https://doi.org/10.29001/2073-8552-2024-39-3-188-198
УДК 616-073.7:004.8:004.658

Объем выборки для оценки диагностической точности программного обеспечения на основе технологий искусственного интеллекта в лучевой диагностике

Т.М. Бобровская[1], Ю.А. Васильев[1, 2], Н.Ю. Никитин[1], А.В. Владзимирский[1, 3], О.В. Омелянская[1], С.Ф. Четвериков[1], К.М. Арзамасов[1, 4]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы (НПКЦ ДиТ ДЗМ),
127051, Российская Федерация, Москва, ул. Петровка, 24, стр. 1

[2] Национальный медико-хирургический Центр имени Н.И. Пирогова Министерства здравоохранения Российской Федерации (НМХЦ им. Н.И. Пирогова Минздрава России),
105203, Российская Федерация, Москва, ул. Нижняя Первомайская, 70

[3] Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет) (Первый МГМУ им. И.М. Сеченова Минздрава России),
119991, Российская Федерация, Москва, ул. Трубецкая, 8, стр. 2

[4] МИРЭА – Российский технологический университет (РТУ МИРЭА),
119454, Российская Федерация, Москва, проспект Вернадского, 78

Аннотация

Введение. Проблема обоснования объема выборки является актуальной для различных научных и практических задач. Однако при всем многообразии существующих на сегодня методов вопрос определения минимального количества исследований для валидации программного обеспечения (ПО) на основе технологий искусственного интеллекта (ТИИ) остается открытым.

Цель: определить минимальное количество исследований, необходимых для проведения валидации ПО на основе ТИИ, для решения задач лучевой диагностики с учетом баланса классов «норма» / «патология».

Материал и методы. Анализировались результаты работы ПО на основе ТИИ на наборе данных из 123 301 уникального анонимизированного маммографического исследования. Оценивались выставленные значения по шкале Bi-RADS: 0 – в случае диагностирования врачом 1-го или 2-го класса Bi-RADS («норма») и 1 – в случае классов Bi-RADS 3, 4, 5 («патология»). Изначально баланс классов в исследовании составлял 89,3% («норма») / 10,7% («патология»). Из общего набора данных случайным образом формировалась выборка заданного объема и баланса классов «норма» / «патология», рассчитывалась площадь под кривой операционной характеристики приемника (AUC ROC). Для статистического обоснования описанные действия повторялись 10 000 раз для всех исследуемых объемов и балансов классов. В результате применения данного алгоритма были получены зависимости средних значений AUC ROC от количества исследований для пяти балансов классов (доля «патологии»: 10, 20, 30, 40 и 50%). Далее был проведен анализ законов распределения и поведения AUC ROC в зависимости от количества исследований.

Результаты. Максимальное значение коэффициента вариации значений AUC ROC для 10% доли «патологии» достигается при количестве исследований, равном 190; для 20% – 80 исследований; для 30% – 120 исследований, для 40% – 110 исследований, а для 50% – 70 исследований.

Заключение. При тестировании ПО на основе ТИИ, а также систем поддержки принятия врачебных решений необходимо учитывать, что количество исследований, отражающих наибольшую неоднородность значений AUC ROC (наибольшее отклонение от среднего значения), различно для разных балансов классов. Баланс классов задается, исходя из возможностей исследователя, а минимальный объем – 190 при доле «патологии» 10%, 80 – при 20%, 120 – при 30%, 110 – при 40%, 70 – при 50%.

Бобровская Татьяна Михайловна, e-mail: BobrovskayaTM@zdrav.mos.ru.

Ключевые слова:	искусственный интеллект; размер выборки; ROC-кривая; статистические методы; валидация; лучевая диагностика.
Финансирование:	данная статья подготовлена авторским коллективом в рамках НИОКР «Разработка платформы повышения качества ИИ-Сервисов для медицинской диагностики» (№ ЕГИСУ: 123031400006-0) 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.
Ресурсное обеспечение:	все расчеты выполнялись на персональных компьютерах 64-разрядной архитектуры с операционными системами Windows 10 с использованием оперативной памяти не менее 16 Гб, объемом жесткого диска не менее 500 Гб.
Соответствие принципам этики:	работа проведена на базе одобренного Комитетом по этике и зарегистрированного в ClinicalTrials исследования (NCT04489992).
Для цитирования:	Бобровская Т.М., Васильев Ю.А., Никитин Н.Ю., Владзимирский А.В., Омелянская О.В., Четвериков С.Ф., Арзамасов К.М. Объем выборки для оценки диагностической точности программного обеспечения на основе технологий искусственного интеллекта в лучевой диагностике. *Сибирский журнал клинической и экспериментальной медицины.* 2024;39(3):188–198. https://doi.org/10.29001/2073-8552-2024-39-3-188-198.

Sample size for assessing a diagnostic accuracy of AI-based software in radiology

Tatiana M. Bobrovskaya[1], Yuriy A. Vasilev[1, 2], Nikita Yu. Nikitin[1], Anton V. Vladzimirskyy[1, 3], Olga V. Omelyanskaya[1], Sergey F. Chetverikov[1], Kirill M. Arzamasov[1, 4]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department (Moscow Center for Diagnostics and Telemedicine),
24, Petrovka str., bld. 1, Moscow, 127051, Russian Federation

[2] Pirogov National Medical and Surgical Center,
70, Nizhnyaya Pervomajskaya str., Moscow, 105203, Russian Federation

[3] L.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University),
8, Trubeckaya str., bld. 2, Moscow, 119991, Russian Federation

[4] MIREA – Russian Technological University, Moscow, Russian Federation,
78, Vernadskogo prospekt, Moscow, 119454, Russian Federation

Abstract

Introduction. Determining the minimum sample size for solving various tasks is an extremely important and at the same time unexplored problem. There are many methods, but most of them are not applicable for AI-based software validation.

Aim: To consider a methodology for determining a balance of classes “norm”/ “abnormality” and propose a statistical approach to determine the data amount necessary for testing AI-based software (validation).

Material and Methods. The results of AI-based software were analyzed using dataset of mammograms. Mammograms were classified by the presence of breast cancer (“abnormality”) and the absence of breast cancer (“norm”). The general set contains 123,301 unique studies. The original balance of classes in the study was “norm” 89.3%/“abnormality” 10.7%. As the results of AI-based software (ML-algorithm), a probability of the presence of pathology in the entire study was taken. The following values were used as empirical data (GT): 0 – in case of Bi-RADS classes 1 or 2 diagnosed by a doctor, and 1 – in case of Bi-RADS classes 3, 4, 5. Each data sample is transferred to AI-based software for processing. Quality metrics are calculated based on its results: AUC ROC. All the described actions were repeated 10,000 times for all the studied balances of “norm”/”abnormality”. Based on the results of AUC ROC calculations, mean values were calculated for different random data series with the same balances. Mean AUC ROC values were subjected to analysis.

Results. A maximum value of the coefficient of variation of AUC ROC values for 10% “abnormality” share is achieved at the number of studies equal to 190; for the 20% share, it is 80 studies; for the 30% share – 120 studies, for the 40% share – 110 studies, and for the 50% share – 70 studies.

Conclusion. Summarizing the conducted study results, it can be concluded that when testing AI-based software, it is necessary to consider that the number of studies reflecting the greatest heterogeneity of AUC ROC values (the largest deviation from the mean value) is different for various class balances. If the purpose of validation is to establish the worst-case behavior of AUC ROC values, then for the studied AI-based software, the “abnormality” share should be 10%, and the number of studies 190. If the validation is carried out under conditions of a limited amount of data, then the “abnormality” share should be 50% and the number of studies equal to 70.

Keywords:	artificial intelligence; statistical methods; sampling; validation; radiology.
Funding:	this paper was prepared by a group of authors as a part of the research and development effort titled “Development of a platform for improving the quality of AI services for clinical diagnostics” (USIS No.: 123031400006-0) in accordance with the Order No. 1196 dated December 21, 2022 “On approval of state assignments funded by means of allocations from the budget of the city of Moscow to the state budgetary (autonomous) institutions subordinate to the Moscow Health Care Department, for 2023 and the planned period of 2024 and 2025” issued by the Moscow Health Care Department.
Resource support:	all calculations were performed on personal computers of 64-bit architecture with Windows 10 operating systems using RAM of at least 16 GB, and a hard disk of at least 500 GB.
Compliance with ethical standards:	the work was carried out on the basis of a study approved by the Ethics Committee and registered with Clinical Trials (NCT04489992).
For citation:	Bobrovskaya T.M., Vasilev Yu.A., Nikitin N.Yu., Vladzimirskyy A.V., Omelyanskaya O.V., Chetverikov S.F., Arzamasov K.M. Sample Size for Assessing a Diagnostic Accuracy of AI-based Software in Radiology. *The Siberian Journal of Clinical and Experimental Medicine*. 2024;39(3):188–198. https://doi.org/10.29001/2073-8552-2024-39-3-188-198.

Введение

Технологии компьютерного зрения и искусственного интеллекта начинают формировать систему поддержки принятия врачебных решений при выявлении патологий у пациентов. В работе большинства алгоритмов компьютерного зрения принято выделять несколько этапов, в частности, предобработку изображения, распознавание (классификацию обнаруженного объекта по различным категориям) и принятие системой решения о наличии интересующего объекта на изображении [1].

Применение программного обеспечения (ПО) на основе технологий искусственного интеллекта (ТИИ) автоматизирует процесс классификации изображений, снижая влияние человеческого фактора на процесс обнаружения объектов (например, «патологий» на медицинских изображениях). Успешное применение ПО на основе ТИИ в приложениях компьютерного зрения было продемонстрировано во многих работах [2, 3]. В частности в [4] было рассмотрено применение нескольких топологий нейронных сетей для классификации рентгенологических снимков по группам «норма» / «патология». Одновременно с этим мы можем наблюдать стремительный рост числа ПО на основе ТИИ, зарегистрированных как медицинское изделие [5].

Одним из крупнейших проектов является Эксперимент по использованию инновационных технологий в области компьютерного зрения для анализа медицинских изображений и дальнейшего применения в системе здравоохранения города Москвы (далее Эксперимент) [6]. Реализация такого масштабного проекта позволила разработать методологию оценки ПО на основе ТИИ с целью так называемой внешней валидации. Внешняя валидация – это оценка качества работы ПО на основе ТИИ на наборе данных, который не использовался при разработке [7]. Внешняя валидация проводится незаинтересованной стороной на независимом наборе данных, что необходимо для объективной оценки обобщаемости и воспроизводимости результатов работы ПО на основе ТИИ [7].

Качество классификации исследований с помощью ПО на основе ТИИ зависит не только от особенностей алгоритмов ИИ, но также от качества и количества данных, на которых проходило обучение [4]. Качество данных определяется наличием технологических дефектов исследования, затрудняющих диагностику [8], а также непосредственно процессом создания наборов данных, включая стратегию разметки, верификации, структуризации данных, квалификации разметчиков и организации процесса создания набора данных в целом [9, 10]. Вопрос количества данных для обучения регулярно освещается в публикациях, однако зачастую указывается количество исследований, но не дается его обоснование [2]. Количество колеблется от нескольких тысяч до полутора миллионов исследований [11]. Такой разброс во многом обусловлен сложностью и стоимостью создания качественного набора данных, включая этические и законодательные аспекты [9, 10]. Еще более остро стоит вопрос количества данных для валидации ПО на основе ТИИ. В работе F. Harrel [12] авторы предложили использовать 100–200 исследований для валидации прогностической регрессионной модели. В более поздних работах [11, 13] также были указаны альтернативные варианты оценки диагностической точности, основанные преимущественно на достижении заданной мощности, в том числе экспериментальные исследования на выборках, значительно превышающих 100 и 200 исследований (более 10 000 исследований) [11]. Однако такой подход не всегда может быть реализован в клинической практике.

В работе [14] авторы предлагают различные способы расчета размера выборки, исходя из показателей калибровочных кривых, площади под ROC-кривой, чистой вы-

годы и достижения заданного доверительного интервала. В этой же работе отмечается важность баланса классов в выборке, однако методики определения предложено не было.

Цель: представить новую методику определения количества исследований, необходимых и достаточных для проведения валидации ПО на основе ТИИ с учетом баланса классов «норма» / «патология».

Материал и методы

Дизайн исследования – ретроспективное обсервационное когортное исследование на базе одобренного Комитетом по этике и зарегистрированного в исследовании ClinicalTrials (NCT04489992).

Набор данных содержит 123 301 уникальное маммографическое исследование, полученное за период с 1 сентября 2021 по 27 декабря 2021 гг. из ЕРИС ЕМИАС (Единый Радиологический Информационный Сервис Единой Медицинской Информационно-Аналитической Системы). Критериями включения были наличие ответа от заданного ПО на основе ТИИ, а также описания заключения от врача-рентгенолога. Критерием исключения являлось отсутствие классификации по Bi-RADS в тексте заключения. Перед использованием данные были предварительно обработаны с целью удаления личной информации пациентов (анонимизация).

Маммографические исследования классифицировались по наличию («патология») и отсутствию («норма») рака молочной железы. Верификация проводилась по текстовым протоколам заключений врачей-рентгенологов с помощью алгоритма естественной обработки языка (MedLabel[1]). Анализировались выставленные значения по шкале Bi-RADS: 0 – в случае диагностирования врачом 1-го или 2-го класса Bi-RADS («норма») и 1 – в случае классов Bi-RADS 3, 4, 5 («патология») [15]. Изначально баланс классов в исследовании составлял «норма» – 89,3% / «патология» – 10,7%.

Производилась оценка результатов работы ПО на основе ТИИ, в качестве которого выступал один из сервисов искусственного интеллекта по направлению «маммография», участвующий в Эксперименте [16]. Валидация проходила в несколько этапов. На первом этапе данные были разделены на две группы – «норма» и «патология». Из разделенных данных случайным образом формировались выборки с балансом классов «норма» / «патология», содержащие «патологию» в количестве 50, 40, 30, 20 и 10%. Минимальная выборка, сформированная случайным образом, содержала 30 исследований, далее размер выборки увеличивался с шагом 10 с учетом сохранения доли «патологии». Максимальный объем изучаемой выборки составил 26 386 (количество исследований с патологией, умноженное на 2) исследований и обусловлен ограничением вычислительных мощностей.

Для каждого баланса классов и объема случайным образом формировались подвыборки 10 000 раз с возвращением (так называемый бутреппинг), для них рассчитывались значения AUC ROC (площадь под кривой операционной характеристики приемника). По результатам работы ПО на основе ТИИ были определены средние значения AUC ROC для различных случайных наборов исследований с одинаковым балансом классов. На рисунке 1 представлена блок-схема описанного алгоритма.

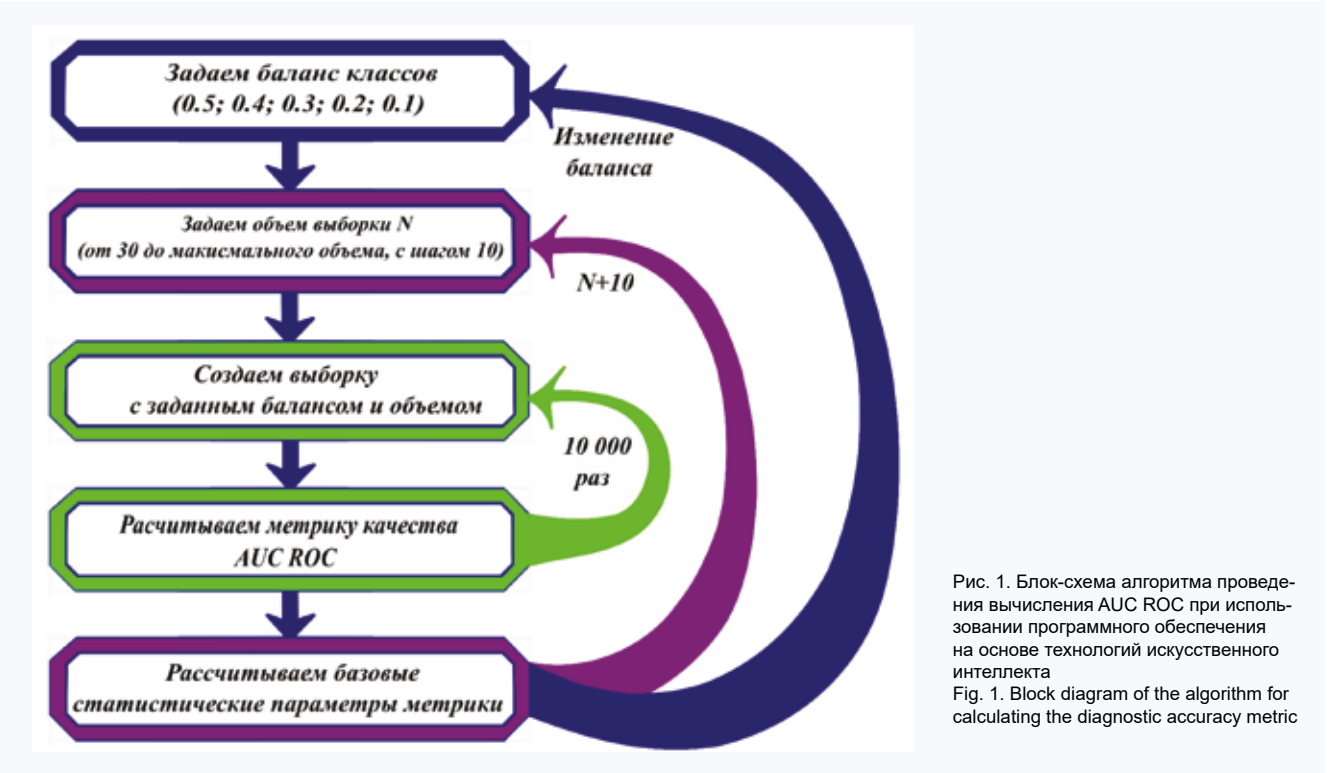

Рис. 1. Блок-схема алгоритма проведения вычисления AUC ROC при использовании программного обеспечения на основе технологий искусственного интеллекта
Fig. 1. Block diagram of the algorithm for calculating the diagnostic accuracy metric

[1] Свидетельство о государственной регистрации программы для ЭВМ № 2020664321 Российская Федерация. MedLabel – автоматизированный анализ медицинских протоколов: № 2020663035: заявл. 27.10.2020: опубл. 11.11.2020 / С.П. Морозов [и др.]; заявитель Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ «НПКЦ ДиТ ДЗМ»).

Средние значения AUC ROC были подвергнуты трем типам анализа:

1. Фурье-анализ значений AUC ROC в зависимости от количества данных.
2. Анализ наиболее близкого теоретического распределения значений AUC ROC посредством применения информационных критериев Акаике и Байеса.
3. Анализ коэффициента вариации в зависимости от количества исследований для установленного наиболее близкого типа распределения AUC ROC.

Анализ наиболее близкого распределения полученных средних значений AUC ROC проводился для 10 различных распределений:

1. Нормального (Гауссово).
2. Логарифмически нормального.
3. Экспоненциального.
4. Пуассона.
5. Коши.
6. Гамма.
7. Логистического.
8. Биноминального.
9. Геометрического.
10. Вейбулла.

Параметры каждого из распределений вычислялись методом максимального правдоподобия. Совокупность описанных выше методов составляет методику определения необходимого и достаточного количества исследований для проведения валидации ПО на основе ТИИ с использованием критерия диагностической точности AUC ROC.

Весь расчет показателей AUC ROC ПО на основе ТИИ и формирование подвыборок из генеральной совокупности осуществлялся на языке Python, версия 3.6. Фурье-анализ и определение наиболее близких типов распределений проводились на программном обеспечении, реализованном на языке R.

Результаты и обсуждение

На рисунке 2 представлены результаты расчета AUC ROC для ПО на основе ТИИ.

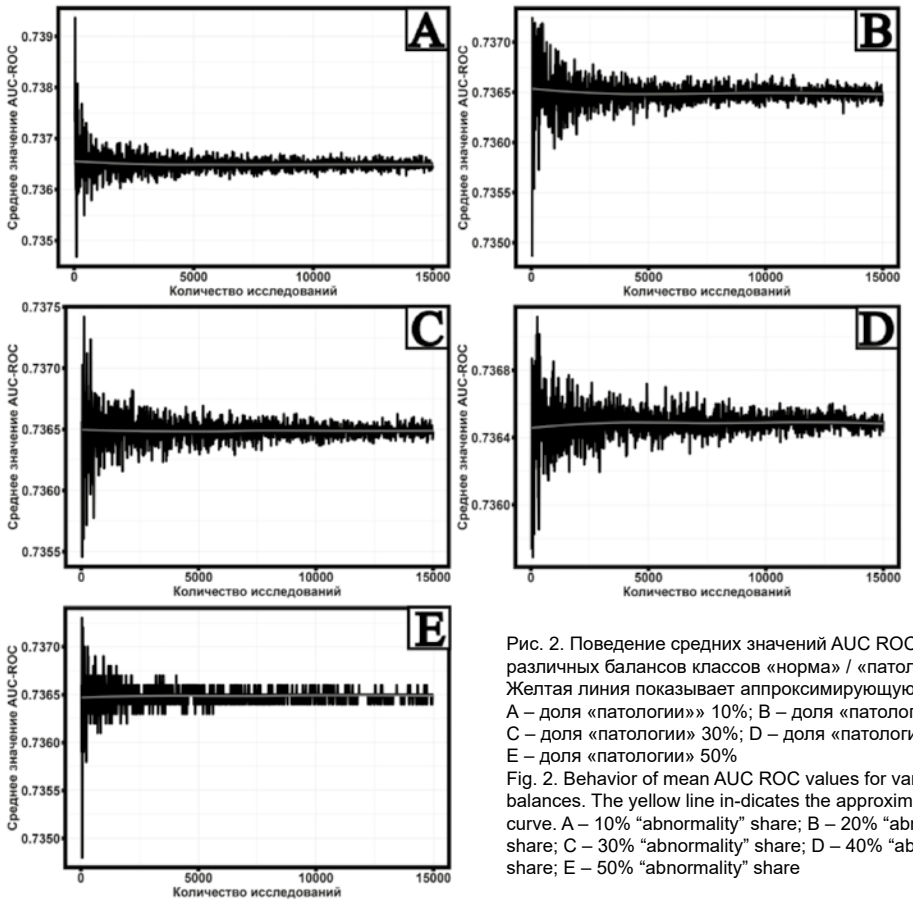

Рис. 2. Поведение средних значений AUC ROC для различных балансов классов «норма» / «патология». Желтая линия показывает аппроксимирующую кривую. A – доля «патологии»» 10%; B – доля «патологии» 20%; C – доля «патологии» 30%; D – доля «патологии» 40%; E – доля «патологии» 50%

Fig. 2. Behavior of mean AUC ROC values for various balances. The yellow line in-dicates the approximating curve. A – 10% “abnormality” share; B – 20% “abnormality” share; C – 30% “abnormality” share; D – 40% “abnormality” share; E – 50% “abnormality” share

Предварительный анализ поведения значений AUC ROC показывает наличие периодической зависимости от количества исследований. Для балансов «норма» / «патология» с долей «патологии» 10, 20% (см. рис. 2a и 2b) наблюдается нисходящий тренд от 30 до 5 000 исследований (желтая линия) и восходящий тренд зависимости значений AUC ROC от количества исследований для баланса с долей «патологии» 40% (см. рис. 2d). Далее эта тенденция меняется на линейную. Полностью линейный тренд зависимости AUC ROC от количества исследова-

ний наблюдается для балансов с долей «патологии»» 30 и 50% (см. рис. 2c и 2e). Представленная зависимость имеет явно выраженный колебательный характер и непрерывна на участке от 0 до N количества исследований. Учитывая выявленный характер зависимости площади под кривой операционной характеристики приемника, можно представить зависимость AUC от количества исследований как

$$\overline{AUC} = F(n) \quad (1)$$

где *F(n)* – некоторая периодическая функция, зависящая от количества исследований.

Если функция *F(n)* непрерывна и интегрируема во всем диапазоне изменения числа исследований, то можно определить спектральную плотность как

$$\widehat{F(n)} = \sum_{j=1}^{N} F_j(n) * \exp(-2\pi i(\gamma, n_j)) \quad (2)$$

где *F(n)* – функция уравнения (1); *n* – количество образцов; *N* – общее количество исследований; *γ* – аргумент спектральной функции:

$$\gamma = Re(\widehat{F(n)})/Im(\widehat{F(n)}) \quad (3)$$

где *Re((F(n))* – вещественная часть спектральной функции; *Im((F(n))* – мнимая часть спектральной функции.

Учитывая результаты, представленные на рисунке 2, и уравнения (2) и (3), был проведен Фурье-анализ средних значений AUC ROC для выявления особенностей в поведении. Результаты вычисления аргумента (3) спектральной функции (2) в зависимости от количества испытаний, полученные с помощью Фурье-анализа, представлены на рисунке 3.

На рисунке 3 для всех балансов можно выделить два основных паттерна поведения главных максимумов и минимумов аргумента спектральной функции AUC ROC. Исключение составляет поведение максимума аргумента спектральной функции AUC ROC баланса классов «норма» / «патология» с долей «патологии» 10%. Значения основных максимумов и минимумов аргумента спектральной функции AUC ROC были подвергнуты дальнейшему анализу на наличие симметрии [17] вида:

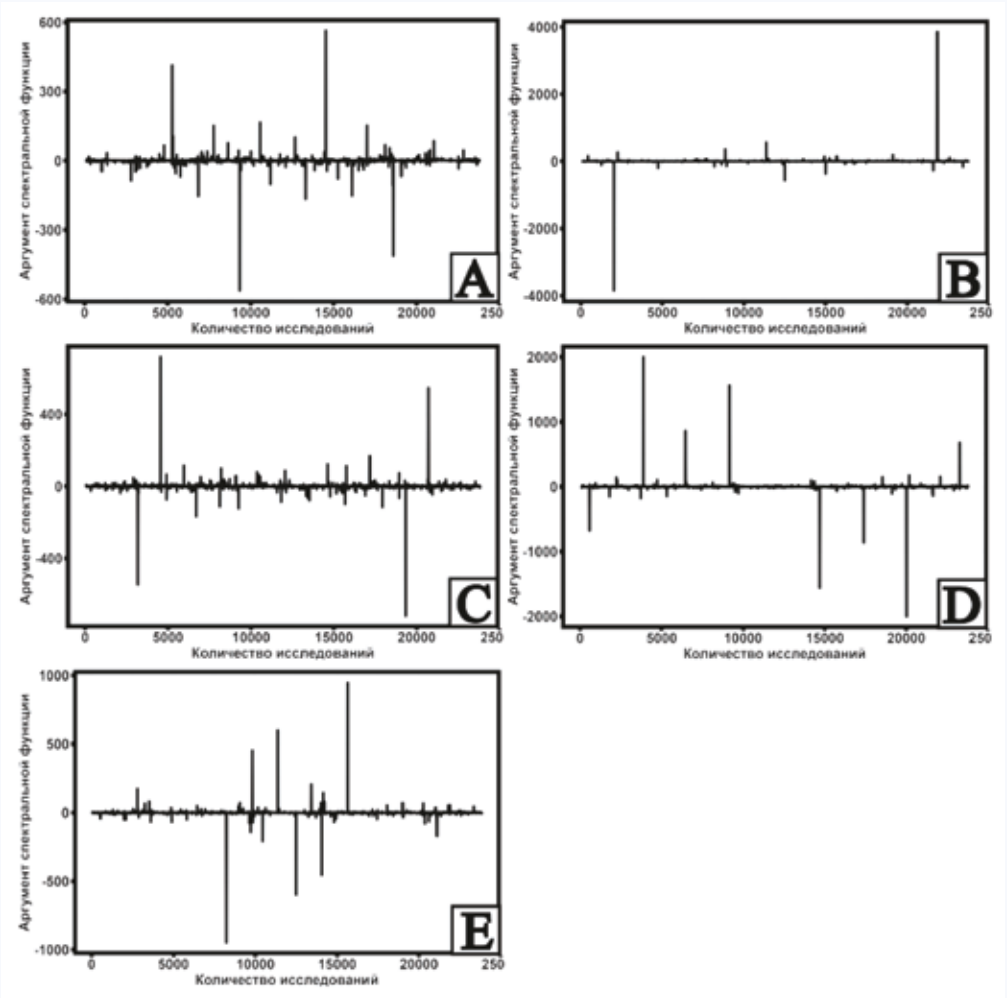

Рис. 3. Зависимость аргумента спектральной функции AUC ROC от количества исследований для разных балансов классов. A – доля «патологии» 10%; B – доля «патологии» 20%; C – доля «патологии» 30%; D – доля «патологии» 40%; E – доля «патологии» 50%
Fig. 3. Dependence of the argument of Fourier spectral function of the AUC ROC on the number of studies in the sample. A – 10% “abnormality” share; B – “20% “abnormality” share; C –30% “abnormality” share; D – 40% “abnormality” share; E – 50% “abnormality” share

$$\gamma(n) + \gamma(n_T - n) = 0 \quad (4)$$

где n_T – точка симметрии аргумента спектральной функции.

На рисунке 4 показана зависимость количества образцов, соответствующих главным максимумам и минимумам аргумента спектральной функции AUC ROC, от доли «патологии» в балансе классов «норма» / «патология».

Синяя линия на рисунке 4 обозначает середину интервала между первыми максимумами и минимумами аргумента спектральных функций AUC ROC. Для всех рассматриваемых долей «патологии» в балансе классов «норма» / «патология» середина интервала соответствует значению 11 940 исследований. Полученное значение является точкой перехода n_T.

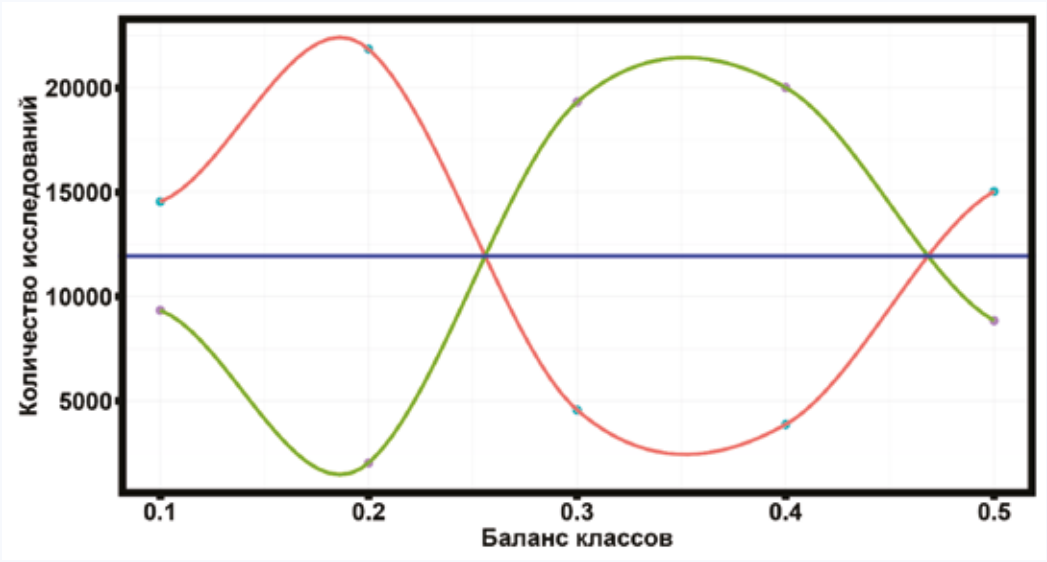

Рис. 4. Зависимость количества исследований, соответствующих главным максимумам и минимумам аргумента спектральной функции AUC ROC, от доли «патологии» в балансе классов «норма» / «патология». Голубые точки с красной линией описывают количество исследований, соответствующих главным максимумам аргумента Фурье образа в зависимости от баланса классов. Фиолетовые точки с салатовой кривой описывают зависимость количества исследований, соответствующих главным минимумам аргумента Фурье образа в зависимости от баланса классов
Fig. 4. Dependence of the number of studies corresponding to the main maxima and minima of the argument of AUC ROC spectral function on the “abnormality” share in the balance of “norm” / “abnormality” classes. Blue dots with red line describe the number of studies corresponding to the main maxima of the Fourier image argument as a function of class balance. Purple dots with a salad curve describe the dependence of the number of studies corresponding to the main minima of the Fourier image argument depending on the balance of classes

Чтобы найти максимальное отклонение от линии тренда (см. рис. 2) среднего показателя точности диагностики слева и справа от точки перехода (11 940 исследований), определяем ближайший тип простого распределения по минимуму критериев Акаике и Байеса. В таблице представлены результаты сравнения распределения значений AUC ROC слева и справа от точки перехода для десяти различных распределений.

Таблица. Типы распределений до и после точки перехода n_T (11 940 исследований)

Table. Types of distributions up to and after transition point n_T (11,940 studies)

№	Доля «патологии» в балансе «норма» / «патология»	Тип распределения до n_T	Тип распределения после n_T
1	0.1	Коши	Нормальное
2	0.2	Коши	Нормальное
3	0.3	Коши	Логистическое
4	0.4	Коши	Логарифмическое нормальное
5	0.5	Коши	Логистическое

Из результатов анализа поведения аргумента спектральной функции AUC ROC и анализа ближайшего теоретического распределения следует, что до точки перехода (11 940 исследований) для всех балансов классов сохраняется один и тот же тип распределения – распределение Коши. После точки перехода (11 940 исследований) тип распределения меняется. Нормальное распределение наблюдается при 10 и 20% «патологии», логистическое – при 30 и 50% «патологии», а логнормальное распределение значений AUC ROC – при 40% «патологии».

Для оценки однородности значений AUC ROC был проведен анализ коэффициента вариации в зависимости от количества исследований (до 11 940 исследований). В случае распределения Коши коэффициент вариации рассчитывался по уравнению:

$$K = \frac{\gamma}{x_0} \quad (4)$$

где γ – масштабный параметр в распределении Коши; x_0 – параметр сдвига в распределении Коши.

На рисунке 5 представлены результаты расчета зависимости коэффициента вариации распределения значений AUC ROC от количества исследований для пяти долей «патология» в балансе классов «норма» / «патология».

Максимальное значение коэффициента вариации значений AUC ROC для 10% доли «патологии» достигается при количестве исследований, равном 190; для 20% – 80 исследований; для 30% – 120 исследований, для 40% – 110 исследований, а для 50% – 70 исследований. Таким образом, формируется гипотеза о возможности следующего применения полученных результатов:

Определение AUC ROC на наборе данных с заданным балансом классов и соответствующим объемом выборки.

Определение доверительного интервала для AUC ROC с помощью метода бутстреппинга.

Использование нижней границы доверительного интервала в качестве порогового значения для принятия решения о допуске ПО на основании ТИИ AUC ROC.

Полученные результаты сопоставимы с данными предыдущего исследования, где оценка качества осуществлялась на основании анализа количества дефектов. Было показано, что оптимальный объем выборки ТИИ на основе ПО составляет 80 исследований [8].

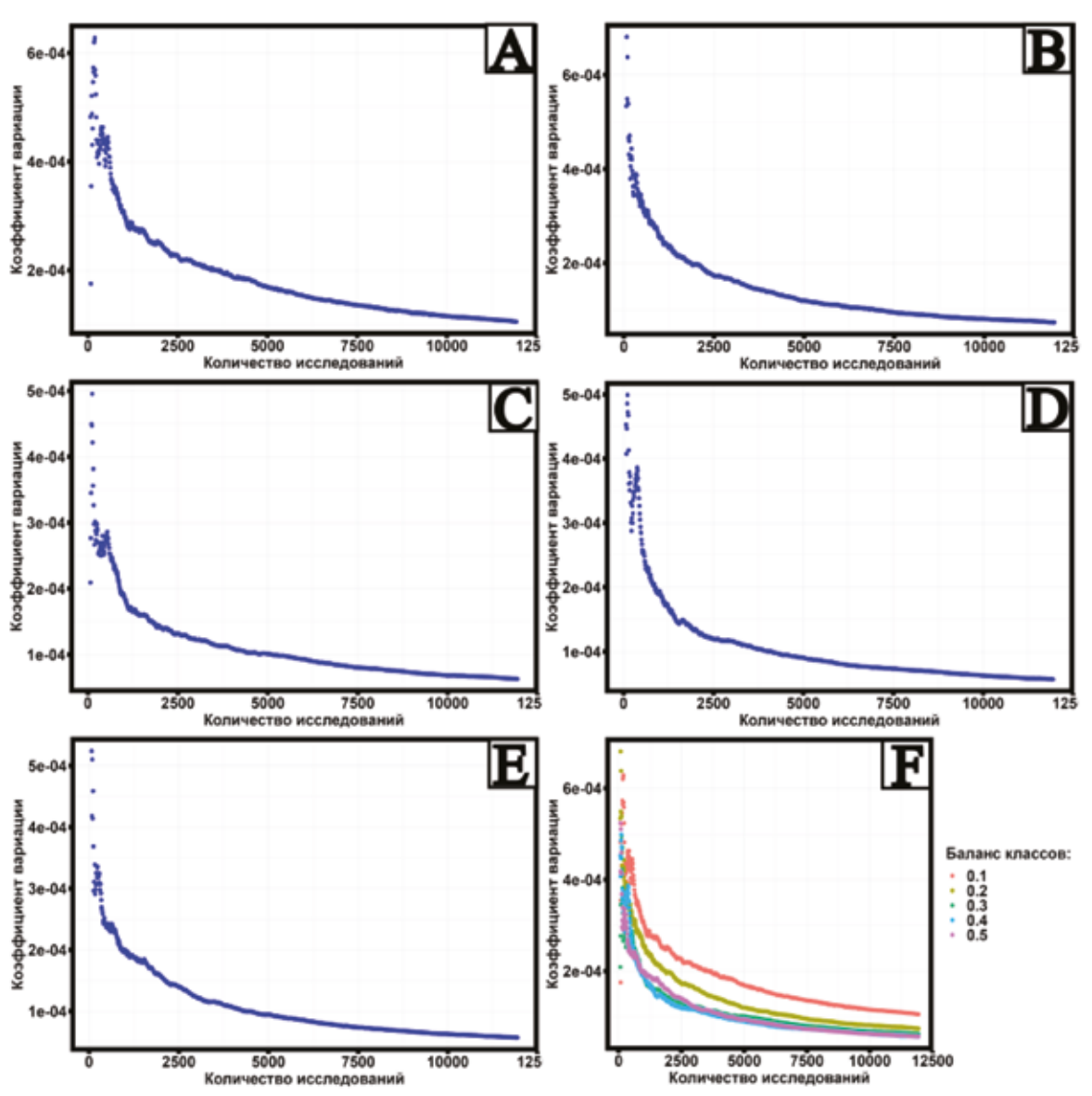

Рис. 5. Коэффициент вариации значений AUC ROC в зависимости от количества исследований для разных балансов классов. A – доля «патологии» 10%; B – доля «патологии» 20%; C – доля «патологии» 30%; D – доля «патологии» 40%; E – доля «патологии» 50%; F – обобщенное представление для всех долей «патологии»

Fig. 5. Coefficient of variation of AUC ROC values depending on the number of studies. A – for 10% “abnormality” share; B – for 20% “abnormality” share; C – for 30% “abnormality” share; D – for 40% “abnormality” share; E – for 50% “abnormality” share; F – a generalized representation for all “abnormality” shares

В представленной работе предложен оригинальный подход к обоснованию необходимого объема данных при заданном балансе классов в исследовании. Ранее в работах встречаются рассуждения о соотношении классов в валидационном наборе данных. В работе [14] предлагается выбирать между сбалансированной выборкой (50/50) и балансом, обусловленным частотой встречаемости целевого признака в популяции. Однако частота встречаемости признака в популяции известна не всегда, она может варьировать с течением времени и в разных популяциях, может быть очень низкой для редко встречающихся патологий. На основании вышеизложенного, логичным решением является задавать баланс классов как постоянную величину и выбирать объем необходимых для валидации данных для заданного баланса классов. Выбор баланса класса зависит от условий, которыми располагает исследователь при создании набора данных, т. е. финансовых, кадровых ресурсов, а также наличия самих исследований в необходимом соотношении и количестве.

Применение преобразования Фурье к колебаниям значений AUC ROC позволило выявить точку перехода, что является своеобразной границей между двумя различными распределениями. Эта граница соответствует значению 11 940 исследований. При использовании меньшего или равного количества исследований значения AUC ROC для всех изученных долей «патологии» в балансе классов «норма» / «патология» распределяются по закону, близкому к распределению Коши. Причем если количество исследований превышало 11 940, то значения AUC ROC имели нормальное распределение для 10 и 20% долей «патологии», логистическое – для 30 и 50%, логарифмически нормальное – для 40% долей «патологии».

Для оценки однородности значений AUC ROC в зависимости от количества исследований был проведен анализ коэффициента вариации распределения Коши, который показал, что наибольшее отклонение от среднего значения AUC ROC наблюдается при доле «патологии» 10% в балансе классов «норма» / «патология» и соответствует количеству исследований, равному 190.

Также следует отметить, что отклонение среднего значения AUC ROC от линии тренда с увеличением количества исследований уменьшается, что свидетельствует о том, что при использовании ПО на основе ТИИ в клинической практике могут демонстрироваться показатели диагностической точности, отличные от полученных при валидационном тестировании. По этой причине на этапе валидации ПО на основе ТИИ необходимо определить максимальные пределы изменения показателей диагностической точности и в дальнейшем проводить регулярный мониторинг его работы [18]. Разработанный подход к определению количества исследований, необходимых для валидации, также может использоваться в этих целях, например, в программной системе мониторинга на основе технологии искусственного интеллекта[1].

Заключение

Обобщая результаты проведенного исследования, можно сделать вывод, что при тестировании ПО на основе ТИИ необходимо учитывать, что количество исследований, отражающих наибольшую неоднородность значений AUC ROC (наибольшее отклонение от среднего значения), различно для разных балансов классов. Баланс классов задается, исходя из возможностей исследователя, а минимальный объем 190 при доле «патологии» 10%, 80 – при 20%, 120 – при 30%, 110 – при 40%, 70 – при 50%. В этом случае будет наблюдаться максимальное отклонение от среднего значения AUC ROC для исследуемого программного обеспечения на основе ТИИ. Полученные результаты можно использовать для валидации ПО на основе ТИИ, а также систем поддержки принятия врачебных решений как при допуске к работе в практической деятельности, так и при дальнейшем мониторинге.

Ограничение исследований

Проведенные исследования были ограничены одной версией ПО на основе ТИИ и долей «патологий» до 50%. В дальнейших исследованиях будет проведен аналогичный анализ для полного баланса классов с долей «патологии» от 0 до 100% с шагом 10% и большего количества версий ПО на основе ТИИ для выявления более общей закономерности.

Литература / References

1. Chervyakov N.I., Lyakhov P.A., Deryabin M.A., Nagornov N.N., Valueva M.V., Valuev G.V. Residue number system-based solution for reducing the hardware cost of a convolutional neural network. *Neurocomputing*. 2020;407:439–453. DOI: 10.1016/j.neucom.2020.04.018.
2. Aggarwal R., Sounderajah V., Martin G., Ting D.S.W., Karthikesalingam A., King D. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit. Med*. 2021;4:65. DOI: 10.1038/s41746-021-00438-z.
3. Тыров И.А., Васильев Ю.А., Арзамасов К.М., Владзимирский А.В., Шулькин И.М., Омелянская О.В. 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;4:76–92.
Tyrov I.A., Vasilev Yu.A., Arzamasov K.M., Vladzimirskyy A.V., Shulkin I.M., Omelyanskaya O.V. 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.
4. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 27–30 June, 2016. IEEE Computer Society; 2015;2016:770–778. DOI: 10.1109/CVPR.2016.90.
5. Гусев А.В., Морозов С.П., Кутичев В.А., Новицкий Р.Э. Нормативно-правовое регулирование программного обеспечения для здравоохранения, созданного с применением технологий искусственного интеллекта, в Российской Федерации. *Медицинские технологии. Оценка и выбор.* 2021;(1):36–45.
Gusev A.V., Morozov S.P., Kutichev V.A., Novitsky R.E. Legal regulation of artificial intelligence software in healthcare in the Russian Federation. *Medical Technologies. Assessment and Choice.* 2021;(1):36–45. (In Russ.) DOI: 10.17116/medtech20214301136.

[1] Свидетельство о государственной регистрации программы для ЭВМ № 2023665713 Российская Федерация. Веб-платформа технологического и клинического мониторинга результатов работы алгоритмов анализа цифровых медицинских изображений: № 2023664691: заявл. 11.07.2023: опубл. 19.07.2023 / Ю.А. Васильев, А.В. Владзимирский, О.В. Омелянская [и др.]; заявитель Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы».

6. Васильев Ю.А., Владзимирский А.В. (ред.) Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента: Монография; 2-е изд., перераб. и дополн. М.: Издательские решения, 2023;376.
Vasilev YU.A., Vladzymyrskyy A.V. (eds.) Komp'yuternoe zrenie v luchevoj diagnostike: pervyj etap Moskovskogo eksperimenta : Monografiya. 2-e izdanie, pererabotannoe i dopolnennoe. Moscow: Izdatel'skie resheniya, 2023;376. (In Russ.).
7. Ramspek C.L., Jager K.J., Dekker F.W., Zoccali C., van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin. Kidney J.* 2021;14(1). DOI: 10.1093/ckj/sfaa188.
8. Четвериков С.Ф., Арзамасов К.М., Андрейченко А.Е., Новик В.П., Бобровская Т.М., Владзимирский А.В. Подходы к формированию выборки для контроля качества работы систем искусственного интеллекта в медико-биологических исследованиях. *Современные технологии в медицине.* 2023;15(2):19–25.
Chetverikov S.F., Arzamasov K.M., Andreichenko A.E., Novik V.P., Bobrovskaya T.M., Vladzimirskyy A.V. Approaches to sampling for quality control of artificial intelligence in biomedical research. *Modern Technologies in Medicine*. 2023;15(2):19–25. (In Russ.). DOI: 10.17691/stm2023.15.2.02.
9. Васильев Ю.А., Бобровская Т.М., Арзамасов К.М., Четвериков С.Ф., Владзимирский А.В., Омелянская О.В. и др. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике. *Менеджер здравоохранения.* 2023;(4):28–41.
Vasilev Y.A., Bobrovskaya T.M., Arzamasov K.M., Chetverikov S.F., Vladzymyrskyy A.V., Omelyanskaya O.V. et al. Medical datasets for machine learning: fundamental principles of standartization and systematization. *Manager Zdravookhranenia*. 2023;(4):28–41. (In Russ.). DOI: 10.21045/1811-0185-2023-4-28-41.
10. Васильев Ю.А., Арзамасов К.М., Владзимирский А.В., Омелянская О.В., Бобровская Т.М., Шарова Д.Е. и др. Подготовка набора данных для обучения и тестирования программного обеспечения на основе технологии искусственного интеллекта: учеб. пособие. М.: Издательские решения; 2024:140. ISBN: 978-5-0062-1244-2.
Vasilev YU.A., Arzamasov K.M., Vladzimirskij A.V., Omelyanskaya O.V., Bobrovskaya T.M. et al. Podgotovka nabora dannyh dlya obucheniya i testirovaniya programmnogo obespecheniya na osnove tekhnologii iskusstvennogo intellekta: Uchebnoe posobie. Moscow: Izdatel'skie resheniya; 2024:140. (In Russ.). ISBN: 978-5-0062-1244-2.
11. Collins G.S., Ogundimu E.O., Altman D.G. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat. Med.* 2016;35(2):214–226. DOI: 10.1002/sim.6787.
12. Harrell F.E., Lee K.L., Mark D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med*. 1996;15(4):361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
13. Vergouwe Y., Steyerberg E.W., Eijkemans M.J.C., Habbema J.D.F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* 2005;58(5):475–483. DOI: 10.1016/j.jclinepi.2004.06.017.
14. Riley R.D., Debray T.P.A., Collins G.S., Archer L., Ensor J., van Smeden M. et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* 2021;40(19):4230–4251. DOI: 10.1002/sim.9025.
15. Breast Imaging Reporting & Data System. American College of Radiology [Internet]. [cited 2024 Jan 23]. URL: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads (16.04.2024).
16. Павлович П.И., Бронов О.Ю., Капнинский А.А., Абович Ю.А., Рычагова Н.И. Сравнительное исследование результатов анализа данных цифровой маммографии системы на основе искусственного интеллекта «Цельс» и врачей-рентгенологов. *Digital Diagnostics.* 2021;2(2S):22–23.
Pavlovich P.I., Bronov O.Y., Kapninsky A.A., Abovich Y.A., Rychagova N.I. Comparative study of the digital mammography data analysis system based on artificial intelligence "Celsus" and radiologists. *Digital Diagnostics*. 2021;2(2S):22–23. (in Russ.). DOI: 10.17816/DD83184.
17. Kashyap R.L. (ed.) Dynamic stochastic models from empirical data: eBook, Vol. 122. Elsiever B.V.; USA: Academic Press, 1976. ISBN: 978-0-12-400550-1.
18. Васильев Ю.А., Владзимирский А.В., Омелянская О.В., Шулькин И.М. Арзамасов К.М., Никитин Н.Ю. Оценка зрелости технологий искусственного интеллекта для здравоохранения: методические рекомендации. Вып. 123. М.: Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы;. 2023:28.
Assessment of maturity of artificial intelligence technologies for healthcare: methodological recommendations; issue 123. Moscow: Scientific and Practical Clinical Centre of Diagnostics and Telemedicine Technologies of the Moscow City Health Department; 2023:28.

Информация о вкладе авторов

Бобровская Т.М. – дизайн эксперимента, разработка ПО, написание рукописи.

Васильев Ю.А. – формирование концепции исследования.

Никитин Н.Ю. – разработка методологии, проведение статистического анализа, написание рукописи.

Владзимирский А.В. – формирование концепции исследования, научное сопровождение.

Омелянская О.В. – администрирование проекта.

Четвериков С.Ф. – предварительный анализ задачи, дизайн эксперимента.

Арзамасов К.М. – сбор данных, дизайн эксперимента, сбор литературных данных.

Все авторы внесли значимый вклад в проведение исследования и подготовку статьи, прочли и одобрили финальную версию статьи перед подачей к публикации.

Information on author contributions

Bobrovskaya T.M. – experiment design, software development, manuscript writing.

Vasilev Yu.A. – formation of the research concept.

Nikitin N.Yu. – methodology development, statistical analysis, manuscript writing.

Vladzimirskyy A.V. – formation of the research concept, scientific support.

Omelyanskaya O.V. – project administration, funding acquisition.

Chetverikov S.F. – preliminary analysis of the problem, design of the experiment.

Arzamasov K.M. – dataset creation, experiment design, collection of literary data

All authors made significant contribution to the research and preparation of the article, read and approved the final version of the article before publication.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Conflicts of Interest: the authors declare no conflict of interest.

Сведения об авторах

Бобровская Татьяна Михайловна, младший научный сотрудник, отдел инновационных технологий, НПКЦ ДиТ ДЗМ, Москва, https://orcid.org/0000-0002-2746-7554.

E-mail: BobrovskayaTM@zdrav.mos.ru.

Васильев Юрий Александрович, канд. мед. наук, директор НПКЦ ДиТ ДЗМ; заведующий кафедрой лучевой диагностики с курсом клинической радиологии, доцент кафедры, НМХЦ им. Н.И. Пирогова Минздрава

Information about the authors

Tatiana M. Bobrovskaya, Junior Research Scientist, Department of Innovative Technologies, Moscow Center for Diagnostics and Telemedicine, Moscow, https://orcid.org/0000-0002-2746-7554.

E-mail: BobrovskayaTM@zdrav.mos.ru.

Yuriy A. Vasilev, Cand. Sci. (Med.), Director of Moscow Center for Diagnostics and Telemedicine; Head of the Department of Radiation Diagnostics with a course of Clinical Radiology; Associate Professor of the

России, Москва, https://orcid.org/0000-0002-0208-5218.
E-mail: VasilevYA1@zdrav.mos.ru.

Никитин Никита Юрьевич, канд. физ.-мат. наук, старший научный сотрудник, отдел медицинской информатики, радиомики и радиогеномики, НПКЦ ДиТ ДЗМ, Москва, https://orcid.org/0000-0002-3193-8320.
E-mail: NikitinNY@zdrav.mos.ru.

Владзимирский Антон Вячеславович, д-р мед. наук, заместитель директора по научной работе, НПКЦ ДиТ ДЗМ; профессор, кафедра информационных и интернет-технологий, Первый МГМУ им. И.М. Сеченова Минздрава России (Сеченовский Университет), Москва, https://orcid.org/0000-0002-2990-7736.
E-mail: VladzimirskijAV@zdrav.mos.ru.

Омелянская Ольга Васильевна, руководитель по управлению подразделениями Дирекции наука, НПКЦ ДиТ ДЗМ, Москва, https://orcid.org/0000-0002-0245-4431.
E-mail: OmelyanskayaOV@zdrav.mos.ru.

Четвериков Сергей Федорович, канд. техн. наук, руководитель сектора, отдел медицинской информатики, радиомики и радиогеномики, НПКЦ ДиТ ДЗМ, Москва, https://orcid.org/0000-0002-3097-8881.
E-mail: tschetserg@yandex.ru.

Арзамасов Кирилл Михайлович, канд. мед. наук, руководитель отдела медицинской информатики, радиомики и радиогеномики, НПКЦ ДиТ ДЗМ; доцент, кафедра технологий искусственного интеллекта, РТУ МИРЭА, Москва, https://orcid.org/0000-0001-7786-0349.
E-mail: ArzamasovKM@zdrav.mos.ru.

✉ **Бобровская Татьяна Михайловна,** e-mail: BobrovskayaTM@zdrav.mos.ru.

Поступила 02.02.2024;
рецензия получена 02.04.2024;
принята к публикации 13.05.2024.

Department, Pirogov National Medical and Surgical Center, Moscow, https://orcid.org/0000-0002-0208-5218.
E-mail: VasilevYA1@zdrav.mos.ru.

Nikita Yu. Nikitin, Cand. Sci. (Phis.-Mat.), Senior Research Scientist, Department of Medical Informatics, Radiomics and Radiogenomics, Moscow Center for Diagnostics and Telemedicine, Moscow, https://orcid.org/0000-0002-3193-8320.
E-mail: NikitinNY@zdrav.mos.ru.

Anton V. Vladzimirskyy, Dr. Sci. (Med.), Deputy Director for Research, Moscow Center for Diagnostics and Telemedicine; Professor, Information and Internet Technology Department, I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, https://orcid.org/0000-0002-2990-7736.
E-mail: VladzimirskijAV@zdrav.mos.ru.

Olga V. Omelyanskaya, Head of Division Management of the Directorate of Science, Moscow Center for Diagnostics and Telemedicine, Moscow, https://orcid.org/0000-0002-0245-4431.
E-mail: OmelyanskayaOV@zdrav.mos.ru.

Sergey F. Chetverikov, Cand. Sci. (Tech.), Head of the Sector of System Development for the Introduction of Medical Intelligent Technologies, Department of Medical Informatics, Radiomics and Radiogenomics, Moscow Center for Diagnostics and Telemedicine, Moscow, https://orcid.org/0000-0002-3097-8881.
E-mail: tschetserg@yandex.ru.

Kirill M. Arzamasov, Cand. Sci. (Med.), Head of the Department of Medical Informatics, Radiomics and Radiogenomics, Moscow Center for Diagnostics and Telemedicine; Associated Professor, Department of Artificial Technology, MIREA – Russian Technological University, Moscow, https://orcid.org/0000-0001-7786-0349.
E-mail: ArzamasovKM@zdrav.mos.ru.

✉ **Tatiana M. Bobrovskaya,** e-mail: BobrovskayaTM@zdrav.mos.ru.

Received 02.02.2024;
review received 02.04.2024;
accepted for publication 13.05.2024.