БОБРОВСКАЯ Т.М.,
ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», Москва, Россия,
e-mail: BobrovskayaTM@zdrav.mos.ru

КИРПИЧЕВ Ю.С.,
ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», Москва, Россия,
e-mail: KirpichevYS@zdrav.mos.ru

САВКИНА Е.Ф.,
ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», Москва, Россия,
e-mail: SavkinaEF@zdrav.mos.ru, ORCID: 0000-0001-9165-0719

ЧЕТВЕРИКОВ С.Ф.,
к.т.н., ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», Москва, Россия,
e-mail: ChetverikovSF@zdrav.mos.ru

АРЗАМАСОВ К.М.,
к.м.н., ГБУЗ «НПКЦ диагностики и телемедицины ДЗМ», Москва, Россия; РТУ МИРЭА, Москва, Россия,
e-mail: ArzamasovKM@zdrav.mos.ru

РАЗРАБОТКА И ВАЛИДАЦИЯ ИНСТРУМЕНТА СТАТИСТИЧЕСКОГО СРАВНЕНИЯ ХАРАКТЕРИСТИЧЕСКИХ КРИВЫХ НА ПРИМЕРЕ РАБОТЫ АЛГОРИТМОВ НА ОСНОВЕ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

DOI: 10.25881/18110193_2023_3_4

Аннотация. *Актуальность: Благодаря Национальной стратегии развития искусственного интеллекта в Российской Федерации активно разрабатываются и внедряются новые технологии на основе искусственного интеллекта, что приводит к появлению большого количества различных практических и научных задач, которые в свою очередь требуют удобных инструментов для их решения. Одним из них является инструмент, предназначенный для ROC-анализа, который был разработан и успешно применялся в рамках проекта «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы». Однако для решения более широкого спектра задач, связанных с аналитикой работы технологий на основе искусственного интеллекта, возникла острая необходимость в разработке модуля сравнения ROC-кривых.*

Цель: реализовать модуль инструмента ROC-анализа по сравнению площади под характеристической кривой с помощью статистических критериев и расчётом p-значения и апробировать его на реальных данных.

Материалы и методы: инструмент реализован на языке Python 3.9. 95% доверительный интервал для ROC-кривых рассчитывался с помощью метода случайных выборок с возвратом (бутстреппинг) и метода Де-Лонг (DeLong). Сравнение площадей под ROC-кривыми осуществлялось с помощью перестановочного теста.

Апробация инструмента осуществлялась на результатах работы 6 алгоритмов на основе технологий искусственного интеллекта на 2 наборах данных. Проводилось попарное сравнение площади под ROC-кривой и полученные результаты сравнивали с результатами анализа тех же данных методом ДеЛонг функции roc.test языка R 3.6.1.

Результаты: p-значения, полученные с помощью перестановочного теста, оказались в большинстве случаев сопоставимы с результатами roc.test, однако в 4 из 30 случаев p-значения принципиально отличались, что приводило к изменениям интерпретации теста.

Обсуждение: различия в результатах, рассчитанных двумя способами, вероятно, обусловлены особенностями используемых методов: ДеЛонг является более консервативным. Также из-за использования метода псевдорандомизации в перестановочном тесте возможна вариативность результатов, что может привести к неопределенности. Кроме того, разработанный инструмент сравнивает наборы данных с одинаковым количеством элементов, что является ограничением его использования, однако возможна дальнейшая его разработка с целью преодоления данного ограничения.

Заключение: был успешно реализован и апробирован модуль сравнения ROC-кривых с помощью статистических критериев с расчётом p-значения.

***Ключевые слова**: искусственный интеллект, ROC-анализ, статистический анализ.*

***Для цитирования**: Бобровская Т.М., Кирпичев Ю.С., Савкина Е.Ф., Четвериков С.Ф., Арзамасов К.М. Разработка и валидация инструмента статистического сравнения характеристических кривых на примере работы алгоритмов на основе технологий искусственного интеллекта. Врач и информационные технологии. 2023; 3: 4-15. doi: 10.25881/18110193_2023_3_4.*

BOBROVSKAYA T.M.,
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: BobrovskayaTM@zdrav.mos.ru

KIRPICHEV Y.S.,
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: KirpichevYS@zdrav.mos.ru

SAVKINA E.F.,
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: SavkinaEF@zdrav.mos.ru

CHETVERIKOV S.F.,
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: ChetverikovSF@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; Russian Technological University, Moscow, Russia, e-mail: ArzamasovKM@zdrav.mos.ru

DEVELOPMENT AND VALIDATION OF A TOOL FOR STATISTICAL COMPARISON OF ROC-CURVES USING THE EXAMPLE OF ALGORITHMS BASED ON ARTIFICIAL INTELLIGENCE TECHNOLOGIES

DOI: 10.25881/18110193_2023_3_4

Abstract. *Background: Due to the National Strategy for the Development of Artificial Intelligence, large-scale digitalization of healthcare is taking place in the Russian Federation, which leads to huge number of various practical and scientific tasks emergence of, which in turn require convenient tools to solve them. ROC analysis tool is one of them, which was developed and successfully applied within the framework of the project «Experiment on the use of innovative technologies in the field of computer vision for the analysis of medical images and further application in the healthcare system of the city of Moscow». However, there is an urgent need for the development of a module comparing ROC-curves in order to solve a wider range of problems related to analytics of the operation of technologies based on artificial intelligence.*

Aim: to implement the ROC analysis tool module for comparing the area under the curve using statistical methods and calculating the p-value, and to test it on real data.

Materials and methods: the tool is implemented in Python 3.9. The 95% confidence interval for ROC curves was calculated using the bootstrapping and the DeLong method. Areas under the ROC curves comparison was carried out using a permutation test.

The testing of the tool was carried out on the 6 algorithms work results on 2 data sets. Area under the ROC curve pairwise comparison was carried out and the results were compared with the same data results analysis, calculated by the DeLong method (roc.test function, R language 3.6.1).

Results: the p-values obtained using the permutation test were in most cases comparable to the roc.test results, however, in 4 out of 30 cases, the p-values differed significantly, which led to changes in the test interpretation.

Discussion: the differences in the results calculated by two separate methods, in our opinion, are due to the peculiarities of the methods used: DeLong method is more conservative. Also, due to the use of the pseudorandomization method in the permutation test, variability of results is possible, which can lead to uncertainty. In addition, the developed tool compares data of the same length, which is a limitation of its use, but its further development is possible for data of different lengths.

Conclusion: the module for comparing ROC curves was successfully implemented and tested using statistical criteria with the calculation of the p-value.

Keywords: *artificial intelligence, ROC analysis, statistical analysis.*

For citation: *Bobrovskaya T.M., Kirpichev Y.S., Savkina E.F., Chetverikov S.F., Arzamasov K.M. Development and validation of a tool for statistical comparison of ROC-curves using the example of algorithms based on artificial intelligence technologies. Medical doctor and information technology. 2023; 3: 4-15. doi: 10.25881/18110193_2023_3_4.*

ОБОСНОВАНИЕ

Разработка и внедрение программного обеспечения на основе технологий искусственного интеллекта (ТИИ) в РФ — одна из главных задач Национальной стратегии развития искусственного интеллекта в Российской Федерации [1]. Масштабная цифровизация всех сфер нашей жизни, в том числе и здравоохранения, приводит к увеличению количества данных и развитию инструментов по работе с ними [2]. Использование ТИИ в медицинской диагностике позволяет оптимизировать работу и снизить нагрузку на врачей. Например, во время пандемии COVID-19, использование алгоритмов на основе ТИИ позволило снизить время обработки протокола заключения компьютерной томографии врачом-рентгенологом [3]. Также исследования показали возможность использования таких алгоритмов в качестве второго чтения, например, при скрининге рака молочной железы [4]. Одним из самых масштабных проектов по внедрению ТИИ в здравоохранение является Московский «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы» (далее Эксперимент) — наглядный пример успешной реализации применения ТИИ в медицинской диагностике [5].

Однако алгоритмы на основе ТИИ далеки от совершенства и требуют постоянного контроля качества, доработки и проверки. Прежде чем допустить такие алгоритмы к применению в медицинских организациях, необходимо провести тщательный анализ работы самого алгоритма, калибровку и оценку качества. Также необходим своевременный мониторинг качества в процессе работы алгоритма, который позволяет оперативно принимать решение о необходимости его перенастройки или вывода его из практической деятельности [6]. Одним из наиболее важных параметров является оценка диагностической точности. Золотым стандартом в оценке диагностических тестов является анализ классификации с помощью ROC-кривых (Receiver Operating Characteristic, характеристическая кривая).

Для реализации задач Эксперимента, а также в других научных исследованиях, требовался удобный и простой инструмент для оценки метрик диагностической точности: чувствительности, специфичности и площади под характеристической кривой ROC AUC (Receiver Operating Characteristic Area Under Curve). Для этого был реализован удобный, простой в использовании инструмент с открытым доступом [7, 8]. Однако недостаточно просто измерить абсолютные величины метрик диагностической точности. Ряд задач требует провести сравнение этих метрик между собой в соответствии с принципами доказательной медицины. Например, в задачах сравнения работы алгоритмов на основе ТИИ между собой или с работой врачей-рентгенологов. Также влияние различных факторов внешней среды может существенно изменить анализируемые данные, что в дальнейшем может привести к ухудшению метрик диагностической точности алгоритмов, а это, в свою очередь, потребует их дообучения, калибровки и нового тестирования для допуска к работе. Так, появление новой коронавирусной инфекции COVID-19 в 2020 году затруднило диагностику злокачественных новообразований легких [9] и, как следствие, мы столкнулись с задачей выявления статистически значимых различий в работе алгоритмов на основе ТИИ, чтобы ответить на вопрос: «Можем ли мы использовать данный алгоритм в новых условиях или необходима его доработка?».

ЦЕЛЬ

Реализовать модуль инструмента ROC-анализа по сравнению площади под характеристической кривой с помощью статистических критериев и расчётом p-значения и апробировать его на реальных данных.

МАТЕРИАЛЫ И МЕТОДЫ

Апробация нового модуля проводилась на данных, полученных в ходе зарегистрированного ранее исследования (NCT04489992), одобренного локальным этическим комитетом, «Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы».

Инструмент статистического анализа и построения характеристических кривых.

Интерфейс и серверная часть была написана с помощью библиотеки Plotly [10]. Для построения характеристических кривых с 95% доверительными интервалами с помощью метода случайных выборок с возвратом (далее «бутстреппинг» от англ. «bootstrapping») была использована библиотека ROC-utils [11]. Для вычислений доверительного интервала методом ДеЛонг была использована библиотека [12] .

Сравнение площадей под характеристической кривой осуществлялось с помощью перестановочного теста [13]. Проверялась нулевая гипотеза об отсутствии статистически значимых различий между двумя ROC AUC.

Валидация инструмента.

С целью апробации работы нового инструмента мы провели ретроспективное обсервационное когортное исследование. Инструмент валидировался на результатах работы 6 алгоритмов на основе ТИИ (далее — ИИ1, ИИ2, ИИ3, ИИ4, ИИ5, ИИ6), полученных в ходе Эксперимента. Критерием для включения этих алгоритмов было соответствие заявленным возможностям определения солидных легочных узлов размером более 6 мм. Критерий исключения — проанализировано менее 90% исследований в каждом наборе данных (НД).

Алгоритмы обрабатывали оба НД в 2021 году одномоментно без каких-либо доработок и изменений версий. Разработчики алгоритмов заявляли, что изменения, характерные для COVID-19, не являются ограничением применения их решений для обнаружения рака легких. Все исследования в каждом НД были обработаны всеми алгоритмами. Далее сравнивались результаты их работы между собой на каждом НД в отдельности с помощью готовой функции roc.test языка R (method = «delong») и с помощью нашего инструмента. Проверялась нулевая гипотеза об отсутствии статистически значимых различий между алгоритмами на основе ТИИ.

Наборы данных.

Для исследования использовались следующие НД.

НД1 представляет собой 82 исследования компьютерной томографии грудной полости с наличием и отсутствием признаков рака легких (бинарная разметка, соотношение классов есть признаки патологии/нет признаков патологии 51/31), проведенных в период с 2015 по 2016, т.е. до начала пандемии COVID-19.

НД2 — 91 исследование компьютерной томографии грудной полости с наличием и отсутствием признаков рака легких (бинарная разметка, соотношение классов есть признаки патологии/нет признаков патологии 47/44), проведенных в 2020 году, в разгар пандемии. Соотношение признаков рака легкого и COVID-19: из 44 исследований без признаков рака легкого 22 исследования с признаками COVID-19, из 47 исследований с признаками рака легкого 29 исследований с признаками COVID-19.

Верификация НД проводилась путем экспертного пересмотра (исследования независимо анализировали 2 врача-рентгенолога, в случае разногласий подключался эксперт с опытом работы более 5 лет по данному направлению), а также с помощью патоморфологического исследования.

Критерии отнесения к классам с признаками/без признаков рака легких при экспертном пересмотре:

- с патологией: хотя бы один солидный или субсолидный узел объемом более 100 куб. мм; если волюметрию выполнить невозможно, использовался наименьший линейный размер узла 6 мм.
- без патологии: нет ни одного узла, подпадающего под указанные условия.

Верификация патологии COVID-19 для НД2 проводилась с помощью экспертного пересмотра, а также с помощью результатов лабораторного тестирования (ПЦР).

Критерии отнесения к классам с признаками/без признаков COVID-19 при экспертном пересмотре:

- с патологией:
  1. Инфильтрация легочной паренхимы по типу матовых стекол с обеих сторон, преимущественно периферической локализации, с или без инфильтрации легочной паренхимы по типу консолидации с положительным признаком воздушной бронхограммы;

2. Инфильтрация легочной паренхимы по типу булыжной мостовой (утолщение междолькового интерстиция на фоне матового стекла) с обеих сторон, преимущественно периферической локализации, с или без инфильтрации легочной паренхимы по типу консолидации с положительным признаком воздушной бронхограммы.

– без патологии: отсутствие вышеперечисленных признаков.

Критериями исключения для обоих НД являлись хирургические вмешательства, артефакты, связанные с пациентом (наложение руки на грудную клетку, ориентация тела, кашель, движения) или некачественное сканирование (артефакты, технические дефекты).

РЕЗУЛЬТАТЫ

Инструмент сравнения характеристических кривых.

Инструмент попарного сравнения результатов работы алгоритмов на основе ТИИ представляет собой специально созданный внутренний веб-инструмент (веб-сайт), на который необходимо загрузить файл с разметкой (уникальный идентификатор и соответствующее ему значение 1 или 0) и результатами, полученными от алгоритма.

После загрузки файлов отображаются их названия и результат сравнения ROC AUC методом перестановочного теста (разница между значениями ROC AUC и p-значение для теста), а также непосредственно сами характеристические кривые. Кроме этого, инструмент оснащен функцией настройки параметров графика (цвет, оптимальное пороговое значение, отображение экспериментальных точек и доверительного интервала). Для характеристических кривых также рассчитывается и отображается площадь под кривой с доверительными интервалами, рассчитанными разными способами. Инструмент находится в открытом доступе по ссылке https://roc-analysis.mosmed.ai.

Результаты сравнения.

Для апробации созданного инструмента мы поставили задачу сравнить работу алгоритмов на основе ТИИ между собой в период до пандемии COVID-19 (НД1) и во время пандемии (НД2). Результаты попарного сравнения ROC AUC алгоритмов на основе ТИИ для выявления новообразований легких представлены в таблице 1. У 8 пар алгоритмов были обнаружены статистически значимые различия до пандемии и у 4 пар алгоритмов — во время пандемии.

Для сравнения работы инструментов расчета статистической значимости мы использовали p-значения, определенные с помощью функции roc.test языка R и перестановочного теста (Таблицы 2 и 3).

Таблица 1 — Результаты попарного сравнения ROC AUC алгоритмов на основе технологий искусственного интеллекта (ИИ1 — ИИ6). На главной диагонали — значения ROC AUC каждого алгоритма (желтый цвет). В ячейках указаны значения разности ROC AUC для каждого алгоритма (ИИi-ИИj). Зеленым цветом обозначены статистически значимые различия, определенные с помощью функции roc.test языка R. Таблица симметрична относительно главной диагонали (со сменой знака). А — результаты сравнения на НД1. Б — результаты сравнения на НД2

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1	0,761	-0,058	-0,166	-0,176	-0,181	-0,220
ИИ2	0,058	0,820	-0,107	-0,117	-0,122	-0,161
ИИ3	0,166	0,107	0,927	-0,010	-0,015	-0,054
ИИ4	0,176	0,117	0,010	0,937	-0,005	-0,044
ИИ5	0,181	0,122	0,015	0,005	0,942	-0,039
ИИ6	0,220	0,161	0,054	0,044	0,039	0,981

А

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1	0,897	0,133	0,122	0,086	0,093	0,013
ИИ2	-0,133	0,764	-0,011	-0,047	-0,040	-0,120
ИИ3	-0,122	0,011	0,775	-0,036	-0,029	-0,109
ИИ4	-0,086	0,047	0,036	0,811	0,007	-0,073
ИИ5	-0,093	0,040	0,029	-0,007	0,804	-0,080
ИИ6	-0,013	0,120	0,109	0,073	0,080	0,884

Б

Таблица 2 — p-значения для сравнения ROC AUC алгоритмов на основе ТИИ, рассчитанные с помощью функции roc.test языка R (таблица симметрична относительно главной диагонали). Красным цветом обозначены значения p, отличия в которых (по сравнению с перестановочным тестом) влияют на интерпретацию результата. А — результаты сравнения на НД1. Б — результаты сравнения на НД2

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1		0,409	0,014	0,005	0,001	0,000
ИИ2	0,409		0,038	0,024	0,017	0,001
ИИ3	0,014	0,038		0,744	0,718	0,080
ИИ4	0,005	0,024	0,744		0,888	0,054
ИИ5	0,001	0,017	0,718	0,888		0,204
ИИ6	0,000	0,001	0,080	0,054	0,204	

А

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1		0,007	0,021	0,107	0,054	0,762
ИИ2	0,007		0,844	0,396	0,291	0,025
ИИ3	0,021	0,844		0,510	0,605	0,008
ИИ4	0,107	0,396	0,510		0,898	0,139
ИИ5	0,054	0,291	0,605	0,898		0,145
ИИ6	0,762	0,025	0,008	0,139	0,145	

Б

Таблица 3 — p-значения для сравнения ROC AUC алгоритмов на основе технологий искусственного интеллекта, рассчитанные с помощью функции перестановочного теста (таблица симметрична относительно главной диагонали). Красным цветом обозначены значения p, отличия в которых (по сравнению с roc.test) влияют на интерпретацию результата. А — результаты сравнения на НД1. Б — результаты сравнения на НД2

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1		0,483	0,002	0,029	0,001	0,000
ИИ2	0,483		0,053	0,252	0,026	0,003
ИИ3	0,002	0,053		0,824	0,738	0,171
ИИ4	0,029	0,252	0,824		0,938	0,372
ИИ5	0,001	0,026	0,738	0,938		0,165
ИИ6	0,000	0,003	0,171	0,372	0,165	

А

j / i	ИИ1	ИИ2	ИИ3	ИИ4	ИИ5	ИИ6
ИИ1		0,008	0,070	0,176	0,031	0,792
ИИ2	0,008		0,834	0,586	0,415	0,023
ИИ3	0,070	0,834		0,568	0,624	0,038
ИИ4	0,176	0,586	0,568		0,924	0,358
ИИ5	0,031	0,415	0,624	0,924		0,151
ИИ6	0,792	0,023	0,038	0,358	0,151	

Б

ОБСУЖДЕНИЕ

Одной из главных задач при разработке инструмента была задача выбора способа статистического сравнения площади под характеристической кривой. На сегодняшний день существует несколько способов статистического анализа метрик диагностической точности [14], а также различные инструменты их реализации, начиная от использования языков программирования и заканчивая готовыми инструментами, не требующими знаний специальных языков [15–17]. Зачастую эти инструменты находятся в закрытом доступе, неудобны в использовании и сложны для понимания, кроме того, в случае самостоятельной реализации кода есть возможность его гибкой настройки под разные задачи. Поэтому мы поставили задачу реализовать удобный, доступный и простой в использовании инструмент. Ранее нами был реализован инструмент

построения ROC-кривых с расчётом доверительных интервалов для ROC AUC по методу ДеЛонга и с помощью бутстреппинга [8], и он успешно использовался в Эксперименте [18]. Однако в рамках задач по сравнению работы алгоритмов на основе ТИИ возникла необходимость в инструменте сравнения, который не только будет определять статистически обоснованные метрики, но и рассчитывать p-значение для сравниваемых метрик. Мы остановили свой выбор на перестановочном тесте [13]. Данный способ наиболее просто реализуется на языке Python, что было одним из важных условий при выборе языка программирования, т.к. наш инструмент изначально разрабатывался на нём. Еще одним аргументом в пользу метода перестановок является то, что он учитывает проблему множественных сравнений [19], что также играет важную роль, т.к. зачастую возникает необходимость сравнения большого количества моделей одновременно.

Для того чтобы оценить работу нашего инструмента, мы сравнивали полученные результаты с данными, рассчитанными с помощью функции roc.test языка R (Таблица 1,2), согласно которым у 8 пар алгоритмов на основе ТИИ имеются статистически значимые различия (p-значение меньше 0,05) до пандемии и у 4х пар алгоритмов — во время пандемии. Однако при расчете с помощью перестановочного теста мы получили несколько принципиальных расхождений, повлиявших на результаты интерпретации теста. Для НД1 p-значения пар ИИ2-ИИ3 и ИИ2-ИИ4 оказались выше 0,05, следовательно, мы не можем сделать вывод о наличии статистически значимой разницы между ними. Для НД2 p-значения пары ИИ1-ИИ3 оказались выше 0,05, а для ИИ1-ИИ5 — ниже, что также меняет интерпретацию результата на отсутствие статистически значимой разницы и ее наличие соответственно.

В целом, при анализе полученных с помощью перестановочного теста p-значений можно отметить их завышение по сравнению с roc.test, но в редких случаях значения совпадают или занижены. По мнению ряда авторов [20] метод ДеЛонга (использовался в функции roc.test) в ряде случаев является чрезмерно консервативным и часто отвергает нулевую гипотезу об отсутствии статистически значимых различий в пользу альтернативной. Кроме того, результат метода ДеЛонга может зависеть от объема и баланса НД, на котором проводится исследование[20], что, вероятнее всего, и является причиной разнородного поведения p-значения при сравнении различных методов.

Стоит отметить тот факт, что перестановочный тест относится к категории симуляционных, поэтому он не обеспечивает полную воспроизводимость результата из-за псевдорандомизации (как и метод бутстрепинга, реализованный, например, в функции roc.test): выборка будет формироваться каждый раз случайным образом, поэтому p-значение будет колебаться в определенных пределах. Это может привести к ситуации неопределенности, когда p-значения будут больше или меньше 0,05 при повторении теста.

Таким образом, нам удалось реализовать модуль инструмента анализа характеристических кривых, позволяющий сравнить между собой площади под характеристическими кривыми, построенными на данных одной длины (количество исследований, которые анализировали алгоритмы на основе ТИИ). Возвращаясь к задаче, которая послужила нам примером для валидации инструмента, мы можем сказать, какие алгоритмы имеют статистически значимые различия. Однако нам бы хотелось найти ответ и на такой вопрос: изменились ли метрики качества работы алгоритмов на основе ТИИ во время пандемии COVID-19? И с помощью функции roc.test мы можем получить ответ, но в случае реализованного нами перестановочного теста мы не можем сравнивать НД с разным количеством элементов.

Разработанный нами инструмент может быть полезен для проведения клинических испытаний систем искусственного интеллекта [21], а также при оценке зрелости решений на основе ТИИ [22]. В дальнейшем мы планируем усовершенствовать наш инструмент и решить данную проблему, например, с помощью достраивания выборки по методу Гиббса [23], а также реализовать новые модули для решения различных задач анализа данных.

ЗАКЛЮЧЕНИЕ

Широкое использование алгоритмов на основе ТИИ приводит к появлению различных задач как для научных исследований, так и для контроля качества результатов их работы, что в свою очередь требует создания удобных инструментов анализа данных. Один из таких инструментов, предназначенный для построения и сравнения характеристических кривых, был реализован и успешно апробирован в данной работе. Однако эта реализация не работает для НД с разным количеством элементов, поэтому в дальнейшем планируется ее доработка.

Источник финансирования. Данная статья подготовлена авторским коллективом в рамках научно-исследовательской работы «Разработка платформы повышения качества ИИ-Сервисов для медицинской диагностики».

Конфликт интересов. Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

Благодарности. Авторы выражают благодарность Никитину Н.Ю. за консультирование по вопросам статистического анализа.

ЛИТЕРАТУРА/REFERENCES

1. Указ Президента Российской Федерации от 10.10.2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации». 2019. Доступ по ссылке: http://www.kremlin.ru/acts/bank/44731. Ссылка активна на 14.02.2023. [Ukaz Prezidenta Rossijskoj Federacii ot 10.10.2019 g. «O razvitii iskusstvennogo intellekta v Rossijskoj Federacii» № 490. Available at: http://www.kremlin.ru/acts/bank/44731/page/1. Accessed 14.02.2023. (In Russ.)]
2. Гусев А.В., Владзимирский А.В., Шарова Д.Е. и др. Развитие исследований и разработок в сфере технологий искусственного интеллекта для здравоохранения в Российской Федерации: итоги 2021 года // Digital Diagnostics. — 2022. — Т.3. — №3. — С.178-194. [Gusev AV, Vladzymyrskyy AV, Sharova DE, et al. Evolution of research and development in the field of artificial intelligence technologies for healthcare in the Russian Federation: results of 2021. Digital Diagnostics. 2022; 3(3): 178-194. (In Russ.)]. doi: 10.17816/DD107367.
3. Морозов С.П., Гаврилов А.В., Архипов И.В. и др. Влияние технологий искусственного интеллекта на длительность описаний результатов компьютерной томографии пациентов с COVID-19 в стационарном звене здравоохранения // Профилактическая медицина. 2022;25(1):14–20. [Morozov SP, Gavrilov AV, Arkhipov IV, et al. Effect of artificial intelligence technologies on the CT scan interpreting time in COVID-19 patients in inpatient setting. Profilakticheskaya Meditsina. 2022; 25(1): 14-20. (In Russ.)] doi: 10.17116/PROFMED20222501114.
4. Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, et al. AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. Radiology. 2021; 300(1): 57-65. doi: 10.1148/RADIOL.2021203555.
5. Морозов С.П., Владзимирский А.В., Ледихова Н.В. и др. Московский эксперимент по применению компьютерного зрения в лучевой диагностике: вовлеченность врачей-рентгенологов // Врач и информационные технологии. 2020. — №4. — С.14-23. [Morozov SP, Vladzymyrskyy AV, Ledikhova NV, et al. Moscow experiment on computer vision in radiology: involvement and participation of radiologists. Vrach i informacionnye tekhnologii. 2020; 4: 14-23. (In Russ.)]
6. Andreychenko AE, Logunova TA, Gombolevskiy VA, et al. A methodology for selection and quality control of the radiological computer vision deployment at the megalopolis scale. medRxiv. 2022: 2022.02.12.22270663. doi: 10.1101/2022.02.12.22270663.
7. 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. С.П. Морозов, А.Е. Андрей-

ченко, С.Ф. Четвериков и др. [Database registration certificate №2022617324 Web-instrument dlya vypolneniya ROC analiza rezul'tatov diagnosticheskih testov: № 2022616046: Appl. 05.04.2022, publ. 19.04.2022. Morozov SP, Andreychenko AE, Chetverikov SF, et al. (In Russ.)]

8. ROC Analysis. Доступно по: https://roc-analysis.mosmed.ai/ Ссылка активна на 12.08.2023. [ROC Analysis. Available at: https://roc-analysis.mosmed.ai/ Accessed 12.08.2023. (In Russ.)]
9. Goncalves S, Fong PC, Blokhina M. Artificial intelligence for early diagnosis of lung cancer through incidental nodule detection in low- and middle-income countries-acceleration during the COVID-19 pandemic but here to stay. Am J Cancer Res. 2022; 12(1): 1.
10. Dash Documentation & User Guide Plotly. Available at: https://dash.plotly.com/docs. Accessed 08.08.2023.
11. roc-utils. Available at: https://github.com/hirsch-lab/roc-utils. Accessed 21.08.2022.
12. Sun X, Xu W. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Process Lett. 2014; 21(11): 1389-1393. doi: 10.1109/LSP.2014.2337313.
13. Pauly M, Asendorf T, Konietschke F. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. Biom J. 2016; 58(6): 1319-1337. doi: 10.1002/BIMJ.201500105.
14. Metz CE. ROC analysis in medical imaging: a tutorial review of the literature. Radiol Phys Technol. 2008; 1(1): 2-12. doi: 10.1007/S12194-007-0002-1/FIGURES/2.
15. Statistical Software. Sample Size Software. NCSS. Available at: https://www.ncss.com/ Accessed 08.02.2023.
16. Goksuluk D, Korkmaz S, Zararsiz G, Karaagaoglu AE. EasyROC: An interactive web-tool for roc curve analysis using r language environment. R Journal. 2016; 8(2): 213-230. doi: 10.32614/RJ-2016-042.

17. ROC Analysis: Online ROC Curve Calculator. Available at: http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html. Accessed 08.02.2023.

18. Artificial intelligence in radiology. Available at: https://mosmed.ai/ai/ Accessed 08.02.2023.

19. Колядин В.Л. Пермутационные критерии как универсальный непараметрический подход к проверке статистических гипотез // Радиоэлектроника и информатика. — 2002. –№3. — С.20. [Kolyadin VL. Permutacionnye kriterii kak universal'nyj neparametricheskij podhod k proverke statisticheskih gipotez. Radioelektronika i informatika. 2002; 3: 20. (In Russ.)]

20. Demler OV, Pencina MJ, D' RB, Sr A. Misuse of DeLong test to compare AUCs for nested models. Published online 2012. doi: 10.1002/sim.5328.

21. Клинические испытания систем искусственного интеллекта (лучевая диагностика) / сост. Ю.А. Васильев, А.В. Владзимирский, Д.Е. Шарова и др. // Серия «Лучшие практики лучевой и инструментальной диагностики». — Вып. 113. — 2-е изд., перераб. и доп. — М.: НПКЦ ДиТ ДЗМ, 2023. — 40 с. [Klinicheskie ispytaniya sistem iskusstvennogo intellekta (luchevaya diagnostika). Vasilyev YA, Vladzymyrskyy AV, Sharova DE, et al. Seriya «Luchshie praktiki luchevoj i instrumental'noj diagnostiki». 2023. 40 p. (In Russ.)]

22. Тыров И.А., Васильев Ю.А., Арзамасов К.М и др. Оценка зрелости технологий искусственного интеллекта для здравоохранения: методология и ее применение на материалах московского эксперимента по компьютерному зрению в лучевой диагностике // Врач и информационные технологии. — 2022. — №4. — С.76-92. [Tyrov IA, Vasilyev YA, Arzamasov KM, et al. Assessment of the maturity of artificial intelligence technologies for healthcare: methodology and its application based on the use of innovative computer vision technologies for medical image analysis and subsequent applicability in the healthcare system of Moscow. Medical doctor and information technology. 2022; 4: 76-92. (In Russ.)] doi: 10.25881/18110193_2022_4_76. 21.

23. Probabilistic Graphical Models: Principles and Techniques — Daphne Koller, Nir Friedman.