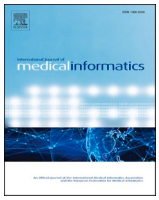
International Journal of Medical Informatics 178 (2023) 105190

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Clinical application of radiological AI for pulmonary nodule evaluation: Replicability and susceptibility to the population shift caused by the COVID-19 pandemic

Yuriy Vasilev [a], Anton Vladzymyrskyy [a,b], Kirill Arzamasov [a], Olga Omelyanskaya [a], Igor Shulkin [a], Darya Kozikhina [a], Inna Goncharova [a], Roman Reshetnikov [a], Sergey Chetverikov [a], Ivan Blokhin [a], Tatiana Bobrovskaya [a,*], Anna Andreychenko [a]

[a] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Healthcare Department, Moscow, Russia
[b] I.M. Sechenov First Moscow State Medical University of the Ministry of Health of the Russian Federation (Sechenov University), 8-2 Trubetskaya str. Moscow, 119991, Russian Federation

ARTICLE INFO

Keywords:
Artificial Intelligence
Medical Image Analysis
Fine-tuning
Lung Cancer
COVID-19
Replicability of medical machine learning

ABSTRACT

Purpose: replicability and generalizability of medical AI are the recognized challenges that hinder a broad AI deployment in clinical practice. Pulmonary nodes detection and characterization based on chest CT images is one of the demanded use cases for automatization by means of AI, and multiple AI solutions addressing this task are becoming available. Here, we evaluated and compared the performance of several commercially available radiological AI with the same clinical task on the same external datasets acquired before and during the pandemic of COVID-19.

Approach: 5 commercially available AI models for pulmonary nodule detection were tested on two external datasets labelled by experts according to the intended clinical task. Dataset1 was acquired before the pandemic and did not contain radiological signs of COVID-19; dataset2 was collected during the pandemic and did contain radiological signs of COVID-19. ROC-analysis was applied separately for the dataset1 and dataset2 to select probability thresholds for each dataset separately. AUROC, sensitivity and specificity metrics were used to assess and compare the results of AI performance.

Results: Statistically significant differences in AUROC values were observed between the AI models for the dataset1. Whereas for the dataset2 the differences of AUROC values became statistically insignificant. Sensitivity and specificity differed statistically significantly between the AI models for the dataset1. This difference was insignificant for the dataset2 when we applied the probability threshold initially selected for the dataset1. An update of the probability threshold based on the dataset2 created statistically significant differences of sensitivity and specificity between AI models for the dataset2. For 3 out of 5 AI models, the update of the probability threshold was valuable to compensate for the degradation of AI model performances with the population shift caused by the pandemic.

Conclusions: Population shift in the data is able to deteriorate differences of AI models performance. Update of the probability threshold together with the population shift seems to be valuable to preserve AI models performance without retraining them.

1. Introduction

The appearance of open-access labeled datasets of medical imaging has led to a rapid increase in the number of AI models in medical diagnostics, in particular in radiology[1]. One of the parameters that characterize the practical applicability and value of AI models is the ability to independently classify and generalize the data received[2]. The generalizability of AI models is an important feature that ensures the trained models to be applied effectively to new data and under new conditions[3–6]. It is known that AI models show the best diagnostic accuracy metrics for the similar patient group and pathological characteristics that were present in the training dataset. However, even

* Corresponding author at: 24 Petrovka Str., bldg. 1, Moscow, 127051, Russia.
E-mail address: BobrovskayaTM@zdrav.mos.ru (T. Bobrovskaya).

https://doi.org/10.1016/j.ijmedinf.2023.105190
Received 19 March 2023; Received in revised form 4 August 2023; Accepted 7 August 2023
Available online 9 August 2023

training the AI model on a dataset based on studies of the target patient population may not be sufficient for the acceptable performance of the AI model during deployment, e.g., [7–9]in the event of the appearance of the previously unaccounted highly prevalent comorbidities. For example, detection of the pulmonary nodules in the presence of the COVID-19 related radiological signs on chest CTs[10]. Particular difficulties can occur for the diagnosis of an early-stage lung cancer that has subsolid/ground glass nodule[11,12]. The latter appear on CT scans very similar to small areas of ground glass opacities that are associated with a mild lung parenchyma involvement of the COVID-19 infection [13]. Thus, it could distort a timely diagnosis of the lung cancer and selection of the proper treatment pathways in the lung cancer care[14]. Thus, lung pulmonary nodule detection task in the presence of the COVID-19 is an obvious population shift for the AI models which impact on the AI performance has to be investigated. Moreover, since in order to assess the severity of lung damage and the subsequent choice of treatment tactics for the COVID-19 patient, chest CTs were widely performed during the pandemic[15], it represents a pivotal opportunity to screen chest CT studies for the lung cancer radiological signs, using the AI models. However, the stability and replicability of diagnostic performance of the existing AI models for lung nodule detection should be first studied and compared for the target population before and during the COVID-19 pandemic. In this study, we compared on the same external data the performances of AI models for lung nodule detection before and during the COVID-19 pandemic.

2. Methods

This retrospective study was designed according to Standards for Reporting Diagnostic accuracy studies (STARD) 2015 guidelines. Approval was obtained from the local ethics committee. Separate informed consent was not required for this retrospective study. An overall study design is shown in Figure S1.

2.1. AI models and datasets

The study included six AI-based commercial models that participated in the research registered in ClinicalTrials (NCT04489992): AI-1 [16], AI-2 [17], AI-3 [18], AI-4 [19], AI-5 [20], AI-6 [21]. The developers announced that all diagnostic metrics (i.e., sensitivity, specificity and AUROC) for a target pathology were greater than 0.81. The target pathology was solid/subsolid nodules with only a solid component is measured larger than 6 mm (greater than100 mm^3) [22]. The criterion for the inclusion of these models was full compliance with the use cases, that is, each declared the possibility of determining the pulmonary nodules larger than 6 mm. The exclusion criterion was in compliance with the analysis of the more than 10% of cases in each dataset.

The details on AI models' development that were provided by the developers are the following:

- AI-1 model was based on U-net and ResNet-50 networks. It was developed analogously to [23]. The datasets used for training and internal validation were LIDC-IDRI[24], NSCLC-Radiomics[25] and a private dataset consisting of 1775 chest CT scans (912 scans contained suspicious lung nodules confirmed by two thoracic radiologists (greater than5 years of experience).
- AI-2 model was based on Faster-RCNN and LSTM networks. During the training, radam optimizer[26] and focal loss[27] were applied. Only a private dataset was used that comprised 1231 chest CT scans (775 with suspicious lung nodules).
- AI-3 model architecture was based on an ensemble of U-net, Densenet and ResNet networks. Luna[28], LNDb[29], NSCLC-Radiomics [25], CTLungCa-500[30] and a private dataset of 3918 chest CT scans were used for training and internal validations.
- AI-4 model was based on Faster-RCNN 3d network. Only a private dataset was used that consisted of 3500 chest CT scans with 5123 suspicious lung nodules confirmed by two experienced radiologists.
- AI-5 model[20] consisted of a combination of U-nets for lung segmentation and lung nodules detection and characterization, focal loss[27] and log loss functions were used during training. For internal validation, a private dataset of 250 chest CT scans (100 with the target pathology) was used. The presence of the suspicious nodules was determined based on a consensus of the experienced thoracic radiologists (greater than10 years of experience).
- AI-6 model: the developer refused to provide information.

Two datasets were used for evaluation of the AI models: Dataset1 (before the COVID-19 pandemic) and Dataset2 (during the COVID-19 pandemic[31]). Both datasets included anonymized unenhanced chest CT studies acquired at the multiple outpatient radiology departments [32] using Toshiba Aquilion 64 CT scanners. Similar scanning and reconstruction parameters were used for each of the different scanners: 120 kVp tube-voltage, 80–500 mAs tube-current (automatically adjusted to achieve noise level of 10 HU for 5.0 mm thick slices), caudocranial direction, pitch of 1.5, slice thickness of 1 mm, FC07 reconstruction kernel, 512 × 512 reconstruction matrix, an average of 300–400 images (slices) per study.

Inclusion criteria of CT studies for both datasets were the following: fully covered lungs, at least 10 mm distance from the lungs to the border of the field of view.

Specific inclusion criteria for Dataset2 was the positive PCR test of a patient with the COVID-19 lung involvement revealed on the CT scans. If a CT scan revealed surgical interventions, patient-related artifacts (hand overlay over chest, body orientation, coughing, movements) or poor-quality scanning (planning, technical defects) this scan was not included in the datasets.

Each CT study in both datasets was marked as "with the target pathology" if it had at least one pulmonary nodule larger than 6 mm, independently confirmed by two experienced radiologists (5 years of experience in thoracic radiology). When neither of two radiologists detected pulmonary nodules greater than 6 mm, the study was marked as "without the target pathology". The studies that did not fulfill one of these two conditions were not included in the datasets.

Dataset1 contained 100 chest CT scans (38 with the target pathology and 62 without the target pathology) acquired before the pandemic of COVID-19 (i.e., before December 2019). Patients characteristics of the dataset1 were the following: median age (IQR) – 61 (52–68), 55 female/36 male. After inclusion of the studies that were processed successfully by all AI models, 91 chest CT scans remained in the dataset1.

Dataset2 contained 100 chest CT scans (50 with the target pathology and 50 without the target pathology), acquired during the COVID-19 pandemic. 32 out of 50 studies with the target pathology had up to 50% lung parenchyma involvement due to the COVID-19, 23 out of 50 studies without the target pathology had up to 50% lung parenchyma involvement due to the COVID-19 (Figure S1). Patients characteristics of the dataset2 were the following: median age (IQR) – 59 (42–72), 47 female/38 male. After inclusion of the studies that were processed successfully by all AI models, 85 chest CT scans remained in the dataset2.

In light of the above, dataset1 and dataset2 are comparable in terms of gender and age composition, but in dataset1 the class balance is slightly biased towards the normal studies (38:53 vs. 40:45). At the same time, dataset2 contains studies with signs of COVID-19-associated pneumonia both in the subgroup with target pathology (28:40) and without target pathology (21:45). Both datasets were collected from 32 outpatient medical facilities equipped with a single model of CT scanner. Thus, we can speak of a fairly homogeneous data composition.

The anonymized studies of both datasets were sent sequentially from PACS to the AI models that were located on external cloud servers. AI models responses were received back in the form of an abnormality

score and DICOM files in case the study was analyzed correctly, or an error message if the study could not be analyzed by AI. When at least one of the AI models processed a study unsuccessfully, the study was excluded from the further analysis.

The AI models returned the abnormality score for the whole study. Area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, net benefit [33] and threshold values were used to evaluated and compare performances of the AI models[34]. Since the datasets were relatively balanced for two classes, ROC analysis was used to select an optimal operating point. The operating point is applied to convert the continuous abnormality score of the AI model to a dichotomous decision: studies with scores equal or above the operating point are interpreted as "with the target pathology" according to the AI model, studies with scores below the operating point are labelled as "without the target pathology". Then a confusion matrix is constructed is order to calculate diagnostic metrics. Sensitivity and specificity were calculated at the operating point of the maximum Youden index[35,36]. The threshold (operating point) values for each AI model were obtained separately for the dataset1 and the dataset2.

The first part of the study aimed to explore whether there were significant differences in the AI models performances on the same datasets (before and during the COVID-19 pandemic). For this purpose, diagnostic accuracy metrics (ROC AUC, sensitivity, specificity, net benefit) of each AI model were compared with each other on dataset1 and dataset2 (Figure S1).

The second part investigated whether the operating point adjustment during the COVID-19 pandemic had a significant impact on the AI models' performance. For this, diagnostic accuracy metrics (sensitivity, specificity, net benefit) were compared on dataset 2 with and without threshold adjustment.

We also analyzed the data obtained using logistic regression to calculate the positive class thresholds [37].

2.2. Statistical analysis

The area under the ROC curve (AUROC) is reported with 95% confidence intervals (CIs 95%). The DeLong method was used to calculate the confidence interval for AUROC[38].

The 95% confidence intervals for the sensitivity and specificity were determined by means of a binominal proportion [39].

ROC AUC were compared using roc.test method = DeLong from R language, with correction for multiple FDR. The McNemar with Yates Correction and False discovery rate (FDR) multiple comparison test was used to compare sensitivity and specificity. The null hypothesis (H_0) of no statistically significant differences in diagnostic accuracy was tested. A p-value less than 0.05 was considered an indicator of the significant difference.

For calculations, we used the web tool for the comprehensive ROC analysis: https://roc-analysis.mosmed.ai.

A complete study design is shown in Figure S1. AI-6 was not included in the final analysis because it did not process studies from the dataset2.

3. Results

Diagnostic accuracy metrics per AI model obtained for two datasets before (dataset1) and during the coronavirus pandemic (dataset2) are shown in Table1. A detailed comparative analysis of the diagnostic accuracy metrics and threshold values is presented below.

3.1. Auroc

On the data acquired before the pandemic (dataset1) a maximal AUROC of 0.97 (95%CI: 0.94–1.0) among all AI models was shown by the AI-1 model (Fig. 1, left column). AUROC of five AI models varied between 0.97 and 0.72 (Table 1). A statistically significant difference (p less than 0.05) between AUROC values occurred between AI-1 vs. AI-2,

Table 1
Diagnostic accuracy metrics obtained in two datasets before (dataset1) and during the coronavirus pandemic (dataset2) with and without threshold update. Note that AUROC is independent from the threshold update and thus is presented only once.

Diagnostic metrics	Dataset1					Dataset2 (with threshold update)					Dataset2 (without threshold update)				
	AI-1	AI-2	AI-3	AI-4	AI-5	AI-1	AI-2	AI-3	AI-4	AI-5	AI-1	AI-2	AI-3	AI-4	AI-5
AUROC (CI 95%)	0.97 (0.94–1.0)	0.72 (0.61–0.83)	0.92 (0.86–0.98)	0.94 (0.89–0.99)	0.78 (0.69–0.87)	0.88 (0.81–0.95)	0.89 (0.81–0.96)	0.80 (0.70–0.89)	0.82 (0.72–0.91)	0.75 (0.65–0.85)	same values as with threshold update				
Sensitivity (CI 95%)	0.97 (0.92–1.0)	0.63 (0.48–0.78)	0.95 (0.88–1.0)	0.95 (0.88–1.0)	0.71 (0.57–0.85)	0.90 (0.81–0.99)	0.92 (0.84–1.0)	1.0 (1.0–1.0)	0.85 (0.74–0.96)	0.73 (0.59–0.86)	0.70 (0.56–0.84)	0.72 (0.59–0.86)	1.0 (1.0–1.0)	0.95 (0.88–1.0)	0.92 (0.84–1.0)
Specificity (CI 95%)	0.98 (0.94–1.0)	0.85 (0.75–0.95)	0.87 (0.78–0.96)	0.79 (0.68–0.90)	0.79 (0.68–0.90)	0.78 (0.66–0.90)	0.73 (0.60–0.86)	0.6 (0.46–0.74)	0.76 (0.63–0.88)	0.73 (0.60–0.86)	0.84 (0.74–0.95)	0.89 (0.80–0.98)	0.53 (0.39–0.68)	0.53 (0.39–0.68)	0.49 (0.34–0.63)
Threshold (operating point)	51	93	34	87	6	16	35	42	97	18	51	93	34	87	6
Net Benefit	0.40 (0.36–0.42)	0.22 (0.12–0.31)	0.33 (0.25–0.40)	0.27 (0.17–0.36)	0.21 (0.10–0.31)	0.28 (0.17–0.40)	0.24 (0.11–0.37)	0.02 (-0.14–0.17)	0.25 (0.13–0.38)	0.21 (0.08–0.34)	0.27 (0.17–0.38)	0.29 (0.21–0.40)	−0.04 (-0.26–0.07)	−0.20 (-0.41–0.04)	−0.21 (-0.44–0)

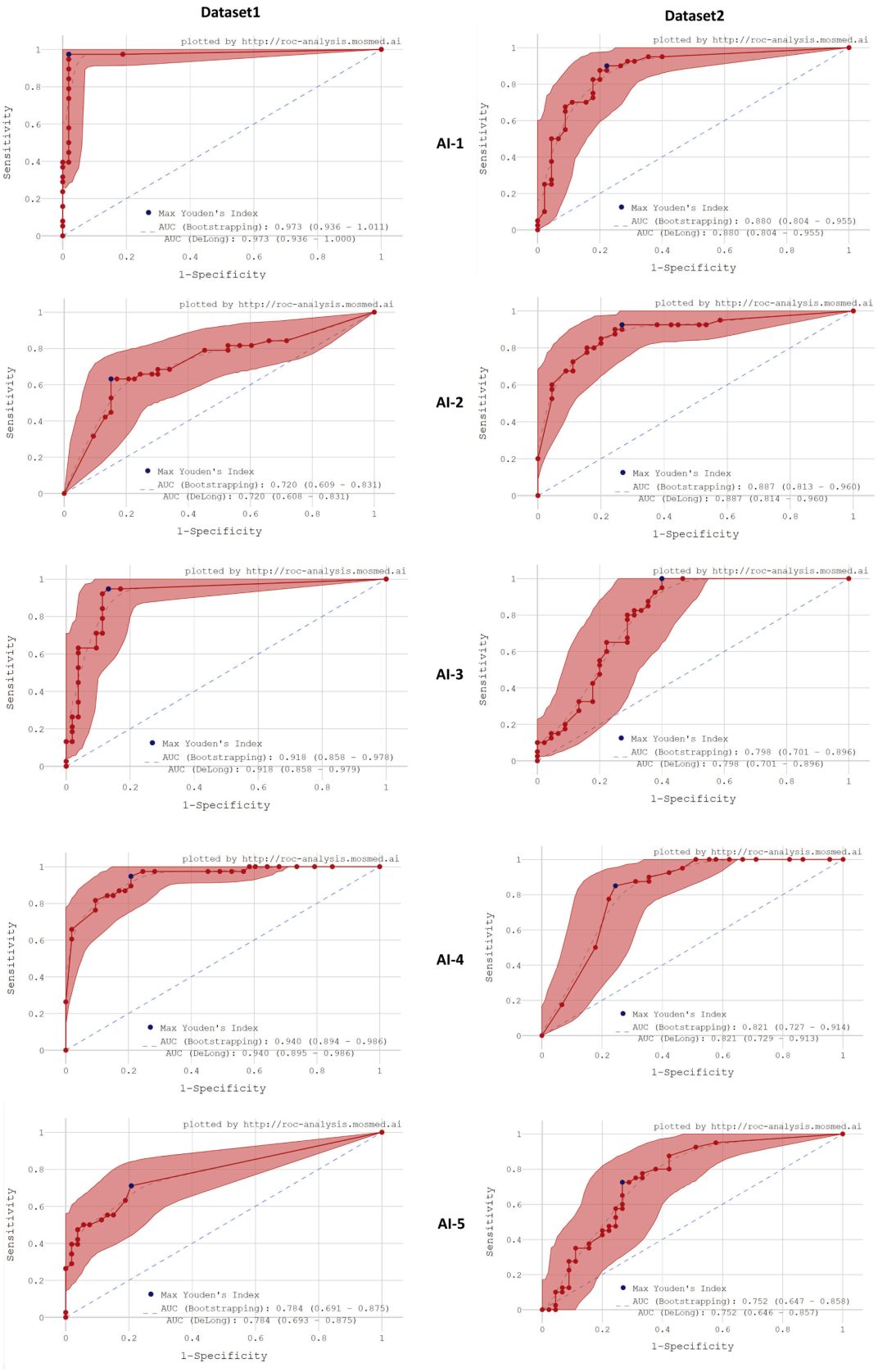


Fig. 1. ROC curves of AI models obtained on the two datasets.

AI-1 vs. AI-5, AI-2 vs. AI-3, AI-2 vs. AI-4, AI-3 vs. AI-5, AI-4 vs. AI-5 pairs (Table 2). On the data acquired during the pandemic (dataset2) a maximal AUROC of 0.89 (95%CI: 0.81–0.96) among all AI models was shown by the AI-2 model (Fig. 1, right column). AUROC of five AI models varied between 0.89 and 0.75 (Table 1). No statistically significant difference between AUROC occurred between five AI-models (Table 3).

Four out of five AI models had lower AUROC values for the dataset1 than for the dataset2, and only AI-2 had a higher AUROC value for the dataset 2 than for the dataset2. However, these difference were not statistically significant (p greater than 0.05).

3.2. Sensitivity and specificity

Sensitivity values ranged among the AI-models between 0.97 and 0.71 for the dataset1, 1.00–0.70 for the dataset2 without the threshold update and 1.00–0.73 for the dataset2 with the threshold update (Table 1). Pairwise statistically significant differences between AI models were observed for each of the three settings (Table 4). However, no pair of AI models remained statistically significant different for all the three settings simultaneously.

Specificity values ranged among the AI-models between 0.98 and 0.79 for the dataset1, 0.89–0.49 for the dataset2 without the threshold update and 0.78–0.60 for the dataset2 with the threshold update (Table1). Pairwise statistically significant differences between AI models were observed for two of the three settings, namely dataset1 and dataset2 with the threshold update (Table 4). One pair of AI models (AI-1 vs. AI-5) remained statistically significant different for these two settings simultaneously.

An example of a study with the target pathology from the dataset2 is shown in Fig. 2. Four out of five AI models detected the lung nodule correctly; however, two of them had additional false positive findings in the same study related to the COVID-19 lung parenchyma changes.

Fig. 3 shows the study without the target pathology from the dataset2, for which all five AI models identified the COVID-19 lesions as a lung nodule. Figs. 4 and 5 show the studies without the target pathology from the dataset1, which majority of the AI models indicated as suspicious (i.e., false positive findings).

3.3. Threshold update

When determining the optimal threshold (operating point) for each AI model, different values were obtained for the dataset1 and the dataset2. The difference in relation to the threshold value obtained on the dataset1 ranged from −12 to + 58 while the threshold may have values in the range from 0 to 100. An impact of the threshold update of the AI models based on the dataset2 in comparison with the setting when the threshold would be determined on the dataset1 only was studied for the sensitivity and specificity values of AI models for the dataset2. Table 5 demonstrates that the threshold update had (1) stabilized specificity values of three AI models between dataset1 and dataset2 (AI-2, AI-4 and AI-5); (2) had no impact for one AI model (AI-3); (3) had stabilized sensitivity values of one AI model (AI-1) between dataset1 and dataset2.

The results of using logistic regression to calculate the positive class thresholds are presented in Supplement (Table S1). They suggest that regression-based thresholds are suboptimal for lung nodule detection in the clinical setting. Therefore, in the future, we will analyze the results obtained by the method of the maximum Youden index.

Table 2
Pairwise AUROC values comparison of AI models for the dataset1. Bold font indicates statistically significant difference. Regular font shows statistically insignificant difference. ROC AUC values are present on the main diagonal

i / *j*	AI-1	AI-4	AI-3	AI-5	AI-2
AI-1	0,97	0,03*	0,05	**0,19**	**0,25**
AI-4	−0,03**	0,94	0,02	**0,16**	**0,22**
AI-3	−0,05	−0,02	0,92	**0,14**	**0,20**
AI-5	**−0,19**	**−0,16**	**−0,14**	0,78	0,06
AI-2	**−0,25**	**−0,22**	**−0,20**	−0,06	0,72

*AI-*j* > AI-*i* – positive value.
**AI-*j* < AI-*i* – negative value.

Table 3
Pairwise AUROC values comparison of AI models for the dataset2 (with updated threshold). Bold font indicates statistically significant difference. Regular font shows statistically insignificant difference. ROC AUC values are present on the main diagonal.

i / *j*	AI-2	AI-1	AI-4	AI-3	AI-5
AI-2	0,89	0,01	0,07	0,09	0,14
AI-1	−0,01	0,88	0,06	0,08	0,13
AI-4	−0,07	−0,06	0,82	0,02	0,07
AI-3	−0,09	−0,08	−0,02	0,8	0,05
AI-5	−0,14	−0,13	−0,07	−0,05	0,75

3.4. Net benefit

Each model's net benefit was calculated for dataset1, dataset 2, and dataset 2 after threshold adjustment (Table 1). For models 4 and 5, the benefit decreased significantly in dataset 2, but returned to the previous values after threshold adjustment. For model 3, the net benefit decreased on dataset2, and threshold adjustment did not increase the benefit. For model 2, the net benefit did not change. For model 1, it decreased, and adjusting the threshold could not affect the benefit.

4. Discussion

Here, we validated externally and compared five radiological AI models. The validation and comparison were performed on the same data and for two settings: before and during the COVID-19 pandemic. A population shift caused by the pandemic resulted in an overall statistically significant degradation of inter- and intra-performances of the majority of included AI models. The significance of the degradation could be decreased though for most of the AI models studied in this work by the prediction threshold update.

The diagnosis of pulmonary nodules within the superimposed COVID-19 related radiological signs is a challenging task even for radiologists due to a number of factors. For example, the concentration of radiologist's attention on the underlying pathology, radiological changes make it difficult to assess comprehensively the lung parenchyma. Therefore, it is valuable to use the AI in conditions of the pandemic and the adaptation of AI systems to the presence of comorbidities[40]. Due to the similarity of the radiological pattern of COVID-19 and early lung cancer[13], AI models often showed false positives in the dataset2 when ground glass was erroneously assessed by the AI model as a suspicion of the pulmonary nodule. A multiclassification deep learning model for diagnosing COVID-19, pneumonia, and other chest diseases could differentiate between lung cancer and COVID-19. Deep learning can possibly differentiate individual types of pathology (multiclass classification) with a high degree of certainty. But can one type of pathology be so well differentiated by AI against the background of another (multilabel classification)? Our recent study showed an acceptable level of diagnostic precision in 3 of 5 studied AI models that had an AUC greater than 0.81[41]. Previous study have already demonstrated the varying effectiveness of AI models in detecting pulmonary nodules in the context of COVID-19[42]. Of the great practical importance is the composition of the datasets on which the AI model was trained. Most studies have been retrospective, using historic data to train models; the true utility comes to the fore in the real-world setting, which

Table 4
Pairwise comparison of sensitivity and specificity values of AI models for the dataset1, dataset2 (without the threshold update) and dataset2 (with the threshold update). Bold font indicates statistically significant difference. Regular font shows statistically insignificant difference.

X vs Y	AI-1 vs AI-2	AI-1 vs AI-3	AI-1 vs AI-4	AI-1 vs AI-5	AI-2 vs AI-3	AI-2 vs AI-4	AI-2 vs AI-5	AI-3 vs AI-4	AI-3 vs AI-5	AI-4 vs AI-5
Dataset1										
Sensitivity	**0,34***	0,02	0,02	**0,26**	**−0,32** **	**−0,32**	−0,08	0***	**0,24**	**0,24**
Specificity	0,13	0,11	**0,19**	**0,19**	−0,02	0,06	0,06	0,08	0,08	0
Dataset2 (with the threshold update)										
Sensitivity	−0,02	−0,10	0,05	0,17	−0,08	0,07	0,19	0,15	**0,27**	0,12
Specificity	0,05	0,18	0,02	0,05	0,13	−0,03	0	−0,16	−0,13	0,03
Dataset2 (without the threshold update)										
Sensitivity	−0,02	**−0,3**	**−0,25**	**−0,22**	**−0,28**	**−0,23**	**−0,2**	0,05	0,08	0,03
Specificity	−0,05	**0,31**	**0,31**	**0,35**	**0,36**	**0,36**	**0,4**	0	0,04	0,04

*X > Y – positive value.
**X < Y – negative value.
***no difference.

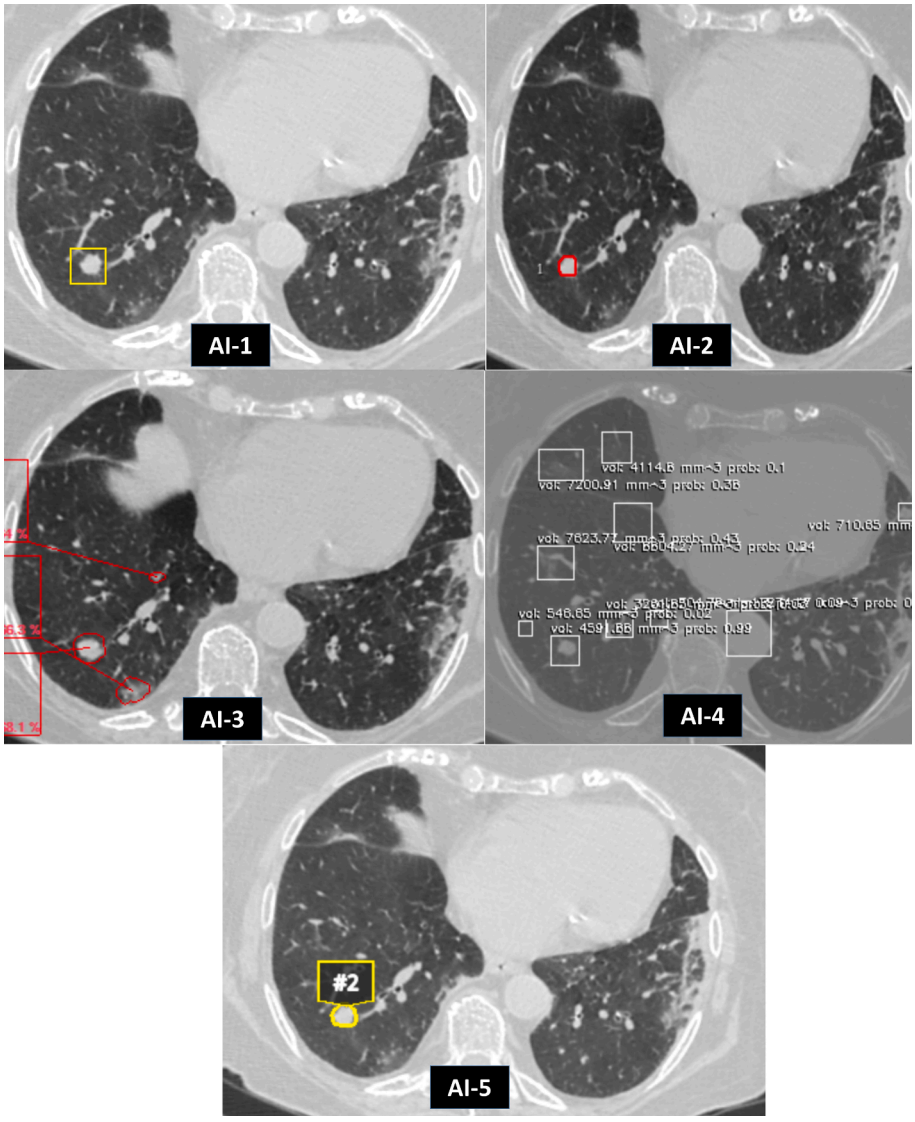


Fig. 2. Case #1 from Dataset2. Case with target pathology – solid lung nodule greater than 6 mm. AI-1 missed a solid nodule of 11 mm. AI-2, AI-5 detected the solid nodule correctly. AI-3, AI-4 – detected the solid nodule correctly, but also there were additional false positive findings associated with Covid-19.

may vastly differ from that experienced in the model training[43]. In our study, unfortunately, we do not reliably know the characteristics of the datasets on which the AI models were trained as we assessed them as the ready-to-use solutions. However, our study led to an important conclusion about the need to fine-tune the AI model to work effectively under the new conditions (COVID-19). According to Susana Goncalves et al., [10] the lack of real-world validation is rapidly being addressed with many studies (ongoing/completed) integrating an external AI

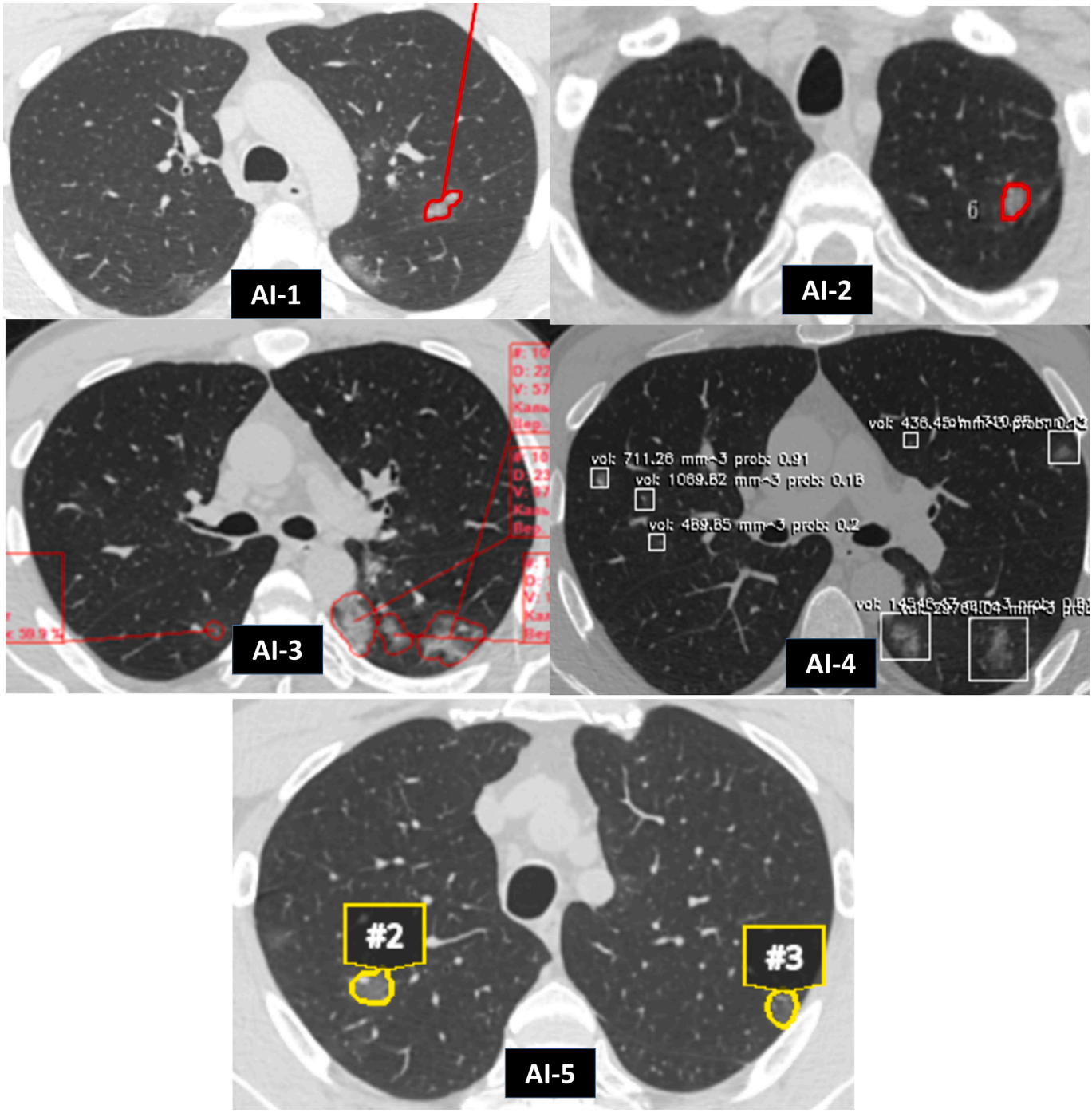


Fig. 3. Case #2 from Dataset2. Case without the target pathology. All AI models identified the changes in the lung due to the COVID-19 as a suspicious lung nodule.

validation in their study design. Without an additional fine-tuning (i.e., the threshold update) of the AI model to the epidemiological situation, we risk deploying AI models which will not have practical value due to the high false positive rate. Our data also showed an increase in net benefit when AI models were fine-tuned, indicating that the model has clinical utility, as the benefits outweigh the harms. This criterion can be used to assess the value of fine-tuning.

In our study, we did not reveal replicability of machine learning for detection of lung nodules in the presence of the population shift due to COVID-19, since AI models showed different trends of changes of their accuracy metrics.

This study has limitations. First, we did not access the influence of increasing the dataset shift on the prediction accuracy of the datasets. However, despite using only two dataset shift points, we have observed the change of performance of the majority of included AI models. Second, the two datasets had slightly different size and balance in terms of norm/pathology cases, female/male proportion, and age distribution. The difference could introduce some unaccounted confounding variables when comparing the AI performance between dataset 1 and 2. Nevertheless, the difference between the datasets is irrelevant when comparing the AI diagnostic accuracy before and after threshold update on the same dataset.

The implications for practice derived from our study reside in the necessity of target pathology datasets update mirroring the change of epidemiological situation. It is especially important when the alternative condition has overlapping radiology features with the target disease.

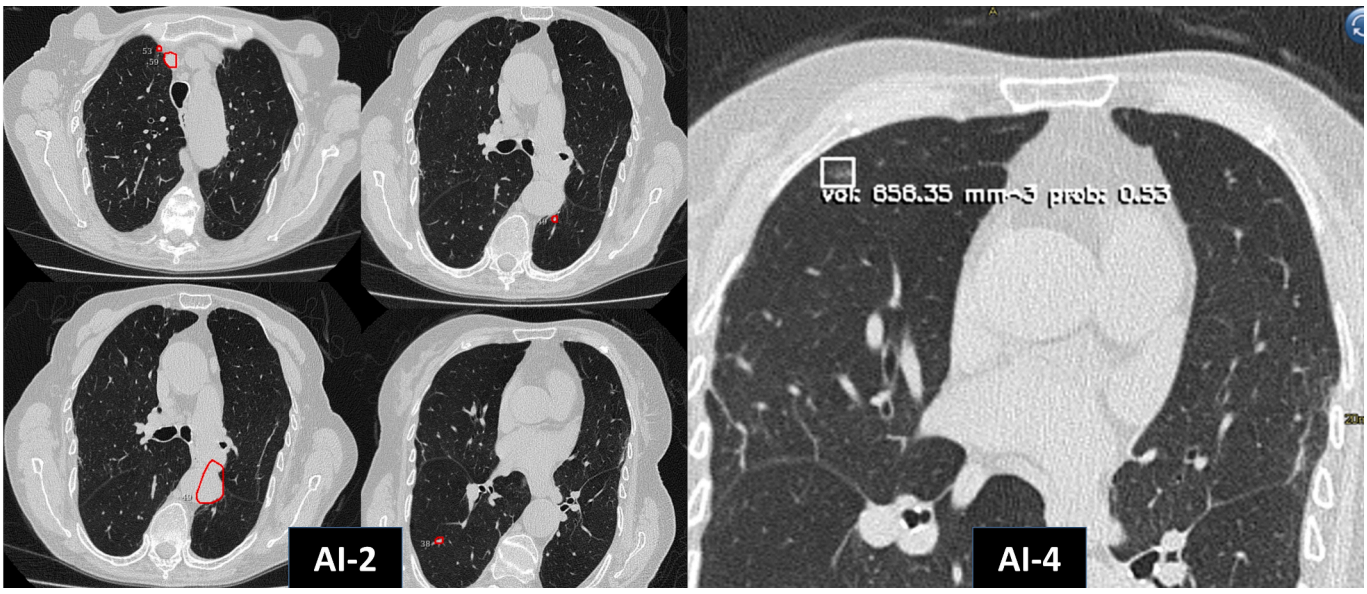


Fig. 4. Case #3 from Dataset1. Case without the target pathology. AI-2: Vessels are marked as the target pathology. AI-4: A false-positive finding of an area of stranded fibrosis. AI-1,3,5 correctly rated the study as "no target pathology".

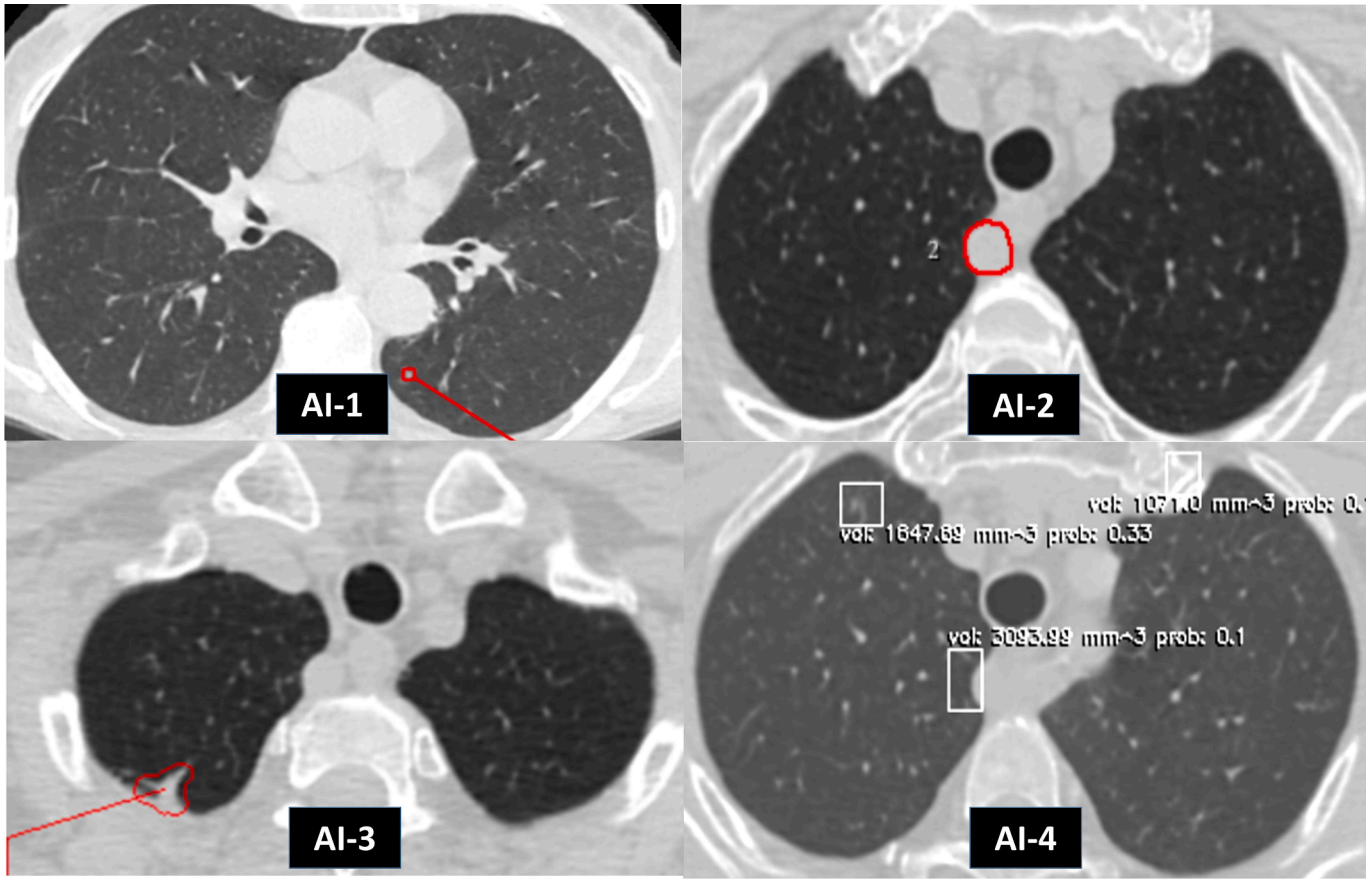


Fig. 5. Case #4 from Dataset1. Case without the target pathology. AI-1: A false positive finding (4 mm nodule). AI-2: A false positive finding (subclavian artery). AI-3: A false positive finding (adhesion). AI-4: A false positive findings (fibrosis, bones, artery fragment).

Such a strategy makes it possible to decrease a possible number of false-positive findings, thereby reducing the burden on the healthcare system caused by unnecessary additional testing of patients with suspected abnormalities.

5. Conclusions

AI models for the diagnosis of pulmonary nodules showed comparable levels of AUROC when analyzing studies acquired during the Covid-19 pandemic compared to the dataset that did not contain Covid-

Table 5
Impact of the threshold update on the AI models' diagnostic accuracy metrics stability for the dataset2 with respect to the dataset1. Bold font indicates statistically significant difference. Regular font shows statistically insignificant difference.

Diagnostic accuracy metrics	Dataset1 vs Dataset2 (with the threshold update)					Dataset1 vs Dataset2 (without the threshold update)				
	AI-1	AI-2	AI-3	AI-4	AI-5	AI-1	AI-2	AI-3	AI-4	AI-5
Sensitivity	0,07*	**−0,29****	−0,05	0,1	−0,02	**0,27**	−0,09	−0,05	0	−0,21
Specificity	**0,2**	0,12	**0,27**	0,03	0,06	0,14	−0,04	**0,34**	**0,26**	**0,3**

*decrease of the value compared to the dataset1 – positive value.
**increase of the value compared to the dataset1 – negative value.

19 related signs. However, the specificity of AI models decreased due to a large number of false positives amid patients with the lung parenchyma changes due to the coronavirus infection if the threshold was not updated during the pandemic. To ensure the optimal diagnostic accuracy metrics during deployment of AI, it is necessary to fine-tune the threshold of the AI models depending on the epidemiological situation and patient population characteristics.

Funding

This paper was prepared by a group of authors as a part of the research and development effort titled "Evidence-based methodologies for sustainable development of artificial intelligence in medical imaging", (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.

Summary table

1. We found that the COVID-19 pandemic has changed the quality of the model: sensitivity, specificity, ROC AUC can either increase or decrease depending on the model.
2. To ensure the optimal diagnostic accuracy metrics during deployment of AI, it is necessary to fine-tune the threshold of the AI models depending on the epidemiological situation and patient population characteristics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge Yury Kirpichev, for developing the web tool for comprehensive ROC analysis and Daria Kokina, Tatiana Logunova, Victor Gombolevskiy for the assistance in the clinical evaluation of AI performance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105190.

References

[1] Li J, Zhu G, Hua C, Feng M, BasheerBennamoun, Li P, et al. A Systematic Collection of Medical Image Datasets for Deep Learning. 2021;
[2] I. Kandel, M. Castelli, How deeply to fine-tune a convolutional neural network: A case study using a histopathology dataset, Appl Sci. 10 (10) (2020 May) 3359.
[3] D. Nguyen, F. Kay, J. Tan, Y. Yan, Y.S. Ng, P. Iyengar, et al., Deep Learning-Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability. Front, Artif Intell. 4 (2021).
[4] H.J. Jang, I.H. Song, S.H. Lee, Generalizability of deep learning system for the pathologic diagnosis of various cancers, Appl Sci. 11 (2) (2021) 808.
[5] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, K. Lekadir, Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study, Artif. Intell. Med. 132 (2022) 102386.
[6] T. Eche, L.H. Schwartz, F.Z. Mokrane, L. Dercle, Toward generalizability in the deployment of artificial intelligence in radiology: Role of computation stress testing to overcome underspecification, Radiol Artif Intell. 3 (6) (2021).
[7] D. Khemasuwan, J.S. Sorensen, H.G. Colt, Artificial intelligence in pulmonary medicine: computer vision, predictive model and COVID-19, Eur Respir Rev 29 (157) (2020) 200181.
[8] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat Med 25 (1) (2019) 44–56.
[9] Bhatnagar S, Cotton T, Brundage M, Avin S, Clark J, Toner H, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation Authors are listed in order of contribution Design Direction. 2018;
[10] S. Goncalves, P.-C. Fong, M. Blokhina, Artificial intelligence for early diagnosis of lung cancer through incidental nodule detection in low- and middle-income countries-acceleration during the COVID-19 pandemic but here to stay, Am J Cancer Res [Internet]. (2022) [cited 2023 Jan 13];12(1):1. Available from: /pmc/articles/PMC8822269/.
[11] B. Botz, T. Radswiki, Fleischner Society pulmonary nodule recommendations, Radiopaedia.org. Radiopaedia.org (2011).
[12] Abuladze LR, Blokhin IA, Gonchar AP, Suchilova MM, Vladzymyrskyy A V., Gombolevskiy VA, et al. CT imaging of HIV-associated pulmonary disorders in COVID-19 pandemic. Clin Imaging [Internet]. 2023 Mar;95:97–106. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0899707123000098.
[13] Y.-J. Zhang, W.-J. Yang, D. Liu, Y.-Q. Cao, Y.-Y. Zheng, Y.-C. Han, R.-S. Jin, Y. u. Han, X.-Y. Wang, A.-S. Pan, J.-Y. Dai, Q.-F. Sun, F.-Q. Zhao, Q.-Y. Yang, J.-H. Zhang, S.-J. Liu, Q. Da, W. Guo, C.-Q. Li, W.-T. Zhang, H. Wu, X.-S. Chen, A.-Q. Ji, J. Xiang, K. Chen, X.-J. Feng, X.-F. Zhang, Q.-Q. Cao, L.e. Qin, J. Li, M. Zhou, Y. Lu, C.-F. Wang, F.-H. Yan, H.-C. Li, J.-M. Qu, COVID-19 and early-stage lung cancer both featuring ground-glass opacities: a propensity score-matched study, Transl Lung Cancer Res. 9 (4) (2020) 1516–1527.
[14] P.J. Mazzone, M.K. Gould, D.A. Arenberg, A.C. Chen, H.K. Choi, F.C. Detterbeck, F. Farjah, K.M. Fong, J.M. Iaccarino, S.M. Janes, J.P. Kanne, E.A. Kazerooni, H. MacMahon, D.P. Naidich, C.A. Powell, S. Raoof, M.P. Rivera, N.T. Tanner, L. K. Tanoue, A. Tremblay, A. Vachani, C.S. White, R.S. Wiener, G.A. Silvestri, Management of lung nodules and lung cancer screening during the COVID-19 Pandemic: CHEST expert panel report, Chest 158 (1) (2020) 406–415.
[15] A. Sharma, I. Ahmad Farouk, S.K. Lal, COVID-19: A review on the novel coronavirus disease evolution, Transmission, Detection Control and Prevention. Viruses. 13 (2) (2021 Feb) 202.
[16] B. Shirokikh, A. Shevtsov, A. Dalechina, E. Krivov, V. Kostjuchenko, A. Golanov, et al., Accelerating 3D medical image segmentation by adaptive small-scale target localization, J Imaging. 7 (2021) 35.
[17] CELSUS Patent №2019610585. 2019.
[18] "Software complex for automatic processing of radiological images "RADLogics Platform" Patent №2020667078. 2020.
[19] Yudin SI. Diagnostic decision support system "AI Diagnostic" Patent №2021615516. 2021.
[20] P.V. Gavrilov, P.G. Roitberg, D.S. Blinov, M.G. Goldin, E.V. Blinova, V.S. Leontiev, I.G. Kamishanskaya, A.A. Dorofeev, V.M. Cheremisin, Y.A. Novokhat'ko, E. V. Sushkov, D.O. Shmatok, A.I. Sokolov, Artificial intelligence-based algorithms in detection and 3D reconstruction of lung nodules on chest computed tomography scans, Oper. khir. 5 (3) (2021) 15.
[21] I.S. Drokin, E.V. Ericheva, O.L. Buhvalov, P.S. Pilyus, T.S. Malygina, V.E. Sinicyn, Experience in the development and implementation of a system for searching for oncological formations using artificial intelligence on the example of X-ray computed tomography of the lungs, Physician Inf Technol. (3) (2019).
[22] J. Bueno, L. Landeras, J.H. Chung, Updated fleischner society guidelines for managing incidental pulmonary nodules: Common questions and challenging scenarios, Radiographics 38 (5) (2018) 1337–1350.
[23] M. Goncharov, M. Pisov, A. Shevtsov, B. Shirokikh, A. Kurmukov, I. Blokhin, V. Chernina, A. Solovev, V. Gombolevskiy, S. Morozov, M. Belyaev, CT-Based COVID-19 triage: Deep multitask learning improves joint identification and severity quantification, Med Image Anal. 71 (2021) 102054.
[24] S.G. Armato, G. McLennan, L. Bidaut, M.F. McNitt-Gray, C.R. Meyer, A.P. Reeves, B. Zhao, D.R. Aberle, C.I. Henschke, E.A. Hoffman, E.A. Kazerooni, H. MacMahon, E.J.R. van Beek, D. Yankelevitz, A.M. Biancardi, P.H. Bland, M.S. Brown, R. M. Engelmann, G.E. Laderach, D. Max, R.C. Pais, D.-Y. Qing, R.Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G.W. Gladish, C.M. Jude, R.

F. Munden, I. Petkovska, L.E. Quint, L.H. Schwartz, B. Sundaram, L.E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallam, M.D. Heath, M.H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B.Y. Croft, L.P. Clarke, The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans, Med Phys. 38 (2) (2011) 915–931.

[25] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun. 5 (1) (2014).

[26] Liu L, Jiang H, He P, Chen W, Liu X, Gao J, et al. On the Variance of the Adaptive Learning Rate and Beyond. arXiv; 2019.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal Loss for Dense Object Detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007.

[28] A.A.A. Setio, A. Traverso, T. de Bel, M.S.N. Berens, C.V.D. Bogaard, P. Cerello, H. Chen, Q.i. Dou, M.E. Fantacci, B. Geurts, R.V.D. Gugten, P.A. Heng, B. Jansen, M.M.J. de Kaste, V. Kotov, J.-H. Lin, J.T.M.C. Manders, A. Sóñora-Mengana, J. C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C.M. Schaefer-Prokop, E.T. Scholten, L. Scholten, M.M. Snoeren, E.L. Torres, J. Vandemeulebroucke, N. Walasek, G.C.A. Zuidhof, B.V. Ginneken, C. Jacobs, Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge, Med Image Anal. 42 (2017) 1–13.

[29] J. Pedrosa, G. Aresta, C. Ferreira, G. Atwal, H.A. Phoulady, X. Chen, R. Chen, J. Li, L. Wang, A. Galdran, H. Bouchachia, K.C. Kaluva, K. Vaidhya, A. Chunduru, S. Tarai, S.P.P. Nadimpalli, S. Vaidya, I. Kim, A. Rassadin, Z. Tian, Z. Sun, Y. Jia, X. Men, I. Ramos, A. Cunha, A. Campilho, LNDb challenge on automatic lung cancer patient management, Med Image Anal. 70 (2021) 102027.

[30] Morosov SP, Kulberg NS, Gombolevskiy, V.A. Ledikhova NV, Sokolina SI, Vladzimirskiy AV, Bardin AS. Tagged results of lung computed tomography scans. RU 2018620500, 2018.

[31] S.P. Morozov, A.E. Andreychenko, I.A. Blokhin, P.B. Gelezhe, A.P. Gonchar, A. E. Nikolaev, N.A. Pavlov, V.Y. Chernina, V.A. Gombolevskiy, MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic, Digit Diagnostics. 1 (1) (2020) 49–59.

[32] N.S. Polishchuk, N.N. Vetsheva, S.P. Kosarin, S.P. Morozov, Kuz'Mina ES., Unified radiological information service as a key element of organizational and methodical work of research and practical center of medical radiology, Radiol - Pract. 1 (2018) 6–17. In Russ.

[33] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med Decis Mak. 26 (6) (2006) 565–574.

[34] H.M. Whitney, K. Drukker, M.L. Giger, Performance metric curve analysis framework to assess impact of the decision variable threshold, disease prevalence, and dataset variability in two-class classification, J Med Imaging. 9 (03) (2022) 1–15.

[35] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1) (1950) 32–35.

[36] M.D. Ruopp, N.J. Perkins, B.W. Whitcomb, E.F. Schisterman, Youden index and optimal cut-point estimated from observations affected by a lower limit of detection, Biom. J. 50 (3) (2008) 419–430.

[37] D.W. Hosmer, S. Lemeshow, Applied logistic regression, 2nd ed., Wiley, N.-Y, 2000, p. 375 p..

[38] X. Sun, W. Xu, Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves, IEEE Signal Process Lett. 21 (11) (2014) 1389–1393.

[39] L.D. Brown, T.T. Cai, A. DasGupta, Interval estimation for a binomial proportion, Stat Sci. 16 (2) (2001 May) 101–133.

[40] P.J. Mazzone, M.K. Gould, D.A. Arenberg, A.C. Chen, H.K. Choi, F.C. Detterbeck, F. Farjah, K.M. Fong, J.M. Iaccarino, S.M. Janes, J.P. Kanne, E.A. Kazerooni, H. MacMahon, D.P. Naidich, C.A. Powell, S. Raoof, M.P. Rivera, N.T. Tanner, L. K. Tanoue, A. Tremblay, A. Vachani, C.S. White, R.S. Wiener, G.A. Silvestri, Management of lung nodules and lung cancer screening during the COVID-19 pandemic: CHEST expert panel report, Radiol Imaging Cancer. 2 (3) (2020) e204013.

[41] Morozov SP, Vladzymyrskyy A V., Klyashtornyy VG, Andreychenko AE, Kulberg NS, Gombolevsky VA, et al. Clinical acceptance of software based on artificial intelligence technologies (radiology). 2019 Aug 1 [cited 2022 Apr 19]; Available from: http://arxiv.org/abs/1908.00381.

[42] D.M. Ibrahim, N.M. Elshennawy, A.M. Sarhan, Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases, Comput Biol Med. 132 (2021).

[43] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2019 Oct) 1–9.