Review Article

Machine learning models for automated interpretation of 12-lead electrocardiographic signals: a narrative review of techniques, challenges, achievements and clinical relevance

Panteleimon Pantelidis[1,2]^, Maria Bampa[1]^, Evangelos Oikonomou[2]^, Panagiotis Papapetrou[1]^

[1]Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden; [2]3rd Department of Cardiology, Thoracic Diseases General Hospital "Sotiria", National and Kapodistrian University of Athens, Athens, Greece

Contributions: (I) Conception and design: P Pantelidis; (II) Administrative support: M Bampa, P Papapetrou, E Oikonomou; (III) Provision of study material or patients: P Pantelidis; (IV) Collection and assembly of data: P Pantelidis, M Bampa; (V) Data analysis and interpretation: P Pantelidis, M Bampa; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Panteleimon Pantelidis, MD, MSc. Department of Computer and Systems Sciences, Stockholm University, Borgarfjordsgatan 12, Kista, Stockholm, Sweden. Email: pan.g.pantelidis@gmail.com.

Background and Objective: Novel advances in machine learning (ML) and its subfield, deep learning (DL), as well as the recent release of large-scale electrocardiogram (ECG) databases, have driven a sharp increase in research related to automated ECG interpretation. This review aims to summarize the recent ML approaches for automatically interpreting standard 12-lead ECG signals.

Methods: We searched 10 indexing databases, for original research in English, referring to the application of ML/DL techniques in 12-lead, raw ECG signal analysis. The retrieved titles were filtered based on their relevance. The results were summarized and reported.

Key Content and Findings: More than 80% of studies integrated a DL approach, while fewer attempts applied a feature extraction method to obtain inputs for training a simple ML classifier. The average diagnostic accuracy was as high as 90%, while several other performance metrics, such as the area under the curve (AUC), F1-score, sensitivity and specificity, were also employed. DL models generally demanded 10-time more samples for training but were capable of better handling multi-class problems. The most frequently involved disease (49% of studies) was myocardial infarction (MI), while atrial fibrillation (AF) was encountered in more than one-third of studies. Various datasets were used for training and testing, constituting either private collections or publicly available databanks [such as the "Physikalisch-Technische Bundesanstalt" (PTB) dataset and datasets derived from the "China Physiological Signal Challenge" and the "Computing in Cardiology Challenge"]. Overall, DL and simpler ML approaches for automated ECG interpretation display unprecedented growth, reaching remarkably high performances.

Conclusions: While such novel tools can support clinicians in reaching reliable diagnoses for life-threatening conditions on the spot, concerns regarding their accountability do exist. Generalizability of the developed approaches is still an issue, possibly mitigable with the extensive deployment of developed models, so as to become massively accessible and validatable. Finally, the observed heterogeneity of the various attempts underlines the need for transparency and reproducibility in the development processes.

Keywords: ECG; electrocardiogram; machine learning (ML); deep learning (DL)

^ ORCID: Panteleimon Pantelidis, 0000-0001-5394-832X; Maria Bampa, 0000-0003-0927-8239; Evangelos Oikonomou, 0000-0001-8079-0599; Panagiotis Papapetrou, 0000-0002-4632-4815.

Received: 10 November 2022; Accepted: 08 May 2023; Published online: 30 May 2023.
doi: 10.21037/jmai-22-94
View this article at: https://dx.doi.org/10.21037/jmai-22-94

Introduction

Since its conception in 1872 by Muirhead, and its formal establishment in 1901 by Einthoven, electrocardiogram (ECG) has been considered a cornerstone of the diagnostic armamentarium in cardiology (1,2). The main concept lies in applying several electrodes onto the patient's skin to capture the voltage signals produced by heartbeat cycles (normal or abnormal), from different angles. The conventional clinical setting involves the application of 10 electrodes, whose inputs are combined to produce a 12-lead electrical signal, namely a collection of 12 time-series signals of voltage (3). These 12 signals represent 12 different "looking angles", six on the coronal plane ("limb" leads) and six on the transverse one ("precordial" leads). To employ ECG as a diagnostic tool, clinicians and researchers are trained to identify the ECG features distinguishably present in each abnormal heart condition. This has led to a rule-based system for identifying ECG compartments (e.g., P wave, QRS complex, ST segment) and their morphological characteristics (e.g., duration, amplitude, shape) and relating them to distinct clinical entities in order to reach a diagnosis. Back in the 1950s, the first attempts to automatically interpret ECGs were based on the same idea (4). Rule-based algorithms were used to delineate ECG signals, detect and extract their components (e.g., the QRS morphology, QRS duration, PR duration) and relate the findings to clinical diseases, according to predefined rules, based on domain knowledge, thus simulating a human examiner (5). However, up until now, such methods have received considerable criticism due to their suboptimal predictive accuracy that might mislead inexperienced clinicians into false conclusions (4).

Today, in the era of machine (ML) and deep learning (DL), new approaches arose as potential solutions to this challenge (6,7). ML and DL techniques have a long history in medical practice. The application of these techniques in medicine began in the 1970s, where statistical and probabilistic models were used to analyze medical data (8). However, it was not until the advent of computer technology and increased computing power that ML and DL began to be used more widely in medical practice. One of the earliest applications of ML in medical practice was in the field of radiology. In the 1980s, researchers began to use ML algorithms to analyze medical images, such as X-rays and computed tomography (CT) scans (9). These algorithms were able to identify patterns and anomalies in the images that were not visible to the human eye, helping radiologists to make more accurate diagnoses. Since then, ML and DL have been used in various medical fields, including oncology, cardiology, neurology, and genomics. DL has made significant advances in medical practice in recent years. DL techniques have been used to develop predictive models for patient outcomes, disease diagnosis, and drug discovery (10).

With respect to automated ECG analysis, the most basic ML approach uses algorithms to extract more abstract features from ECG data, often not interpretable by the human eye. After applying some dimensionality reduction technique, such as Discrete Fourier Transformation (DFT) or Discrete Wavelet Transformation (DWT), these features (for example, frequency coefficients of a frequency spectrum, derived by DFT, or wavelet coefficients produced by DWT) are then "fed" to a simple ML classifier, such as a k-Nearest Neighbor (kNN) or a support vector machine (SVM) model to establish a relationship used for classification (11). The most recent approach, representing an "end-to-end" solution, is the application of DL with neural networks (NN) (12,13). In this case, raw ECG data are roughly "dumped" to an NN that "learns" the complex parameters mediating the relationship between data points and the associated different clinical entities (labels). Various NN architectures with different components can be applied for this task. For example, long-short term memory (LSTM) structures, which re-use previously seen information during later processing, and residual NN layers that preserve raw information by enabling it to "bypass" intermediate analysis steps, are useful techniques for that purpose. On top of that, many approaches often involve extensive data preprocessing and frequently employ combinations of the aforementioned techniques, usually by applying a feature extraction method, and then serving the extracted features as inputs to an NN model.

Rationale and objective

Considering the remarkable variations among potential

models, including simpler ML techniques (such as SVM, kNN) and purely DL models with various configurations, the high heterogeneity of different approaches becomes evident. Additionally, the orientation of the task at hand can vary significantly. ECG interpretation models can be utilized for many purposes and in many different settings, ranging from binary classification of specific entities [for example, atrial fibrillation (AF) against normal] using few-lead, long-duration settings (such as Holter devices), up to multi-label classification into several diseases, from the standard 12-lead ECG setting (14-16). This great abundance of different models and focuses, along with the recent release of several large-scale, high-quality, annotated ECG datasets (17-20), has led to an explosive growth in relevant projects, which, in turn, generates the need for summarizing them, assessing the current state in the field, and fostering a solid ground for future research. Aligning with this need, this narrative review constitutes a comprehensive summary of advances in the field of ML/DL-enabled automated ECG interpretation, focusing on the clinically relevant, standard 12-lead ECG setting, and providing evidence for the predictive performance of the various approaches. We present this article in accordance with the Narrative Review reporting checklist (available at https://jmai.amegroups.com/article/view/10.21037/jmai-22-94/rc).

Methods

Eligibility criteria

We included titles relevant to automated interpretation of the ECG, using ML or DL techniques. Not prototype studies or studies not in English were excluded. Additionally, studies using fewer than 12 leads as initial input (for example, 3-lead Holter signals) were also excluded. We followed this approach to narrow down the spectrum of included settings to only the 12-lead standard ones and avoid heterogeneity in application fields. Studies that used initial inputs other than raw ECG signals (for example, heartbeat features such as RR interval) were not included. Similarly, models that produced classification outputs (labels) other than clinical entities (for example, heartbeat features such as PR prolongation) or performed the binary classification "normal/abnormal" were excluded to ensure the clinical usability of the studies included. Studies that did not report metrics for their predictive performance were also excluded.

Search strategy and study selection

We searched PubMed, Scopus, Google Scholar, IEEE Xplore Digital Library, Microsoft Academic, dblp—Computer Science Bibliography, ACM Digital Library, arXiv, medRxiv and bioRxiv indexing databases for relevant titles, from January 2016 up to January 2021 (*Table 1*). To implement the aforementioned inclusion criteria, an appropriate set of search terms was used, depending on each specific database. The following search query is an indicative example for PubMed: ("ECG" OR "electrocardiogram") AND (("pattern" OR "signal" OR "automatic") AND ("recognition" OR "detection" OR "prediction" OR "analysis" OR "reading" OR "diagnosis")) AND ("algorithm" OR "machine learning" OR "ML" OR "deep learning" OR "DL" OR "neural networks" OR "CNN" OR "RNN" OR "time series" OR "series analysis"). After screening and filtering the retrieved titles, based on content eligibility, paradigmatic studies were shortlisted and presented.

Data extracted

The algorithm type employed, its configuration/architecture, the dataset(/s) used, the number and type of labels involved and the performance metrics applied, were the main features extracted. Moreover, the dataset size, as well as the subset percentage allocated for training, were documented. The performance measures are reported for testing on unseen external data or on a hold-out fold, and if none of these approaches were followed, then they represent the average cross-validation scores. For studies applying a two-step classification model [for example, myocardial infarction (MI) *vs.* normal, followed by the type of myocardial infarction], the performance measure is reported for the final step. Further information about the ECG databanks, such as the recording frequency and the length of their waveforms, was obtained. The availability of the source code and the data used is also reported.

Algorithm types

The majority of studies (35 studies, 81%) employed various DL techniques, either as complete "end-to-end" solutions, or in conjunction with other approaches, such as feature extraction techniques (*Table 2*). Although falling into the same algorithmic category, these studies display great heterogeneity in the architecture of their NN

Table 1 The search strategy summary

Items	Specification
Date of search	January 2021
Databases and other sources searched	PubMed, Scopus, Google Scholar, IEEE Xplore Digital Library, Microsoft Academic, dblp—Computer Science Bibliography, ACM Digital Library, arXiv, medRxiv and bioRxiv
Search terms used	Terms related to ECG: ("ECG" OR "electrocardiogram"), pattern recognition: (("pattern" OR "signal" OR "automatic") AND ("recognition" OR "detection" OR "prediction" OR "analysis" OR "reading" OR "diagnosis")), and ML/DL algorithms and models: ("algorithm" OR "machine learning" OR "ML" OR "deep learning" OR "DL" OR "neural networks" OR "CNN" OR "RNN" OR "time series" OR "series analysis")
Timeframe	From 2016 to 2021
Inclusion and exclusion criteria	Prototype studies in English, relevant to automated interpretation of the ECG, using ML or DL techniques. Studies using fewer than 12 leads as initial input were excluded. Studies that used initial inputs other than raw ECG signals, as well as those with models producing classification outputs (labels) other than clinical entities or performing binary classification into "normal/abnormal", were also excluded. Studies not reporting performance metrics were not included
Selection process	Two authors independently screened and filtered the retrieved titles, based on content eligibility. In case of discrepancies, a third author was involved and the decision was reached by majority voting

ECG, electrocardiogram; ML, machine learning; DL, deep learning; CNN, convolutional neural network; RNN, recurrent neural network.

(Table S1). Both convolutional and recurrent NN (CNN, RNN; respectively) are widely used, with CNN present in 28 studies (80%) and RNN in 13 (37%). A usual pattern suggests RNN for parallel "temporal" analysis (across each lead) and CNN for "spatial" analysis (aggregating outputs from all previously analyzed leads), although CNN can be used for "temporal" analysis as well (21-28). Gated recurrent units (GRU) and LSTM layers were shown to display superior performance, against the vanilla version of RNN (29). Quite often is the use of bidirectional RNN, GRU and LSTM structures (9 studies, 26%), in order to retain information in both directions (back and forth) during processing (25,27,30-36). The ReLu activation function is typically employed after each convolution in hidden layers, while Softmax (multi-label classification) and Sigmoid (binary classification and multi-label classification with multiple labels per entry) activation functions are used in the final layer. Pooling (average, max or custom) and batch normalization layers are widely applied several times in intermediate steps, to reduce the dimensionality, improve performance and speed-up the whole training process. Residual blocks, with skip connections, are found in 12 studies (34%) in order to preserve the information intact, in parallel with its processing through other layers, before concatenating the two outputs (22,26,37-46). Attention layers and inception blocks are also encountered. Attention mechanisms are quite often (10 studies, 22%) as a means for focusing on important intermediate inputs and improving performance (23,25,26,30-32,36,37,47,48). On the contrary, inception is more rarely applied (3 studies, 7%) in order to reduce overfitting and computational cost (24,44,48). The exclusive use of more traditional ML techniques is also found in some studies. These often involve a technique for abstract feature extraction, most usually DFT, DWT or a similar variation, and then "feed" the extracted features to train a simple ML classifier, such as kNN or SVM (49-53). Another frequent pattern is the combinatorial employment of both feature extraction techniques (usually a DWT or a DFT approach) and the use of these extracted features for training an NN (37,43,45,54-57). Data preprocessing techniques are also frequently encountered, with noise reduction and signal normalization commonly applied. Moreover, many studies prefer not to use the whole, repetitive ECG signals, but slice the long ones to achieve equal lengths, or segment them into separate beats and select representative ones. A detailed presentation of different algorithm architectures for each study, along with relevant comments, is provided in Table S1.

Datasets

As shown in *Table 2*, 12 (28%) projects use an internal dataset for training and testing, usually derived from the clinical environment of the researchers. The remaining

Table 2 Summary of study characteristics

Study	Technique	Dataset	Total number of samples	Labels	Performance metric[†]
Adedinsewo *et al.*, 2020	DL	Int	N/A	LVSD, no-LVSD	0.859
Attia, Kapa, *et al.*, 2019	DL	Int	97,829	LVSD, no-LVSD	0.857
Attia, Noseworthy, *et al.*, 2019	DL	Int	649,931	AF, non-AF (future onset)	0.794
Cai *et al.*, 2020	DL	Int, CPSC2018, MLws	16,557	AF, normal	0.994
Chen *et al.*, 2020	DL	CPSC2018	9,831	AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, normal	0.97
Darmawahyuni *et al.*, 2019	DL	PTB	12,359	MI, normal	0.976
Feng *et al.*, 2019	Mi	CCDD	620	SRa, ST, PAC, PVC, RBBB, LVHV, STc, Tc, Lad, normal	0.742
Fu *et al.*, 2020	DL	PTB	760,128	AMI, ALMI, ASMI, IMI, ILMI	0.629
Han & Shi, 2019	sML	PTB	33,586	MI, non-MI	0.927
Jafarian *et al.*, 2020	DL	PTB	5,968	AMI, ALMI, ASMI, IMI, ILMI, IPLMI, normal	1
Jia *et al.*, 2019	DL	1stChiECG	13,500	AF, I-AVB, RBBB, LAFB, PVC, PAC, ER, Tc, normal	0.872 (f1)
Jo *et al.*, 2021	DL	Int, PTB-XL, ChapECG, CinC2017	169,369	AF, non-AF	0.993
Khawaja, 2018	n.a.	Int	n.a.	Hypertrophy, MI, normal	0.746
Li *et al.*, 2020	DL	Int	7,000	AF, I-AVB, RBBB, LAFB, PVC, PAC, ER, Tc, normal	0.928
Liu J *et al.*, 2019	sML	PTB	104	MI, normal	0.816
Liu W *et al.*, 2020	DL	PTB	64,350	AMI, ASMI, ALMI, IMI, ILMI, normal	0.931
Lu *et al.*, 2020	DL	CCDD	143,092	SVT, VT, MI, LVH, LAtH, normal	0.878
Makimoto *et al.*, 2020	DL	PTB	289	MI, non-MI	0.81
Megahed *et al.*, 2019	DL	PTB	549	AMI, ALMI, ASMI, ASLMI, IMI, ILMI, IPLMI, IPMI, LMI, PMI, PLMI, normal	0.993 (f1)
Mostayed *et al.*, 2018	DL	CPSC2018	107,414	AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, normal	0.739 (f1)
Oh *et al.*, 2017	sML	PTB, INCART	16,099	CAD, non-CAD	0.997
Oppelt *et al.*, 2020	Mi	CinC2020	49,539	CinC2020 classes	0.724
Padhy & Dandapat, 2017	sML	PTB	9,946	AMI, ALMI, ASMI, IMI, ILMI, normal	0.981
Prabhakararao & Dandapat, 2020	DL	PTB, STAFFIII	14,260	EMI, AcMI, CMI, non-MI, normal	0.978
Raghunath *et al.*, 2020	DL	Int	1,151,037	AF, non-AF (future onset)	0.83 (auc)
Ribeiro *et al.*, 2020	DL	Int	2,322,513	I-AVB, RBBB, LBBB, SB, AF, ST	0.926 (f1)
Sigurthorsdottir *et al.*, 2020	DL	CinC2020	43,135	CinC2020 classes	0.573
Strodthoff & Strodthoff, 2019	DL	PTB	53,489	AMI, IMI, normal	0.933 (sen)
Tison *et al.*, 2019	Mi	PTB, Int	36,186	PAH, HyC, CA, MVP, normal	0.87 (auc)

Table 2 (*continued*)

Table 2 (*continued*)

Study	Technique	Dataset	Total number of samples	Labels	Performance metric[†]
Tripathy & Dandapat, 2016	sML	PTB	68	BBB, MI, HMD, normal	0.861
Tripathy & Dandapat, 2017	sML	PTB	68	BBB, MI, HMD, normal	0.984
Tripathy *et al.*, 2019a	Mi	PTB	17,100	AMI, ALMI, IMI, ILMI, IPLMI, normal	0.998
Tripathy *et al.*, 2019b	Mi	PTB	174	MI, non-MI	0.997
Wang C *et al.*, 2019	Mi	1stChiECG	14,000	AF, I-AVB, RBBB, LAFB, PVC, PAC, ER, Tc, normal	0.863 (f1)
Wang HM *et al.*, 2019	DL	PTB	61,074	AMI, IMI, normal	0.953 (auc)
Yao *et al.*, 2020	DL	CPSC2018	9,831	AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, normal	0.852
Yuan & Xing, 2019	Mi	1stChiECG	14,000	AF, I-AVB, RBBB, LAFB, PVC, PAC, ER, Tc, normal	0.879 (f1)
Zhang X *et al.*, 2019	DL	PTB	54,753	AMI, ALMI, ASMI, IMI, ILMI, normal	0.998
Zhang X *et al.*, 2020	DL	Int	277,807	PAC, AF, Afl, PVC, asystole, oMI, MI, LVHV, STc, hyperK, Tc, LVH, I-AVB, II-AVB, LBBB, RBBB, VpES, normal	0.95
Zhang D *et al.*, 2021	DL	CPSC2018	9,831	AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, normal	0.966
Zhang J *et al.*, 2020	DL	CPSC2018	9,831	AF, I-AVB, LBBB, RBBB, PAC, PVC, STD, STE, normal	0.868
Zhang J *et al.*, 2021	sML	PTB	60,766	AMI, ALMI, ASMI, ASLMI, IMI, ILMI, IPLMI, IPMI, LMI, PMI, PLMI, normal	0.994
Zhu *et al.*, 2020	DL	Int	180,112	Several including I-AVB, II-AVB, VT, PAC, AF, WPW, normal	0.983 (auc)

[†], represents accuracy for most studies, except when otherwise indicated, according to the following notion: (f1), F1-score; (auc), area under the curve; (sen), sensitivity (for multi-label classification models the average values are provided). DL, deep learning; sML, simple (non-DL) machine learning; Mi, Mixed methods (more details on the algorithms used in each approach are provided in Table S1); Int, internal database (private collection); CPSC2018, China Physiological Signal Challenge 2018; MLws, Mason-Likar wearable ECG 12-lead system; PTB, Physikalisch-Technische Bundesanstalt ECG database; CCDD, Chinese Cardiovascular Disease Database; 1stChiECG, First China ECG Intelligent Competition; PTB-XL, new release of PTB ECG database; ChapECG; Chapman ECG database; CinC2020/2017, PhysioNet/Computing in Cardiology 2020 or 2017; INCART, St. Petersburg INCART 12-lead Arrhythmia Database; STAFIII, STAFF III Database in PhysioNet; N/A, not applicable; n.a., not available; LVSD, left ventricular systolic dysfunction; AF, atrial fibrillation; I-AVB, 1st degree atrioventricular block; LBBB, left BBB; RBBB, right BBB; BBB, bundle branch block; PAC, premature atrial contraction; PVC, premature ventricular contraction; STD, ST-segment depression; STE, ST-segment elevation; MI, myocardial infarction; SRa, sinus-rhythm arrhythmia; ST, sinus tachycardia; LVHV, left ventricle high voltage; STc, ST-segment change; Tc, T-wave change; Lad, left axis deviation; AMI, anterior MI; ALMI, anterolateral MI; ASMI, anteroseptal MI; IMI, inferior MI; ILMI, inferolateral MI; IPLMI, inferoposterolateral MI; LAFB, left anterior fascicular block; ER, early repolarization; SVT, supraventricular tachycardia; VT, ventricular tachycardia; LVH, left ventricular hypertrophy; LAtH, left atrial hypertrophy; IPMI, inferoposterior MI; PMI, posterior MI; PLMI, posterolateral MI; CAD, coronary artery disease; EMI, early progression of MI; AcMI, acute MI; CMI, chronic MI; SB, sinus bradycardia; PAH, pulmonary arterial hypertension; HyC, hypertrophic cardiomyopathy; CA, cardiac amyloidosis; MVP, mitral valve prolapse; HMD, heart muscle defect; Afl, atrial flutter; oMI, old MI; hyperK, hyperkalemia; II-AVB, 2nd degree atrioventricular block; VpES, ventricular pre-excitation syndrome; WPW, Wolf-Parkinson-White syndrome.

Table 3 Most frequently used datasets

Dataset	Year of release	Size	Length (s)	Frequency (Hz)	Label size	Comments and availability
PTB Diagnostic ECG Database	2004	549	~115	1,000	9 (with additional specific details for each case)	http://www.physionet.org/content/ptbdb/1.0.0/
China Physiological Signal Challenge 2018	2018	9,831	6 to 60	500	9	https://2018.icbeb.org/Challenge.html
The First China ECG Smart Competition	2019	14,000	9 to 90	500	9	Refers to the "Rematch" phase; Questionable availability (not included on the webpage); http://mdi.ids.tsinghua.edu.cn/
China Cardiovascular Disease Database	2010	179,130		500	378	Availability upon request; http://www.ecgdb.com/
Computing in Cardiology Challenge 2020	2020	20,674	6 to 60/10	500	111 disease labels (according to SNOMED CT)	Includes 5 sub-datasets, from which only 2 are included here (CinC201, Georgia Challenge); https://physionetchallenges.org/2020/

Size refers to the number of ECG records included. Length refers to each ECG record. SNOMED Clinical Terms (https://www.snomed.org/). PTB, Physikalisch-Technische Bundesanstalt; ECG, electrocardiogram; SNOMED CT, Systematized Nomenclature of Medicine Clinical Terms.

studies employ publicly available datasets, while a few use datasets offered in the context of relevant contests (25,26,30,32,35,37,39,41,43,45,47,58). Five (12%) projects employ combinations of two (23,55,59), three (58) or four (41) datasets, by using separate datasets for development and testing, or by mixing them and then splitting them from scratch for the different building stages. The "Physikalisch-Technische Bundesanstalt" (PTB) ECG dataset (60), present in 21 studies (49%), is the most widely used source of data. Its current version was released in 2004, consisting of 549 ECGs by 290 different individuals, belonging to nine different classes. The second most frequent databank is the "China Physiological Signal Challenge 2018" dataset, adopted by six studies (14%). The "First China ECG Smart Competition" dataset, the "China Cardiovascular Disease Database" (CCDD) and the data from "Computing in Cardiology/Physionet Challenge 2020" follow, with two associated projects each. More details on the most frequent data sources are provided in *Table 3*. Finally, some datasets, such as the "Chapman ECG" dataset, the "St. Petersburg INCART 12-lead Arrhythmia Database" and the PTB-XL database are exploited by only one study each. PTB-XL, released in April 2020, is the most recent, large, publicly available ECG dataset, with more than 21,000 ECG samples of 10 seconds length, obtained from 18,885 individuals. The ECG samples are classified roughly into five super-classes (MI, ST-/T-change, conduction disturbance, hypertrophy and normal) with many secondary subcategories, accompanied by a rich collection of demographic and clinical information. A usual approach is to segment the ECG lines into several frames or even single heartbeats, so that a large number of datapoints is generated. The most frequent algorithm for QRS detection and beat separation is the Pan-Tompkins algorithm (61). The average dataset size across all studies is 158,490, with an average size for the training set equal to 134,053. Models based on DL or mixed techniques (DL combined with other methods) demanded more subjects for training (158,577 on average) than simpler ML models (an average of 14,940 samples).

Labels

There are great variations among studies concerning the clinical entities used as labels (Table S2). The most widely included pathology is the MI (62,63), encountered in 21 (49%) of the studies. In nine of them, the model also focuses on the localization of the infarction (inflicted area). AF is the second most popular disease, included in 16 (37%) of the studies. Moreover, while most studies attempt to identify the ongoing heart condition from the ECG, a few studies claim to provide predictive value for future onset of AF (22,24) or other conditions (21). Different combinations of these clinical entities are used in each study (often underlined by the employed, already annotated, dataset), with an average of seven different disease classes per study. Detailed information on the included class labels is provided

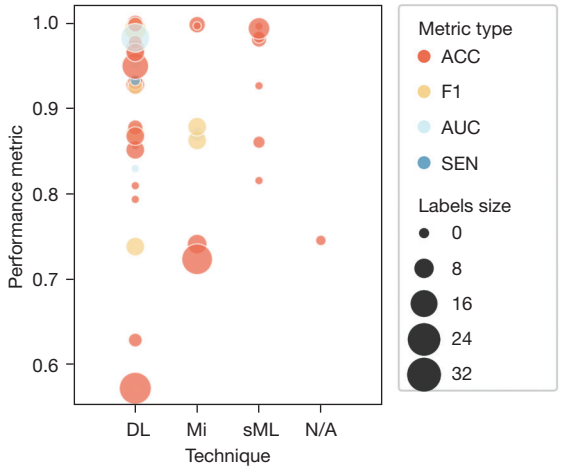

Figure 1 Algorithm types (x-axis) along with the most frequently used performance metrics (y-axis for metric value, dot color for metric type) and label sizes (dot size). ACC, accuracy; F1, F1-score; AUC, area under the curve; SEN, sensitivity; DL, Exclusively Deep Learning methods; Mi, Mixed methods; sML, Simpler Machine Learning methods; N/A, not available.

in *Table 2*.

Performance metrics

In the present review, only studies providing metrics for their predictive performance were included. Most of them apply several different measures, such as accuracy, area under the curve (AUC), F1-score, sensitivity, specificity, positive and negative predictive value (PPV and NPV; respectively), while two studies adopt custom metrics (25,37). The most widely used performance metric is accuracy, namely the percentage of correct to overall attempted diagnoses, present in 32 (74%) of the studies. The average, unweighted accuracy is 89.3% [standard deviation (SD) =11.5%], although performance can vary significantly with respect to several factors, such as the algorithm type, the algorithm configuration, the train/test sets and the label size and balance (*Figure 1*). While the mean specificity, representing the ability of the model to correctly distinguish normal samples, was 91.2%, the mean sensitivity, namely the model's performance in identifying positive cases, was fairly lower at 89.3%. The mean AUC was 91.6%, while average F1-score was 84%. Finally, the mean NPV was found higher than the average PPV (94% and 85.7%, respectively). A detailed report of all the performance measures for each study can be found in Table S3. Apart from providing evidence for their predictive performance, four studies took a step forward and compared the performance of their algorithms against that of benchmark diagnostic techniques or human evaluators, concluding that the automated algorithms can outperform the existing, traditional diagnostic approaches (38,46,64,65).

Additional information

Five out of the 43 studies were available only as not peer-reviewed preprints, indexed in arXiv, medRxiv and bioRxiv (24,25,35,37,39). Most studies do not provide instructions for accessing the source code, with only 8 (19%) offering explicit directions (21,24,30,35,38,39,46,55). Of the overall 12 studies using non-publicly available datasets, only 4 of them (33%) provided instructions for accessing the source data (21,24,38,46). Finally, according to the 16 studies (37%) offering information on the hardware infrastructure used, multi-core central (CPU) and graphic (GPU) processing units are used, with random access memory (RAM) capacity ranging from 8 to 224 Gigabytes (22,24,26,27,29,31-33,36,40,48,49,51,54,58,65).

Discussion

Summary of evidence and insights

The explosive growth in research is evident, even when restricting the included models by applying narrow eligibility criteria (aforementioned criteria, including 12-lead only models, raw ECG as inputs, clinical entities as output and other). Relevant studies display significant heterogeneity in many aspects. At first, several different performance metrics are used across studies, hampering a comparative analysis at a certain extent. Moreover, there are numerous different datasets employed, comprising either publicly available databanks or private data collections, which, in turn, also determine the associated classification labels. Most importantly, different projects vary greatly with regard to the algorithmic types they use. While 28 studies (65%) employ a strictly DL approach, 7 studies (16%) follow more traditional approaches with simple ML classifiers based on extracted features as inputs. Finally, another 7 works (16%) employ mixed techniques (feature extraction and NN), while 1 study does not report the underlying algorithm in detail.

This trend toward DL approaches can plausibly be attributed to the recent release of numerous publicly

available ECG databanks, leading the way for developing trainable, "data-hungry" DL models. From the MIT-BIH database (two-lead signals by 47 subjects) published in 2005 (66) and the PTB database (549 12-lead records from 290 subjects) released in 2004 (60), developers can now exploit tens of thousands of annotated, complete 12-lead ECG lines, obtained by various contemporary databanks (17,18,20,67). Recent ECG analysis contests, such as the "PhysioNet/Computing in Cardiology", the "First China ECG Smart Competition" and the "China Physiological Challenge", have fostered new attempts and boosted related research. Moreover, modern advances in the DL field, such as attention layers, residual blocks, bidirectional LSTM/GRU and other structures, as well as the improvement of computer hardware, have also made feasible the analysis of large chunks of high-dimensional ECG data. Although deep NN usually demand considerably larger data sizes than simpler ML techniques, they can achieve remarkably high performances, even when dealing with multi-class problems (many clinical entities). On the contrary, simpler models are usually constrained to fewer classes in order to reach a satisfactory performance (e.g., binary classification between AF/non-AF). Indeed, strictly DL and mixed models use an average of 158,577 samples for training, while the same figure for simpler ML techniques (no NN) is 14,940 (these values were obtained by multiplying the whole dataset size with the percentage used for training, after any segmentation of ECG lines into frames/heartbeats had already taken place during pre-processing). However, the former models are able to "ingest" many more labels (7.7 on average), while the latter usually support only binary or few-label classification (4.6 labels on average). This intuitive tradeoff among algorithm type, dataset size, label size and predictive performance can be seen in *Figure 1*. Subsequently, DL approaches seem to prevail, despite their computational and data demands, partially due to the current existence of abundant data (public databases, digitization and automated storage of ECGs by novel ECG machines), but also due to the ongoing increase in computational power.

A critical point not to be missed is the interpretation of model performance in the medical scene. Although a slight difference of 0.5% in sensitivity might seem trivial as a number, when applied to the diagnosis of patients with a severe condition, for example, a MI, it means that 5 persons out of 1,000 will go undiagnosed when suffering from a possibly fatal heart attack. Therefore, sensitivity, in particular, is a crucial measure that should not be overlooked, in the case of a reassuringly high accuracy, F1-score or other metric. On the other hand, sacrificing PPV for the sake of sensitivity, will unnecessarily impose an extra burden on the already heavy workload of the clinicians, as it might accumulate overdiagnosed, healthy individuals in the emergency rooms. This example demonstrates the practical significance of optimizing diagnostic algorithms when applying them in such situations. Achieving a golden balance, but also excellence in particular metrics, is of absolute importance, but still seems an unreached goal that cultivates doubt among clinicians. Consequently, clinicians seem to consult automated ECG analysis reports for reaching a diagnosis, but still not to trust it over human judgment (4,68).

Generalizability and transparency are major concerns here, as they generally are in ML/DL development projects (69). Although novel algorithms have achieved impressive performance records, even in multi-label problems, most of them are predominantly trained and tested in the same, common databases, and predominantly in publicly available ones (*Table 2*). This raises the issue of external validity when dealing with real data from the worldwide pool of patients. In essence, such models might display satisfactory performance within the study setting, but may prove incapable of predicting accurately external data, as they might differ from those used for training in various aspects that had not been accounted for. One of the main reasons of limited diagnostic accuracy and compromised generalizability is the quality of the ECG signals. Raw 12-lead signals can be affected by several factors such as muscle movement, electrode placement, and interference from other devices, which can affect the accuracy of the model (70). Additionally, the quality of the signal can be affected by noise and artifacts, which can result in incorrect feature extraction and subsequently lower performance (71). Furthermore, the lack of diversity in the dataset used for training ML models can impact their external validity, making it difficult to generalize the results to other patient populations (21). Finally, the choice of ML algorithm can also impact the accuracy. For instance, DL-based models have shown great potential in ECG analysis (72), but they require significant computational power and large datasets.

Careful consideration of these factors is necessary to design accurate and reliable ML models for automated ECG analysis of raw 12-lead signals in cardiac diagnosis. Additionally, another possible solution to the problem of low generalizability could be the massive deployment of

developed models through convenient means, e.g., mobile applications, so that they become accessible to clinicians worldwide. This could facilitate their validation by independent researchers, testing them on their own private collections of real ECG data. Furthermore, policies that compromise transparency should also be flagged. Only 33% of the projects using internal datasets, provided instructions on how to obtain the anonymized data, while only 19% of the overall studies offered guidance for accessing the source code. Preventing the specialized reader from examining and understanding the "inner workings" of the proposed models, hinders their further improvement and restricts scholars to comparing different approaches merely by their self-reported metrics, depriving them of the opportunity for an in-depth analysis. These facts indicate that transparency and reproducibility are problems still to be addressed.

Strengths and limitations

We assume that this review can add value to the existing knowledge since it summarizes the novel advances in ECG automated analysis models, focused on the narrow domain of 12-lead standard ECG practice. Its reproducibility is enhanced by explicitly reporting all the analysis steps, as well as by the adoption of the PRISMA standards. Insights on the application of different algorithmic approaches, along with their requirements and their benefits (trade-off among data size, predictive performance and label size "ingestion") are provided and justified with evidence. We deem that this review fosters a solid ground for future research, by attempting a golden balance between technical aspects useful for data analysts/developers, and clinical viewpoints, meaningful to healthcare experts. In this report, we decided to narrow down the spectrum of examined applications to the standard 12-lead ECG setting. While integrating more subfield orientations [for example, Holter signals with long durations but few leads (16), or even single-lead wearable devices destined for patient self-monitoring (14,15)] would possibly make the provided information more inclusive, it would result in a report of limited practical usability and applicability, as the outcomes would stem from highly heterogenous underlying settings. Despite any variations among the model structures, the studies included in this review apply to the same setting, enabling their outcomes to be directly comparable. Finally, by applying inclusive search terms and scanning a variety of databases, we esteem that our findings are as up-to-date as possible.

However, this work is better understood within the spectrum of its limitations. A general drawback of reviews in this field is that they quickly become outdated. Indeed, the ability to build such models with only a commodity computer and adequate data, as minimal requirements, has led to a remarkable bloom in related research. "New waves" of attempts are typically triggered with release of every new database. To handle this enormous amount of entries across "time" and "space", we restricted our search to studies published within the past 5 years, that focused specifically on 12-lead ECG models. In this way, we ensured that only contemporary approaches are included, we enhanced the clinical usability of the results and secured their longer "life-span", however, with the inevitable cost of possibly omitting neighboring subdomains, such as Holter signal analyses, that might offer additional insights. Nevertheless, comparative analyses of single- *vs.* multi-lead models have shown a slight, yet clinically important, difference in performance that favors the latter (30,58). Additionally, the considerable, multilevel heterogeneity in studies (among algorithms, datasets, metrics and labels) makes it difficult to construct a standardized way for grouping and comparing studies, not to mention identifying any superior approach. It seems that there is no "one-to-rule-them-all" solution, but many "try-and-error", domain-specific attempts. Furthermore, the presentation of the very fine details of every model's structure was impossible, leading to some loss of possibly significant information from the developer's point of view, especially regarding the deep NN architecture (for example, window size, stride and other aspects of every convolution layer). To compensate for that, we created an illustrated, semantic representation of the basic structure of each model, presented in Table S1, and we also provided full references to all related studies. Finally, we were not able to reproduce the workflow and results of each model, since very few studies provided access to their source code and data. Given this, the need for policies fostering reproducibility and transparency must be highlighted.

Conclusions

Despite having a long-standing presence since the 1950s, automated ECG analysis exhibits unprecedented growth nowadays. Modern analytic techniques (mainly employing DL approaches), newly released, publicly available datasets, and the improvements in computational power, may be considered as the driving factors of this progress. Several studies, experimenting with novel approaches in this field, report performance measures close to 100%. Although

encouraging, these outcomes should be interpreted with caution, always minding the underlying heterogeneity of relevant attempts, as well as any generalizability, transparency and reproducibility issues.

Overcoming these issues is still a challenge and a call for future research. Deployable models, for example through mobile or web applications, could not only sharply increase the usability of the various models, but also validate their high predictive value on real-life ECG data, obtained from the worldwide medical community. Room for improvement also exists in the number and type of clinical entities employed, as classification labels. In order to serve supportively, or even to substitute clinicians, in diagnosing heart conditions from ECG, automated models must not only provide high performance metrics for their tasks, but also adopt tasks of true value. This means that binary classification between individuals presenting with a specific disease (e.g., AF) and those who do not, particularly when this disease is easily detectable by the human eye, is of limited clinical value. On the contrary, models that can distinguish among a variety of different conditions, especially when they are typically hard to spot (e.g., ST-segment depression that might have various underlying causes), could be much more appreciated and utilized by clinicians. Fortunately, modern ECG databanks, with adequately large numbers of entries, and rich, nearly complete annotations, exist to support this aim and bring us one step closer to fully automated, reliable ECG interpretation.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://jmai.amegroups.com/article/view/10.21037/jmai-22-94/rc

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-22-94/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-22-94/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Birse RM. Muirhead, Alexander (1848–1920), Electrical Engineer. Oxford University Press, 2004.
2. Rivera-Ruiz M, Cajavilca C, Varon J. Einthoven's string galvanometer: the first electrocardiograph. Tex Heart Inst J 2008;35:174-8.
3. de Luna AB. ECGs for Beginners. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014.
4. Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms: Benefits and Limitations. J Am Coll Cardiol 2017;70:1183-92.
5. Fotiadis D, Likas A, Michalis L, et al. Electrocardiogram (ECG): Automated Diagnosis. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2006.
6. Lyon A, Mincholé A, Martínez JP, et al. Computational techniques for ECG analysis and interpretation in light of their contribution to medical advances. J R Soc Interface 2018;15:20170821.
7. Deepak HA, Vijayakumar T. Review of ECG Signal Classification Using Deep Learning and Traditional Methods. International Journal of Scientific & Technology Research 2020;9:5683-90.
8. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994;1:8-27.
9. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018;319:1317-8.
10. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.
11. Abreu-Lima C, de Sá JP. Automatic classifiers for the

interpretation of electrocardiograms. Rev Port Cardiol 1998;17:415-28.
12. Xu SS, Mak MW, Cheung CC. Towards End-to-End ECG Classification With Raw Signal Extraction and Deep Neural Networks. IEEE J Biomed Health Inform 2019;23:1574-84.
13. Parvaneh S, Rubin J, Babaeizadeh S, et al. Cardiac arrhythmia detection using deep learning: A review. J Electrocardiol 2019;57S:S70-4.
14. Ravanshad N, Rezaee-Dehsorkh H, Lotfi R, et al. A level-crossing based QRS-detection algorithm for wearable ECG sensors. IEEE J Biomed Health Inform 2014;18:183-92.
15. Sharma J, Kumar V, Ayub S, et al. Uniform Sampling of ECG Waveform of MIT-BIH Normal Sinus Rhythm Database at Desired Intervals. Int J Comput Appl 2012;50:6-9.
16. Kim H, Yazicioglu RF, Merken P, et al. ECG signal compression and classification algorithm with quad level vector for ECG holter system. IEEE Trans Inf Technol Biomed 2010;14:93-100.
17. Wagner P, Strodthoff N, Bousseljot RD, et al. PTB-XL, a large publicly available electrocardiography dataset. Sci Data 2020;7:154.
18. Zheng J, Zhang J, Danioko S, et al. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. Sci Data 2020;7:48.
19. Liu F, Liu C, Zhao L, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. J Med Imaging Health Inform 2018;8:1368-73.
20. Zhang J, Wang L, Liu X, et al. Chinese Cardiovascular Disease Database (CCDD) and Its Management Tool. 2010 IEEE International Conference on BioInformatics and BioEngineering. Philadelphia, PA, USA: IEEE, 2010:66-72.
21. Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70-4.
22. Attia ZI, Noseworthy PA, Lopez-Jimenez F, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019;394:861-7.
23. Prabhakararao E, Dandapat S. Myocardial Infarction Severity Stages Classification From ECG Signals Using Attentional Recurrent Neural Network. IEEE Sensors Journal 2020;20:8711-20.
24. Raghunath S, Pfeifer JM, Ulloa-Cerna AE, et al. Deep Neural Networks Can Predict Incident Atrial Fibrillation from the 12-Lead Electrocardiogram and May Help Prevent Associated Strokes. medRxiv 2020. doi: https://doi.org/10.1101/2020.04.23.20067967.
25. Sigurthorsdottir H, Van Zaen J, Delgado-Gonzalo R, et al. ECG Classification with a Convolutional Recurrent Neural Network. Computing in Cardiology 2020;47:3-6.
26. Yao Q, Wang R, Fan X, et al. Multi-Class Arrhythmia Detection from 12-Lead Varied-Length ECG Using Attention-Based Time-Incremental Convolutional Neural Network. Information Fusion 2020;53:174-82.
27. Zhang X, Li R, Dai H, et al. Localization of Myocardial Infarction with Multi-Lead Bidirectional Gated Recurrent Unit Neural Network. IEEE Access 2019;7:161152-66.
28. Khawaja A. A Novel Algorithm for Full-Automatic ECG Interpretation and Diagnostics. 2018 Computing in Cardiology Conference (CinC). Maastricht: IEEE, 2018.
29. Darmawahyuni A, Nurmaini S, Sukemi, et al. Deep Learning with a Recurrent Network Structure in the Sequence Modeling of Imbalanced Data for ECG-Rhythm Classifier. Algorithms 2019;12:118.
30. Chen TM, Huang CH, Shih ESC, et al. Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model. iScience 2020;23:100886.
31. Fu L, Lu B, Nie B, et al. Hybrid Network with Attention Mechanism for Detection and Location of Myocardial Infarction Based on 12-Lead Electrocardiogram Signals. Sensors (Basel) 2020;20:1020.
32. Zhang J, Liu A, Gao M, et al. ECG-based multi-class arrhythmia detection using spatio-temporal attention-based convolutional recurrent neural network. Artif Intell Med 2020;106:101856.
33. Liu W, Wang F, Huang Q, et al. MFB-CBRNN: A Hybrid Network for MI Detection Using 12-Lead ECGs. IEEE J Biomed Health Inform 2020;24:503-14.
34. Lu P, Xi H, Zhou B, et al. A New Multichannel Parallel Network Framework for the Special Structure of Multilead ECG. J Healthc Eng 2020;2020:8889483.
35. Mostayed A, Luo J, Shu X, et al. Classification of 12-Lead ECG Signals with Bi-directional LSTM Network. arXiv 2018. arXiv:1811.02090.
36. Zhang X, Gu K, Miao S, et al. Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system. Cardiovasc Diagn Ther 2020;10:227-35.

37. Oppelt MP, Riehl M, Kemeth FP, et al. Combining Scatter Transform and Deep Neural Networks for Multilabel Electrocardiogram Signal Classification. Computing in Cardiology 2020;47:1-4.
38. Ribeiro AH, Ribeiro MH, Paixão GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 2020;11:1760.
39. Zhang D, Yang S, Yuan X, et al. Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram. iScience 2021;24:102373.
40. Jafarian K, Vahdat V, Salehi S, et al. Automating Detection and Localization of Myocardial Infarction Using Shallow and End-to-End Deep Neural Networks. Applied Soft Computing 2020;93:106383.
41. Jo YY, Cho Y, Lee SY, et al. Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. Int J Cardiol 2021;328:104-10.
42. Strodthoff N, Strodthoff C. Detecting and interpreting myocardial infarction using fully convolutional neural networks. Physiol Meas 2019;40:015001.
43. Wang C, Yang S, Tang X, et al. A 12-Lead ECG Arrhythmia Classification Method Based on 1D Densely Connected CNN. In: Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. MLMECH CVII-STENT 2019. Cham: Springer, 2019:72-9.
44. Wang HM, Zhao W, Jia DY, et al. Myocardial Infarction Detection Based on Multi-lead Ensemble Neural Network. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:2614-7.
45. Yuan B, Xing W. Diagnosing Cardiac Abnormalities from 12-Lead Electrocardiograms Using Enhanced Deep Convolutional Neural Networks. In: Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. MLMECH CVII-STENT 2019. Cham: Springer, 2019:36-44.
46. Zhu H, Cheng C, Yin H, et al. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. Lancet Digit Health 2020;2:e348-57.
47. Jia D, Zhao W, Li Z, et al. An Ensemble Neural Network for Multi-label Classification of Electrocardiogram. In: Machine Learning and Medical Engineering for Cardiovascular Health and Intravascular Imaging and Computer Assisted Stenting. MLMECH CVII-STENT 2019. Cham: Springer, 2019:20-7.
48. Li D, Wu H, Zhao J, et al. Automatic Classification System of Arrhythmias Using 12-Lead ECGs with a Deep Neural Network Based on an Attention Mechanism. Symmetry 2020;12:1827.
49. Han C, Shi L. Automated interpretable detection of myocardial infarction fusing energy entropy and morphological features. Comput Methods Programs Biomed 2019;175:9-23.
50. Liu J, Zhang C, Ristaniemi T, et al. Detection of Myocardial Infarction from Multi-lead ECG using Dual-Q Tunable Q-Factor Wavelet Transform. Annu Int Conf IEEE Eng Med Biol Soc 2019;2019:1496-9.
51. Padhy S, Dandapat S. Third-Order Tensor Based Analysis of Multilead ECG for Classification of Myocardial Infarction. Biomed Signal Process Control 2017;31:71-8.
52. Tripathy RK, Dandapat S. Detection of Cardiac Abnormalities from Multilead ECG using Multiscale Phase Alternation Features. J Med Syst 2016;40:143.
53. Tripathy RK, Dandapat S. Automated detection of heart ailments from 12-lead ECG using complex wavelet sub-band bi-spectrum features. Healthc Technol Lett 2017;4:57-63.
54. Feng N, Xu S, Liang Y, et al. A Probabilistic Process Neural Network and Its Application in ECG Classification. IEEE Access 2019;7:50431-9.
55. Tison GH, Zhang J, Delling FN, et al. Automated and Interpretable Patient ECG Profiles for Disease Detection, Tracking, and Discovery. Circ Cardiovasc Qual Outcomes 2019;12:e005289.
56. Tripathy RK, Bhattacharyya A, Pachori RB. Localization of Myocardial Infarction from Multi-Lead ECG Signals Using Multiscale Analysis and Convolutional Neural Network. IEEE Sensors Journal 2019;19:11437-48.
57. Tripathy RK, Bhattacharyya A, Pachori RB. A Novel Approach for Detection of Myocardial Infarction From ECG Signals of Multiple Electrodes. IEEE Sensors Journal 2019;19:4509-17.
58. Cai W, Chen Y, Guo J, et al. Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. Comput Biol Med 2020;116:103378.
59. Oh SL, Adam M, Tan JH, et al. Automated Identification of Coronary Artery Disease from Short-Term 12 Lead Electrocardiogram Signals by Using Wavelet Packet Decomposition and Common Spatial Pattern Techniques. J Mech Med Biol 2017;17:1740007.
60. Bousseljot R, Kreiseler D, Schnabel A. Nutzung Der EKG-Signaldatenbank CARDIODAT Der PTB Über Das Internet. Biomedical Technology (Berl) 1995;40:317-8.
61. Pan J, Tompkins WJ. A real-time QRS detection

algorithm. IEEE Trans Biomed Eng 1985;32:230-6.
62. Zhang J, Liu M, Xiong P, et al. A Multi-Dimensional Association Information Analysis Approach to Automated Detection and Localization of Myocardial Infarction. Eng Appl Artif Intell 2021;97:104092.
63. Megahed M, Jain U, Leasure M, et al. Localization of Myocardial Infarction from 12 Lead ECG Empowered with Novel Machine Learning. In: 2019 3rd International Symposium on Computer Science and Intelligent Control. Association for Computing Machinery, 2019.
64. Adedinsewo D, Carter RE, Attia Z, et al. Artificial Intelligence-Enabled ECG Algorithm to Identify Patients With Left Ventricular Systolic Dysfunction Presenting to the Emergency Department With Dyspnea. Circ Arrhythm Electrophysiol 2020;13:e008437.
65. Makimoto H, Höckmann M, Lin T, et al. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. Sci Rep 2020;10:8445.
66. Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20:45-50.
67. The First Edition of the Artificial Intelligence Competition of Cardiovascular Disease Diagnosis 2019. Available online: http://mdi.ids.tsinghua.edu.cn
68. Smulyan H. The Computerized ECG: Friend and Foe. Am J Med 2019;132:153-60.
69. Macfarlane PW, Mason JW, Kligfield P, et al. Debatable issues in automated ECG reporting. J Electrocardiol 2017;50:833-40.
70. Rahman S, Karmakar C, Natgunanathan I, et al. Robustness of electrocardiogram signal quality indices. J R Soc Interface 2022;19:20220012.
71. Tripathi PM, Kumar A, Komaragiri R, et al. A Review on Computational Methods for Denoising and Detecting ECG Signals to Detect Cardiovascular Diseases. Arch Computat Methods Eng 2022;29:1875-914.
72. Hannun AY, Rajpurkar P, Haghpanahi M, et al. Publisher Correction: Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019;25:530.

doi: 10.21037/jmai-22-94
Cite this article as: Pantelidis P, Bampa M, Oikonomou E, Papapetrou P. Machine learning models for automated interpretation of 12-lead electrocardiographic signals: a narrative review of techniques, challenges, achievements and clinical relevance. J Med Artif Intell 2023;6:6.

Supplementary

Table S1 Details on the architecture of each model, along with relevant comments

Study	Architecture[†]	Details and comments[‡]
(64)	N/A	Previously trained and validated algorithm by Attia, 2019; LVSD validated with echo; Superior to NT-proBNP (AUC: 0.80)
(21)	I-{CBReP}x6CBRe{DBReDr}x2So-O	Explicit "Temporal – spatial" architecture; LVSD validated with echo; Patients without diagnosed LVSD but positive for model-screening, had a 4-fold higher risk for future LVSD
(22)	I-{{{BRCBC}//{CP}}x3Dr}x3CBReDrDSo-O	Explicit "Temporal – spatial" architecture; Future AF by current SR ECG; Combined AF/Atrial flutter vs. other labels; AF validated by trained personnel
(58)	I-{C//C//C}{SE}Cx2{SE}Cx4{SE}Cx6{SE}Cx4PDSi-O	{SE}: I-{PDReDSi}//{}-O; AF validated by expert cardiologists; Also two other experiments: AF vs. non-AF, AF vs. normal vs. non-AF
(30)	I-{CCP}x5GbABDSi-O	Adequate performance also for 476 patients with combined arrhythmias; Slightly worse for single-lead ECG
(29)	LSTM structure	Also compared against GRU and vanilla RNN models; No details about other blocks; LSTM seems superior to vanilla RNN and GRU alternatives; ACC is balanced (BACC); 549 ECGs, further segmented to 12,359 data points
(54)	I-DTW-DDSo-O	10 cluster centers by 40 typical samples; Superior to other traditional methods; Low training - High testing time
(31)	I-A{{CReBPDrCReBPDrA}//{GbBDrA}}BDrD-O	First detect MI (ACC=0.96), then locate it (ACC=0.63); Metrics for inter-patient analysis (intra-patient: ACC=0.99); Beat-based
(49)	WT - feature selection - SVM	Beat-based
(40)	I-{CRe//}x12C{CReBCReB}x3CBPDSo-O	Also compared against I-WT-PCAx4-3layerNN-O; Results slightly lower for hierarchical classification (WT-PCA-NN); 549 ECGs divided into 5,968 segments
(47)	I-{CBReP{Db}x3CBReP{Db}x3CBReP{Db}x3CBReP{Db}x3ALSo}//{{CBReP{Db}x3CBReP{Db}x3CBReP{Db}x3CBReP{Db}x3}//{CBReP{Db}x9}}-O	{Db}: I-{CBRe{{CBRe}//{ABRe}}-O; Complex, ensemble structure;
(41)	I-{CBP{Res}x4FDDrD}//2-DDrDDrDSi-O	{Res}: I-{CBCB}//{}Si-O; One module for irregularity and one for P detection (in parallel); Developed with internal data, validated on external data; Performance reported for PTB-XL
(28)	n.a.	Not provided; Copyrighted algorithms for each step (not provided); Validated on CSE
(48)	I-CPCP{Inc}x2P{Inc}x5P{Inc}x2P{G}x2DrSo-O	{Inc}: I-{CBR//CCBRe//CCBRe//PCBRe}-O; Also tested against external databases (custom set of 6,500 and PhysioNet set of 500 samples); Metrics from the external PhysioNet databank
(50)	WT - 228-feature selection (Relief) - KNN	Rule-based; Also compared against other classifiers (SVM, ANN, DA) with suboptimal results
(33)	I-{CBPCBPCBPP}//12-Lb//12-D-O	Beat-based
(34)	I-{{CBReP}x3DDrRe}//{Lb}-{DDrRe}x3So-O	Lb only for lead II
(65)	I-CBReP{CReP}x3DrDBReDrDSi-O	Superior to cardiologists' performance; Better results against lower resolution ECG images or fewer leads
(63)	n.a.	No details on how signals are segmented into 1 sec intervals
(35)	I-Lbx2DSo-O	
(59)	WT delineator (4 -level) - split - {CSP} - kNN	{CSP}: Custom trainable filter for extracting features; Beat-based; Training of the CSP filter (2-fold) and the kNN classifier (10-fold cross-validation); Train/test sets formed after WT
(37)	I-CP{Res}x3CAPDDSi-O	{Res}: I-{{{Sc}B}//{CB{Sc}BCB}Sw}-O; {Sc}: Custom module, similar to DWT; Dr added in between layers; Modified metric (CinC2020 guidelines)
(51)	WT - SVD - SVM	Segmented to beats to form frames; The ACC for detecting MI only is 0.953
(23)	I-{GA}//12ADSo-O	{GA} for every lead, then A for inter-lead mixing
(24)	I-{CReB{I}x4CReBP}//8{DRe}x3DSi-O	{I}: {CReB}//3P; Patients positive for algorithm result had a 7.2-fold higher risk for developing AF over 30 years; Downsized to 8 from 12 leads
(38)	I-CBRe{Res}x4DSi-O	{Res}: I-{{PC}//{CBReDrC}}{{BReDr}//{}}-O; No "normal" class is reported; Sigmoid activation to account for multi-labeled cases; Outperforms human evaluators (F1-score)
(25)	I-{{MB}Si{MB}P}//8-{{GbDr}//{Dr}}-ReDrABReDrDSi-O	{MB}: {{CReCRe}//{CReDr}}; Modified metric (CinC2020 guidelines); Downsized to 8 from 12 leads
(42)	I-{CP}x6So-O	

Table S1 (*continued*)

https://dx.doi.org/10.21037/jmai-22-94

Table S1 (*continued*)

Study	Architecture[†]	Details and comments[‡]
(55)	I-{CNN}-HMM-heuristic-GBM-O	{CNN}: CNN with C, B, D layers for delineation (features of ECG); HMM (Hidden Markov Model) and a heuristic used for optimizing the delineator; CNN finds meaningful segments (e.g.,: P wave) and features [725] which are "fed" to a GBM (Gradient Boosted Machine) classifier; CNN trained on 170 manually annotated ECGs
(52)	Phase alteration features (WT) - fuzzy kNN	
(53)	WT/FT - SVM	
(56)	I-FT/WT-CRePDDSo-O	Beat-based
(57)	WT/FT - NN	
(43)	I-{segm}-{C{{DB}BErCDrP}x3{DB}BReP}//10-DrFDSi-O	{segm}: Segmentation to 10 pieces of 10 sec each (with overlapping); {DB}: I-{{BReCDr}//{}}-{{BReCDr}//{}}-O
(44)	I-{Net1}//{Net2}//{Net3}-Ensemble-O	{Net1}, {Net2}, {Net3} represent 3 different networks ensembled: They have different architectures (including CBRe, P, D, etc. layers) and receive 12 single-lead and one 12-lead signals.; Beat-based; Averaged metrics for each class
(26)	I-{CoB}//12-Lx2-ASo-O	{CoB}: I-CCPCCPCCCPCCCPCCCP-O; Beat-based
(45)	I-{{FE}//{CBRe{CBReCBRe}x16{P//P}}}-DSo-O	{FE}: Morphological feature (QRS, RR, etc.) extraction
(27)	I-Gb//8ReDSo-O	Beat-based; Downsized to 8 from 12 leads
(36)	I-{{CRe}x3Dr}x5-LbReDrABReDrD-O	
(73)	I-CBReP{Res}x4PDSi-O	{Res}: I-{{CBReDrCB}//{}}Re-O
(32)	I-{C{As}{At}}x5-GbPSo-O	{As}: I-{P//P}DD{P//P}Si-O; {At}: I-{P//P}CSi-O; "Spatio-temporal" attention mechanisms
(62)	WT - PFA - Bagged Tree	Tensorization with WT (3rd order) to produce 36 features
(46)	I-CBReP{{CoB}P{IDEN}}x4PF{DReDr}x2Si-O	{CoB}: I-{{CBReCBRe}//{CB}}Re-O; {IDEN}: I-{{CBRe}//{}}Re-O; Outperforms human experts

[†], I, input; O, output; C, convolutional layer; F, flatten layer; G, gated recurrent unit (GRU) layer; Gb, bidirectional GRU layer; L, long-short term memory (LSTM) layer; Lb, bidirectional LSTM layer; B, batch normalization layer; A, attention layer; Re, ReLu activation layer; P, pooling layer; D, dense (fully connected) layer; Dr, dropout layer; So, Softmax activation; Si, Sigmoid activation; x, number of layer/module consecutive repetitions, e.g., Cx3 means 3 convolutional layers one after the other; {}, wraps a module/block; //, in parallel (if accompanied by a number, e.g., //3, it indicates the number of same modules running in parallel); QRS, QRS complex (ECG feature); HMM, Hidden Markov model; kNN, k-nearest neighbor; SVM, support vector machine; GBM, Gradient Boosting Machine; FT, fourier transformation variant; WT, wavelet transformation variant; n.a., not available. [‡], MI, myocardial infarction; PCA, principal component analysis; NN, neural network; CNN, convolutional NN; RNN, recurrent NN; RR, RR interval (ECG feature); Sw, Swish activation (x*Si(x)); sec, seconds; EF, ejection fraction; LVSD, left ventricle systolic dysfunction; echo, echocardiogram; proBNP, pro B-type natriuretic peptide (heart failure biomarker); AUC, area under the curve; AF, atrial fibrillation; SR, sinus rhythm; CSE, Common Standards for Electrocardiography database; PCinC2020, Physionet/Computing in Cardiology challenge 2020; ACC, accuracy.

https://dx.doi.org/10.21037/jmai-22-94

Table S2 Clinical conditions used as classification labels across studies

Abbreviation	Clinical condition	Frequency
MI	Myocardial infarction	21
AMI	Anterior MI	10
IMI	Inferior MI	10
ALMI	Anterolateral MI	8
ILMI	Inferolateral MI	8
ASMI	Anteroseptal MI	7
IPLMI	Inferoposterolateral MI	4
PMI	Posterior MI	2
IPMI	Inferoposterior MI	2
PLMI	Posterolateral MI	2
oMI	old MI	1
EMI	Early progression of MI	1
AcMI	Acute MI	1
CMI	Chronic MI	1
AF	Atrial fibrillation	16
I-AVB	1st degree atrioventricular block	12
RBBB	Right BBB	12
PAC	Premature atrial contraction	11
PVC	Premature ventricular contraction	11
LBBB	Left BBB	7
Tc	T-wave change	6
STD	ST-segment depression	5
STE	ST-segment elevation	5
LAFB	Left anterior fascicular block	4
ER	Early repolarization	4
LVSD	Left ventricular systolic dysfunction	2
II-AVB	2nd degree atrioventricular block	2
BBB	Bundle branch block	2
ST	Sinus tachycardia	2
LVHV	Left ventricle high voltage	2
STc	ST-segment change	2
HMD	Heart muscle defect	2
LVH	Left ventricular hypertrophy	2
VT	Ventricular tachycardia	2
SRa	Sinus-rhythm arrhythmia	1
Lad	Left axis deviation	1
WPW	Wolf-Parkinson-White syndrome	1
Afl	Atrial flutter	1
hyperK	Hyperkalemia	1
VpES	Ventricular pre-excitation syndrome	1
SVT	Supraventricular tachycardia	1
LAtH	Left atrial hypertrophy	1
–	Hypertrophy	1
CAD	Coronary artery disease	1
PAH	Pulmonary arterial hypertension	1
HyC	Hypertrophic cardiomyopathy	1
CA	Cardiac amyloidosis	1
MVP	Mitral valve prolapse	1
–	asystole	1

Frequencies pertain to the number of appearances of each clinical condition in the included studies. These clinical terms are stated as reported in each study. The MI superclass includes all the specific MI subcategories (indented).

https://dx.doi.org/10.21037/jmai-22-94

Table S3 Detailed performance metrics reported in each study

Study	AUPRC	AUC	ACC	F1	SEN	SPE	PPV	NPV
(64)		0.89	0.86		0.74	0.87	0.4	0.97
(21)		0.93	0.86		0.86	0.86		
(22)		0.87	0.79	0.39	0.79	0.8		
(58)			0.99	0.99	0.99	0.99		
(30)		0.91	0.97	0.84				
(29)			0.98	0.96	0.98	0.98	0.96	
(54)			0.74	0.76	0.75	0.74		
(31)			0.63		0.64	0.63		
(49)			0.93					
(40)			1		1	1		
(47)				0.87				
(41)		1	0.99		1	0.99	0.91	1
(28)			0.75		0.87	0.92		
(48)			0.93	0.9	0.9	0.98	0.9	
(50)			0.82		0.79	0.88		
(33)			0.93		0.94	0.86	0.97	
(34)			0.88		0.86	0.88		
(65)		0.88	0.81	0.82	0.86	0.76	0.79	0.85
(63)				0.99	0.99		0.99	
(35)				0.74				
(59)			1		1	1		
(37)			0.72					
(51)			0.98					
(23)			0.98		0.98	0.99		
(24)	0.21	0.83						
(38)				0.93	0.93	1	0.92	
(25)			0.57					
(42)					0.93	0.9	0.94	
(55)		0.87						
(52)			0.86					
(53)			0.98					
(56)			1					
(57)			1		1	1		
(43)				0.86				
(44)		0.95			0.96	0.95		
(26)			0.85	0.81	0.8		0.83	
(45)				0.88				
(27)			1					
(36)			0.95					
(73)		0.97	0.97	0.81	0.81		0.82	
(32)			0.87	0.84				
(62)			0.99		1	1		
(46)		0.98		0.89	0.87	1		

AUPRC, area under the precision-recall curve; AUC, area under the curve; ACC, accuracy; F1, F1-score; SEN, sensitivity; SPE, specificity; PPV, positive predictive value; NPV, negative predictive value.

References

73. Zhang D, Yuan X, Zhang P. Interpretable Deep Learning for Automatic Diagnosis of 12-Lead. arXiv:2010.10328.

https://dx.doi.org/10.21037/jmai-22-94