nature medicine

Article

https://doi.org/10.1038/s41591-025-03516-x

Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography

Received: 6 May 2024

Accepted: 16 January 2025

Published online: 10 February 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Developments in ambulatory electrocardiogram (ECG) technology have led to vast amounts of ECG data that currently need to be interpreted by human technicians. Here we tested an artificial intelligence (AI) algorithm for direct-to-physician reporting of ambulatory ECGs. Beat-by-beat annotation of 14,606 individual ambulatory ECG recordings (mean duration = 14 ± 10 days) was performed by certified ECG technicians (n = 167) and an ensemble AI model, called DeepRhythmAI. To compare the performance of the AI model and the technicians, a random sample of 5,235 rhythm events identified by the AI model or by technicians, of which 2,236 events were identified as critical arrhythmias, was selected for annotation by one of 17 cardiologist consensus panels. The mean sensitivity of the AI model for the identification of critical arrhythmias was 98.6% (95% confidence interval (CI) = 97.7–99.4), as compared to 80.3% (95% CI = 77.3–83.3%) for the technicians. False-negative findings were observed in 3.2/1,000 patients for the AI model versus 44.3/1,000 patients for the technicians. Accordingly, the relative risk of a missed diagnosis was 14.1 (95% CI = 10.4–19.0) times higher for the technicians. However, a higher false-positive event rate was observed for the AI model (12 (interquartile range (IQR) = 6–74)/1,000 patient days) as compared to the technicians (5 (IQR = 2–153)/1,000 patient days). We conclude that the DeepRhythmAI model has excellent negative predictive value for critical arrhythmias, substantially reducing false-negative findings, but at a modest cost of increased false-positive findings. AI-only analysis to facilitate direct-to-physician reporting could potentially reduce costs and improve access to care and outcomes in patients who need ambulatory ECG monitoring.

In recent years, there have been rapid developments in ambulatory electrocardiogram (ECG) technology that enable markedly increased use of ambulatory ECG monitoring. At the same time, the importance of detecting brief, infrequent arrhythmias, particularly atrial fibrillation (AF), has been recognized[1,2]. Longer ECG recording duration and frequency lead to higher detection rates of arrhythmia[3–6], and extended ECG monitoring is recommended for patients with syncope[7,8] and individuals in whom screening for AF to prevent new-onset or recurrent stroke could be beneficial[9]. The number of patients that may benefit from rhythm monitoring is also growing, particularly with evidence that short-duration subclinical AF[10] may benefit from anticoagulation[1,11]. With the increasing availability of lower-cost devices, longer-term monitoring capabilities and the emergence of direct-to-consumer devices that provide irregular pulse notifications and record single-lead ECG intermittently, there has come a deluge of heart rhythm monitoring data that requires analysis[12,13]. Given the worldwide shortages of healthcare workers[14], this increased workload may overburden human ECG technician resources, possibly reducing the quality of heart rhythm annotations[15–18], leading to misdiagnosis, delayed treatment and adverse patient outcomes.

e-mail: linda.johnson@med.lu.se

https://doi.org/10.1038/s41591-025-03516-x

Table 1 | Performance of DeepRhythmAI and ECG technicians compared to the consensus panel of cardiologists for critical arrhythmias

	Accuracy (95% CI), %		True-positive rate/ sensitivity, % (95% CI)		True-negative rate/ specificity, % (95%CI)		PPV, % (95% CI)		NPV, %(95% CI)		F1 score, %	
	AI	Technician	AI	Technician	AI	Technician	AI	Technician	AI	Technician	AI	Technician
Overall average critical arrhythmias	98.1 (97.9–98.2)	98.4 (98.1–98.5)	**98.6 (97.7–99.4)**	80.3 (77.3–83.3)	98.1 (97.9–98.2)	**99.2 (99.0–99.3)**	71.3 (68.5–73.9)	**82.7 (79.4–85.6)**	**99.9 (99.9–100)**	99.1 (98.9–99.2)	82.7 (80.9–84.5)	81.5 (79.0–83.6)
VT ≥ 10 s	98.2 (98.1–98.3)	99.5 (99.4–99.6)	**98.0 (94.8–100)**	64.4 (54.9–73.8)	98.2 (98.1–98.3)	**99.8 (99.7–99.8)**	27.2 (22.8–32.3)	**67.7 (58.2–76.6)**	**99.98 (99.96–100)**	99.7 (99.6–99.8)	42.6 (37.1–48.6)	**66.0 (57.4–73.2)**
AF ≥ 30 s	97.2 (96.5–97.9)	97.4 (96.6–98.0)	**99.1 (97.7–100)**	90.5 (86.8–94.0)	96.9 (96.2–97.7)	**98.4 (97.8–98.9)**	82.3 (77.8–86.8)	88.9 (84.7–92.6)	**99.9 (99.7–100)**	98.6 (98.0–99.2)	90.0 (87.1–92.7)	89.7 (86.7–92.3)
SVT ≥ 30 s	97.4 (97.1–97.9)	96.1 (95.5–96.7)	**97.3 (94.9–99.1)**	62.9 (56.6–69.3)	97.4 (97.0–97.9)	98.1 (97.7–98.4)	70.6 (65.9–75.7)	65.8 (59.3–72.2)	**99.8 (99.7–99.9)**	97.8 (97.2–98.3)	**81.8 (78.3–75.2)**	64.3 (58.7–69.8)
Asystole ≥ 3.5 s	98.5 (98.2–98.7)	**99.2 (99.0–99.4)**	**100 (100–100)**	80.6 (75.0–86.0)	98.4 (98.2–98.6)	**99.8 (99.7–99.9)**	65.7 (60.5–70.4)	**91.2 (87.8–95.6)**	**100 (100–100)**	99.4 (99.2–99.6)	79.2 (75.4–82.6)	85.8 (82.1–89.5)
Third-degree AV block	99.3 (99.2–99.4)	99.5 (99.3–99.6)	**96.4 (92.5–99.2)**	52.6 (44.0–61.6)	99.3 (99.2–99.4)	**99.9 (99.8–99.9)**	51.2 (44.6–48.2)	**76.3 (67.1–85.4)**	**100 (99.9–100)**	99.6 (99.5–99.7)	66.9 (61.2–72.8)	62.2 (53.9–70.0)

The bold values denote nonoverlapping CIs between methods. NPV, negative predictive value; PPV, positive predictive value.

While it has widely been predicted that artificial intelligence (AI) will replace humans in some areas[19], the nearest examples in healthcare are in mammography, where AI can replace a second physician reader for mammograms[20–23], and in pathology, where AI tools improve pathologist accuracy and efficiency[24,25]. Implementation of an AI model that uses ECGs to alert physicians to high-risk hospitalized patients was recently shown to reduce mortality[26], and several machine learning-based models that use ECG data to predict arrhythmia have been developed[27,28]. AI holds considerable promise for arrhythmia diagnostics as it can rapidly analyze a large amount of data at low cost, provide consistent annotations without risk of mental fatigue and provide results in near real time[29]. Previous studies indicate that AI algorithms can be trained to detect and accurately classify arrhythmias on resting ECG and ambulatory ECG recordings[30,31], but no study has evaluated the role of AI in performing scanning and technical annotation of ambulatory ECG and providing results that can then be forwarded for physicians to review. Because AI-only reporting would mean that large amounts of ECG data would never be seen by a healthcare professional, such an AI model would need to have excellent negative predictive value for critical arrhythmias without generating unacceptable rates of false-positive annotations that would require physician review.

We designed the DeepRhythmAI for autonoMous Analysis of RhyThm INvestigatIon (DRAI MARTINI) study to test the DeepRhythmAI model for direct-to-physician reporting of ambulatory ECG data. The aim was to report on the performance of the DeepRhythmAI compared to technician analysis of ambulatory ECG data, including absolute rates of false-negative and false-positive detection for both the AI model and ECG technicians.

Results

The study population consisted of 14,606 patients (mean age = 65.5 ± 10 years, 42.8% males), who were monitored for a mean of 14 ± 10 days (Extended Data Fig. 1). Monitoring indications were provided through the device for 14,596 patients and are reported in Extended Data Table 1. The most common monitoring indications were palpitations, syncope, dizziness and examination for AF.

Critical arrhythmias

The AI model had superior sensitivity for the primary endpoint of false-negative findings (all instances of the arrhythmia missed for the full recording) of critical arrhythmia (98.6% (95% confidence interval (CI) = 97.7–99.4) versus 80.3% (95% CI = 77.3–83.3%); Table 1). This category includes ≥30 s of AF, ≥30 s of supraventricular tachycardia (SVT), sinus arrest/asystole events lasting ≥3.5 s, third-degree AV block of any duration and ≥10 s of ventricular tachycardia (VT) ≥ 120 beats per minute. The AI model analysis had 3.2 false negatives per 1,000 patients, compared to 44.3 per 1,000 for technicians (Fig. 1), resulting in a relative risk (RR) of a false-negative finding of critical arrhythmias of 14.1 (95% CI = 10.4–19.0) for technician analysis compared to DeepRhythmAI model analysis. Extended Data Table 2 reports these results for individual arrhythmias. The lower false-negative rate with the AI model was observed in both males and females (Extended Data Fig. 2). In a sensitivity analysis where misclassifications between critical arrhythmias were not considered AI or technician false negatives, we saw largely unchanged results—2.3 false-negative findings per 1,000 patients using the AI model and 39.4 per 1,000 patients for technicians (RR = 16.9 (95% CI = 12.0–23.9); Extended Data Fig. 3). This RR for false-negative findings over the full recording increased with increasing monitoring duration (RR = 7.8 (95% CI = 3.1–19.8) for 1–2 days of monitoring, RR = 9.1 (95% CI = 3.9–21.1) for 3–7 days of monitoring and RR = 17.9 (95% CI = 11.9–26.9) for ≥8 days of monitoring). Overall, the negative predictive value for critical arrhythmias was 99.9% (95% CI = 99.9–100%) for the AI model compared to 99.1% (95% CI = 98.9–99.2) for technicians, and the AI model had superior negative predictive values for all individual critical arrhythmia classes (Table 1). The AI model detection rates of true-positive VTs, SVTs, asystoles and third-degree AV blocks were substantially higher than the technicians, and the AI model detected numerically more AF events (Fig. 2). Episode durations for false-negative events are reported in the Extended Data Table 3.

DeepRhythmAI model analysis resulted in more false-positive findings of asystoles, third-degree AV block and ≥10 s VT (Fig. 3). In sensitivity analyses when misclassifications between critical arrhythmias were not considered false positives, the total false-positive event rate over the full recordings was 6.3% for the AI model and 2.3% for technicians (Extended Data Fig. 4), corresponding to 12 (interquartile range (IQR) = 6–74) false-positive events per 1,000 patient days of recording for AI and 5 (IQR = 2–153) per 1,000 patient days of recording for technicians. Panel classifications of patients for whom strips were extracted are reported in Fig. 4. The duration of false-positive detections by the AI model and technicians is reported in Extended Data Table 3.

Full confusion matrix statistics for individual critical arrhythmias for both the AI model and technicians compared to panel annotations are reported in Table 1. DeepRhythmAI model analysis was superior in terms of sensitivity but had lower specificity for ≥10 s VT, asystole and third-degree AV block. The AI model analysis had similar positive predictive value to technicians for AF and sustained SVTs but lower

https://doi.org/10.1038/s41591-025-03516-x

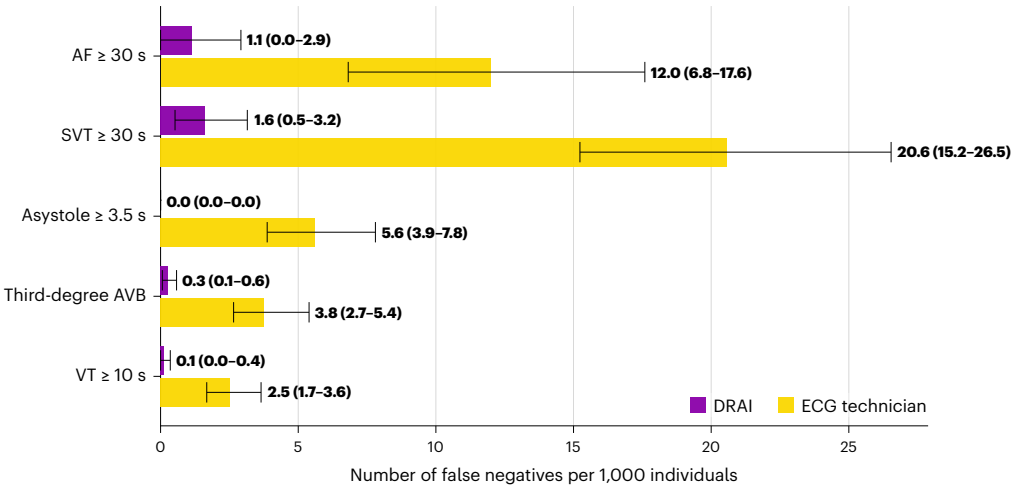


Fig. 1 | False-negative critical arrhythmias per 1,000 patients by AI and technician analysis. Error bars represent 95% CIs derived using bootstrapping. AVB, AV block.

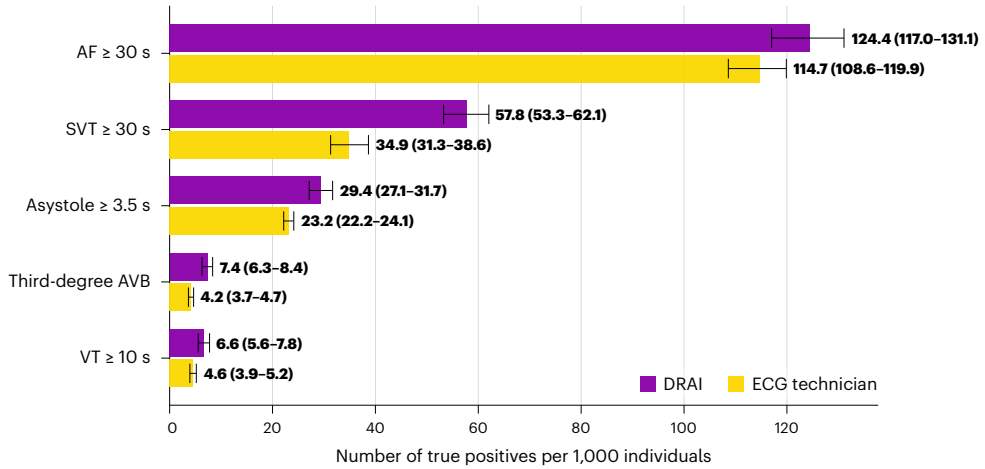


Fig. 2 | True-positive critical arrhythmias per 1,000 patients by AI and technician analysis. Error bars represent 95% CIs derived using bootstrapping.

positive predictive values for sustained VT, third-degree AV block and asystoles. The overall F1 score, which is the harmonized mean of positive predictive value and sensitivity, was similar for the AI model and technicians. However, the F1 scores for AI were superior for sustained SVT, and the F1 score for technicians was better for VT.

Noncritical arrhythmias

Noncritical arrhythmias included premature atrial complexes and premature ventricular complexes, second-degree AV block, pauses of 2.0–3.5 s, VT episodes <10 s, idioventricular/accelerated idioventricular rhythms, SVT episodes ≤30 s and ectopic atrial rhythm. Results for these rhythm classes are reported in Table 2. The AI model had superior sensitivity for all noncritical arrhythmias and a superior F1 score for pauses and idioventricular/accelerated idioventricular rhythms but lower specificity for all noncritical arrhythmias except SVT episodes <30 s and ectopic atrial rhythms.

Discussion

This large, carefully adjudicated analysis demonstrates that the DeepRhythmAI model could safely replace technician interpretation of ambulatory ECG recordings, with an impressive sensitivity for critical arrhythmias and a modest increase of false-positive detections. The DeepRhythmAI model had a negative predictive value for critical arrhythmias that exceeded 99.9% and, compared to technicians, resulted in 17 times fewer patients with a missed diagnosis of a critical arrhythmia. This was at a cost of 2.4 times more false-positive detections, which for critical arrhythmias occurred once every 6 recordings for AI and once every 14 recordings for technicians. Considering that the DeepRhythmAI model performance exceeds the benchmarks of 99% negative predictive value and 70% positive predictive value that guidelines have recommended for accepting a single high-sensitive troponin to rule out major adverse cardiovascular events[32–36], we consider DeepRhythmAI model-only analysis to be safe for the analysis of ambulatory ECG data.

The current study differs fundamentally from previous studies of AI for arrhythmia classification in that we evaluate the use of AI as the only reader for the majority of the health data, with physician confirmation only of AI model-selected episodes. This may be necessary for the management of the rising volume of ECG that will need to be accurately adjudicated without missing critical events. The sample size in terms of annotated strips in this study is 6–16 times larger than previous studies[30,31], and the patient population negative predictive value, absolute false-positive and false-negative rates for AI-only analysis have never been reported before. These data are necessary to determine whether an AI can safely be used for direct-to-physician reporting and have not been shown in previous studies evaluating AI for arrhythmia diagnostics. Direct-to-physician reporting of ambulatory ECG results could unburden strained healthcare environments and result in an appropriate expansion of access, which should result in more equitable access to testing and subsequent care. We used a

https://doi.org/10.1038/s41591-025-03516-x

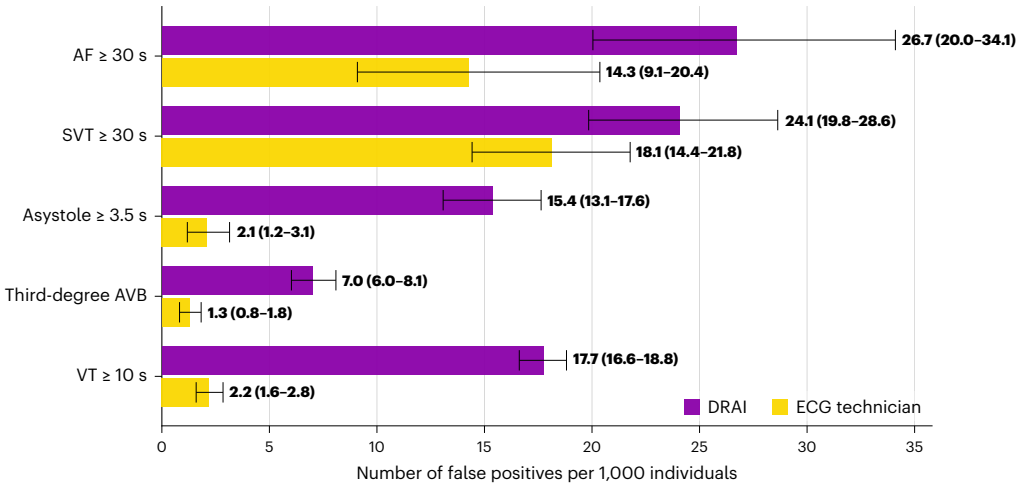


Fig. 3 | False-positive critical arrhythmias per 1,000 patients by AI and technician analysis. Error bars represent 95% CIs derived using bootstrapping.

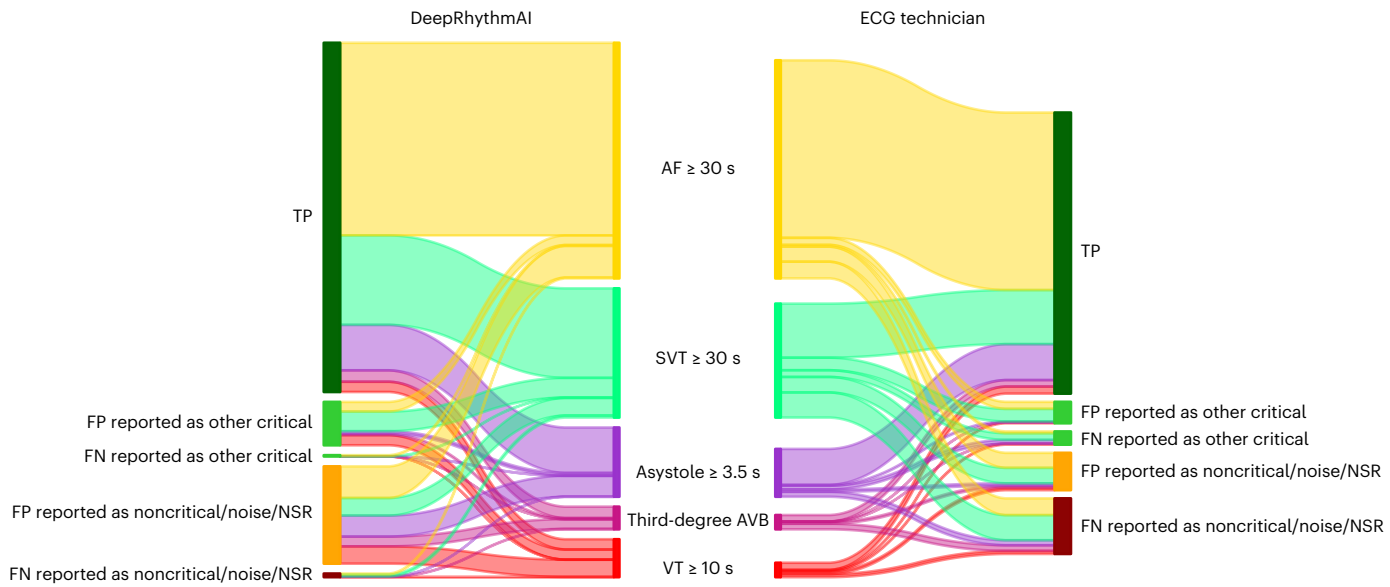


Fig. 4 | Diagnoses of patients with critical arrhythmias by DeepRhythmAI and ECG technicians. Sankey diagram showing arrhythmic event durations for critical arrhythmias as detected by each of the two methods. Cardiologist panel annotations are used to classify DeepRhythmAI and ECG technician annotations into TP, FP or FN. For FP and FN detections, we also report whether these were annotated by the cardiologist panels as another critical arrhythmia class or as a noncritical arrhythmia/noise or NSR. TP, true positives; FP, false positives; FN, false negatives; NSR, normal sinus rhythm.

large, unselected clinical patient population to estimate how the use of the DeepRhythmAI model analysis instead of ECG technician analysis would affect the accurate detection and false-positive rates, using the beat-to-beat classification of a large and representative sample of arrhythmic events. Due to our sampling strategy, the measures of sensitivity that we report are not directly comparable to the sensitivity reported in selected rhythm strips in previous studies. We report as false negatives only patients in whom a diagnosis was missed for the full duration of the recording (that is, 14 ± 10 days of monitoring), arguably a more relevant evaluation metric. With this in mind, the AI model we evaluated had better sensitivity for all critical arrhythmias that were evaluated in both this study and a study assessing a deep neural network architecture for rhythm classification of single-lead ECGs[31], a study evaluating a convolutional neural network for rhythm classification of 12-lead ECGs[30] and a study comparing a deep neural network with physician over-reading of the full ECG to an electrophysiologist review of a traditional Holter system[37]. While the technician sensitivity in this study is low, this finding is in line with previous studies that show a low average accuracy in ECG interpretation for technicians[38].

The large difference in false-negative findings using the DeepRhythmAI model and technician analysis could be dependent on factors related to algorithms and factors related to causes of human error. The higher rate of technician false negatives is likely in part to be due to limitations of features-based algorithms compared to AI models, but because technician work also includes scanning the ECG manually and assessing heart rate trends, there could also be effects of time pressure, information overload[15,17] and other factors related to limits in human perception and memory[16,18], which do not affect AI models. Thus, with increasing data volume that will require analysis, the AI model increasingly outperforms technician interpretation, giving consistent annotations not subject to fatigue. Rhythm analysis by technicians depends on correctly identifying and retaining in memory a large number of visual features; for example, a single capture beat in a wide complex tachycardia is pathognomonic for VT, but the human working memory has a fixed upper limit, and high information loads, such as in the

https://doi.org/10.1038/s41591-025-03516-x

Table 2 | Performance of DeepRhythmAI and ECG technicians compared to the consensus panel of cardiologists for noncritical arrhythmias

	Accuracy, % (95% CI)		Sensitivity, % (95% CI)		Specificity, % (95% CI)		PPV, % (95% CI)		NPV, % (95% CI)		F1 score, %	
	AI	Technician	AI	Technician	AI	Technician	AI	Technician	AI	Technician	AI	Technician
Second-degree AV block	92.3 (91.7–92.9)	**96.7 (95.7–97.4)**	**100 (100–100)**	38.6 (30.9–46.1)	91.9 (91.3–92.5)	**99.5 (99.3–99.7)**	41.9 (37.3–46.3)	**77.8 (68.9–87.0)**	**100 (100–100)**	97.1 (96.1–97.9)	59.1 (54.3–63.3)	51.6 (43.7–59.2)
2.0–3.5 s pauses	**95.0 (93.7–96.3)**	90.3 (87.4–92.5)	**97.8 (95.1–100.0)**	48.8 (40.1–57.7)	94.5 (93.0–96.0)	**97.8 (96.7–98.8)**	78.0 (71.9–84.3)	80.5 (70.4–89.3)	**99.5 (99.0–100.0)**	91.3 (88.1–93.6)	**86.8 (82.8–90.5)**	60.8 (52.0–68.0)
VT 3 beats, 10 s	81.9 (78.6–85.2)	84.7 (80.0–88.7)	**100.0 (100.0–100.0)**	58.3 (48.8–67.7)	75.8 (72.6–79.3)	**95.7 (93.6–97.8)**	58.5 (50.9–66.1)	**84.8 (76.8–92.4)**	**100.0 (100.0–100.0)**	84.7 (79.2–89.0)	73.8 (67.5–79.6)	69.1 (61.3–76.7)
AIVR	85.5 (82.6–88.4)	81.3 (77.2–84.8)	**100.0 (100.0–100.0)**	52.5 (42.4–62.0)	81.1 (78.1–84.3)	**90.1 (87.5–92.4)**	61.6 (53.8–69.2)	61.6 (50.6–71.1)	**100.0 (100.0–100.0)**	86.2 (81.5–90.2)	**76.2 (70.0–81.8)**	56.7 (47.3–64.6)
IVR	93.0 (92.0–94.1)	92.8 (90.4–94.6)	**100.0 (100.0–100.0)**	29.7 (22.1–38.2)	92.4 (91.4–93.5)	**98.9 (98.4–99.4)**	54.0 (47.5–61.1)	**72.9 (60.0–85.7)**	**100.0 (100.0–100.0)**	93.6 (91.0–95.4)	**70.1 (64.4–75.9)**	42.2 (32.9–51.1)
SVT 3 beats, 30 s	79.1 (73.1–84.6)	76.9 (71.0–82.3)	**100.0 (100.0–100.0)**	90.3 (84.2–96.0)	55.8 (49.6–62.9)	62.9 (56.4–70.4)	71.6 (63.4–79.1)	71.8 (63.5–79.4)	**100.0 (100.0–100.0)**	86.2 (77.7–94.1)	83.5 (77.6–88.4)	80.0 (73.9–85.1)
EAR ≥ 3 beats	**83.0 (78.7–87.6)**	64.8 (59.4–70.2)	**99.1 (96.8–100.0)**	56.6 (45.9–67.4)	70.1 (65.0–75.9)	68.9 (64.7–73.2)	**72.7 (65.7–79.7)**	48.0 (38.0–58.0)	**99.0 (96.5–100.0)**	75.8 (68.7–82.2)	**83.8 (78.9–88.7)**	51.9 (43.0–60.5)

The bold values denote nonoverlapping CIs between methods. AIVR, accelerated idioventricular rhythm; IVR, idioventricular rhythm; EAR, ectopic atrial rhythm.

analysis of ambulatory ECG recordings, can lead to reduced accuracy decision quality[15,16].

Some limitations in study design should be considered. First of all, the technicians, but not the physician panels or the AI model, had access to clinical information such as monitoring indication, age and sex, which may have introduced a bias in favor of the technicians. At the same time, while the technicians were performing their analysis during paid clinical work hours, the cardiologist panels were performing their analysis as part of a research protocol, and therefore the panel annotations do not exactly represent a clinical workflow. Panel cardiologists may have been either more or less careful than they would have been with clinical patients, which could have introduced misclassification bias. We have not differentiated between second-degree AV block types 1 and 2, and we do not report subgroups by monitoring indication. Because monitoring indications were entered through the device, the absence of a reported indication should not be interpreted as a lack of that indication. The false-negative events in the study were patients in whom all episodes of arrhythmia were missed for the entire recording duration by one method, but at least one was detected by the other. While we consider this to be a robust method for false-negative estimation, it is possible that there are additional arrhythmic events that were undetected by both the AI model and technicians. If any arrhythmias were missed by both methods, this would imply a lower sensitivity and negative predictive value for both technicians and the AI model but not affect the results showing a superior sensitivity and negative predictive value for the AI model compared to technicians. It is also important to point out that, while the technicians were aided by a Food and Drug Administration-approved algorithm and also performed a manual review and reannotation of the data, their use of a different algorithm may have yielded different results. The underlying ECG data were recorded by a device providing leads II and III. However, the use of devices with nonstandard lead configurations and single-lead recording is becoming more prevalent. The results cannot be generalized to other AI algorithms, and the DeepRhythmAI model may have different performances on other signals, although, in view of the accuracy that the DeepRhythmAI model demonstrated in this study, the model could be tested on other ECG recording signals in the future. Finally, while we used an unselected patient population and extracted a large representative sample of relevant arrhythmic episodes for evaluation, some evaluation metrics that we report, such as the negative predictive value, are dependent on the population prevalence of arrhythmia, which may differ between different populations and may change over time.

Direct-to-physician reporting of leads II and III ambulatory ECG recordings using the DeepRhythmAI model would result in 17 times fewer missed diagnoses of critical arrhythmias than usual care with technician annotation and has a negative predictive value exceeding 99.9%. This would be at a cost of seven extra false-positive findings per 1,000 patient days of recording. AI analysis may substantially reduce labor costs and could potentially report results in near real time.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-025-03516-x.

References

1. Healey, J. S. et al. Apixaban for stroke prevention in subclinical atrial fibrillation. *N. Engl. J. Med.* **390**, 107–117 (2024).
2. Svennberg, E. et al. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. *Lancet* **398**, 1498–1506 (2021).
3. Diederichsen, S. Z. et al. Comprehensive evaluation of rhythm monitoring strategies in screening for atrial fibrillation: insights from patients at risk monitored long term with an implantable loop recorder. *Circulation* **141**, 1510–1522 (2020).
4. Steinhubl, S. R. et al. Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed atrial fibrillation: the mSToPS randomized clinical trial. *JAMA* **320**, 146–155 (2018).
5. Dziubinski, M. et al. Diagnostic yield is dependent on monitoring duration. Insights from a full-disclosure mobile cardiac telemetry system. *Kardiol. Pol.* **80**, 49–55 (2022).
6. Reiffel, J. A. et al. Rhythm monitoring strategies in patients at high risk for atrial fibrillation and stroke: a comparative analysis from the REVEAL AF study. *Am. Heart J.* **219**, 128–136 (2020).
7. Carrington, M. et al. Clinical applications of heart rhythm monitoring tools in symptomatic patients and for screening in high-risk groups. *Europace* **24**, 1721–1729 (2022).
8. Barrett, P. M. et al. Comparison of 24-hour Holter monitoring with 14-day novel adhesive patch electrocardiographic monitoring. *Am. J. Med.* **127**, 95 (2014).

9. Schnabel, R. B. et al. Searching for atrial fibrillation poststroke: a white paper of the AF-SCREEN international collaboration. *Circulation* **140**, 1834–1850 (2019).
10. Healey, J. S. et al. Subclinical atrial fibrillation and the risk of stroke. *N. Engl. J. Med.* **366**, 120–129 (2012).
11. Linz, D. et al. Longer and better lives for patients with atrial fibrillation: the 9th AFNET/EHRA consensus conference. *Europace* **26**, euae070 (2024).
12. Brandes, A. et al. Consumer-led screening for atrial fibrillation: frontier review of the AF-SCREEN international collaboration. *Circulation* **146**, 1461–1474 (2022).
13. Svennberg, E. et al. How to use digital devices to detect and manage arrhythmias: an EHRA practical guide. *Europace* **24**, 979–1005 (2022).
14. Mathieu, B. et al. The global health workforce stock and distribution in 2020 and 2030: a threat to equity and 'universal' health coverage? *BMJ Glob. Health* **7**, e009316 (2022).
15. Hahn, M., Lawson, R. & Lee, Y. G. The effects of time pressure and information load on decision quality. *Psychol. Market.* **9**, 365–378 (1992).
16. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390**, 279–281 (1997).
17. Phillips-Wren, G. & Adya, M. Decision making under stress: the role of information overload, time pressure, complexity, and uncertainty. *J. Decis. Syst.* **29**, 213–225 (2020).
18. Miller, G. A. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**, 81–97 (1956).
19. Obermeyer, Z. & Emanuel, E. J. Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
20. Lång, K. et al. Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol.* **24**, 936–944 (2023).
21. Gilbert, F. J. et al. Single reading with computer-aided detection for screening mammography. *N. Engl. J. Med.* **359**, 1675–1684 (2008).
22. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
23. Yu, F. et al. Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
24. Eloy, C. et al. Artificial intelligence-assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. *Virchows Arch.* **482**, 595–604 (2023).
25. Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
26. Lin, C. S. et al. AI-enabled electrocardiography alert intervention and all-cause mortality: a pragmatic randomized clinical trial. *Nat. Med.* **30**, 1461–1470 (2024).
27. Johnson, L. S. et al. Can 24 h of ambulatory ECG be used to triage patients to extended monitoring? *Ann. Noninvasive Electrocardiol.* **28**, e13090 (2023).
28. Attia, Z. I. 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* **394**, 861–867 (2019).
29. Siontis, K. C., Noseworthy, P. A., Attia, Z. I. & Friedman, P. A. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nat. Rev. Cardiol.* **18**, 465–478 (2021).
30. Zhu, H. et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit. Health* **2**, e348–e357 (2020).
31. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
32. Lowry, M. T. H. et al. Troponin in early presenters to rule out myocardial infarction. *Eur. Heart J.* **44**, 2846–2858 (2023).
33. Chapman, A. R. et al. Association of high-sensitivity cardiac troponin I concentration with cardiac outcomes in patients with suspected acute coronary syndrome. *JAMA* **318**, 1913–1924 (2017).
34. Than, M. et al. What is an acceptable risk of major adverse cardiac event in chest pain patients soon after discharge from the Emergency Department?: a clinical survey. *Int. J. Cardiol.* **166**, 752–754 (2013).
35. Sandoval, Y. et al. Myocardial infarction risk stratification with a single measurement of high-sensitivity troponin I. *J. Am. Coll. Cardiol.* **74**, 271–282 (2019).
36. Collet, J.-P. et al. 2020 ESC guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: the task force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). *Eur. Heart J.* **42**, 1289–1367 (2020).
37. Fiorina, L. et al. Evaluation of an ambulatory ECG analysis platform using deep neural networks in routine clinical practice. *J. Am. Heart Assoc.* **11**, e026196 (2022).
38. Tahri Sqalli, M. et al. Understanding cardiology practitioners' interpretations of electrocardiograms: an eye-tracking study. *JMIR Hum. Factors* **9**, e34058 (2022).

L. S. Johnson[1,2] ✉, P. Zadrozniak[3], G. Jasina[3], A. Grotek-Cuprjak[3], J. G. Andrade[4], E. Svennberg[5,6], S. Z. Diederichsen[7], W. F. McIntyre[2,8], S. Stavrakis[9], J. Benezet-Mazuecos[10], P. Krisai[11], Z. Iakobishvili[12,13], A. Laish-Farkash[12], S. Bhavnani[14], E. Ljungström[15], J. Bacevicius[16], N. L. van Vreeswijk[17], M. Rienstra[17], R. Spittler[18], J. A. Marx[18], A. Oraii[2], A. Miracle Blanco[10], A. Lozano[10], I. Mustafina[9,19], S. Zafeiropoulos[20,21], R. Bennett[4], J. Bisson[22], D. Linz[23,24], Y. Kogan[12], E. Glazer[12], G. Marincheva[12], M. Rahkovich[12], E. Shaked[12], M. H. Ruwald[25], K. Haugan[26], J. Węcławski[3], G. Radoslovich[27],

S. Jamal[27,28], A. Brandes[29,30], P. T. Matusik[31,32], M. Manninger[33], P. B. Meyre[11], S. Blum[11], A. Persson[1,34], A. Måneheim[1,34], P. Hammarlund[35], A. Fedorowski[1,5,36], T. Wodaje[5,36], C. Lewinter[5,36,37], V. Juknevicius[16], R. Jakaite[16], C. Shen[14], T. Glotzer[27,28], P. Platonov[15,38], G. Engström[1], A. P. Benz[2,18] & J. S. Healey[2,8]

[1]Department of Clinical Sciences, Malmö, Lund University, Lund, Sweden. [2]Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada. [3]Medicalgorithmics S.A., Warsaw, Poland. [4]Vancouver General Hospital, University of British Columbia, Vancouver, British Columbia, Canada. [5]Karolinska Institutet, Stockholm, Sweden. [6]Department of Medicine Huddinge, Karolinska University Hospital, Stockholm, Sweden. [7]Department of Cardiology, Copenhagen University Hospital—Rigshospitalet, Copenhagen, Denmark. [8]Department of Medicine, McMaster University, Hamilton, Ontario, Canada. [9]University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA. [10]Cardiology Department Hospital Universitario La Luz, Madrid, Spain. [11]Department of Cardiology and Cardiovascular Research Institute Basel, University Hospital Basel, University of Basel, Basel, Switzerland. [12]Department of Cardiology, Assuta Ashdod University Hospital, Ben-Gurion University of the Negev, Ashdod, Israel. [13]Department of Cardiology, Clalit Health Services, Tel Aviv Jaffa District, Israel. [14]Division of Cardiology, Scripps Clinic, San Diego, CA, USA. [15]Arrhythmia Clinic, Skåne University Hospital, Lund, Sweden. [16]Clinic of Heart and Vessel Diseases, Institute of Clinical Medicine, Faculty of Medicine, Vilnius University, Vilnius, Lithuania. [17]Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [18]Department of Cardiology, University Medical Center Mainz, Johannes Gutenberg-University Mainz, Mainz, Germany. [19]Department of Internal Diseases, Bashkir State Medical University, Ufa, Russia. [20]Feinstein Institutes for Medical Research at Northwell Health, Manhasset, NY, USA. [21]Department of Cardiology, University Hospital of Zurich, Zürich, Switzerland. [22]Department of Cardiology, Centre hospitalier de l'Université de Montréal—Université de Montréal, Montréal, Quebec, Canada. [23]Department of Cardiology, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University Medical Centre, Maastricht, The Netherlands. [24]Faculty of Health and Medical Sciences, Department of Biomedical Sciences, University of Copenhagen, Copenhagen, Denmark. [25]Department of Cardiology, Gentofte Hospital, Hellerup, Denmark. [26]Department of Cardiology, Zealand University Hospital, Roskilde, Denmark. [27]Hackensack University Medical Center, Hackensack, NJ, USA. [28]Hackensack Meridian School of Medicine, Nutley, NJ, USA. [29]Department of Cardiology, Esbjerg Hospital—University Hospital of Southern Denmark, Esbjerg, Denmark. [30]Department of Regional Health Research, University of Southern Denmark, Esbjerg, Denmark. [31]Department of Electrocardiology, Institute of Cardiology, Faculty of Medicine, Jagiellonian University Medical College, Kraków, Poland. [32]St. John Paul II Hospital, Kraków, Poland. [33]Division of Cardiology, Department of Medicine, Medical University of Graz, Graz, Austria. [34]Department of Clinical Physiology, Skåne University Hospital, Malmö, Sweden. [35]Department of Cardiology, Helsingborg Hospital, Helsingborg, Sweden. [36]Department of Cardiology, Karolinska University Hospital, Stockholm, Sweden. [37]University of Glasgow, University of Glasgow, Institute of Wellbeing, Glasgow, UK. [38]Department of Clinical Sciences, Lund University, Lund, Sweden. ✉e-mail: linda.johnson@med.lu.se

Methods

Data source

The source population for this study is an unselected patient population of 14,606 individuals, consisting of a random sample of patients who had been monitored in the United States for clinical indications between 2016 and 2019. Recording durations varied from 1 to 31 days. The dataset consisted of 211,010 days of ambulatory monitoring collected in these patients using PocketECG (Medicalgorithmics). PocketECG is a full-disclosure ECG device with limb lead configuration (leads II and III) and a sampling rate of 300 samples per second. The device can record and transmit ECG signals for up to 31 days. The patients were referred by 1,079 different physicians from 166 clinics, and the recordings were analyzed in clinical practice at an independent diagnostic testing facility by one of 167 certified ECG technicians working with a features-based algorithm using adaptive beat morphology template generation and comparison so that each QRS complex in the recording was annotated beat-to-beat by the ECG technician. ECG technician work was extensive and included a review of the whole ECG recording and verification of all events detected by the algorithm, including pauses and asystoles, all bradycardia events, all missed heartbeats or second- and third-degree AV blocks, all ventricular and supraventricular arrhythmias and all episodes detected as AF. In this process, artifacts and electrode dysfunction were re-annotated. The technicians also inspected all regions of the recording marked as having a 'patient-triggered symptom' flag and reviewed the recording at the time of the fastest, slowest and average minutely heart rate. They were aided in this process by software that allowed them to manually inspect heart rate trends for irregularities, filter beats by heart rate and group beats into morphologies. At the end of the review, episodes were selected for inclusion in a report to physicians.

Before inclusion in the study, all data were anonymized, and the Ethics Review Board of Sweden has therefore waived the need for approval (decision 2019-03227). As such, the Ethics Review Board did not consider that informed consent was necessary.

DeepRhythmAI

The DeepRhythmAI model (v3.1; Medicalgorithmics) is a proprietary mixed network ensemble for rhythm classification. The network performs QRS and noise detection, beat classification and rhythm identification using several algorithms based on convolutional neural networks and transformer architecture with custom-built components[39–42]. The main network components for QRS detection and rhythm classification have been pretrained on 1,716,141 5-min-long ECG strips and fine-tuned on 60,549 ≤30 s ECG strips. These were extracted from 69,706 anonymized clinical long-term recordings. Algorithm internal validation was performed using 15,188 ≤30 s strips from 12,330 additional separate patient recordings. A high-level flowchart of the algorithm is presented in the Extended Data Fig. 5. The preprocessing involves selecting desired ECG channels from input data, scaling the signal amplitude according to the input analog–digital conversion values and resampling to a frequency of 300 Hz. A deep learning model predicts the probability of QRS complex presence and signal readability, extracts signal features and predicts the probability of QRS complex presence and readable signal[39]. This output, together with the preprocessed signal, is passed to an ensemble combined from models of two structures. The first is intended for the analysis of information from a wide context and has a hybrid architecture of the convolutional neural network and transformer encoder layers[40]. The second is a pure-transformer implementation based on Vision Transformer[41], allowing for a superior interpretation of signal within a relatively narrow window. Additionally, a specialized classifier was developed for the detection of asystole events.

The QRS complex detector uses custom residual modules inspired by MobileNetV2.[42] Each module consists of the following three one-dimensional convolutional layers: a pointwise convolution to expand feature dimension; a convolutional layer with a kernel length of 3 and variable dilation rates; a pointwise convolution to reduce feature dimensions to their original size. The dilation rate doubles in each residual module during the first half of the model and then progressively decreases to a rate of 1 at the output layer. A final linear layer converts the output features into probabilities of QRS complex presence and signal readability for each sample. Thresholding and morphological operations are subsequently applied to extract QRS positions and identify nondiagnostic ranges. The wide-context architecture comprises a series of submodules. Initially, features are extracted from heart rate trends, calculated based on QRS detections, using the same architecture as the QRS detector (excluding the final linear layer). Another submodule extracts features for each sample of the preprocessed ECG signal using residual modules from the QRS detector but with a fixed dilation rate progression. The signal is downsampled using strided convolutional layers. Subsequently, windows of downsampled features are extracted, and two-dimensional strided convolutional layers are applied, resulting in features for each beat. The resulting features are processed using transformer encoder layers, augmented by an additional convolutional layer inserted between the linear layers in the fully connected blocks. Finally, the features are converted to logits for each QRS complex class using two pointwise convolutional layers.

The signal-detail architecture is based on transformer encoder layers that process ECG signals split into patches. A linear layer embeds each patch. The transformer layers process the embedded patches, and logits for each QRS complex class are calculated using a linear layer. Only the patches containing QRS complexes are selected for predictions. The asystole filter module shares the same architecture as the wide-context model but is trained with hyperparameters and a dataset tailored to the asystole detection task.

We used the same dataset for training the QRS complex and noise detector and the main components of the heartbeat classification ensemble (three wide-context models and three signal-detail models). Data augmentation techniques tailored to each of these tasks, like noise artifact generation or synthesis of heartbeats with rare features, were used to enhance training dataset diversity and mitigate overfitting. In addition to that, a classifier specializing in the interpretation of asystole events was developed by feeding to a single model with wide-context analysis architecture a carefully selected 11,670 strips with asystole or sinus arrest and 20,292 strips with noise or electrode dysfunction. The training process of this model encompassed methods from supervised and self-supervised learning domains. The ensemble model output is averaged or replaced by the asystole filter model output (for heartbeats with RR interval greater than the sinus arrest threshold of 2 s) to provide the probabilities of QRS complex classes. Finally, the heartbeat types that are the final output of the DeepRhythmAI model are translated to heart rhythm types. Optimization was performed using the AdamW algorithm. Models were internally evaluated by measuring the root mean squared error metric based on sensitivity, precision and F1 score calculated from predictions and ground truth of internal validation/test strips, following the methodology provided by the International Electrotechnical Commission 60601-2-47 standard[43].

The ECG recordings used in this study had never been presented to the DeepRhythmAI model or any AI model from which the DeepRhythmAI model was derived, but as part of the study protocol, we analyzed the entire raw ECG signal data from these same recordings using the DeepRhythmAI model to provide detection and beat-to-beat classification of all heartbeats.

Definition of critical and noncritical arrhythmias

Selection of representative arrhythmic episodes. Our strip selection method was designed to not introduce any bias toward using

ECG signals with less baseline noise or arrhythmic events presenting with typical ECG diagnoses. We did this by automation; fully random individual recordings were searched by an algorithm for the presence of arrhythmic events of each rhythm class, and 34-s strips containing arrhythmia events according to either the AI model annotations, the ECG technician annotations or both were selected, at a maximum of one per method and arrhythmia class per patient. The automated selection script ran until a total of 500 strips each had been selected for each of the critical arrhythmias and 250 strips each had been selected for the noncritical rhythm classes, or all recordings had been searched and no more arrhythmias were found. The number of individual recordings that had to be searched to yield the strips for each rhythm class was considered the source population size for that class. The strip selection is described in Extended Data Fig. 6. In addition to the critical and noncritical rhythm classes, we included sinus rhythm, sinus bradycardia and unreadable signals due to noise or electrode dysfunction to evaluate the AI model performance for these signals and to ensure that the physician annotators would be provided with a differentiated sample in which they did not know which strips would contain critical arrhythmias. In total, we selected 5,245 strips, of which 2,240 were critical arrhythmias, and after errors in uploading ten of these to the annotation platform, we had 5,235 strips, of which 2,236 were critical arrhythmias.

Consensus panel annotations. All 34-s strips were annotated beat-to-beat by 17 panels consisting of three expert annotators each—≥2 board-certified cardiologists and additionally including board-certified clinical physiologists ($n = 2$) or final-year cardiology residents. The physicians on the panels performed the annotation independently of AI and technician annotations and were blinded to the strip selection criteria. Strips were randomly distributed among panels and presented in random order and were annotated using a custom-built software platform in which QRS complex tags, without beat type classifications, as detected by the AI model, were present. We used DeepRhythmAI model-detected QRS complexes for strips detected by both the AI model and the technicians to minimize bias; technicians in clinical practice may not have bothered to correct QRS tags for all instances of arrhythmia, and differential methodology for strips could have resulted in unblinding. The QRS tags were highly concordant. For QRS complexes that resulted in technician false negatives, there was a 98% overlap between the AI model and fixed features algorithm QRS positions. Physician annotators were asked to identify the beat type for each QRS complex according to an annotation manual (Supplemental Note), correct any mistaken QRS position placements, add any missed QRS complexes and mark areas that were unreadable due to poor signal or electrode dysfunction. Each physician annotated the entire strip beat by beat, and all discrepancies on the beat level were resolved by panel consensus. The resulting gold-standard annotations were compared to the beat-to-beat annotations of the AI model and technicians according to prespecified acceptance criteria, where we considered arrhythmic events to be concordant with the panel annotation in case of ≥80% overlap in beat type and duration with the panel annotation for all sustained tachyarrhythmias and 90% overlap in duration for asystole events and pauses. For second- or third-degree AV block, we considered the presence of any such event within the strip to be a concordant annotation, and for ECG technicians, we also considered annotation of an unspecified 'missed beat' to be a concordant annotation for second-degree AV block. Single ectopic atrial and ventricular beats were considered concordant within ±45 samples (150 ms). Noise annotations were considered concordant if within 80% of the panel annotation as regards duration. Minor discrepancies between the AI/technician annotations and consensus panel annotations, on the beat-to-beat level, were thus allowed, for example, low numbers of supraventricular beats or beats with unknown beat types within AF episodes.

Statistics. The primary analysis compares the frequency of false-negative, true-positive and false-positive critical arrhythmias per 1,000 individual patients over the full duration of the recordings for technicians and the AI model, along with full confusion matrix statistics for the AI model and technician performance compared to panel annotations. As a result of the sampling strategy, false negatives were only reported in patients in whom all instances of an arrhythmia type were missed for the entire duration of the recording. True-positive events were defined as episodes detected by the AI model or technician, with correct annotations according to the independent gold-standard consensus panel annotation. Descriptive statistics are reported as mean ± s.d. CIs were derived using bootstrapping with 1,000 replications. Definitions for the confusion matrix statistics are reported in the Extended Data Table 4. We also performed subanalyses where misclassifications of critical arrhythmias were not considered false-negative or false-positive events because these events would have been reported to physicians. In these analyses also, we did not consider second-degree AV block to be a false-positive finding. For the analyses of total false-positive and false-negative findings of critical arrhythmias, the prevalence of all arrhythmias was weighted to the full population size according to the proportion of the population queried. Nonoverlapping CIs were considered evidence of the superiority of one method over the other. All analyses were performed in Python, except for the calculations of RR, which were done in Stata version 17.0 for Mac, using two-sided Fisher's exact P values. Analyses were performed by L.S.J. and G.J., with involvement from the steering group, according to prespecified plans. The study steering group (L.S.J., J.S.H., A.P.B. and A.M.) met regularly throughout the conduct of the study without the presence of Medicalgorithmics employees.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data that supports the findings of this study are derived from patient ECGs and are not publicly available due to privacy concerns but will be made available after a request for access to the corresponding author for the purpose of reviewing the study results and at the cost of a data preparation fee. No requests that include a commercial interest will be approved. Data are located in controlled access data storage at Medicalgorithmics. A response to a request to access the data can be expected within 2 months.

Code availability

The code for the DeepRhythmAI model is not available due to its proprietary nature. The code used for statistical analyses will be made available upon request to the corresponding author for the purpose of reviewing the results in the paper, and a response can be expected within 2 months.

References

39. Oord, A. V. D. et al. WaveNet: a generative model for raw audio. Preprint at *arXiv* 10.48550/arXiv.1609.03499 (2016).
40. Vaswani, A. et al. Attention is all you need. Preprint at *arXiv* 10.48550/arXiv.1706.03762 (2017).
41. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at *arXiv* 10.48550/arXiv.2010.11929 (2020).
42. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L. C. MobileNetV2: inverted residuals and linear bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 4510–4520 (IEEE, 2018).
43. Young, B. New standards for ECG equipment. *J. Electrocardiol.* **57**, S1–S4 (2019).

Acknowledgements
The authors acknowledge the contributions of M. Dziubinski, whose previous AI development work was foundational for the models developed for this study, and M. Rosenqvist, who gave feedback on the study results. We also acknowledge the contributions of the engineers at Medicalgorithmics, who have built the technological platforms that have been used in the study, M. Kulesza for his supervision of engineering work and the ECG technicians who have curated the datasets for training of the DeepRhythmAI. Medicalgorithmics hosted a study meeting for discussion of results and provided the technology, data curation and statistical analyses for the study, as well as paying consulting fees to L.S.J. The study was designed independently by L.S.J. before receiving consulting fees from Medicalgorithmics. J.W., P.Z. and G.J., who are current (P.Z. and G.J.) or former (J.W.) Medicalgorithmics employees, conceived of and wrote the algorithm for strip selection. The steering committee for the study did not include any Medicalgorithmics employees. Besides providing information on algorithm architecture and training, Medicalgorithmics was not involved in the writing of the paper and did not suggest any changes to the draft. L.S.J. has primarily been funded by the Swedish Heart and Lung Foundation (grant 20210343), the Swedish Research Council (grant 2022-00903) and the Swedish Society for Medical Research.

Author contributions
L.S.J. conceived the study and was the principal investigator. The study was designed by L.S.J., J.W., P.Z. and G.J., with the assistance of a steering committee also consisting of J.S.H., A.B. and A.M. P.Z., J.W. and G.J. have designed and built the DeepRhythmAI model. A.G.C. has curated the data and kept the study database with G.J. and P.Z. L.S.J. and J.S.H. drafted the paper, which was critically revised by all other authors, in particular T.G., E.S., J.A., S.S., W.F.M., D.L., P.K., P.P., M.M. and S.Z.D. The expert consensus panels that annotated the selected ECG strips consisted of L.S.J., J.G.A., E.S., S.Z.D., W.F.M., S.S., J.B.M., P.K., Z.I., A.L.F., S.B., E.L., J.B., N.L.v.V, M.R., R.S., J.A.M., A.O., A.M.B., A.L., I.M., S.Z., R.B., J.B., D.L., Y.K., E.G., G.M., M.R., E.S., M.H.R., K.H., G.R., S.J., A.B., P.T.M., M.M., P.B.M., S.B., A.P., P.H., A.F., T.W., C.L., V.J., R.J., C.S., T.G., P.P. and J.S.H.

Funding
Open access funding provided by Lund University.

Competing interests
L.S.J. receives consulting fees from Medicalgorithmics. P.Z., G.J. and A.G.C. are Medicalgorithmics employees. J.W. is a former Medicalgorithmics employee. J.S.H. has research grants and speaking fees from BMS/Pfizer, Boehringer Ingelheim, Boston Scientific, Novartis, Medtronic and Servier. A.P.B. has received speaker fees from Bristol Myers Squibb and AstraZeneca, and participates in an educational program supported by Boston Scientific (Fellowship Herzrhythmus). E. Svennberg is supported by the Stockholm County Council (clinical researcher appointment), the Swedish Research Council (DNR 2022-01466), the Swedish Heart and Lung Foundation and CIMED and has received institutional remunerations for lectures from Abbott, AstraZeneca, Bristol-Myers Squibb-Pfizer and Johnson & Johnson. P.H. reports lecture fees from Pfizer and Boehringer Ingelheim. M.H.R. reports speaker fees from Cardiofocus and Boston Scientific. M. Rienstra reports consultancy fees from Bayer (OCEANICAF national PI) and InCarda Therapeutics (RESTORE-SR national PI) to the institution. P.K. reports speaker fees from BMS/Pfizer and research grants from the Swiss National Science Foundation, Swiss Heart Foundation, Foundation for Cardiovascular Research Basel and Machaon Foundation. S.Z.D. is a part-time employee of Vital Beats and has received consultancy fees from Cortrium, Acesion Pharma and Bristol-Myers Squibb-Pfizer, lecture fees from Bayer and Bristol-Myers Squibb-Pfizer and travel grants from Abbott and Boston Scientific. M.M. received speaker fees/honoraria from Abbott, Bayer Healthcare, Biosense Webster, Biotronik, Amomed, AOP Orphan, Boston Scientific, Daiichi Sankyo and BMS/Pfizer and research grants from Biosense Webster and Abbott. P.T.M. has received speech honoraria from Boehringer Ingelheim and participated in educational activities, which were supported by cardiovascular implantable electronic device manufacturers. J. Bacevicius holds a patent for the TeltoHeart technology, consults Teltonika Telemedic and has received travel grants from Abbot and Biosense Webster. V.J. has received speaker fees from BMS/Pfizer, Bayer, Boehringer and Servier. A.B. has received research grants from Theravance, the Zealand Region, the Canadian Institutes of Health Research, the European Union Interreg 5A Programme, the Danish Heart Foundation, the Independent Research Fund Denmark and a lecture honorarium from Bristol Myers Squibb outside the submitted work. T.G. serves on an advisory board for Medtronic and Boston Scientific and has received speaking honoraria from Medtronic, Boston Scientific and Abbott.

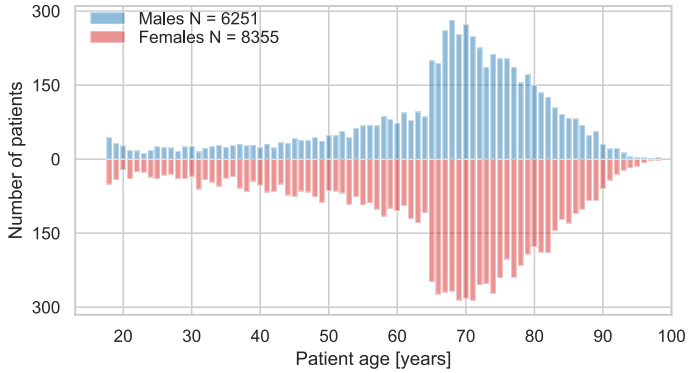
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41591-025-03516-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-025-03516-x.

Correspondence and requests for materials should be addressed to L. S. Johnson.

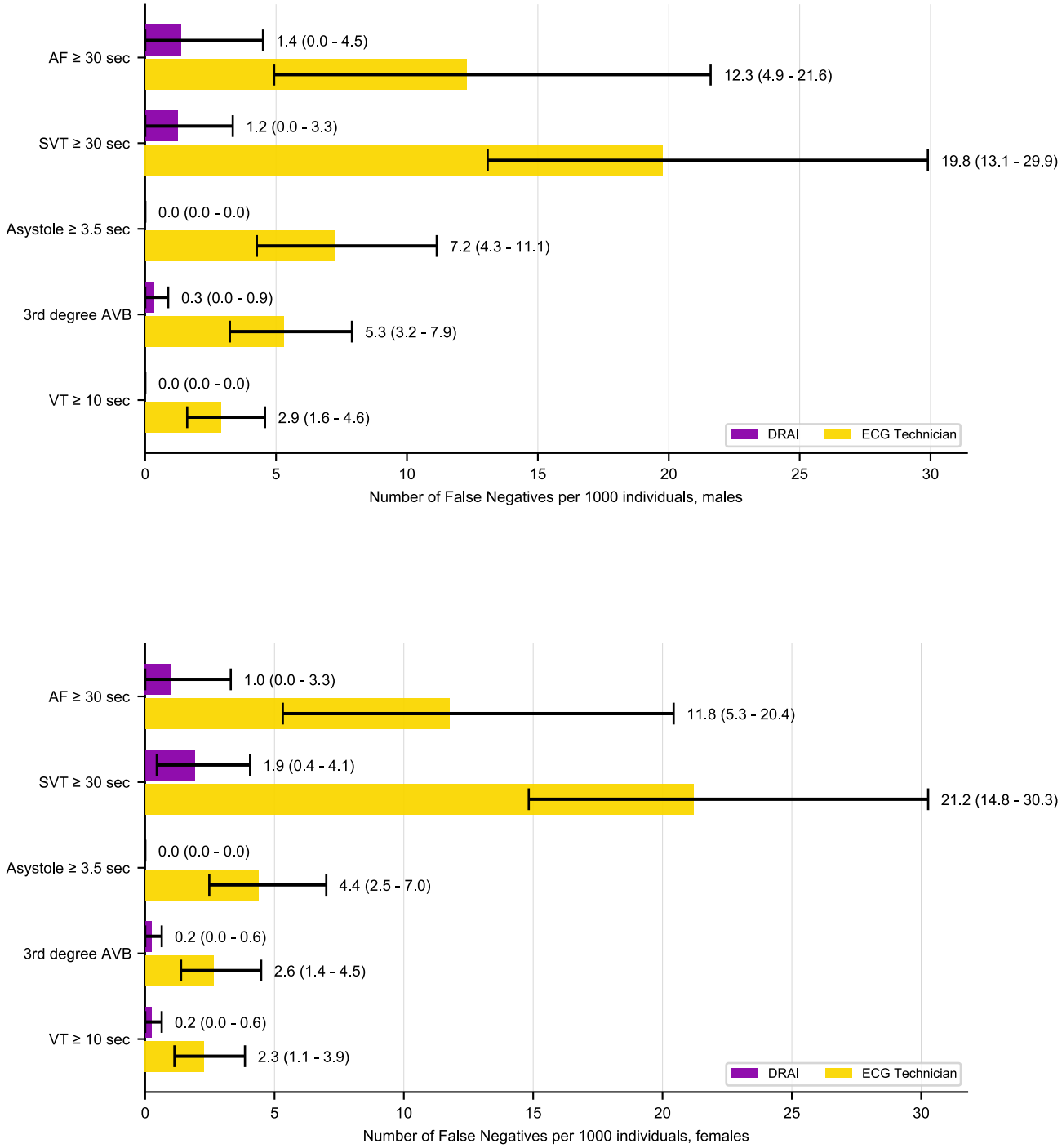
Peer review information *Nature Medicine* thanks Axel Bauer, Albert Rogers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

https://doi.org/10.1038/s41591-025-03516-x



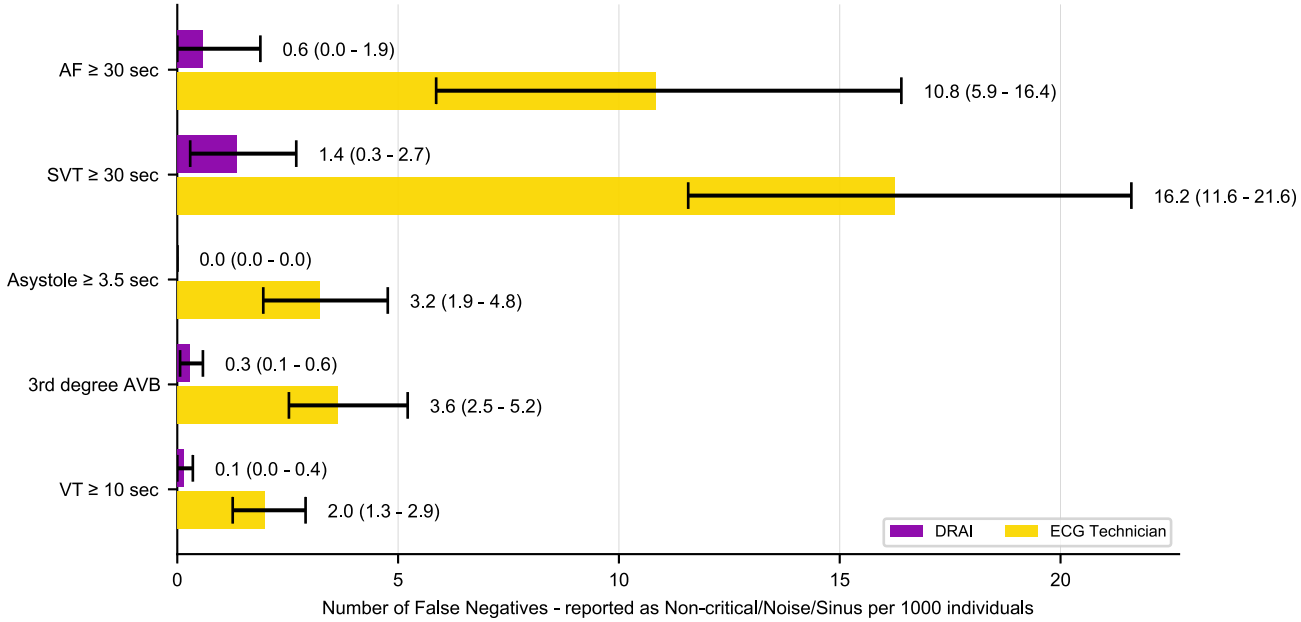
Extended Data Fig. 1 | Age and sex distribution. Age and sex distribution of the patient sample included in the analyses.

https://doi.org/10.1038/s41591-025-03516-x



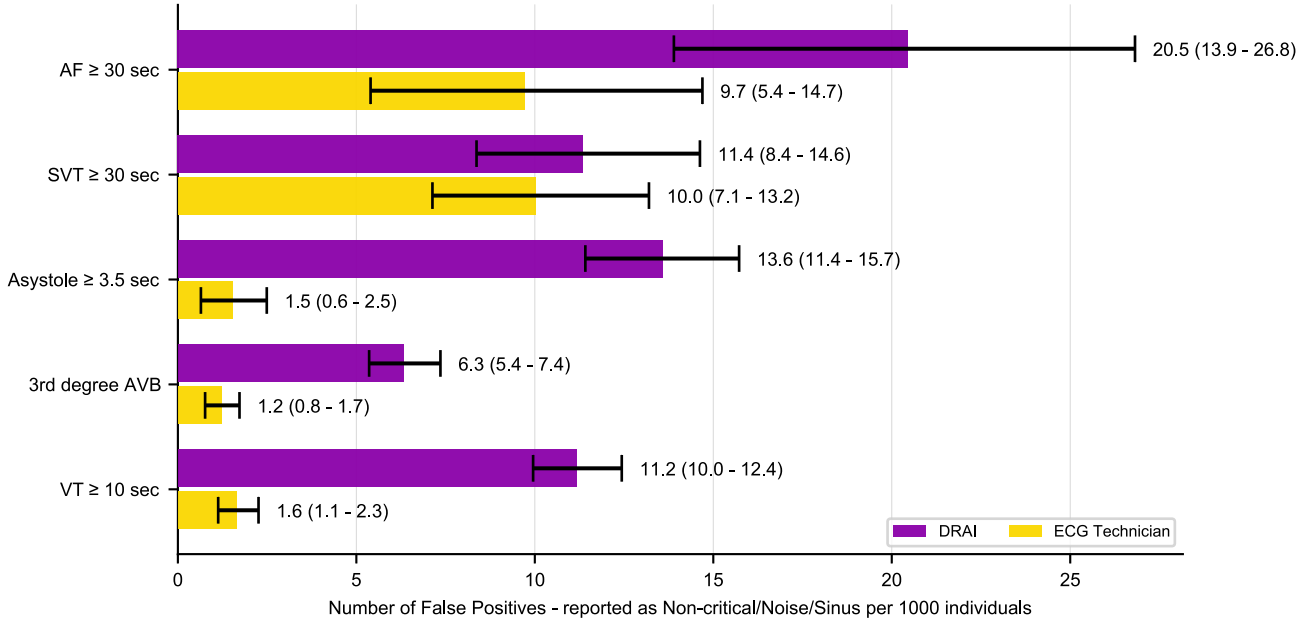
Extended Data Fig. 2 | False-negative findings in males and females. Error bars denote 95% confidence intervals and were derived using bootstrapping.

https://doi.org/10.1038/s41591-025-03516-x



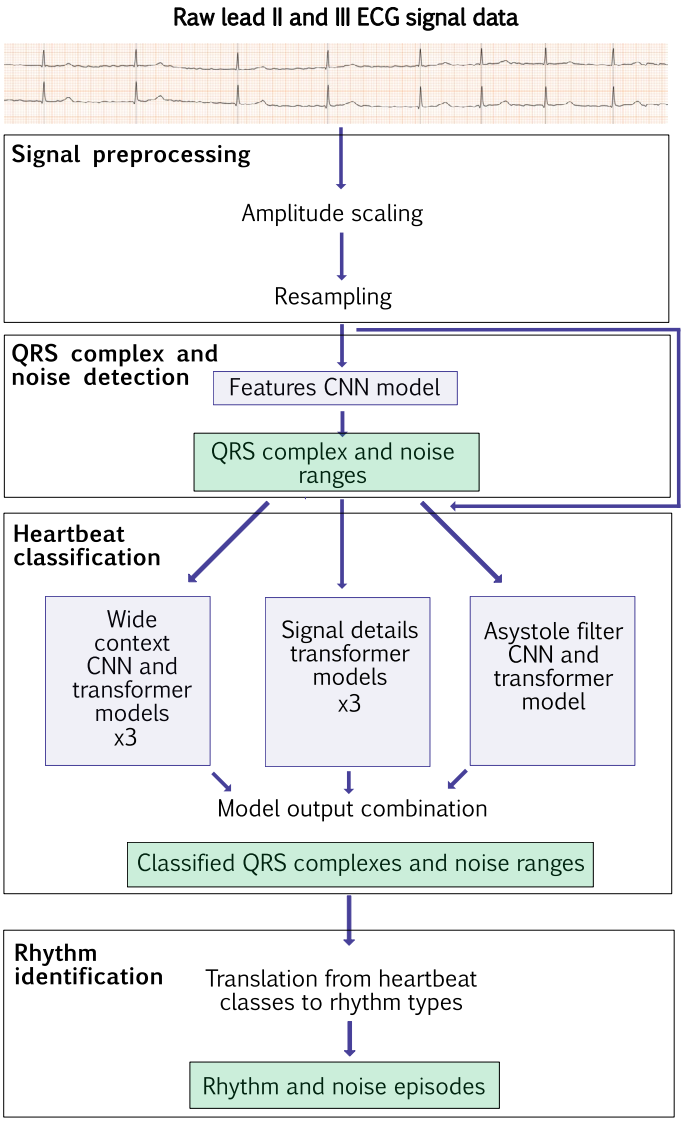
Extended Data Fig. 3 | False-negative findings, excluding events reported as other critical arrhythmias. Error bars denote 95% confidence intervals and were derived using bootstrapping.

https://doi.org/10.1038/s41591-025-03516-x



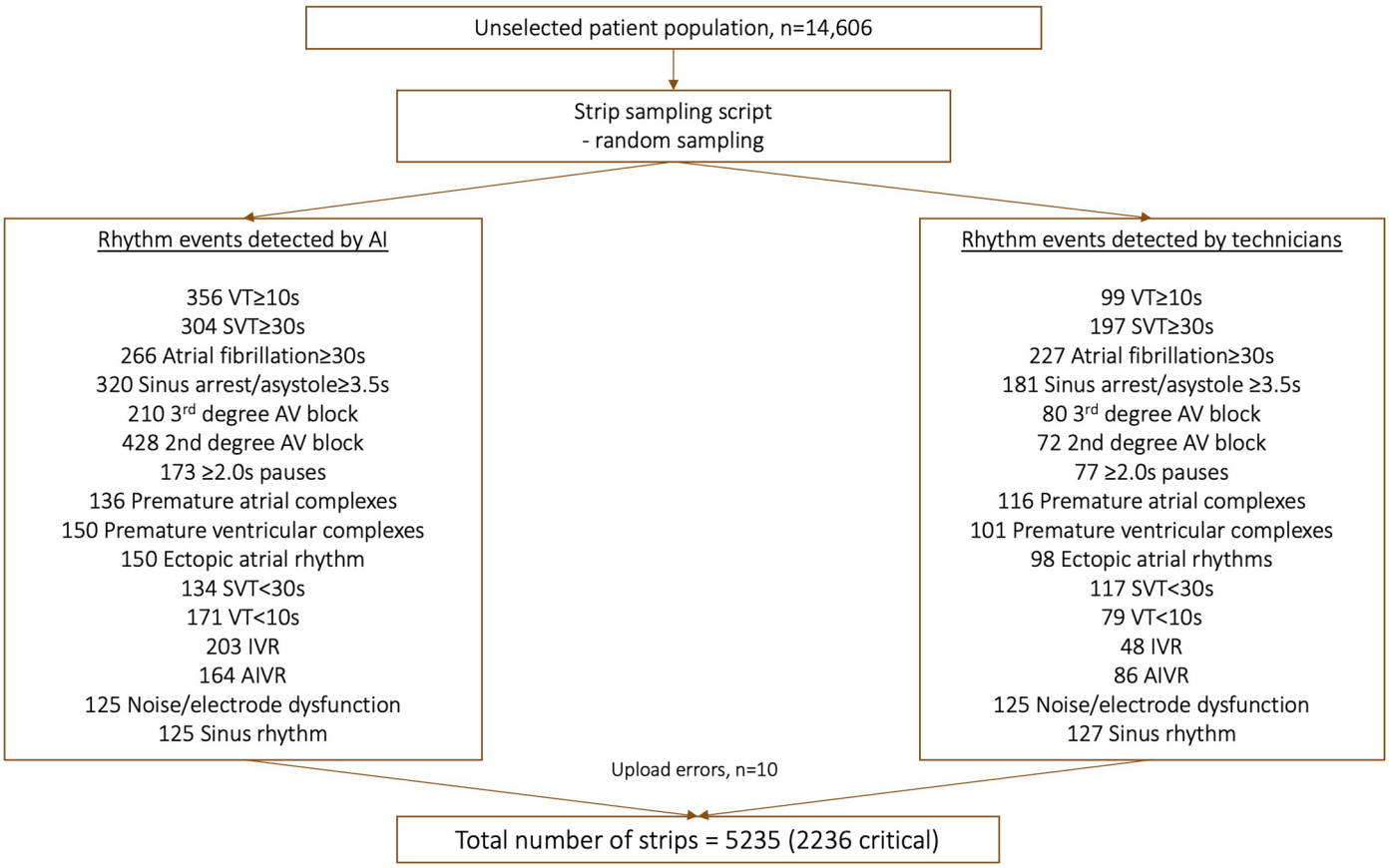
Extended Data Fig. 4 | False-positive findings of critical arrhythmias. Error bars denote 95% confidence intervals and were derived using bootstrapping.

https://doi.org/10.1038/s41591-025-03516-x



Extended Data Fig. 5 | Schematic overview of the DeepRhythmAI model. The raw ECG signal (in timestamp + mV format) is pre-processed and fed to a single CNN classifier model that identifies the QRS complexes and segments of noisy (non-diagnostic) signals in the raw ECG data. The network components downstream to this module are fed both raw signal and the QRS/noise module output. This combined signal and QRS/noise data are processed by ensemble of a total of 7 models with both wide context (HR trend and morphology of beats) and narrow context (signal details). The wide context module is an ensemble of three custom deep neural network models with both CNN and transformer layers. The narrow context module is an ensemble of three transformer models all based on Vision Transformer ideas but with custom adaptations to 1D multichannel ECG signal. The output from these models is then combined with a wide context asystole filter that has the same architecture as wide context models but with hyperparameters tuned for asystole detection. The asystole filter overrides and replaces the other probabilities when asystoles are detected; otherwise, the output probabilities are averaged.

https://doi.org/10.1038/s41591-025-03516-x



Extended Data Fig. 6 | Description of the strip selection process. VT, ventricular tachycardia; AF, atrial fibrillation; SVT, supraventricular tachycardia; AIVR, accelerated idioventricular rhythm; IVR, idioventricular rhythm; EAR, ectopic atrial rhythm.

https://doi.org/10.1038/s41591-025-03516-x

Extended Data Table 1 | Monitoring indications reported on device

Indication	N	Age, mean±SD	% males
Palpitations	7026	61.7±17.3	36.3
Syncope/collapse	2231	68.5±16.7	43.7
Dizziness and giddiness	1500	69.0±14.9	45.7
Atrial fibrillation, paroxysmal	1481	71.7±10.7	49.6
Atrial fibrillation, unspecified	1222	71.3±11.7	51.6
Transient Ischemic Attack	1035	71.7±10.3	43.1
Dyspnea	946	66.8±15.2	39.1
Tachycardia	811	59.9±19.7	39.0
Bradycardia	890	70.2±15.3	55.1
Atrial flutter	651	68.6±12.2	56.5
Premature ventricular complexes	477	64.5±16.7	33.3
Persistent AF	382	62.1±17.2	49.0
Stroke	322	70.9±10.5	43.2
Other	222	61.0±19.6	41.4

Several answers per patient possible.

https://doi.org/10.1038/s41591-025-03516-x

Extended Data Table 2 | Relative risks of false-negative findings for technicians compared to artificial intelligence

	Relative risk	95% confidence intervals	p
Any critical arrhythmia	14.1	10.4-19.0	<0.0001
AF≥30s	10.2	6.2-16.8	<0.0001
SVT ≥30s	12.5	8.2-18.9	<0.0001
Asystole≥3.5s	82.0	11.0-589.0	<0.0001
3rd degree AV block	13.8	5.0-37.9	<0.0001
VT≥10s	18.5	4.5-76.7	<0.0001

AI analysis resulted in no false negatives for asystoles ≥3.5 seconds, results are calculated with the addition of 1 false negative for AI in this category. Two-sided *P* values were derived using Fischer's exact test. VT, ventricular tachycardia; AF, atrial fibrillation; SVT, supraventricular tachycardia.

https://doi.org/10.1038/s41591-025-03516-x

Extended Data Table 3 | Duration and heart rate of false-negative and false-positive findings by either method

False negative findings

	AI					Technician				
	n	25th %	median	75th %	mean HR, bpm	n	25th %	median	75th %	mean HR, bpm
Atrial fibrillation	2	33.95	33.97	33.98	137.12	21	33.65	33.82	34.0	94.74
Sustained SVT	6	33.93	33.97	34.0	129.39	76	33.75	33.93	34.0	120.89
Asystole≥3.5 s	-	-	-	-	-	40	3.62	3.76	4.01	-
3rd degree AV block	4	4.79	5.75	6.89	36.43	55	1.54	2.24	4.77	41.57
Sustained VT ≥10 s	2	13.79	13.92	14.04	124.4	37	11.48	14.11	19.51	150.24
2nd degree AV block	-	-	-	-	-	89	1.01	1.20	1.35	52.08
Pause 2.0-3.4 s	3	2.06	2.07	2.16	-	65	2.02	2.05	2.14	-
VT 3 beats-9.9 s	-	-	-	-	-	48	1.53	1.69	1.90	153.48
Accelerated idioventricular rhythm	-	-	-	-	-	48	1.94	2.22	2.69	98.49
Idioventricular rhythm	-	-	-	-	-	83	2.67	3.03	3.65	65.84
SVT <30 s	-	-	-	-	-	9	1.77	1.99	2.19	153.48
Ectopic atrial rhythm	1	3.33	3.33	3.33	96.83	36	2.08	2.72	3.71	87.61

False positive findings

	AI					Technician				
	n	25th %	median	75th %	mean HR, bpm	n	25th %	median	75th %	mean HR, bpm
Atrial fibrillation	47	32.35	33.77	33.99	101.62	25	33.75	33.88	34.00	107.47
Sustained SVT	89	33.77	33.90	34.0	121.41	67	33.81	33.99	34.0	123.19
Asystole≥3.5 s	110	3.81	4.74	5.53	-	15	3.58	4.14	8.12	-
3rd degree AV block	102	1.25	1.71	3.18	46.78	19	3.84	6.42	8.62	42.38
Sustained VT ≥10 s	259	13.36	18.20	33.76	148.75	32	12.84	15.96	27.88	136.58
2nd degree AV block	248	1.06	1.36	1.58	49.15	16	2.49	5.94	14.40	43.83
Pause 2.0-3.4 s	38	2.07	2.14	2.35	-	15	1.98	2.02	2.42	-
VT 3 beats-9.9 s	71	1.50	1.73	2.16	155.7	12	1.73	2.33	2.60	135.75
Accelerated idioventricular rhythm	63	1.90	2.50	3.86	100.9	33	1.91	2.27	3.09	101.26
Idioventricular rhythm	93	2.77	3.49	5.13	62.76	13	3.70	3.90	4.85	65.43
SVT <30 s	38	1.72	1.89	2.49	123.09	33	1.66	2.20	2.67	132.44
Ectopic atrial rhythm	41	2.08	2.38	3.78	88.45	51	2.03	2.67	4.16	90.42

All durations are presented as median (interquartile range) in seconds, and all heart rates are presented as means. SVT, supraventricular tachycardia; VT, ventricular tachycardia; HR, heart rate.

https://doi.org/10.1038/s41591-025-03516-x

Extended Data Table 4 | Confusion matrix definitions

	For AI	For technician
True Positives (TP), per 1000 patients	Patients queried, in whom AI detected an arrhythmia and panel annotation was consistent	Patients queried in whom technicians detected an arrhythmia and panel annotation was consistent
True Negatives (TN), per 1000 patients (TN)	Patients queried in whom no arrhythmia was found by AI, excluding patients in whom technicians found an arrhythmia that had a consistent panel annotation	Patients queried in whom no arrhythmia was found by technicians, excluding patients in whom AI found an arrhythmia that had a consistent panel annotation
False Positive (FP), per 1000 patients (FPR)	Patients queried for an arrhythmia in whom AI detected an arrhythmia for which panel annotation was not consistent	Patients queried for an arrhythmia in whom technicians detected an arrhythmia for which panel annotation was not consistent
False Negative (FN), per 1000 patients (FNR)	Patients queried in whom no arrhythmia was found by AI, but technician found an arrhythmia confirmed by the panel	Patients queried in whom no arrhythmia was found by technician, but AI found an arrhythmia confirmed by the panel
Accuracy	$(TP_{AI} + TN_{AI})/(TP_{AI} + TN_{AI} + FP_{AI} + FN_{AI})$	$(TP_{tech} + TN_{tech})/(TP_{tech} + TN_{tech} + FP_{tech} + FN_{tech})$
True Positive rate (TPR) Sensitivity	$TP_{AI} / (TP_{AI} + FN_{AI})$	$TP_{tech} / (TP_{tech} + FN_{tech})$
True Negative Rate (TNR) Specificity	$TN_{AI} / (TN_{AI} + FP_{AI})$	$TN_{tech} / (TN_{tech} + FP_{tech})$
Positive Predictive Value (PPV)	$TP_{AI} / (TP_{AI} + FP_{AI})$	$TP_{tech} / (TP_{tech} + FP_{tech})$
Negative Predictive Value (NPV)	$TN_{AI} / (TN_{AI} + FN_{AI})$	$TN_{tech} / (TN_{tech} + FN_{tech})$
F1 Harmonic mean of PPV and TPR	$2* ((PPV_{AI} * TPR_{AI})/(PPV_{AI} + TPR_{AI})$	$2* ((PPV_{tech} * TPR_{tech})/(PPV_{tech} + TPR_{tech})$

Describes how the confusion matrix statistics are calculated from the annotated episodes.

nature portfolio

Corresponding author(s): Linda Johnson

Last updated by author(s): 25 November 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data extraction from ECGs was performed in Python v3.8.12, using custom written code and numpy 1.22.3, pandas 1.1.0, and torch 1.10.2 +cu113
Data analysis	Statistical analyses were mainly performed in Python v3.8.12, using custom written code and numpy 1.22.3, pandas 1.1.0, and torch 1.10.2 +cu113. Incidence rate ratios were calculated in Stat for Mac v 17.0 using epitab commands.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The study data that support the findings of this study are not publicly available due to their potentially sensitive nature but will be made available from the corresponding author for the purpose of reviewing the study results, at the cost of a reasonable fee for data preparation. No requests that include a commercial

interest will be approved. Data are located in controlled access data storage at MEDICALgorithmics. A response to a request to access the data can be expected within two months.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We report sub-analyses by sex for the most important study outcomes
Reporting on race, ethnicity, or other socially relevant groupings	We have no information regarding race, ethnicity or any other socially relevant grouping.
Population characteristics	The paper reports on an unselected patient population that has been referred to and undergone mobile cardiac telemetry in the US. The mean age is 65.5±10 years, and 42.8% of the population was male.
Recruitment	Patients were not specifically recruited for this study - they were included if they had undergone mobile cardiac telemetry in the US for a clinical indication, between 2016-2019, after anonymisation of the data.
Ethics oversight	The Ethics Review Board of Sweden has waived the need for an ethical approval, considering the anonymised nature of the data, decision number 2019-03227

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used the maximal population size of unselected patient recordings never before seen by the algorithm that was available to us (n=14,606 patient recordings). The number of strips annotated was pre-determined based on three criteria: 1. The maximal sample size that we could annotate using gold standard methodology, assuming that each annotator would spend a maximum of 40 hours to annotate the data, and the number of physicians in our network that were willing to contribute to the project. 2. What we considered to be a reasonable number of arrhythmic events in order to capture a diverse sample of electrographic presentations (minimum 250, but with a larger sample for critical arrhythmia events). 3. The maximum number of arrhythmic events found in the population sample of 14,606 patients that was available to you.
Data exclusions	No data was excluded
Replication	No external dataset was available to us for replication
Randomization	Arrhythmic episodes included in the study were randomly selected within each individual. No randomization was performed at the patient level - all ECG monitoring sessions were analysed with both DeepRhythmAI and ECG technician analysis (usual care)
Blinding	The annotators were blinded to the methods by which the strips included in the study were selected (except LSJ, A Benz, and JH who designed the study) and all annotators were blinded to whether they were annotating an AI or technician selected strip as well as to what the AI or technician annotations were.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.