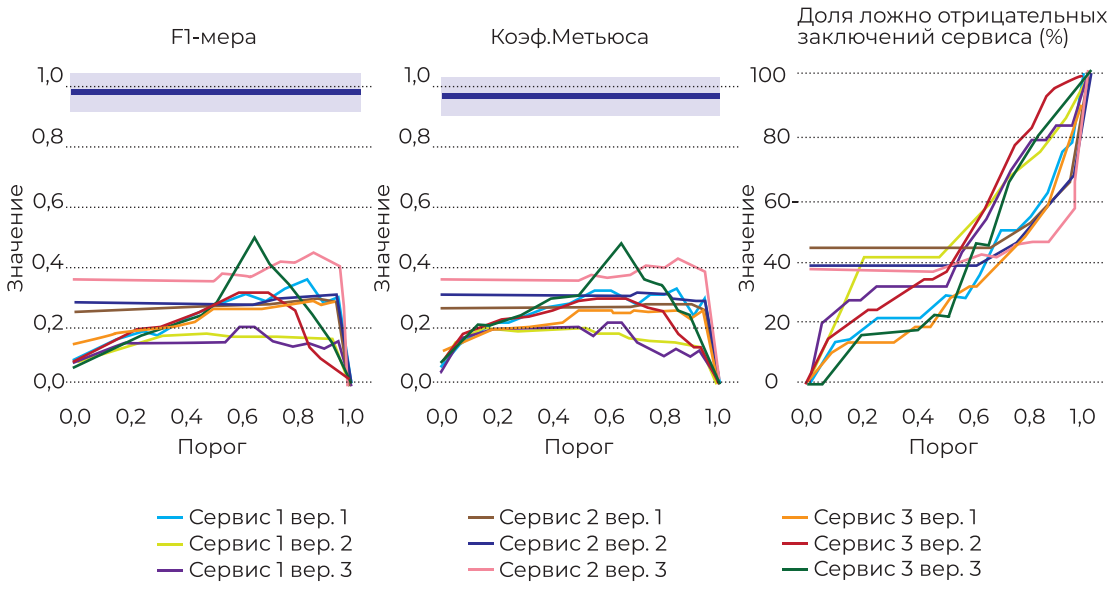

Рисунок 2 — Показатели F1-меры, коэффициента корреляции Метьюса, доли ложноотрицательных заключений в зависимости от порога срабатывания ИИ-сервисов и их отношение к доверительному интервалу интерпретатора (синий цвет).

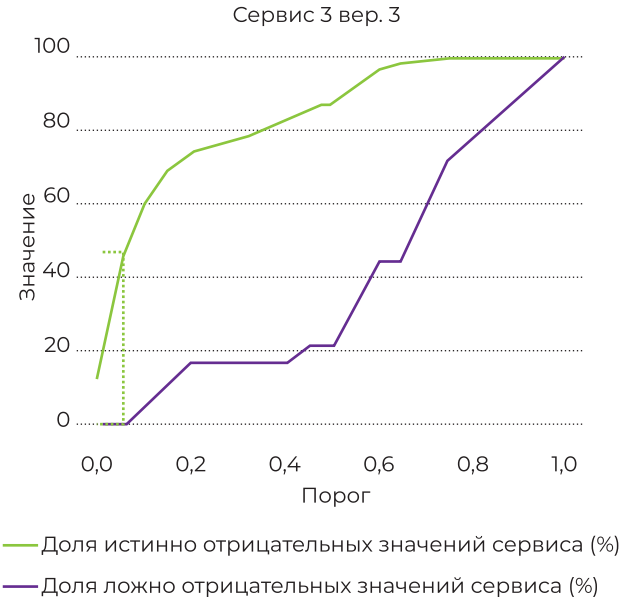

Рисунок 3 — Показатели ложно отрицательных значений и истинно отрицательных значений для сервиса 3 версии 3.

Оценка разницы в совершаемых сервисами ошибках (ложноотрицательных и ложноположительных заключениях) на общих данных с помощью теста Макнемара представлена в таблице 3. Выявлено, что во всех случаях, кроме одного, при сравнении сервиса 1 версии 1 и сервиса 2 версии 1, ИИ-сервисы совершают различные ошибки на одних и тех же данных.

ОБСУЖДЕНИЕ

В настоящем исследовании мы изучили работу трех ИИ-сервисов по направлению ФЛГ, каждый из которых был представлен тремя версиями. Доработка ПО на основе алгоритмов глубокого машинного обучения является важной задачей, цель которой — повышение удобства и качества работы ИИ-сервиса. В процессе доработки решений можно было наблюдать изменение метрик диагностической точности, однако они не всегда являлись оптимальными и сбалансированными.

Дисбаланс классов может оказывать сильное влияние на метрики оценки качества алгоритмов машинного обучения, такие как точность (Accuracy), специфичность и некоторые другие [20], делая акцент на исследованиях без патологии, приближая показатели к 1, что может в конечном итоге привести к ложным выводам. Выбор качественной метрики является серьезной проблемой при оценке алгоритмов классификации при дисбалансе классов в наборе данных, что приводит к разработке и внедрению новых видов оценки [20–23].

В анализируемых данных Эксперимента имеется выраженный дисбаланс по количеству ФЛГ исследований с патологией и без: доля исследований с патологическими изменениями составляет 3,4%. Это, в свою очередь, не позволяет нам корректно сравнить результаты настоящей работы с результатами, опубликованными в других работах. Только в работе Liz H. и соавторов делается акцент на дисбаланс в наборе данных, при этом количество исследований без патологии равно 47,4% [24].

Также дисбаланс по количеству ФЛГ с патологией и без усложняет возможности классификации, так как алгоритмы в первую очередь обучены на выявление патологических изменений, а при увеличении количества патологий, оцениваемых сервисами, увеличивается количество и изменчивость признаков, которые характерны для данных изменений [24]. Это приводит к высокому количеству ложных заключений, т.к. сервис будет искать патологические признаки там, где их нет. Так, при оценке ИИ-сервисов с помощью теста МакНемара (табл. 3) отмечается, что они совершают различные виды ошибок на одних и тех же данных вероятнее всего из-за различной интерпретации обрабатываемых данных.

Большое количество ложноотрицательных значений может стать существенным препятствием для внедрения данных ИИ-сервисов в медицинскую практику в качестве автономных медицинских изделий. Если ИИ-сервис будет давать околонулевое количество ложноотрицательных значений, то такой сервис можно внедрить в повседневную практику в качестве фильтра, отсекающего «нормальные» ФЛГ и оставляющего для интерпретации врача исследования,

Таблица 3 — Сравнение F1-меры и ошибок, которые совершают ИИ-сервисы на пересекающихся исследованиях

	Сервис 2 вер. 1	Сервис 2 вер. 3	Сервис 3 вер. 2	Сервис 3 вер. 3
Сервис 1 вер. 1	0,28 / 0,33 (450) p = 0,330	–	0,15 / 0,28 (883) p<0,010	–
Сервис 1 вер. 2	–	–	0,15 / 0,27 (958) p<0,010	–
Сервис 1 вер. 3	–	0,18 / 0,39 (1189) p<0,010	–	0,20 / 0,31 (469) p<0,010
Сервис 2 вер. 3	–	–	–	0,34 / 0,27 (556) p<0,010

Примечание: значения представлены в формате F1-мера алгоритма строки / F1 алгоритма столбца (количество общих исследований), p-значение для теста МакНемара.

где предположительно есть патология. Исследования с «нормой» в ложноположительных заключениях ИИ-сервисов будут отсекаться уже врачами. Чем больше доля истинно определенной сервисом нормы, тем меньше будет нагрузка на квалифицированный врачебный персонал при внедрении данного ИИ-сервиса.

Одной из возможных причин ложных заключений ИИ-сервисов является неоптимальный порог срабатывания ИИ-сервиса, но он сильно зависит от внутренней архитектуры ПО, и часть, в которой происходит определение наличия патологии по порогу срабатывания для данных сервисов, нам не известна. Поэтому стоит предполагать, что при определенном изменении порога срабатывания доля ложноотрицательных значений будет уменьшаться, как представлено на рис. 2, что также можно отметить в работах других исследователей [25, 26]. Для версии 3 сервиса 3, следует отметить, что при пороге срабатывания 0,05 правильно определяется 46,8% исследований с нормой, при этом ложноотрицательные оценки отсутствуют. Для остальных сервисов также можно подобрать такой порог срабатывания, при котором доля ложноотрицательных значений будет равна нулю, что позволит рассматривать возможность внедрения данных сервисов в рутинную практику и снизить нагрузку на врачей лучевой диагностики, сконцентрировав их навыки на описании исследований с патологией.

ЗАКЛЮЧЕНИЕ

В ходе Московского эксперимента по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений были продемонстрированы возможности ИИ-сервисов, основанных на алгоритмах глубокого машинного обучения, в качестве ассистентов врачей рентгенологов. Количество ложноотрицательных срабатываний для ИИ-сервисов при текущих настройках анализируемых продуктов, в том числе связанное с выраженным дисбалансом нормальных и патологических ФЛГ, является препятствием для автономного их внедрения. Были выявлены методы оптимизации ИИ-сервисов. Одним из них является изменение порога срабатывания сервиса, что позволяет в отдельных случаях добиться безошибочного исключения 46,8% исследований с нормой и снизить нагрузку на врачей-рентгенологов почти в два раза. Но данный метод оптимизации сильно органичен ввиду закрытости архитектур алгоритмов глубокого машинного обучения ИИ-сервисов и требует дополнительного изучения.

Тесное сотрудничество разработчиков ИИ-сервисов и конечных потребителей, врачей, а также дальнейшие исследования в области оптимизации, и, возможно, переориентации в методах обучения алгоритмов на более точный поиск нормальных ФЛГ исследований, позволит в дальнейшем автономно использовать ИИ-сервисы в медицинской практике.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Финансирование. Данная статья подготовлена авторским коллективом в рамках научно-исследовательской работы «Научные методологии устойчивого развития технологий искусственного интеллекта в медицинской диагностике» (№ ЕГИСУ: 123031500004-5) в соответствии с Приказом от 21.12.2022 г. №1196 «Об утверждении государственных заданий, финансовое обеспечение которых осуществляется за счет средств бюджета города Москвы государственным бюджетным (автономным) учреждениям подведомственным Департаменту здравоохранения города Москвы, на 2023 год и плановый период 2024 и 2025 годов» Департамента здравоохранения города Москвы.

ЛИТЕРАТУРА/REFERENCES

1. World Heath Organization. Mortality and global health estimates. Available at: https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates. Accessed Mar 28. 2023.
2. Sung H, Ferlay J, Siege RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021; 71(3): 209-249. doi: 10.3322/caac.21660.
3. Федеральная служба государственной статистики. Число умерших по основным классам причин смерти. Доступно по: https://rosstat.gov.ru/storage/mediabank/demo24-1_2021.xls. Ссылка активна на 28.03.2023. [Federal State Statistics Service. Chislo umershikh po osnovnym klassam

prichin smerti Available at: https://rosstat.gov.ru/storage/mediabank/demo24-1_2021.xls. Accessed Mar 28. 2023. (In Russ.)]

4. Состояние онкологической помощи населению России в 2021 году. / Под ред. Каприна А.Д., Старинского В.В., Шахзадовой А.О. — М.: МНИОИ им. П.А. Герцена, 2022. [Sostoyanie onkologicheskoi pomoshchi naseleniyu Rossii v 2021 godu. Kaprin AD, Starinsky VV, Shakhzadova AO, editors. M.: MNIOI im. P.A. Gertsena; 2022. (In Russ.)]
5. Синицын В.Е., Тюрин И.Е., Митьков В.В. Временные согласительные методические рекомендации Российского общества рентгенологов и радиологов (РОРР) и Российской ассоциации специалистов ультразвуковой диагностики в медицине (РАСУДМ) «Методы лучевой диагностики пневмонии при новой коронавирусной инфекции COVID-19» (версия 2) // Вестник рентгенологии и радиологии. — 2020. — Т.101. — №2. — С.72-89. [Sinitsyn VE, Tyurin IE, Mitkov VV. Consensus Guidelines of Russian Society of Radiology (RSR) and Russian Association of Specialists in Ultrasound Diagnostics in Medicine (RASUDM) «Role of Imaging (X-ray, CT and US) in Diagnosis of COVID-19 Pneumonia» (version 2). Journal of radiology and nuclear medicine. 2020; 101(2): 72-89. (In Russ.)] doi: 10.20862/0042-4676-2020-101-2-72-89.
6. Colman J, Zamfir G, Sheehan F, et al. Chest radiograph characteristics in COVID-19 infection and their association with survival. Eur J Radiol Open. 2021; 8: 100360. doi: 10.1016/j.ejro.2021.100360.
7. ACR. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection. Available at: https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection. Accessed Mar 28. 2023.
8. Wong HYF, Lam HYS, Fong AH, et al. Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19. Radiology. 2020; 296(2): E72-E78. doi: 10.1148/radiol.2020201160.
9. Морозов С.П., Проценко Д.Н., Сметанина С.В. и др. Лучевая диагностика коронавирусной болезни (COVID-19): организация, методология, интерпретация результатов: методические рекомендации. — М.: ГБУЗ «НПКЦ ДиТ ДЗМ», 2020. [Morozov SP, Protsenko DN, Smetanina SV, et al. Luchevaya diagnostika koronavirusnoi bolezni (COVID-19): organizatsiya, metodologiya, interpretatsiya rezul'tatov: metodicheskie rekomendatsii. M.: GBUZ «NPKC DiT DZM»; 2021. (In Russ.)]
10. Segal B, Rubin DM, Rubin G, Pantanowitz A. Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs. SN Comput Sci. 2021; 2(4): 321. doi: 10.1007/s42979-021-00720-7.
11. Rahman T, Chowdhury MEH, Khandakar A, et al. Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray. Applied Sciences. 2020; 10(9): 3233. doi: 10.3390/app10093233.
12. Gazda M, Plavka J, Gazda J, Drotár P. Self-Supervised Deep Convolutional Neural Network for Chest X-Ray Classification. IEEE Access. 2021; 9: 151972-151982. doi: 10.1109/ACCESS.2021.3125324.
13. Wu JT, Wong KCL, Gur Y, et al. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. JAMA Netw Open. 2020; 3(10): e2022779. doi: 10.1001/jamanetworkopen.2020.22779.
14. 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 (Basel). 2023; 11(12): 1684. doi: 10.3390/healthcare11121684.
15. Huang XM, Yang BF, Zheng WL, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. BMC Health Serv Res. 2022; 22(1): 260. doi: 10.1186/s12913-022-07655-6.
16. Романовсков Ю.Ф., Коновалов В.К., Колмогоров В.Г. Заочная консультация рентгенологических исследований в Алтайском крае // Digital Diagnostics. — 2021. — Т.2. — №1S. — С.26-27. [Romanovskov YF, Konovalov VK, Kolmogorov VG. Correspondence consultation of X-ray examinations in the Altai Territory. Digital Diagnostics. 2021; 2(1S): 26-27. (In Russ.)] doi: 10.17816/DD20211s26.

17. Приказ Департамента здравоохранения города Москвы от 19.02.2020 №142 «Об утверждении Порядка и условий проведения эксперимента на использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы». Доступно по: https://mosmed.ai/documents/9/Приказ_департамента_здравоохранения_города_Москвы_от_19.02.202015142.pdf_lFf9_slECRah.pdf. Ссылка активна на 28.03.2023. [Order of the Moscow Department of Health №142 of 19 February 2020. «Ob utverzhdenii Poryadka i usloviy provedeniya eksperimenta na ispol'zovaniyu innovatsionnykh tekhnologiy v oblasti komp'yuternogo zreniya dlya analiza meditsinskikh izobrazheniy i dal'neyshego primeneniya v sisteme zdravookhraneniya goroda Moskvy». Available at: https://mosmed.ai/documents/9/Приказ_департамента_здравоохранения_города_Москвы_от_19.02.202015142.pdf_lFf9_slECRah.pdf. Accessed Mar 28. 2023. (In Russ.)]
18. Васильев Ю.А., Владзимирский А.В., Арзамасов К.М., и др. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента: Монография. 2-е издание, переработанное и дополненное. — М.: Издательские решения, 2023. [Vasilev YA, Vladzymyrskyy AV, Arzamasov KM, et al. Computer vision in radiology: the first stage of the Moscow experiment: Monograph. 2nd edition. M.: Izdatelskie resheniya; 2023. (In Russ.)]
19. Свидетельство РФ о государственной регистрации программы для ЭВМ №2022617324. Морозов С.П., Андрейченко А.Е., Четвериков С.Ф., и др.Веб-инструмент для выполнения ROC анализа результатов диагностических тестов: №2022616046/19.04.2022. [Certificate RUS of state registration of a computer program №022617324. Morozov SP, Andreichenko AE, Chetverikov SF, et al. Web-based tool for performing ROC analysis of diagnostic test results: №2022616046/04/19/2022. (In Russ.)]
20. Luque A, Carrasco A, Martín F, Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019; 91: 216-231. doi: 10.1016/j.patcog.2019.02.023.
21. Mortaz E. Imbalance accuracy metric for model selection in multi-class imbalance classification problems. Knowledge-Based Systems. 2020; 210: 106490. doi: 10.1016/j.knosys.2020.106490.
22. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009; 45: 427-437. doi: 10.1016/j.ipm.2009.03.002.
23. Chicco D, Warrens MJ, Jurman G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen's Kappa and Brier Score in Binary Classification Assessment. IEEE Access. 2021; 9: 78368-78381. doi: 10.1109/ACCESS.2021.3084050.
24. Liz H, Huertas-Tato J, Sánchez-Montañés M, et al. Deep learning for understanding multilabel imbalanced Chest X-ray datasets. Future Generation Computer Systems. 2023; 144: 291-306. doi: 10.1016/j.future.2023.03.005.
25. Minaee S, Kafieh R, Sonka M, et al. Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. Med Image Anal. 2020; 65: 101794. doi: 10.1016/j.media.2020.101794.
26. Bharodiya AK, Atul MG. An Improved Segmentation Algorithm For X-Ray Images Based On Adaptive Thresholding Classification. International Journal of Scientific & Technology Research. 2019; 8(9): 1617-1623.