



Validation of an automated artificial intelligence system for 12-lead ECG interpretation

Robert Herman, MD ^{a,b,c,*}, Anthony Demolder, MD, PhD ^c, Boris Vavrik, MSc ^c, Michal Martonak, MSc ^c, Vladimir Boza, MSc, PhD ^{c,d}, Viera Kresnakova, MSc, PhD ^{c,e}, Andrej Iring, MSc ^c, Timotej Palus, Ing ^c, Jakub Bahyl, MSc ^c, Olivier Nelis, MSc ^b, Monika Beles, RN ^b, Davide Fabbriatore, MD ^b, Leor Perl, MD ^f, Jozef Bartunek, MD, PhD ^b, Robert Hatala, MD, PhD ^{g,**}

^a Department of Advanced Biomedical Sciences, University of Naples Federico II, Naples, Italy

^b Cardiovascular Centre Aalst, Aalst, Belgium

^c Powerful Medical, Bratislava, Slovakia

^d Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Bratislava, Slovakia

^e Department of Cybernetics and Artificial Intelligence, Technical University of Kosice, Kosice, Slovakia

^f Department of Cardiology, Rabin Medical Center, Petah Tikvah, Israel

^g Department of Arrhythmia and Pacing, National Institute of Cardiovascular Diseases, Bratislava, Slovakia

ARTICLE INFO

Keywords:

Artificial intelligence
Computerized electrocardiogram
Rhythm analysis
Acute coronary syndrome
Conduction abnormality
Machine learning

ABSTRACT

Background: The electrocardiogram (ECG) is one of the most accessible and comprehensive diagnostic tools used to assess cardiac patients at the first point of contact. Despite advances in computerized interpretation of the electrocardiogram (CIE), its accuracy remains inferior to physicians. This study evaluated the diagnostic performance of an artificial intelligence (AI)-powered ECG system and compared its performance to current state-of-the-art CIE.

Methods: An AI-powered system consisting of 6 deep neural networks (DNN) was trained on standard 12-lead ECGs to detect 20 essential diagnostic patterns (grouped into 6 categories: rhythm, acute coronary syndrome (ACS), conduction abnormalities, ectopy, chamber enlargement and axis). An independent test set of ECGs with diagnostic consensus of two expert cardiologists was used as a reference standard. AI system performance was compared to current state-of-the-art CIE. The key metrics used to compare performances were sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score.

Results: A total of 932,711 standard 12-lead ECGs from 173,949 patients were used for AI system development. The independent test set pooled 11,932 annotated ECG labels. In all 6 diagnostic categories, the DNNs achieved high F1 scores: Rhythm 0.957, ACS 0.925, Conduction abnormalities 0.893, Ectopy 0.966, Chamber enlargement 0.972, and Axis 0.897. The diagnostic performance of DNNs surpassed state-of-the-art CIE for the 13 out of 20 essential diagnostic patterns and was non-inferior for the remaining individual diagnoses.

Conclusions: Our results demonstrate the AI-powered ECG model's ability to accurately identify electrocardiographic abnormalities from the 12-lead ECG, highlighting its potential as a clinical tool for healthcare professionals.

Introduction

The standard 12-lead electrocardiogram (ECG) is a comprehensive

diagnostic method readily available at the first point of contact, allowing for a rapid assessment of a wide spectrum of cardiac abnormalities. ECG interpretation is a comprehensive process that requires considerable

* Correspondence to: R Herman, Department of Biomedical Sciences, University of Naples Federico II, C.so Umberto I, 40, 80138 Naples, Italy.

** Correspondence to: R Hatala, Department of Arrhythmia and Pacing, National Institute of Cardiovascular Diseases, Pod Krásnou horkou 1, 833 84 Bratislava, Slovakia.

E-mail addresses: robi.herman@gmail.com (R. Herman), hatala@nusch.sk (R. Hatala).

<https://doi.org/10.1016/j.jelectrocard.2023.12.009>

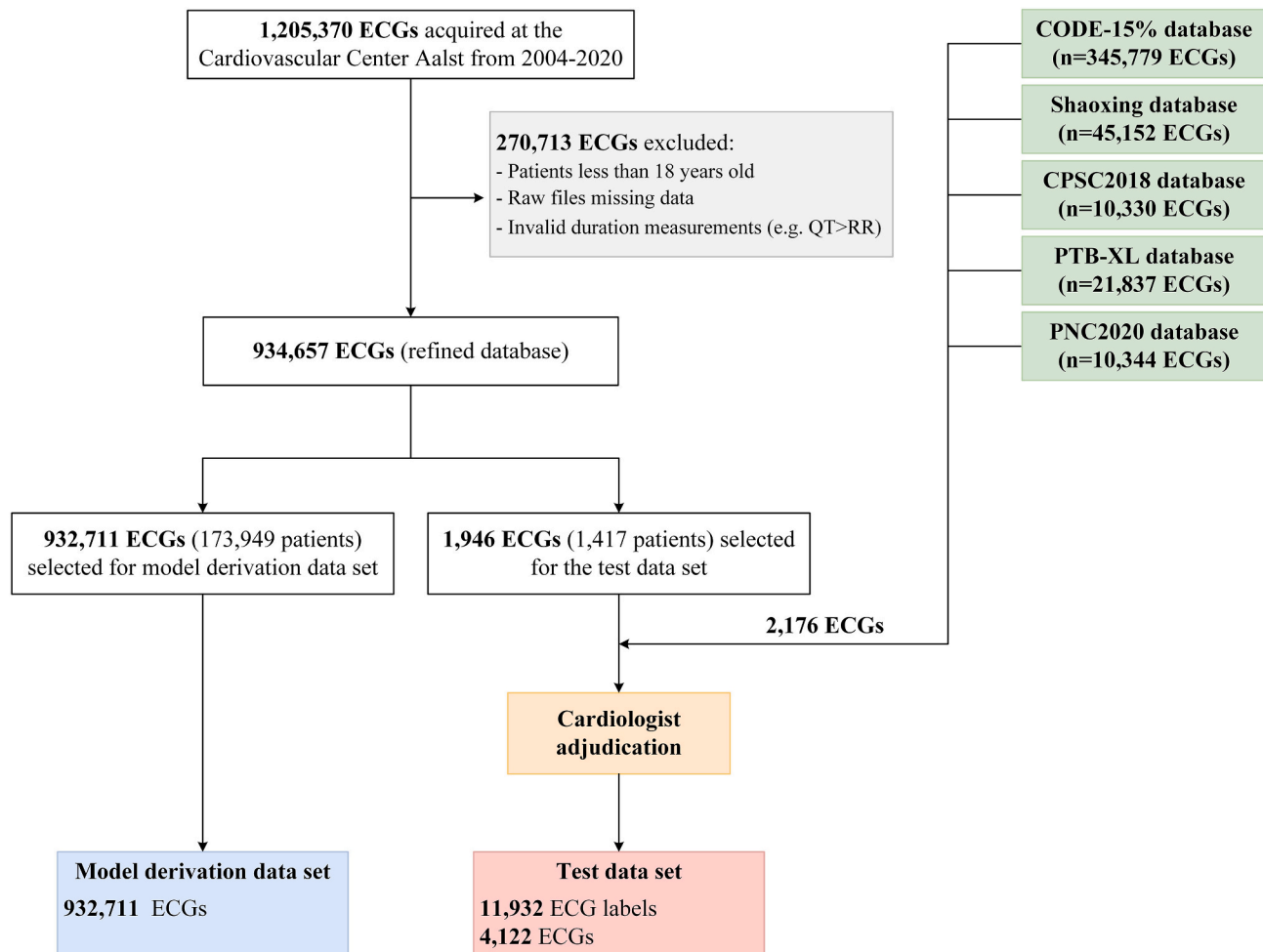


Fig. 1. PRISMA Flowchart of the data set creation.

expertise and training. It has been shown that physicians at all training levels have deficiencies in ECG interpretation, even after educational interventions. An extensive meta-analysis reports a median accuracy for ECG interpretation of 54% across all healthcare professional training levels, with non-cardiologist practicing physicians achieving 68.5% accuracy [1].

Computerized interpretation of the electrocardiogram (CIE) was introduced to improve ECG interpretation accuracy and reduce inter-rater variability. Despite improvements in diagnostic performance over time, traditional CIE often falls short in providing accurate interpretations, as evidenced by reported accuracy of 54% for interpreting non-sinus rhythms and error rates of 40.7% and 75.0% for diagnosing acute coronary syndrome or higher degree AV block, respectively [2–4]. Hence, CIE is viewed with skepticism and physicians are advised against relying on CIE to drive clinical decision-making [5]. It has been shown that the over-reading physician's interpretation is significantly influenced by CIE and errors are often overlooked [6–9], advocating more accurate ECG interpretation support.

Artificial intelligence (AI) developed using large databanks of ECG data showed promising results in landmark studies [10–13]. Although, their interpretation remained limited to a select number of supported ECG diagnoses and were not validated on adequate sample sizes, preventing wide-scale adoption of AI-powered ECG interpretation. We pursued the development of an AI-powered system that provides accurate detection of major ECG abnormalities addressing the limitations of traditional CIE. We hypothesized that this AI system will surpass the performance of CIE evaluated in a large external dataset.

Methods

Study design

This retrospective study followed three key steps: (1) development of an AI-powered system composed of multiple deep neural networks (DNNs) detecting essential diagnostic patterns on 12-lead ECGs (“AI system development”); (2) evaluation of the AI system performance on a separate, independent dataset of ECGs (“test set”); (3) comparison of the AI system to a state-of-the-art CIE algorithm. The study was approved by the local ethics committee for human research and complied with the Declaration of Helsinki.

Model development data

A total of 1,205,370 ECGs acquired between 01/1/2004 to 31/12/2020 were extracted from the Cardiovascular Center Aalst data vault. After filtering out 270,713 ECGs (patients younger than 18 years at the time of acquisition, raw files missing data, invalid duration measurements), the refined ECG data set (934,657 ECGs) was used to sample the model derivation data set and the test set (Fig. 1). Raw waveform data from standard 10 s 12-lead ECGs recorded at a sampling rate of 500 Hz (GE Healthcare, Milwaukee, WI, USA) was extracted. There was no preprocessing performed on the waveform data (no additional filtering, resampling or other techniques). For the model derivation data set, the original diagnostic statements (automatically generated by the Marquette 12SL algorithm [version 2005], GE Healthcare, Milwaukee, WI,

USA) were mapped into 20 diagnostic patterns (Appendix 1, Supplemental Table 1). The term “ECG label” refers to a diagnostic label assigned to any given ECG, either positive (indicating the presence of a specific pattern), or negative (indicating its absence). For the ECG duration measurements model, the standard ECG measurements automatically calculated by the Marquette 12SL (heart rate, P-wave duration, PR-interval, QRS, duration, and QT interval) were used for model development.

Primary and secondary outcomes

The primary outcome was the ability of an AI system to accurately detect 20 essential diagnostic patterns and standard ECG measurements on 12-lead ECGs included in the test set. The AI system performance on each diagnostic pattern was assessed using a majority vote by two expert cardiologists as the reference standard. ECG measurements evaluation followed the International Electrotechnical Commission (IEC) 60,601–2-25:2015 standard, which mandates reporting measurement results on expert-annotated ECGs from the Common Standards for Electrocardiography (CSE) database [14,15]. Secondary outcomes included the comparison of AI system performance to a state-of-the-art CIE (Marquette 12SL algorithm [version 2005], GE Healthcare) on the subset of 1946 (47.8%) of ECGs in the test set with available Marquette 12SL diagnostic statements.

Model development

The AI system consists of two components: one detecting the diagnostic patterns (diagnostic component) and another for the ECG duration measurements (measurements component). Our approach for ECG diagnostics leverages the power of multiple DNNs. A random 5 s segment was chosen for each lead from the 10s 12-lead raw ECG (sampled at 500 Hz) as input for each of the 6 DNNs. For DNN testing, the first 5 s for limb leads and the last 5 s for precordial leads of the standard 6 × 2 ECG format were used. For the DNN model development, a randomly selected 5 s segment from the 10s 12-lead raw ECG, sampled at 500 Hz, was used for each lead. For model testing, the first 5 s for limb leads and the last 5 s for precordial leads were used, mirroring the standard 6 × 2 ECG format. This method was intentionally chosen to provide an equivalent informational basis for both the DNN and the cardiologists annotating the test set. The DNN architecture comprises two key components: feature extraction and classification (Supplemental Fig. 1). The feature extraction component consists of 15 Convolutional layers, designed to extract features from leads. The second component, classification, combines all extracted features and processes them through 3 fully connected layers, interspersed with dropouts. Analysis of each lead and integration of the knowledge gained mimics the analytical approach of human experts to make a final diagnosis. The network utilizes the Adam optimizer, ReLU activation functions, and Dropout for regularization. The training phase was terminated once the model's performance on the tuning dataset ceased to improve, also known as early stopping.

Test set creation

For model testing, an independent test set was derived by collecting ECGs from multiple sources to ensure robust performance evaluation and demonstrate generalizability. In addition to ECGs sampled from the Cardiovascular Center Aalst ECG data vault, external data sources included the CODE-15% database (Telehealth Network of Minas Gerais), PTB-XL (Physikalisch Technische Bundesanstalt), Shaoxing database (Shaoxing People's Hospital and Ningbo First Hospital), PNC2020 (PhysioNet/Computing in Cardiology Challenge 2020), CPSC2018 (China Physiological Signal Challenge 2018). A minimum requirement of 50 positive and 50 negative ECG labels for each diagnostic pattern in our test set was set, taking into consideration the test set sizes used in

Table 1
Patient demographics of the model derivation data set and testing data set.

Parameter	Model derivation data set	Test set
Counts		
Number of ECGs, n	932,711	4122
Number of unique patients, n	173,949	3593
Age group		
≤25, n (%)	14,909 (1.6%)	149 (3.6%)
26–40, n (%)	39,246 (4.2%)	353 (8.6%)
41–60, n (%)	209,443 (22.5%)	948 (23.0%)
61–80, n (%)	458,436 (49.2%)	1912 (46.4%)
≥81, n (%)	154,499 (16.6%)	709 (17.2%)
Missing info, n (%)	56,178 (6.0%)	51 (1.2%)
Sex		
Male, n (%)	487,721 (52.3%)	2342 (56.8%)
Female, n (%)	352,948 (37.8%)	1774 (43.0%)
Missing info, n (%)	92,042 (9.9%)	6 (0.1%)
Data origin		
Cardiovascular Center Aalst ECG data vault, n (%) ^a	932,711 (100%)	1946 (47.2%)
CODE-15%, n (%)	0 (0%)	1067 (25.9%)
Shaoxing, n (%)	0 (0%)	551 (13.4%)
CPSC2018, n (%)	0 (0%)	206 (5.0%)
PTB-XL, n (%)	0 (0%)	180 (4.4%)
PNC2020, n (%)	0 (0%)	172 (4.2%)

ECG, electrocardiogram; n, number.
^a State-of-the-art CIE (GE Marquette 12SL) diagnostic statements available.

previous studies for guidance [10–13]. A major vote of two expert cardiologists with extensive experience in electrocardiography was considered the reference standard. If the two expert cardiologists agreed, the shared diagnosis was considered as the reference standard. In case of disagreement, the ECGs were removed from the test set to ensure sufficient ECG quality for interpretation. There was no patient overlap between the model derivation data set and the test set. An expanded methodology for the test set creation is available in Appendix 2 of the Supplementary files.

Statistical analysis

Metrics used for statistical evaluation of diagnostic performance include sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score, all with 95% confidence intervals (CI). For the DNN, the CI was derived using the Wald 95% CI method [16]. Performance metrics are presented as average values (95% CI), grouped per diagnostic category. The benchmark aimed to determine whether the point average of the DNN is above, below or within the CI of the comparator. All code was written in Python programming language and executed within the environment of AWS Sagemaker, either as a Processing job or a Development job. The development was conducted on a g4dn.4xlarge machine, equipped with 64 GiB of memory and 225 GB NVMe SSD storage. Analysis and visualization of the results were performed in Python using the standard data-science libraries Pandas, NumPy, Scipy, Scikit-Learn, Matplotlib, Seaborn libraries.

Results

Study population

A total of 932,711 ECGs from 173,949 unique patients were used for model development (model derivation set) (Fig. 1). The test set consisted of 4122 unique ECGs with 11,932 annotated ECG labels (1946

Table 2
Full diagnostic performance of the DNN on the full annotated test set ($N = 4122$ ECGs with 11,932 ECG labels).

Diagnostic pattern	Positive samples	Negative samples	Sensitivity	Specificity	PPV	NPV	F1 score
Rhythm			0.957 (0.952–0.962)	0.989 (0.987–0.992)	0.957 (0.952–0.962)	0.989 (0.987–0.992)	0.957 (0.952–0.962)
Sinus rhythm	389	772	0.990 (0.984–0.996)	0.969 (0.959–0.979)	0.941 (0.928–0.955)	0.995 (0.990–0.999)	0.965 (0.954–0.975)
Paced rhythm	275	886	0.978 (0.970–0.987)	0.986 (0.980–0.993)	0.957 (0.946–0.969)	0.993 (0.988–0.998)	0.968 (0.957–0.978)
Atrial fibrillation	192	969	0.948 (0.935–0.961)	0.998 (0.995–1.000)	0.989 (0.983–0.995)	0.990 (0.984–0.996)	0.968 (0.958–0.978)
Atrial flutter	159	1002	0.969 (0.959–0.979)	0.995 (0.991–0.999)	0.969 (0.959–0.979)	0.995 (0.991–0.999)	0.969 (0.959–0.979)
Other rhythm	146	1015	0.829 (0.807–0.850)	0.993 (0.988–0.998)	0.945 (0.932–0.958)	0.976 (0.967–0.985)	0.883 (0.865–0.902)
ACS			0.930 (0.912–0.948)	0.957 (0.943–0.971)	0.920 (0.901–0.939)	0.963 (0.949–0.976)	0.925 (0.907–0.943)
Suspected ST-elevation ACS (STEMI)	162	230	0.994 (0.986–1.000)	0.926 (0.900–0.952)	0.904 (0.875–0.934)	0.995 (0.989–1.000)	0.947 (0.925–0.969)
Suspected Non-ST-elevation ACS (NSTEMI)	110	282	0.836 (0.800–0.873)	0.982 (0.969–0.995)	0.948 (0.927–0.970)	0.939 (0.915–0.963)	0.889 (0.858–0.920)
Conduction abnormalities			0.864 (0.852–0.876)	0.968 (0.962–0.974)	0.925 (0.916–0.934)	0.939 (0.931–0.948)	0.893 (0.882–0.904)
Left bundle branch block	189	349	0.947 (0.928–0.966)	0.994 (0.988–1.000)	0.989 (0.980–0.998)	0.972 (0.958–0.986)	0.968 (0.953–0.983)
Right bundle branch block	178	360	0.994 (0.988–1.000)	0.942 (0.922–0.961)	0.894 (0.868–0.920)	0.997 (0.992–1.000)	0.941 (0.922–0.961)
Left anterior fascicular block	155	321	0.910 (0.884–0.935)	0.975 (0.961–0.989)	0.946 (0.926–0.967)	0.957 (0.939–0.975)	0.928 (0.904–0.951)
Left posterior fascicular block	162	314	0.821 (0.787–0.855)	0.978 (0.964–0.991)	0.950 (0.930–0.970)	0.914 (0.888–0.939)	0.881 (0.852–0.910)
2nd degree AV block Mobitz type I (Wenckebach)	87	417	0.747 (0.709–0.785)	0.971 (0.957–0.986)	0.844 (0.812–0.876)	0.948 (0.929–0.968)	0.793 (0.757–0.828)
Higher degree AV block	185	319	0.708 (0.668–0.748)	0.947 (0.927–0.966)	0.885 (0.857–0.913)	0.848 (0.817–0.880)	0.787 (0.751–0.823)
Ectopy	184	184	0.940 (0.916–0.964)	0.995 (0.987–1.000)	0.994 (0.987–1.000)	0.943 (0.920–0.967)	0.966 (0.948–0.985)
Chamber enlargement			0.991 (0.984–0.998)	0.928 (0.909–0.946)	0.954 (0.939–0.969)	0.985 (0.977–0.994)	0.972 (0.960–0.984)
Suspected atrial enlargement	270	133	0.996 (0.990–1.000)	0.932 (0.908–0.957)	0.968 (0.950–0.985)	0.992 (0.983–1.000)	0.982 (0.969–0.995)
Suspected ventricular hypertrophy	171	157	0.982 (0.968–0.997)	0.924 (0.895–0.952)	0.933 (0.906–0.960)	0.980 (0.964–0.995)	0.957 (0.935–0.979)
Axis			0.897 (0.880–0.914)	0.966 (0.956–0.976)	0.897 (0.880–0.914)	0.966 (0.956–0.976)	0.897 (0.880–0.914)
Normal axis	54	248	0.990 (0.979–1.000)	0.864 (0.826–0.903)	0.791 (0.745–0.837)	0.994 (0.986–1.000)	0.879 (0.843–0.916)
Left axis deviation	91	211	0.747 (0.698–0.796)	0.995 (0.988–1.000)	0.986 (0.972–0.999)	0.901 (0.868–0.935)	0.850 (0.810–0.890)
Right axis deviation	54	248	0.907 (0.875–0.940)	0.992 (0.982–1.000)	0.961 (0.939–0.983)	0.980 (0.964–0.996)	0.933 (0.905–0.961)
Extreme axis deviation	103	199	0.963 (0.942–0.984)	0.996 (0.989–1.000)	0.981 (0.966–0.996)	0.992 (0.982–1.000)	0.972 (0.953–0.991)

ACS, acute coronary syndrome; AV, atrioventricular; ECG, electrocardiogram; NSTEMI, Non-ST-elevation myocardial infarction; STEMI, ST-elevation myocardial infarction.

ECGs [47.8%] from the Cardiovascular Center Aalst cohort enriched with 2176 ECGs [52.8%] from external sources). Study population demographics are summarized in [Table 1](#).

DNN model performance

Results of diagnostic performance on the test set are provided in [Table 2](#) and [Fig. 2](#). In all 6 diagnostic categories, the DNNs achieved high F1 scores: Rhythm 0.957, acute coronary syndromes (ACS) 0.925, Conduction abnormalities 0.893, Ectopy 0.966, Chamber enlargement 0.972, and Axis 0.897. The DNN performance yielded high F1 scores for each individual rhythm pattern and the ability of the DNN to identify atrial fibrillation achieved nearly perfect performance ([Fig. 3, panel A](#)), as evidenced by the sensitivity 0.948 (95% CI 0.935–0.961), specificity 0.998 (95% CI 0.995–1.000), PPV 0.989 (95% CI 0.983–0.995), NPV 0.990 (95% CI 0.984–0.996), and F1 score 0.968 (95% CI 0.958–0.978). In the detection of ST-elevation ACS, the DNN achieved sensitivity 0.994

(95% CI 0.986–1.000), specificity 0.926 (95% CI 0.900–0.952), PPV 0.904 (95% CI 0.875–0.934), NPV 0.995 (95% CI 0.989–1.000), and F1 score 0.947 (95% CI 0.925–0.969) ([Fig. 3, Panel B](#)). Results on the external subset of the testing dataset are comparable, indicating that the DNN generalizes well beyond the Aalst ECG data cohort ([Supplemental Table 2](#)).

For the ECG measurements, average differences (with their acceptable thresholds) for P wave duration, PR interval, QRS duration, QT interval, and RR interval were 9.2 ms (± 10 ms), -1.1 ms (± 10 ms), 3.0 ms (± 10 ms), -4.1 ms (± 25 ms) and -0.3 ms (± 25 ms), respectively. The automated measurements passed all criteria proposed by the CSE standards. [Table 3](#) provides the results of automated measurements performed on expert-annotated ECGs from the CSE database.

Benchmarking

Comparison to a current state-of-the-art CIE was performed on 1946

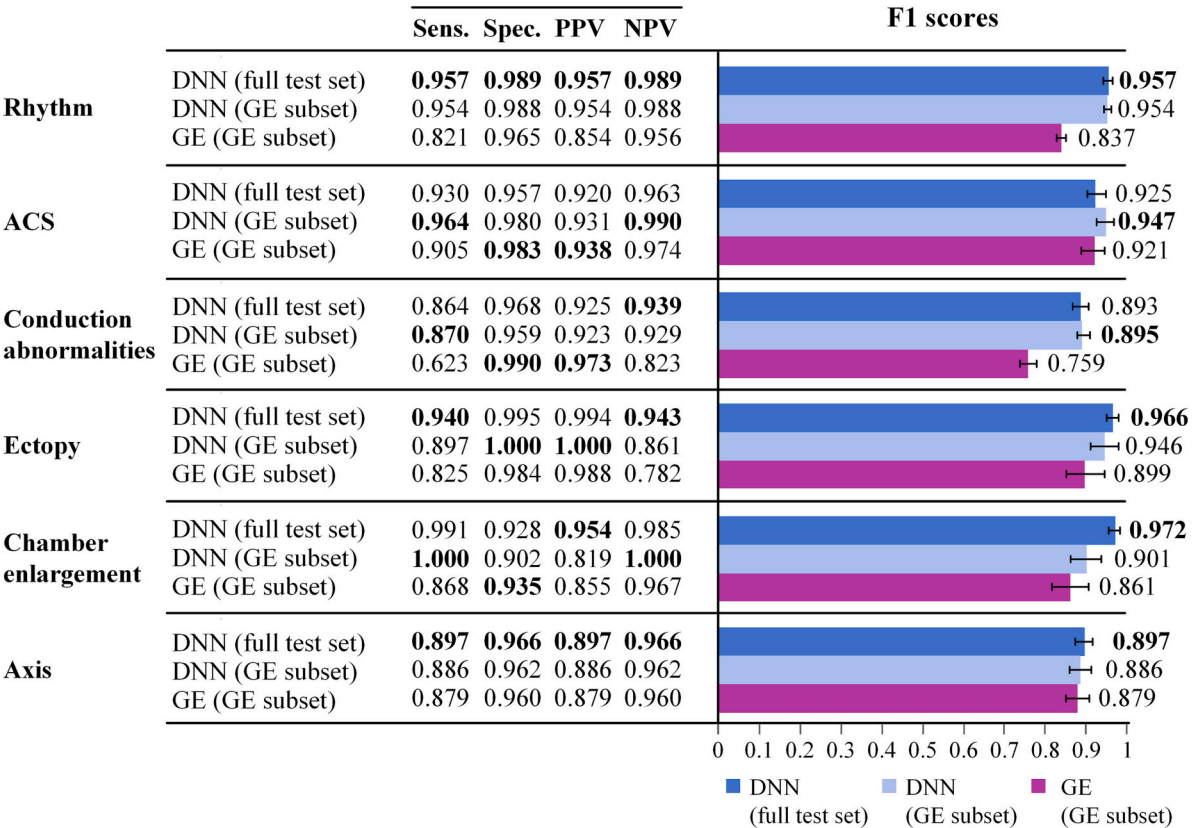


Fig. 2. Diagnostic performance comparison between DNN and GE Marquette 12SL. Bold values represent the highest performance for that diagnosis. ACS, acute coronary syndrome; DNN, deep neural network; GE, General Electric Marquette 12SL algorithm; NPV, negative predictive value; PPV, positive predictive value; Sens, sensitivity; Spec, specificity.

ECGs (47.8% of entire test set) with available Marquette 12SL diagnostics statements. In this comparison, the DNNs achieved significantly higher F1 scores than the CIE for 13 out of the 20 diagnostic patterns (Table 4). For the overall rhythm assessment, the F1 score of the DNNs was superior to the CIE (0.954 vs. 0.837, $P < 0.05$), surpassing the CIE for each individual rhythm pattern. Compared to the CIE, the DNN reduced false negatives by 41.7% and increased true positives by 5.7% for the diagnosis of AF. Likewise, the DNN reduced false negatives to 0 and increased true positives by 19.2% for the diagnosis of STEMI compared to the CIE. For the diagnosis of conduction abnormalities, DNN performance was significantly greater compared to the CIE (F1 score 0.895 vs. 0.759, $P < 0.05$), with the exception of LBBB (equal performance). In the three diagnostic patterns exhibiting the lowest CIE performance, namely LPFB, 2nd degree AV block Mobitz type I, and Higher degree AV block, the DNNs showcased a significant enhancement in F1-score performance (0.901 vs. 0.495; 0.768 vs. 0.690; 0.813 vs. 0.526, respectively, $P < 0.05$ for all). The DNN was non-inferior compared to the CIE in the diagnostic class of ACS, ectopy, chamber enlargement and cardiac axis. For each individual diagnostic pattern, the AI system performance was either significantly better ($P < 0.05$) or non-inferior compared to the CIE, as adjudicated by the F1 scores.

Discussion

In this study, we present a novel AI-powered system composed of multiple DNNs detecting 20 essential ECG patterns and standard ECG measurements and compared its performance to state-of-the-art CIE. Trained on over 900,000 ECGs, the diagnostic performance of the AI system surpassed the state-of-the-art CIE for 13 out of 20 evaluated diagnostic patterns and was non-inferior for the remaining. These findings emphasize the potential of a DNN approach to substantially

improve the accuracy of computerized ECG interpretation in clinical practice.

ECG interpretation plays a critical role in the primary diagnosis and management of cardiac patients at the first point of contact. However, accuracy varies widely across all training levels of physicians and remains suboptimal even after training intervention [1,17]. Current CIE approaches do not reliably address this, with incorrect interpretations in up to 33% cases, leading to unnecessary diagnostic testing and initiation of inappropriate treatment. Furthermore, existing AI approaches often do not support an adequate number of diagnostic patterns, omit quantifying standard ECG measurements, lack validation in sizeable, external ECG datasets and do not report a comprehensive range of metrics potentially overestimating their effectiveness. [10–13].

Our present study bears several methodological strengths. First, we present a comprehensive AI system consisting of 6 DNNs detecting 20 essential diagnostic patterns and 5 measurements. The F1 scores of the AI system for identifying the diagnostic patterns were excellent and all measurements passed the threshold requirements. Second, we have validated the performance of DNNs on a large, independent and diverse test data set of ECGs. Third, the AI system performance was compared to a current state-of-the-art (GE Marquette 12SL). The DNNs demonstrated robust performance in ECGs where the CIE achieved sub-optimal performance. By effectively reducing false positives without increasing false negatives, implementation of AI-based ECG interpretation could greatly reduce the burden posed on cardiologists, who would only be required to provide their expertise for more intricate cases (Such as those where the numeric DNN output is close to the threshold for the diagnostic pattern).

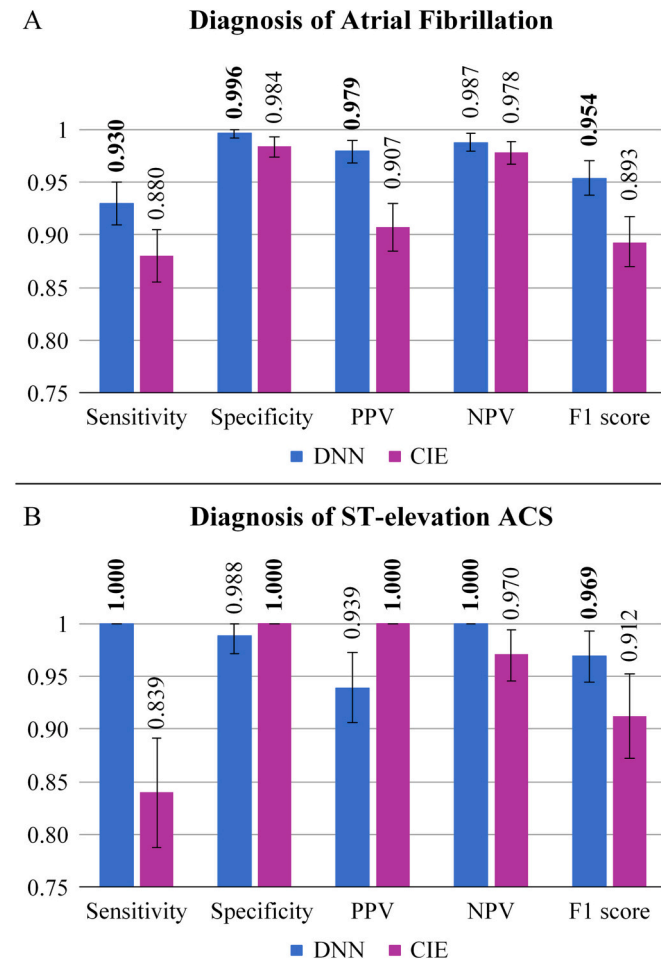


Fig. 3. Diagnostic performance of the DNN and GE Marquette 12SL for detection of atrial fibrillation and suspected ST-elevation acute coronary syndrome. Bold values represent the highest performance for that diagnosis. ACS, acute coronary syndrome; DNN, deep neural network; GE, General Electric Marquette 12SL algorithm; NPV, negative predictive value; PPV, positive predictive value.

Clinical implications

Our study has important clinical implications for ECG interpretation at the first point of patient contact. Fusing identified diagnostic patterns and measurements, the AI system can detect up to 38 clinically relevant diagnoses, 4 cardiac axes and 5 ECG measurements (*Supplemental Table 1*), thereby recommending comprehensive ECG interpretation. The AI system showed excellent performance especially in detecting AF, offering a valuable tool to improve early detection of AF in primary care. Likewise, this approach has the potential to accelerate management in emergency care through rapid pre-clinical diagnosis of STEMI. These benefits are relevant in the context of ever-increasing burden of cardiovascular diseases on the healthcare system [18], providing the

opportunity to improve timely and efficient referral with favorable impact on clinical outcomes. With the addition of new clinically validated data, DNNs have the potential to enhance continuously, improving their reliability.

Limitations

Following limitations should be acknowledged. One limitation of our study is the exclusion of ECGs from the dataset in instances of disagreement between the two expert cardiologists. While this approach was adopted to ensure sufficient ECG quality and maintain reliability of the dataset, this may have resulted in removal of more complex cases. Although the test dataset was enriched by external sources to ensure a diverse range of ECGs, further independent validation is needed to demonstrate generalizability of the AI system. A direct one-to-one comparison to existing AI models was not possible because of the unavailability of their source code or dataset availability for external validation environment. The selection of diagnostic patterns in this study is not all-inclusive, some patterns such as signs of old infarction were not included in the scope of the AI system. While some ECGs had multiple abnormalities present, performance analysis of the AI system on these concomitant ECG findings would introduce a high degree of complexity which was beyond the intended scope of this study. Further research is needed to examine the prospective efficacy of the AI-powered system and healthcare professional adherence to AI-based interpretation.

Conclusion

The ability of an AI-powered ECG model to identify and learn features and patterns from a large amount of ECG data has significantly attenuated the rate of misdiagnosis, exceeding current state-of-the-art CIE. As such, the algorithm's ability to accurately identify cardiac abnormalities from the 12-lead ECG showcases its utility as a clinical decision-support tool for healthcare professionals.

Tweet proposal

AI-powered #ECG interpretation system accurately identifies cardiac abnormalities and outperforms a traditional state-of-the-art CIE in a large 12-lead ECG dataset! This #AI diagnostic tool could be a valuable asset to healthcare professionals in the detecting cardiac disease at the first point of contact. @RobertHermanMD.

Ethical committee approval

The study was approved by the local ethics committee for human research and complied with the Declaration of Helsinki.

CRediT authorship contribution statement

Robert Herman: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. **Anthony**

Table 3
Results of ECG measurements on the CSE Multilead dataset.

Measurement	Number of ECGs	Average diff. (ms)	SD (ms)	Acceptable average diff. (ms) ^a	Acceptable SD (ms) ^a	Adjudication
P wave duration	92	9.2	7.3	±10	15	Pass
PR interval	92	−1.1	7.4	±10	10	Pass
QRS duration	92	3.0	6.2	±10	10	Pass
QT interval	92	−4.1	11.0	±25	30	Pass
RR interval	92	−0.3	7.6	±25	30	Pass

CSE, Common Standards of Electrocardiography; SD, standard deviation.

^a Criteria proposed by the Common Standards of Electrocardiography (CSE).

Table 4

Diagnostic performance of the DNN and the GE Marquette 12SL algorithm on the subset with available GE statements ($N = 1946$ ECGs; 47% of entire test set). Bold values denote a significantly better performance for that specific metric.

Diagnostic pattern	Counts		Sensitivity		Specificity		PPV		NPV		F1 score	
	Positive samples	Negative samples	DNN	GE	DNN	GE	DNN	GE	DNN	GE	DNN	GE
Rhythm			0.954 (0.946–0.961)	0.821 (0.807–0.834)	0.988 (0.985–0.992)	0.965 (0.958–0.971)	0.954 (0.946–0.961)	0.854 (0.842–0.866)	0.988 (0.985–0.992)	0.956 (0.949–0.963)	0.954 (0.946–0.961)	0.837 (0.824–0.850)
Sinus rhythm	147	500	0.980 (0.969–0.990)	0.993 (0.987–1.000)	0.976 (0.964–0.988)	0.928 (0.908–0.948)	0.923 (0.903–0.944)	0.802 (0.772–0.833)	0.994 (0.988–1.000)	0.998 (0.994–1.000)	0.950 (0.934–0.967)	0.888 (0.863–0.912)
Paced rhythm	242	405	0.979 (0.968–0.990)	0.702 (0.667–0.738)	0.983 (0.973–0.993)	0.998 (0.994–1.000)	0.971 (0.958–0.984)	0.994 (0.988–1.000)	0.988 (0.979–0.996)	0.849 (0.821–0.876)	0.975 (0.963–0.987)	0.823 (0.794–0.853)
Atrial fibrillation	100	547	0.930 (0.910–0.950)	0.880 (0.855–0.905)	0.996 (0.992–1.000)	0.984 (0.974–0.993)	0.979 (0.968–0.990)	0.907 (0.885–0.930)	0.987 (0.979–0.996)	0.978 (0.967–0.989)	0.954 (0.938–0.970)	0.893 (0.870–0.917)
Atrial flutter	54	593	0.963 (0.948–0.978)	0.833 (0.805–0.862)	0.993 (0.987–1.000)	0.993 (0.987–1.000)	0.929 (0.909–0.948)	0.918 (0.897–0.939)	0.997 (0.992–1.000)	0.985 (0.976–0.994)	0.945 (0.928–0.963)	0.874 (0.848–0.899)
Other rhythm	104	543	0.875 (0.850–0.900)	0.788 (0.757–0.820)	0.991 (0.983–0.998)	0.924 (0.904–0.945)	0.948 (0.931–0.965)	0.667 (0.630–0.703)	0.976 (0.965–0.988)	0.958 (0.943–0.973)	0.910 (0.888–0.932)	0.722 (0.688–0.757)
ACS			0.964 (0.946–0.983)	0.905 (0.875–0.934)	0.980 (0.966–0.994)	0.983 (0.971–0.996)	0.931 (0.906–0.956)	0.938 (0.914–0.962)	0.990 (0.980–1.000)	0.974 (0.958–0.99)	0.947 (0.925–0.970)	0.921 (0.894–0.948)
Suspected ST-elevation ACS (STEMI)	31	161	1.000 (1.000–1.000)	0.839 (0.787–0.891)	0.988 (0.972–1.000)	1.000 (1.000–1.000)	0.939 (0.906–0.973)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	0.970 (0.946–0.994)	0.969 (0.944–0.993)	0.912 (0.872–0.952)
Suspected Non-ST-elevation ACS (NSTEMI)	53	139	0.943 (0.911–0.976)	0.943 (0.911–0.976)	0.971 (0.948–0.995)	0.964 (0.938–0.990)	0.926 (0.889–0.963)	0.909 (0.868–0.950)	0.978 (0.958–0.999)	0.978 (0.957–0.999)	0.935 (0.900–0.970)	0.926 (0.889–0.963)
Conduction abnormalities			0.870 (0.852–0.887)	0.623 (0.598–0.648)	0.959 (0.949–0.969)	0.990 (0.985–0.995)	0.923 (0.909–0.937)	0.973 (0.965–0.981)	0.929 (0.916–0.942)	0.823 (0.804–0.843)	0.895 (0.880–0.911)	0.759 (0.737–0.781)
Left bundle branch block	126	139	0.929 (0.898–0.960)	0.897 (0.860–0.933)	0.986 (0.971–1.000)	0.993 (0.983–1.000)	0.983 (0.968–0.999)	0.991 (0.980–1.000)	0.938 (0.909–0.967)	0.914 (0.880–0.948)	0.955 (0.930–0.980)	0.942 (0.913–0.970)
Right bundle branch block	90	175	0.989 (0.976–1.000)	0.844 (0.801–0.888)	0.954 (0.929–0.979)	0.989 (0.976–1.000)	0.918 (0.884–0.951)	0.974 (0.955–0.993)	0.994 (0.985–1.000)	0.925 (0.893–0.957)	0.952 (0.926–0.978)	0.905 (0.869–0.940)
Left anterior fascicular block	79	132	0.886 (0.843–0.929)	0.582 (0.516–0.649)	0.962 (0.936–0.988)	0.992 (0.981–1.000)	0.933 (0.900–0.967)	0.979 (0.959–0.998)	0.934 (0.900–0.967)	0.799 (0.745–0.853)	0.909 (0.870–0.948)	0.730 (0.670–0.790)
Left posterior fascicular block	74	137	0.865 (0.819–0.911)	0.338 (0.274–0.402)	0.971 (0.948–0.994)	0.985 (0.969–1.000)	0.941 (0.909–0.973)	0.926 (0.891–0.961)	0.930 (0.896–0.964)	0.734 (0.674–0.793)	0.901 (0.861–0.942)	0.495 (0.428–0.563)
2nd degree AV block Mobitz type I (Wenckebach)	53	196	0.717 (0.661–0.773)	0.547 (0.485–0.609)	0.959 (0.935–0.984)	0.990 (0.977–1.000)	0.826 (0.779–0.873)	0.935 (0.905–0.966)	0.926 (0.894–0.959)	0.890 (0.851–0.929)	0.768 (0.715–0.820)	0.690 (0.633–0.748)
Higher degree AV block	100	149	0.760 (0.707–0.813)	0.360 (0.300–0.420)	0.926 (0.894–0.959)	0.993 (0.983–1.000)	0.874 (0.832–0.915)	0.973 (0.953–0.993)	0.852 (0.808–0.896)	0.698 (0.641–0.755)	0.813 (0.764–0.861)	0.526 (0.464–0.588)
Ectopy	97	62	0.897 (0.850–0.944)	0.825 (0.766–0.884)	1.000 (1.000–1.000)	0.984 (0.964–1.000)	1.000 (1.000–1.000)	0.988 (0.970–1.000)	0.861 (0.807–0.915)	0.782 (0.718–0.846)	0.946 (0.910–0.981)	0.899 (0.852–0.946)
Chamber enlargement			1.000 (1.000–1.000)	0.868 (0.823–0.912)	0.902 (0.863–0.941)	0.935 (0.902–0.967)	0.819 (0.769–0.870)	0.855 (0.809–0.901)	1.000 (1.000–1.000)	0.941 (0.910–0.972)	0.901 (0.861–0.940)	0.861 (0.816–0.907)
Suspected atrial enlargement	39	62	1.000 (1.000–1.000)	0.949 (0.906–0.992)	0.919 (0.866–0.972)	0.952 (0.910–0.993)	0.886 (0.824–0.948)	0.925 (0.874–0.976)	1.000 (1.000–1.000)	0.967 (0.932–1.000)	0.940 (0.893–0.986)	0.937 (0.889–0.984)
Suspected ventricular hypertrophy	29	91	1.000 (1.000–1.000)	0.759 (0.682–0.835)	0.890 (0.834–0.946)	0.923 (0.875–0.971)	0.744 (0.665–0.822)	0.759 (0.682–0.835)	1.000 (1.000–1.000)	0.923 (0.875–0.971)	0.853 (0.790–0.916)	0.759 (0.682–0.835)
Axis			0.886 (0.859–0.913)	0.879 (0.851–0.907)	0.962 (0.946–0.978)	0.960 (0.943–0.976)	0.886 (0.859–0.913)	0.879 (0.851–0.907)	0.962 (0.946–0.978)	0.960 (0.943–0.976)	0.886 (0.859–0.913)	0.879 (0.851–0.907)
Normal axis	34	98	1.000 (1.000–1.000)	0.971 (0.942–0.999)	0.878 (0.822–0.933)	0.908 (0.859–0.957)	0.739 (0.664–0.814)	0.786 (0.716–0.856)	1.000 (1.000–1.000)	0.989 (0.971–1.000)	0.850 (0.789–0.911)	0.868 (0.811–0.926)
Left axis deviation	35	97	0.657 (0.576–0.738)	0.714 (0.637–0.791)	0.990 (0.972–1.000)	0.959 (0.925–0.993)	0.958 (0.924–0.992)	0.862 (0.803–0.921)	0.889 (0.835–0.943)	0.903 (0.852–0.953)	0.780 (0.709–0.850)	0.781 (0.711–0.852)
Right axis deviation	31	101	0.968 (0.938–0.998)	1.000 (1.000–1.000)	0.990 (0.973–1.000)	0.980 (0.956–1.000)	0.968 (0.938–0.998)	0.939 (0.899–0.980)	0.990 (0.973–1.000)	1.000 (1.000–1.000)	0.968 (0.938–0.998)	0.969 (0.939–0.998)
Extreme axis deviation	32	100	0.938 (0.896–0.979)	0.844 (0.782–0.906)	0.990 (0.973–1.000)	0.990 (0.973–1.000)	0.968 (0.938–0.998)	0.964 (0.933–0.996)	0.980 (0.956–1.000)	0.952 (0.915–0.988)	0.952 (0.916–0.989)	0.900 (0.849–0.951)

AV, atrioventricular; DNN, deep neural network; GE, General Electric Marquette 12SL algorithm; NSTEMI, Non-ST-elevation myocardial infarction; STEMI, ST-elevation myocardial infarction.

Demolder: Methodology, Validation, Visualization, Writing – review & editing. **Boris Vavrik:** Conceptualization, Formal analysis, Methodology, Software. **Michal Martonak:** Data curation, Formal analysis, Methodology, Software, Visualization. **Vladimir Boza:** Formal analysis, Methodology, Software, Visualization. **Viera Kresnakova:** Software, Visualization. **Andrej Iring:** Software, Visualization. **Timotej Palus:** Data curation, Formal analysis, Methodology, Software. **Jakub Bahyl:** Data curation, Software, Visualization. **Olivier Nelis:** Data curation. **Monika Beles:** Data curation. **Davide Fabbriatore:** Validation. **Leor Perl:** Supervision, Validation. **Jozef Bartunek:** Conceptualization, Supervision, Validation, Writing – review & editing. **Robert Hatala:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

Dr. Herman is the Co-founder and Chief Medical Officer of Powerful Medical and supported by a research grant from the CardioPaTh PhD Program. Michal Martonak, Jakub Bahyl, Andrej Iring, Vladimir Boza and Anthony Demolder are employees of Powerful Medical. Other authors report no conflict of interest.

Acknowledgements

The authors would like to express the appreciation to the clinical experts, study team, data scientists and AI engineers supporting the data collection, processing, and validation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jelectrocard.2023.12.009>.

References

- [1] Cook DA, Oh SY, Pusic MV. Accuracy of physicians' electrocardiogram interpretations: a systematic review and meta-analysis. *JAMA Intern Med* Nov 1 2020;180(11):1461–71. <https://doi.org/10.1001/jamainternmed.2020.3989>.
- [2] Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* Dec 1 2016;176(12):1860–1. <https://doi.org/10.1001/jamainternmed.2016.6001>.
- [3] Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *J Electrocardiol* Sep-Oct 2007;40(5):385–90. <https://doi.org/10.1016/j.jelectrocard.2007.03.008>.
- [4] Guglin ME, Thatai D. Common errors in computer electrocardiogram interpretation. *Int J Cardiol* Jan 13 2006;106(2):232–7. <https://doi.org/10.1016/j.ijcard.2005.02.007>.
- [5] Kadish AH, et al. ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography: a report of the ACC/AHA/ACP-ASIM task force on clinical competence (ACC/AHA Committee to develop a clinical competence statement on electrocardiography and ambulatory electrocardiography) endorsed by the International Society for Holter and noninvasive electrocardiology. *Circulation* Dec 18 2001;104(25):3169–78 [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/11748119>.
- [6] Martinez-Losas P, et al. The influence of computerized interpretation of an electrocardiogram reading. *Am J Emerg Med* Oct 2016;34(10):2031–2. <https://doi.org/10.1016/j.ajem.2016.07.029>.
- [7] Novotny T, et al. The role of computerized diagnostic proposals in the interpretation of the 12-lead electrocardiogram by cardiology and non-cardiology fellows. *Int J Med Inform* May 2017;101:85–92. <https://doi.org/10.1016/j.ijmedinf.2017.02.007>.
- [8] Bogun F, et al. Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med* Nov 1 2004;117(9):636–42. <https://doi.org/10.1016/j.amjmed.2004.06.024>.
- [9] Anh D, Krishnan S, Bogun F. "Accuracy of electrocardiogram interpretation by cardiologists in the setting of incorrect computer analysis," (in eng). *J Electrocardiol* Jul 2006;39(3):343–5. <https://doi.org/10.1016/j.jelectrocard.2006.02.002>.
- [10] Ribeiro AH, et al. Author correction: automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* May 1 2020;11(1):2227. <https://doi.org/10.1038/s41467-020-16172-1>.
- [11] Hannun AY, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med* Jan 2019;25(1):65–9. <https://doi.org/10.1038/s41591-018-0268-3>.
- [12] Zhu H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *Lancet Digit Health* Jul 2020;2(7):e348–57. [https://doi.org/10.1016/S2589-7500\(20\)30107-2](https://doi.org/10.1016/S2589-7500(20)30107-2).
- [13] Hughes JW, et al. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA Cardiol* Nov 1 2021;6(11):1285–95. <https://doi.org/10.1001/jamacardio.2021.2746>.
- [14] Medical electrical equipment – Part 2–25: Particular requirements for the basic safety and essential performance of electrocardiographs. ANSI/AAMI/IEC 60601–2-25:2011(R)2016; Medical electrical equipment – Part 2–25: Particular requirements for the basic safety and essential performance of electrocardiographs. 2023.
- [15] Willems JL, et al. A reference data base for multilead electrocardiographic computer measurement programs. *J Am Coll Cardiol* Dec 1987;10(6):1313–21. [https://doi.org/10.1016/s0735-1097\(87\)80136-5](https://doi.org/10.1016/s0735-1097(87)80136-5).
- [16] Wallis S. "Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods," (in English). *J Quant Linguist* Aug 1 2013;20(3):178–208. <https://doi.org/10.1080/09296174.2013.799918>.
- [17] Breen CJ, Kelly GP, Kernohan WG. ECG interpretation skill acquisition: a review of learning, teaching and assessment. *J Electrocardiol* Jul-Aug 2022;73:125–8. <https://doi.org/10.1016/j.jelectrocard.2019.03.010>.
- [18] Townsend N, et al. Epidemiology of cardiovascular disease in Europe. *Nat Rev Cardiol* Feb 2022;19(2):133–43. <https://doi.org/10.1038/s41569-021-00607-3>.