Computers in Biology and Medicine 172 (2024) 108235

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review

Pedro A. Moreno-Sánchez [a,*], Guadalupe García-Isla [b], Valentina D.A. Corino [b], Antti Vehkaoja [a], Kirsten Brukamp [c], Mark van Gils [a], Luca Mainardi [b]

[a] *Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*
[b] *Department of Electronics Information and Bioengineering, Politecnico di Milano, Italy*
[c] *Protestant University Ludwigsburg, Ludwigsburg, Germany*

ARTICLE INFO

Keywords:
Cardiovascular diseases (CVD)
ECG
Deep learning
Machine learning
Trustworthy Artificial Intelligence
Explainable Artificial Intelligence (XAI)
Bias
Ethical, Legal and societal implications (ELSI)

ABSTRACT

Cardiovascular diseases (CVD) are a leading cause of death globally, and result in significant morbidity and reduced quality of life. The electrocardiogram (ECG) plays a crucial role in CVD diagnosis, prognosis, and prevention; however, different challenges still remain, such as an increasing unmet demand for skilled cardiologists capable of accurately interpreting ECG. This leads to higher workload and potential diagnostic inaccuracies. Data-driven approaches, such as machine learning (ML) and deep learning (DL) have emerged to improve existing computer-assisted solutions and enhance physicians' ECG interpretation of the complex mechanisms underlying CVD. However, many ML and DL models used to detect ECG-based CVD suffer from a lack of explainability, bias, as well as ethical, legal, and societal implications (ELSI). Despite the critical importance of these Trustworthy Artificial Intelligence (AI) aspects, there is a lack of comprehensive literature reviews that examine the current trends in ECG-based solutions for CVD diagnosis or prognosis that use ML and DL models and address the Trustworthy AI requirements. This review aims to bridge this knowledge gap by providing a systematic review to undertake a holistic analysis across multiple dimensions of these data-driven models such as type of CVD addressed, dataset characteristics, data input modalities, ML and DL algorithms (with a focus on DL), and aspects of Trustworthy AI like explainability, bias and ethical considerations. Additionally, within the analyzed dimensions, various challenges are identified. To these, we provide concrete recommendations, equipping other researchers with valuable insights to understand the current state of the field comprehensively.

1. Introduction

Cardiovascular diseases (CVD) refer to a range of conditions that affect the heart and blood vessels, including coronary artery disease, heart failure, stroke, and peripheral artery disease. CVD is one of the leading causes of death worldwide, accounting for an estimated 17.9 million deaths each year (i.e. 32% of all global deaths) [1]. In addition to the significant impact on mortality, CVD also results in significant morbidity and reduced quality of life, with symptoms ranging from chest pain and shortness of breath to cognitive impairment and mobility limitations [2]. Despite significant advances in the diagnosis, prognosis, and prevention of CVD, challenges remain in tackling the complex nature of these conditions. Overall, dealing with the growing burden of CVD requires a multi-faceted approach, where research efforts must also continue to expand the understanding of the complex mechanisms underlying CVD, as well as develop innovative approaches to address these conditions.

The electrocardiogram (ECG) is a signal that can be used as a diagnostic tool by capturing the electrophysiological activity of the heart. It is widely used in clinical medicine, offering crucial information necessary for the identification and treatment of various CVDs [3]. The usefulness of the ECG is not limited to acute care settings but also extends to various other areas including primary care for outpatients, home care, preoperative evaluations, athletic screenings, telemedicine, and self-monitoring. With the growing number of ECG recording procedures being conducted, the demand for skilled cardiologists to interpret and analyze the resulting data continues to rise substantially. This has led to increased workloads and financial pressures, ultimately creating an

* Corresponding author. (
E-mail address: pedro.morenosanchez@tuni.fi (P.A. Moreno-Sánchez).

https://doi.org/10.1016/j.compbiomed.2024.108235
Received 11 August 2023; Received in revised form 7 February 2024; Accepted 25 February 2024
Available online 28 February 2024

environment that contributes to physician burnout and reporting inaccuracies [4].

Since its introduction more than 50 years ago, computer-assisted interpretation of the ECG has become an essential component of clinical workflows, supplementing physician interpretation. Conventional approaches rely on computer-aided detection and measurement of predetermined ECG features such as waves, segments, and time-intervals, followed by rule-based classification of their normal or abnormal status. However, the accuracy of traditional models for computer-assisted ECG interpretation is suboptimal, which can be attributed to outdated classification rules and their vulnerability to imperfect tracings. Data-driven approaches such as Artificial Intelligence (AI) models have been recently employed to tackle this issue – so far with mixed results.

AI has become a promising tool for building computer-aided diagnosis systems capable of classifying individuals with specific symptoms, e.g., either as having a disease or being healthy [5,6]. AI research and development is a multidisciplinary field that combines principles from mathematics and computer science to create systems capable of learning from example data and existing knowledge to perform tasks with increasing performance. It encompasses various subfields, among which are Machine Learning (ML) and Deep Learning (DL). ML enables the construction of data-driven models that are adept at performing tasks such as classification, regression, and clustering. Shallow ML methods (logistic regression, random forest, support vector machine, K-nearest neighbors) often use feature engineering, where domain expertise is used to extract relevant features from raw data for effective, and explainable, model training. Deep Learning (DL), a specialized subset of ML, uses multiple hidden layers that can implement more complex processing steps for more demanding tasks. Provided enough training data are available, these networks may autonomously learn to transform raw data through a series of non-linear operations, thereby highlighting features that are crucial for task-specific objectives like classification and regression. The architecture of deep networks allows also the handling of large volumes of unstructured data such as free text [7]. Recent studies in clinical cardiology have demonstrated that ML and DL, using combined modalities, are better equipped to predict cardiovascular or all-cause mortality compared to the individual use of individual clinical or imaging modalities [7].

Although ML and DL have become popular in the information technology industry, their integration into the CVD field has been much more restrained. This is due to a set of constraints and requirements that go beyond the mere technical performance of the algorithms, such as data collection processes (e.g., representativeness of the population, sample size), adoption in the existing medical workflow, external validation, or (un)fairness in predictions made [8]. Modern ML and DL tools can provide accurate predictions, but their “black box” behavior can be problematic in understanding their decision logic. This may hinder the acceptance of ML/DL tools by clinicians, who must interpret the tools’ outputs in their decision-making process. Therefore, eXplainable Artificial Intelligence (XAI) provides insights into how an AI system arrives at its decisions, making it easier for healthcare experts to understand, thereby enhancing the clinical adoption and acceptance of AI models [9]. Despite its objective-data-driven nature, ML in healthcare faces challenges such as biases and ethical concerns (fairness, data governance, etc.) requiring.

Despite the recognized significance of addressing understandability, bias, and other ethical, legal, and social implications (ELSI) in ML/DL models for healthcare, there remains a notable gap in comprehensive literature reviews that specifically focus on these aspects in CVD prediction models. This systematic review aims to bridge this gap by offering, to the best of our knowledge, the first in-depth analysis that holistically examines ECG-based data-driven models for CVD prediction, with a particular focus on Trustworthy AI aspects [10]. Initially, the review categorizes existing works based on the type of CVD diseases predicted, the nature of data, and dataset sizes used as inputs. The novelty of our work is further underscored by an extensive analysis of DL model architectures, performance metrics, and a critical evaluation of their strengths and weaknesses. A cornerstone of our review is the emphasis on explainability and ethical considerations in healthcare applications. We meticulously explore and scrutinize the limited yet pivotal works addressing these issues, making this a central element of our analysis. Finally, we present a discussion that identifies various challenges and shortcomings in the field, followed by concrete recommendations to address these issues. With this systematic review, we offer a multifaceted analysis of ML/DL models that utilize ECG for CVD prediction, equipping other researchers with valuable insights to comprehensively understand the current state of the field. In addition, we provide a tabular representation of most representative works that employ DL in this area, as well as works that implement XAI to be used as a quick reference that encapsulate the state-of-the-art in the field guiding future research endeavors.

The remainder content of this paper is structured as follows: Section II describes the use of the PRISMA methodology to perform the review. Section III details the analysis of the areas proposed as strategic research lines for the review, describing and discussing the main trends identified. Section IV presents a comprehensive discussion of the review findings in terms of challenges and shortcomings, and provides recommendations to address them. The conclusions drawn from this study are exposed in Section V.

2. Review methodology

To perform an exhaustive analysis of the field that complies with medical literature review standards, we have adopted the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) methodology [11,12]. A comprehensive detailing of the items listed by PRISMA can be found in Supplementary Table S1.

The research and review questions guiding this study were agreed upon by the authors following multidisciplinary discussions around the clinical endpoints investigated in the ERA PerMed research project “Personalised Prognostics and Diagnostics for Improved Decision Support in Cardiovascular Diseases (PerCard)” [13]. The formulated research questions (RQ) are as follows: RQ1) Which CVD, and diagnostic or prognosis scenarios are being targeted by the data-driven models found in literature?; RQ2) What types of CVD-related data are used as input to these models?; RQ3) Which ECG features are employed for training the models?; RQ4)What are the data-driven models predominantly used for CVD prediction, and what is their performance?; RQ5) How is explainability addressed in the context of these data-driven models’ output?; RQ6) Do the reviewed works take into account ethical considerations?

Prior to collecting the works for the literature survey, we established eligibility criteria including papers published from 2007 until February 2023, written in English, with an availability of full papers, and published by peer-reviewed scientific journals or presented at renowned international conferences in the field, such as Computing in Cardiology [14], IEEE-BHI [15], and IEEE EMBC [16]. We consider research works published during the last 15 years, as this period has witnessed unprecedented advancements in ML and DL, especially due to the advent of big data, improved algorithms such as convolutional neural networks, and increased computational power [17]. Additionally, for this review on data-driven approaches, the studies must include AI and ML techniques to process ECG and optionally other clinical and patient data. The clinical output of these studies must have been focused on the automated diagnosis or prognosis of CVD.

To ensure a comprehensive and targeted literature search, multiple databases were utilized. Google Scholar, and Web of Science served as the primary general literature databases, selected for their extensive coverage across multiple disciplines. To focus on publications specifically within the medical domain, the search was augmented with articles from PubMed and Scopus. Additionally, IEEE Xplore was initially included to capture specialized articles published in the conference proceedings of interest. The search query employed in the databases is included in Table 1.

The process of selecting papers for analysis was carried out in two phases. In the first phase, three independent reviewers (P.M.S., G.G.I., V. D.A.C.) assessed whether the search results fulfilled the eligibility criteria. In the second phase, four independent reviewers (P.M.S., G.G.I., V. D.A.C, A.V.) reviewed the titles and abstracts of the screened results, ensuring that the papers focused on processing ECG signals through AI/ML/DL algorithms for prediction tasks, and include a performance validation. As an optional criterion, this second phase also considered papers that included data inputs other than ECG and analyzed the explainability or bias of the results. In case of disagreement, the reviewers reached a consensus on which articles to screen full text through discussion. Finally, six independent reviewers (P.M.S., G.G.I., V. D.A.C., A.V., L.M., K·B.) were assigned an equal number of papers to read in full, extract and collect the information by using a shared common form that reflect the different areas of analysis, which are shown in Table 2.

Following the PRISMA instructions, the effect measure of interest in this literature survey was defined as the performance of the AI/ML models utilized in the studies. Specifically, the metrics of accuracy or AUROC were considered as the main effect measures.

As outlined in Table 2, this systematic review content has been organized by grouping the different areas into bigger categories of synthesis pertaining to AI-based solutions for predicting CVD using ECG. These categories and their subsequent areas are aligned with the research questions as well as the search terms and inclusion criteria employed when scrutinizing the search results.

3. Results

3.1. Screening and selection of related works

In accordance with the predefined search criteria, the eligible references underwent a rigorous selection process consisting of various screening and identification phases. The number of references, along with the specific requirements at each stage, are presented in Fig. 1. By utilizing the search terms in the designated databases, we initially identified a total of 637 references, which were further filtered down to 243 during the screening phase. Subsequently, we scrutinized the title and abstract of the screened papers based on the inclusion criteria related to content, resulting in a final set of 101 papers that were selected for full-text reading and subsequent analysis.

3.2. Study design

1) Diagnosis and prognosis of CVD

The majority of the reviewed studies had a diagnostic objective (78%), with only a modest percentage (15%) having a prognostic scope. The remaining 7% of papers focused on different objectives, such as ECG quality assessment, noise filtering, blood pressure estimation, or physical activity detection. Studies with a diagnostic objective present models aimed to support clinicians to identify the existence of a certain occurring condition in the heart (e.g., atrial fibrillation, arrhythmias, etc.). On the other hand, prognostic studies pursue estimating the onset of a future certain heart condition or event (e.g., myocardial infarction, or stroke) that is not manifest at the time of the prediction.

Table 2
Areas of analysis considered in the review with corresponding research questions RQ1-RQ6.

Macro categories of analysis	Area of analysis	Research question
Study design	Type of CVD examined (e.g., arrhythmia, acute myocardial infarction)	RQ1
	Target of AI decision support (e.g., diagnosis, prognosis, risk stratification)	RQ1
	Data used by the AI models (e.g., ECG, clinical data, images, other biosignals)	RQ2
	Number of participants/data items (e.g., patient records, heartbeats, ECG signals) in the train/test sets	RQ2
	Type of study performed (e.g., retrospective or prospective)	RQ2
Methodological approaches	Type of ECG features processed: (e.g., raw signals with one or more leads, hand-crafted time- or frequency features from signals)	RQ3
	DL/ML methods used (e.g., Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), ensemble classifiers, shallow ML classifiers)	RQ4
	Model performance results and performance metrics used (e.g., accuracy, sensitivity, specificity, area under the receiver operating characteristic curve (AUROC))	RQ4
	Type of validation (e.g., cross-validation, hold-out test set, external validation)	RQ4
Trustworthiness	Explainable AI techniques considered (e.g., SHapley Additive exPlanations (SHAP) values, feature importance, class activation mapping)	RQ5
	Any reported variable being subjected to any type of bias, and any Ethical, Legal, and Societal Implications (ELSI)	RQ6

Regarding the distribution of the different CVD addressed in the reviewed works, the majority of the reviewed studies, as shown in Fig. 2 a, are focused on arrhythmias [18–55], myocardial infarction [40,42,46,56–74], and conduction disorders [18,20–22,36–39,42,46–48,56,57,64,69,70,74–78], comprising 26%, 15%, and 15% of the total, respectively. Additionally, 11% of the papers aimed to develop beat classifiers [36,48,78–91], while 9% focused on ST or T wave alterations [20–22,37–39,47,56,64,69,70,74,92], 8% on hypertrophy [19,37,56,64,69,70,73,74,93–95], and another 7% on coronary artery disease [71,72,95–102]. Notably, a relatively small proportion of studies, approximately 9 % of the total, assessed other CVD diseases or cardiovascular-related issues such as estimation of blood pressure, hypertension, biological age via ECG, ejection fraction, wellness condition, mortality prediction, or ECG quality assessment [24,103–114].

Most of the diagnostic studies were focused on arrhythmia at 29%,

Table 1
Search Query used for collecting research papers.

```
(((“Cardiovascular disease” OR CVD) AND ((risk OR diagnosis OR prognosis) OR (recurren* coronary event*) OR (MACE OR major adverse cardiac events) OR (angiography) OR
   (postoperative atrial fibrillation)))
 AND
 ((ECG OR Electrocardiogram)
   OR ((ECG OR Electrocardiogram) AND (digital biomarkers OR biosignal features))
   OR ((ECG OR Electrocardiogram) AND (multivar*)))
 AND
 ((“Machine Learning” OR “Deep Learning” OR “Unsupervised Learning” OR “Artificial Intelligence” OR “supervised Learning” OR “semi-supervised Learning")
   OR ((“Machine Learning” OR “Deep Learning” OR “Unsupervised Learning” OR “Artificial Intelligence” OR “supervised Learning” OR “semi-supervised Learning”) AND
   ((“Explainable AI” OR XAI OR Interpretab* OR Explainab*) OR (Bias)))))
```

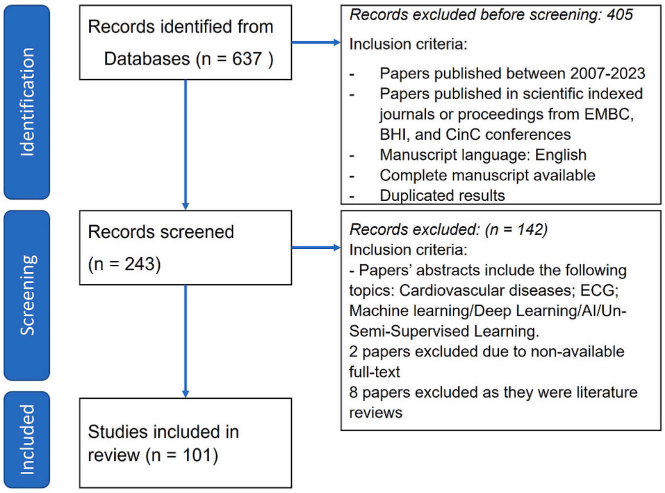
Identification
Screening
Included
Records identified from Databases (n = 637)
Records excluded before screening: 405
Inclusion criteria:
- Papers published between 2007-2023
- Papers published in scientific indexed journals or proceedings from EMBC, BHI, and CinC conferences
- Manuscript language: English
- Complete manuscript available
- Duplicated results
Records screened (n = 243)
Records excluded: (n = 142)
Inclusion criteria:
- Papers' abstracts include the following topics: Cardiovascular diseases; ECG; Machine learning/Deep Learning/AI/Un-Semi-Supervised Learning.
2 papers excluded due to non-available full-text
8 papers excluded as they were literature reviews
Studies included in review (n = 101)

Fig. 1. Identification, screening, and inclusion of studies for the review following the PRISMA method.

followed by studies on conduction disorder at 18% and myocardial infarction at 13%, beat classification at 12%, ST/T-wave alterations at 10%, hypertrophy at 9% and coronary artery disease at 7%. Notably, one study specifically examined post-COVID-19 patients' diagnosis and low-resynchronization therapy success. With respect to prognostic studies, the majority (28%) was focused on myocardial infarction, followed by stroke at 12% and heart failure at 11%. The remaining prognostic studies addressed miscellaneous objectives. Fig. 2 b shows a graphical distribution of the different CVD diseases and their relation with the clinical goal, i.e., diagnosis or prognosis identified in the reviewed studies.

The differences in the volume of studies targeting diagnosis instead of prognosis could be caused by the publicly available data. While diagnosis or monitoring based on ECG signals can be developed using just ECG signals regardless of their length, a predictive model offering any CVD prognosis requires clinical data and ECG recordings previous to the disease occurrence or adverse outcome development. The time interval from the ECG signals recording to the target event occurrence (i.e., mortality, arrhythmia development, MI, HYP, etc.) plays an essential role in the development and validation of any predictive model. Most current prognostic models are presented without any information regarding the stage of development of the disease. If no information is given regarding the duration between diagnosis and the available ECG recordings, no insight can be provided regarding at which stage of the disease the model is able to predict it. Therefore, in most cases, it would be more proper to talk about "monitoring" or "classifiying" rather than performing prognosis. Nevertheless, in contrast with prognosis studies, even without this information it is possible to develop a classifier able to distinguish between a pathological ECG recording and a healthy one.

CVD prognosis is of similar or higher importance than diagnosis, as long as we are not referring to a critical disease diagnosis that require immediate intervention, as lifestyle changes and prophylactic measures could prevent fatal events. Taking the example of AF, predictive models could prevent cryptogenic strokes or further myocardial tissue deterioration into the actual development of the arrhythmia. Most CVD are irreversible, thus, predictive and prognostic models potentially have a high clinical and socio-economic impact.

2) Study design and sample size

All the studies in the reviewed literature have been retrospectively organized i.e., we did not find any study that was set up to prospectively validate an AI decision support tool, neither for diagnosis nor for prognosis. Therefore, none of the studies also included intervention based on the use of reported AI-based tools and are thus all observational.

Several studies have used large or very large datasets for the analysis, such as 740,000 patients by Han et al. in Ref. [97], 277,807 12-lead ECGs by Zhang et al. [42], 175,943 12-lead ECGs in the study by Jin and Dong [36], 140,000 12-lead ECGs recordings used by Zhou et al. [55], 88,597 patients by Chen et al. [24], 71,741 patients by Chang et al. [106], 64,196 patients by Yang et al. [45], 56,793 patients by Diamant et al. [77], 51,579 patients by Liang et al. [21], and 42,511 in the study by Sakli et al. [22]. Also, twelve other papers among the total 103 reviewed had at least 10,000 patients in their datasets [19,20,24,32,42,62,64,70,74,100,108,114], while 13 papers account for between 1000 and 10,000 [23,35,41,49,52–54,61,68,94,110,113,115], and 20 between 100 and 1000 [43,60,63,66,71,72,75,90,92,93,95,99,101–104,111,112,116,117]. However, a large part of the studies (23) had less than 100 subjects [18,26–28,31,40,44,50,51,57,60,65,80,81,83,86–89,91,96,105,107]. The distribution of the reviewed works according to the

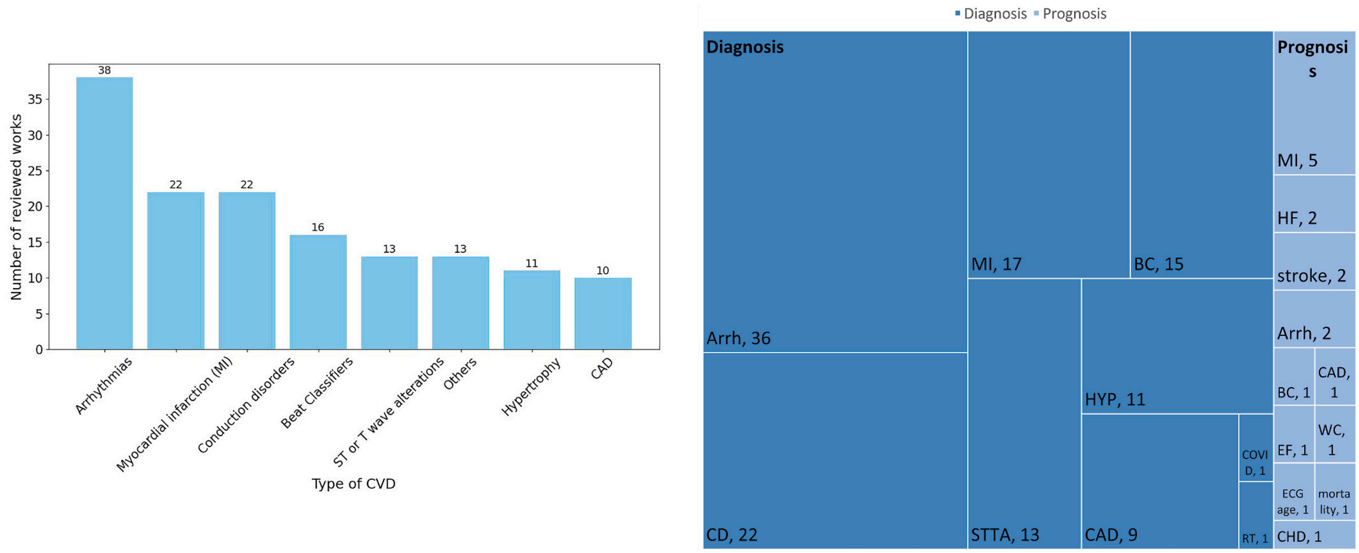

Fig. 2. Number of reviewed works categorized by (a) type of CVD, (b) and clinical goals (diagnostic, prognostic). MI: myocardial infarction, Arrh: Arrhythmia, CD: Conduction Disorder, BC: Beat Classifier, STTA: ST/T alterations, HYP: Hypertrophy, CAD: Coronary Artery disease, RT: Resynchronization therapy, HF: Heart Failure, WC: Wellness condition, EF: Ejection Fraction, CHD: Coronary Heart Disease.

number of participants recruited is depicted in Fig. 3. It is questionable whether less than 100 subjects would allow adequate sample size for training and optimizing the structure of a deep learning model, especially when data augmentation techniques have not been widely used, and whether the obtained results are generalizable.

The most widely used database is MIT-BIH [25–28,31,36,43,44,50,50,52,54,80,84,86–90,105,114], which includes datasets like such as Normal Sinus Rhythm (18 EC G recordings), Malignant Ventricular ectopy (22 EC G recordings), CU Ventricular Tachycardia (35 EC G recordings), Supraventricular Arrhythmia Database (48 EC G recordings), and Atrial Fibrillation database (23 EC G recordings). In addition, other open databases such as PTB-XL [22,38,43,63,64,67,69–74,111], CPSC2018 [21,22,38,39,47,49,53,118], Physionet CinC [33,35,35,38,38,52,53,109], St. Petersburg [71,72,84,90], or MIMIC-II [104,113] were frequently used. Most studies using proprietary datasets had less than one hundred subjects. Using open datasets is beneficial as it enables fast development cycles and more straightforward comparing the results with what has been obtained by other researchers. On the other hand, publicly available datasets may not include all the desired variables, and even developing a tool that could be evaluated and used in a specific setting, there may be poor applicability of that open-access dataset to the local population. Also, there may be variations in how data is collected between countries, which may influence the data, e.g., in case of blood pressure and blood test values.

It is worth noting that some of the reviewed papers handled the data as data segments, and not as subjects. This could imply a potential bias since the ECG segments belong to a same patient while are wrongly treated as independent observation [18,24,50,51,56,62,64,67,71,74,80,94]. Therefore, there is a clear research gap in the definition of evaluation or validation practices of how AI decision support tools trained with retrospective datasets could be estimated to perform in prospective research settings.

3) Study data

According to the reviewed studies, the majority (80%) of AI models utilized ECG signals as the only input, while 20 % used also additional information. Additional data sources included imaging, clinical background data, and other biosignals. Imaging data were used in only a small proportion of studies, with CT, echocardiography, and MRI images included in 2, 4, and 1 study respectively [23,100], [24,93,102,116], [23]. Clinical information, ranked by frequency of use, included demographic data [20,23,27,31,40,60,67,68,97,99,102,106–108,113,119,120], anthropometry [23,40,60,68,97,102,106,108], biochemistry [23,24,60,102,106], clinical measurements [60,102,103,106,108,119], lifestyle [23,68,102,102,107], clinical history [23,102,106,119] and medication [23,119]. The most included demographic information was age [20,23,27,40,60,67,68,97,102,106,116], followed by sex [27,40,60,67,68,97,102,106], and ethnicity [23,60,68]. Other added signals were the photoplethysmogram PPG (5 studies) [31,99,104,113,120], and arterial blood pressure (ABP) (3 studies) [23,68,102]. The different pieces of information considered in the related works are shown graphically in Fig. 4.

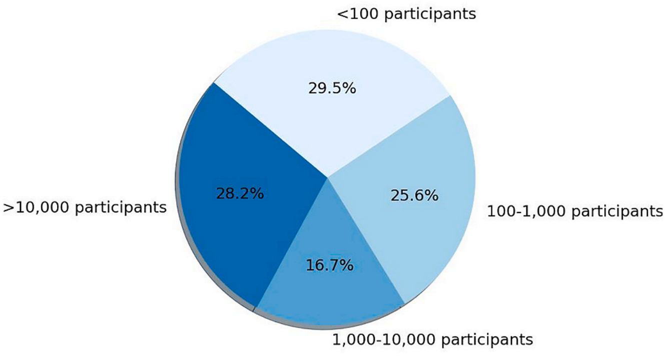

Fig. 3. Distribution of studies according to the number of subjects available in the database.

The limitation of using solely ECG signals as input likely primarily stems from the constraints posed by data availability rather than by the relevance of the other type of data, as most public databases contain only ECG signals, that are relatively easy to obtain. The development of models using exclusively ECGs has the advantage that ECG collection is a cheap and accessible process that forms part of routine clinical procedures. Therefore, these models could easily be introduced into a clinical setting. Multi-modal data gathering is time- and resource consuming, thus, public databases play a pivotal role in democratizing research efforts, allowing research groups lacking extensive resources to construct AI models using multi-center data. In addition, they set a common framework for comparison between studies. However, the use of additional other data types can help make more powerful models, better interpret the AI model performance, and target more complex decision support scenarios. Complete open-access databases integrating ECGs, clinical history, imaging data, and biochemistry could open the field for more complete, reliable, and complex studies.

3.3. Methodological approach

1) ECG signal processing, leads, and features

In the field of ECG signal analysis, there is an emerging trend toward adopting end-to-end approaches, i.e., using a DL model without so-called feature engineering to extract summary descriptors (features) from the raw data, such as statistical descriptors, clinically inspired features etc, but rather input raw data ‘as is’. However, many studies employ feature extraction methods in conjunction with deep learning (DL) techniques, aiming to facilitate the interpretation of the classification or diagnostic performance of the developed models, which is important in healthcare

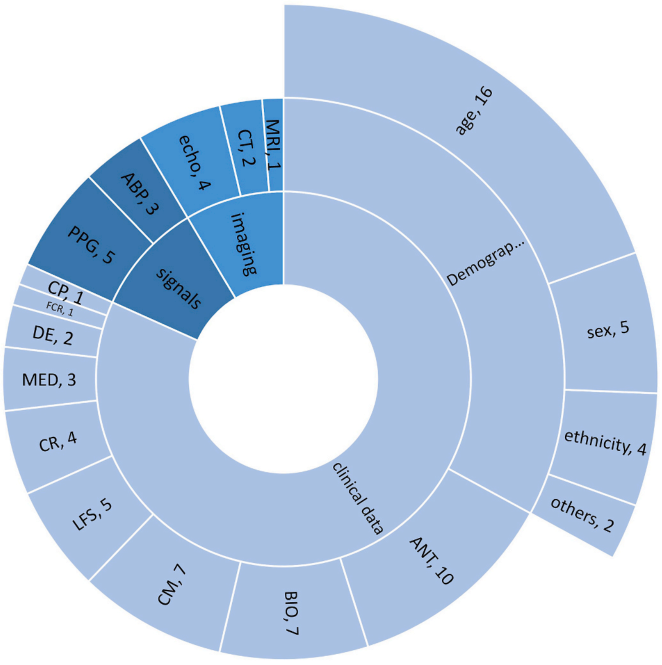

Fig. 4. Distribution of data considered as input in prediction models reviewed. The figure displays the number of papers that consider the specific type of data (e.g., age, sex, PPG). The numbers are not exclusive, i.e. one study may contain several types of data. CM: Clinical measurement, MED: Medication, DE: disease-related events, ANT: anthropometry, CP: clinical procedures, FCR: Family clinical record, CR: Clinical record, BIO: Biochemistry, LFS: Lifestyle.

applications. In this literature review, we identified that 58 studies (56%) [3,4,7,18,20]– [25,27,29,30,33,36,39]– [43,46,47,49,51]– [53,57,59, 62,64,69]– [72,74,75,77,78,80,82,84,85,87,88,90]– [92,96,101,102, 104,105,107,112,114,115,118,121] exclusively utilized raw ECG signals for AI model development, while 46 studies (44%) incorporated feature extraction. The average length of ECG signals used in these studies was 34 s, with a standard deviation of 105 s and a median of 8 s. Concerning the number of ECG leads used in the reviewed studies, 41 studies used all 12 leads [19,21–24,32,36,38,39,41,42,45,47,49,53–58,62,63,63,64, 67–70,74–76,92–94,97,101,103,106,108,109,115], 35 studies used a single lead [18,20,26,31,35,43,44,46,51,52,60,61,65,71–73,78,79,81, 83,86,91,95,96,99,101,102,104,105,107,111,113,116,118,122], and 21 studies used multiple leads (less than 12). Among the limb leads, Lead II was the most frequently utilized, while precordial lead V1 was also commonly employed. An overall comparison of the use of the different ECG's leads is shown in Fig. 5.

From the studies calculating ECG features to fit in the prediction algorithms, we can divide the most commonly extracted ECG features into three main categories: a) statistical features, b) morphological features, and c) RR-based features.

Statistical features quantify the signal fluctuations through the measurements of certain mathematical properties of the histograms of ECG values [26,28,32,35,39,41,65,73,76,86,95,111,113,117,123]. They include minimum, maximum, and mean histogram values, standard deviations, kurtosis, and skewness measurements. These features may be computed on the raw ECG traces or after the application of some transformations. In most cases, the adopted transformation is the Wavelet Transform, and the statistical features are extracted on signal approximations/details at different scales [34,63,73,111]. The other adopted transformations were the Fast Fourier Transform [95], the Legendre Polynomial Transform [110], and the Variational Mode Decomposition [63]. A second class of statistical measurements is based on the computation of parameters related to the non-linear dynamic characterization of the ECG such as Entropies (Shannon, Approximate, etc …) [73,111,113] or fractal dimensions computation (Higuci fractal dimension, Kats fractal dimension, etc …) computation [73,113]. This second class of indexes provides information on the sequences of samples evidencing repetitive signal patterns or long-term correlations. As before, these indexes were computed directly on the ECG leads or after some transformation.

Morphological features describe the classical ECG waveforms in terms of amplitudes, area and waveform polarity on each lead and/or by combinations of leads to derive the electrical axis (P axis, QRS-Axis and T-axis) [19,28,32,41,45,55,56,58,61,67,68,92–95,101,102,106,108, 117,124]. Among the morphological features, we have also included waveform durations (*P*-QRS and T durations) and time intervals among the main waveforms (PR, QT, QTc interval). The use of these features is recommended as they carry a direct interpretation in terms of the electrical behavior of the heart, and are well-known to clinical experts, who are typically the end-users of the provided tools. Therefore, these features can be employed for the (explainable) prediction of some arrhythmic (AF, AV block, Tachycardia, Ectopic beats, etc.) or ischemic patterns.

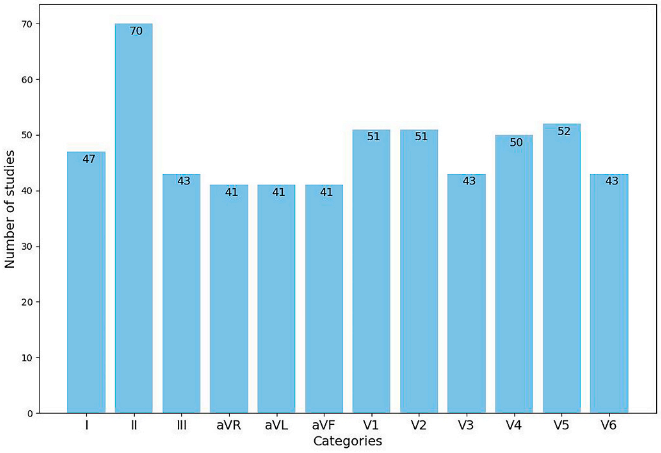

Fig. 5. Overall view of ECG leads used in reviewed works.

RR-based features describe the variability of the RR interval series computed between the R-peaks of two consecutive normal QRS complexes (the *N*N-interval, NNI) and are also called Heart Rate Variability (HRV) features. They can be clustered into three groups: time domain, frequency domain and time-frequency domain features. Common RR time-domain parameters include the mean of NN intervals (Mean-NNI), median of the successive difference between NN intervals (Median-NNI), range NNI (Range-NNI), PNNI-50 (percentage of successive NN interval greater than 50 ms) and standard deviation of the NN intervals (STD-NNI). Commonly used HRV features derived from the NN intervals may also include RMSSD (root mean square of successive differences of NN intervals), CVNNI (Co-efficient of variation equal to the ratio of standard deviation of the NN intervals divided by mean NN interval) and CVSD (Coefficient of variation of successive difference) equal to the root mean square NN intervals divided by mean NN interval [26,31,32,34,34,35, 41,45,54,56,60,63,65,67,68,83,89,92,95,101,103,108,116]. Frequency-domain features are obtained after computation of the NN spectrum and include quantification of spectral power in specific bands: absolute or normalized power in High frequency (HF:>0.15 Hz), Low Frequency (LF: 0.05–0.15 Hz), Very Low Frequency (VLF: <0.04 Hz), or their ratios LF/HF (ratio of low frequency and high-frequency power) [45,60,61,66,95,101,103,105,113,116]. Finally, the time-frequency parameters include the computation of statistical features (histograms-based or Entropies) after a time-frequency transformation (most frequently a Wavelet transform) of the original NN series [39,44,124].

The ML approach based on ECG features extraction has the advantage of feeding the ML model with domain-specific, physiological knowledge of the bioelectric phenomena within the heart and their reflections on recorded ECG traces (either manifesting in morphological, rhythmic, or variability changes). Indeed, it is nowadays possible to build 3D computational models of the heart and the torso and simulate ECG traces in a realistic way also including pathological situations at different degrees of severity. Models can be restricted to atria activity [125], ventricular activity [126], and its manifestation [127,128] or extended to a whole heart model [129]. Irrespective of the model used, these approaches lead to the definition of domain-specific (or pathology-specific) ECG hallmarks to be used for detecting the pathology and its manifestations. Such features can support the adoption of human-defined features or help to identify new ones.

ML models adopting these domain-specific features inherently have a lower structural complexity, which also implies shorter model training times (improving computational efficiency) and the possibility to perform training on smaller ECG datasets. In general, these models are less prone to overfitting because the ECG features are physiologically grounded and are expected to consistently emerge, as significant, in various databases. Conversely, deep-learning models are the result of the mining of a specific database, and they constantly need verification on several datasets to minimize the risk of overfitting. The advantages of a feature-based approach would obviously vanish if a battery of ECG features is blindly computed and used to feed a network without any domain-specific motivation [37,94,95]; an approach not unlike a 'fishing trip', which unfortunately happened in many of the studies in reviewed papers. Finally and importantly, feature-specific ML models, especially when trained on smaller and well-curated datasets, are highly interpretable and it is easier to identify the driving factors that have led to model predictions [130].

When feature-based ML models are employed, a potential issue emerges, since features must be calculated from the ECG signal and any random or systematic error in their computation will propagate to the output and may limit the accuracy or generalizability of the model. This observation should focus the researcher's attention on the methods for feature pre-processing and feature screening to critically assess, ideally in close interaction with a domain expert, the presence of outlier or deviant values in the data, before including them in the model development, but this aspect is rarely addressed in the examined works [32, 58]. Moreover, the algorithm used for computing the features is also part of the solution and does thereby affect the performance of the ML decision making algorithm. Data curation, pre-processing, feature extraction algorithms and ML models are connected parts in the decision support pipeline, and if they are disentangled, the observed performance on the development set may not be guaranteed to be representative of real-life performance on new data. This is another aspect which is seldom addressed in the papers.

2) ML and DL algorithms and their performance in CVD predictions

In the reviewed research works, 14 different ML or DL algorithms were employed. DL architectures were slightly more common than shallow ML algorithms, accounting for 59% of the studies. When we refer to studies that use DL, we are considering architectures such as Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM), the combination of CNN and LST, (CNN + LSTM), Recurrent Neural Networks (RNN) and Gate Recurrent Unit Network (GRU) network. Conversely, other algorithms employed in the reviewed studies, such as MLP, have been excluded from the DL group due to a low number of layers used (from 1 to 3), which falls outside the concept of deep architecture.

DL models offer the advantage of dealing efficiently with raw ECG signals in various information domains such as time and frequency domains, providing a potentially more comprehensive and deeper learning of 'hidden' or 'embedded' information by capturing complex data representations [40,52,62,81,92,104]. Another common characteristic of all the presented DL methods is their ability to preserve temporal variation of the signal for both short and long-term learning, which is considered necessary for efficient classification of time-series data. DL may also use transfer learning, which means that models trained with ECG signals can serve as basis for application to other cardiovascular signals such as PPG and vice versa [31,62]. This review has also identified DL architectures used as end-to-end approaches, which eliminate the need to preprocess the ECG signal before fitting it into the model classifier [49,53,57,62,81,96,104], and preventing the model from any random or systematic error in calculating features that could be propagated to the output and limit the model's accuracy.

As depicted in Fig. 6, the CNN is the most frequently used architecture in the works reviewed since the convolution operation, which constitutes the basis of this architecture, is a classical well-known and computationally efficient technique in signal processing for signal enhancement [131]. The use of CNN is extensive, and it can be applied across different stages of the CVD prediction pipeline ranging from an end-to-end architecture that performs the classification directly from the ECG signal, to the extraction of features from ECG signals that are then fit to other classifier algorithms. In this context, a hybrid combination of CNN and LSTM is often found in the reviewed works, where the former is typically dedicated to extracting ECG features and the latter performs the actual classification task [44,62,85,114]. Moreover, while most papers that use CNNs consider one-dimensional input per lead of the ECG signal (i.e., signal intensity in volts over time), other papers leverage the well-known capabilities of CNNs for image processing, and transform the ECG signal into a time-frequency spectrogram (that is then treated as a 2-dimensional image) that is used as input, achieving a better performance than using the 1-D timeseries signal [40,44,52,62] since it allows a time-frequency domain learning. This 2D spectral-longitudinal modeling approach offers a more general representation that can overcome issues of variability in sampling rates, noise sensitivity and ECG monitoring devices from different manufacturers.

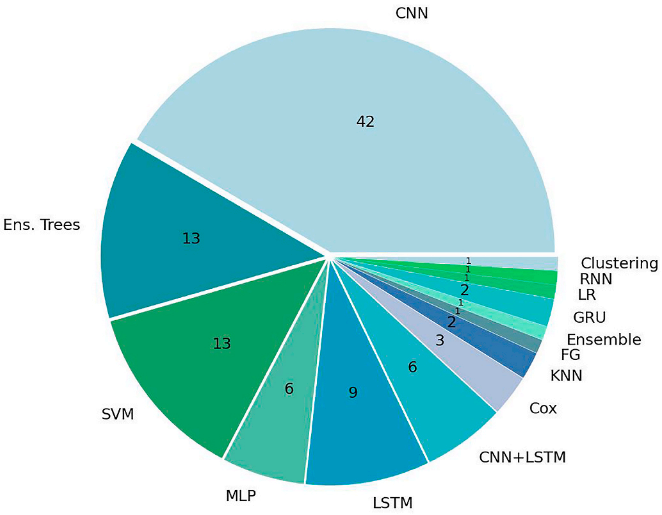

Fig. 6. Distribution of ML algorithms in the reviewed bibliography (CNN: Convolutional Neural Networks, Ens. Trees: Ensemble Trees, SVM: Support Vector Machine, MLP: MultiLayer Perceptron, LSTM: Long-Short Term Memory, Cox: proportional hazards model and regression, KNN: K-Nearest Neighbor, FG: Factor graph, Ensemble: Ensemble of Random Forest, SVM, KNN and Boosting, GRU: Gated-Recurrent Unit, LR: Linear Regression).

A range of DL algorithms distinct from CNN, such as LSTM, GRU and RNN, are used for classification purposes in a significant number of studies [27,35,52,61,64,74,89,90,101,101,114]. These algorithms, thanks to their ability to learn from temporally dependent information of fixed or variable-length sequences, offer substantial classification performance in detecting patterns in the time domain of the ECG signal.

Due to the dominance of the CNN algorithm and other DL approaches such as LSTM in the reviewed research works, a further analysis of these methods was conducted (see Supplementary Table 2). The most representative works in the different areas of the CVD field are shown in Table 3. This table describes the classification or regression goal, the DL architecture, performance, input data of the models, and the strengths and limitations identified by the authors.

While DL is used as the prediction algorithm in many cases, other methods such as, Ensemble Trees (13%) [32,42,45,56,58,65,67,74, 93–95,101,111], and Support Vector Machines (13%) [19,21,26,34,35, 54,59,61,76,86,90,102,116], and Multilayer perceptron (6%) [60,63, 66,78,79,122] are also employed. These algorithms can be applied to either hand-crafted ECG features, or to CNN-extracted ECG features. Apart from these crafted ECG features, shallow ML allows the combination with other types of data (clinical data, demographics) as shown in section B.3. DL models could also operate on this type of structured, tabular data (i.e. clinical data and demographics combined with ECG features), however, there is no significant improvement over shallow ML models because, in these cases, the data complexity is not high enough to leverage the capabilities of DL [131].

The use of these less complex and shallow ML algorithms, despite a theoretical decrease in performance when having ECG signals as inputs, brings some other advantages that make them preferable over the DL. Shallow ML algorithms do not require of large datasets for training, unlike complex DL architectures, allowing researchers to develop models with modest dataset sizes. The computing capabilities demanded by DL are much larger than other shallow ML algorithms, making the latter suitable for integrating predictive models into low-power embedded devices. Another important aspect to consider that can make ML algorithms preferable over DL is the 'black-box' paradigm that DL suffers from. Even though some ML algorithms (ensemble trees, SVM) are not considered fully transparent, we can obtain explanations about their decisions by analyzing the intrinsic logic. In addition, CNNs are vulnerable to adversarial perturbations of input data (even minimal

Table 3

List of most representative reviewed works that employed Deep Learning architectures to develop the models. The table contains the CVD goal targeted by the models together with the type of supervised learning problem (classification/regression), the characteristic of the deep learning architecture, the model's performance, the input data used by the model, and different strengths and limitations highlighted by the paper's authors. MAE: Mean Absolute Error, ACC: Accuracy, SENS: Sensitivity, PREC: Precision, avg: average, Bal ACC: Balanced accuracy, SPEC: Specificity, AUC: Area Under Curve Receiver Operating Characteristic, MI: Myocardial Infarction.

Author	Model's CVD goal	Network Architecture	Performance	Input	Outcomes (Strengths/Limitations)
Baek et al. [104]	Systolic/Diastolic blood pressure prediction (Regression)	1D convolution neural network composed by stacking 4 groups of one extraction block and then one concentration block for processing both time and frequency. The outputs are combined into a new convolutional layer. Extraction blocks are based on dilated convolution, meanwhile concentrating blocks are on strided convolution.	MAE (systolic) 5.32, (diastolic) 3.38	Raw ECG (PPG) signals	*Strengths*: End-to-end approach that does not require hand-made features. Flexibility to handle different raw signal inputs (PPG/ECG, time, frequency). *Limitations*: Data employed from MIMIC II presents a low sampling rate, below the minimum recommended (1000 Hz). Need to validate performance with a higher sampling rate. Results are subject to population bias in terms of age and BP values due to data collected in the ICU.
Butun et al. [96]	Automated detection of coronary artery disease (Classification)	1D – capsule network (CapsNet). A former CapNet for processing image data (Primary Caps with 32 different capsules) is updated to work with ECG signal by stacking additional layers (ECG Caps with two caps). A decoder is connected to the ECG Caps for reconstruction loss of the ECG signal.	ACC: 0.994, (2-s), 0.986 (5-s)	ECG signal of 2 s Long and 5 s Long	*Strengths*: End-to-end adaptation of CapsNet to a 1D signal (ECG) domain to detect ECG segments associated with coronary artery diseases in 2s and 5s duration. The 1 d-CADCapsNet shows high performance for relatively small data mainly based on captured information of R peaks and T-wave signals. More robust than CNN to adversarial samples. *Limitations*: Due to its architecture, the 1D-CapsNet presents a computation cost higher than a CNN
Dai et al. [57]	CVD diagnosis (classification)	A CNN variant of Residual Network (ResNet) that introduces shortcut connections between consecutive convolutional layers, besides the original data flow. 12-lead ECG signals that are segmented in different intervals (1s for training, and 2s and 3s for testing)	ACC: 0.998 (3 s), SENS: 0.995 (3 s), SPEC: 0.999 (3 s)	12-leads ECG signal	*Strengths*: End-to-end approach only applying simple min-max normalization to short-term duration ECG signal as preprocessing. Low computational complexity of data preparation. Use of Focal Loss as a loss function to address data imbalance. *Limitations:* The dataset employed presents issues due to varied sampling frequency as well as disease subject imbalance. The short-term duration approach avoids detecting more specific CVD diseases.
Haleem [40]	CVD diagnosis (classification)	A two-stage multiclass algorithm, where the first stage performs ECG segmentation based on Convolutional Bidirectional Long Short-Term Memory neural networks with an attention mechanism for segmenting the ECG signal. The second stage is based on a time adaptive (that controls the duration of the ECG time) Convolutional Neural networks applied to the 2-D ECG spectrogram beats extracted from the first stage for several time intervals.	ACC (average): 0.989	2-leads ECG signal	*Strengths*: A time-adaptive CVD detector via ECG beats segmentation through an automated approach and controlled by time. The use of time-adaptive spectrograms as input for classification improves the CVD detection accuracy. *Limitations*: the use of a smaller amount of training data produces overfitting problems when using 2D-CNN architectures along with a large amount of memory to handle spectrograms of the time adaptive ECG beat extraction. A down-sampling made to address different ECG sampling frequencies and data imbalance might affect the model's accuracy.
Jangra et al. [82]	Arrhythmia Diagnosis (Classification)	*O*–WCNN: improvement over traditional CNN models by implementing a multi-channel model to concatenate spectral and spatial feature maps; and a structural unit composed of a depth-wise separable convolution layer followed by activation and batch normalization layers	ACC (avg): 0.994	3-beat ECG	*Strengths:* An effective integration of spectral and spatial features that allows an efficient utilization of model parameters in depth-wise separable convolution layers. This optimized multichannel 1-D CNN outperforms arrhythmia classification of other state-of-art methods. *Limitations:* The approach requires a DWT decomposition and further reconstruction (not an end-to-end approach). The use of spectral features requires a greater computation complexity than other state-of-the-art methods. Compromised generalizability due to the presence of signals in training and testing sets belonging to the same patient.
Li et al. [20]	Arrhythmia Diagnosis (Classification)	DeepECG system based on transfer learning of Inception-V3 with 11 inception modules (2D ECG images).	Bal ACC: 0.984, SENS: 0.954, SPEC: 0.967	1-lead ECG signal (Lead II)	*Strengths*: Analysis of ECG images (pdf format) via DCNN without requiring the 1D ECG signal providing higher scalability

(*continued on next page*)

Table 3 (*continued*)

Author	Model's CVD goal	Network Architecture	Performance	Input	Outcomes (Strengths/Limitations)
		Each inception module includes 4 convolution groups composed by pooling with a single 1 × 1 convolution, 1 × 1 convolution with three 3 × 3 convolutions, 1 × 1 convolution with three 5 × 5 convolutions and a single 1 × 1 convolution. The results of 4 convolution groups will be concatenated.			due to the greater availability of ECG images. The method is suitable for potential deployment in handheld devices where a photo of the ECG works for CVD diagnosis. *Limitations*: Performance is subject to the quality of ECG images due to a lack of preprocessing noise filtering of the images.
Ma et al. [29]	Atrial Fibrillation detection (Classification)	A Deep Neural Network composed of dilated casual convolutional blocks. Each block is composed of a dilated causal convolution, a weight normalization, an activation function, and a dropout layer. In addition, a convolutional block joins the shortcut connection.	ACC: 0.986, SEN: 0.987 SPEC: 0.990	1-lead ECG signal	*Strengths*: The employment of dilated causal convolution effectively improves the training speed and classification accuracy. The method can be appropriate for real-time diagnosis of ECG signals. *Limitations*: The dataset presents a high imbalance of the atrial flutter (2% of the subjects) which might affect its performance in a real setting,
Ramesh et al. [31]	Atrial Fibrillation detection (Classification)	One-dimensional deep convolutional neural network that uses HRV-derived features and utilizes the knowledge transfer paradigm for cross-domain generalizability by training a model on ECG databases and adapting the developed model for PPG signals-based AF classification.	ACC: 0.955, SENS: 0.945, SPEC: 0.960, F1: 0.933, AUC: 0.953	ECG and PPG signals	*Strengths*: The method proposes supports potentially a seamless adaptation of gold-standard ECG-trained models for non-ambulatory AF detection with consumer wearable devices which is sensor-agnostic. A method based on transfer learning to address challenges associated with PPG datasets availability. *Limitations*: The dataset employed shows overlapping between each class's samples which affects its reliability. The models do not consider spectral and non-linear HRV features that would account for a more robust representation of each class.
Tadesse et al. [62]	Myocardial Infarction identification and occurrence (Classification)	DeepMI: Recurrent neural networks (a Dense-LSTM). First, a transfer learning module to obtain deep features, and then three diagnosis modeling techniques: spectral and longitudinal and joint spectral-longitudinal.	AUC:0.967 (normal MI), 0.829(acute MI), 0.686 (recent MI) and 0.738 (old MI).	12-leads ECG signal	*Strengths*: The joint spectral-temporal modeling allows for overcoming issues of variability in sampling rates and ECG device specifications. Model prediction based on a large-scale dataset containing >323 k data samples. The end-to-end framework provides flexibility for different levels of multi-lead ECG fusion and performs feature extraction via transfer learning. *Limitations*: The data fusion spectral-temporal seems ineffective in distinguishing the leads which might lead to independent modeling for each lead.
Tan et al. [85]	cardiovascular monitoring for COVID-19 patients (Classification)	CNN combined with LSTM. CNN is used for feature extraction and LSTM for the classification of AF type. Data enhancement is carried out through SMOTE	ACC: 0.992, SENS: 0.977, SPEC: 0.995	ECG signal from wearable device	*Strengths*: The use of 5G plus Flink framework to ensure low latency and high throughput of ECG signal data transmission in wearable devices. The combination of CNN and LSTM provides better generalization. *Limitations*: Insufficient number of heartbeat types in the dataset that leads to the use of SMOTE that might imply a deviation in the AF prevalence
Vijayarangan et al. [118]	R Peak detection in noisy ECG (Classification)	RPnet: A novel application of the Unet (Encoder-Decoder) combined with Inception and Residual blocks to perform the extraction of R-peaks from an ECG (it has been adapted from the INcResU-Net network)	F1: 0.9837	1-lead ECG signal	*Strengths*: The R-peaks in the ECG can be obtained through minimal post-processing due to the distance map generated by the architecture. *Limitations*: Relatively poor predictive power at lower SNR levels. The computational complexity is significant which avoids the real-time usage.
Zhang et al. [39]	Arrhythmia Diagnosis (Classification)	1D-CNN network consisting of 34 layers. 4 stacked residual blocks are used to extract deep features. Within each residual block, there are two 1D convolutional (Conv1d) layers, two batch normalization (BatchNorm1d) layers, 1 dropout (Dropout) layer, and two rectified linear unit (ReLU) activation layers.	ACC (avg): 0.966, SENS (avg): 0.812, PREC (avg):0.821, F1 (avg): 0.813, AUC: (avg) 0.97	12-leads ECG signal	*Strengths*: Explainability analysis of the model's prediction to both patient and population levels. Detection of the top-performing leads for the diagnostics classes which are I, aVR, and V5. *Limitations*: Likely population bias since data was collected from China hospitals, thus, further validation is needed to test the model's robustness.
Meng [30]	Arrhythmia Diagnosis (Classification)	A deep CNN takes the input as a two-dimensional image and implements 3 convolutional layers that contain 32	ACC: 0.856	8 leads ECG signal	*Strengths*: The translation of the starting point after leads filtering to increase the training sample improves the accuracy.

(*continued on next page*)

Table 3 (*continued*)

Author	Model's CVD goal	Network Architecture	Performance	Input	Outcomes (Strengths/Limitations)
		feature surfaces. Finally, two fully connected layers are adopted for binary classification.			The performance of V5 is higher than the other leads. *Limitations*: Due to the limitation of the dataset employed, only 4 arrhythmia classes prediction is made causing certain deviation in the results.
Zhou et al. [114]	ECG quality assessment (Classification)	Two CNN branches (2 convolutional layers each, 1st 128 and 32 filters, 2nd 64 and 16 filters) and an LSTM branch (2 layers, 200 and 100 units each). The three branches are concatenated into a dense layer. Conditional Generative Adversarial Network (GAN) is used for data augmentation	ACC: 0.971, SENS: 0.986, SPEC: 0.964;	1-lead ECG signal	*Strengths*: The system shows better generalization ability due to non-relying on manual feature extraction and rules for decision-making. The use of CGAN for data augmentation outperforms down sampling strategies by generating ECG signals with a larger diversity. *Limitations*: The ECG quality assessment proposed does not consider specific applications, and the same ECG can be acceptable/unacceptable for different purposes (e.g., HRV time-domain analysis/AF detection). The quality assessment only consider 1 EC G lead which makes it suitable for bedside monitors or wearable devices but not for CVD detection which requires 12 leads. The system cannot quantify the quality through a signal-to-noise ratio.
Zhu et al. [70]	Normal and Abnormal ECG detection (Classification)	A CNN-FWS network that combines three convolutional neural networks trained independently (CNN) and recursive feature elimination based on feature weights (FW-RFE). A final fully connected layer concatenates each RFE output module.	SENS: 0.889, F1: 0.902	12-leads ECG signal	*Strengths:* By selecting the most relevant features while eliminating unrelated and redundant features, the diagnostic efficiency of ECG abnormalities is improved. The methods show significant robustness when the amount of data decreases. *Limitations:* The presence of noise features in the ECG might affect the diagnostic capability of the model.
Doldi et al. [92]	Identification of congenital and often concealed LQTS (Classification)	CNN network based on four XceptionTime modules are connected in series to capture both the temporal and spatial information of the multivariate signal. The model is based on a simultaneous analysis of multiple leads and different-sized kernels to address both long and short-time intervals.	BalACC: 0.911	12-leads ECG signal	*Strengths*: In contrast to other same-goal studies, the model validation has used a more generalizable and comorbidly diseased control cohort. The model shows a significant stability over 25-fold CV in detecting LQTS patients including a large amount with a concealed phenotype. *Limitations*: Being a single-center validation, the model needs external validation by other heart rhythm clinics. Additionally, there could be potential selection bias in the control cohort since the selection was based on a missing suspicion for LQTS without genetic testing.
Diamant et al. [77]	Impaired Heart Rate Recovery (HRR) detection (Regression)	A 1-dimensional CNN based on the DenseNet architecture	Pearson correlation with actual model r5: 0,48	3-leads ECG signal	*Strengths*: The model achieves an estimation of impaired HRR independently associated with future clinical outcomes, including new-onset diabetes and all-cause mortality. The ECG resting could be used as a proxy for screening HRR which usually requires exercise provocation. *Limitations:* The correlation with HRR actual and predicted is modest, leaving the model subject to misclassification. The model was developed with 3-leads, thus, using 12-leads the performance might improve. The sample used could be subject to selection bias due to participants were recruited based on clinical risk factor assessment that impedes some of them from undergoing exercise testing.
Toma et al. [44]	Arrhythmia Diagnosis (Classification)	The model proposes a novel parallel cross-convolutional recurrent neural network consisting of two branches: a recurrent neural network (RNN) and a 2D CNN for temporal characteristics and spatial features that take CWT scalogram and segmented ECG sample.	ACC:0.997	1-lead ECG signal (lead II)	*Strengths:* The model presents significant robustness through the cross-function of the features that makes suitable for imbalanced ECG signal classification, especially for the AAMI standard-based classes, which are largely skewed. The model can effectively learn temporal

(*continued on next page*)

Table 3 (*continued*)

Author	Model's CVD goal	Network Architecture	Performance	Input	Outcomes (Strengths/Limitations)
					characteristics and rich spatial information of raw ECG signals. *Limitations*: The model presents a high training time since the number of training parameters is 34 million.
Yoo [49]	Arrhythmia Diagnosis (Classification)	1-D CNN (xECGNet) based on Attention Branch Network (ABN) which is fine-tuned with L2-Norm to reflect features of all concurrent ground truth (GT) labels through multi-loss optimization. The fine-tunning adds to the objective function the L2-norm between the model-produced attention map and a reference map created by averaging the response maps of all GT labels.	ACC (multilabel class): 0.846, F1-score: 0.812	12-leads ECG signal	*Strengths*: Addressing two main problems in medical AI: multilabel classification (concurrent arrhythmias) and explainability by using ABN with attention maps fine-tunning. *Limitations*: The use of zero-padded inputs for training might affect the efficiency to detect arrhythmia in real-life
Yao [50]	Arrhythmia Diagnosis (Classification)	BiLSTM-Treg algorithm composed by a BiLSTM to select the optimal heartbeat segment length, and next, a tree regularization model to optimize the BiLSTM and improve classification performance.	ACC (avg): 0.993	1-lead ECG signal (lead II)	*Strengths*: the Tree regularization outperform traditional L1 and L2 regularization, and provide an explainability component to the network by denoting the feature's relevance. *Limitations*: the interpretability information is limited since it is based on specific signal points
Radhakrishnan [52]	Atrial Fibrillation Diagnosis (Classification)	2D deep convolutional BiLSTM to detect and classify AF episodes using the time-frequency images of ECG signals as inputs. The network also consists of four convolution layers, one BiLSTM layer, and three fully-connected layers	ACC:0.991, SENS: 0.991, SPEC: 0.991, (Overall metrics over 6 categories)	2-leads ECG signal	*Strengths*: two-fold detection of Normal-AF and terminating and non-terminating AF. Use of 2D CNN to decipher subtle temporal and spatial correlation of ECG signal in the time-frequency plane. This approach is computationally faster than multiscale fusion-based deep CNN. *Limitations*: The terminating-non terminating AF accuracy can be improved since only 20 EC G records have been used

such as a single pixel in an ECG) that might become a barrier for widespread adoption and implementation as opposed to shallow ML algorithms that are based their decision in previously calculated features.

In the present analysis of the ML/DL algorithms performance, it should be noted that the results are dependent on the CVD and the clinical goal (diagnosis or prognosis), thus, a direct comparison between studies is not feasible. Nevertheless, certain insights can be gained by examining the performance metrics used and their values. Over 90% aimed to address classification problems, whereas regression (7%) or clustering (1%) problems were less frequent. Consequently, performance metrics such as accuracy, sensitivity, specificity, precision, F1-score, and AUROC were predominantly employed. Studies dealing with regression problems utilized a variety of metrics, including mean absolute error (MAE), *C*-index value, Sum of Square Distance (SSD), Maximum absolute square (MAD), Percentage of root distance (PRD), and Cosine similarity.

In the reviewed works aimed at CVD, classification accuracy emerges as the most prevalent metric, used in 70 out of 101 studies. However, its exclusive use in 25 out of these 65 studies [18,19,25,26,40,43,44,54,56,65–67,78,79,81,81,82,90,92,96,101,105,111,115–117], potentially leads to an incomplete model assessment. This is because metrics like sensitivity and specificity, which provide critical insights into false positives and negatives, are often overlooked. This oversight is particularly concerning given the frequency of imbalanced datasets in CVD research ([20–22,38,38,39,43,44,47,49,53,54,57,58,64,71,80,81,84,92,102,118]), yet only a few studies ([20,92]) employ balanced accuracy. The second most used metric is sensitivity (54%), followed closely by specificity (35%), and more sporadically F1-score (26%) and precision (18%). Only 16 works [23,24,30,31,36,38,39,45,58,59,62,64,76,93,97,110] include AUROC in their metrics, despite the comprehensive information it can offer. Given these findings, we advocate for a more holistic approach in selecting performance metrics. Metrics such as accuracy and precision, while informative, can be influenced by systemic limitations like class imbalance. Thus, a comprehensive utilization of other widely used metrics such as recall/sensitivity, specificity, f1-score, AUROC, Positive Prediction Value (PPV), and Negative Predictive Value (NPV), ensures a more rounded and accurate assessment of the model's performance in varying clinical contexts. In addition, incorporating additional metrics including Matthews Correlation Coefficient (MCC), Area Under the Precision-Recall Curve (AUPRC), Cohen's Kappa, and Youden's Index can significantly enhance the evaluation of model effectiveness.

An important factor affecting the comparability of the results, and especially the evaluation of their generalizability, is the type of validation used. The most widely used cross-validation approaches in the analyzed papers were 5-fold, e.g. Refs. [31,41,42,46,49,53,73,77,81,92,94,96], and 10-fold, e.g. Refs. [26,32,44,45,52,57,58,63,71,72,82,91,102,111,123], while leave-one-subject-out cross-validation (LOOCV) was not used despite its relevance for patient-centric data, especially when limited data is available. LOOCV is critical in medical datasets where individual patient variability is significant since it ensures a comprehensive evaluation of the model's performance across diverse patient cases. Several studies, e.g. Refs. [24,33,34,38,40,42,44,45,49,51,53,56,57,61,75,80,97,106,108,110,114], used a separate hold-out testing dataset for validation where the hold-out set was separated from the original data before performing the training and cross-validation phase. However, we note that overoptimistic generalization capabilities are not often mitigated since an independent dataset for testing is not employed. Even more realistic results can be obtained if the final model is tested with data from a completely independent setting, e.g., by using external retrospective datasets or by performing a prospective data collection. An external retrospective dataset was used in studies, e.g. Refs. [33,34,38,42,97,110,114], however, none of the studies performed testing with a prospective clinical investigation. An additional critical aspect, often overlooked, is the treatment of different ECG data segments as independent observations, even if they were recorded from the same person. This may, in the worst case, result in the situation that data from the same person is present in both in the training as well as validation set. There is a significant lack of reporting on this aspect that might lead to biased model performance. Only a few studies explicitly address a patient-wise grouping for training and testing sets [77,92].

3.4. Trustworthiness of the AI models

1) Explainability aspects

In the majority of the reviewed papers, the algorithms employed by the prediction models rely on techniques which suffer from a lack of transparency in the sense of understanding their decision-making logic, such as CNN, LSTM, or to a lesser extent SVM, Ensemble Trees. In response to this challenge, the application of XAI offers mechanisms to address this barrier by improving the understandability of the AI models that will revolve around the adoption by clinical experts. However, despite the importance of XAI for those CVD prediction models, we have identified a low number of works tackling XAI (18 out of 101 works) [34,39,41,45,49,51,53,58,76,77,93,101–103,106,108,109].

In addition, it is important to highlight that the XAI techniques employed in the reviewed works are mostly post-hoc methods, i.e., they are applied on an already trained model to interpret its prediction. This is due to the fact that the majority of ML/DL algorithms used in these works are considered non-transparent requiring ancillary techniques to explain their decision-making processes. The 18 works implementing XAI techniques in the context of CVD account for the use of specific XAI techniques depending on the algorithms used. As a result, among the post-hoc XAI techniques employed, model-specific techniques such as Class Activation Mapping (CAM) [41,49,53,106,109] and Saliency Maps [77] for CNN, or Feature Importance for Random Forest [108] were utilized. The model-agnostic XAI post-hoc technique such as SHAP was implemented in seven of the works for any of the ML algorithms mentioned above [39,45,45,58,76,93,103].

From the perspective of the models' input data, the use of the XAI techniques can be divided into two categories: (i) the relevance of the ECG-calculated features used in the AI model that contribute to the prediction, (ii) and the relevance of those ECG signals segments or leads that influence the AI model's decision. Regarding the relevance of the features, these explanations, expressed quantitatively, are derived either from those model-specific XAI methods that are implicit in ensemble trees [108], or also from applying the SHAP method to CNN [58,76,93]. When using this feature relevance information, the authors focused on providing general explanations aimed at describing how the features contribute to the decisions of the whole model. However, only Ibrahim et al. [58] show local explanations by offering information about the relevance of the features for individual predictions. On the other hand, those explanations concerning ECG signals or leads which use SHAP [39,103] or Class activation mapping [49,53,106,109] are offered exclusively for local predictions. Thus, the type of input data explanation seems to condition the generalizability of the explanations, since when raw ECG signals are used at prediction models' entrance, XAI can only provide individual explanations over a single subject's ECG recording. Table 4 shows a detailed view of the explainability information provided about the models' logic in the works that consider XAI.

There are some limitations associated with the different XAI techniques. For instance, for those based on heatmaps that localize the regions of the ECG signal used by the model in its classification output, apart from localization accuracy, the presentation of the heatmap might not be well understood by clinicians and fail to help with making an evidence-based diagnosis. In addition, even though individual explanations help to understand the relations made by the prediction models between specific input instances and their corresponding outputs, these explanations are only valid for a single input instance and sometimes lack stability [132].

As said, the number of works tackling XAI to address the black-box paradigm brought by DL models remains quite low with roughly 18% of the reviewed research works. This highlights the need for further research into XAI in the context of CVD, as it could potentially increase the clinical relevance and acceptance of AI models in this field. Additionally, we have also identified a significant absence of proper validation of the XAI techniques used, which assesses aspects such as acceptability, usefulness, correctness, etc., while only Wang et al. [51] and Yoo et al. [49] include cardiologists' feedback in their models. This shortcoming could present a significant limitation to the eventual clinical adoption of the AI models aimed at CVD. The evaluation of the explainable results becomes essential to guarantee the eventual use of the AI models by clinicians since it revolves around the trustworthiness experienced by the professionals, to advance adopting and applying the results in their clinical practice. Moreover, in the general XAI domain, most researchers claim their explanations to be sufficient to understand the analyzed black-box model without considering validating their methods by human experts in the field [133]. Thus, researchers must be aware of the importance of addressing this issue, and propose different explainability evaluation approaches, preferably considering the clinical expert in the loop or applying generic metrics available in the XAI research literature [134].

2) Bias and Ethical aspects

In the reviewed papers, the term bias, which carries a negative connotation, is mentioned in a general sense, as technical bias, as social bias, as prejudices, and in the context of discrimination. Most often, references to bias occur because of technical reasons, i.e., objective differences between true values and estimates. Bias concerning technical circumstances extends to model fit in general (with underfitting and overfitting), to classification, sampling, selection, and to data challenges, such as incomplete or missing data and imbalances between majority and minority class instances. Twenty eight (28) papers refer to bias as part of their development work or model descriptions [23,31,33,37,40,43,57,60–62,64,69,70,72–74,76,82,86,92,99,100,107,109,113,117,118,122]. In contrast, social bias is mentioned less frequently (n = 12 papers) [4,7,23,39,60,62,65,105,117,119,121,124]. Social bias refers to the representation of subjective cognitive bias from humans when using AI applications. Psychological distortions of perception and judgment are not relevant for the selected publications themselves because the authors mainly use data from databases and do not collect extensive new data. Three papers specifically mention bias due to individual human input [62,105,119], such as by annotations. References to social prejudices [124] and discrimination [7] occur rarely. Two papers [23,60] actively consider multiethnic populations as their target groups. Bias in the sense of social prejudices has only rarely been recognized as a limiting factor for the practical use of AI tools in the health care system. According to the investigations in this study, bias is more frequently considered as a technical term, with relationship to models and calculations. The social aspects of bias are more infrequently addressed.

Ethical, legal, and social implications (ELSI) include a wide range of topics, for example ethical values, specific legal norms, human interaction, user orientation in design, and societal development. These aspects are relevant for the acceptance and acceptability of technology in society and deserve greater attention than what they have received thus far. ELSI are specifically addressed only in a small minority of the reviewed papers. Five papers mention some ELSI [7,18,105,119,124], and only one paper refers to them within a dedicated section. One conclusion from this dedicated ethical section on ELSI states that the misuse of patient data should be avoided. Misuse may, e.g., occur when patient data are extracted from hospital procedures with the intention of monetization by private companies. The expanding power of algorithms poses a threat to measures of de-identification [124]. Only one study specifically mentions the participation of physicians during the algorithmic processes for personalization purposes [105]. The papers do not

Table 4

List of reviewed works where XAI application is reported. The table contains the ML algorithm and XAI technique utilized in the model, the type of XAI technique (model intrinsic/ model agnostic) the purpose of the XAI explanation (local and/or global), and the information provided by the XAI techniques. CNN: Convolutional Neural Networks, RF: Random Forest, SVM: Support Vector Machines, XGBoost: Extreme Gradient Boosting.

Authors	ML algorithm	XAI Technique	Model intrinsic/ Agnostic	Global/Local explainability	Explainability information provided
Chang et al. [106]	CNN	Class activation mapping (CAM) and attention mechanism	Intrinsic	Local	The ECG segment is highlighted for each ECG-age prediction, indicating the importance of each lead (unimportant/important), the contribution of each position, and whether the ECG segments mean younger or older rhythm.
Liu et al. [109]	CNN	Gradient CAM (Grad-CAM)	Intrinsic	Local	Grad-CAM allows finding which parts of the multi-lead ECG dominate the output score by showing the log odds for the different leads as well as the precise localization of abnormalities in the ECG.
Fayyazifar et al. [41]	CNN	Grad-CAM	Intrinsic	Local	The application of Grad-CAM highlights the principal ECG regions of the 12 leads that the model uses to extract the discriminative information to diagnose wide QRS complex tachycardia in an individual prediction.
Diamant et al. [77]	CNN	Saliency maps	Intrinsic	Global	Saliency maps demarcate the areas of the ECG waveform of greatest influence on heart rate recovery predictions. The displayed saliency maps represent the average of 200 individuals and are grouped based on resting heart rate.
Agrawal et al. [103]	CNN	SHAP	Agnostic	Local and Global	XAI information is provided at two levels: i) patient-wise (local explanations) by highlighting the ECG regions that positively contribute to the classification of a post-COVID; ii) lead-wise (global explanations), where the importance of each lead for each class (patient healthy/post-COVID) is given.
Zhang et al. [39]	CNN	SHAP	Agnostic	Local and Global	Explainability information is provided two-fold: i) at the patient-level by showing the specific ECG region associated with the classification of one out of ten types of arrhythmias. Furthermore, the use of SHAP allows providing explanations of misclassification cases. Ii) at the population-level, by employing the additive character of SHAP to show the contribution rate of ECG leads towards each diagnostic class.
Gorodeski et al. [108]	RF	Feature importance	Intrinsic	Global	The importance of the different ECG features along with patient demographics and clinical data is calculated based on their minimal depth across the Random Forest’s trees. Setting a threshold for the minimal depth allows obtaining the most 20 important out of 499 variables.
Angelaki et al. [93]	RF	SHAP	Agnostic	Global	SHAP provides the features’ importance in both binary and multiclass classification scenarios. Furthermore, feature interaction plots are presented to illustrate the interaction effect between specific pairs of features, such as age-sex, QTc duration-hypertension, and BMI/SL - S V5, in both binary and multiclass classification tasks.
Villa et al. [76]	SVM	SHAP	Agnostic	Global	For each feature, the SHAP values quantify its contribution to the prediction of a fragmented/non-fragmented signal. The XAI results indicate some issues in the datasets concerning overlapping between the ECG signal classes.
Wang et al. [34]	SVM	Surrogate model (decision tree)	Agnostic	Global	The Decision Tree Classification algorithm is used as a surrogate model to evaluate the importance of the ECG features in the diagnosis of atrial fibrillation.
Ibrahim et al. [58]	XGBoost	SHAP	Agnostic	Local and Global	Global explainability results show the features with the greatest influence and how their values affect the Acute Myocardial Infarction (AMI) diagnosis. A local explanation example is also shown regarding both cases with low and high probabilities of AMI.
Yang et al. [45]	XGBoost	SHAP	Agnostic	Global	The top ten most important features that contribute to the prediction of atrial fibrillation (AF)are quantitatively expressed as the global average of the SHAP values across all samples.
Yoo et al. [49]	CNN	Class activation mapping	Intrinsic	Local	The ECG segment is highlighted according to the attention maps’ information given by the network by showing where the model attends to upon multilabel ECG classification. The explainability output is evaluated by a group of cardiologists.
Yao et al. [50]	CNN	Decision Tree	Agnostic	Global	The tree regularization model is leveraged to build a simulated decision tree that offers decision information about the ECG signal’ points which are considered as model’s feature.
Wang et al. [51]	CNN	Human-in-the-Loop	Agnostic	Global	Human-machine collaborative knowledge representation where cardiologists adjust the human-interpretable part of the human-machine collaborative knowledge representation by observing the waveforms shape-changing characteristics of the ECG signal. The cardiologists can adjust the basis generated by the encoder providing an interpretable output decision.
Wang et al. [53]	CNN	Class activation mapping	Intrinsic	Local	The attention mechanism highlights the segment of the ECG that correspond to an anormal pattern associated to the arrhythmia.
Alizadehsani et al. [102]	SVM	Feature importance	Intrinsic	Global	The features selected by the best performing model to classify Coronary Artery Disease (CAD) are shown which are designated by the Weight-by-SVM method.

systematically acknowledge the relevance of ethical norms in society and the legal frameworks for new technologies, such as medical devices. While separate literature sources on ELSI for AI in health care exist, an integration of reflection on ELSI into more technically and clinically inclined studies, such as those under review here, may be worthwhile in order to arrive at acceptable applications through a responsible development process.

4. Discussion

CVD detection through the use of ML and DL brings promising results and is an active research area. ECG has become a crucial input in prediction models for CVD diseases that tackle a data-driven approach. The findings of this literature review demonstrate that ML and DL models show great potential in aiding the diagnosis and prognosis of CVD by utilizing the ECG as the primary input for the prediction and underscoring a significant advancement in the application of AI technologies in healthcare. The utilization of ECGs as a reliable and informative physiological variable may be a key factor contributing to the diagnosis effectiveness [121]. In particular, the algorithms used to diagnose arrhythmia and conduction disorders demonstrate encouraging outcomes compared to other non-ML diagnostic methods. Al Hinai et al. [3] show that DL models appeared to outperform other common ML models such as support vector machines, random forests, and logistic regression. Thus, initiatives to implement ML into ECG analysis systems should consider DL as a favorable approach due to their capabilities in persevering temporal variation of the signal which can occur both within the beats and over the beats. The implementation of DL algorithms for ECG analysis has the potential to enable clinicians and health care personnel to detect, based on short and long-term learning, previously unrecognized cardiac conditions that may have gone undetected or been diagnosed much later via specialist evaluations or echocardiography. Timely identification of CVD can result in earlier initiation of treatment and improved outcomes, while delayed or missed diagnosis can lead to poorer health outcomes.

In recent years there have been other literature review works intending to compile the latest trends in applying ML and DL techniques for the prediction of CVD. Ebrahimi et al. [135] and Musa et al. [136] offer an extensive systematic literature review of DL for ECG arrhythmia classification focusing exclusively on the DL architectures and the characteristics of the datasets used by the reviewed works to diagnose different manifestations of arrhythmias. Siontis et al. [119] also focus solely on DL architectures but extend the inputs to NLP solutions and also consider wearable and mobile ECG technologies. However, the inclusion of shallow ML algorithms is not addressed in these works despite their adequacy in certain situations as described in section C.2, especially when the amount of data cases is limited, which in practical healthcare settings often is the case. Other aspects distinctively considered in this manuscript as critical areas, such as explainable AI and ELSI, are notably unexplored by these literature reviews. In their review, Somani et al. [131] point out the lack of providing in the reviews an intuitive understanding for clinicians as well as a clinical perspective. Ayano et al. [133] provide the only literature review that specifically surveys the application of XAI in the models for classifying CVD using ECG. Similar to our work, they identify the challenges and limitations of the XAI techniques such as the need to validate the performance of explanations. However, despite carrying out a comprehensive review of the works that use XAI techniques, they only assess a limited period of 5 years and do not offer a detailed analysis of the DL methods employed by the work identified. Therefore, our contribution to the current literature is based on offering, to the best of our knowledge, the first systematic review that undertakes a holistic analysis across multiple dimensions of ECG-based data-driven models to predict CVD, encompassing publications from the last 15 years. We explore several areas of analysis that align with other literature reviews, such as the type of CVD diseases predicted, the nature and size of the input data, or the types of ECG features used in the algorithms. Nevertheless, the core of this paper presents a thorough and distinctive analysis where the ML/DL architectures are meticulously examined together with those works that have addressed key factors for the clinical adoption of these prediction models, such as explainability, bias, as well as ethical, legal, and societal issues. This holistic approach provides a vital reference for researchers, facilitating a comprehensive and nuanced understanding of the current landscape in this rapidly evolving field.

This review also highlights a key finding, which is that the most accurate results in ECG analysis have been consistently attained through the use of DL models that innovatively combine CNN with other neural networks. These models can effectively learn different types of functions in a single network, thereby leveraging the strengths of each component. One key advantage of end-to-end DL models is, provided enough good quality training data is available, their ability to automatically learn discriminative features from complex and heterogeneous inputs, such as ECG signals or radiographic images, without requiring prior definition, extraction, and processing of relevant features (i.e., feature engineering). Shallow ML models necessitate pre-processing and feature engineering, which can be a multi-step process that risks overlooking potentially informative features, but they don't need so much data for training and have a more understandable logic in their decisions. Nevertheless, in these latter cases, CNN can be used for feature engineering highlighting the versatility and significance of this type of DL network in CVD detection.

Despite the advantages that ML/DL bring to the field of ECG-based prediction models for CVDs, AI developers together with clinicians must carry out a careful design of the model where aspects such as computing capabilities, noise and spurious features, and algorithm accuracy must be addressed together to guarantee a minimal technological acceptance of the models in their ultimate usage. Additionally, both actors must confront a set of other non-technical aspects to achieve a satisfactory adoption of the ML/DL solutions in the clinical routine such as assessing the added value of the complex AI models over existing simpler models, evaluating the suitability of AI models in the existing clinical workflow, or considering the use of reporting guidelines (TRIPOD) to document the prediction model [8].

Additionally, we have identified several challenges in the development of AI systems aimed at CVD predictions, which could be classified into dataset issues, and ML/DL model development issues [102]. Concerning the issues with the datasets employed to build the models, the literature reports different shortcomings concerning data collection. For instance, the lack of information for many geographical regions impacts on the generalization capability of the models to tackle regional and racial differences, which can lead to unfair and biased predictions or even discrimination due to specific traits of the subjects. Other kinds of issues in data collection might arise due to, for instance, the use of devices from various manufacturers with different preprocessing methods and data handling approaches, or the restriction to a specific subset of a general CVD population (in-hospital, home care, etc.), which lead to the risk of overfitting and poor generalization. While models are frequently developed using top-tier databases with carefully acquired ECGs and thoroughly characterized patients, their performance may falter when applied to ECGs from everyday clinical settings in the real world. To circumvent these issues, the quality of the dataset must be prioritized because even the most refined model adjustments cannot compensate for a dataset of insufficient quality. Therefore, external validation in geographically diverse multicenter populations with multi-vendor ECG systems would be crucial not only for improving the generalizability performance of the model but also for enhancing the model uptake. Thus, the representativeness of the population targeted by the prediction model must include those individuals with an 'atypical presentation' in order to avoid bias in the predictive performance measures.

Continuing with dataset issues, in a significant number of research (30%) the sample sizes are quite low with less than 100 subjects, which limits the models' performances and questions their reliability and generalizability capabilities. These datasets were usually private and collected with the purpose of proving the research questions of one specific study. However, their small size might not cover a needed heterogeneity among the study participants to facilitate the deployment and adoption of the model up of the model in a wider population. When working with small datasets, DL methods tend to overfit their performance to the training data and struggle to generalize to unseen test data. Thus, in such scenarios, simpler, shallow techniques often yield superior results due to their ability to handle smaller datasets more effectively. Additionally, Typically, DL is computationally demanding with a significant memory requirement [137] implying issues and constraints on their deployment on low-power embedded devices.

Many open datasets offer a substantial number of subjects which allows the development of several prediction models (also more complex ones) to tackle different clinical endpoints of the CVD. While large datasets offer the advantage of increased statistical power, they also introduce the potential for greater heterogeneity within the sample. This is an important consideration for the internal validity of individual studies. Studies in the field of CVD prediction models often employ diverse sampling methods, inclusion criteria, and data collection techniques. As a result, the datasets used in different studies can vary significantly in both the number of samples and the features collected. This inter-study heterogeneity makes it particularly challenging to directly compare or benchmark the performance of different algorithms across multiple studies. Such variations introduce biases or inconsistencies that are not easily accounted for, complicating meta-analyses or comparative evaluations. Therefore, while large sample sizes within a study can indeed lead to a more diverse and potentially more representative sample, the heterogeneity across different studies poses a significant obstacle to the objective comparison of CVD prediction models.

All in all, to accelerate research efforts and enhance the precision and dependability of predictions, it seems imperative to prioritize the enlargement of publicly accessible annotated ECG datasets that meet requirements concerning dealing appropriately with data collection bias. This will provide researchers with the necessary source data to carry out comprehensive investigations. Additionally, the process of labeling the target outcome in those datasets needs to be ensured in terms of reliability, replicability, and independence. Nevertheless, the process of annotating data is essential but costly, which may contribute to the slow pace of collecting high-quality data.

Most of the reviewed works focus only on ECG signal characteristics and still exclude other important characteristics of patient data that might carry useful information for CVD prediction (e.g., age, gender, patient history, laboratory tests, imaging results, -omics etc.). This review has also identified a considerable amount of research that considers ECG as the sole information source. However, other works attempt to enhance the performance of their models by incorporating additional data sources such as clinical and demographic data, other biosignals, or medical images. Apart from the potential improvement in model performance, the latter approach presents challenges in integrating different data sources, necessitating the consideration of alternative DL approaches (e.g., perceiver networks [138] or ensembles combining DL and handcrafted features [139]).

Another shortcoming identified in the review concerning model development issues is the validity of the results when employing cross-validation in prediction models that use ECG segments as ‘independent inputs’, ignoring that they may come from the same subject. It is crucial to ensure that random splitting of the data avoids including segments from the same subject in both training and testing sets. The latter would introduce a bias that badly affects the performance of the model with unseen data.

To promote the performance comparison of the different models developed, the implementation of a benchmark environment would help to track their progress [121]. For instance, the works reviewed have not followed a common approach to the use of certain performance metrics, especially concerning classification problems. Therefore, the necessity of agreeing with a set of metrics to express the models’ results appears as an important point to achieve a comprehensive analysis. We advocate for using other metrics than accuracy alone, such as, sensitivity, specificity, recall, F1-score, AUROC, MCC or Kappa’s score - particularly when the dataset presents a significant imbalance in its target outcomes. In addition, to avoid falling into an optimism bias trap of predictive performance of the model, it is essential to test it through rigorous internal and external validation procedures. Furthermore, since there is no unified framework for benchmarking the performance of models across different institutions, creating an open platform that facilitates the sharing of ideas, datasets, and pre-trained model weights is challenging. However, it can pave the way for collaboration, breaking down the apparent barriers to institutional development.

Despite the development of a DL model, the informative features that contribute to its performance may remain opaque or “hidden”, which can make it challenging for clinicians to interpret or apply them in a non-computer-assisted setting. This lack of transparency could significantly impact the eventual adoption of the AI models by clinical experts who require explanations of the results. Additionally, regulatory organizations have become aware of this transparency issue and have included explainability as a requirement in regulations such as the EU AI Act [140] for AI models development. This measure aims to protect users against high-risk AI systems that might potentially pose significant harm to people’s health, safety, and fundamental rights, (e.g., those used in healthcare). The advent of XAI seems appropriate to overcome these interpretability barriers, making the models and their outputs more accessible to the users and helping them to identify which features contribute most to making predictions and exploring causal relations between features and clinical outcomes. Despite the importance of considering the understandability and trustworthiness of ML models to increase their adoption in the clinical routine, to the best of our knowledge, only a few research works have considered these aspects in models aimed at supporting CVD diagnosis or prognosis. Additionally, even when XAI is considered, another challenge arises like the absence of standardized measures to evaluate the performance of the interpretability method, which hinders the clinicians from selecting the best XAI technique for a particular problem as well as the researchers from comparing and improving the limitations of the techniques. The current metrics can be broadly classified into qualitative and quantitative. Qualitative metrics involve human-in-the-loop assessments, where the experts (i.e., clinicians) evaluate the correctness of the explanation by comparing it with clinical findings or assessing the suitability of the explanations to the clinical case. However, the quantitative metrics do not tend to involve the expertise of clinicians, and most researchers claim their proposed techniques sufficiently explain the prediction by giving quantitative metrics that are not reflected upon by human experts in the field or compared to a ground truth [141]. Moreover, to leverage the advantages that XAI offers and improve its adoption, there should be a user-driven design of the explainable information given by the XAI techniques to address the specific clinicians’ needs concerning the understanding of the AI model’s outputs and its application to the decision-making process for CVD diagnosis or prognosis.

Another challenge identified in this review highlights the importance of addressing issues related to transparency, bias, and other kind of aspects related to trustworthy AI or ELSI in the ML/DL models aimed at CVD. The prevailing focus on the classification and regression performance of the models, coupled with the neglect of these critical aspects pose significant barriers to their adoption in clinical practice and raises concern regarding regulation issues in their eventual deployment in the healthcare systems. Thus, identifying any kind of source of bias (algorithmic, unrepresentative data, sampling, prejudice, or discriminatory) and ensuring that patient data are handled ethically and securely in model development and deployment are essential to ensure that these models are both accurate and equitable in their predictions. In addition, since the model development often accounts for data exchange between research teams, there should be concerns for the protection of sensitive patient data subject to cyber-attacks or other threats. In practice, we suggest that ML/DL models aimed at predicting CVD should be aligned with the various Trustworthy AI requirements defined by the EU in their ethical guidelines [10] which accounts for aspects such as i) *"Human agency and oversight"* by adopting a human-in- the-loop multidisciplinary approach involving CVD domain experts and AI developers throughout the entire project life, and by evaluating and refining the human-AI interaction and oversight as a part of the co-development and integration process that foster the adoption of the models in the clinical routine; ii) *"Technical robustness and safety"* by implementing safety mechanisms (error-correcting codes, redundancy, data validation and filtering, fallback plans) that prevent and protect tools and model developed from adversarial attacks, recovering the well-functioning status in case of vulnerability, and continuous monitoring through robustness tests to identify and mitigate any weaknesses in the models; iii) *"Privacy and data governance"* where privacy regulations like EU General Data Protection Regulation (GDPR) [142] and data strategies principles such as Findability, Accessibility, Interoperability, and Reuse (FAIR) [143] are implemented; iv) *"Transparency"* by applying explainable AI techniques combined with transparency-by-design approaches in the development of the models as well as fostering the traceability of the decision within the models' inner logic, and communicating to the models' users (clinicians, patients) about AI-assisted decisions; v) *"Diversity, non-discrimination, and fairness"* by inspecting the data collection and curation paying special attention to any sources of bias and discrimination that could alter the functioning of the model, and by assessing its predictions to ensure the avoidance of any kind of bias, unfairness, and inaccuracies; vi) *"Accountability"* by enabling affected users to audit the logic behind the model decisions and implement redress strategies to mitigate potential harmful decisions made by the models.

Finally, the regulatory frameworks for integrating data-driven ECG diagnoses into direct clinical care are beginning to take shape. The hurdles for approval by regulatory authorities could differ from those for devices or drugs. They may be influenced by a clinician's capacity to review and interpret the AI-generated ECG results.

As limitations of this review, it should be noted that only papers published in 2007 or later were included, potentially resulting in the exclusion of some important earlier works. However, as we emphasized in the eligibility criteria definition, by focusing on the last 15 years, we are confident that the biggest and latest advances in ML and DL in the field of ECG-based CVD detection have been captured in this review. While efforts were made to provide clear guidelines for data collection based on the papers, there may be minor discrepancies in the interpretation of the information presented by the reviewers, which could potentially impact the review's findings. These limitations should be taken into account when considering the implications and generalizability of the review's results.

5. Conclusions

CVD currently possess a significant impact on worldwide healthcare systems. Timely diagnosis and prognosis of CVD could prevent adverse outcomes suffered by patients such as mortality, morbidity, and decrease of quality of life. ECG-related information analysis offers crucial information necessary for the identification and treatment of various CVD, but it requires skilled cardiologists to interpret and analyze the recorded data acquired. This systematic literature review highlights the significant potential of machine learning (ML) models in aiding the diagnosis and prognosis of CVD by utilizing ECG data as the main input. The reviewed works demonstrate promising results, particularly when using DL approaches such as CNN and LSTM networks. However, several important aspects require attention for the successful adoption and deployment of ML models in clinical practice. These include: lack of addressing explainability in the model's decision output, potential bias in data collection due to the location of subjects population, or type/brand of ECG monitoring equipment employed, limited validation with external datasets to ensure generalizability of the models, or a non-comprehensive use of performance evaluation metrics. This literature review introduces a novel focus on critical areas such as transparency, bias mitigation, as well as ethical, legal, and social considerations. The application of XAI techniques and meeting Trustworthy AI requirements can address these barriers by providing insights into the model's reasoning and by making it more accessible, robust, fair and understandable for clinical experts. Future research should focus on approaching XAI and TrustworthyAI requirements by validating the XAI results including the CVD experts, and addressing bias issues to ensure the ethical and responsible deployment of ML models in the healthcare domain. This review considered works published from 2007 onwards. By focusing on the last 15 years we pursued collecting the biggest and latest advances in ML and DL in the field of ECG-based CVD detection that also correspond with the biggest breakthrough of the AI algorithms in the health field.

Summary

Globally, cardiovascular diseases (CVD) stand as a principal cause of mortality, contributing substantially to morbidity and diminished life quality. The electrocardiogram (ECG) is integral to diagnosing, prognosticating, and preventing CVD, yet challenges persist, notably the escalating need for skilled cardiologists for precise ECG interpretation. This demand results in increased workloads and potential diagnostic errors. In response, machine learning (ML) and deep learning (DL) techniques have been developed to enhance computer-aided solutions and aid clinicians in deciphering the intricate mechanisms of CVD via ECG analysis. Nonetheless, these ML and DL models for ECG-based CVD detection often contend with issues of explainability, bias, as well as ethical, societal, and legal implications (ELSI). Despite the critical importance of these Trustworthy Artificial Intelligence (AI) aspects, there is a notable absence of exhaustive literature reviews focusing on the latest developments in ML and DL models for ECG-based CVD diagnosis and prognosis, particularly those addressing the essentials of Trustworthy AI. This review aims to bridge this knowledge gap by providing a systematic review to undertake a holistic analysis across multiple dimensions of these data-driven models. Following a set of defined review questions and adhering to the PRISMA methodology, a total of 101 research studies were analyzed across different perspectives such as the type of CVD diseases predicted, the nature of data, and dataset sizes used as inputs Moreover, this review provides a thorough analysis of the DL models, detailing their architectures, performance metrics, and identifying their main strengths and limitations. Explainability and ethical aspects are essential in healthcare applications, and are thus part of the core of this review; we include a dedicated section to describe the sparse works that have addressed these issues. For each analyzed area, various challenges are identified, and we discuss, from different perspectives, how they impact the use of the models in diagnosing and prognosticating CVD. Finally, a thorough discussion is provided highlighting a range of challenges and limitations within this domain, and we subsequently provide specific suggestions to tackle these concerns. Within the discussion, considering that the adoption of these models by healthcare professionals hinges on fulfilling Trustworthy AI requirements, we suggest various approaches to address principles such as explainability, bias, robustness, fairness, as well as human agency and oversight. This systematic review delivers a comprehensive examination of ML/DL models using ECG for predicting CVD, providing fellow researchers with essential insights to thoroughly grasp the field's present status. Furthermore, it includes a tabulated summary of the most notable studies employing DL in this realm, along with those integrating explainable AI (XAI) to be used as a quick reference.

CRediT authorship contribution statement

Pedro A. Moreno-Sánchez: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Data curation, Conceptualization. **Guadalupe García-Isla:** Writing – review & editing, Writing – original draft, Visualization, Formal analysis, Data curation. **Valentina D.A. Corino:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization. **Antti Vehkaoja:** Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Kirsten Brukamp:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Mark van Gils:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Luca Mainardi:** Writing – review & editing, Writing – original draft, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by the project PerCard (Personalised Prognostics and Diagnostics for Improved Decision Support in Cardiovascular Diseases) in ERA PerMed supported by the Research Council of Finland (decision number 351846), under the frame of ERA PerMed; by the Federal Ministry of Education and Research, Germany [i.e. "Bundesministerium für Bildung und Forschung (BMBF)"], grant number 01KU2211; and by the Fondazione Regionale della Ricerca Biomedica (FRRB) under the frame of ERA PerMed. L.M. is an additional member of the PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108235.

References

[1] Cardiovascular diseases (CVDs), (n.d.). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed April 17, 2023).

[2] M. Amini, F. Zayeri, M. Salehi, Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017, BMC Publ. Health 21 (2021) 401, https://doi.org/10.1186/s12889-021-10429-0.

[3] G. Al Hinai, S. Jammoul, Z. Vajihi, J. Afilalo, Deep learning analysis of resting electrocardiograms for the detection of myocardial dysfunction, hypertrophy, and ischaemia: a systematic review, European Heart Journal. Digital Health 2 (2021) 416–423, https://doi.org/10.1093/ehjdh/ztab048.

[4] A. Chang, L. Cadaret, K. Liu, Machine learning in electrocardiography and echocardiography: technological advances in clinical cardiology, Curr. Cardiol. Rep. 22 (2020), https://doi.org/10.1007/s11886-020-01416-9.

[5] N. Lei, X. Zhang, M. Wei, B. Lao, X. Xu, M. Zhang, H. Chen, Y. Xu, B. Xia, D. Zhang, C. Dong, L. Fu, F. Tang, Y. Wu, Machine learning algorithms' accuracy in predicting kidney disease progression: a systematic review and meta-analysis, BMC Med. Inf. Decis. Making 22 (2022) 205, https://doi.org/10.1186/s12911-022-01951-1.

[6] J. Qezelbash-Chamak, S. Badamchizadeh, K. Eshghi, Y. Asadi, A survey of machine learning in kidney disease diagnosis, Machine Learning with Applications 10 (2022) 100418, https://doi.org/10.1016/j.mlwa.2022.100418.

[7] S. Al'Aref, K. Anchouche, G. Singh, P. Slomka, K. Kolli, A. Kumar, M. Pandey, G. Maliakal, A. van Rosendael, A. Beecy, D. Berman, J. Leipsic, K. Nieman, D. Andreini, G. Pontone, U. Schoepf, L. Shaw, H. Chang, J. Narula, J. Bax, Y. Guan, J. Min, Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging, Eur. Heart J. 40 (2019) 1975, https://doi.org/10.1093/eurheartj/ehy404.

[8] M. van Smeden, G. Heinze, B. Van Calster, F.W. Asselbergs, P.E. Vardas, N. Bruining, P. de Jaegere, J.H. Moore, S. Denaxas, A.L. Boulesteix, K.G. M. Moons, Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease, Eur. Heart J. 43 (2022) 2921–2930, https://doi.org/10.1093/eurheartj/ehac238.

[9] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, L. Cilar, Interpretability of machine learning based prediction models in healthcare, WIREs Data Mining Knowl Discov 10 (2020), https://doi.org/10.1002/widm.1379.

[10] Ethics guidelines for trustworthy AI | Shaping Europe's digital future, (n.d.). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed August 17, 2021).

[11] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (2021) n71, https://doi.org/10.1136/bmj.n71.

[12] M.J. Page, D. Moher, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, J.E. McKenzie, PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, BMJ 372 (2021) n160, https://doi.org/10.1136/bmj.n160.

[13] PerCard project |Tampere Universities, PerCard Project (2022). https://projects.tuni.fi/percard/. (Accessed 27 February 2023).

[14] CinC – Computing in Cardiology, (n.d.). https://cinc.org/(accessed February 27, 2023).

[15] IEEE BHI-BSN-2022 – IEEE BHI-BSN-2022 Conference, (n.d.). https://bhi-bsn-2022.org/(accessed February 27, 2023).

[16] EMBC 2023, (n.d.). https://embc.embs.org/2023/(accessed February 27, 2023).

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Commun. ACM 60 (2017) 84–90, https://doi.org/10.1145/3065386.

[18] O. Yildirim, P. Plawiak, R. Tan, U. Acharya, Arrhythmia detection using deep convolutional neural network with long duration ECG signals, Comput. Biol. Med. 102 (2018) 411–420, https://doi.org/10.1016/j.compbiomed.2018.09.009.

[19] Yanting Shen, Yang Yang, S. Parish, Zhengming Chen, R. Clarke, D.A. Clifton, Risk prediction for cardiovascular disease using ECG data in the China kadoorie biobank, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2016, 2016, pp. 2419–2422, https://doi.org/10.1109/EMBC.2016.7591218.

[20] C. Li, H. Zhao, W. Lu, X. Leng, L. Wang, X. Lin, Y. Pan, W. Jiang, J. Jiang, Y. Sun, J. Wang, J. Xiang, DeepECG, Image-based electrocardiogram interpretation with deep convolutional neural networks, Biomed. Signal Process Control 69 (2021), https://doi.org/10.1016/j.bspc.2021.102824.

[21] Y. Liang, S. Yin, Q. Tang, Z. Zheng, M. Elgendi, Z. Chen, Deep learning algorithm Classifies heartbeat events based on electrocardiogram signals, Front. Physiol. 11 (2020), https://doi.org/10.3389/fphys.2020.569050.

[22] N. Sakli, H. Ghabri, B.O. Soufiene, F.A. Almalki, H. Sakli, O. Ali, M. Najjari, ResNet-50 for 12-lead electrocardiogram automated diagnosis, Comput. Intell. Neurosci. 2022 (2022), https://doi.org/10.1155/2022/7617551, 7617551–7617551.

[23] J. Bundy, S. Heckbert, L. Chen, D. Lloyd-Jones, P. Greenland, Evaluation of risk prediction models of atrial fibrillation (from the multi-Ethnic study of Atherosclerosis [MESA]), Am. J. Cardiol. 125 (2020) 55–62, https://doi.org/10.1016/j.amjcard.2019.09.032.

[24] H. Chen, C. Lin, W. Fang, Y. Lou, C. Cheng, C. Lee, C. Lin, Artificial intelligence-enabled electrocardiography predicts left ventricular dysfunction and future cardiovascular outcomes: a retrospective analysis, J. Personalized Med. 12 (2022), https://doi.org/10.3390/jpm12030455.

[25] Y. Cheng, Y. Ye, M. Hou, W. He, T. Pan, Multi-label arrhythmia classification from fixed-length Compressed ECG segments in real-time wearable ECG monitoring, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2020, 2020, pp. 580–583, https://doi.org/10.1109/EMBC44109.2020.9176188.

[26] R. Devi, V. Kalaivani, Machine learning and IoT-based cardiac arrhythmia diagnosis using statistical and dynamic features of ECG, J. Supercomput. 76 (2020) 6533–6544, https://doi.org/10.1007/s11227-019-02873-y.

[27] M. Hammad, A. Iliyasu, A. Subasi, E. Ho, A. Abd El-Latif, A multitier deep learning model for arrhythmia detection, IEEE Trans. Instrum. Meas. 70 (2021), https://doi.org/10.1109/TIM.2020.3033072.

[28] P.-Y. Hsu, C.-K. Cheng, Arrhythmia classification using deep learning and machine learning with features extracted from waveform-based signal processing, in: Annual International Conference of the IEEE Engineering in Medicine and Biology SocietyIEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2020, 2020, pp. 292–295, https://doi.org/10.1109/EMBC44109.2020.9176679.

[29] H. Ma, C. Chen, Q. Zhu, H. Yuan, L. Chen, M. Shu, An ECG signal classification method based on dilated causal convolution, Comput. Math. Methods Med. 2021 (2021), https://doi.org/10.1155/2021/6627939, 6627939–6627939.

[30] Y. Meng, G. Liang, M. Yue, Deep Learning-Based Arrhythmia Detection in Electrocardiograph, SCIENTIFIC PROGRAMMING 2021, 2021, https://doi.org/10.1155/2021/9926769.

[31] J. Ramesh, Z. Solatidehkordi, R. Aburukba, A. Sagahyroon, Atrial fibrillation classification with smart wearables using short-term heart rate variability and

deep convolutional neural networks, Sensors 21 (2021), https://doi.org/10.3390/s21217233.

[32] T. Rieg, J. Frick, H. Baumgartl, R. Buettner, Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms, PLoS One 15 (2020) e0243615–e0243615, https://doi.org/10.1371/journal.pone.0243615.

[33] B. Tutuko, S. Nurmaini, A. Tondas, M. Rachmatullah, A. Darmawahyuni, R. Esafri, F. Firdaus, A. Sapitri, AFibNet: an implementation of atrial fibrillation detection with convolutional neural network, BMC Med. Inf. Decis. Making 21 (2021), https://doi.org/10.1186/s12911-021-01571-1.

[34] L. Wang, Z. Yan, Y. Yang, J. Chen, T. Yang, I. Kuo, P. Abu, P. Huang, C. Chen, S. Chen, A classification and prediction hybrid model construction with the IQPSO-SVM algorithm for atrial fibrillation arrhythmia, Sensors 21 (2021), https://doi.org/10.3390/s21155222.

[35] X. Zhang, M. Jiang, W. Wu, V. de Albuquerque, Hybrid feature fusion for classification optimization of short ECG segment in IoT based intelligent healthcare system, Neural Comput. Appl. (n.d.). https://doi.org/10.1007/s00521-021-06693-1.

[36] L.-P. Jin, J. Dong, Ensemble deep learning for biomedical time series classification, Comput. Intell. Neurosci. 2016 (2016), 6212684–6212684.

[37] D. Wang, Q. Meng, D. Chen, H. Zhang, L. Xu, Automatic detection of arrhythmia based on multi-resolution representation of ECG signal, Sensors 20 (2020), https://doi.org/10.3390/s20061579.

[38] X. Xie, H. Liu, D. Chen, M. Shu, Y. Wang, Multilabel 12-lead ECG classification based on leadwise grouping multibranch network, IEEE Trans. Instrum. Meas. 71 (2022), https://doi.org/10.1109/TIM.2022.3164141.

[39] D. Zhang, S. Yang, X. Yuan, P. Zhang, Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram, iScience 24 (2021), https://doi.org/10.1016/j.isci.2021.102373, 102373–102373.

[40] M. Haleem, R. Castaldo, S. Pagliara, M. Petretta, M. Salvatore, M. Franzese, L. Pecchia, Time adaptive ECG driven cardiovascular disease detector, Biomed. Signal Process Control 70 (2021), https://doi.org/10.1016/j.bspc.2021.102968.

[41] N. Fayyazifar, G. Dwivedi, D. Suter, S. Ahderom, A. Maiorana, O. Clarkin, S. Balamane, N. Saha, B. King, M.S. Green, M. Golian, B.J.W. Chow, A novel convolutional neural network structure for differential diagnosis of wide QRS complex tachycardia, Biomed. Signal Process Control 81 (2023) 104506, https://doi.org/10.1016/j.bspc.2022.104506.

[42] X. Zhang, K. Gu, S. Miao, X. Zhang, Y. Yin, C. Wan, Y. Yu, J. Hu, Z. Wang, T. Shan, S. Jing, W. Wang, Y. Ge, Y. Chen, J. Guo, Y. Liu, Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system, Cardiovasc. Diagn. Ther. 10 (2020) 227–235, https://doi.org/10.21037/cdt.2019.12.10.

[43] H.M. Rai, K. Chatterjee, S. Dashkevych, The prediction of cardiac abnormality and enhancement in minority class accuracy from imbalanced ECG signals using modified deep neural network models, Comput. Biol. Med. 150 (2022) 106142, https://doi.org/10.1016/j.compbiomed.2022.106142.

[44] T.I. Toma, S. Choi, A parallel cross convolutional recurrent neural network for automatic imbalanced ECG arrhythmia detection with continuous wavelet transform, Sensors 22 (2022), https://doi.org/10.3390/s22197396.

[45] H.-W. Yang, C.-Y. Hsiao, Y.-Q. Peng, T.-Y. Lin, L.-W. Tsai, C. Lin, M.-T. Lo, C.-M. Shih, Identification of patients with potential atrial fibrillation during Sinus Rhythm using isolated P wave characteristics from 12-lead ECGs, J. Personalized Med. 12 (2022) 1608, https://doi.org/10.3390/jpm12101608.

[46] Y. Deng, Z. Gao, S. Xu, P. Ren, Y. Wen, Y. Mao, Z. Li, ST-Net: synthetic ECG tracings for diagnosing various cardiovascular diseases, Biomed. Signal Process Control 61 (2020), https://doi.org/10.1016/j.bspc.2020.101997.

[47] Z. Ge, X. Jiang, Z. Tong, P. Feng, B. Zhou, M. Xu, Z. Wang, Y. Pang, Multi-label correlation guided feature fusion network for abnormal ECG diagnosis, Knowl. Base Syst. 233 (2021), https://doi.org/10.1016/j.knosys.2021.107508.

[48] A. Tyagi, R. Mehra, Intellectual heartbeats classification model for diagnosis of heart disease from ECG signal using hybrid convolutional neural network with Goa, SN Appl. Sci. 3 (2021), https://doi.org/10.1007/s42452-021-04185-4.

[49] J. Yoo, Y. Jin, B. Ko, M.-S. Kim, K-labelsets method for multi-label ECG signal classification based on se-resnet, Appl. Sci. 11 (2021), https://doi.org/10.3390/app11167758.

[50] J. Yao, R. Li, S. Shen, W. Zhang, Y. Peng, G. Chen, Z. Wang, Combining Rhythm information between heartbeats and BiLSTM-treg algorithm for intelligent beat classification of arrhythmia, Journal of Healthcare Engineering 2021 (2021), https://doi.org/10.1155/2021/8642576.

[51] J. Wang, R. Li, R. Li, B. Fu, C. Xiao, D.Z. Chen, Towards interpretable arrhythmia classification with human-machine collaborative knowledge representation, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 68 (2021) 2098–2109, https://doi.org/10.1109/TBME.2020.3024970.

[52] T. Radhakrishnan, J. Karhade, S.K. Ghosh, P.R. Muduli, R.K. Tripathy, U. R. Acharya, AFCNNet: automated detection of AF using chirplet transform and deep convolutional bidirectional long short term memory network with ECG signals, Comput. Biol. Med. 137 (2021), https://doi.org/10.1016/j.compbiomed.2021.104783.

[53] R. Wang, J. Fan, Y. Li, Deep multi-scale fusion neural network for multi-class arrhythmia detection, IEEE Journal of Biomedical and Health Informatics 24 (2020) 2461–2472, https://doi.org/10.1109/JBHI.2020.2981526.

[54] D. Kong, J. Zhu, S. Wu, C. Duan, L. Lu, D. Chen, A novel IRBF-RVM model for diagnosis of atrial fibrillation, Comput. Methods Progr. Biomed. 177 (2019) 183–192, https://doi.org/10.1016/j.cmpb.2019.05.028.

[55] F.-Y. Zhou, L.-P. Jin, J. Dong, Premature ventricular contraction detection combining deep neural networks and rules inference, Artif. Intell. Med. 79 (2017) 42–51, https://doi.org/10.1016/j.artmed.2017.06.004.

[56] S. Smigiel, ECG classification using orthogonal matching pursuit and machine learning, Sensors 22 (2022) 4960, https://doi.org/10.3390/s22134960.

[57] H. Dai, H.-G. Hwang, V.S. Tseng, Convolutional neural network based automatic screening tool for cardiovascular diseases using different intervals of ECG signals, Comput. Methods Progr. Biomed. 203 (2021), https://doi.org/10.1016/j.cmpb.2021.106035, 106035–106035.

[58] L. Ibrahim, M. Mesinovic, K. Yang, M. Eid, Explainable prediction of acute myocardial infarction using machine learning and shapley values, IEEE Access 8 (2020) 210410–210417, https://doi.org/10.1109/ACCESS.2020.3040166.

[59] T. Khan, K. Kadir, S. Nasim, M. Alam, Z. Shahid, M. Mazliham, Proficiency assessment of machine learning classifiers: an implementation for the prognosis of breast tumor and heart disease classification, Int. J. Adv. Comput. Sci. Appl. 11 (2020) 560–569.

[60] N. Nayan, H. Ab Hamid, M. Suboh, R. Jaafar, N. Abdullah, N. Yusof, M. Hamid, N. Zubiri, A. Arifin, S. Abd Daud, M. Kamaruddin, A. Jamal, Cardiovascular disease prediction from electrocardiogram by using machine learning, INTERNATIONAL JOURNAL OF ONLINE AND BIOMEDICAL ENGINEERING 16 (2020) 34–48, https://doi.org/10.3991/ijoe.v16i07.13569.

[61] J. Park, E. Urtnasan, S. Kim, K. Lee, A prediction model of incident cardiovascular disease in patients with sleep-disordered breathing, Diagnostics 11 (2021), https://doi.org/10.3390/diagnostics11122212.

[62] G. Tadesse, H. Javed, K. Weldemariam, Y. Liu, J. Liu, J. Chen, T. Zhu, DeepMI: deep multi-lead ECG fusion for identifying myocardial infarction and its occurrence-time, Artif. Intell. Med. 121 (2021), https://doi.org/10.1016/j.artmed.2021.102192.

[63] W. Zeng, J. Yuan, C. Yuan, Q. Wang, F. Liu, Y. Wang, Classification of myocardial infarction based on hybrid feature extraction and artificial intelligence tools by adopting tunable-Q wavelet transform (TQWT), variational mode decomposition (VMD) and neural networks, Artif. Intell. Med. 106 (2020), https://doi.org/10.1016/j.artmed.2020.101848.

[64] S. Virgeniya, E. Ramaraj, A novel deep learning based gated recurrent unit with Extreme learning machine for electrocardiogram (ECG) signal recognition, Biomed. Signal Process Control 68 (2021), https://doi.org/10.1016/j.bspc.2021.102779.

[65] I. Campero Jurado, A. Fedjajevs, J. Vanschoren, A. Brombacher, Interpretable assessment of ST-segment deviation in ECG time series, Sensors 22 (2022), https://doi.org/10.3390/s22134919.

[66] W. Zeng, C. Yuan, Myocardial infarction detection using ITD, DWT and deterministic learning based on ECG signals, Cogn. Neurodynamics (n.d.). https://doi.org/10.1007/s11571-022-09870-7.

[67] D. Chumachenko, M. Butkevych, D. Lode, M. Frohme, K.J.G. Schmailzl, A. Nechyporenko, Machine learning methods in predicting patients with suspected myocardial infarction based on short-time HRV data, Sensors 22 (2022) 7033, https://doi.org/10.3390/s22187033.

[68] S.S. Kumar, S. Al-Kindi, N. Tashtish, V. Rajagopalan, P. Fu, S. Rajagopalan, A. Madabhushi, Machine learning derived ECG risk score improves cardiovascular risk assessment in conjunction with coronary artery calcium scoring, Front. Cardiovasc. Med. 9 (2022) 976769, https://doi.org/10.3389/fcvm.2022.976769.

[69] R. Krishnaswamy, B. Sivakumar, B. Viswanathan, F. Al-Wesabi, M. Obayya, A. Hilal, Intelligent biomedical electrocardiogram signal processing for cardiovascular disease diagnosis, CMC-COMPUTERS MATERIALS & CONTINUA 71 (2022) 255–268, https://doi.org/10.32604/cmc.2022.021995.

[70] J. Zhu, J. Lv, D. Kong, CNN-FWS: a model for the diagnosis of normal and abnormal ECG with feature adaptive, Entropy 24 (2022), https://doi.org/10.3390/e24040471.

[71] V. Jahmunah, E. Ng, T. San, U. Acharya, Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals, Comput. Biol. Med. 134 (2021), https://doi.org/10.1016/j.compbiomed.2021.104457.

[72] O. Lih, V. Jahmunah, T. San, E. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, O. Faust, U. Acharya, Comprehensive electrocardiographic diagnosis based on deep learning, Artif. Intell. Med. 103 (2020), https://doi.org/10.1016/j.artmed.2019.101789.

[73] A. Mohsin, O. Faust, Automated characterization of cardiovascular diseases using wavelet transform features extracted from ECG signals, J. Mech. Med. Biol. 19 (2019), https://doi.org/10.1142/S0219519419400098.

[74] S. Karthik, M. Santhosh, M. Kavitha, A. Paul, Automated deep learning based cardiovascular disease diagnosis using ECG signals, Comput. Syst. Sci. Eng. 42 (2022) 183–199, https://doi.org/10.32604/csse.2022.021698.

[75] B. Krzowski, J. Rokicki, R. Glowczynska, N. Fajkis-Zajaczkowska, K. Barczewska, M. Masior, M. Grabowski, P. Balsam, The use of machine learning algorithms in the evaluation of the effectiveness of resynchronization therapy, JOURNAL OF CARDIOVASCULAR DEVELOPMENT AND DISEASE 9 (2022), https://doi.org/10.3390/jcdd9010017.

[76] A. Villa, B. Vandenberk, T. Kentta, S. Ingelaere, H. Huikuri, M. Zabel, T. Friede, C. Sticherling, A. Tuinenburg, M. Malik, S. Van Huffel, R. Willems, C. Varon, A machine learning algorithm for electrocardiographic fQRS quantification validated on multi-center data, Sci. Rep. 12 (2022), https://doi.org/10.1038/s41598-022-10452-0.

[77] N. Diamant, P. Di Achille, L.-C. Weng, E.S. Lau, S. Khurshid, S. Friedman, C. Reeder, P. Singh, X. Wang, G. Sarma, M. Ghadessi, J. Mielke, E. Elci, I. Kryukov, H.M. Eilken, A. Derix, P.T. Ellinor, C.D. Anderson, A.A. Philippakis,

P. Batra, S.A. Lubitz, J.E. Ho, Deep learning on resting electrocardiogram to identify impaired heart rate recovery, Cardiovasc Digit Health J 3 (2022) 161–170, https://doi.org/10.1016/j.cvdhj.2022.06.001.

[78] Z. Jin, Y. Sun, A.C. Cheng, Predicting cardiovascular disease from real-time electrocardiographic monitoring: an adaptive machine learning approach on a cell phone, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2009, 2009, pp. 6889–6892, https://doi.org/10.1109/IEMBS.2009.5333610.

[79] B. Baraeinejad, M. Shayan, A. Vazifeh, D. Rashidi, M. Hamedani, H. Tavolinejad, P. Gorji, P. Razmara, K. Vaziri, D. Vashaee, M. Fakharzadeh, Design and implementation of an ultralow-power ECG patch and smart cloud-based platform, IEEE Trans. Instrum. Meas. 71 (2022), https://doi.org/10.1109/TIM.2022.3164151.

[80] J. Gao, H. Zhang, P. Lu, Z. Wang, An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset, Journal of Healthcare Engineering 2019 (2019), https://doi.org/10.1155/2019/6320651, 6320651–6320651.

[81] M. Jangra, S. Dhull, K. Singh, ECG arrhythmia classification using modified visual geometry group network (mVGGNet), J. Intell. Fuzzy Syst. 38 (2020) 3151–3165, https://doi.org/10.3233/JIFS-191135.

[82] M. Jangra, S. Dhull, K. Singh, A. Singh, X. Cheng, O-Wcnn: an optimized integration of spatial and spectral feature map for arrhythmia classification, Complex & intelligent systems (n.d.). https://doi.org/10.1007/s40747-021-00371-4.

[83] L. Niu, C. Chen, H. Liu, S. Zhou, M. Shu, A deep-learning approach to ECG classification based on adversarial domain adaptation, Healthcare 8 (2020), https://doi.org/10.3390/healthcare8040437.

[84] T. Romdhane, H. Alhichri, R. Ouni, M. Atri, Electrocardiogram heartbeat classification based on a deep convolutional neural network and focal loss, Comput. Biol. Med. 123 (2020), https://doi.org/10.1016/j.compbiomed.2020.103866.

[85] L. Tan, K. Yu, A. Bashir, X. Cheng, F. Ming, L. Zhao, X. Zhou, Toward real-time and efficient cardiovascular monitoring for COVID-19 patients by 5G-enabled wearable medical devices: a deep learning approach, Neural Comput. Appl. (n.d.). https://doi.org/10.1007/s00521-021-06219-9.

[86] X. Tang, Z. Ma, Q. Hu, W. Tang, A real-time arrhythmia heartbeats classification algorithm using parallel delta modulations and rotated linear-kernel support vector machines, IEEE Trans. Biomed. Eng. 67 (2020) 978–986, https://doi.org/10.1109/TBME.2019.2926104.

[87] H. Wang, H. Shi, X. Chen, L. Zhao, Y. Huang, C. Liu, An improved convolutional neural network based approach for automated heartbeat classification, J. Med. Syst. 44 (2019), https://doi.org/10.1007/s10916-019-1511-2.

[88] L. Wang, X. Zhou, Y. Xing, M. Yang, C. Zhang, Clustering ECG heartbeat using improved semi-supervised affinity propagation, IET Softw. 11 (2017) 207–213, https://doi.org/10.1049/iet-sen.2016.0261.

[89] T. Wang, C. Lu, Y. Sun, M. Yang, C. Liu, C. Ou, Automatic ECG classification using continuous wavelet transform and convolutional neural network, Entropy 23 (2021), https://doi.org/10.3390/e23010119.

[90] W. Yang, Y. Si, D. Wang, G. Zhang, A novel approach for multi-lead ECG classification using DL-CCANet and TL-CCANet, Sensors 19 (2019), https://doi.org/10.3390/s19143214.

[91] Y. Zhang, J. Yu, Y. Zhang, C. Liu, H. Li, IEEE, A convolutional neural network for identifying premature ventricular contraction beat and right bundle branch block beat, pp. 158–162, https://doi.org/10.1109/SNSP.2018.00037, 2018.

[92] F. Doldi, L. Plagwitz, L.P. Hoffmann, B. Rath, G. Frommeyer, F. Reinke, P. Leitz, A. Buscher, F. Guner, T. Brix, F.K. Wegner, K. Willy, Y. Hanel, S. Dittmann, W. Haverkamp, E. Schulze-Bahr, J. Varghese, L. Eckardt, Detection of patients with congenital and often concealed long-QT syndrome by novel deep learning models, J. Personalized Med. 12 (2022), https://doi.org/10.3390/jpm12071135.

[93] E. Angelaki, M. Marketou, G. Barmparis, A. Patrianakos, P. Vardas, F. Parthenakis, G. Tsironis, Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: an ECG-based approach, J. Clin. Hypertens. 23 (2021) 935–945, https://doi.org/10.1111/jch.14200.

[94] M. Rahman, M. Abul Kashem, A. Nayan, M. Akter, F. Rabbi, M. Ahmed, M. Asaduzzaman, Internet of things (IoT) based ECG system for rural health care, Int. J. Adv. Comput. Sci. Appl. 12 (2021) 470–477.

[95] H. Zhang, X. Wang, C. Liu, Y. Li, Y. Liu, P. Li, L. Yao, J. Wang, Y. Jiao, A method for detecting coronary artery stenosis based on ECG signals, J. Mech. Med. Biol. 21 (2021), https://doi.org/10.1142/S0219519421500032.

[96] E. Butun, O. Yildirim, M. Talo, R. Tan, U. Acharya, 1D-CADCapsNet: one dimensional deep capsule networks for coronary artery disease detection using ECG signals, PHYSICA MEDICA-EUROPEAN JOURNAL OF MEDICAL PHYSICS 70 (2020) 39–48, https://doi.org/10.1016/j.ejmp.2020.01.007.

[97] C. Han, K. Kang, T. Kim, J. Uhm, J. Park, I. Jung, M. Kim, S. Bae, H. Lim, D. Yoon, Artificial intelligence-enabled ECG algorithm for the prediction of coronary artery calcification, FRONTIERS IN CARDIOVASCULAR MEDICINE 9 (2022), https://doi.org/10.3389/fcvm.2022.849223.

[98] T. Kim, Effects of a Pharmacological Activator of Ca2+-Activated K+ Channels (KCa2/3) on Recovery from Acute Myocardial Infarction in Mice, 2021.

[99] F. Miao, X. Wang, L. Yin, Y. Li, A wearable sensor for arterial stiffness monitoring based on machine learning algorithms, IEEE Sensor. J. 19 (2019) 1426–1434, https://doi.org/10.1109/JSEN.2018.2880434.

[100] R. Shadmi, V. Mazo, O. Bregman-Amitai, E. Elnekave, IEEE, FULLY-CONVOLUTIONAL DEEP-LEARNING BASED SYSTEM FOR CORONARY CALCIUM SCORE PREDICTION FROM NON-CONTRAST CHEST CT, 2018, pp. 24–28. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8363515.

[101] L. Yao, C. Liu, P. Li, J. Wang, Y. Liu, W. Li, X. Wang, H. Li, H. Zhang, Enhanced automated diagnosis of coronary artery disease using features extracted from QT interval time series and ST-T waveform, IEEE Access 8 (2020) 129510–129524, https://doi.org/10.1109/ACCESS.2020.3008965.

[102] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, U. Acharya, Machine learning-based coronary artery disease diagnosis: a comprehensive review, Comput. Biol. Med. 111 (2019), https://doi.org/10.1016/j.compbiomed.2019.103346.

[103] A. Agrawal, A. Chauhan, M.K. Shetty G.M, P.M.D. Gupta, A. Gupta, ECG-iCOVIDNet: interpretable AI model to identify changes in the ECG signals of post-COVID subjects, Comput. Biol. Med. 146 (2022), https://doi.org/10.1016/j.compbiomed.2022.105540, 105540–105540.

[104] S. Baek, J. Jang, S. Yoon, End-to-End blood pressure prediction via fully convolutional networks, IEEE Access 7 (2019) 185458–185468, https://doi.org/10.1109/ACCESS.2019.2960844.

[105] R. Bie, G. Zhang, Y. Sun, S. Xu, Z. Li, H. Song, Smart assisted diagnosis solution with multi-sensor Holter, Neurocomputing 220 (2017) 67–75, https://doi.org/10.1016/j.neucom.2016.06.074.

[106] C. Chang, C. Lin, Y. Luo, Y. Lee, C. Lin, Electrocardiogram-based heart age estimation by a deep learning model provides more information on the incidence of cardiovascular disorders, FRONTIERS IN CARDIOVASCULAR MEDICINE 9 (2022), https://doi.org/10.3389/fcvm.2022.754909.

[107] X. Fan, Y. Zhao, H. Wang, K. Tsui, Forecasting one-day-forward wellness conditions for community-dwelling elderly with single lead short electrocardiogram signals, BMC Med. Inf. Decis. Making 19 (2019), https://doi.org/10.1186/s12911-019-1012-8.

[108] E. Gorodeski, H. Ishwaran, U. Kogalur, E. Blackstone, E. Hsich, Z. Zhang, M. Vitolins, J. Manson, J. Curb, L. Martin, R. Prineas, M. Lauer, Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women the women's health initiative, CIRCULATION-CARDIOVASCULAR QUALITY AND OUTCOMES 4 (2011), https://doi.org/10.1161/CIRCOUTCOMES.110.959023, 521-U80.

[109] G. Liu, X. Han, L. Tian, W. Zhou, H. Liu, ECG quality assessment based on hand-crafted statistics and deep-learned S-transform spectrogram features, Comput. Methods Progr. Biomed. 208 (2021), https://doi.org/10.1016/j.cmpb.2021.106269.

[110] P. Myers, B. Scirica, C. Stultz, Machine learning improves risk stratification after acute coronary syndrome, Sci. Rep. 7 (2017), https://doi.org/10.1038/s41598-017-12951-x.

[111] J.S. Rajput, M. Sharma, R.S. Tan, U.R. Acharya, Automated detection of severity of hypertension ECG signals using an optimal bi-orthogonal wavelet filter bank, Comput. Biol. Med. 123 (2020), https://doi.org/10.1016/j.compbiomed.2020.103924.

[112] F. Romero, D. Pinol, C. Vazquez-Seisdedos, DeepFilter: an ECG baseline wander removal filter using deep learning techniques, Biomed. Signal Process Control 70 (2021), https://doi.org/10.1016/j.bspc.2021.102992.

[113] U. Senturk, K. Polat, I. Yucedag, A non-invasive continuous cuffless blood pressure estimation using dynamic Recurrent Neural Networks, Appl. Acoust. 170 (2020), https://doi.org/10.1016/j.apacoust.2020.107534.

[114] X. Zhou, X. Zhu, K. Nakamura, M. Noro, Electrocardiogram quality assessment with a generalized deep learning model assisted by conditional generative adversarial networks, LIFE-BASEL 11 (2021), https://doi.org/10.3390/life11101013.

[115] Y.-J. Chen, C.-L. Liu, V.S. Tseng, Y.-F. Hu, S.-A. Chen, Large-scale classification of 12-lead ECG with deep learning, pp. 1–4, in: 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2019, https://doi.org/10.1109/BHI.2019.8834468.

[116] H. Kim, M. Ishag, M. Piao, T. Kwon, K. Ryu, A data mining approach for cardiovascular disease diagnosis using heart rate variability and images of carotid arteries, SYMMETRY-BASEL 8 (2016), https://doi.org/10.3390/sym8060047.

[117] K. Rjoob, R. Bond, D. Finlay, V. McGilligan, S. J Leslie, A. Rababah, A. Iftikhar, D. Guldenring, C. Knoery, A. McShane, A. Peace, Reliable deep learning-based detection of misplaced chest electrodes during electrocardiogram recording: algorithm development and validation, JMIR Medical Informatics 9 (2021) e25347, https://doi.org/10.2196/25347.

[118] S. Vijayarangan V. R, B. Murugesan, P. Sp, J. Joseph, M. Sivaprakasam, RPnet: a Deep Learning approach for robust R Peak detection in noisy ECG, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, vol. 2020, 2020, pp. 345–348, https://doi.org/10.1109/EMBC44109.2020.9176084.

[119] K. Siontis, P. Noseworthy, Z. Attia, P. Friedman, Artificial intelligence-enhanced electrocardiography in cardiovascular disease management, Nat. Rev. Cardiol. 18 (2021) 465–478, https://doi.org/10.1038/s41569-020-00503-2.

[120] M. Sharma, J. Rajput, R. Tan, U. Acharya, Automated detection of hypertension using physiological signals: a review, Int. J. Environ. Res. Publ. Health 18 (2021), https://doi.org/10.3390/ijerph18115838.

[121] G. Castelyn, L. Laranjo, G. Schreier, B. Gallego, Predictive performance and impact of algorithms in remote monitoring of chronic conditions: a systematic review and meta-analysis, Int. J. Med. Inf. 156 (2021), https://doi.org/10.1016/j.ijmedinf.2021.104620.

[122] S. Siddiqui, A. Athar, M. Khan, S. Abbas, Y. Saeed, M. Khan, M. Hussain, Modelling, simulation and optimization of diagnosis cardiovascular disease using

computational intelligence approaches, J. Med. Imaging Health Inform. 10 (2020) 1005–1022, https://doi.org/10.1166/jmihi.2020.2996.

[123] X. Liu, H. Wang, Z. Li, L. Qin, Deep learning in ECG diagnosis: a review, Knowl. Base Syst. 227 (2021), https://doi.org/10.1016/j.knosys.2021.107187.

[124] W. Ben Ali, A. Pesaranghader, R. Avram, P. Overtchouk, N. Perrin, S. Laffite, R. Cartier, R. Ibrahim, T. Modine, J. Hussin, Implementing machine learning in interventional cardiology: the benefits are worth the trouble, Frontiers in Cardiovascular Medicine (2021) 1775.

[125] A. Ferrer, R. Sebastián, D. Sánchez-Quintana, J.F. Rodríguez, E.J. Godoy, L. Martínez, J. Saiz, Detailed anatomical and electrophysiological models of human atria and torso for the simulation of atrial activation, PLoS One 10 (2015) e0141573, https://doi.org/10.1371/journal.pone.0141573.

[126] Electromechanical models of the ventricles | American Journal of Physiology-Heart and Circulatory Physiology, (n.d.). https://journals.physiology.org/doi/full/10.1152/ajpheart.00324.2011 (accessed November 2, 2023).

[127] R. Sassi, L.T. Mainardi, An estimate of the dispersion of repolarization times based on a biophysical model of the ECG, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 58 (2011) 3396–3405, https://doi.org/10.1109/TBME.2011.2166263.

[128] L. Mainardi, R. Sassi, Some theoretical results on the observability of repolarization heterogeneity on surface ECG, J. Electrocardiol. 46 (2013) 270–275, https://doi.org/10.1016/j.jelectrocard.2013.02.011.

[129] R. Moss, E.M. Wülfers, S. Schuler, A. Loewe, G. Seemann, A fully-coupled electro-mechanical whole-heart computational model: influence of cardiac contraction on the ECG, Front. Physiol. 12 (2021). https://www.frontiersin.org/articles/10.3389/fphys.2021.778872. (Accessed 2 November 2023).

[130] A. Mincholé, J. Camps, A. Lyon, B. Rodríguez, Machine learning in the electrocardiogram, J. Electrocardiol. 57 (2019) S61–S64, https://doi.org/10.1016/j.jelectrocard.2019.08.008.

[131] S. Somani, A.J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J.K. De Freitas, N. Naik, R. Miotto, G.N. Nadkarni, J. Narula, E. Argulian, B.S. Glicksberg, Deep learning and the electrocardiogram: review of the current state-of-the-art, EP Europace 23 (2021) 1179–1191, https://doi.org/10.1093/europace/euaa377.

[132] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C.A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, B. Bischl, General pitfalls of model-agnostic interpretation methods for machine learning models, in: A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, W. Samek (Eds.), xxAI - beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Springer International Publishing, Cham, 2022, pp. 39–68, https://doi.org/10.1007/978-3-031-04083-2_4.

[133] Y.M. Ayano, F. Schwenker, B.D. Dufera, T.G. Debelee, Interpretable machine learning techniques in ECG-based heart disease classification: a systematic review, Diagnostics 13 (2023) 111, https://doi.org/10.3390/diagnostics13010111.

[134] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: challenges and prospects, arXiv:1812.04608 [Cs], http://arxiv.org/abs/1812.04608, 2019. (Accessed 11 January 2021).

[135] Z. Ebrahimi, M. Loni, M. Daneshtalab, A. Gharehbaghi, A review on deep learning methods for ECG arrhythmia classification, Expert Syst. Appl. X 7 (2020) 100033, https://doi.org/10.1016/j.eswax.2020.100033.

[136] N. Musa, A.Y. Gital, N. Aljojo, H. Chiroma, K.S. Adewole, H.A. Mojeed, N. Faruk, A. Abdulkarim, I. Emmanuel, Y.Y. Folawiyo, J.A. Ogunmodede, A.A. Oloyede, L. A. Olawoyin, I.A. Sikiru, I. Katb, A systematic review and Meta-data analysis on the applications of Deep Learning in Electrocardiogram, J. Ambient Intell. Hum. Comput. 14 (2023) 9677–9750, https://doi.org/10.1007/s12652-022-03868-z.

[137] M. Loni, A. Zoljodi, S. Sinaei, M. Daneshtalab, M. Sjödin, NeuroPower: designing energy efficient convolutional neural network architecture for embedded systems, in: I.V. Tetko, V. Kůrková, P. Karpov, F. Theis (Eds.), Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation, Springer International Publishing, Cham, 2019, pp. 208–222, https://doi.org/10.1007/978-3-030-30487-4_17.

[138] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, J. Carreira, Perceiver: General Perception with Iterative Attention, 2021, https://doi.org/10.48550/arXiv.2103.03206.

[139] G. Garcia-Isla, F.M. Muscato, A. Sansonetti, S. Magni, V.D.A. Corino, L. T. Mainardi, Ensemble classification combining ResNet and handcrafted features with three-steps training, Physiol. Meas. 43 (2022) 094003, https://doi.org/10.1088/1361-6579/ac8f12.

[140] EU AI Act: First Regulation on Artificial Intelligence, News | European Parliament, 2023. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. (Accessed 15 June 2023).

[141] E. Pietilä, P.A. Moreno-Sánchez, When an explanation is not enough: an overview of evaluation metrics of explainable AI systems in the healthcare domain, in: A. Badnjević, L. Gurbeta Pokvić (Eds.), MEDICON'23 and CMBEBIH'23, IFMBE Proceedings, Springer Nature Switzerland, Cham, 2024, pp. 573–584. https://doi.org/10.1007/978-3-031-49062-0_60.

[142] M. Mourby, K. Ó Cathaoir, C.B. Collin, Transparency of machine-learning in healthcare: the GDPR & European health law, Comput. Law Secur. Rep. 43 (2021) 105611, https://doi.org/10.1016/j.clsr.2021.105611.

[143] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C.T. Evelo, C. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R.W.W. Hooft, M. Imming, K. G. Jeffery, R. Kaliyaperumal, M.G. Kersloot, C.R. Kirkpatrick, T. Kuhn, I. Labastida, B. Magagna, P. McQuilton, N. Meyers, A. Montesanti, M. van Reisen, P. Rocca-Serra, R. Pergl, S.-A. Sansone, L.O.B. da Silva Santos, J. Schneider, G. Strawn, M. Thompson, A. Waagmeester, T. Weigel, M.D. Wilkinson, E. L. Willighagen, P. Wittenburg, M. Roos, B. Mons, E. Schultes, FAIR principles: interpretations and implementation considerations, Data Intelligence 2 (2020) 10–29, https://doi.org/10.1162/dint_r_00024.

PEDRO A. MORENO-SÁNCHEZ received the B.S. degree in telecommunication engineering, the M.S. degree in telemedicine and bioengineering, and the Ph.D. degree in biomedical engineering from the Technical University of Madrid, in 2007, 2008, and 2014, respectively. Since 2022, he has been a Postdoctoral Research Fellow with Tampere University, Finland. His research interests include digital health and in the application of artificial intelligence and machine learning to develop clinical prediction models. Currently, his research area is centered in explainable and trustworthy AI in healthcare.

GUADALUPE GARCÍA-ISLA obtained her PhD in Biomedical Engineering by the Politecnico di Milano University (POLIMI) as an early stage researcher of the Marie Skłodowska-Curie MY-ATRIA international training network. She is currently a postdoctoral researcher at POLIMI in the Bioimaging, Biosignals and Bioinformatics laboratory (B3Lab) in the Department of Electronics, Information and Bioengineering and a teaching assistant in the Applied Artificial Intelligence in Biomedicine and Biomedical Signal Processing Laboratory subjects in the Biomedical Engineering master's degree in POLIMI. Her research field includes the study of cardiac arrhythmia, especially atrial fibrillation, artificial intelligence applied to biomedical data and biomedical signal processing.

VALENTINA D.A. CORINO is Associate Professor at the Department of Electronics, Information and Bioengineering at Politecnico di Milano, Italy. Her research activity lies in the field of biomedical signal and signal processing. Her main interests are in signals of cardiovascular origin, especially regarding supraventricular arrhythmias to assess for example autonomic tone, drug response. Regarding image processing, her research lies in radiomics applied to magnetic resonance images and computed tomography of oncological and cardiac patients to assess treatment response and survival analysis. She is responsible of the LEGO lab (DigitaL tEchnologies for imaGing and sensOrs), a laboratory dedicated to digital technologies, imaging and sensoring, and part of the joint research center between Politecnico di Milano and IRCCS Centro Cardiologico Monzino.

ANTTI VEHKAOJA received the D.Sc. (Tech.) degree in automation science and engineering from the Tampere University of Technology, Tampere, Finland, in 2015. He is currently an Associate Professor (tenure track) of Sensor Technology and Biomeasurements at the Faculty of Medicine and Health Technology, Tampere University. His research interests include embedded measurement technologies for physiological monitoring and related signal processing and data analysis methods. He has authored more than 100 scientific articles in the field of biomedical engineering.

KIRSTEN BRUKAMP, MD, MS, MA, is a professor of health sciences at Protestant University Ludwigsburg in Germany. Her research focuses on the use and implementation of health technologies as well as ethical and social implications. Her background is in medicine, philosophy, and cognitive science.

MARK VAN GILS is Professor of Digital Healthcare, leading the research group Decision Support for Health, at the Faculty of Medicine and Health Technology at Tampere University, Finland. His over 25-year long career in biomedical data analysis has included AI-driven patient monitoring research during the 1990's at Eindhoven University of Technology, the Netherlands, a wide scope of research and research leadership activities (including Research Professorship) in the area of digital health at VTT Technical Research Center of Finland, and an Adjunct Professorship at Aalto University in Espoo, Finland. During his career, he has worked tightly with renowned university hospitals and health tech companies, and he has obtained extensive experience in coordinating multi-disciplinary international research projects. His professional interests are in data-driven decision support for health and wellbeing, with special attention to addressing real-life challenges in realizing actual uptake and impact of technical solutions.

LUCA MAINARDI is Full Professor in Bioelectromagnetism and Biomedical Signal Processing at the Department of Electronics, Information and Bioengineering of the Politecnico di Milano. He is the Director of the Bachelor and Master Programme in Biomedical Engineering at Politecnico di Milano and the Co-Chair of the recently established Joint-Research Facility center between Politecnico di Milano and the Auxologico Hospital. Prof. Mainardi is fellow of the EAMBES and member of the Board of Computing in Cardiology. His research activity is in the field of biomedical signal and image processing, and biomedical system modeling. He studies and develops methods for time-frequency analysis and non-linear analysis of cardiovascular signals and series, with interest in the investigation of atrial fibrillation. He is also interested in advanced biomedical image processing techniques for features extraction, image registration and Radiomics for oncology application. He was the Coordinator of the Marie-Curie EU project MY-ATRIA a "MultidisciplinarY training network for ATrial fibRillation monItoring, treatment and progression.