Health Innovation Journal - Global

To Cite: Tahir HN, Ullah N, Tahir S, Alaklabi S, Rizvi SB, Tahir M, Ali Y. Diagnostic accuracy of artificial intelligence for paroxysmal atrial fibrillation detection using ECGs: a systematic review and meta-analysis. HIJ Global. 2026;1(1): 52-63.

Received: 03 November 2025
Accepted: 16 March 2026
Published: 19 March 2026

Correspondence to:
Naseer Ullah
khannasir965@gmail.com

Diagnostic Accuracy of Artificial Intelligence for Paroxysmal Atrial Fibrillation Detection Using ECGs: A Systematic Review and Meta-Analysis

Hasan Nawaz Tahir[1], Naseer Ullah[2*], Shahnawaz Tahir[3], Saad Alaklabi[4], Syeda Bushra Rizvi[5], Mursala Tahir[6], Yousaf Ali[7]

ABSTRACT

Background: Paroxysmal atrial fibrillation (PAF) can cause stroke, weaken the heart muscle, and lead to blood clot formation. The diagnostic method for PAF is electrocardiography (ECG), which requires physician interpretation. However, the increasing use of artificial intelligence (AI) in healthcare has enabled automated diagnosis of cardiac arrhythmias. PAF, often underdiagnosed owing to its transient nature, can now potentially be detected by AI models applied to ECGs. This study evaluated the diagnostic performance of AI models in detecting PAF.

Methods: PubMed, Cochrane Library, and Google Scholar were searched, and studies were included if they assessed the diagnostic accuracy of AI models for PAF. Sensitivity and specificity were the primary outcomes for assessing diagnostic accuracy, and subgroup analyses were performed based on sample size.

Results: A total of 956,664 ECGs from eight studies were assessed using AI models. The pooled sensitivity and specificity of the AI models for PAF detection were 0.8585 (95% CI: 0.85, 0.86) and 0.9771 (95% CI: 0.97, 0.98), respectively. Subgroup analysis revealed higher sensitivity in smaller-sample-size studies (0.91, 95% CI: 90.3, 91.92) compared with larger-sample-size studies (0.8579, 95% CI: 85.69, 85.89). Conversely, larger-sample-size studies demonstrated higher specificity (0.9775, 95% CI: 97.73, 97.77) than smaller-sample-size studies (0.90, 95% CI: 89.42, 90.69).

Conclusion: These findings suggest that sample size may influence the diagnostic accuracy of AI models. AI models exhibit high diagnostic accuracy for detecting PAF, offering an efficient alternative for clinical diagnosis.

Keywords: Artificial Intelligence; Deep Learning; Algorithm; Atrial Fibrillation; Smart Watch; Wearable Devices; ECG.

INTRODUCTION

Paroxysmal atrial fibrillation (PAF) is a type of atrial fibrillation characterized by self-terminating, irregular, and rapid heart rhythms, accounting for 25-62% of all cases of atrial fibrillation [1, 2]. It is often asymptomatic and can be undiagnosed until serious complications occur. Approximately 15%–20% of stroke patients have AF [3]. Early detection and management of PAF are crucial for decreasing these risks and improving patient outcomes. Traditional PAF detection methods depend heavily on manual interpretation, which can be subject to human error, particularly in areas where human resources are scarce. This highlights the urgent need for reliable and accessible diagnostic methods.

The era of artificial intelligence has brought about significant advancements in the medical field, particularly in the analysis of diagnostic tests. AI models have shown promising results in automating the detection of cardiac arrhythmias [4], offering the potential for increased accuracy and efficiency. These models are trained on large datasets to recognize patterns indicative of PAF, thus reducing their dependence on manual interpretation. AI not only enhances diagnostic precision, but also reduces the workload of healthcare professionals, enabling a more consistent and faster diagnosis.

Despite these potential benefits, the diagnostic accuracy of AI models for detecting PAF remains a critical area of investigation. Limited previous studies have reported a sensitivity and specificity of > 80% for AF [5-7], but there is no meta-analysis evaluating PAF separately, highlighting the need for evaluation of AI performance for PAF. This systematic review and meta-analysis aimed to assess the pooled sensitivity and specificity of AI models for PAF detection. Additionally, it explores the impact of sample size on the diagnostic performance of artificial intelligence, providing insights into the conditions under which AI models perform optimally.

METHODS

Search Strategy

A systematic search of PubMed, Cochrane Library, and Google Scholar was performed to identify studies evaluating artificial intelligence for PAF detection published from January 2014 to November 2024. The search terms included “artificial intelligence”, “deep learning”, “machine learning”, “ECG”, “paroxysmal atrial fibrillation”, and “atrial fibrillation”. Studies were included if they reported sensitivity and specificity data.

Inclusion Criteria

Observational and validation studies were included if they reported sensitivity and specificity outcomes and evaluated artificial intelligence screening for PAF in patients aged 20–95 years diagnosed with PAF. Studies were excluded if they did not report the outcomes of interest (sensitivity and specificity), if the authors did not respond, or if the full text was unavailable.

Study Selection

Two independent reviewers initially screened the titles and abstracts, and full texts were obtained for those that met the inclusion criteria or were uncertain. Full texts were then reviewed by the same reviewers, and any disagreements were resolved through mutual discussion or with the help of another reviewer. The selection process is documented in the PRISMA flow diagram shown in Fig. 1.

Data Extraction

The data extracted for AI-based detection of PAF included study design, publication year, study setting, country, ECG monitoring device, algorithm, age, sample size (small sample < 10,000, large sample > 10,000), number of ECGs, and the values for this study (true positive, true negative, false positive, and false negative). Data extraction was performed independently by two reviewers to minimize bias, and disagreements were resolved through discussion or input from a third reviewer. Data were stored in Excel and analyzed using Review Manager 5.4 and R. The Newcastle-Ottawa Scale was used to assess the risk of bias in observational studies, and the QUADAS-2 tool was used for validation studies.

Quality Assessment

Quality assessment was performed independently by two reviewers to evaluate different types of bias, including selection, comparability, exposure, index test, reference standard, and flow and timing bias. The risk-of-bias assessment tools used were QUADAS-2 for validation studies and the Newcastle-Ottawa Scale for observational studies, to report the strength of evidence for the results of this meta-analysis. The Newcastle-Ottawa Scale was evaluated using a scoring system in

which < 4 stars indicated high risk, 4-6 stars moderate risk, and 7-9 stars low risk.

Statistical Analysis

Statistical analyses were conducted using Review Manager 5.4 and R. Pooled sensitivity and specificity were calculated using a bivariate random-effects model. Heterogeneity among the studies was assessed using the I^2 statistic. Subgroup analyses were performed to investigate the potential impact of sample size and AI model type on diagnostic accuracy. All results were reported with 95% confidence intervals (CIs), and a p-value < 0.05 was considered statistically significant.

RESULTS

Study Characteristics

Eight studies were included in this meta-analysis, comprising 220,348 participants across 6 studies, with 2 studies [8, 9] not reporting their number of participants; sample sizes ranged from 152 to 142,310 across different countries. A total of 956,664 ECGs from the 8 studies were tested using different AI models for PAF detection. Five studies used the CNN model [8-12], while different algorithms were used in one study each: DNN [13], EfficientNet-V2 network [14], XGBoost [15], and SVM [9]. Two of the 8 studies were validation studies [9, 13], and six were retrospective observational studies [8, 10-12, 14, 15], as given in **Table 1**.

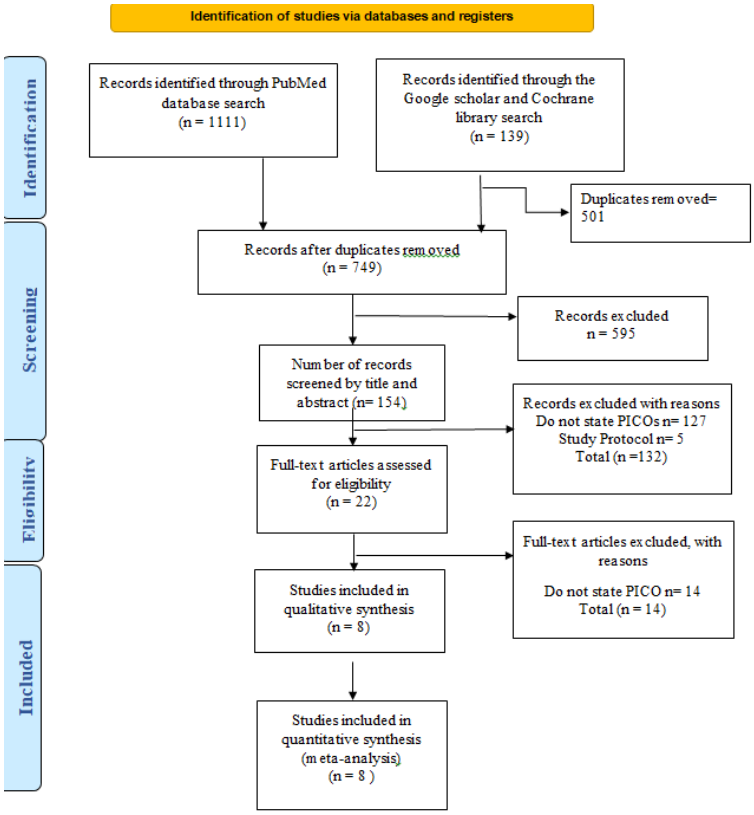

Fig. 1. PRISMA flow diagram for study selection.

Table 1. Characteristics of included studies.

Author	Year	Study Design	Study Setting	ECG Monitoring Device	Algorithm	Country	Age (Mean ± SD)	N	ECGs (n)
Gruwez et al., 2023	2023	Validation study	Inpatient and outpatient	12-lead ECG	DNN	United States	Not mentioned	142,310	494,042
Gilon et al., 2023	2023	Retrospective observational study	Outpatient	12-lead ECG	1D CNN and XGBoost	Belgium	72 ± 11 years	152	167
Raghunath et al., 2023	2023	Retrospective observational study	Community-based	KardiaMobile 6L device	CNN and standard dense neural network layers	United States	58.14 ± 3.1	73,861	267,469
Chen et al., 2024	2024	Retrospective observational study	Public dataset	Single-lead ECG	CNN	China	Not mentioned	Not mentioned	185,204
Zhou et al., 2024	2024	Retrospective observational study	Hospital	12-lead ECG	EfficientNet-V2 network	China	62.8 ± 13.9	2,192	5,688
Kisohara et al., 2021	2020	Retrospective observational study	Community-based	24-hour ambulatory Holter ECG	CNN	Japan	Not mentioned	215	215
Wang et al., 2022	2022	Validation study	Public dataset	Lead II of long-term dynamic ECG signals	SVM and CNN	China	Not mentioned	Not mentioned	1,436
Jiang et al., 2023	2023	Retrospective observational study	Hospital setting	12-lead ECG	CNN	China	58.8	1,618	2,443

CNN: convolutional neural network; DNN: deep neural network; SVM: support vector machine; ECG: electrocardiogram.

a. Only test datasets used for external validation were included in the meta-analysis; datasets used for training and internal validation were excluded. b. Sensitivity and specificity were calculated from AI model outputs. c. External validation of AI models was conducted in clinical settings such as hospital-based, outpatient, or community-based screening programs, as specified in the individual studies.

Sensitivity

Artificial intelligence diagnostic accuracy was analyzed for the diagnosis of PAF. In the included studies, the SROC curves show variable sensitivity and specificity, particularly for small sample sizes, in Fig. 2 and Fig. 3. Most of the studies exhibit a narrow CI (mainly larger-sample-size studies), and all of the small-sample-size studies indicate a wider CI in sensitivity, as do 2 studies in specificity [10, 14], contributing to variable results in small sample sizes with consistent results in large sample sizes.

Among all the studies, sensitivity ranged from 0.25 [10] to 0.95 [11], with a pooled sensitivity of 0.82 (95% CI: 0.74, 0.87). There is notable consistency of the outcomes in most of the studies, indicating good performance of AI-based screening of PAF, as shown in **Table 2**, Fig. 4.

Specificity

Specificity for the studies ranged between 0.57 [10] and 1.00 [11], with a pooled effect of 0.95 (95% CI: 0.89, 0.98), suggesting good performance. Most of the studies show consistent results with little variability, indicating good performance for the diagnosis of PAF, as shown in **Table 2**, Fig. 5.

Overall, the diagnostic accuracy of artificial intelligence, as shown in the SROC plots, indicates robust diagnostic capability for detecting PAF, mainly in larger models, with most of the studies showing higher sensitivity and specificity, as can be seen in Fig. 2 and Fig. 3.

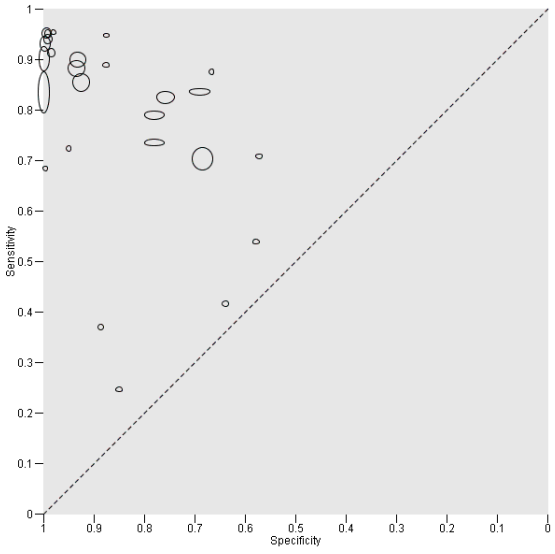

Fig. 2. Overall SROC plot.

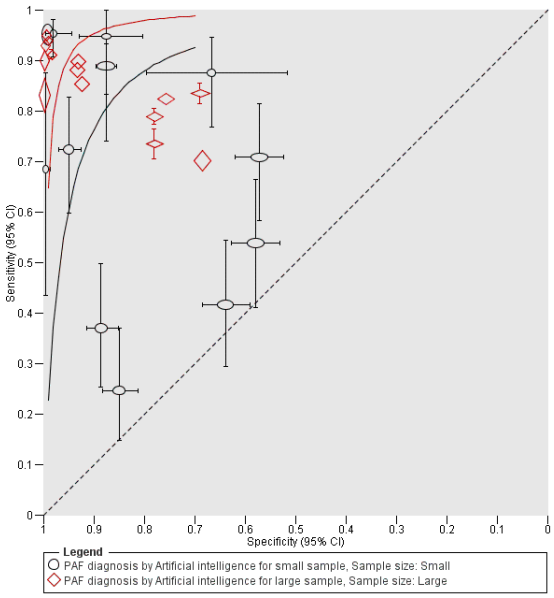

Fig. 3. SROC plot for subgroup analysis.

Table 2. Results for outcomes.

Study	Model	TP	TN	FP	FN	Sensitivity (95% CI)	Specificity (95% CI)
Gruwez et al., 2023	DNN for AF prevalence 3%	632	21,607	6,096	228	0.73 [0.70, 0.76]	0.78 [0.78, 0.78]
	DNN for AF prevalence 9%	2,028	20,271	5,717	542	0.79 [0.77, 0.80]	0.78 [0.77, 0.79]
	DNN for AF prevalence 30%	6,756	15,419	4,926	1,442	0.82 [0.82, 0.83]	0.76 [0.75, 0.76]
	DNN for AF prevalence 5%	1,091	17,121	7,695	215	0.84 [0.81, 0.86]	0.69 [0.68, 0.70]
Gilon et al., 2023	XGBoost	143	147	3	7	0.95 [0.91, 0.98]	0.98 [0.94, 1.00]
	CNN	3,898	4,063	33	198	0.95 [0.94, 0.96]	0.99 [0.99, 0.99]
Raghunath et al., 2023	Alivecor KardiaMobile	23,210	15,034	6,934	9,821	0.70 [0.70, 0.71]	0.68 [0.68, 0.69]
Chen et al., 2024	Model M1	23,779	40,753	3,334	4,057	0.85 [0.85, 0.86]	0.92 [0.92, 0.93]
	Model M2	25,027	41,086	3,001	2,809	0.90 [0.90, 0.90]	0.93 [0.93, 0.93]
	Model M3	24,547	41,192	2,895	3,289	0.88 [0.88, 0.89]	0.93 [0.93, 0.94]
Zhou et al., 2024	AI-pECG model	56	32	16	8	0.88 [0.77, 0.94]	0.67 [0.52, 0.80]
Kisohara et al., 2021	CNN for SWL 10	158,073	903,219	930	31,584	0.83 [0.83, 0.84]	1.00 [1.00, 1.00]
	CNN for SWL 20	79,173	458,259	692	8,750	0.90 [0.90, 0.90]	1.00 [1.00, 1.00]
	CNN for SWL 50	31,620	184,063	671	2,369	0.93 [0.93, 0.93]	1.00 [1.00, 1.00]
	CNN for SWL 100	15,838	92,160	527	807	0.95 [0.95, 0.95]	0.99 [0.99, 0.99]
	CNN for SWL 200	7,944	45,759	423	509	0.94 [0.93, 0.94]	0.99 [0.99, 0.99]
	CNN for SWL 500	3,230	18,006	281	308	0.91 [0.90, 0.92]	0.98 [0.98, 0.99]
	CNN	47	399	21	18	0.72 [0.60, 0.83]	0.95 [0.92, 0.97]

Study	Model	TP	TN	FP	FN	Sensitivity (95% CI)	Specificity (95% CI)
Jiang et al., 2023	Apple	35	243	177	30	0.54 [0.41, 0.66]	0.58 [0.53, 0.63]
	Base-AF2	46	240	180	19	0.71 [0.58, 0.81]	0.57 [0.52, 0.62]
	CAAP-AF	16	357	63	49	0.25 [0.15, 0.37]	0.85 [0.81, 0.88]
	DR-FLASH	27	268	152	38	0.42 [0.29, 0.54]	0.64 [0.59, 0.68]
	MB-LATER	24	372	48	41	0.37 [0.25, 0.50]	0.89 [0.85, 0.91]
Wang et al., 2022	SVM	13	486	2	6	0.68 [0.43, 0.87]	1.00 [0.99, 1.00]
Total		407,423	1,921,607	44,966	67,164	0.8585 (0.85, 0.86)	0.9771 (0.97, 0.98)

CI: confidence interval; TP: true positive; TN: true negative; FP: false positive; FN: false negative; SWL: segment window length; DNN: deep neural network; SVM: support vector machine.

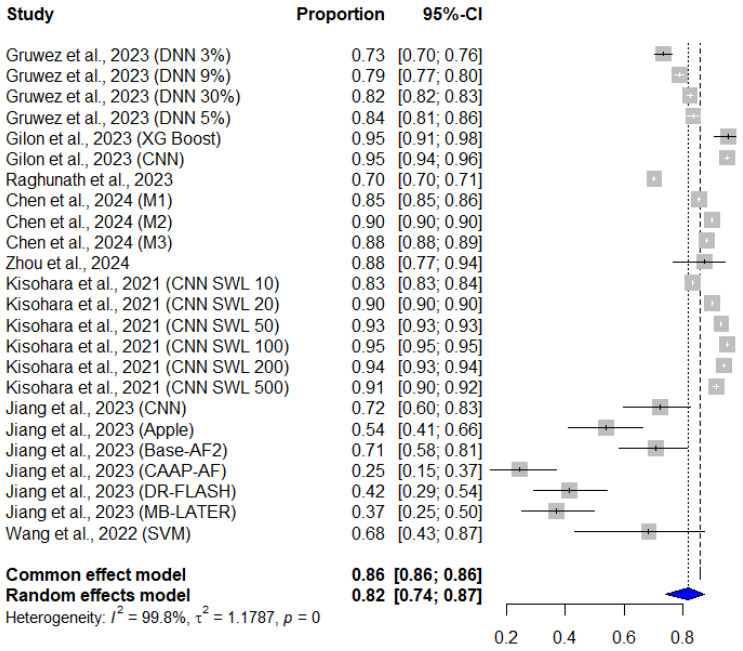

Fig. 4. Forest plot for overall sensitivity.

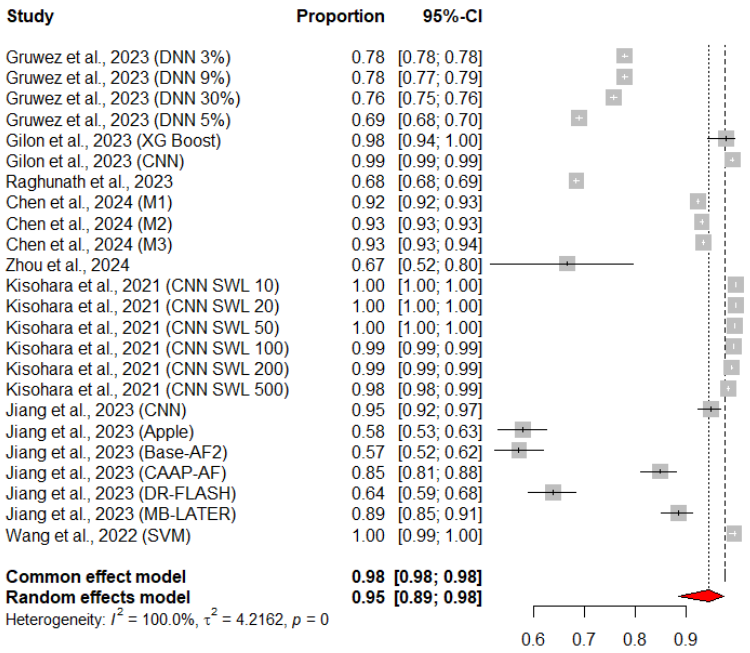

Fig. 5. Forest plot for overall specificity.

Subgroup Analysis

In the subgroup analysis, small-sample-size studies showed sensitivity ranging from 0.25 [10] to 0.88 [14], with a pooled effect of 0.71 (95% CI: 0.51, 0.85), while studies with a large sample size showed results between 0.70 [12] and 0.95 [11], with a pooled sensitivity of 0.87 (95% CI: 0.83, 0.90), indicating good performance of AI-based diagnostic accuracy for large-sample-size studies. There is notable variability in some small-sample-size studies and consistency in most large-sample-size studies, as shown in **Table 2**, Fig. 6, and Fig. 8.

Specificity for small sample sizes showed a range from 0.57 [10] to 1.00 [9], with a pooled specificity of 0.91 (95% CI: 0.76, 0.97), while large sample sizes showed robust specificity, with a range between 0.68 [12] and 1.00 [11] and a pooled effect of 0.96 (95% CI: 0.90, 0.99), suggesting good performance. This indicates that larger sample sizes enhance specificity, potentially due to reduced variability and more robust AI model training, as shown in Fig. 7 and Fig. 9.

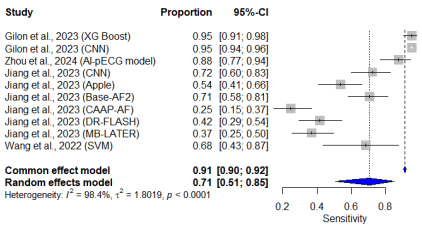

Fig. 6. Sensitivity, small sample size.

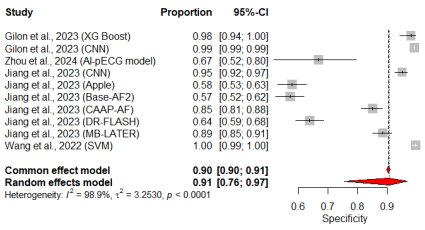

Fig. 7. Specificity, small sample size.

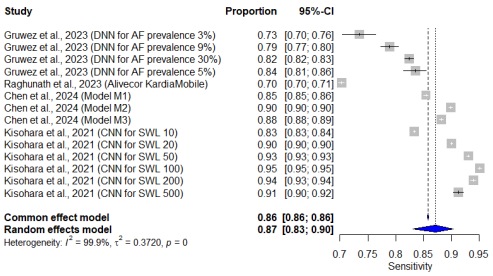

Fig. 8. Sensitivity, large sample size.

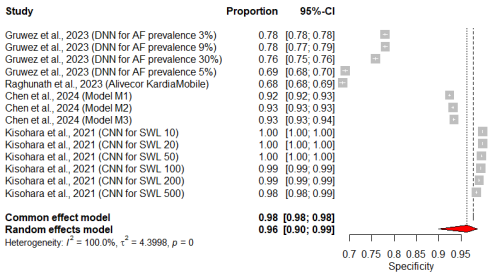

Fig. 9. Specificity, large sample size.

Risk of Bias

A total of 8 included studies were assessed for risk of bias. Two of the eight studies were validation studies, assessed using the QUADAS-2 tool [9, 13]; one study showed low risk in all domains, while the second study showed moderate risk in the first and fourth domains, with an overall unclear risk, as shown in Fig. 10 and Fig. 11.

Six studies were retrospective observational studies [8, 10-12, 14, 15], assessed using the Newcastle-Ottawa Scale; five studies showed a low risk of bias [10-12, 14, 15], while one study [8] showed an overall high risk of bias, as can be seen in **Table 3**.

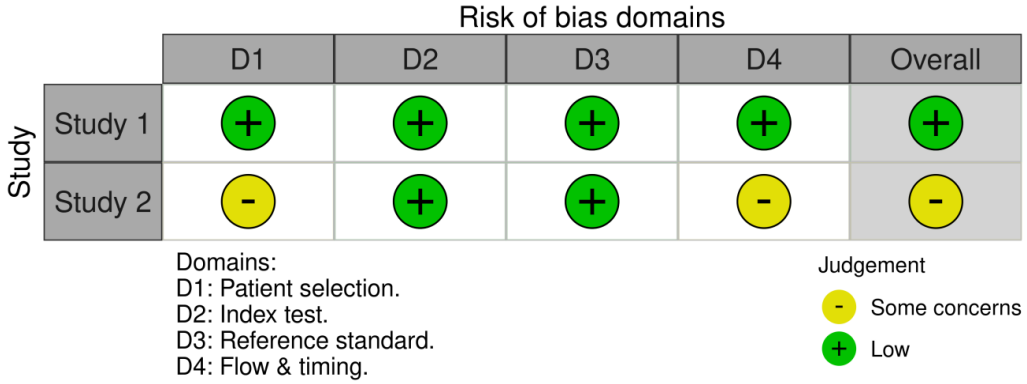

Fig. 10. Risk-of-bias assessment traffic-light plot for the QUADAS-2 tool.

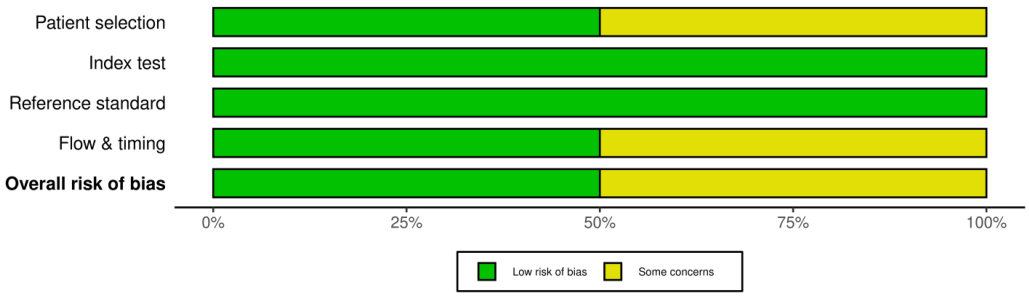

Fig. 11. Risk-of-bias assessment summary plot for the QUADAS-2 tool.

Table 3. Asterisk rating in observational studies according to the Newcastle-Ottawa Scale (NOS) tool.

Study	Adequacy of Selection				Comparability	Outcome Assessment		Total Score
	Is the Case Definition Adequate?	Representativeness of the Cases	Selection of Controls	Definition of Controls	Comparability of Cases and Controls on the Basis of Design or Analysis	Assessment of Exposure	Non-Response Rate	
Raghunath et al., 2023	*	*	*	*	**	*		7.0/9.0
Chen et al., 2024				*	*	*		3.0/9.0
Zhou et al., 2024	*	*		*	**	**		7.0/9.0
Kisohara et al., 2021	*	*	*	*	*	**		7.0/9.0
Jiang et al., 2023	*	*		*	**	**	*	8.0/9.0
Gilon et al., 2023	*	*	*	*	**	**		8.0/9.0

*NOS: Newcastle-Ottawa Scale; **: indicates two stars in the NOS for the comparability and outcome assessment domains; *: indicates one star for selection, comparability, or outcome assessment based on NOS guidelines for cohort studies. Scoring system: < 4 for high risk, 4-6 for moderate risk, and 7-9 for low risk.*

DISCUSSION

This systematic review and meta-analysis suggest that artificial intelligence models using electrocardiograms exhibit high diagnostic accuracy for detecting paroxysmal atrial fibrillation. The overall pooled sensitivity and specificity demonstrate robust performance of AI in diagnosing PAF. These findings are significant in the context of early detection, which is critical for preventing complications such as stroke and heart failure associated with PAF.

The subgroup analysis reveals differences in diagnostic performance across sample sizes. Small-sample-size studies show more variability in sensitivity, whereas larger-sample-size studies exhibit more consistent results, pointing to the influence of sample size on AI performance. Small-sample studies may face limitations in data diversity, potentially affecting the model's ability to generalize across different populations. Conversely, larger-sample-size studies likely contribute to the robustness and reliability of AI-based diagnosis, as reflected in the consistent results. These results indicate that AI models can be a promising tool for PAF diagnosis in clinical use.

LIMITATIONS AND IMPLICATIONS

Several limitations are noted in this meta-analysis. First, the majority of included studies are observational, which may introduce biases related to study design and patient selection. Second, the included studies vary in terms of AI models, sample size, and study settings, which may contribute to the high heterogeneity in the results.

Significant implications for clinical practice and future research arise from the high diagnostic accuracy of AI models, which can be integrated into routine clinical use to assist in the early detection of PAF, particularly in limited-resource settings where manual interpretation of ECGs may be challenging. Future research should focus on standardizing AI models, including training data and performance metrics. Additionally, large-scale, multi-centre validation studies are needed to further clarify the clinical utility of AI models.

CONCLUSIONS

Artificial intelligence demonstrates high diagnostic accuracy in detecting PAF using ECGs. Sample size plays a crucial role in diagnostic accuracy, emphasizing that large-sample validation studies should be conducted to ensure effectiveness across diverse clinical settings. Development, standardization, and broader validation of artificial intelligence models are essential to optimize their diagnostic performance and facilitate widespread implementation in healthcare.

LIST OF ABBREVIATIONS

PAF: paroxysmal atrial fibrillation. AI: artificial intelligence.

AUTHOR AFFILIATIONS

[1]Department of Community Medicine, College of Medicine, Dawadmi, Shaqra University, Saudi Arabia;
[2]Department of Community Medicine, Khyber Medical College, Peshawar, Pakistan;
[3]Dow University of Health Sciences, Karachi, Pakistan;
[4]College of Science and Humanities, Dawadmi, Shaqra University, Saudi Arabia;
[5]Department of General Surgery, Jinnah Postgraduate Medical Center, Karachi, Pakistan;
[6]Department of Community Medicine, Liaquat National Hospital and Medical College, Karachi, Pakistan;
[7]College of Medicine, Dawadmi, Shaqra University, Saudi Arabia.

Ethical Considerations

There is no need for ethical approval, as this is a systematic review and meta-analysis of published data.

Funding

No funding was received for this study.

REFERENCES

1. Lip GY, Hee FL. Paroxysmal atrial fibrillation. QJM. 2001;94(12):665-678.
2. UpToDate. Paroxysmal Atrial Fibrillation. 2025. Available from: https://www.uptodate.com/contents/paroxysmal-atrial-fibrillation
3. American Heart Association. What Is Atrial Fibrillation (AFib or AF)? 2025. Available from: https://www.heart.org/en/health-topics/atrial-fibrillation/what-is-atrial-fibrillation-afib-or-af
4. Nagarajan VD, et al. Artificial intelligence in the diagnosis and management of arrhythmias. Eur Heart J. 2021;42(38):3904-3916.
5. Manetas-Stavrakakis N, et al. Accuracy of artificial intelligence-based technologies for the diagnosis of atrial fibrillation: a systematic review and meta-analysis. J Clin Med. 2023;12(20).
6. Menezes Junior ADS, et al. A scoping review of the use of artificial intelligence in the identification and diagnosis of atrial fibrillation. J Pers Med. 2024;14(11).
7. Attia ZI, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861-867.
8. Chen W, et al. Achieving real-time prediction of paroxysmal atrial fibrillation onset by convolutional neural network and sliding window on R-R interval sequences. Bioengineering (Basel). 2024;11(9).
9. Wang Y, et al. A two-step method for paroxysmal atrial fibrillation event detection based on machine learning. Math Biosci Eng. 2022;19(10):9877-9894.
10. Jiang J, et al. An artificial intelligence-enabled ECG algorithm for predicting the risk of recurrence in patients with paroxysmal atrial fibrillation after catheter ablation. J Clin Med. 2023;12(5).
11. Kisohara M, et al. Optimal length of R-R interval segment window for Lorenz plot detection of paroxysmal atrial fibrillation by machine learning. Biomed Eng Online. 2020;19(1):49.
12. Raghunath A, et al. Artificial intelligence-enabled mobile electrocardiograms for event prediction in paroxysmal atrial fibrillation. Cardiovasc Digit Health J. 2023;4(1):21-28.

13. Gruwez H, et al. Detecting paroxysmal atrial fibrillation from an electrocardiogram in sinus rhythm: external validation of the AI approach. JACC Clin Electrophysiol. 2023;9(8 Pt 3):1771-1782.

14. Zhou Y, et al. Screening tool for paroxysmal atrial fibrillation based on a deep-learning algorithm using printed 12-lead electrocardiographic records during sinus rhythm. Rev Cardiovasc Med. 2024;25(7):242.

15. Gilon C, et al. IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database. Sci Data. 2023;10(1):714.