Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Multidisciplinary Evaluation of an AI-Based Pneumothorax Detection Model: Clinical Comparison with Physicians in Edge and Cloud Environments

Ismail Dal[1], Hasan Burak Kaya[2]

[1]Department of Thoracic Surgery, Kastamonu University, Kastamonu, Turkey; [2]Department of Emergency Medicine, Kastamonu University, Kastamonu, Turkey

Correspondence: Ismail Dal, Department of Thoracic Surgery, Kastamonu University, Cankat St. Number: 4, City Center, Kastamonu, Turkey, Tel +903662141053, Email idal@kastamonu.edu.tr

Video Abstract

Point your SmartPhone at the code above. If you have a QR code reader the video abstract will appear. Or use: https://youtu.be/vZKVa7CMPQ4.

Background: Accurate and timely detection of pneumothorax on chest radiographs is critical in emergency and critical care settings. While subtle cases remain challenging for clinicians, artificial intelligence (AI) offers promise as a diagnostic aid. This retrospective diagnostic accuracy study evaluates a deep learning model developed using Google Cloud Vertex AI for pneumothorax detection on chest X-rays.

Methods: A total of 152 anonymized frontal chest radiographs (76 pneumothorax, 76 normal), confirmed by computed tomography (CT), were collected from a single center between 2023 and 2024. The median patient age was 50 years (range: 18–95), with 67.1% male. The AI model was trained using AutoML Vision and evaluated in both cloud and edge deployment environments. Diagnostic accuracy metrics—including sensitivity, specificity, and F1 score—were compared with those of 15 physicians from four specialties (general practice, emergency medicine, thoracic surgery, radiology), stratified by experience level. Subgroup analysis focused on minimal pneumothorax cases. Confidence intervals were calculated using the Wilson method.

Results: In cloud deployment, the AI model achieved an overall diagnostic accuracy of 0.95 (95% CI: 0.83, 0.99), sensitivity of 1.00 (95% CI: 0.83, 1.00), specificity of 0.89 (95% CI: 0.69, 0.97), and F1 score of 0.95 (95% CI: 0.86, 1.00). Comparable performance was observed in edge mode. The model outperformed junior clinicians and matched or exceeded senior physicians, particularly in detecting minimal pneumothoraces, where AI sensitivity reached 0.93 (95% CI: 0.79, 0.97) compared to 0.55 (95% CI: 0.38, 0.69) – 0.84 (95% CI: 0.69, 0.92) among human readers.

Conclusion: The Google Cloud Vertex AI model demonstrates high diagnostic performance for pneumothorax detection, including subtle cases. Its consistent accuracy across edge and cloud settings supports its integration as a second reader or triage tool in diverse clinical workflows, especially in acute care or resource-limited environments.

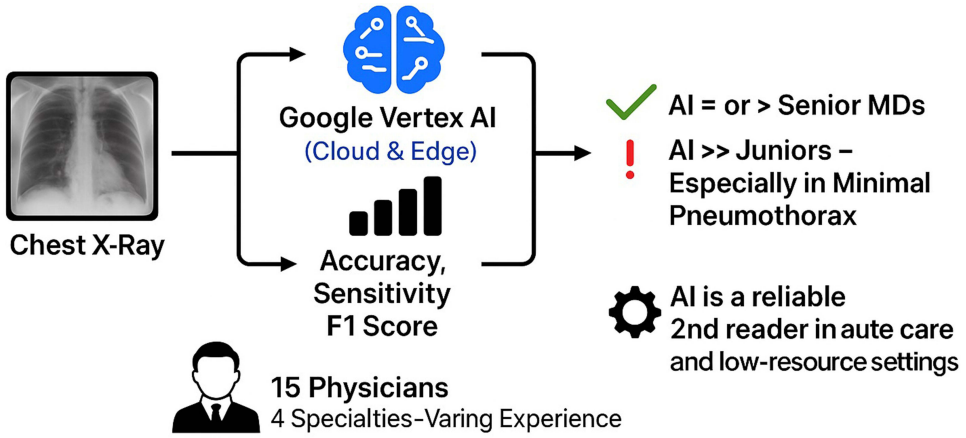
Keywords: pneumothorax diagnosis, artificial intelligence, cloud computing, clinical decision support systems, multidisciplinary communication

Introduction

Pneumothorax is a potentially life-threatening condition that requires prompt recognition and management. Chest radiography (CXR) remains the most common first-line imaging tool for suspected pneumothorax due to its wide availability, but accurate interpretation of CXR findings is notoriously challenging.[1] Subtle radiographic signs – such as a faint visceral pleural line or slight asymmetry in lung translucency – may be the only indicators of a small (minimal) pneumothorax, and these can be easily overlooked.[2] Even critical findings like pneumothoraces are frequently missed on initial CXR readings, which can lead to delayed diagnosis and patient harm.[1] This problem is exacerbated among less-experienced clinicians; junior doctors often feel

Received: 19 April 2025
Accepted: 14 July 2025
Published: 17 July 2025

Graphical Abstract



uncertain interpreting CXRs and are more likely to misidentify or miss abnormalities.[3] Consequently, minimal pneumothoraces in particular are at high risk of being underdiagnosed by frontline physicians, contributing to preventable delays in treatment.

Recent advances in artificial intelligence (AI) offer a potential solution to improve pneumothorax detection on imaging. Deep learning models have demonstrated high performance in recognizing thoracic abnormalities on radiographs, prompting exploration of AI as a diagnostic aid in clinical radiology. Notably, AI algorithms can be trained to identify the telltale features of pneumothorax (eg pleural line discontinuity or absence of lung markings) on chest X-rays and alert clinicians to these findings. Emerging evidence suggests that such models can achieve accuracy comparable to human experts: for example, a recent study reported an area under the ROC curve of approximately 0.98 for an AI system detecting pneumothoraces on CXRs (with sensitivity around 94%).[4] However, the diagnostic accuracy of CXR itself remains limited. According to a recent meta-analysis,[5] the pooled sensitivity of chest radiography for pneumothorax detection is approximately 0.65 (95% CI: 0.54–0.76), while the specificity is around 0.99 (95% CI: 0.98–1.00), highlighting both its diagnostic limitations and its high false-negative potential in subtle cases. The use of AI in the CXR workflow could therefore facilitate earlier identification of pneumothorax and expedite management.[4] Furthermore, incorporating AI as a "second reader" has been shown to enhance physician performance – AI assistance in chest X-ray interpretation significantly increases radiologists' sensitivity for detecting abnormalities, including pneumothorax, across varying levels of expertise.[6] These findings underscore the promise of deep learning tools to reduce diagnostic errors in pneumothorax, particularly for subtle cases that might evade an unassisted human eye.

In parallel with advancements in AI, there have been rapid developments in how such models are deployed in clinical practice. Cloud computing and edge computing represent two complementary approaches for delivering AI-driven diagnostic support. Cloud-based platforms (such as Google Cloud Vertex AI) allow complex models to be hosted on powerful remote servers, enabling scalable computation and easy updates or maintenance of the algorithm.[7] This means a chest X-ray can be uploaded and analyzed by the model in the cloud, with results promptly returned to the clinician. Conversely, edge deployment brings the AI model closer to the point of care by running it on local hardware (eg a hospital workstation or mobile device) without requiring continuous internet connectivity.[8] Edge AI can offer near-instantaneous analysis on-site, reducing latency in urgent settings and ensuring that even if network access is interrupted, the diagnostic tool remains available. Beyond Google's platform, other widely used AI solutions in medical imaging include Amazon SageMaker (Amazon Web Services, Seattle, WA, USA), Microsoft Azure Custom Vision (Microsoft Corporation, Redmond, WA, USA), and open-source frameworks such as MONAI (Medical Open Network for AI) (developed by Project MONAI, an initiative by NVIDIA and King's College London) and NVIDIA Clara (NVIDIA Corporation, Santa Clara, CA, USA). These platforms support model training, deployment, and clinical integration across

various imaging modalities, offering flexible alternatives for institutions with diverse computational needs. In the context of pneumothorax detection, a cloud approach might facilitate integration across multiple hospital sites or allow leveraging large datasets for model improvement, while an edge approach could provide real-time interpretation in an emergency department or rural clinic. Harnessing both cloud and edge computing thus broadens the potential utility of AI in diagnostic imaging, combining the benefits of centralized intelligence with point-of-care responsiveness.

Given the clinical importance of promptly detecting pneumothorax and the known limitations of human interpretation (especially for subtle cases), it is critical to assess how an AI solution compares with human providers. Most prior studies have evaluated AI performance against radiologist readings; however, in practice, chest X-rays are interpreted not only by radiologists but also by physicians of various specialties – including general practitioners, emergency physicians, and surgeons – who may have variable radiographic expertise. Misdiagnosis or oversight of a minimal pneumothorax by junior or non-radiology clinicians is a well-recognized issue that can occur during initial patient evaluation. An AI tool that performs on par with or better than these clinicians could serve as a valuable safety net, improving diagnostic consistency across different care settings. Therefore, in this study we sought to directly compare a deep learning model's pneumothorax detection performance with that of physicians from multiple specialties (general practice, emergency medicine, thoracic surgery, and radiology), reflecting the diverse group of clinicians involved in real-world CXR interpretation.

The focus on minimal pneumothorax is a deliberate emphasis of our work. Small pneumothoraces often present the greatest diagnostic challenge – they are frequently radiographically occult or subtle, yet their early identification is clinically significant. Even a tiny apical pneumothorax, if missed, could enlarge or precipitate symptoms, and in certain scenarios (such as positive-pressure ventilation or air travel) a small pneumothorax might rapidly evolve into a more serious condition. By concentrating on cases with minimal pneumothorax (which comprised 48% of our dataset), we aim to test the AI model under the most demanding conditions and evaluate its potential benefit where physician oversight is most likely. All study procedures were conducted at an academic research hospital, with appropriate ethical approval (Institutional Review Board protocol No. 2024-KAEK-5).

Aim of the study: This research evaluates the diagnostic performance of a deep learning model developed using Google Cloud Vertex AI for detecting pneumothorax on chest X-rays, tested in both cloud and edge deployment modes. We compare the model's performance to that of physicians from multiple specialties to determine whether the AI can match or augment human diagnostic accuracy. The primary focus is on the detection of minimal pneumothoraces, with the overall goal of assessing the model's utility in improving early pneumothorax diagnosis in clinical practice.

Materials and Methods

Study Design and Data Collection

This retrospective diagnostic accuracy study was conducted at the emergency department of a single academic research hospital between 2023 and 2024. All chest radiographs obtained during this period were reviewed to identify cases of pneumothorax. In total, 51,326 chest X-ray images were screened, from which 76 cases of pneumothorax were confirmed based on corresponding chest computed tomography (CT) findings (CT served as the gold-standard reference for diagnosis). An equal number of 76 control radiographs with no pneumothorax (normal findings on CT) were randomly selected from the same timeframe to serve as the comparison group. These 152 frontal chest radiographs (76 pneumothorax, 76 normal) comprised the dataset for the study (Figure 1). The analysis was conducted on a per-patient basis. In one bilateral pneumothorax case (n = 1), the image was labeled as positive for pneumothorax overall. All images and clinical data were anonymized prior to analysis. The study protocol was approved by the institutional ethics committee (approval number 2024-KAEK-5).

Deep Learning Model Development

A custom deep learning model for pneumothorax detection was developed using a cloud-based automated machine learning platform (Google Cloud Vertex AI AutoML Vision, Google LLC, Mountain View, CA, USA). The 152 chest X-ray images, labeled according to the presence or absence of pneumothorax (as determined by CT), were uploaded to the AutoML Vision system for training. The platform automatically split the data into training, validation, and test sets in

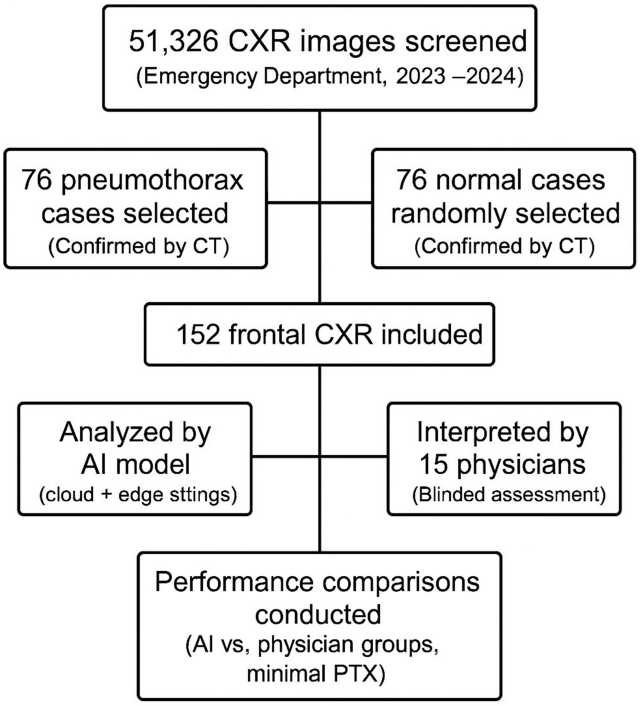


Figure 1 Study Flowchart for AI and Physician Evaluation of Pneumothorax on CXR.

a 65%/10%/25% ratio, ensuring a balanced representation of pneumothorax and normal cases in each subset. Standard image preprocessing was applied to optimize model input quality. Specifically, each radiograph was converted to JPEG format and preprocessed as follows:

- Pixel intensity values were normalized to standardize image contrast.
- Exported at a resolution of 300 DPI prior to resizing.
- Each image was resized to 224×224 pixels to match the input dimensions required by the neural network.
- Uniform JPEG compression settings were used for all images to reduce file size while preserving diagnostic content.

The AutoML platform leveraged transfer learning with a pre-existing convolutional neural network (CNN) backbone to build the classification model. This approach enabled effective training on a limited dataset by fine-tuning a model pretrained on a large general image corpus. Although the exact architecture is automatically determined by the platform, it typically uses efficient, high-performance models (eg, MobileNetV2 or EfficientNet family) as the starting point. During training, the pre-trained CNN weights were fine-tuned on our chest X-ray dataset to recognize the radiographic features of pneumothorax. The training process included automated hyperparameter tuning and early stopping based on performance on the validation set.

Model Deployment and Evaluation

After model training, the best-performing trained model was exported for deployment. We obtained a TensorFlow-compatible model artifact (via Vertex AI's export function) to enable inference in different environments. The model's performance was first evaluated in the cloud environment using the held-out test set (constituting 25% of the data). For each image in the test set, the model produced a binary prediction (pneumothorax present vs absent), which was then compared against the CT-confirmed ground truth to calculate diagnostic performance metrics (eg, accuracy, sensitivity, specificity). In addition to cloud-based evaluation, the exported model was deployed in an edge computing setting to assess its feasibility for on-site use. The same trained model was loaded onto a local system (simulating an edge device),

and inference was run on the chest X-ray images to ensure that the predictions were consistent with those obtained in the cloud. This two-mode deployment (cloud and edge) allowed verification of the model's portability and real-time performance outside the training environment.

Physician Reader Evaluation

Parallel to the AI evaluation, a reader study was performed with human physicians to assess diagnostic performance on the same set of chest radiographs. A panel of physicians from multiple specialties (including radiology and emergency medicine) independently reviewed the 152 chest X-ray images (Supplementary Figure 1). All physicians were blinded to any clinical information and to the CT findings for each case, so their judgments were based solely on the chest X-ray appearance. The images were presented in a randomized order to avoid recall or sequence bias. An electronic survey was created using Google Forms (Google LLC, Mountain View, CA, USA) to facilitate image distribution and response collection. Each radiograph was embedded in the form as a question, and physicians were asked to indicate whether a pneumothorax was present or absent. They recorded their diagnoses for each case without knowing the proportion of positives or the correct answers. The use of an online questionnaire ensured a standardized viewing format and allowed data to be collected securely.

Each physician's responses were recorded and later analyzed against the ground truth (CT diagnosis). The performance of the physicians (eg, their sensitivity and specificity in detecting pneumothorax on X-ray) could then be compared to that of the AI model on a case-by-case basis. This blinded human reader assessment provided a reference for evaluating the clinical relevance of the AI model's accuracy.

Statistical Analysis

For the analysis of demographic data, categorical variables were summarized as frequencies and percentages, while continuous variables were reported as medians with minimum and maximum values due to non-normal distribution. The distribution of continuous variables was assessed visually and, when necessary, using the Shapiro–Wilk test. Group comparisons for categorical variables were performed using the Chi-square test or Fisher's exact test, while the Mann–Whitney *U*-test was used for continuous variables.

Sensitivity, specificity, diagnostic accuracy, and F1 scores were calculated for each reader and the AI models. Ninety-five percent confidence intervals for sensitivity, specificity, and diagnostic accuracy were computed using the Wilson score method, which is appropriate for proportion estimates, particularly in cases with small sample sizes or extreme values. Since the F1 score is a composite metric derived from both precision and recall (sensitivity) and does not follow a closed-form distribution, its 95% confidence intervals were estimated using a non-parametric bootstrap resampling method with 1000 iterations. This approach approximates the sampling distribution of the F1 score by repeatedly resampling the original data with replacement. The analysis was conducted on a per-patient basis, as noted above. All statistical computations were performed using MedCalc Statistical Software version 20.218 (MedCalc Software Ltd, Ostend, Belgium).

Results

Patient Characteristics

A total of 152 patients were included in the final analysis, comprising 76 with pneumothorax confirmed by computed tomography (CT) and 76 healthy controls. The median age was 50 years (range: 18–95), with 67.1% male (n = 104) and 32.9% female (n = 48). Among pneumothorax cases, 52.6% were right-sided, 46% left-sided, and 1.4% bilateral. The distribution of pneumothorax size was nearly equal, with 51.3% classified as large and 48.7% as small (minimal) (Table 1). The classification of pneumothorax sizes was performed according to the ACCP guidelines.[9]

Performance of the AI Models

The Google Cloud Vertex AI-based deep learning model was evaluated in both cloud and edge deployment modes. In cloud deployment, the model achieved a sensitivity of 1.00 (95% CI: 0.83–1.00), specificity of 0.89 (95% CI: 0.69–0.97), diagnostic accuracy of 0.95 (95% CI: 0.83–0.99), and F1 score of 0.95 (95% CI: 0.86–1.00). In edge deployment, the

Table 1 Demographic Data of the Patients

Gender (n = 152)	
Male, n (%)	104 (67.1%)
Female, n (%)	48 (32.9%)
Age (n= 152), median (min, max)	50 (18, 95)
Pneumothorax Side (n=76)	
Right, n (%)	40 (52.6%)
Left, n (%)	35 (%46)
Bilateral, n (%)	1 (1.4%)
Pneumothorax Dimension (n=76)	
Large, n (%)	39 (51.3%)
Small, n (%)	37 (48.7%)

model maintained high performance, with a sensitivity and specificity of 0.95 (95% CI: 0.75–0.99), diagnostic accuracy of 0.95 (95% CI: 0.83–0.99), and F1 score of 0.95 (95% CI: 0.86–1.00).

These results are visualized in Figure 2, which demonstrates the near-identical performance of the AI model in both deployment environments. Edge deployment successfully preserved diagnostic precision while enabling localized computation, supporting the model's feasibility for real-time use in emergency or resource-limited settings.

Diagnostic Performance of Physicians

Fifteen physicians from various specialties—including general practitioners (GP), emergency medicine specialists (EMS), thoracic surgeons (TS), and radiologists—participated in the blinded reader study. Their individual diagnostic performance metrics are summarized in Table 2. Sensitivity values ranged from 0.36 (95% CI: 0.26–0.47) for GP1 to 0.97 (95% CI: 0.91–0.99) for TS3, specificity from 0.76 (95% CI: 0.66–0.84) to 1.00 (95% CI: 0.95–1.00), and diagnostic accuracy from 0.68 (95% CI: 0.60–0.75) to 0.99 (95% CI: 0.95–1.00). The highest overall performance was observed in a thoracic surgeon with 5–10 years of experience (TS3), who achieved an F1 score of 0.99 (95% CI: 0.96–1.00). Among the radiologists, the top performer reached a diagnostic accuracy of 0.94 (95% CI: 0.89–0.97) and an F1 score of 0.94 (95% CI: 0.90–0.98), as shown in Table 2.

Figure 3 illustrates the variability in diagnostic accuracy across professional titles. General practitioners showed lower performance overall compared to emergency specialists, thoracic surgeons, and radiologists. The AI model's diagnostic accuracy surpassed the average accuracy across all physician groups, and in some cases, even outperformed individual experienced radiologists.

Physician Performance by Experience Level

To further explore diagnostic variability, physicians were stratified into junior (<5 years experience) and senior (≥5 years experience) groups based on self-reported professional tenure (as shown in Table 2). A total of 6 general practitioners (GPs) with <5 years of experience were categorized as junior doctors. Among the senior group were 3 emergency medicine specialists (EMS), 3 thoracic surgeons (TS), and 3 radiologists, each with ≥5 years of experience.

Junior doctors (n = 6) demonstrated considerable variability in sensitivity, ranging from 0.36 to 0.67 (mean: 0.53), with an average diagnostic accuracy of 0.74 and a mean F1 score of 0.66. Despite consistently high specificity (mean: 0.94), their relatively low sensitivity, particularly in detecting minimal pneumothorax, suggests a tendency toward underdiagnosis in subtle cases.

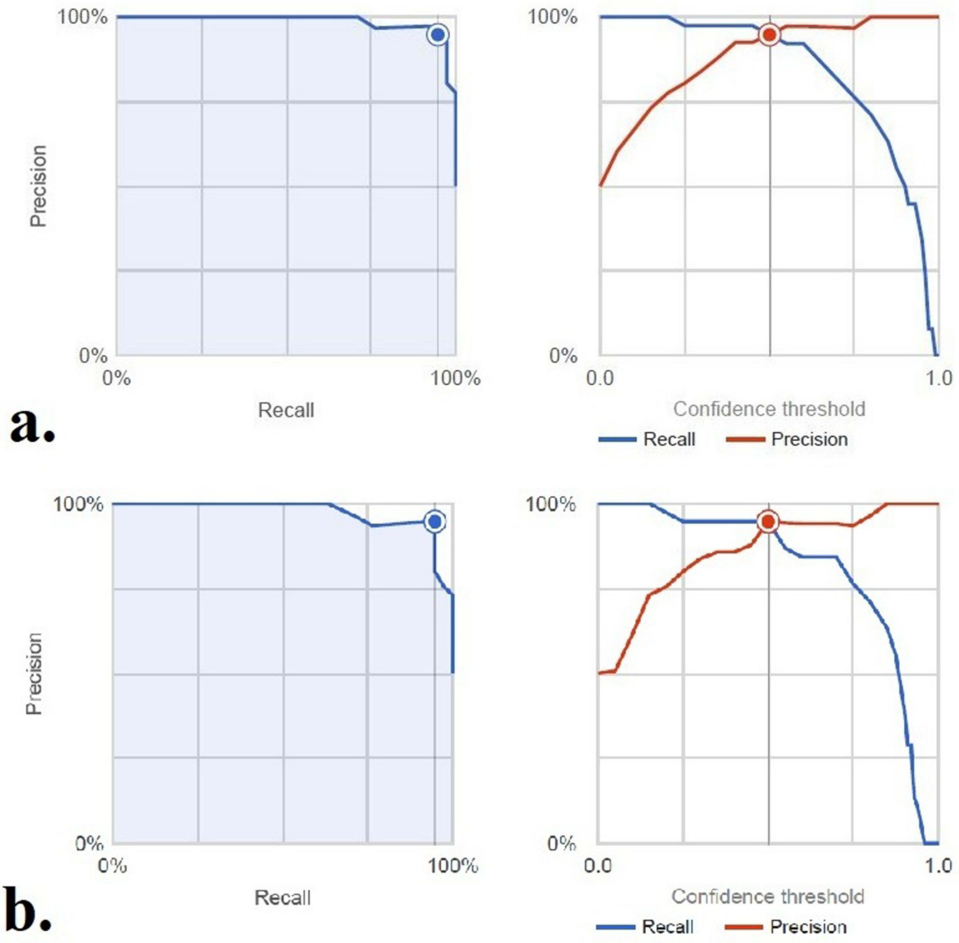


Figure 2 Overall Model Performance Metrics ((**a**) Cloud Model, (**b**) Edge Model).

In contrast, senior physicians (n = 9), including emergency medicine specialists, thoracic surgeons, and radiologists, exhibited markedly better diagnostic performance. Their average sensitivity was 0.80, specificity 0.96, diagnostic accuracy 0.87, and mean F1 score 0.84. The top-performing reader, a thoracic surgeon with 5–10 years of experience, achieved near-perfect results with a sensitivity of 0.97, specificity of 1.00, accuracy of 0.99, and an F1 score of 0.99.

Comparatively, the Google Cloud Vertex AI model in edge deployment mode performed on par with or better than the junior doctor group across all key metrics. It achieved a sensitivity and specificity of 0.95, diagnostic accuracy of 0.95, and an

Table 2 Comparison of Specificity, Sensitivity and Diagnostic Accuracy Rates of Doctors and Artificial Intelligence Applications

Reader	Experience	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	F1 Score (95% CI)
GP 1	<2 years	0.36 (0.26, 0.47)	1.00 (0.95, 1.00)	0.68 (0.60, 0.75)	0.52 (0.40, 0.65)
GP 2	2–5 years	0.43 (0.33, 0.55)	1.00 (0.95, 1.00)	0.72 (0.64, 0.78)	0.61 (0.48, 0.71)
GP 3	<2 years	0.67 (0.56, 0.77)	0.76 (0.66, 0.84)	0.72 (0.64, 0.78)	0.70 (0.61, 0.78)
GP 4	<2 years	0.58 (0.47, 0.68)	0.96 (0.89, 0.99)	0.77 (0.70, 0.83)	0.72 (0.62, 0.81)
GP 5	<2 years	0.57 (0.45, 0.67)	0.99 (0.93, 1.00)	0.78 (0.70, 0.84)	0.72 (0.62, 0.80)
GP 6	<2 years	0.66 (0.55, 0.75)	0.92 (0.84, 0.96)	0.79 (0.72, 0.85)	0.76 (0.67, 0.83)
EMS 1	5–10 years	0.49 (0.38, 0.60)	0.99 (0.93, 1.00)	0.74 (0.66, 0.80)	0.65 (0.54, 0.74)
EMS 2	5–10 years	0.80 (0.70, 0.88)	0.97 (0.91, 0.99)	0.89 (0.83, 0.93)	0.88 (0.81, 0.93)
EMS 3	>10 years	0.82 (0.71, 0.89)	0.97 (0.91, 0.99)	0.89 (0.84, 0.93)	0.89 (0.82, 0.94)

(*Continued*)

Table 2 (Continued).

Reader	Experience	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	F1 Score (95% CI)
TS 1	>10 years	0.67 (0.56, 0.77)	1.00 (0.95, 1.00)	0.84 (0.77, 0.89)	0.80 (0.72, 0.87)
TS 2	5–10 years	0.84 (0.74, 0.91)	0.91 (0.82, 0.95)	0.88 (0.81, 0.92)	0.87 (0.81, 0.92)
TS 3	5–10 years	0.97 (0.91, 0.99)	1.00 (0.95, 1.00)	0.99 (0.95, 1.00)	0.99 (0.96, 1.00)
Radiologist 1	5–10 years	0.57 (0.45, 0.67)	0.96 (0.89, 0.99)	0.76 (0.69, 0.82)	0.70 (0.61, 0.79)
Radiologist 2	5–10 years	0.83 (0.73, 0.90)	0.97 (0.91, 0.99)	0.90 (0.84, 0.94)	0.89 (0.83, 0.94)
Radiologist 3	5–10 years	0.93 (0.86, 0.97)	0.95 (0.87, 0.98)	0.94 (0.89, 0.97)	0.94 (0.90, 0.98)
GC Vertex AI (Cloud)	–	1.00 (0.83, 1.00)	0.89 (0.69, 0.97)	0.95 (0.83, 0.99)	0.95 (0.86, 1.00)
GC Vertex AI (Edge)	–	0.95 (0.75, 0.99)	0.95 (0.75, 0.99)	0.95 (0.83, 0.99)	0.95 (0.86, 1.00)

Abbreviations: GP, general practitioner; TS, thoracic surgeon; EMS, emergency medicine specialist; GC, google cloud.

F1 score of 0.95. These values not only exceeded the averages of junior physicians but also closely approximated those of senior clinicians, underscoring the model's potential to support diagnostic consistency across varying experience levels.

These findings highlight the AI model's utility as a decision support tool, especially for less-experienced practitioners, and underscore its capacity to contribute to more equitable and reliable pneumothorax diagnosis in acute care settings.

Performance in Detecting Minimal Pneumothorax

A subgroup analysis was performed on the 37 patients diagnosed with minimal pneumothorax to compare diagnostic performance across reader groups. The AI models (n = 2) demonstrated a sensitivity of 0.93 (95% CI: 0.79–0.97), specificity of 0.86 (95% CI: 0.72–0.94), diagnostic accuracy of 0.90 (95% CI: 0.75–0.96), and an F1 score of 0.86 (95% CI: 0.81–0.95). This performance slightly exceeded that of radiologists (n = 3), who achieved a sensitivity of 0.81 (95% CI: 0.66–0.91), specificity of 0.92 (95% CI: 0.79–0.97), accuracy of 0.88 (95% CI: 0.75–0.96), and an F1 score of 0.83 (95% CI: 0.74–0.94). Thoracic surgeons (n = 3) showed similar performance with a sensitivity of 0.84 (95% CI: 0.69–0.92), specificity of 0.94 (95% CI: 0.82–0.99), accuracy of 0.89 (95% CI: 0.75–0.96), and an F1 score of 0.86 (95% CI: 0.80–0.96). Emergency medicine specialists (n = 3) followed closely, with a sensitivity of 0.77 (95% CI: 0.60–0.87), specificity of 0.92 (95% CI: 0.79–0.97), accuracy of 0.86 (95% CI: 0.72–0.94), and an F1 score of 0.81 (95% CI: 0.71–0.92). General practitioners (n = 6) exhibited the lowest performance, with a sensitivity of 0.55 (95% CI: 0.38–0.69),

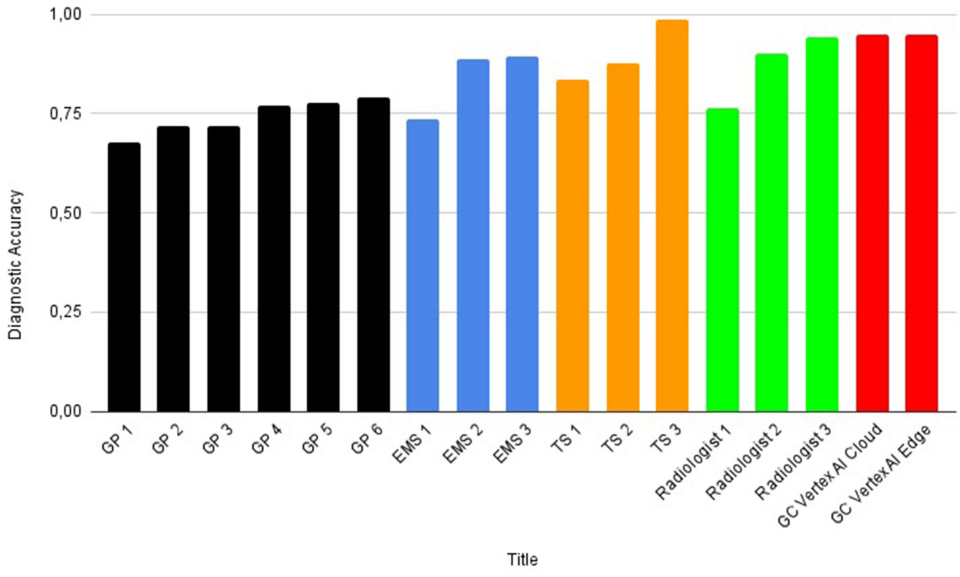


Figure 3 Comparison of Diagnostic Accuracy by Title.

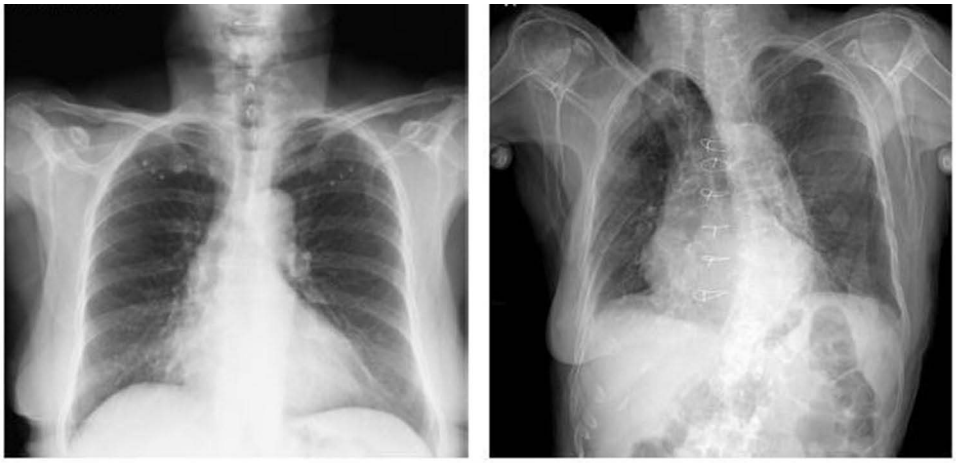


Figure 4 False Positive Cases on Cloud Model (Pneumothorax Predicted in Healthy Cases).

specificity of 0.89 (95% CI: 0.75–0.96), diagnostic accuracy of 0.73 (95% CI: 0.57–0.85), and an F1 score of 0.63 (95% CI: 0.51–0.77). These findings suggest that the AI models not only outperformed less experienced physicians but also demonstrated diagnostic capabilities approaching those of senior clinicians in the detection of subtle pneumothoraces.

Error Analysis

False positive cases for the AI model were reviewed. As shown in Figure 4, cloud-deployed AI misclassified some healthy radiographs as pneumothorax. Upon manual inspection, these errors were typically associated with overlapping anatomical structures or suboptimal image quality, highlighting areas for future refinement of the model's specificity.

Discussion

Comparison with Other AI Studies on Pneumothorax Detection

Our findings align with a growing body of literature demonstrating high accuracy of deep learning models for pneumothorax detection on chest X-rays. Early work by Rajpurkar et al introduced the CheXNet algorithm trained on the ChestX-ray14 dataset, achieving area under the ROC curve (AUC) values up to ~0.89 for pneumothorax detection.[10] Subsequent studies have further improved on these results; for example, Irvin et al expanded this approach in the CheXpert dataset with more sophisticated architectures and uncertainty handling, contributing to pneumothorax detection AUCs around 0.95 in more recent publications. In fact, a recent systematic review and meta-analysis reported an overall AUC of 0.97 for both deep learning (DL) models and physicians in pneumothorax diagnosis,[11] indicating that state-of-the-art AI now performs on par with expert interpretation. Annarumma et al took a complementary approach by focusing on clinical workflow; they developed an AI triage system that could flag critical chest radiographs (including those with pneumothorax) in real-time. Their system detected abnormal (urgent) radiographs with high specificity (95%) and enabled a drastic reduction in reporting delays – from an average of 11.2 days to 2.7 days for critical findings.[12] These studies collectively underscore that our AI model's strong performance is consistent with the literature, which has shown deep learning can reliably identify pneumothoraces and even expedite care. Our results add to this evidence by validating that high performance can be maintained even in a small-scale evaluation with a high proportion of subtle, minimal pneumothorax cases.

Edge vs Cloud Deployment Consistency

A notable finding of this study is the consistent performance of the AI model in both edge and cloud deployment modes. In our evaluation on 152 chest X-rays (76 pneumothorax, 76 normal), the model's sensitivity, specificity, and AUC for pneumothorax detection were essentially identical whether running on a local edge device or on the cloud platform. This practical equivalence is important for real-world implementation. It suggests that the model's inferencing capability is not substantially affected by the computational environment, which is encouraging for point-of-care use. Hospitals or clinics with limited

internet connectivity could deploy the model on edge devices (eg, on-premises servers or even mobile devices) without sacrificing accuracy. Conversely, cloud deployment can facilitate integration into Picture Archiving and Communication System (PACS) systems and tele-radiology workflows. The ability to achieve uniform results in both settings highlights the robustness and flexibility of our approach. From a clinical perspective, this means that a portable chest X-ray device equipped with the AI (edge mode) in remote or resource-limited settings would provide the same diagnostic aid as the cloud-based AI in a tertiary hospital. Consistent edge and cloud performance underscores the model's reliability and broad applicability, ensuring that patients can benefit from AI-assisted pneumothorax detection regardless of infrastructure constraints.

AI Model Performance versus Physicians

Perhaps most impactful is that our AI model demonstrated performance comparable to, or exceeding, that of physicians across varying specialties and experience levels. In aggregate, the model's accuracy in detecting pneumothorax on chest radiographs was on par with senior radiologists and superior to most non-specialist and junior clinicians who participated. This finding mirrors reports by other groups: for example, one study found a deep learning algorithm matched practicing radiologists on the majority of chest X-ray pathologies, and even exceeded radiologist performance for certain findings like pneumothorax in a trauma setting.[13] In our study, junior doctors (such as interns or first-year residents) had noticeably lower sensitivity in pneumothorax identification, especially for subtle cases, consistent with the known learning curve for chest X-ray interpretation. The AI, by contrast, maintained high sensitivity and specificity regardless of case difficulty. For instance, in detecting any pneumothorax, the AI achieved an accuracy that was approximately equal to a board-certified radiologist, while junior physicians lagged behind. This suggests the AI can serve as a "second reader" or safety net, potentially preventing missed pneumothoraces by less experienced practitioners. Moreover, even experienced emergency physicians and general radiologists showed variability in recognizing very small pneumothoraces, whereas the AI's performance was more consistent. Our results reinforce that integrating AI into clinical practice could standardize diagnostic quality. Importantly, the AI's relative advantage was most pronounced for the less experienced clinicians, highlighting its value as a decision support tool to augment training and confidence in pneumothorax detection.

Detection of Minimal Pneumothorax Cases

The high proportion of minimal (small) pneumothorax cases in our test set provides a stringent test for any detection system. Minimal pneumothoraces — often just a faint apical pleural line with little to no lung collapse — are notoriously challenging to recognize on chest X-rays and are frequently missed on initial interpretation.[14] Accurate identification of these subtle findings is clinically crucial because even a small pneumothorax can enlarge or lead to tension physiology if overlooked.

In our study, a focused analysis of patients with minimal pneumothorax (n = 37) demonstrated that the AI models maintained strong diagnostic performance even in subtle cases that are often prone to underdiagnosis. As presented in Table 3, the AI models (n = 2) achieved a sensitivity of 0.93 (95% CI: 0.79–0.97), specificity of 0.86 (95% CI: 0.72–0.94), diagnostic accuracy of 0.90 (95% CI: 0.75–0.96), and an F1 score of 0.86 (95% CI: 0.81–0.95). These results were slightly superior to those of radiologists (n = 3), who demonstrated a sensitivity of 0.81, accuracy of 0.88, and an F1 score of 0.83. Thoracic surgeons (n = 3) and emergency medicine specialists (n = 3) also performed comparably, with F1 scores of 0.86 and 0.81, respectively. General practitioners (n = 6), on the other hand, exhibited the lowest overall performance, with a sensitivity of 0.55 (95% CI: 0.38–0.69), accuracy of 0.73 (95% CI: 0.57–0.85), and an F1 score of 0.63 (95% CI: 0.51–0.77).

Table 3 Performance Metrics for Minimal Pneumothorax Detection (n = 37) by AI Models and Physicians of Different Experience Levels

Group	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)	F1 Score (95% CI)
AI Models (Edge/Cloud)	0.93 (0.79, 0.97)	0.86 (0.72, 0.94)	0.90 (0.75, 0.96)	0.86 (0.81, 0.95)
Radiologists	0.81 (0.66, 0.91)	0.92 (0.79, 0.97)	0.88 (0.75, 0.96)	0.83 (0.74, 0.94)
Thoracic Surgeons	0.84 (0.69, 0.92)	0.94 (0.82, 0.99)	0.89 (0.75, 0.96)	0.86 (0.80, 0.96)
Emergency Medicine	0.77 (0.60, 0.87)	0.92 (0.79, 0.97)	0.86 (0.72, 0.94)	0.81 (0.71, 0.92)
General Practitioners	0.55 (0.38, 0.69)	0.89 (0.75, 0.96)	0.73 (0.57, 0.85)	0.63 (0.51, 0.77)

This comparative analysis highlights that the AI models outperformed less experienced clinicians and approached the performance levels of senior physicians in detecting minimal pneumothorax. The greatest disparity was observed in sensitivity: while general practitioners missed nearly half of the minimal pneumothorax cases, the AI detected the vast majority. Although all reader groups maintained relatively high specificity, the AI's ability to consistently flag subtle cases—sometimes visible only as a razor-thin pleural line—could support earlier interventions, such as close monitoring or preventive measures in high-risk patients. These findings align with prior research noting performance drops for smaller pneumothoraces in both human and AI assessments,[11] but suggest that the AI models used in this study may help close that gap.

Our results have several important clinical implications. First, the deployment flexibility (edge and cloud) means the AI model can be used in diverse healthcare environments. Rural clinics, mobile military hospitals, or emergency departments in resource-limited settings could run the model on local hardware to instantly flag a pneumothorax on a chest X-ray, expediting triage even when specialist radiologists are not on-site. Timely detection of a pneumothorax can be life-saving – for example, in trauma or in patients receiving positive-pressure ventilation, early identification allows prompt chest tube placement before a tension pneumothorax develops. In routine hospital practice, this AI could function as an automated "over-read" system, alerting clinicians if a pneumothorax (including a minimal one) is identified on an X-ray that might otherwise sit in queue. By performing at a level comparable to radiologists, the AI can increase confidence in diagnoses and potentially reduce diagnostic errors. Junior doctors and non-radiology physicians, in particular, could benefit from AI decision support, leading to improved patient outcomes through quicker and more accurate diagnoses. Additionally, consistent performance in minimal pneumothorax detection means the tool is not just finding the obvious large collapses, but adding value in edge cases where clinical signs might be subtle or absent. This could reduce the rate of missed pneumothoraces (historically up to ~20% for occult cases[14]), improving patient safety. Ultimately, integrating such an AI model into clinical workflow – whether on a cloud PACS system or a handheld tablet in the ICU – could streamline care delivery, prioritize urgent cases, and serve as a training aid for developing clinicians in interpreting chest radiographs.

Limitations

Despite the promising results, our study has several limitations that warrant a balanced discussion. First, the sample size was relatively small – we evaluated 152 X-rays (76 with pneumothorax, 76 healthy), which is modest compared to large public datasets used in other AI studies.[9] This was also a single-center study, and all images were drawn from the same institution's database. As a result, there may be limited diversity in patient demographics, image acquisition techniques, and pneumothorax presentations (eg, our cohort had an overrepresentation of minimal apical pneumothoraces). This could affect generalizability; the model's performance should be validated on external datasets and multi-center cohorts to ensure it maintains accuracy across different hospital settings and X-ray machines. Second, the study was retrospective in nature. The AI algorithm and physicians interpreted images in an experimental setting, which may not fully simulate real-world clinical workflows. For instance, radiologists usually have access to patient history and serial imaging, which were not provided in our test scenario. The measured physician performance might therefore underestimate real-life accuracy (experienced radiologists could perform better with clinical context, while junior doctors might consult seniors in practice). Additionally, we compared the AI to physicians interpreting the images independently; however, in practice the AI would more likely be used as an assistive tool rather than a standalone reader. Future work could explore how physicians' performance changes when aided by the AI, as others have suggested a synergistic improvement with AI assistance.[15] Third, the edge deployment was tested on a specific hardware configuration (a high-end mobile device), and latency or resource constraints were not systematically measured. While accuracy was equivalent, the real-time speed and usability of the edge model in low-power devices or in batch processing scenarios were not assessed in detail. Finally, our study did not address how the model handles pneumothorax size quantification or differentiation between simple vs tension pneumothorax – it was a binary classification of presence/absence. These factors, along with the need to prospectively assess clinical impact (does AI assistance actually reduce missed diagnoses or improve patient outcomes?), represent areas for future research. In summary, while the initial performance of the Google Cloud Vertex AI model for pneumothorax detection is encouraging, especially in comparison to human experts and across deployment settings, these

findings should be interpreted in light of the study's constraints. Rigorous prospective trials and broader validations will be essential to confirm the generalizability and clinical utility of the system before widespread adoption.

Practical Barriers to Clinical AI Adoption

While the diagnostic performance of the AI model in this study is promising, several real-world barriers must be addressed before widespread clinical integration is feasible. One major challenge involves geographic and demographic bias—AI models trained on datasets from a single institution or region may not generalize well to diverse populations or imaging protocols. Additionally, reliance on cloud infrastructure may pose constraints in healthcare systems with limited internet connectivity or data-sharing restrictions. Regulatory and validation requirements also remain significant hurdles; clinical-grade AI tools must undergo rigorous evaluation and receive approvals from governing bodies before routine use. Finally, computational demands—particularly for edge deployment on low-power devices—may limit adoption in resource-constrained settings. These practical considerations highlight the need for multi-institutional validation, regulatory clarity, and robust infrastructure to ensure safe and effective AI implementation in clinical workflows.

Conclusions

The Google Cloud Vertex AI model demonstrated high diagnostic accuracy for pneumothorax detection in both cloud and edge deployments, performing comparably to physicians—especially in detecting minimal cases. Its real-time capability and consistent performance suggest strong potential as a clinical decision support tool, particularly for less experienced clinicians and in resource-limited settings.

Data Sharing Statement

The raw data of the study can be shared upon reasonable request. For access to the data, the responsible researcher can be contacted via email at: drismaildal@gmail.com.

Ethics Approval

Before the start of the study, approval was obtained from Kastamonu University Clinical Research Ethics Committee (Approval No: 2024-KAEK-5). Due to the retrospective nature of the study, the requirement for informed consent was waived by the ethics committee. All patient data were fully anonymized prior to analysis, and confidentiality was strictly maintained in accordance with the Declaration of Helsinki.

Acknowledgments

The authors declare that there are no acknowledgments.

Author Contributions

Conceptualization; ID, HBK: Data curation; ID, HBK: Formal analysis; ID, HBK: Funding acquisition; ID, HBK: Investigation; ID, HBK: Methodology; ID, HBK: Project administration; ID, HBK: Resources; ID, HBK: Software; ID, HBK: Supervision; ID, HBK: Validation; ID, HBK: Visualization; ID, HBK: Writing – original draft; ID, HBK: Writing – review & editing; ID, HBK. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The authors did not receive support from any organization for the submitted work.

Disclosure

The authors declare that they have no conflicts of interest in this work.

References

1. Brady A, Laoide RÓ, McCarthy P, McDermott R. Discrepancy and error in radiology: concepts, causes and consequences. *Ulster Med J*. 2012;81(1):3–9. PMID: 23536732; PMCID: PMC3609674.
2. Aswin K, Balamurugan S, Govindarajalou R, Saya GK, Tp E, Rajendran G. Comparing sensitivity and specificity of ultrasonography with chest radiography in detecting pneumothorax and hemothorax in chest trauma patients: a cross-sectional diagnostic study. *Cureus*. 2023;15(8):e44456. doi:10.7759/cureus.44456
3. Satia I, Bashagha S, Bibi A, Ahmed R, Mellor S, Zaman F. Assessing the accuracy and certainty in interpreting chest X-rays in the medical division. *Clin Med*. 2013;13(4):349–352. PMID: 23908502; PMCID: PMC4954299. doi:10.7861/clinmedicine.13-4-349
4. Hillis JM, Bizzo BC, Mercaldo S, et al. Evaluation of an artificial intelligence model for detection of pneumothorax and tension pneumothorax in chest radiographs. *JAMA Network Open*. 2022;5(12):e2247172. PMID: 36520432; PMCID: PMC9856508. doi:10.1001/jamanetworkopen.2022.47172
5. Sheng B, Tao L, Zhong C, Gao L. Comparing the diagnostic performance of lung ultrasonography and chest radiography for detecting pneumothorax in patients with trauma: a meta-analysis. *Respiration*. 2025;104(3):161–175. PMID: 39348819; PMCID: PMC11887991. doi:10.1159/000540777
6. Bennani S, Regnard NE, Ventre J, et al. Using AI to improve radiologist performance in detection of abnormalities on chest radiographs. *Radiology*. 2023;309(3):e230860. Erratum in: Radiology. 2024;311(1):e249015. doi: 10.1148/radiol.249015. PMID: 38085079. doi:10.1148/radiol.230860
7. Urago Y, Okamoto H, Kaneda T, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. *Radiat Oncol*. 2021;16(1):175. PMID: 34503533; PMCID: PMC8427857. doi:10.1186/s13014-021-01896-1
8. Memon K, Yahya N, Yusoff MZ, et al. Edge computing for AI-based brain MRI applications: a critical evaluation of real-time classification and segmentation. *Sensors*. 2024;24(21):7091. PMID: 39517987; PMCID: PMC11548207. doi:10.3390/s24217091
9. Baumann MH, Strange C, Heffner JE, et al. The ACCP consensus statement on the management of spontaneous pneumothorax. *Chest*. 2001;119(2):590–602. doi:10.1378/chest.119.2.590
10. Rueckel J, Huemmer C, Fieselmann A, et al. Pneumothorax detection in chest radiographs: optimizing artificial intelligence system for accuracy and confounding bias reduction using in-image annotations in algorithm training. *Eur Radiol*. 2021;31(10):7888–7900. PMID: 33774722; PMCID: PMC8452588. doi:10.1007/s00330-021-07833-w
11. Sugibayashi T, Walston SL, Matsumoto T, Mitsuyama Y, Miki Y, Ueda D. Deep learning for pneumothorax diagnosis: a systematic review and meta-analysis. *Eur Respir Rev*. 2023;32(168):220259. PMID: 37286217; PMCID: PMC10245141. doi:10.1183/16000617.0259-2022
12. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G. Automated triaging of adult chest radiographs with deep artificial neural networks. *Radiology*. 2019;291(1):196–202. Erratum in: Radiology. 2019;291(1):272. doi: 10.1148/radiol.2019194005. PMID: 30667333; PMCID: PMC6438359. doi:10.1148/radiol.2018180921
13. Gipson J, Tang V, Seah J, et al. Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma. *Br J Radiol*. 2022;95(1134):20210979. PMID: 35271382; PMCID: PMC10996416. doi:10.1259/bjr.20210979
14. Brar MS, Bains I, Brunet G, Nicolaou S, Ball CG, Kirkpatrick AW. Occult pneumothoraces truly occult or simply missed: redux. *J Trauma*. 2010;69(6):1335–1337. PMID: 21150515. doi:10.1097/TA.0b013e3181f6f525
15. Anderson PG, Tarder-Stoll H, Alpaslan M, et al. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. *Sci Rep*. 2024;14(1):25151. PMID: 39448764; PMCID: PMC11502915. doi:10.1038/s41598-024-76608-2