DOI: https://doi.org/10.17816/DD625967

Evaluating the performance of artificial intelligence-based software for digital mammography characterization

Yuri A. Vasilev[1,2], Alexander V. Kolsanov[3], Kirill M. Arzamasov[1], Anton V. Vladzymyrskyy[1,4], Olga V. Omelyanskaya[1], Serafim S. Semenov[1], Lubov E. Axenova[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] National Medical and Surgical Center named after N.I. Pirogov, Moscow, Russia;
[3] Samara State Medical University, Samara, Russia;
[4] Sechenov First Moscow State Medical University, Moscow, Russia

ABSTRACT

BACKGROUND: Digital screening mammography is a key modality for early detection of breast cancer, reducing mortality by 20–40%. Many artificial intelligence (AI)-based services have been developed to automate the analysis of imaging data.

AIM: The aim of the study was to compare mammography assessments using three types of AI services in multiple versions with radiologists' conclusions.

MATERIALS AND METHODS: Binary mammography scoring scales were compared with several types and versions of AI services regarding diagnostic accuracy, Matthews correlation coefficient, and maximum Youden's index.

RESULTS: A comparative analysis showed that the use of a binary scale for evaluating digital mammography affects the number of detected abnormalities and accuracy of AI results. In addition, diagnostic accuracy was found to be threshold dependent. AI Service 1 in version 3 had the best performance, as confirmed by most diagnostic accuracy parameters.

CONCLUSION: Our results can be used to select AI services for interpreting mammography screening data. Using Youden's index maximization to set up an AI service provides a balance of sensitivity and specificity that is not always clinically relevant.

Keywords: malignant tumors of breast; digital mammography; artificial intelligence services; diagnostic accuracy; Youden's index.

To cite this article:
Vasilev YA, Kolsanov AV, Arzamasov KM, Vladzymyrskyy AV, Omelyanskaya OV, Semenov SS, Axenova LE. Evaluating the performance of artificial intelligence-based software for digital mammography characterization. *Digital Diagnostics.* 2024;5(4):695–711. DOI: https://doi.org/10.17816/DD625967

Received: 24.01.2024 **Accepted:** 10.10.2024 **Published online:** 02.12.2024

DOI: https://doi.org/10.17816/DD625967

Оценка производительности программного обеспечения на основе технологии искусственного интеллекта при описании цифровых маммографических исследований

Ю.А. Васильев[1,2], А.В. Колсанов[3], К.М. Арзамасов[1], А.В. Владзимирский[1,4], О.В. Омелянская[1], С.С. Семёнов[1], Л.Е. Аксёнова[1]

[1] Научно-практический клинический центр диагностики и телемедицинских технологий, Москва, Россия;

[2] Национальный медико-хирургический Центр имени Н.И. Пирогова, Москва, Россия;

[3] Самарский государственный медицинский университет, Самара, Россия;

[4] Первый Московский государственный медицинский университет имени И.М. Сеченова, Москва, Россия

АННОТАЦИЯ

Обоснование. Цифровая скрининговая маммография — это основной инструмент для раннего выявления злокачественных новообразований молочной железы, позволяющий снизить смертность на 20–40%. На сегодняшний день разработано множество сервисов на основе искусственного интеллекта (ИИ), позволяющих автоматизировать анализ таких исследований.

Цель — сравнить результаты оценки цифровых маммографических исследований, выполненной тремя типами ИИ-сервисов в нескольких версиях, с заключениями врачей-рентгенологов.

Материалы и методы. Проведено сравнение бинарных шкал оценки маммографических исследований и нескольких типов и версий ИИ-сервисов по показателям диагностической точности, коэффициенту Мэтьюса и максимальному индексу Юдена.

Результаты. Сравнительный анализ показал, что выбор бинарной шкалы для оценки цифрового маммографического исследования влияет на количество выявляемых случаев патологии и точность результатов ИИ-сервисов. Кроме того, обнаружена зависимость показателей диагностической точности от порогового значения. Наилучшей производительностью обладает ИИ-сервис 1 в версии 3, что подтверждается большинством показателей диагностической точности.

Заключение. Полученные нами результаты могут быть полезны при выборе ИИ-сервисов для интерпретации данных скрининговой маммографии. Настройка ИИ-сервиса методом максимизации индекса Юдена позволяет получать сбалансированные значения чувствительности и специфичности, что не всегда целесообразно с клинической точки зрения.

Ключевые слова: злокачественные новообразования молочной железы; цифровая маммография; сервисы искусственного интеллекта; показатели диагностической точности; индекс Юдена.

Как цитировать:

Васильев Ю.А., Колсанов А.В., Арзамасов К.М., Владзимирский А.В., Омелянская О.В., Семёнов С.С., Аксёнова Л.Е. Оценка производительности программного обеспечения на основе технологии искусственного интеллекта при описании цифровых маммографических исследований // Digital Diagnostics. 2024. Т. 5, № 4. С. 695–711. DOI: https://doi.org/10.17816/DD625967

Рукопись получена: 24.01.2024 **Рукопись одобрена:** 10.10.2024 **Опубликована online:** 02.12.2024

DOI: https://doi.org/10.17816/DD625967

基于人工智能技术的软件在描述数字乳房造影检查中的性能评估

Yuri A. Vasilev[1,2], Alexander V. Kolsanov[3], Kirill M. Arzamasov[1], Anton V. Vladzymyrskyy[1,4], Olga V. Omelyanskaya[1], Serafim S. Semenov[1], Lubov E. Axenova[1]

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies, Moscow, Russia;
[2] National Medical and Surgical Center named after N.I. Pirogov, Moscow, Russia;
[3] Samara State Medical University, Samara, Russia;
[4] Sechenov First Moscow State Medical University, Moscow, Russia

摘要

论证。数字乳房造影筛查是早期发现乳腺恶性肿瘤的主要工具，可将死亡率降低20~40%。目前，已开发出许多基于人工智能（AI）的服务来自动分析此类检查。

目的 — 比较三种人工智能服务在不同版本中进行的乳房造影检查评估结果与放射科医生的意见。

材料和方法。比较了乳房造影检查二元评估量表与多种类型和版本的AI服务在诊断准确性指标、马修斯系数和最大尤登指数等方面的差异。

结果。比较分析表明，评估数字乳房造影检查的二元评估量表的选择会影响检测到的病理病例数量和AI服务结果的准确性。此外，还发现了诊断准确性指标对阈值的依赖性。版本3中的AI服务1实现了最佳性能，大多数诊断准确性指标都证实了这一点。

结论。我们的研究结果可能有助于选择AI服务来解读乳房造影筛查数据。通过最大化尤登指数来设置AI服务，可以获得灵敏度和特异性的平衡值，但从临床角度来说，并不总是合理的。

关键词：乳腺恶性肿瘤；数字乳房造影；人工智能服务；诊断准确性指标；尤登指数。

引用本文:
Vasilev YA, Kolsanov AV, Arzamasov KM, Vladzymyrskyy AV, Omelyanskaya OV, Semenov SS, Axenova LE. 基于人工智能技术的软件在描述数字乳房造影检查中的性能评估. *Digital Diagnostics.* 2024;5(4):695–711. DOI: https://doi.org/10.17816/DD625967

收到: 24.01.2024 **接受:** 10.10.2024 **发布日期:** 02.12.2024

BACKGROUND

In X-ray radiography, digital mammography is the primary diagnostic tool and the sole method of breast cancer screening. Screening lowers cancer mortality by 20%–40% by enabling much earlier diagnosis of cancer-associated abnormal mammary gland changes [1]. As artificial intelligence (AI) evolves, novel AI-based systems and services are being introduced for the automated analysis of digital mammography images [2–4]. Certain studies indicate that AI services facilitate reliable diagnosis and can even outperform radiologists. This has especially been demonstrated in the detection of breast cancer signs at early stages and/or when fibroglandular breast tissue is the predominant site of abnormality. Other studies, however, indicate that radiologists are still able to interpret mammograms more accurately than AI systems[5]. Machine learning models constitute the core functional elements of AI services that are responsible for the detection and segmentation of areas of interest with anomalous changes, data processing and classification, and generating predictions or solutions based on these data. To compare machine learning models, diagnostic accuracy parameters such as sensitivity (Sens) and specificity (Spec) are computed, and the area under the curve (AUC) is analyzed [6, 7].

A true value must be chosen and contrasted with the outcomes of the AI service in order to evaluate the AI performance. Calculations are typically performed relative to the model's output and the gold standard, which is based on the findings of additional studies [8, 9]. Moreover, AI outcomes can be evaluated by comparing them with a physician's opinion [10, 11]. The ability to fine-tune AI systems is their primary benefit. However, the accuracy testing of software that produces probabilistic data instead of conventional binary data is critical for the deployment and use of AI services in medicine.

A threshold that differentiates the probabilities deemed "abnormal" from those considered "normal" must be established in order to interpret the probabilistic data. The optimal probability threshold depends on the purpose and intended application of the AI service. Since probability distributions for imbalanced data tend to shift toward the "normal" category [12], setting a threshold at 0.5 may not be the optimal option. Effective cancer detection and a reduction in false positives require a balance between a machine learning model's sensitivity and specificity. Using Youden's index to maximize the sum of the Sens and Spec values is a commonly employed technique [7]. Additionally, Chen et al. proposed a technique for comparing the maximum Youden's index values for several diagnostic tests [13]. Given that the subpar performance of AI systems in medical diagnosis can be linked to serious risks, a comprehensive assessment of the capabilities and operational constraints of such AI systems is required.

AIM

This study aimed to evaluate how various AI service versions interpret digital mammography findings in comparison to the conclusions of radiologists.

MATERIALS AND METHODS

Study design

This was a multicenter, observational, cross-sectional study. Fig. 1 illustrates the dataset production chart for the analysis as well as the study design.

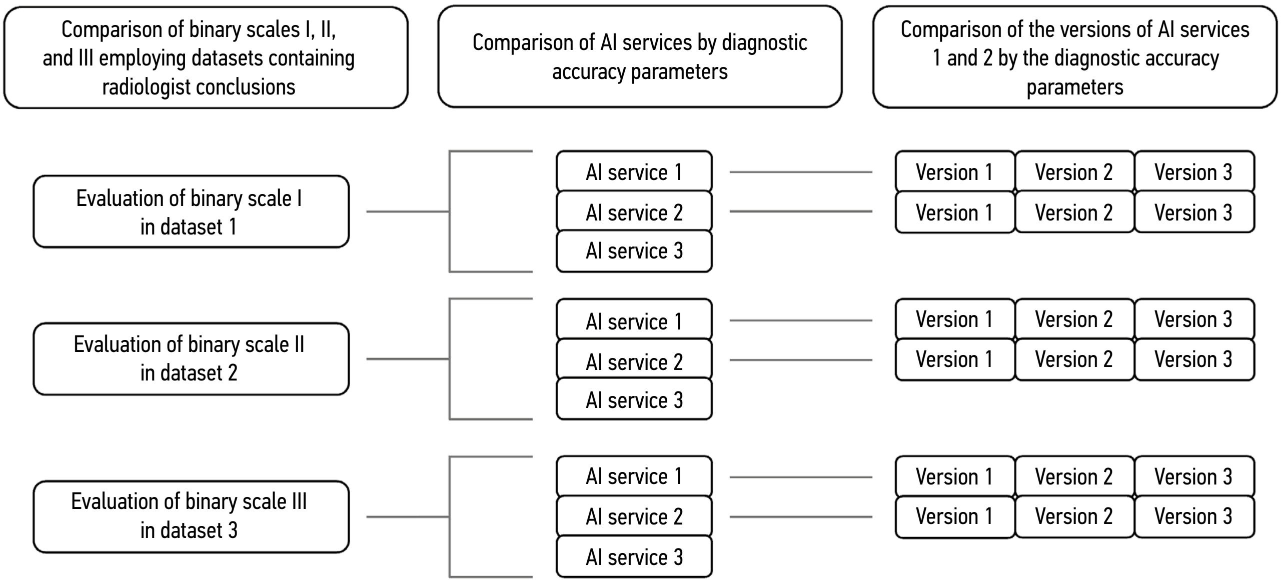
Comparison of binary scales I, II, and III employing datasets containing radiologist conclusions
Comparison of AI services by diagnostic accuracy parameters
Comparison of the versions of AI services 1 and 2 by the diagnostic accuracy parameters
Evaluation of binary scale I in dataset 1
Evaluation of binary scale II in dataset 2
Evaluation of binary scale III in dataset 3
AI service 1
AI service 2
AI service 3
Version 1
Version 2
Version 3

Fig. 1. Study design and generation of datasets for analysis. AI service, artificial intelligence service.

DOI: https://doi.org/10.17816/DD625967

Eligibility criteria

Inclusion criteria. Female patients (independent of age or comorbidities) who received digital mammograms performed between July 22, 2020, and December 29, 2022, with DICOM pictures and other data available for analysis using an AI service, were included in the study.

Exclusion criteria:

1. Insufficient data in the medical records for at least one evaluated AI service to process;
2. Technically flawed images that impede correct interpretation (e.g., artifacts, partially missing data);
3. Incompleteness of the metadata essential for the analysis.

Additional information. Examinations in patients with breast implants and those who had received radiotherapy were not categorized into separate subgroups, and their numbers in the sample were not reported.

Study setting

The study sample included examination results from 123 outpatient healthcare facilities of the Moscow Healthcare Department. The study comprised 531 radiologists who specialized in mammography. All participating radiologists analyzed the examinations performed at the healthcare facilities of the Moscow Healthcare Department. The results of the AI service were compared to the true value obtained from a radiologist's conclusion for each examination. Each radiologist reported an average of 1,250 examinations during the study period.

Data generation and analysis

The accuracy of the AI service results was evaluated using radiologist opinions from medical records as reference values. The conclusions were presented in accordance with categories 1–6 of the Breast Imaging Reporting and Data System (BI-RADS), independently for each mammary gland. Three binary diagnostic scales were utilized based on the probability of cancer according to BI-RADS: Scale I classified BI-RADS categories 1–2 as "normal" and categories 3–6 as "abnormal"; Scale II classified BI-RADS categories 1–3 as "normal" and categories 4–6 as "abnormal"; and Scale III classified BI-RADS categories 1–2 as "normal" and categories 4–6 as "abnormal" (BI-RADS category 3 was not considered in this case).

The study assessed three AI services: TrioDM-MT® (Medical Technologies Ltd., Russia) (AUC 0.90; specificity 0.85; sensitivity 0.83; accuracy 0.84); Celsus® (Medical Screening Systems LLC, Russia) (AUC 0.96); and Lunit INSIGHT MMG® (Lunit Inc., Republic of Korea) (AUC 0.96; sensitivity 0.89 when assessed together with a radiologist) [2–4]. The AI service results for every mammography test were displayed as probabilities, with 0% denoting a low likelihood of cancer and 100% denoting a high probability of cancer. The trade names of the AI services are anonymized and randomized further in the text.

During data preprocessing, lines lacking radiologist descriptions and/or AI service results were excluded. Moreover, test results from male patients, female patients aged <40 years or >100 years, and examinations in which the radiologist's conclusion did not correspond to BI-RADS categories 1–6 or any AI service mentioned above were excluded from the dataset.

After performing data preprocessing for each mammography examination, diagnostic accuracy parameters were determined, including AUC, sensitivity (Sens), specificity (Spec), accuracy (Acc), positive predictive value (PPV), false negative rate (FNR), case detection rate (CDR), abnormal interpretation rate (AIR), Matthews correlation coefficient (MCC), and Youden's index (J). Table 1 presents a description of each parameter, along with diagnostic scales where the maximum values of these parameters were observed.

Table 1. Descriptions of diagnostic accuracy parameters and diagnostic scales with the highest values of these parameters

Parameter	Description	Diagnostic scale
AUC	Area Under the Curve: represents the ability to differentiate between classes; is not sensitive to class imbalance	II and III
Sens	Sensitivity: represents the ability to detect the "abnormal" class	III
Spec	Specificity: represents the ability to detect the "normal" class	II
Acc	Accuracy: represents the proportion of correctly classified objects in the total number of objects in the sample; sensitive to class imbalance	II
PPV	Positive Predictive Value: represents the consistency of the detected "abnormal" class with a true abnormality	I
AIR	Abnormal Interpretation Rate: the proportion of examinations classified as "abnormal" and requiring additional diagnostic procedures; represents the highest number of false positives	I
CDR	Case Detection Rate: represents the detection of abnormalities irrespective of the total number of false positives	I
FNR	False Negative Rate: represents the number of "abnormal" cases not detected by an AI service	I
MCC	Matthews Correlation Coefficient: evaluates the quality of classification, considering all four elements of the error matrix; not sensitive to class imbalance	I
J	Youden's index	-

AI services were updated during the study, including fine-tuning and other modifications. Every update matched a change in the version of the AI service. The study only addressed the modifications made to the AI service core that influenced the diagnostic accuracy parameters. Consequently, three iterations of AI service 1 and AI service 2 were identified, each of which represented a subsequent model modification and was used during different periods. Since AI service 3 had not undergone any major changes, different versions were not considered.

The optimal probability threshold was ascertained using the AUC and the maximum Youden's index value. The calculations were performed employing a web tool developed by the Center for Diagnostics and Telemedicine (Moscow)[1]. Youden's index was calculated using the following formula:

$$J = Sens - Spec - 1 \tag{1}$$

where Sens = sensitivity; Spec = specificity.

The binary AI service results were calculated using the threshold. The outcomes of the AI service were then compared with the radiologist's judgments using the following parameters:

- TP, the number of true positives;
- TN, the number of true negatives;
- FP, the number of false positives;
- FN, the number of false negatives.

The resulting TP, TN, FP, and FN values were utilized to calculate the following AI service accuracy parameters (Table 1) [14]:

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i) * (y_i + y_{i+1}) \tag{2}$$

where x = X-axis values (e.g., false positives); y = Y-axis values (e.g., true positives); n = total number of points on a curve; and i = current point index.

$$Sens = \frac{TP}{TP + FN} \tag{3}$$

$$Sens = \frac{TP}{TP + FN} \tag{4}$$

$$Acc = \frac{TP + TN}{FP + FN + TP + TN} \tag{5}$$

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

$$AIR = \frac{TP + FP}{TP + TN + FP + FN} \times 1000 \tag{7}$$

$$CDR = \frac{TP}{TP + TN + FP + FN} \times 1000 \tag{8}$$

$$FNR = \frac{FN}{FN + TP} \tag{9}$$

$$MCC = \frac{TP \times TN + FP + FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

To model the ratios calculated in datasets 1–3, 100 data samples (1,000 samples each) with a category 0 ("normal") to category 1 ("abnormal") ratio of 9:1 (Scale I), 33:1 (Scale II), and 31:1 (Scale III) were created, and the accuracy parameters and confidence intervals were calculated using bootstrapping.

Ethical review

This srudy was part of the Experiment on the Use of Innovative Computer Vision Technologies for Analysis of Medical Images in the Moscow Healthcare System (Moscow Experiment) (Protocol No. NCT04489992 of February 21, 2020) previously approved by the local ethics committee.

Statistical analysis

The study compared the accuracy of breast cancer detection for three binary scales based on radiologist's conclusions and three AI services. The nonparametric Kolmogorov-Smirnov test was used to determine if the resulting datasets were normally distributed.

The significance of the differences between the maximum Youden's index values for various types and versions of AI services was assessed in accordance with the method outlined by Chen et al. [13]. The variance (Var) of the difference

[1] S.P. Morozov, A.E. Andreychenko, S.F. Chetverikov, et al. Certificate of state registration of computer program No. 2022617324 Russian Federation. Web tool for performing ROC analysis of diagnostic test results: No. 2022616046: declared 04/05/2022: published 04/19/2022. Access mode: https://roc-analysis.mosmed.ai/ Date of access: 08/20/2023 EDN: ECMPNH

between two independent Youden's index values was ascertained using the following formula:

$$Var\left(J_1 - J_2\right) = Var\left(J_1\right) + Var\left(J_2\right) \tag{11}$$

where J = Youden's index; Var is calculated using the following formula:

$$Var\left(J\right) = Spec^2 \times Var\left(Sens\right) + Sens^2 \times Var\left(Spec\right) \tag{12}$$

where Spec = specificity; Sens = sensitivity.

Thus, the formula is as follows:

$$Var\left(J_1 - J_2\right) = Spec_1^2 \times Var\left(Sens_1\right) + Sens_1^2 \times \times Var\left(Spec_1\right) + Spec_2^2 \times Var\left(Sens_2\right) + Sens_2^2 \times \times Var\left(Spec_2\right) \tag{13}$$

The statistical test and two-sided confidence interval for the difference between two independent Youden's index (J) values were computed based on the central limit theorem:

$$Z = \frac{J_1 - J_2}{\sigma_{D_J}} = \frac{J_1 - J_2}{\sqrt{Var\left(J_1 - J_2\right)}} \tag{14}$$

$$d \pm Z_{\alpha/2} \times \sigma_{D_J} = d \pm Z_{\alpha/2} \times \sqrt{Var\left(J_1 - J_2\right)} \tag{15}$$

where Z = standard normal random value that represents the variation of the difference from zero, in standard deviations; Var = variance; d = difference between two Youden's index values; σ_{D_J} = standard deviation of the difference between Youden's index values.

P-values < 0.05 were considered statistically significant. The confidence interval was determined to be 95%. Calculations were performed using the Python libraries Pandas, Matplotlib, Seaborn, Scikit-learn, NumPy, and Statistics (stats) (Python Software Foundation, version 3.11.0).

RESULTS

Comparison of binary diagnostic scales based on radiologist conclusions

Based on a radiologist's evaluation of the normal distribution of BI-RADS categories 1–6, the distribution of these parameters was considered to be nonnormal. The distribution histograms for the BI-RADS categories are illustrated in Fig. 2. Peaks in the graph correspond to the most probable categories. In this case, the greatest peak corresponds to BI-RADS category 2 (benign), indicating that most examinations included in the sample did not reveal any aberrant cancer-associated changes.

Datasets 1 and 2 included 663,606 examinations, while dataset 3 comprised 618,947 examinations. The number of "normal" cases in datasets 1, 2, and 3 was 64,100, 19,441, and 19,441, respectively, while the number of "abnormal" cases was 599,506, 644,165, and 599,506, respectively. Thus, the incidence of cancer in the evaluated data sample was 9.66% for binary scale I and 2.9% for binary scales II and III (Fig. 3). Comprehensive details about the datasets are presented in Table 2 and Table 3.

To evaluate the agreement between radiologist decisions and AI service outcomes, diagnostic accuracy measures were calculated (Table 1 and Table 4). AUC for Scale I significantly

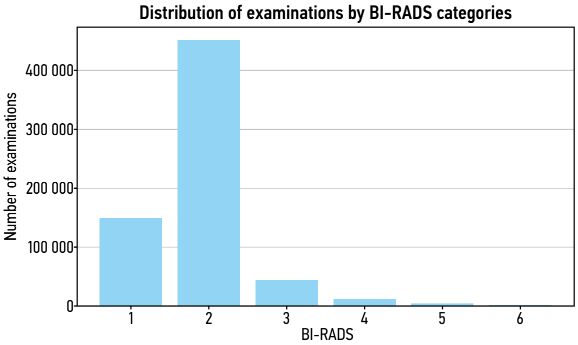

Fig. 2. Distribution of BI-RADS categories 1–6 allocated by a radiologist when analyzing digital mammography findings for the assessed datasets: X-axis values = BI-RADS categories 1–6; Y-axis values = number of examinations.

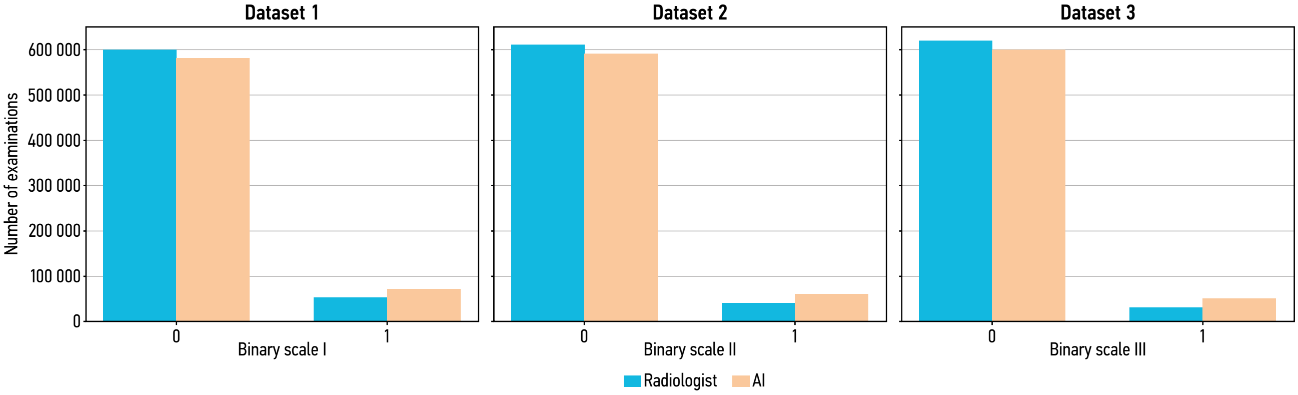

Fig. 3. Comparison of the distribution of categories 0–1 allocated by radiologists and an artificial intelligence service for three binary scales: X-axis values = binary scales I–III; Y-axis values = number of examinations.

DOI: https://doi.org/10.17816/DD625967

Table 2. Number of normal and abnormal cases in datasets 1–3

	“Normal”	“Abnormal”	All examinations	Amount of “normal” per “abnormal” case
Scale I	599,506	64,100	663,606	9
Scale II	644,165	19,441	663,606	33
Scale III	599,506	19,441	618,947	31

Table 3. Number of examinations in the datasets (2020–2022)

Scales	I–II							III					
Number of examinations	663,606							618,947					
Services	1			2			3	1			2		
Number of examinations	545,362			108,763			9481	508,929			101,654		
Versions	1	2	3	1	2	3	-	1	2	3	1	2	3
Number of examinations	90,949	212,968	241,445	4922	46,851	56,990	-	83,828	198,231	226,870	4711	43,687	53,256

Table 4. Artificial intelligence services and their versions with the highest accuracy parameters compared to diagnostic scales

Parameter	Scale	AI service	AI service 1 version number	AI service 2 version number
AUC	I	1	3	1 and 3
	II	1	3	1 and 2
	III	1	3	1 and 2
Sens	I	1 and 2	2 and 3	2 and 3
	II	1 and 2	3	2
	III	1 and 2	3	2
Spec	I	1	3	1
	II	1	1 and 3	1
	III	1	1	1
Acc	I	1	3	1
	II	1	3	1
	III	1	1	1
PPV	I	1	3	1
	II	1	3	1
	III	1	2 and 3	1
AIR	I	3	1	2
	II	3	2	2
	III	3	3	2
CDR	I	1 and 2	2 and 3	2 and 3
	II	1 and 3	3	2
	III	1 and 3	3	2
FNR	I	3	1	1
	II	2	1	1 and 3
	III	2	1	1 and 3
MCC	I	1	3	1
	II	1	3	1
	III	1	3	1
Youden's index	I	1	3	1
	II	1	3	2
	III	1	3	2

Note. AI service, artificial intelligence service.

differed from that for Scales II and III (there were no differences in AUC between the latter). Moreover, Scale III had a higher sensitivity (Sens), whereas Scale II demonstrated a higher specificity (Spec). Scale I displayed the highest abnormal interpretation rate (AIR) and false negative rate (FNR), whereas Scale II exhibited the lowest AIR and FNR. Scale I demonstrated the highest consistency level measured using MCC, as well as the highest PPV and CDR (Table 5).

Comparison of AI services with each other and with scales based on radiologist conclusions

Fig. 4 displays the probability distributions of anomalous changes for AI services 1–3. Scales II and III exhibited the most comparable distribution of the AI service probabilities. The distribution for the “normal” category demonstrated a right shift, particularly for AI services 2 and 3. The distribution for the “abnormal” category exhibited a left shift for AI services 1, 2, and 3.

The performance of AI services 1, 2, and 3 was evaluated and compared using the same diagnostic accuracy metrics. AI service 1 was the most consistent with the radiologist’s conclusions in determining the “normal” and “abnormal” categories, whereas the highest abnormal interpretation rate (AIR) and false negative rate (FNR) were noted for AI services 2 and 3 (Table 6).

Comparison of AI service versions with each other

Additionally, diagnostic accuracy metrics were measured to evaluate the different versions of AI services 1 and 2 (Table 7 and Table 8). The majority of the diagnostic accuracy parameters varied depending on the scale; however, some of these variations were nonsignificant. Consequently, the optimal version of AI services cannot be identified.

Table 5. Diagnostic accuracy parameters evaluated for the artificial intelligence service results (AI service 1)

Parameter	Binary diagnostic scale		
	I	II	III
Threshold	62	74	68
AUC	0.659 [0.654; 0.663]	0.726 [0.717; 0.735] *	0.738 [0.730; 0.745] *
Sens	0.569 [0.560; 0.578]	0.626 [0.609; 0.644]	0.679 [0.664; 0.694] *
Spec	0.748 [0.746; 0.751]	0.826 [0.823; 0.828] *	0.796 [0.793; 0.798]
Acc	0.730 [0.728; 0.733]	0.820 [0.817; 0.822] *	0.792 [0.790; 0.795]
PPV	0.201 [0.198; 0.204] *	0.100 [0.098; 0.103]	0.099 [0.097; 0.101]
AIR	283.280 [280.57; 285.988] *	187.810 [185.199; 190.421]	219.24 [216.787; 221.693]
CDR	56.870 [55.960; 57.780] *	18.790 [18.273; 19.307]	21.730 [21.252; 22.208]
FNR	0.431 [0.422; 0.440] *	0.374 [0.356; 0.391]	0.321 [0.306; 0.336]
MCC	0.211 [0.205; 0.217] *	0.198 [0.190; 0.205]	0.202 [0.196; 0.209]

Note. The data are presented as means [95% CI]; *, significant differences between Scales I and II, I and III, and II and III (CIs do not overlap).

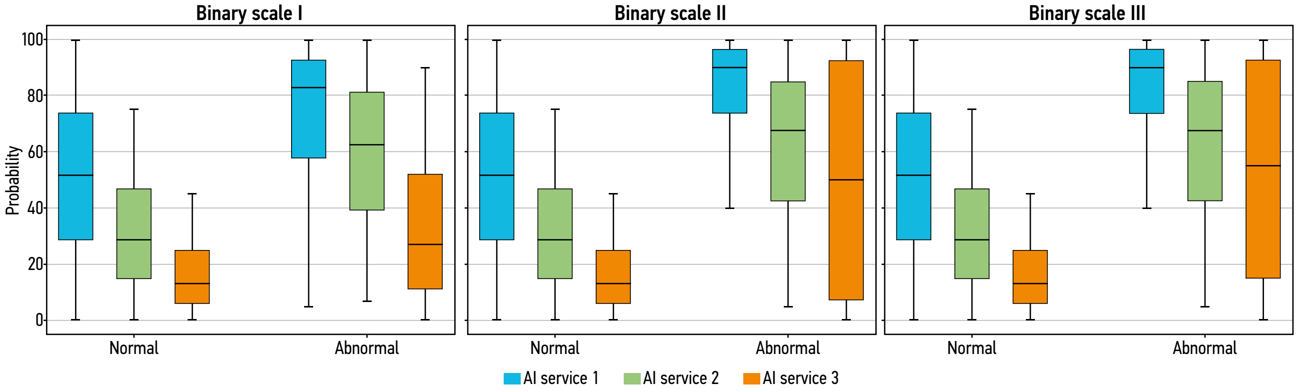

Fig. 4. Distribution of the outcomes of three artificial intelligence services when analyzing the three datasets: X-axis values = artificial intelligence services; Y-axis values = probability; the data are presented as follows: central line = median; edges of the "box" = first (Q1) and third (Q3) quartiles; "whiskers" = minimum and maximum.

Table 2 presents the assessment results for the scales, services, and versions with the highest diagnostic accuracy parameters.

In addition to the diagnostic accuracy parameters, the maximum Youden's index value was employed to contrast the performance of different AI service types and versions. This parameter assesses how well an AI service balances sensitivity and specificity. When assessed using the highest Youden's index value, AI service 1 performed best (see Table 2). When comparing the versions of AI service 1, version 3 exhibited the best diagnostic accuracy parameters. However, in terms of Youden's index, version 1 of AI service 1 had the best results for Scale I, whereas version 2 performed best for Scales II and III. Significant differences were observed for all results.

DISCUSSION

The study examined the binary diagnostic scales for digital mammography, three AI services, and three versions of AI services 1 and 2. The standard diagnostic accuracy parameters and Youden's index were ascertained.

The ratio of detected cancer cases to the total number of examinations performed was 0.10, 0.03, and 0.03 for Scales I, II, and III, respectively. Because the risk of undiagnosed abnormalities must be taken into account, variations in diagnostic scales are essential for screening. For example, BI-RADS category 3 necessitates an additional examination; based on it, certain individuals may be placed in categories with a higher level of malignancy. Thus, employing Scale II, which classifies BI-RADS category 3 as class 1 ("abnormal"), lowers the probability of undetected abnormalities.

Notably, when using Scale I, the incidence of abnormalities was neither influenced by the presence of BI-RADS category 3 in the "abnormal" group nor its complete absence from the dataset. However, including this category in the "normal" group significantly elevated the estimated cancer incidence in the screened population. Moreover, in BI-RADS category 3, Scale I exhibits a similar pattern for AUC, with Scales II and III exhibiting no significant variations.

When comparing the three AI services on the basis of diagnostic accuracy parameters, it was discovered that AUC, specificity, sensitivity, accuracy, PPV, CDR, and MCC

Table 6. Diagnostic accuracy parameters evaluated for the results of three artificial intelligence services compared to radiologist conclusions

Parameter	Binary scale	AI service 1	AI service 2	AI service 3
Threshold	I	64	32	10
	II	75	44	20
	III	74	44	20
AUC	I	0.671 [0.666; 0.676] *	0.647 [0.641; 0.652]	0.597 [0.592; 0.602]
	II	0.750 [0.740; 0.759] *	0.698 [0.689; 0.708]	0.713 [0.704; 0.722]
	III	0.755 [0.746; 0.763] *	0.708 [0.699; 0.717]	0.720 [0.713; 0.728]
Sens	I	0.610 [0.600; 0.620] *	0.602 [0.591; 0.613] *	0.578 [0.569; 0.587]
	II	0.687 [0.669; 0.706] *	0.609 [0.591; 0.627]	0.676 [0.658; 0.693] *
	III	0.689 [0.672; 0.705] *	0.616 [0.598; 0.633]	0.679 [0.664; 0.693] *
Spec	I	0.731 [0.728; 0.734] *	0.691 [0.688; 0.694]	0.616 [0.613; 0.619]
	II	0.812 [0.810; 0.815] *	0.788 [0.785; 0.790]	0.749 [0.746; 0.752]
	III	0.821 [0.819; 0.824] *	0.800 [0.798; 0.803]	0.762 [0.759; 0.765]
Acc	I	0.719 [0.717; 0.722] *	0.682 [0.679; 0.685]	0.612 [0.609; 0.615]
	II	0.809 [0.806; 0.811] *	0.782 [0.779; 0.785]	0.747 [0.744; 0.750]
	III	0.817 [0.815; 0.819] *	0.794 [0.792; 0.797]	0.760 [0.757; 0.762]
PPV	I	0.202 [0.199; 0.205] *	0.178 [0.175; 0.181]	0.143 [0.141; 0.145]
	II	0.102 [0.099; 0.105] *	0.082 [0.079; 0.084]	0.077 [0.075; 0.079]
	III	0.113 [0.110; 0.116] *	0.093 [0.090; 0.095]	0.086 [0.084; 0.088]
AIR	I	302.800 [299.851; 305.749]	338.140 [335.251; 341.029]	403.640 [400.720; 406.560] *
	II	202.600 [200.315; 204.885]	224.330 [221.624; 227.036]	263.320 [260.371; 266.269] *
	III	195.130 [192.823; 197.437]	212.940 [210.594; 215.286]	251.890 [249.124; 254.656] *
CDR	I	61.040 [60.041; 62.039] *	60.210 [59.113; 61.307] *	57.780 [56.889; 58.671]
	II	20.620 [20.062; 21.178] *	18.280 [17.735; 18.825]	20.270 [19.750; 20.790] *
	III	22.040 [21.507; 22.573] *	19.700 [19.131; 20.269]	21.720 [21.260; 22.180] *
FNR	I	0.390 [0.380; 0.400]	0.398 [0.387; 0.409]	0.422 [0.413; 0.431] *
	II	0.313 [0.294; 0.331]	0.391 [0.373; 0.409] *	0.324 [0.307; 0.342]
	III	0.311 [0.295; 0.328]	0.384 [0.367; 0.402] *	0.321 [0.307; 0.336]
MCC	I	0.223 [0.217; 0.230] *	0.186 [0.179; 0.193]	0.118 [0.113; 0.124]
	II	0.212 [0.204; 0.220] *	0.163 [0.155; 0.170]	0.165 [0.158; 0.172]
	III	0.227 [0.219; 0.234] *	0.179 [0.171; 0.187]	0.179 [0.173; 0.185]

Note. AI service, artificial intelligence service; the data are presented as means [95% CI]; *, significant differences between services 1 and 2, 1 and 3, and 2 and 3 (CIs do not overlap).

were greater for AI service 1 than for all diagnostic scales. AI service 3 demonstrated the highest AIR, whereas AI services 2 and 3 exhibited the highest FNR, depending on the binary scale. In general, these results indicate the superior performance of AI service 1. When comparing the maximum Youden's index values, AI service 1 demonstrated superior accuracy to all the diagnostic scales. However, while the assessment using Youden's index revealed significant differences between all scales and services, the assessment using bootstrap confidence intervals exhibited no significant differences in CDR and sensitivity between certain services.

Similar to binary scales, when selecting AI services and their versions, it is crucial to consider their intended use. For example, when the early detection of radiological signs of cancer is required, sensitivity (Sens) will be the primary diagnostic accuracy parameter because an AI service needs to identify as many true positive (abnormal) cases as feasible. Another crucial parameter is the FNR, which should be minimized to lower the possibility of undetected abnormalities. This study indicates that AI services 1 and 2 are the most sensitive. However, AI service 2 displayed the highest FNR for Scales II and III. Versions 3 and 2 of AI services 1 and 2, respectively, performed best in this study.

Table 7. Diagnostic accuracy parameters assessed for the results of three versions of artificial intelligence service 1 compared to radiologist conclusions

Scale	AI service	AI service version	Threshold	AUC	Sens	Spec	Acc
I	1	1	29	0.633 [0.627; 0.638]	0.617 [0.607; 0.627]	0.648 [0.645; 0.652]	0.645 [0.642; 0.648]
		2	68	0.674 [0.669; 0.679]	0.643 [0.634; 0.653] *	0.705 [0.702; 0.708]	0.699 [0.696; 0.702]
		3	66	0.702 [0.697; 0.706] *	0.647 [0.638; 0.656] *	0.757 [0.754; 0.760] *	0.746 [0.743; 0.748] *
II	1	1	57	0.684 [0.674; 0.693]	0.525 [0.506; 0.545]	0.842 [0.840; 0.844] *	0.833 [0.830; 0.835]
		2	78	0.750 [0.741; 0.760]	0.683 [0.664; 0.702]	0.817 [0.815; 0.819]	0.813 [0.811; 0.815]
		3	79	0.782 [0.774; 0.791] *	0.725 [0.708; 0.742] *	0.840 [0.837; 0.842] *	0.836 [0.834; 0.839] *
III	1	1	57	0.701 [0.692; 0.711]	0.548 [0.530; 0.566]	0.855 [0.853; 0.857] *	0.845 [0.843; 0.847] *
		2	78	0.756 [0.746; 0.766]	0.678 [0.659; 0.698]	0.833 [0.831; 0.836]	0.828 [0.826; 0.831]
		3	75	0.789 [0.783; 0.796] *	0.758 [0.745; 0.770] *	0.821 [0.819; 0.823]	0.819 [0.817; 0.821]

Scale	AI service	AI service version	PPV (95% CI)	AIR (95% CI)	CDR (95% CI)	FNR (95% CI)	MCC (95% CI)
I	1	1	0.163 [0.161; 0.166]	378.070 [375.137; 381.003] *	61.690 [60.650; 62.730]	0.383 [0.373; 0.393] *	0.164 [0.157; 0.171]
		2	0.195 [0.193; 0.198]	329.660 [326.815; 332.505]	64.340 [63.384; 65.296] *	0.357 [0.347; 0.366]	0.223 [0.216; 0.229]
		3	0.228 [0.225; 0.231] *	283.71 [280.851; 286.569]	64.670 [63.768; 65.572] *	0.353 [0.344; 0.362]	0.269 [0.263; 0.275] *
II	1	1	0.093 [0.090; 0.097]	169.010 [166.841; 171.179]	15.760 [15.175; 16.345]	0.475 [0.455; 0.494] *	0.167 [0.158; 0.176]
		2	0.104 [0.101; 0.106]	198.140 [195.926; 200.354] *	20.500 [19.931; 21.069]	0.317 [0.298; 0.336]	0.214 [0.206; 0.222]
		3	0.123 [0.120; 0.126] *	177.240 [174.831; 179.649]	21.750 [21.234; 22.266] *	0.275 [0.258; 0.292]	0.253 [0.245; 0.260] *
III	1	1	0.111 [0.108; 0.115]	158.040 [155.998; 160.082]	17.540 [16.969; 18.111]	0.452 [0.434; 0.470] *	0.195 [0.186; 0.203]
		2	0.119 [0.115; 0.122] *	183.040 [180.595; 185.485]	21.710 [21.078; 22.342]	0.322 [0.302; 0.341]	0.233 [0.224; 0.242]
		3	0.123 [0.121; 0.125] *	197.480 [195.148; 199.812] *	24.240 [23.826; 24.654] *	0.242 [0.230; 0.255]	0.256 [0.250; 0.262] *

Note. AI service, artificial intelligence service; the data are presented as means [95% confidence interval]; *, confidence intervals do not overlap, indicating significant differences.

Table 8. Diagnostic accuracy parameters assessed for the results of three versions of artificial intelligence service 2 compared to radiologist conclusions

Scale	AI service	AI service version	Threshold	AUC	Sens	Spec	Acc
I	2	1	32	0.660 [0.655; 0.665] *	0.592 [0.581; 0.602]	0.728 [0.725; 0.731] *	0.714 [0.712; 0.71] *
		2	30	0.647 [0.642; 0.652]	0.621 [0.611; 0.631] *	0.673 [0.670; 0.676]	0.668 [0.665; 0.671]
		3	32	0.654 [0.649; 0.660] *	0.619 [0.609; 0.629] *	0.690 [0.687; 0.693]	0.683 [0.68; 0.685]
II	2	1	44	0.712 [0.702; 0.721] *	0.587 [0.568; 0.606]	0.837 [0.834; 0.839]*	0.829 [0.827; 0.832] *
		2	42	0.708 [0.700; 0.716] *	0.653 [0.637; 0.669] *	0.763 [0.760; 0.765]	0.759 [0.757; 0.762]
		3	44	0.688 [0.679; 0.696]	0.582 [0.566; 0.598]	0.793 [0.790; 0.795]	0.786 [0.784; 0.789]
III	2	1	44	0.709 [0.701; 0.718] *	0.576 [0.560; 0.593]	0.843 [0.841; 0.845] *	0.834 [0.832; 0.836] *
		2	39	0.722 [0.714; 0.731] *	0.688 [0.672; 0.705] *	0.756 [0.753; 0.759]	0.754 [0.751; 0.756]
		3	44	0.698 [0.689; 0.706]	0.587 [0.570; 0.604]	0.808 [0.806; 0.810]	0.801 [0.799; 0.803]

Scale	AI service	AI service version	PPV	AIR	CDR	FNR	MCC
I	2	1	0.195 [0.192; 0.198] *	304.02 [301.225; 306.815]	59.190 [58.143; 60.237]	0.408 [0.398; 0.419] *	0.209 [0.202; 0.216] *
		2	0.174 [0.172; 0.177]	356.470 [353.548; 359.392] *	62.110 [61.135; 63.085] *	0.379 [0.369; 0.389]	0.184 [0.178; 0.191]
		3	0.182 [0.179; 0.184]	341.050 [338.332; 343.768]	61.880 [60.877; 62.883]*	0.381 [0.371; 0.391]	0.195 [0.189; 0.202]
II	2	1	0.100 [0.097; 0.104] *	175.920 [173.412; 178.428]	17.600 [17.029; 18.171]	0.413 [0.394; 0.432] *	0.190 [0.181; 0.199] *
		2	0.079 [0.077; 0.080]	249.890 [247.158; 252.622] *	19.590 [19.115; 20.065] *	0.347 [0.331; 0.363]	0.164 [0.158; 0.170]
		3	0.080 [0.078; 0.082]	218.500 [216.045; 220.955]	17.470 [16.991; 17.949]	0.418 [0.402; 0.434] *	0.155 [0.148; 0.162]
III	2	1	0.108 [0.105; 0.111] *	170.700 [168.638; 172.762]	18.440 [17.907; 18.973]	0.424 [0.407; 0.440] *	0.196 [0.188; 0.204] *
		2	0.085 [0.083; 0.088]	258.390 [255.703; 261.077] *	22.030 [21.511; 22.549] *	0.312 [0.295; 0.328]	0.179 [0.172; 0.186]
		3	0.092 [0.089; 0.095]	204.570 [202.244; 206.896]	18.790 [18.255; 19.325]	0.413 [0.396; 0.430] *	0.173 [0.165; 0.180]

Note. AI service, artificial intelligence service; the data are presented as means [95% confidence interval]; *, confidence intervals do not overlap, indicating significant differences.

DOI: https://doi.org/10.17816/DD625967

Both PPV, which verifies that the majority of positive findings are actual abnormalities, and specificity are essential for minimizing false positives and achieving the maximum possible interpretation accuracy. AI service 1 was the most effective in these cases. Differences in the specificity and PPV were found between various AI service 1 versions for different scales.

A decrease in AIR cuts down the time spent by radiologists on additional data interpretation, as long as the AI service classification is unambiguous and reliable. In this study, AI service 1 exhibited the lowest AIR for all scales.

Two parameters must be considered to examine the accuracy of “normal” and “abnormal” case classification for AI services and their versions. These include the accuracy (Acc) parameter, where a high value denotes the degree to which both groups can be classified correctly, and the Matthews correlation coefficient, which examines the overall performance of a classifier, considering all components of the error matrix. The study found that version 3 of AI service 1 achieved the highest overall categorization accuracy. Notably, the AI service results were compared with radiologists’ conclusions allocated to a specific class, resulting in certain limitations because it is vital to understand the diagnostic accuracy parameters used by a radiologist. A reference dataset, where the correct value is established based on histology findings, can be used to evaluate the diagnostic accuracy of radiologists’ conclusions. Such a study has already been conducted and verified the high diagnostic accuracy of radiologist conclusions (AUC 0.928) [15]. Lower AUC values for the evaluated AI services were found in our study, indicating the necessity of updating, which was subsequently completed between 2020 and 2022. Furthermore, it is important to consider the sensitivity and specificity, which are inferior to those of radiologist conclusions [15]. The optimization of a particular parameter’s settings was not covered in this study. Importantly, the sensitivity settings in the AI services can vary for the same AUC value. For example, a sensitivity of approximately 100% removes the risk of undetected abnormalities but raises the number of false positives. We intend to perform a thorough examination of the AI service fine-tuning in the future to enhance the sensitivity and specificity.

The primary usage of AI services in mammography is for the initial reading, which will augment the accuracy of breast cancer diagnosis [16] by improving the sensitivity. Alternatively, AI services can be used for image sorting when the sensitivity is close to 100%. In this scenario, radiologists will submit the examinations as electronic medical records right away and won't need to explain the ones that an AI service has deemed "normal." A recent study exhibited promising results for this approach in the autonomous sorting of fluorography findings [17]. Numerous benign changes that may potentially necessitate attention and additional testing may make this scenario less effective in mammography.

Study limitations

This paper covers the data collected during the first three years of the large-scale Experiment on the Use of Innovative Computer Vision Technologies for Analysis of Medical Images in the Moscow Healthcare System [18]. It does not address the optimal settings of AI services. One limitation of this study is that AUC may be inadequate for evaluating the performance of AI services in a clinical setting because specific thresholds are not always applicable in real-world practice. Moreover, the sensitivity (Sens) and specificity (Spec) do not account for the population-wide prevalence of the disease. Thus, in future research, we plan to employ various techniques of assessing AI service efficacy in a clinical setting, as well as to use histological verification findings as true values. Furthermore, this study only included mammography examinations with AI service results; mammograms where an AI service failed to produce results were not evaluated. Moreover, this study did not ascertain AI service performance in patients with a foreign body in the breast (breast implants) or those with radiotherapy-induced changes. However, such cases are highly relevant to practice, and additional research is warranted.

CONCLUSION

This study discovered that the method for developing a “normal/abnormal” binary scale affects the diagnostic accuracy parameters of various types and versions of AI services. Significant discrepancies between the accuracy parameters of AI services and diagnostic scales were identified by Youden’s index, and the clinical setting determines which parameters should be utilized in the comparative evaluation of AI services. Using Youden’s index maximization to set up an AI service provides a balance of sensitivity and specificity that is not necessarily clinically significant.

ADDITIONAL INFORMATION

Funding source. This article was prepared by a group of authors as a part of the research and development effort titled "Scientific methodologies for sustainable development of artificial intelligence technologies in medical diagnostics" (USIS No.: 123031500004-5) 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.

Competing interests. The authors declare that they have no competing interests.

Authors’ contribution. All authors made a substantial contribution to the conception of the work, acquisition, analysis, interpretation of data for the work, drafting and revising the work, final approval of the version to be published and agree to be accountable for all aspects of the work. Yu.A. Vasiliev, A.V. Vladzimirskyy, O.V. Omelyanskaya, A.V. Kolsanov — research concept; K.M. Arzamasov — planning and directing the research; S.S. Semenov — data analysis; L.E. Axenova — data analysis, text writing.

REFERENCES

1. Seely JM, Alhassan T. Screening for breast cancer in 2018-what should we be doing today? *Curr Oncol.* 2018;25(suppl 1):S115–S124. doi: 10.3747/co.25.3770

2. Artificial intelligence in mammography screening. *Clinical applications, issues and directions for development* [Internet; cited 20 August 2023]. Available from: https://www.itmportal.ru/upload/iblock/69e/7q981uhfaxjhcntal0exngxtq43xeth2/2.2.3.-Kandoba-ITM_AI-2022.pdf (in Russ.)

3. Celsus — *AI-software for analysis of X-ray and CT studies.* Mammography [Internet; cited 20 Aug 2023]. Available from: https://celsus.ai/products-mammography/

4. Kim HE, Kim HH, Han BK, et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health.* 2020;2(3):e138–e148. doi: 10.1016/S2589-7500(20)30003-0

5. Yoon JH, Strand F, Baltzer PAT, et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis. *Radiology.* 2023;307(5):e222639. doi: 10.1148/radiol.222639

6. Zhou X-H, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine.* NJ: John Wiley & Sons, Inc.; 2011. doi: 10.1002/9780470906514

7. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb).* 2016;26(3):297–307. doi: 10.11613/BM.2016.034

8. Schaffter T, Buist DSM, Lee CI, et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms. *JAMA Netw Open.* 2020;3(3):e200265. doi: 10.1001/jamanetworkopen.2020.0265

9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature.* 2020;577(7788):89–94. doi: 10.1038/s41586-019-1799-6

10. Nam JG, Kim M, Park J, et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs. *Eur Respir J.* 2021;57(5):2003061. doi: 10.1183/13993003.03061-2020

11. Sakhnov SN, Axenov KD, Axenova LE, et al. Development of a cataract screening model using an open dataset and deep machine learning algorithms. *Fyodorov Journal of Ophthalmic Surgery.* 2022;(S4):13–20. EDN: VEGPAW doi: 10.25276/0235-4160-2022-4S-13-20

12. King G, Zeng L. Logistic Regression in Rare Events Data. *Political Analysis.* 2001;9(2):137–163. doi: 10.1093/oxfordjournals.pan.a004868

13. Chen F, Xue Y, Tan MT, Chen P. Efficient statistical tests to compare Youden index: accounting for contingency correlation. *Stat Med.* 2015;34(9):1560–1576. doi: 10.1002/sim.6432

14. Vasiliev YuA, Vladzimirsky AV, Sharova DE, et al. *Clinical trials of artificial intelligence systems (radiation diagnostics).* Moscow: State budgetary healthcare institution of the city of Moscow «Scientific and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Department». 2023. 40 p. (In Russ.) EDN: PUIJLD

15. Arzamasov KM, Vasilev YuA, Vladzymyrskyy AV, et al. The use of computer vision for the mammography preventive research. *The Russian Journal of Preventive Medicine.* 2023;26(6):117–123. EDN: YBKHPS doi: 10.17116/profmed202326061117

16. Vasilev YuA, Tyrov IA, Vladzymyrskyy AV, et al. Double-reading mammograms using artificial intelligence technologies: A new model of mass preventive examination organization. *Digital Diagnostics.* 2023;4(2):93–104. EDN: VRIEOH doi: 10.17816/DD321423

17. Vasilev YuA, Tyrov IA, Vladzymyrskyy AV, et al. A New Model of Organizing Mass Screening Based on Stand-Alone Artificial Intelligence Used for Fluorography Image Triage. *Public Health and Life Environment — PH&LE.* 2023;31(11):23-32. EDN: SYIQBX doi: 10.35627/2219-5238/2023-31-11-23-32

18. Vladzimirskyy AV, Vasilev YuA, Arzamasov KM, et al. *Computer vision in radiology: the first stage of the Moscow experiment.* Moscow: Izdatel’skie resheniya; 2022. (In Russ.) EDN: FOYLXK

СПИСОК ЛИТЕРАТУРЫ

1. Seely J.M., Alhassan T. Screening for breast cancer in 2018-what should we be doing today? // Curr Oncol. 2018. Vol. 25, Suppl. 1. P. S115–S124. doi: 10.3747/co.25.3770

2. Кандоба В.И. Искусственный интеллект в скрининговой маммографии. Клиническое использование, проблемы и направления развития [интернет]. Режим доступа: https://www.itmportal.ru/upload/iblock/69e/7q981uhfaxjhcntal0exngxtq43xeth2/2.2.3.-Kandoba-ITM_AI-2022.pdf Дата обращения: 20.08.2023

3. Цельс. Система поддержки принятия врачебных решений на базе технологий искусственного интеллекта для анализа цифровых медицинских изображений. Маммография [интернет]. Режим доступа: https://celsus.ai/products-mammography/ Дата обращения: 20.08.2023

4. Kim H.E., Kim H.H., Han B.K., et al. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study // Lancet Digit Health. 2020. Vol. 2, N 3. P. e138–e148. doi: 10.1016/S2589-7500(20)30003-0

5. Yoon J.H., Strand F., Baltzer P.A.T., et al. Standalone AI for Breast Cancer Detection at Screening Digital Mammography and Digital Breast Tomosynthesis: A Systematic Review and Meta-Analysis // Radiology. 2023. Vol. 307, N 5. ID: e222639. doi: 10.1148/radiol.222639

6. Zhou X.-H., Obuchowski N.A., McClish D.K. Statistical Methods in Diagnostic Medicine. NJ: John Wiley & Sons, Inc.; 2011. doi: 10.1002/9780470906514

7. Habibzadeh F., Habibzadeh P,, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results // Biochem Med (Zagreb). 2016. Vol. 26, N 3. P. 297–307. doi: 10.11613/BM.2016.034

8. Schaffter T., Buist D.S.M., Lee C.I., et al. Evaluation of Combined Artificial Intelligence and Radiologist Assessment to Interpret Screening Mammograms // JAMA Netw Open. 2020. Vol. 3, N 3. ID: e200265. doi: 10.1001/jamanetworkopen.2020.0265

9. McKinney S.M., Sieniek M., Godbole V., et al. International evaluation of an AI system for breast cancer screening // Nature. 2020. Vol. 577, N 7788. P. 89–94. doi: 10.1038/s41586-019-1799-6

10. Nam J.G., Kim M., Park J., et al. Development and validation of a deep learning algorithm detecting 10 common abnormalities on chest radiographs // Eur Respir J. 2021. Vol. 57, N 5. ID: 2003061. doi: 10.1183/13993003.03061-2020

11. Сахнов С.Н., Аксенов К.Д., Аксенова Л.Е., и др. Разработка модели скрининга катаракты с использованием открытого набора данных и алгоритмов глубокого машинного обучения. Офтальмохирургия. 2022. № S4. С. 13–20. EDN: VEGPAW doi: 10.25276/0235-4160-2022-4S-13-20

12. King G., Zeng L. Logistic Regression in Rare Events Data // Political Analysis. 2001. Vol. 9, N 2. P. 137–163. doi: 10.1093/oxfordjournals.pan.a004868

13. Chen F., Xue Y., Tan M.T., Chen P. Efficient statistical tests to compare Youden index: accounting for contingency correlation // Stat Med. 2015. Vol. 34, N 9. P. 1560–1576. doi: 10.1002/sim.6432

14. Васильев Ю.А., Владзимирский А.В., Шарова Д.Е., и др. Клинические испытания систем искусственного интеллекта (лучевая диагностика). Москва: Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы», 2023. 40 с. EDN: PUIJLD

15. Арзамасов К.М., Васильев Ю.А., Владзимирский А.В., и др. Применение компьютерного зрения для профилактических исследований на примере маммографии // Профилактическая медицина. 2023. Т. 26, № 6. С. 117–123. EDN: YBKHPS doi: 10.17116/profmed202326061117

16. Васильев Ю.А., Тыров И.А., Владзимирский А.В., и др. Двойной просмотр результатов маммографии с применением технологий искусственного интеллекта: новая модель организации массовых профилактических исследований // Digital Diagnostics. 2023. Т. 4, № 2. С. 93–104. EDN: VRIEOH doi: 10.17816/DD321423

17. Васильев Ю.А., Тыров И.А., Владзимирский А.В., и др. Новая модель организации массовых профилактических исследований, основанная на автономном искусственном интеллекте для сортировки результатов флюорографии // Здоровье населения и среда обитания — ЗНиСО. 2023. Т. 31, № 11. С. 23–32. EDN: SYIQBX doi: 10.35627/2219-5238/2023-31-11-23-32

18. Владзимирский А.В., Васильев Ю.А., Арзамасов К.М., и др. Компьютерное зрение в лучевой диагностике: первый этап Московского эксперимента. Москва: Издательские решения, 2022. EDN: FOYLXK

AUTHORS' INFO

*** Kirill M. Arzamasov,** MD, Cand. Sci. (Medicine);
address: 24 bldg. 1 Petrovka str., 127051, Moscow, Russia;
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

Yuriy A. Vasilev, MD, Cand. Sci. (Medicine);
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: VasilevYA1@zdrav.mos.ru

Alexander V. Kolsanov, MD, Dr. Sci. (Medicine), Professor;
ORCID: 0000-0002-4144-7090;
eLibrary SPIN: 2028-6609;
e-mail: a.v.kolsanov@samsmu.ru

Anton V. Vladzymyrskyy, MD, Dr. Sci. (Medicine), Professor;
ORCID: 0000-0002-2990-7736;
eLibrary SPIN: 3602-7120;
e-mail: VladzimirskijAV@zdrav.mos.ru

Olga V. Omelyanskaya;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: OmelyanskayaOV@zdrav.mos.ru

Serafim S. Semenov;
ORCID: 0000-0003-2585-0864;
eLibrary SPIN: 4790-0416;
e-mail: SemenovSS3@zdrav.mos.ru

Lubov E. Axenova;
ORCID: 0000-0003-0885-1355;
eLibrary SPIN: 7705-6293;
e-mail: AksenovaLE@zdrav.mos.ru

ОБ АВТОРАХ

*** Арзамасов Кирилл Михайлович,** канд. мед. наук;
адрес: Россия, 127051, Москва, ул. Петровка, д. 24, стр. 1;
ORCID: 0000-0001-7786-0349;
eLibrary SPIN: 3160-8062;
e-mail: ArzamasovKM@zdrav.mos.ru

Васильев Юрий Александрович, канд. мед. наук;
ORCID: 0000-0002-5283-5961;
eLibrary SPIN: 4458-5608;
e-mail: VasilevYA1@zdrav.mos.ru

Колсанов Александр Владимирович, д-р мед. наук, профессор;
ORCID: 0000-0002-4144-7090;
eLibrary SPIN: 2028-6609;
e-mail: a.v.kolsanov@samsmu.ru

Владзимирский Антон Вячеславович, д-р мед. наук, профессор;
ORCID: 0000-0002-2990-7736;
eLibrary SPIN: 3602-7120;
e-mail: VladzimirskijAV@zdrav.mos.ru

Омелянская Ольга Васильевна;
ORCID: 0000-0002-0245-4431;
eLibrary SPIN: 8948-6152;
e-mail: OmelyanskayaOV@zdrav.mos.ru

Семёнов Серафим Сергеевич;
ORCID: 0000-0003-2585-0864;
eLibrary SPIN: 4790-0416;
e-mail: SemenovSS3@zdrav.mos.ru

Аксёнова Любовь Евгеньевна;
ORCID: 0000-0003-0885-1355;
eLibrary SPIN: 7705-6293;
e-mail: AksenovaLE@zdrav.mos.ru

* Corresponding author / Автор, ответственный за переписку