Archives of Academic Emergency Medicine. 2025; 13(1): e86

OPEN ACCESS

REVIEW ARTICLE

The Role of Artificial Intelligence in Diagnosing Pulmonary Embolism: A Systematic Review and Meta-analysis

Alireza Farzaei[1,2†], Fateme Hajzeinolabedini[3†], Babak Sharif Kashani[4], Mohamad Sadegh Keshmiri[4], Alireza Khodayari Javazm[5], Yasaman Farzaei[3], Mohamad Fereidooni[3], Behrouz Emamjomeh[4], Amir Nezami-asl[1*]

1. AJA university of medical sciences, school of medicine, Tehran, Iran
2. Department of Cardiology, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3. Kermanshah university of medical sciences, school of medicine, Kermanshah, Iran
4. Department of Cardiology, National Research Institute of Tuberculosis and Lung Diseases, Shahid Beheshti University of Medical Sciences, Tehran, Iran
5. Department of Cardiology, School of Medicine, Alborz university of medical sciences, Karaj, Iran
6. Tehran Heart Center Hospital, Tehran University of Medical Sciences, Tehran, Iran

Received: October 2025; Accepted: November 2025; Published online: 30 December 2025

Abstract: **Introduction:** Missed or delayed diagnosis of pulmonary embolism (PE) is associated with increased morbidity, mortality, and longer hospitalizations. This study aimed to evaluate the diagnostic accuracy of Artificial Intelligence (AI) models in detecting PE across imaging. **Methods:** We systematically searched PubMed/MEDLINE, Scopus, Embase and Web of Science from inception to 1 January 2025 without language or regional limits. After removing duplicate results, the remaining records were screened through titles/abstracts, and two reviewers independently assessed full texts. Risk of bias was evaluated in duplicate with the QUADAS-2 tool. Pooled sensitivity, specificity, positive and negative likelihood ratios, diagnostic odds ratio and area under the ROC curve were calculated with random effects models in STATA 17. Heterogeneity was quantified with Cochran's Q and I^2, while we explored its sources using subgroup analyses (for categorical moderators) and meta regression (for continuous moderators). Publication bias was assessed with Deeks' funnel plot and trim and fill, and we examined robustness through leave one out sensitivity analyses. **Results:** A total of 1,432 records were identified through database searches, with 654 duplicates removed. After screening titles and abstracts of 787 articles, 256 full-text articles were assessed for eligibility, and 28 studies met the inclusion criteria. Internal validation phases included 43,330 participants (4,866 PE-positive, 38,463 PE-negative), while external validation phases comprised 3,588 participants (1,699 PE-positive, 1,889 PE-negative). In the internal validation phase, the pooled sensitivity and specificity of AI in PE diagnosis across imaging were 0.91 (95% confidence interval (CI): 0.88–0.95; I^2=9%) and 0.94 (95% CI: 0.86–0.98; I^2=99.78%), respectively. The positive likelihood ratio (PLR) was 16.08, and the negative likelihood ratio (NLR) was 0.09, both statistically significant (P < 0.001). The pooled diagnostic odds ratio (DOR) was 163.55 (95% CI: 71.30-375.14, I^2: 96.1), and the area under the curve (AUC) was 0.95 (95% CI: 0.93 to 0.97), indicating excellent accuracy. In external validation, the pooled sensitivity and specificity were slightly lower at 0.89 (95% CI: 0.79–0.95; I^2=95.60%) and 0.88 (95% CI: 0.80–0.93; I^2=91.48%), respectively. The DOR was 59.65 (95% CI: 23.53 to 151.17, I^2: 89.6) and AUC was 0.94 (95% CI: 0.92 to 0.96, I^2: 89.6). There was no significant publication bias detected. **Conclusion:** AI models achieved high diagnostic accuracy in detecting PE through imaging. However, this performance tends to decrease from internal to external validation, highlighting limitations in generalizability. Additionally, substantial heterogeneity was observed across studies, as indicated by high I^2 values, which should be considered when interpreting the pooled estimates.

Keywords: Pulmonary Embolism; Artificial Intelligence; Machine Learning; Deep Learning; Computed Tomography Angiography; Meta-analysis

Cite this article as: Farzaei A, Hajzeinolabedini F, Sharif Kashani B, et al. The Role of Artificial Intelligence in Diagnosing Pulmonary Embolism: A Systematic Review and Meta-analysis. Arch Acad Emerg Med. 2025; 13(1): e86. https://doi.org/10.22037/aaem.v13i1.2720.

1. Introduction

Pulmonary embolism (PE), a life-threatening disease caused by the obstruction of pulmonary arteries, is one of the leading causes of cardiovascular death worldwide. Approximately 10–30% of patients with PE are estimated to die within 30 days of disease onset. Patients with high-risk PE, in which

***Corresponding Author:** Amir Nezami-asl; AJA University of Medical Sciences, School of Medicine, Tehran, Iran. Tel: 00989123479863, Email: nezamiaslamir@gmail.com, ORCID: https://orcid.org/0000-0003-3458-4422.
†These authors contributed equally to this article.

blood flow is blocked by more than 50%, are more likely to die within 90 days (1,2). In order to prevent fatal outcomes, it's crucial to get a diagnosis as quickly and accurately as possible. Computed tomography pulmonary angiography (CTPA) serves as the gold standard for diagnosing PE, offering high sensitivity and specificity. Despite this, interpreting these imaging studies requires significant expertise, is time-consuming, and can be subject to inter-observer variability. Challenges are increased by limited access to expert radiologists in resource-constrained settings. Researchers have shown that CTPA has a missed diagnosis rate of 14% (3–5).

Missed or delayed diagnoses are associated with significant clinical consequences, including increased morbidity, mortality, longer hospitalizations, and elevated healthcare costs. In inpatient settings, PE was missed on abdominal CT scans in 81.8% of patients who were later diagnosed via chest CT scan. Pooled results indicate that in emergency department settings, 27.5% of PE cases are initially misdiagnosed, while 53.6% are misdiagnosed in inpatient settings. Furthermore, autopsy studies have found that 37.9% of ICU patients who died were found to have undiagnosed PE. These alarming figures illustrate the diagnostic challenges clinicians face (6,7).

AI can play an important role in the early and more precise thromboembolic diseases diagnosis. In this context, AI encompasses both traditional machine learning methods—such as support vector machines and random forests—and deep learning approaches, particularly convolutional neural networks (CNNs), which are highly effective in analyzing medical imaging data. These algorithms are capable of detecting subtle patterns in imaging studies, which can improve PE diagnosis by automatically identifying emboli in CTPAs, ventilation-perfusion scans, and other imaging modalities (8,9).

However, while original studies demonstrated promising results, the evidence remains inconsistent. Some studies report very low sensitivity and specificity, while others demonstrate much higher diagnostic performance. Furthermore, most prior reviews have been narrative and did not include pooled accuracy metrics or formally assess between-study heterogeneity (10–12).

This highlights the need for a systematic meta-analysis to quantitatively synthesize existing evidence and evaluate the diagnostic robustness and generalizability of AI models for PE detection. So, this study aimed to assess the diagnostic accuracy of any AI model, regardless of type, in detecting PE across various imaging techniques.

2. Methods

2.1. Study design and setting

This study is conducted in accordance with the Cochrane Handbook for Systematic Reviews of Interventions and the results are reported according to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) 2020 guidelines. The protocol of this study is registered in PROSPERO (registration code: CRD42024627713). Studies that have assessed the role of any AI model, including deep learning models (e.g., convolutional neural networks, recurrent neural networks), machine learning models, or hybrid AI systems, for the diagnosis of PE from medical images were included.

2.2. Eligibility criteria of primary studies

We included primary studies that were observational and had assessed patients' radiologic lung images (either of non-contrast or intravenous (IV)-contrast chest computed tomography (CT) scans and pulmonary CT angiograms). Therefore, cross-sectional, case-control, prospective and retrospective cohort designs were included. Clinical trials, experimental and quasi experimental studies, protocols, reviews, editorials, comments, books or book chapters and letters were excluded. No language restrictions were applied. All studies conducted on patients without restriction on their age, gender, and ethnicity were included. Either radiologic reports or ICD-10 codes of PE were considered as reference standard.

2.3. Search strategy

PubMed/Medline, SCOPUS, EMBASE and Web of Science (WOS) were searched from their inception to January 1, 2025. The references of included studies and previous relevant systematic reviews and meta-analysis were searched manually in order to ensure inclusion of most relevant studies. We did not restrict our search by language or geographical area. The detailed search syntax of all databases is presented in supplementary table 1.

2.4. Screening, selection, and data extraction process

All records retrieved from the databases were first imported into EndNote (version 20) for initial duplicate removal. The deduplicated records were then imported into Rayyan for title and abstract screening, followed by full-text selection. Then title and abstract of primary studies were screened by A.F. in the next step, the full-text of potentially eligible articles were retrieved and two independent reviewers (A.F. and F.H.) performed the selection process. Any disagreement during this process was resolved by consensus or a third reviewer (A.N.A.).

Based on our study objectives we designed a data extraction checklist. Then, required data were extracted by Y.F. and M.F., independently, from original studies and any disagreements in the extracted data were resolved through consensus. Data regarding first author name, country of study, the year of study, the AI model, imaging modality, imaging device, mean and standard devotion (SD) of age of participants, the percentage of female participants, the total number of participants, and the number of patients with and without PE. When available, true positive, true negative, false positive, and false negative counts were extracted directly

from the tables or confusion matrices to calculate diagnostic performance measures. If these values were not provided and only summary receiver operating characteristic (SROC) plots or likelihood ratios were reported, we used the "top-left method" to derive sensitivity and specificity values from the SROC plot. These values were then rounded to the nearest whole number, when necessary, especially if they were presented as decimal values.

2.5. Quality assessment (risk of bias assessment)

A risk of bias (ROB) checklist was designed and two reviewers (B.S.K. and M.K.) independently assessed the included articles. The differences between opinions were resolved through consultation with a third expert author or consensus. QUADAS-2, consists of four key domains: Patient selection, index test, reference standard and flow and timing. We considered some adjustments for AI diagnostic studies using QUADAS-2 tool. In particular, we modified the signaling questions in the index test domain to address concerns related to the "black box" nature of AI algorithms. These algorithms often lack interpretability, making it challenging to fully assess the decision-making process behind the test results. As such, we considered how the transparency of the AI models was reported, including whether the studies provided sufficient detail on model development, training, and validation procedures to evaluate the risk of bias accurately. Also, we evaluated whether studies ensured that radiologists interpreting the reference standard were blinded to the AI-generated outputs, and vice versa. This is crucial to maintain independence between the index and reference tests, minimizing potential bias. Each is assessed in terms of risk of bias and the first three in terms of concerns regarding applicability. Each of the seven items was assessed and rated as low ROB, high ROB, or unclear ROB. The overall quality (or ROB) status was evaluated in accordance with the QUADAS-2 guidelines.

2.6. Statistical analysis

We used random effects meta-analysis for pooling diagnostic accuracy metrics due to the substantial heterogeneity observed across the included studies in study characteristics and performance. To present the diagnostic performance forest plots were used. Heterogeneity among studies, with thresholds based on the guidelines of Higgins et al. was assessed using I^2 statistics (0 to 25%: Mild heterogeneity; 25 to 50%: Moderate heterogeneity; 50 to 75%: Severe heterogeneity; 75 to 100%: Very severe heterogeneity) (13). The primary measure of our study was diagnostic performance including area under curve–receiver operating characteristic curve (AUC–ROC) and other diagnostic accuracy measures (positive likelihood ratio (PLR), negative likelihood ratio (NLR), sensitivity, specificity, accuracy, and diagnostic odds ratio (DOR)). I^2 statistics and Q Cochrane test were used for evaluation of heterogeneity. In order to find potential sources of heterogeneity and effective variables we performed subgroup analysis (on categorical variables) and meta regression (on continuous variables). For publication or reporting bias evaluation we used Deeks' funnel plot and trim and fill method. The leave-one-out method was used to examine the effect of each primary study and the overall measure as a sensitivity analysis approach.

All statistical analysis were performed using STATA 17 (StataCorp. College Station, TX, USA).

3. Results

3.1. Characteristics of included studies

Overall, 1432 records were found trough search in electronic databases. Among them 654 duplicate records were removed. Title and abstract of 787 articles were evaluated and then in the selection phase, 256 full-text articles were independently evaluated. Finally, 28 articles were found eligible for our study. Out of the 28 included studies, 22 reported only internal validation, 2 reported only external validation, and 4 reported both internal and external validation results. The flowchart of our study is presented in figure 1.

Features of included studies are listed in table 1. The distribution of studies' countries was: 2 studies in China, 6 studies in USA, 1 in Brazil, 2 in India and others in European countries (Turkey, Netherlands, Switzerland, France, Finland and Germany). Most of the models were trained on CTPA, 2 on non-contrast CT images, 2 on IV-contrast CT images, and 1 on ventilation-perfusion scans. Study design of all included articles was retrospective observational. Some of the studies have used commercial AI models such as Aidoc and CINA-PE. Others have used various deep learning models and algorithms. The imaging devices used in studies were from GE, Siemens, Philips and Canon companies.

The mean age of participants was 61.07 ± 16.6 and 49.65% of participant were female. The total number of participants in internal validation phase was 43330 (4866 of them were PE positive and 38.463 of them were negative). In the external validation phase the total number of patients was 3588, among them 1699 were positive for PE and 1889 were PE negative.

3.2. Diagnostic accuracy of AI models in internal validation

The combined diagnostic accuracy measure of the internal validation phase of studies is presented in figure 2. The sensitivity and specificity of AI models in diagnosing PE through medical images were 0.91 (95% confidence interval (CI): 0.88 to 0.94, p <0.001, I^2: 94.75) and 0.94 (95% CI: 0.86 to 0.98, p <0.001, I^2: 99.78), respectively. Their positive diagnostic likelihood ratio was 16.08 (95% CI: 6.41 to 40.35, p <0.001, I^2: 99.77) and their negative diagnostic likelihood ratio was 0.09 (95% CI: 0.07 to 0.12, p <0.001, I^2: 95.57). Moreover, the pooled diagnostic odds ratio of the models was 163.55 (95% CI: 71.30 to 375.14, p <0.001, I^2: 96.1) and their pooled AUC was 0.95 (95% CI: 0.93 to 0.97). All of the measures were sta-

tistically significant.

3.3. Diagnostic accuracy of AI models in external validation

Figure 3 displays the overall diagnostic accuracy metrics from the external validation phase across the studies. In detecting PE using medical images, the sensitivity of models was 0.89 (95% CI: 0.79 to 0.95, p <0.001, I^2: 95.60) and their pooled specificity was 0.88 (95% CI: 0.80 to 0.93, p <0.001, I^2: 91.48). The positive diagnostic likelihood ratio and negative diagnostic likelihood ratio were 7.19 (95% CI: 4.35 to 11.89, p <0.001, I^2: 85.77) and 0.12 (95% CI: 0.006 to 0.26, p <0.001, I^2: 95.4), respectively. Finally, the combined AUC of the models was 0.94 (95% CI: 0.92 to 0.96, p <0.001, I^2: 89.6) and their pooled diagnostic odds ratio was 59.65 (95% CI: 23.53 to 151.17, p <0.001, I^2: 89.6).

3.4. Subgroup analysis and meta-regression

Due to considerable heterogeneity, subgroup analyses were carried out on imaging devices (GE, Siemens, Canon, Philips, Multiple devices, and studies that did not report their device) and AI model (CNN, Recurrent Neural Network (RNN), Feedforward Neural Network (FNN) and Hybrid and Specialized Architectures). The term "multiple devices" refers to studies in which patients were scanned using more than one CT angiography system within a single center, or in multicenter studies where different imaging devices were used across participating sites. Results are presented in figure 4.

CNN models achieved a greater diagnostic performance than other groups with a DOR value of 294.68 (95% CI: 102.8 to 844.63, p < 0.001, I^2=96.2%). Moreover, in comparison of devices, the DOR of studies that have used multiple devices showed higher DOR value 567.53 (95% CI: 167.10 to 1927.53, p < 0.001, I^2=94%).

In the next step, we performed meta-regression to evaluate the effect of mean age, percentage of female participants, and the prevalence of PE in each study on the diagnostic performance of models. Mean age and percentage of female participants did not show a significant effect on DOR (P-values: 0.211 and 0.630, respectively). However, the prevalence of PE in the included studies had a statistically significant negative association with diagnostic accuracy (coefficient = –0.055, 95% CI: -0.097 to -0.012, P-value: 0.013). This indicates that studies with higher PE prevalence tended to report lower diagnostic performance of AI-based tools. Results are presented in table 2 and supplementary figure 1. According to the small number of eligible studies in the external validation phase, subgroup analysis and meta-regression were not feasible. Therefore, these analyses were conducted only on internal validation phase data.

3.5. Risk of bias assessment

The detailed report of risk of bias of the included studies is presented in figure 5. In risk of bias section, all of the studies had appropriately described the reference standard and how it was conducted and interpreted and had appropriate interval between index test and reference standard; therefore, they had high quality in flow and timing and reference standard domains. In the index test domain, 2 studies were judged to be at high risk of bias, and 1 study had unclear risk. In the patient selection domain, 3 studies were at high risk, and 2 studies had unclear risk. Regarding applicability concerns, 4 studies were rated as having high concern in the patient selection domain, 2 studies in the index test domain, and 1 study in the reference standard domain.

3.6. Publication bias and sensitivity analysis

We used Deeks' funnel plot and trim and fill method for better evaluation of the potential effect of publication bias. Trim-and-fill analysis did not impute any study and the funnel plot appeared symmetric. While reliability of Deeks' funnel plot in fewer than 10 studies is limited, no significant publication bias was observed (Figure 6 and supplementary figure 2).

In order to assess robustness of our results we used leave-one-out method. No single study had considerable effect on the overall measures. Results are presented in supplementary figure 3.

4. Discussion

This meta-analysis aimed to examine the diagnostic accuracy of AI models in PE detection using medical images. The results of the internal validation phase of the present meta-analysis show that the AI models have remarkable accuracy in this regard. The overall sensitivity and specificity of the models were high (0.92 and 0.95, respectively), and near or slightly above the performance range of radiologists (65–90% sensitivity and 97–99% specificity) (14,15). The high levels of positive likelihood ratio and negative likelihood ratio indicate that AI models can help clinicians in both detecting and ruling out the PE. These findings suggest that AI systems have the potential to enhance diagnostic accuracy, particularly in high-throughput environments.

In clinical practice, AI may serve as a complementary tool to assist radiologists, thereby improving workflow efficiency and reducing oversight. In settings with limited radiologist availability, validated AI systems might also be used as preliminary screening tools to flag high-risk cases. However, AI should not be viewed as a replacement for expert clinical judgment, particularly in complex or ambiguous cases.

There was high heterogeneity between the included studies. Therefore, we performed subgroup analysis and meta-regression to identify the source of heterogeneity. The results showed that the mean age of the participants and the percentage of female participants played no significant role in heterogeneity. However, a significant negative association was found between PE prevalence and the diagnostic performance of AI models. This finding suggests that as the prevalence of PE increases, the diagnostic accuracy of AI systems declines. One possible explanation is that high-prevalence

settings, such as specialized or tertiary care centers, may include more complex or atypical cases, which are more challenging for AI algorithms to correctly classify. These results highlight the importance of considering disease prevalence when evaluating and deploying AI diagnostic tools in diverse clinical settings.

Moreover, AI models' architectures played a significant role in heterogeneity, with studies using CNN models showing the best accuracy. CNNs' superiority over other architectures likely reflects their inherent efficiency for image pattern recognition. Fortunately, we did not have any considerable publication bias and sensitivity analysis proves the robustness of our findings. Given the number of subgroup analyses performed, there is an inherent risk of type I error due to multiple comparisons. So, these results should be interpreted cautiously and the findings are intended to generate hypotheses for future researches.

In external validation, we found that AI models are still valuable and have high diagnostic accuracy. Nevertheless, they experience a decrease in performance, which is to be expected. Considering the variability in imaging protocols, patient demographics, and institutional practices, this slight decline highlights the need for more studies to evaluate external validation of AI models. Additionally, the limited number of studies in the external validation phase reduces the precision and reliability of this estimate and shows that these findings should be interpreted cautiously. Further external validation studies are needed to confirm the generalizability of AI models in diverse clinical settings.

The high degree of heterogeneity observed may be attributed to several factors. First, the prevalence of PE varied across studies—from as low as 1% to as high as 97%—which likely influenced diagnostic metrics. Second, differences in imaging acquisition protocols and preprocessing techniques may have introduced variability in how input data were handled by AI models. These factors, alongside differences in model architecture, highlight the need for standardized protocols in future diagnostic AI studies.

It is important to note that the majority of AI models included in this meta-analysis were developed on CTPA, which is the standard imaging modality for PE diagnosis in most clinical settings. However, this limits the external validity of our findings to other modalities such as ventilation-perfusion (V/Q) scans or non-contrast CT, which are often used in patients with contraindications to contrast agents or in pregnancy. Further research is needed to assess the diagnostic performance of AI models across diverse imaging techniques to ensure broader clinical applicability. AI has rapidly emerged as a valuable tool in the diagnosis and risk stratification of PE. Some studies have demonstrated a 23% improvement in detecting subsegmental emboli using AI compared to manual interpretation. Moreover, deep learning models, such as convolutional neural networks (CNNs), have been used to calculate thrombus ratios and volumes, providing predictive insights into patient risk levels, particularly for hemodynamically stable individuals. These models often outperform traditional scoring systems in stratifying high-risk patients. Additionally, AI-assisted image analysis integrated with clinical data has achieved up to 95% accuracy in predicting PE severity based on thrombus characteristics and perfusion deficits (16–19).

While AI models have shown good potential in PE detection, a couple of limitations still prevent their widespread clinical adoption. One key challenge is the lack of standardization across AI algorithms, which affects their generalizability and performance consistency in varying clinical settings. Models trained on specific datasets may not perform as well when applied to different populations or imaging protocols. As evidenced by our results, the performance of models dropped in the external validation phase. There is also concern regarding over-reliance on AI outputs, particularly when false positives or false negatives occur, which could delay appropriate treatment or lead to unnecessary interventions (20–22).

In addition, interpreting complex findings beyond simple embolism detection presents significant challenges. Clinical images often contain anatomical structures, physiological variations, and imaging artifacts that can confound AI algorithms' interpretation and lead to false positives or negatives. Distinguishing between acute and chronic PE, identifying alternative diagnoses mimicking PE, and assessing the clinical significance of detected emboli require careful human clinical judgment and field expertise that may not be fully captured by AI systems alone (17,23,24). Our results suggest that AI models have the potential to assist clinicians in diagnosing PE, showing acceptable accuracy measures. However, further studies should be conducted to evaluate their impact on several items, including image reading time for suspected PE, time to diagnosis, and time to treatment following the implementation of AI models as assistive tools.

Despite the promising diagnostic performance of AI models, several practical barriers may limit their immediate clinical adoption. Integration into existing picture archiving and communication systems (PACS) can be technically complex and resource-intensive. Moreover, most AI tools require regulatory approval (e.g., by the FDA), which involves rigorous validation beyond academic settings. Additionally, the introduction of AI-generated alerts could contribute to alert fatigue, particularly in high-throughput radiology departments. Addressing these challenges will be crucial for the safe and effective integration of AI into routine clinical practice.

The strength of our study is the large number of included studies, robustness of our results in sensitivity analysis, absence of publication bias, and meta-analysis on both internal and external validation phases.

However, it is important to keep the following points in mind when interpreting our results. Firstly, some studies did not report their model performance completely, so they were not included in the meta-analysis. Further, the lack of demographic data and potential confounders in the original

studies prevented us from conducting subgroup analyses on more variables.

So, we suggest that future researchers report the detailed performance of their AI model, including the number of people in true positive and negative groups, as well as the number of people in false positive and negative groups. It will also be helpful for future assessments if we have more demographic information.

Next, we intended to include studies that had considered either radiologic reports or ICD-10 codes as reference standards. We acknowledge that these represent different diagnostic benchmarks and may introduce variability in outcome classification due to differences in clinical documentation, image interpretation, and coding practices. However, all our included studies used radiologist reports as the reference standard, so this potential source of variability did not affect our analysis.

Future research should focus on prospective, multicenter trials that compare AI-assisted interpretation to standard radiologist workflows in real-world clinical settings. Additionally, exploration of AI as a triage tool may offer substantial workflow benefits, especially in emergency or resource-constrained environments.

Caution must be exercised regarding over-reliance on AI outputs, as false positives or false negatives may lead to delayed treatment or unnecessary interventions. The safe and effective integration of AI into medical practice requires that it support, not replace, clinical judgment. Ultimately, AI should be integrated into clinical workflows as a supportive tool—enhancing, but not replacing, clinician expertise and judgment.

5. Limitations

A limitation of our meta-regression and subgroup analysis was that some potentially influential covariates were not consistently reported across included studies. This limited our ability to fully explore sources of heterogeneity. We recommend that future studies provide detailed reporting of such variables to facilitate more robust moderator analyses in future meta-analyses.

Additionally, the meta-regression analyses included a relatively small number of studies, which may limit the statistical power to detect significant associations, particularly when assessing multiple continuous covariates. As such, null findings should be interpreted cautiously, and future studies with larger sample sizes and studies are needed to assess these exploratory associations.

Moreover, studies that assessed the model's performance on datasets other than those on which it was developed were relatively few. Consequently, the power of our meta-analysis on this item was low, and the source of heterogeneity could not be evaluated methodologically. Future research should focus on this aspect and evaluate the generalizability of AI models.

6. Conclusions

Our findings indicate that AI models achieved high diagnostic accuracy in detecting PE based on medical imaging. These results support AI's utility as a valuable assistive tool in clinical decision-making for PE diagnosis. This performance tends to decrease from internal to external validation, highlighting limitations in generalizability. Additionally, substantial heterogeneity was observed across studies, as indicated by high I^2 values, which should be considered when interpreting the pooled estimates. However, the relatively limited number of studies with external validation highlights the need for further research to assess the generalizability and robustness of these models across diverse populations and imaging settings.

7. Declarations

7.1. Acknowledgments

The authors would like to express their gratitude toward Taleghani Hospital Shahid Beheshti University of Medical Sciences, Tehran, Iran for their support.

7.2. Authors' contributions

Concept and Design: AF, AN, BSK; Data Curation: AF, FH, YF, MF; Formal Analysis: AF; Project Administration and Validation: AN, BSK, MSK, AKJ; Supervision: AN; Writing: AF, FH, BSK, MSK, AKJ, YF, MF, BE, AN; Review and Editing: AF, FH, BSK, AN. All authors read and approved the final version of manuscript.

7.3. Data Availability

Data are available on reasonable request from the first or corresponding author.

7.4. Funding Statement

There were no fundings dedicated to this project.

7.5. Declaration of Interests

The authors have declared that they have no conflict of interest.

7.6. Using artificial intelligence chatbots

Artificial intelligence-based chatbots were only used for language editing and improvement of grammar. The authors take full responsibility for the content, accuracy, and integrity of the manuscript. AI tools were not used for data analysis, interpretation of results, or generation of scientific conclusions.

References

1. Beckman MG, Hooper WC, Critchley SE, Ortel TL. Venous thromboembolism: a public health concern. American journal of preventive medicine. 2010;38(4):S495–S501.

2. Barnes GD, Muzikansky A, Cameron S, Giri J, Heresi GA, Jaber W, et al. Comparison of 4 acute pulmonary embolism mortality risk scores in patients evaluated by pulmonary embolism response teams. JAMA network open. 2020;3(8):e2010779–e.
3. Peris M, López-Nuñez JJ, Maestre A, Jimenez D, Muriel A, Bikdeli B, et al. Clinical characteristics and 3-month outcomes in cancer patients with incidental versus clinically suspected and confirmed pulmonary embolism. European Respiratory Journal. 2021;58(1).
4. Barco S, Valerio L, Ageno W, Cohen AT, Goldhaber SZ, Hunt BJ, et al. Age-sex specific pulmonary embolism-related mortality in the USA and Canada, 2000–18: an analysis of the WHO Mortality Database and of the CDC Multiple Cause of Death database. The Lancet Respiratory Medicine. 2021;9(1):33–42.
5. Lucassen WA, Beenen LF, Büller HR, Erkens PM, Schaefer-Prokop CM, van den Berk IA, et al. Concerns in using multi-detector computed tomography for diagnosing pulmonary embolism in daily practice. A cross-sectional analysis using expert opinion as reference standard. Thrombosis research. 2013;131(2):145–9.
6. Lim KY, Kligerman SJ, Lin CT, White CS. Missed pulmonary embolism on abdominal CT. American Journal of Roentgenology. 2014;202(4):738–43.
7. Kwok CS, Wong CW, Lovatt S, Myint PK, Loke YK. Misdiagnosis of pulmonary embolism and missed pulmonary embolism: A systematic review of the literature. Health Sciences Review. 2022;3:100022.
8. Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. Neurocomputing. 2021;444:92–110.
9. Kshatri SS, Singh D. Convolutional neural network in medical image analysis: a review. Archives of Computational Methods in Engineering. 2023;30(4):2793–810.
10. Wiklund P, Medson K, Elf J. Incidental pulmonary embolism in patients with cancer: prevalence, underdiagnosis and evaluation of an AI algorithm for automatic detection of pulmonary embolism. European Radiology. 2023;33(2):1185–93.
11. Stamate E, Piraianu A-I, Ciobotaru OR, Crassas R, Duca O, Fulga A, et al. Revolutionizing cardiology through artificial intelligence—Big data from proactive prevention to precise diagnostics and cutting-edge treatment—A comprehensive review of the past 5 years. Diagnostics. 2024;14(11):1103.
12. Mohanarajan M, Salunke PP, Arif A, Gonzalez PMI, Ospina D, Benavides DS, et al. Advancements in Machine Learning and Artificial Intelligence in the Radiological Detection of Pulmonary Embolism. Cureus. 2025;17(1).
13. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. bmj. 2003;327(7414):557–60.
14. Cheikh AB, Gorincour G, Nivet H, May J, Seux M, Calame P, et al. How artificial intelligence improves radiological interpretation in suspected pulmonary embolism. European radiology. 2022;32(9):5831–42.
15. Teigen CL, Maus TP, Sheedy 2nd P, Stanson AW, Johnson CM, Breen JF, et al. Pulmonary embolism: diagnosis with contrast-enhanced electron-beam CT and comparison with pulmonary angiography. Radiology. 1995;194(2):313–9.
16. Xi L, Xu F, Kang H, Deng M, Xu W, Wang D, et al. Clot ratio, new clot burden score with deep learning, correlates with the risk stratification of patients with acute pulmonary embolism. Quantitative Imaging in Medicine and Surgery. 2024;14(1):86–97.
17. Liu W, Liu M, Guo X, Zhang P, Zhang L, Zhang R, et al. Evaluation of acute pulmonary embolism and clot burden on CTPA with deep learning. European radiology. 2020;30(6):3567–75.
18. Lanza E, Ammirabile A, Francone M. nnU-Net-based deep-learning for pulmonary embolism: detection, clot volume quantification, and severity correlation in the RSPECT dataset. European Journal of Radiology. 2024;177:111592.
19. Diaz-Lorenzo I, Alonso-Burgos A, Friera Reyes A, Pacios Blanco RE, de Benavides Bernaldo de Quiros MdC, Gallardo Madueño G. Current role of CT pulmonary angiography in pulmonary embolism: A state-of-the-art review. Journal of Imaging. 2024;10(12):323.
20. Li L, Peng M, Zou Y, Li Y, Qiao P. The promise and limitations of artificial intelligence in CTPA-based pulmonary embolism detection. Frontiers in Medicine. 2025;12:1514931.
21. Li Y, Zhang L, Liu H, Li Y, Liu Z. Research progress of artificial intelligence and machine learning in pulmonary embolism. Frontiers in Medicine. 2025;12:1577559.
22. Naser AM, Vyas R, Morgan AA, Kalaiger AM, Kharawala A, Nagraj S, et al. Role of Artificial Intelligence in the Diagnosis and Management of Pulmonary Embolism: A Comprehensive Review. Diagnostics. 2025;15(7):889.
23. Mudrik A, Efros O. Artificial Intelligence and Venous Thromboembolism: a narrative review of applications, benefits, and limitations. Acta Haematologica. 2025.
24. Huang S-C, Kothari T, Banerjee I, Chute C, Ball RL, Borus N, et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ digital medicine. 2020;3(1):61.

Table 1: Characteristics of the included studies (all studies had retrospective observational design)

First author, Year, Country	Validation phase $^{\delta}$	Artificial intelligence			Image type	Age	Female n (%)	PE prevalence	Participants		
		Model	Categorization	Device					Total	PE +	PE _
Wu, 2025, China	Both	Deep learning (YOLOv8)	CNN	GE	CTPA	63.7±13.28	950 (47.5)	50%	2000	1000	1000
Graeve, 2025, Switzerland	Internal	Deep CNN (AIDOC)	CNN	Canon	Contrast CT	69 (NR)	794 (40.4)	2.2%	1964	44	1920
Zhu, 2024, China	Both	3D CNN (EMB-Net)	CNN	GE	CTPA	62.9±16.8	94 (57.7)	20.2%	163	33	130
Vallée, 2024, France	Internal	CINA-PE v1.0.3	Hybrid	Siemens	CTPA	NR	NR	15.8%	196	31	165
Tulum, 2024, Turkey	Internal	MLP	FNN	NR	CT + Perfusion	NR	NR	54%	37	20	17
Da Silva, 2024, Brazil	Both	RNN-LSTM	RNN	GE/Toshiba	CTPA	NR	NR	56.8%	160	91	69
Langius-Wiffen, 2024, Netherlands	External	CNN	CNN	Philips	CTPA	60.9±16.6	128 (54)	30.6%	238	73	165
Langius-Wiffen, 2024, Netherlands	External	3D CNN	CNN	Philips	Mono CTPA	NR	NR	35%	114	40	74
Koul, 2024, India	Internal	EfficientNetV2M	CNN	NR	X-ray / CT	NR	NR	51.4%	877	451	426
Kahraman, 2024, Sweden	Both	3D U-Net	CNN	Multi-vendor	CTPA	NR	NR	18.8%	679	128	551
Ayobi, 2024, USA	Internal	CINA-PE v1.0.5	Hybrid	Multi-vendor	CTPA	62.5±16.5	519 (43.2)	14.7%	1203	178	1025
Ammari, 2024, France	Internal	CINA-PE	Hybrid	GE/Siemens	CTPA	60.86±12.44	1060 (34.8)	1.11%	3047	37	3010
Zaazoue, 2023, USA	Internal	Deep CNN (AIDOC)	CNN	NR	CTPA	64 (53–74)	939 (63)	35.2%	1491	526	965
Topff, 2023, Netherlands	Internal	Deep CNN (AIDOC)	CNN	NR	Contrast CT	62±12	5605 (47.9)	1.22%	11702	143	11559
Grenier, 2023, France	Internal	CINA-PE v1.0.3	Hybrid	Multi-vendor	CTPA	62.5±16.3	181 (47)	48%	387	186	201
Huhtanen, 2022, Finland	Internal	InceptionResNet V2	CNN	NR	CTPA	NR	NR	47.5%	204	97	107
Ben Cheikh, 2022, France	Internal	Deep CNN (AIDOC)	CNN	GE	CTPA	65.4±18.9	144 (57.3)	75.3%	252	190	62
Batra, 2022, USA	Internal	Deep CNN (AIDOC)	CNN	NR	Contrast CT	53.2±14.5	19 (52.4)	97.2%	37	36	1
Ajmera, 2022, India	Internal	U-Net + Xception	Hybrid	Philips	CTPA	49.7±17.4	105 (42.2)	21.9%	251	55	196
Wildman-Tobriner, 2021, USA	Internal	Coread AI	Hybrid	GE/Siemens	Non-contrast CT	58.1±25.1	3269 (51.1)	1.23%	6398	79	6319
Schmuelling, 2021, Switzerland	Internal	Deep CNN (AIDOC)	CNN	NR	CTPA	64.5±17.9	234 (56.9)	22.5%	412	93	318
Müller-Peltzer, 2021, Germany	Internal	Syngo.via	Hybrid	Siemens	CTPA	NR	NR	14.8%	1229	182	1047
Weikert, 2020, Switzerland	Internal	ResNet	CNN	Multi-vendor	CTPA	NR	NR	15.8%	1465	232	1233
Wittenberg, 2010, Netherlands	Internal	ML algorithm	FNN	Philips	CTPA	57±24.8	140 (50.4)	24.4%	278	68	210
Reinartz, 2006, Germany	Internal	Registration algorithm	Hybrid	Siemens	V/Q scan	NR	NR	41.5%	53	22	31

Data are presented as mean ± standard deviation, median (interquartile range (IQR)), or number (%). PE: Pulmonary Embolism; CTPA: Computed Tomography Pulmonary Angiography; DL: deep learning; NR: not reported; DCNN: Deep Convolutional Neural Network; RNN-LSTM: Recurrent Neural Network - Long Short-Term Memory; FNN: Feedforward Neural Network; IV: intravenous; V/Q: ventilation -perfusion.$: AIDOC is a commercial AI platform; #: The CINA-PE model is a commercial AI tool specifically designed to assist in the diagnosis of Pulmonary Embolism; $^{\delta}$: 22 studies reported only internal validation ,2 reported only external validation.
Four studies reported both internal and external validation.

Table 2: Meta-regressions for different variables

Variable	Coefficient (95% CI)	P	I^2 (%)	Heterogeneity (%)
Internal validation				
Mean age	0.044 (-0.028 to 0.117)	0.211	36.91	0.00
Female percentage	-0.012 (-0.066 to 0.041)	0.630	41.16	0.00
Prevalence of PE	-0.055 (-0.097 to -0.012)	0.013	95.25	21.07

CI: confidence interval; PE: pulmonary embolism.

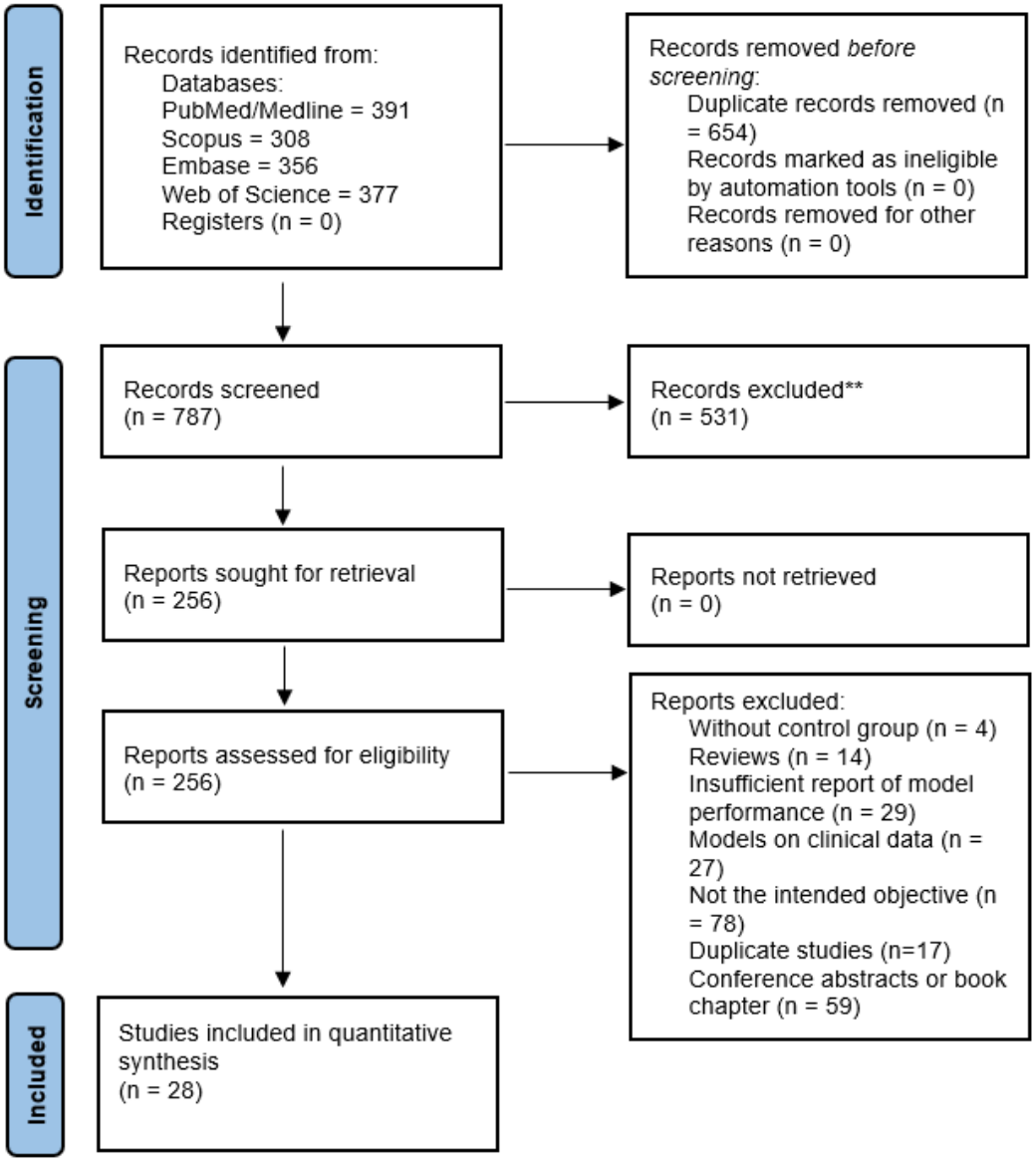

Figure 1: Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) flow diagram of the literature search and study selection process.

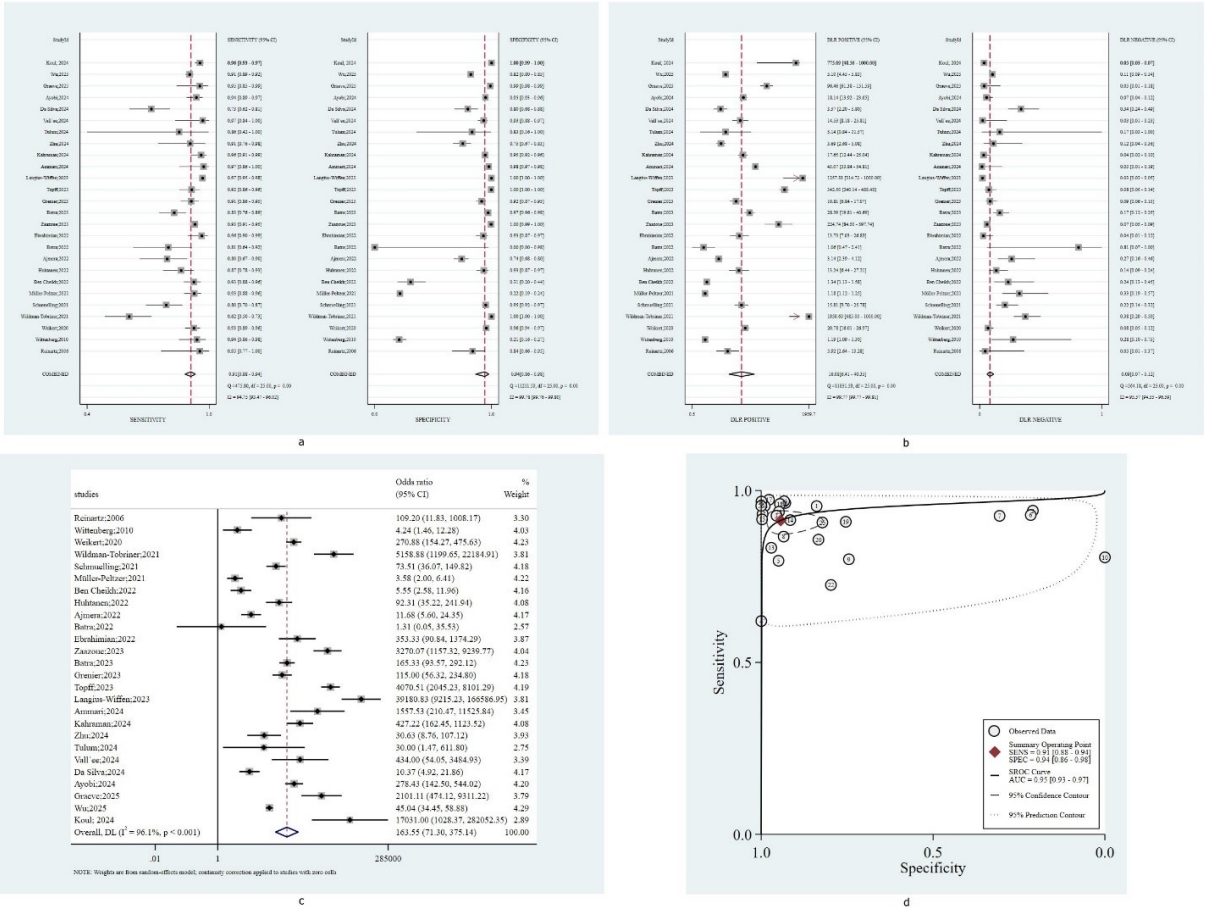

Figure 2: Forest plot (random-effects model) depicting diagnostic accuracy measures of artificial intelligence (AI) models for pulmonary embolism (PE) in internal validation phase: (a) Sensitivity and Specificity; (b) Positive and Negative likelihood ratio; (c) Diagnostic Odds Ratio (DOR); (d) Summary Receiver Operating Curve (SROC). CI: Confidence Interval; DLR: Diagnostic Likelihood Ratio; AUC: Area Under the Curve; SENS: sensitivity; SPEC: specificity.

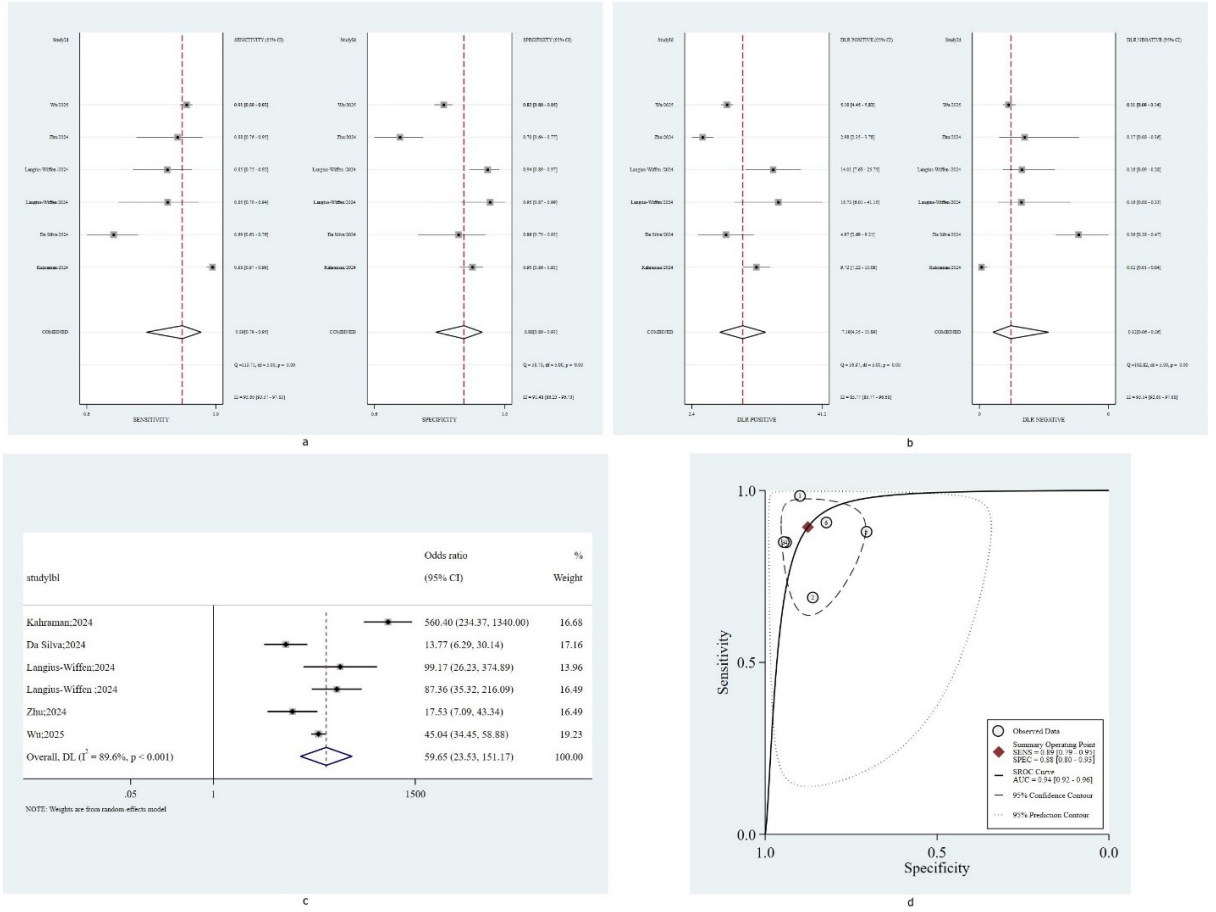

Figure 3: Forest plot (random-effects model) depicting diagnostic accuracy measures of artificial intelligence (AI) models for pulmonary embolism (PE) in external validation phase: (a) Sensitivity and Specificity; (b) Positive and Negative likelihood ratio; (c) Diagnostic Odds Ratio (DOR); (d) Summary Receiver Operating Curve (SROC). CI: Confidence Interval; DLR: Diagnostic Likelihood Ratio; AUC: Area Under the Curve; SENS: sensitivity; SPEC: specificity.

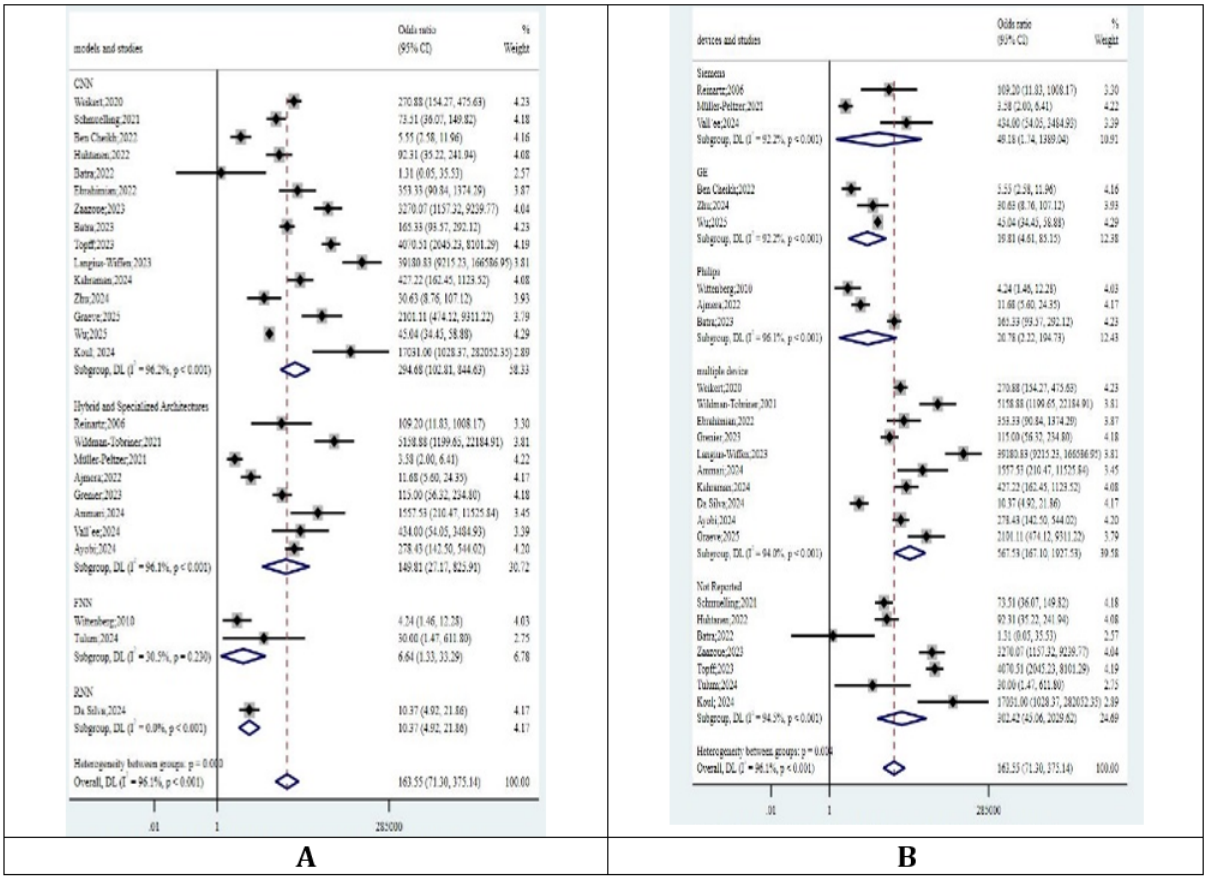

Figure 4: Subgroup Forest plot (random-effects model) depicting diagnostic accuracy measures of artificial intelligence (AI) models for pulmonary embolism (PE) in internal validation phase: (A) AI models, (B) devices used for imaging. CI: Confidence Interval; CNN: Convolutional Neural Networks; FNN: Feedforward Neural Network; RNN: Recurrent Neural Network.

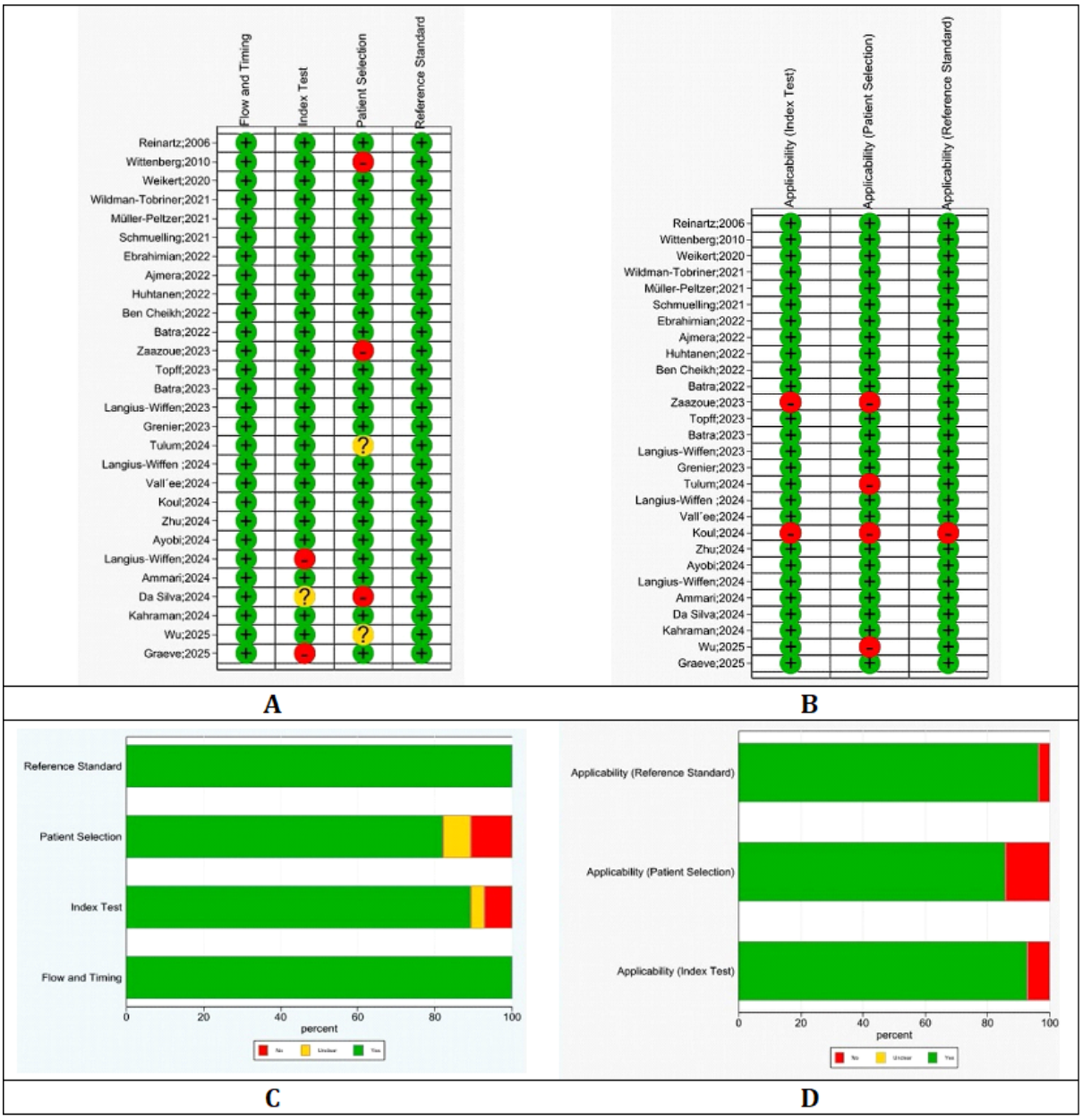

Figure 5: Risk of bias (ROB) and applicability domains of included studies using QUADAS-2 tool (A) detailed ROB domain results; (B) detailed applicability domain results; (C) pooled ROB domain results; (D) pooled applicability domain results.

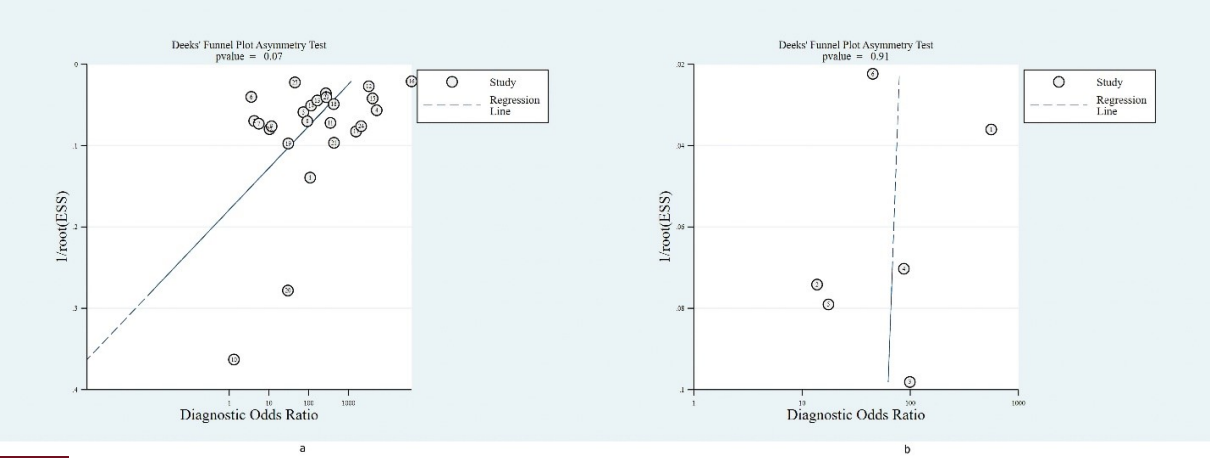

Figure 6: Deeks' Funnel plot for publication bias assessment (a) internal validation phase; (b) external validation phase.