Diagnostic and prognostic models based on electrocardiograms for rapid clinical applications

ABSTRACT:
Leveraging artificial intelligence (AI) for the analysis of electrocardiograms (ECG) has the potential to transform diagnosis and estimate the prognosis of not only cardiac but, increasingly, non-cardiac conditions. In this review, we summarize clinical studies and AI-enhanced ECG-based clinical applications in the early detection, diagnosis, and estimating prognosis of cardiovascular diseases (CVD) in the last five years (2019-2023). With advancements in deep learning and the rapid increased use of ECG technologies, a large number of clinical studies have been published. However, a majority of these studies are single-center, retrospective, proof-of-concept studies that lack external validation. Prospective studies that progress from development toward deployment in clinical settings account for <15% of the studies. Successful implementations of ECG-based AI applications that have received approval from the Food and Drug Administration (FDA) have been developed through commercial collaborations, with about half of them being for mobile or wearable devices. The field is in its early stages, and overcoming several obstacles is essential, such as prospective validation in multi-center large datasets, addressing technical issues, bias, privacy, data security, model generalizability, and global scalability. This review concludes with a discussion of these challenges and potential solutions. By providing a holistic view of the state of AI in ECG analysis, this review aims to set a foundation for future research directions, emphasizing the need for comprehensive, clinically integrated, and globally deployable AI solutions in CVD management.

Introduction:
The Electrocardiograms (ECG) has long been a cornerstone in the diagnostic and prognostic assessment of cardiovascular diseases (CVD), the leading cause of death globally.[1] The significance of ECGs in clinical diagnosis stems from their simplicity, low cost, and non-invasive nature, making them indispensable in detecting and managing a range of cardiac conditions, including arrhythmias, myocardial infarction (MI), and coronary artery disease (CAD). This paper offers a comprehensive review of the use of Artificial Intelligence/Machine Learning (AI/ML) in enhancing ECG-based diagnostic and prognostic methods, particularly focusing on the use of deep learning (DL) techniques and their application in clinical settings. Recent advancements in computational models and ML algorithms have significantly enhanced the potential of ECGs. Unlike traditional ECG analysis, which is limited by human expertise, reliance on knowledge-based features and decision-making rules, DL techniques facilitate more informative feature extraction,[2] demonstrating superior performance in disease detection and prediction. The integration of AI/ML with ECG analysis has revolutionized the field, facilitating the development of sophisticated diagnostic and prognostic models.[3,4]

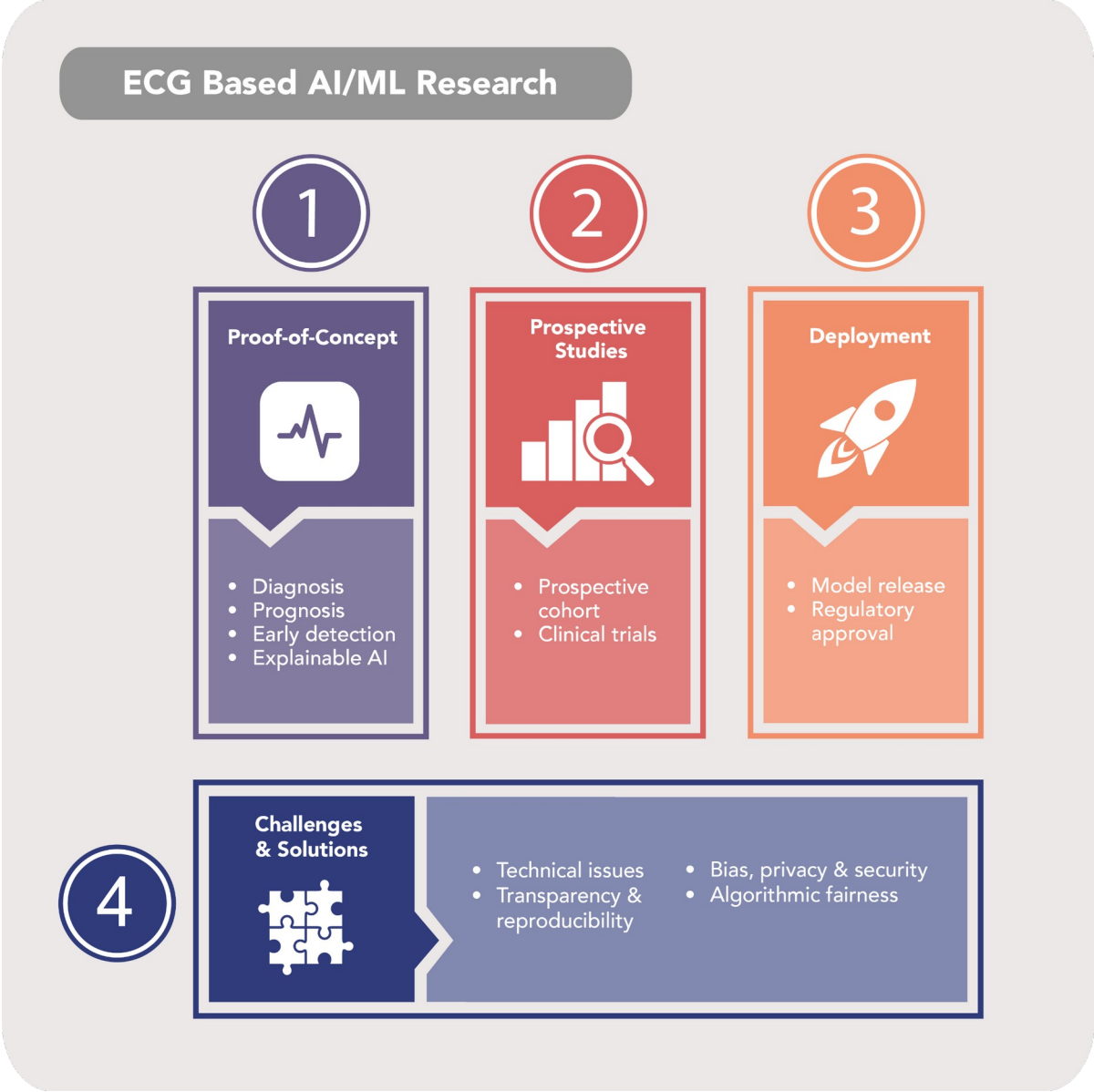

Figure 1: AI/ML ECG based models development to deployment process in clinical applications.

Existing review papers[5–10] primarily focus on DL and ML strategies for detecting or predicting specific diseases; and none have comprehensively addressed prospective studies and deployment of AI/ML-based ECG models in clinical settings. Our review aims to bridge this gap and includes the following four sections (Figure 1): the first section is focused on AI/ML models for ECG analysis developed as proof-of-concept studies using retrospectively collected datasets, with an emphasis on their roles in diagnosis, prognosis, and early detection of cardiac abnormalities; the second section includes both cohort studies and clinical trials focused on prospective evaluation of ECG-based AI/ML algorithms; the third section includes AI/ML ECG software or models that are deployed in the real world and have received regulatory approval; and the fourth and final section discusses some of the major challenges, including technical

issues, transparency and reproducibility, bias, privacy and security, and algorithmic fairness, and potential mitigation strategies in the development to deployment of ECG-based AI/ML models in healthcare settings. We hope this review highlights the combined potential of AI and ECG to improve clinical practice and inform the future direction of AI/ML applications in healthcare.

1: RETROSPECTIVE, PROOF-OF-CONCEPT STUDIES

The availability of DL methods has revolutionized ECG analysis by enabling direct feature extraction from raw data, eliminating the need for manual feature selection.[11] This advancement enhances diagnostic and prognostic capabilities, as DL algorithms can detect complex patterns within large datasets more effectively than traditional methods.[5] The number of publications on AI/ML-based ECG models for CVD detection and prediction has increased substantially over the last five years. (Figure 2).

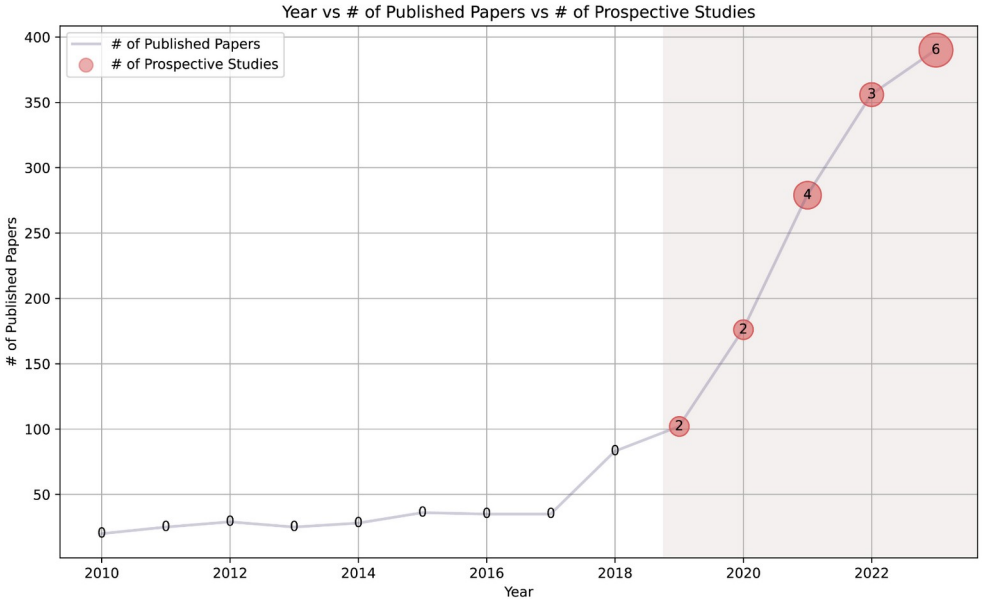

Figure 2: Yearly trends in AI/ML ECG model publications and prospective studies.

1.1: Early detection models of CVDs

One of the most remarkable uses of AI/ML models for ECG analysis is for early disease detection.[12–14] AI models are shown to be proficient in recognizing some early indicators of cardiac dysfunction, especially for atrial fibrillation (AF),[15] heart failure (HF),[16] and aortic stenosis (AS)[13] —conditions that, although potentially severe, can be effectively managed if detected early. Furthermore, these models show promise in forecasting future cardiac events, such as the onset of AF, a common yet serious heart rhythm condition that poses an increased risk of stroke.[17,18]

Early detection of AF is crucial, but conventional methods often have a low detection yield.[19] Majority[15,17,20–27] of the early detection methods use DL based models while focusing on AF detection during normal sinus rhythm (NSR). For example, Attia et al.[15] demonstrated that the AI-ECG algorithm can identify AF, whereas Lee et al.[24] use DL methods with attention mechanisms for early detection of AF based on P-wave location. Recently, Budaraju et al.[25] stacked ML models to obtain insight into important clinical ECG aspects for early AF prediction.

In addition, some papers[12,28–32] have focused on early detection of ST-elevation myocardial infarction (STEMI), mitral regurgitation (MR), pulmonary hypertension (PH), ventricular premature complex (VPC), left ventricular ejection fraction (LVEF), hypocalcemia and hypercalcemia. For example, Zhao et al.[12] developed an AI model for early detection of STEMI, demonstrating capabilities akin to cardiologist-level diagnosis, whereas Kashou et al.[28] have shown that the AI-ECG algorithm can serve as a rapid screening tool for early diagnosis of LVEF in a low resource setting.

1.2: ECG based Diagnostic Models

The ECG's compatibility with DL approaches allows for the development of models that can interpret signals beyond human capabilities, identifying subtle indicators of conditions such as left ventricular dysfunction,[33,34] silent AF, and hypertrophic cardiomyopathy (HCM),[35] as well as physiological traits like age and sex.[36,37] AI applications in ECG analysis promise more rapid and cost-effective cardiovascular phenotyping, directing further investigations based on detailed health assessments. The heatmaps in Figure 3 illustrate the distribution of papers categorized by disease type and ECG leads, indicating that studies focusing on arrhythmias and heart failure with 12-lead ECGs are the predominant topics within the research landscape. Many diagnostic studies aim to support clinicians to detect and classify ECG abnormalities[38–51] including AF, Atrial flutter (AFL), Supraventricular Arrhythmia, Ventricular Arrhythmia, Left bundle branch block (LBBB), and Right bundle branch block (RBBB). Diagnostic classification tasks are varied from binary disease classification to multilabel classification of different cardiac arrhythmias. For example, Ribeiro et al.[52] applied the DL model to classify six ECG abnormalities and compared results with the cardiologist's validation with a high specificity of 99%. Several other studies[53–60] have focused on heart attack, heart valvular disease, CAD, Congenital heart disease (CHD) and related conditions. For example, Jahmunah et al.[57] developed an AI based automatic tool to detect normal, CAD, myocardial infarction (MI) and congestive heart failure (CHF) classes using DL models whereas Elias et al.[53] use AI/ML ECG-based models to detect moderate or severe AS, aortic regurgitation (AR) in combination with MR. While Grogan et al.[61] developed an AI ECG model for cardiac amyloidosis detection, Kwon et al.[54] developed an ensemble model to detect AS using ECGs. Recently, Kalmady et al.[62] used DL and extreme gradient boosting to classify 15 different CVD diagnoses simultaneously using ECG traces and measurements, respectively. Interestingly, DL models demonstrate strong discriminative potential in population-based ECG datasets, across a wide spectrum of diseases

including non-cardiovascular conditions such as mental, neurological, metabolic and infectious conditions.[63]

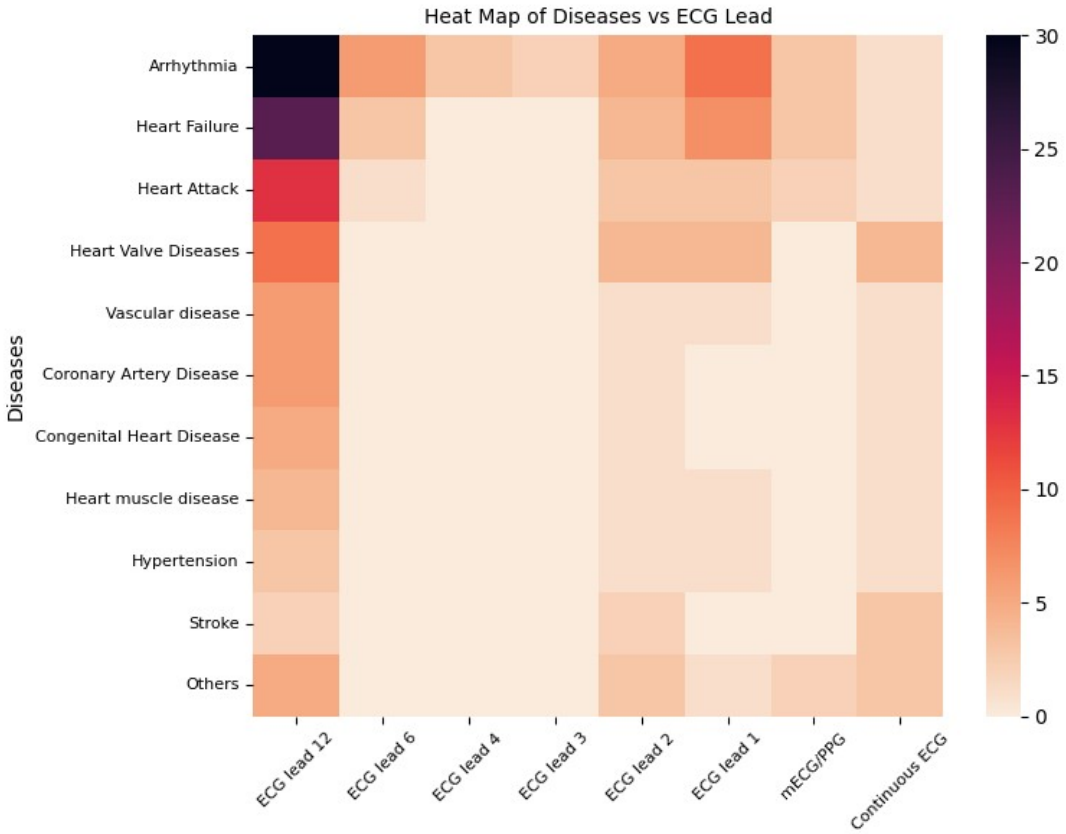

Figure 3: Number of papers studying various cardiac disease categories by ECG data type.

1.3: ECG based prognostic models:

While the volume of research on AI/ML-based ECG models for CVD prediction and detection has surged in the last five years, prognostic studies in this area have grown at a slower pace. The majority[34,64–70] of the prognostic studies have focused on heart failure including Left ventricular (LV) dysfunction, MI, followed by mortality prediction. Among studies focused on LV dysfunction,[34,64–67] Vaid et al.[34] developed AI-enhanced ECG-based predictive tools for prognosis LV and RV dysfunction, enhancing how patients are prioritized for intervention. De la Garza et al.[66] used ECG and C5.0 algorithm to predict Echo-LVH, and Mahayni et al.[67] developed an ECG-based AI algorithm that predicts severe ventricular dysfunction to predict long-term mortality among patients with left ventricular ejection fraction (LVEF) > 35% undergoing valve and/or coronary bypass surgery. Furthermore, Dutta et al.[71] proposed a two-layer CNN demonstrating balanced class-specific performance, especially in large, imbalanced ECG datasets, for coronary heart disease (CHD). Kwon et al.[72] implemented a DNN that learned a model that used ECG data to predict patients' 12- and 36-month mortality following acute heart failure. Van de Leur et al.[73] developed models that performed well in predicting in-hospital all-cause mortality of patients with COVID-19 with pre-trained DNN using age, sex, and the raw ECG waveforms. Raghunath et al.[74] predicted 1-year all-cause mortality from ECG voltage–time

traces with custom-designed DL architecture that utilized CNNs using five branches to accommodate varying durations of ECG acquisition across the groups of leads. Sun et al.[75] developed ECG-based ML models to predict 30-day, 1-year, and 5-year mortality risk among patients presenting to an ED or hospital at the population level. Lima et al.[76] developed a DNN model for mortality risk prediction in broader patient populations.

Additionally, recent works[77–80] have introduced novel DL models for detailed prognostic predictions, including long-term clinical outcomes and risk identification for CVD. Some papers[81–83] employed techniques such as pre-training and transfer learning for various prognostic predictions, from acute heart failure to the impacts of COVID-19 on patient mortality. Nademi et al.[79] and Sun et al.[80] developed a supervised feature extractor based on a pre-trained diagnostic DL model to predict patient-specific individual survival distributions using ECG. Wouters et al.[77] developed FactorECG, an end-to-end explainable DL model that can accurately predict long-term clinical outcomes, including death, left ventricular assist device implantation, heart transplantation, and HF hospitalization. In addition, a few studies have used wearable or mobile device single lead ECG to predict left ventricular hypertrophy (LVH),[84] and STEMI.[40]

1.4: Data, Methods, and Explainability

Access to large volumes of digitized ECG data is a critical and necessary step in the development of AI/ML models. To date, the evolution of AI/ML ECG techniques has been supported by the availability of public datasets,[85] primarily the MIT-BIH,[86] data released as part of Physionet challenges,[87] and PTB databases.[88] as well as private/proprietary datasets (Appendix A). Recent studies show that the success of AI/ML models in identifying CVDs depends on the data size and data quality. Seo et al.[89] demonstrated the data dependency of ECG-based AI/ML models for detecting atrial fibrillation, highlighting performance variations when models trained on one dataset are tested on external datasets. While many public datasets are small in size and lack external validation data, recently published papers used privately curated large datasets. For example, a few studies have used standard 12-lead large datasets (more than 1 million ECGs) for the diagnosis and prognosis task, including 2,322,513 ECG from 1,676,384 patients in Brazil;[90] 2,015,808 ECG from 260,065 patients in Alberta;[75] 1,576,581 ECGs from 449,380 patients in Mayo clinic in US[91] and 1,169,662 ECG from 253,397 patients by Geisinger, USA.[74] Similarly, the largest dataset for smartwatch or wearables is from multinational eHealth study[92] with 3,144,331 ECG from 66,788 patients.

The effectiveness of DL models in interpreting ECG data is influenced by both the quality and type of ECG data collected, which can vary from digital to paper-based formats and include wearable or remote ECG devices. These devices, which can record ECGs in formats ranging from single to 12 leads, heart rate variability (HRV), and photoplethysmography (PPG), offer simplified, patient-operated methods for cardiac diagnostics and management, thus enhancing healthcare outcomes and accessibility. As illustrated in Figure 3, the standard 12-lead ECG is most commonly used in research, while single-lead ECGs have been used to reduce computational demands[93–102] and few other studies[39,103–108] have utilized ECG images.

Additionally, several studies[109–111] have explored the use of heart rate variability features and ECG data extracted from the mobile app for detecting heart abnormalities and stress, further demonstrating the versatility and adaptability of ECG data in advancing cardiac care through AI/ML technologies.

A diverse array of methods have been employed in developing ECG-based AI/ML algorithms, spanning from conventional ML techniques like Logistic Regression (LR) and Support Vector Machine (SVM), to more recent advancements like gradient boosting trees, and cutting-edge DL models. The selection of algorithm methodology hinges on several factors including the size of the dataset for training and testing, the nature of the data format (e.g., summarized ECG measurements versus ECG signals or images), and the specific prediction task such as classification, regression, or time-to-event modeling.

Conventional algorithm methodologies are typically favored for smaller datasets with a limited number of features. For instance, He et al.[112] utilized SVM to detect new-onset postoperative atrial fibrillation with a dataset comprising 100 patients. Gradient boosting techniques such as XGBoost exhibit superior performance in tabular datasets containing ECG measurements and patient characteristics[75,106,113] Meanwhile, DL models prove effective for semi- or unstructured datasets encompassing digitized ECG tracings or scanned ECG images.[39,75,103,105,114] DL methods leverage automatic feature extraction and often outperform traditional ML methods, particularly in standard or large datasets. Various types of deep convolutional neural networks (CNNs) such as ResNet,[41,75,89,115–117] EfficientNet,[39,118,119] DenseNet,[120] multi-scale CNNs,[121–123] and attention based transformers (cite) have been applied in ECG-based AI studies. Additionally, temporal or sequential learning algorithms like Long Short-Term Memory networks (LSTMs),[56,124,125] Bi-LSTM with Attention,[103,126–129] and hybrid combinations of CNNs and LSTMs[129,130] are also utilized. Methodological innovations extend to areas such as transfer learning,[46,83,131] cloud-based frameworks for real-time analysis,[45,48] and generative AI techniques including autoencoders[43,49,132] and Generative Adversarial Networks (GANs).[133,134] Many of these methods are coupled with solutions for addressing data imbalance, such as Synthetic Minority Over-sampling Technique (SMOTE)[117] or multi-center data sharing issues such as Federated learning (FL).[135] While some studies focus on binary classifications with two classes,[114,136] others tackle multilabel classifications encompassing several classes.[103,137] Additionally, some employ personalized survival distributions for predicting time to death in censored datasets.[79,80,138,139]

Explainable AI (XAI) serves as a crucial bridge between the complex decision-making processes of AI models and the intuitive understanding required by users, particularly in critical fields like healthcare.[140,141] It intends to ‘open’ the AI "black box," enhancing transparency and trust.[141] In the context of medical image classification, XAI can pinpoint the image regions most pivotal in influencing its predictions.[142] This also fosters generation of hypothesis for future studies by identifying relevant features. XAI's integration into healthcare ML/DL models promotes trust, acceptance, and continual performance improvements.[143] However, less than a third of studies

use an ECG-based XAI model to facilitate physicians' comprehension of the ML model's results. The Gradient-weighted Class Activation Mapping (GradCAM)[144] appears to be the most employed XAI method (Figure 4). In addition to GradCAM, Saliency maps,[145] Heat maps, SHapley Additive exPlanations (SHAP),[53,146] and Local Interpretable Model-agnostic Explanations (LIME)[147] are frequently used, each contributing uniquely to the field of ECG-based XAI. SHAP is valuable in healthcare for elucidating predictions on patient outcomes such as mortality and admissions For example, Ibrahim et al.[146] used SHAP to identify highly contributing features from ECG in prediction of acute MI. GradCAM, effective in identifying key areas in ECG traces, enhances interpretative capabilities at individual patient level.[148–150] GradCAM also has been used in diverse ECG formats for AF classification, MI classification or mortality prediction in various CV conditions.[120,151,152] Conversely, LIME offers model-agnostic explainability, which has been utilized in heartbeat classification.[147] In algorithms for detecting valvular disease, saliency maps have been employed to emphasize the segments of the ECG that influenced the model's decision in chosen samples.[13,54,145]

Figure 4: Distribution of popular XAI methods used in the published ECG based AI/ML papers.

2: PROSPECTIVE STUDIES

To comprehensively evaluate the practicality and efficacy of emerging AI/ML models based on ECG in healthcare, it is crucial to undertake a series of planned prospective studies. Compared to the number of proof-of-concept or retrospective studies papers published in recent years, only 17 papers are prospective studies, of which 15 are cohort studies and two are clinical trials. Among the two trials, the EAGLE trial[153,154] assessed an AI algorithm for identifying left ventricular dysfunction in a large-scale study involving over 100 clinical teams and 24,000 patients across nearly 50 primary care practices. This trial evaluated the algorithm's effectiveness in detecting LVEF and assessed how clinicians interpret and act upon AI-generated information. Meanwhile, Noseworthy et al.[155] conducted a non-randomized interventional trial which demonstrated that AI-guided screening significantly improves AF detection.

Among the 15 observational or cohort studies, three studies evaluated AI's capability in detecting low LVEF, and others concentrated on AF, AMI, HF, and other abnormalities. In the LVEF context, significant contributions have been made. For example, Attia et al.[156] validate a DL algorithm that predicts an LVEF less than or equal to 35% based on the 12-lead ECG in a large prospective cohort. Similarly, Bachtiger et al.[157] conducted an observational, prospective, multicentre study for LVEF to interpret single-lead ECG input with an ECG-enabled stethoscope. Additionally, Sangha et al.[107] developed and externally validated a DL model that

identifies LV systolic dysfunction from ECG images. These approaches represent automated and accessible screening strategies for LV systolic dysfunction, particularly in low-resource settings.

Another prospective study includes Bumgarner et al.[158] which demonstrated that integrating an AI algorithm with a wearable ECG recorder enhances AF detection across various conditions in a single-center dataset where two separate electrophysiologists and physician teams interpreted and compared the results. Similarly, Lee et al.[159] developed a predictive model for AF in patients with acute ischemic stroke, effectively validating it with another dataset. Zhao et al.[161] and Carpretz et al.[162] have developed ML models for differentiating cardiac conditions using surface ECG characteristics and predicting AMI or death in emergency department patients, respectively. These models were validated with small external data (n=100) and a prospective cohort (n=50), respectively. Surendra et al.[163] introduced a CNN-based screening tool for HF detection, emphasizing digital ECG's potential at a population level and evaluating it with a single center prospective cohort. Liu et al.[103] introduced the aggregation attention multilabel electrocardiogram classification model (AA-ECG), capable of identifying cardiac abnormalities in raw ECG images with image-level annotations, which was validated in a two-site prospective study. It involved creating large-scale, real-world ECG datasets annotated by experts and comparing the model's performance against seven classifiers for 27 primary categories. Bouzid et al.[164] made noteworthy advancements in detecting culprit lesions using temporal and spatial ECG features with a random forest classifier, surpassing traditional ST amplitude measurements. In subsequent prospective studies,[165,166] they identified key ECG features for acute coronary syndrome (ACS) detection and demonstrated the efficacy of random forest classifiers in diagnosing non-ST-elevation ACS with out-of-hospital ECGs. In the wearable space, Poh et al.[167] developed a medical-grade continuous AF monitoring diagnostic tool for wrist-worn devices. Similarly, Fu et al.[160] conducted a clinical study and demonstrated that wearing a dynamic ECG recorder integrated with an AI algorithm can detect AF effectively in different postures and after exercises. In addition, Giudicessi et al.[168] developed a DL model using smartphone-enabled electrodes for accurate QTc interval prediction, offering a cost-effective alternative for screening long QT syndrome. These studies underscore AI's role in enhancing diagnostic and prognostic accuracy in diverse clinical scenarios.

3: CLINICAL DEPLOYMENT

Experts predict that AI/ML will play a pivotal role in the diagnosis, estimating prognosis, management, and treatment of a diverse array of medical conditions.[169] While the proof-of-concept performance data of AI/ML ECG models is encouraging, their true value will be measured by their tangible contributions to improving clinical practices and patient outcomes.[170] So far, several AI/ML models have been tested in various clinical applications, including algorithms to identify LV dysfunction,[153] concomitant silent AF, or the risk of near-term AF.[4,171] Prior to deployment in the real world clinical settings, AI software or models need to be approved by regulatory bodies such as the Health Canada, the Food and Drug Administration

(FDA) and the European Union Medical Device Regulation (EU-MDR). Regulatory agencies often classify an AI algorithm as 'Software as a Medical Device' (SaMD), which is widely recognized as software designed for clinical purposes but not incorporated as a component within a physical medical device.[172–175]

Since 2019, the FDA has authorized a growing number of SaMD (marketed via 510(k) clearance, granted De Novo request, or premarket approval) that use AI/ML in healthcare settings.[176,177] Between 2008 and October 2023, 77 AI/ML based SaMD received FDA approval in cardiovascular field, with only 19 SaMD being specifically ECG-based.[178] Appendix B provides a comprehensive list of FDA-approved ECG-based AI/ML algorithms, devices, and mobile applications deployed in clinical settings.

Among the current ECG-based AI/ML SaMDs, ten models are clinician-facing applications for medical facilities such as hospitals, clinics, or doctor's offices, whereas nine models are exclusively for mobile apps or mHealth devices. Most of the SaMD (n=18) assist with diagnostic tasks to detect, identify, or assess heart rhythms, while one algorithm is intended to aid in screening for LVEF less than or equal to 40% in adults at risk for heart failure. A majority (7/10) of the clinical-facing applications are for assessing abnormal heart rhythms and arrhythmia detection, focusing on AF, including asystole, bradycardia, atrial tachycardia, ventricular tachycardia, NSR, and artifact. Another clinician facing application is the AI-ECG Tracker, which is used for QRS detection, Supraventricular and Ventricular Ectopic Beat detection, QRS feature extraction, interval measurement and heart rate measurement. Similarly, the Analytic for Hemodynamic Instability (AHI) software is used by healthcare professionals managing adult inpatients who are receiving continuous physiological monitoring with ECG. Most of the mHealth device applications (n = 7 out of 9) focus on diagnosing irregular heart rhythms, e.g., AF. The rest of the applications assess cardiac activity and generate reports for clinicians to review.

4: CHALLENGES AND POTENTIAL SOLUTIONS

Even with the exponential growth of AI/ML within the field of cardiology, the number of models deployed in real-world clinical settings remains limited due to numerous challenges.[179] Some of the significant challenges in deploying AI/ML models in daily clinical practice is the trustworthiness, reliability, and regulation of such technologies. Moreover, while AI/ML methods are beneficial for capturing intricate patterns in data, they can pose challenges regarding interpretability, generalization, regularization, robustness, stability, transparency, and optimization.[180] Healthcare providers need to understand the high-level decision-making process of AI models to trust and effectively use them in clinical practice.

Simplifying complex models without compromising performance, ensuring privacy, security and maintaining regulatory compliance is a formidable challenge.[181] A summary of some of the challenges and potential solutions for ECG-based AI/ML research are presented in Table 1, and discussed below.

4.1: Data Quality and Size**:**

The success of AI/ML models in ECG analysis relies critically on high-quality data, which must be both accurate and representative of the target clinical population. Variability in data source, sample size, type, and format, notably between mobile and conventional ECG devices, can lead to inconsistencies due to format differences.[182] Ensuring data is free from errors, biases, and inconsistencies is crucial. Although most studies use standard 12 lead ECG (Figure 3), models developed from high-quality databases and well-characterized patient cohorts may underperform when applied to ECGs from routine clinical practice due to real-world variability. The challenge lies in gathering comprehensive datasets of diverse populations, including various age groups, ethnicities, and underlying health conditions. Another major challenge is handling missing data, expert data annotation, and verification.[5] Standard ECG data, being inherently multi-dimensional, have prompted some studies to focus on reducing dimensionality to boost algorithm performance.[56] The presence of nonlinearities and complex transformations within ML algorithms can challenge the traceability of source data and its processing - which underscores the importance of code sharing and, whenever feasible, sharing datasets.

Challenges and potential Solutions			
	Challenges	**Potential solutions**	**Phase**
1	Data Quality and Size	Digital voltage-time series traces instead of scanned images - preferably standard 12 lead for clinical settings, collected from a diverse population, including various age groups, ethnicities, and underlying health conditions. Ensure expert data annotation and verification.	Development, Validation
2	Data Imbalance	Weight balancing, SMOTE, Focal loss, N-folds, ECG generation, augmentation, balancing loss functions etc.	Development
3	Model Reproducibility	Follow standard reporting frameworks such as STARD-AI, TRIPOD-AI, CLAIM and open codebase for model replication.	Development, Validation
4	Robustness	A well balanced, high quality data with good real-world representation, data augmentation, synthetic data, noise induction, cross domain data	Development, Validation
5	Transparency, Lack of explainability	Standard XAI model and develop new methods for model interpretability	Development, Validation, Deployment
6	Generalizability and Bias	Socio-demographically diverse datasets, multi-center data; multi-national cohorts, sensitivity analysis for out-of-distribution shift or covariate shift, prospective	Development, Validation,

		studies and clinical trials; risk of bias assessment tool for prediction models such as PROBAST–AI[183]	Deployment
7	Privacy and security	Differential privacy, decentralized distributed models, federated learning, swarm learning, blockchain technology	Development, Validation, Deployment
8	Regulatory Compliance intellectual property concerns	Adaptive standards and streamlined regulatory approval process such as health Canada guidelines, EU guidance, FDA guidelines.	Deployment
9	Human factors, uptake by healthcare professionals, personal liability	Train and educate healthcare professionals, integration with workflow, and the design of user interfaces; transparent, clinically validated AI system.	Deployment
10	Scalability	Comprehensive policy, logistics, technical, and financial planning from the government body.	Deployment

Table 2: Challenges and potential solutions for ECG based AI/ML research.

4.2 Data Imbalance:

Data imbalance in terms of class proportions refers to a situation where the training size available for class of interest is disproportionately smaller compared to other classes.[184] Models trained on imbalanced datasets with numerous features yet sparse observations per feature risk overfitting, leading to poorer performance on external datasets. Few papers have discussed the problem and adopted various data level or algorithm level approaches to handle data imbalance issues. Approaches such as weight balancing,[57] synthetic ECG generation,[43] focal loss,[185] Synthetic Minority Oversampling Technique (SMOTE),[117] and N-folds[117,130,186] are most frequently used.

4.3 Model Reproducibility

A lack of standardized scientific reporting and external reproducibility has hampered the integration of DL models in ECG analysis.[187] The inherent complexity of these models, coupled with numerous variations in development and design, necessitates a precise and detailed description of methods. Yet, the field suffers from inconsistent methodological reporting and a lack of standardization, as evident in existing studies.[188] Standards for Reporting of Diagnostic Accuracy Study-AI (STARD-AI) and Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis-AI (TRIPOD-AI) are standard guideline for AI/ML specific reporting.[189] Similarly, the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) provides comprehensive guidelines for the wide-ranging use of AI in medical imaging, particularly focusing on aspects of model development.[190] It's unclear to what extent existing studies have strictly complied with established guidelines. Standard definitions, including the characterization of cohorts used for assessment of model performance vary across publications,

leading to ambiguity in discerning models that have undergone external testing from those prone to bias or overfitting.

Figure 5: Histogram of increasing number of AI-ECG publications with external validation. Data and source code availability for reproducibility and standardized scientific reporting.

Moreover, unlike traditional clinical predictive models where the publication of prognostic formulas is mandatory for clinical use,[191] the reproducibility and external testing of ECG AI/ML models are significantly constrained. Increasingly, the models' external validation are included as part of the studies, with most of them demonstrating validity in independent, publicly available datasets (Figure 5 – left panel). However, source code and accompanying documentation necessary for replicating model training and testing are often missing (Figure 5 - right panel). Some publications explicitly state that the codes are proprietary, while others are ambiguous, with several indicating potential code sharing upon request and others providing no information on code availability.

4.4 Robustness:

Ensuring that the ECG-based AI model is robust when exposed to the changing environment expected during deployment is another major challenge. For example, the study of Han et al.[192] demonstrated that subtle perturbations, barely noticeable on an ECG, could mislead a model, highly accurate in diagnosing AF, into incorrectly identifying NSR as AF with considerable certainty, despite the ECG appearing unchanged to a human expert. Establishing efficient data pipelines is essential for the seamless flow and processing of ECG data.[193] This involves the integration of data collection, preprocessing, and transformation processes that are scalable and secure. The challenge is to create efficient data pipelines that facilitate the rapid processing of large volumes of data while ensuring the integrity and confidentiality of sensitive health information.[10] In addition, the computational demands of AI/ML models, especially for those processing large datasets or employing complex algorithms, are substantial, and only a limited number of studies,[194] have implemented and validated optimized frameworks while considering low-resource devices and computation power for real-time ECG analysis.

4.5 Transparency and Lack of Explainability

Although existing models such as GradCAM or saliency maps partially interpret model decisions,[195] no perfect method is available for explainability. Integration of ECG-based AI/ML models into clinical workflow risk automation bias and over-reliance. Lack of explainability may produce algorithmic aversion, causing clinicians to distrust AI recommendations and necessitating improved transparency for clinical adoption. While XAI methods can interpret model outcomes at a high level, the utility of current techniques for model explainability in clinical tasks is still in question.[140,196] In particular, the underlying algorithms of commercial products pose significant challenges in assessing model failures due to intellectual property concerns. The 'black box' nature of these algorithms, characterized by their use of millions of parameters and intricate fine-tuning processes, exacerbates this issue.[197] So, developing a model that explains the model output is required for clinical applications.

4.6 Generalizability and Bias

Lack of standardized digital ECG acquisition across clinical and consumer settings as well as the challenge of poor-quality data limits the model generalizability of the ECG-based AI/ML models in deployment environments. For example, current AI tools developed for digitized ECG traces may not translate well to the analysis of ECGs stored in scanned image formats, restricting their use in certain clinical scenarios.[4] Current research is probing algorithmic failure during out-of-distribution (OOD) shifts, leading to inaccurate classifications and uncertainties, highlighting the complexity of maintaining model robustness across varying data distributions.[198] For example, Vranken et al.[199] investigated uncertainty estimation in DNN-based ECG classification, emphasizing the importance of accurate uncertainty estimation for quality control in clinical deep learning applications.

To avoid biases and ensure the generalizability of the AI models, prospective validation with multisite and multinational data is critical before incorporation into clinical practice.[53] Only a few studies, such as Ulloa et al.[200] validated their model across 10 clinical sites and with external temporal data testing sets. Yet, as AI models transition across sites, their performance can diminish, underscoring the importance of designing algorithms that maintain efficacy across diverse populations.[201] Predictive models must mirror the characteristics of the entire study population to ensure that disease representations and interventions are unbiased and universally applicable. However, AI technologies often carry inherent biases, potentially exacerbating health equity disparities.[202] Studies have demonstrated that AI algorithms can perpetuate racial bias, with specific populations receiving preferential treatment based on skewed risk scores. The key question is whether these models maintain consistent diagnostic performance irrespective of racial or ethnic backgrounds.[203] Only a handful of paper[4,204] have validated their model with racially diverse cohorts, but validation of their AI-ECG model with multisite or multi-national data is still pending. Noseworthy et al.[205] demonstrated that ECG characteristics vary by race, and the generalizability of the models can be affected by patient selection, with variable performance among diverse ethnic, racial, age, and sex groups. Harmon et al.[206] used temporal evaluation as well as evaluated the algorithms with respect to age, sex, race, and ethnicity. Sun et al showed

the performance of mortality prediction varied across patients with different primary diagnosis.[75] The potential of AI-ECG algorithms to mirror and intensify existing racial and ethnic disparities presents a significant challenge in clinical implementation.[3] Furthermore, data used to train these algorithms can lack diversity, with women, minority groups, and specific socio-economic sections underrepresented, reflecting real-world biases in healthcare outcomes. The PROBAST-AI tool, designed for evaluating the risk of bias and the applicability of diagnostic and prognostic prediction model studies, offers a structured approach to mitigate bias in AI-driven healthcare predictions.

4.7: Privacy and Security:

Privacy and security concerns are heightened in healthcare data due to the potential for re-identification, propelled by sophisticated analysis algorithms and vast datasets.[207,208] Traditional privacy measures like pseudonymization and anonymization may falter, as evidenced by studies demonstrating high re-identification rates using demographic attributes.[208] As medical data often require handling personal or sensitive information, ensuring privacy becomes a complex challenge. However, innovative approaches like differential privacy[209,210] introduce random noise to datasets, safeguarding individual identities while maintaining data utility for algorithm training.

Additionally, privacy-preserving AI technologies are evolving towards decentralized, distributed systems to ensure data security and protection of sensitive patient data.[211–213] These systems, such as federated learning and Swarm Learning,[214] keep data localized, reducing central vulnerabilities and promoting cooperative learning without a centralized command center. For example, Goto et al.[135] use federated learning approaches to detect HCM and to differentiate it from other cardiac conditions using ECGs with robust generalizability across multi-institute and multinational cohorts. Vikhyat et al[210] use federated learning-based ECG models which allow collaborative model training without sharing data between multisite hospitals in Canada. Further advancements in cryptographic techniques, like blockchain technology,[215] homomorphic encryption, and secure multi-party computation (SMPC),[211] provide robust layers of data protection. These methods, including cryptographic protocols, have shown promise in sensitive fields like cardiology, genetic sequencing, in maintaining participant privacy to a high degree.[216] Establishing a rigorous ethical framework and robust regulatory measures to address these challenges is essential, ensuring that AI/ML models in healthcare provide equitable, secure, and unbiased clinical solutions.

4.8: Ethical and legal considerations

Incorporating AI into clinical practice presents significant ethical and legal challenges,[217] necessitating robust ethical frameworks and regulatory guidelines for its application to guarantee the integrity of data and models.[218,219]

Although Health Canada, the FDA and EU-MDR have become increasingly active in advancing legislative initiatives on the ethical aspects of AI, the unique adaptability of AI in ECG-based diagnosis or prognosis poses distinct challenges.[9] Recent studies[220] have revealed that supposedly de-identified data can still contain extractable patient information. For example, ECG signals could be sufficient to identify an individual, especially from the wearable ECG device or mobile[220] The FDA and EU-MDR apply rigorous guidelines to medical technologies, mandating that applications like the Apple Watch IRNF App and Samsung Watch ECG Monitor app, which pose varying levels of risk to user privacy, comply with established regulatory standards. Similarly, Health Canada released a draft premarket guidance for machine learning-enabled medical devices recently which guides model design for post-market device and model monitoring. However, a robust legal framework addressing AI ECG-based clinical decision-making is yet to be established.[221] Additionally, compliance with data protection laws such as The Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR) is critical, especially given the sensitivity of patient data used in AI models. Collaboration among AI developers, healthcare providers, and regulatory authorities is crucial in shaping clear, adaptive standards and streamlined approval processes, ensuring AI's safe and ethical utilization in clinical care.

4.9. Professional liability and Human factors:

The issue of professional liability may arise as AI influences clinical decision-making, potentially obscuring accountability in cases where AI-assisted decisions result in adverse outcomes.[222] This situation is complicated by the reliance of less experienced clinicians on AI, which can lead to diagnostic discrepancies and increased liability risks.[222,223] The opaque nature of AI 'black box' models, where the decision-making process remains unclear, further complicates responsibility attribution between clinicians and AI developers. For example, athlete participation in screening with ECG helps identify cardiovascular abnormalities that elevate risk effectively. Yet, institutions fear that standardizing such screenings could increase physician and institutional liability, especially if an undetected abnormality leads to an athlete's adverse event during competition.[224] Thus, a transparent, clinically validated AI system that enhances patient outcomes is essential for overcoming skepticism and ensuring clinician acceptance.

Human factors, such as the acceptance of clinicians, integration with workflow, and the design of user interfaces, are critical for the successful deployment of AI/ML models in healthcare.[225] It is crucial to have comprehensive and ongoing training programs to enable healthcare staff to adopt clinical applications and familiarize with AI systems. For instance, Sandhu et al.[226] have reported that integrating AI can complicate clinical workflows and team dynamics, especially in high-pressure environments, thus highlighting the need for training and digital literacy among healthcare staff. Overcoming these socio-technical challenges is essential for bridging the gap between AI's potential and its practical application in clinical care.

4.10: Scalability and Global Implementation:

Due to limited resources and diverse regulatory environments, AI/ML healthcare solutions face scalability and deployment challenges in low- and middle-income countries (LMIC)s.[227] However, the reliance on specialized training for quality ECG diagnostics emphasizes the need for AI/ML implementation in these regions, where human expertise may often be lacking. FDA-approved wearable AI/ML ECG tools offer a viable solution for monitoring cardiac patients at high risk in LMICs, particularly in remote areas with a scarcity of subspecialized cardiologists. Case studies, such as Krones and Walker's work in rural Brazil using a ResNet model and XGBoost for heart disease detection,[228] alongside initiatives like Wadhwani AI[229] in India, Ubenwa AI[230] in Nigeria, demonstrate the global effort to utilize AI in addressing healthcare challenges, including cardiovascular and neurological disorders. However, deploying AI/ML in LMICs faces challenges such as limited access to necessary devices like smartphones and wearables, affordability and availability of high-speed internet, infrastructure and government support systems, training barriers, and more. The global disparity in physician distribution, the accessibility, and affordability is another major challenge, and the overall quality of healthcare services is linked directly to the economic status of these regions, influencing the scalability and effectiveness of AI healthcare solutions.[231] Adopting AI-enhanced ECG solutions in LMICs requires comprehensive policy, logistics, technical, and financial support planning.

CONCLUSION

AI/ML ECG models are increasingly demonstrating value for diagnosis, prognosis and early detection purposes and have captivated the attention of the healthcare community as they can directly impact patient care. Despite the significant advancement of AI, the successful development and deployment of AI/ML-enhanced ECG-based models into clinical applications remains relatively sparse, indicating that its full potential to enhance patient outcomes is yet to be fully realized. This review examines representative studies at various stages of workflow from proof-of-concept to FDA-approved deployed models for clinical applications and sheds light on the essential considerations and challenges for deploying AI/ML ECG models in clinical applications. The reporting of existing proof-of-concept publications in ECG deep learning is inconsistent, often lacking scientific reporting, and being primarily tested on internal or single-site data. Our findings support the need for standardized pipeline and evaluation criteria to bridge the gap between development of innovative AI/ML ECG-based models and their practical, ethical implementation in healthcare settings. In addition, compliance with standardized scientific reporting guidelines, alongside testing with external datasets, could significantly enhance the field's credibility and reproducibility. However, the impact of AL/ML ECG-based models on cardiology is poised to expand significantly. To ensure that investments in AI translate into meaningful clinical benefits rather than leading to disillusionment, it is crucial to maintain a balanced approach to developing and implementing these technologies. Standardized protocols, guidelines and regulations on end-to-end pipelines from development to clinical deployment can pave the way for ECG based AI/ML to fully realize its promise in enhancing patient health and shaping the future of cardiology.

Acknowledgments
Dr. Kaul holds the Canadian Institutes of Health Research Sex and Gender Science Chair and a Heart & Stroke Chair in Cardiovascular Research. We sincerely thank Ellen Pyear at the Canadian VIGOUR Centre, for assistance with the images used in this paper and Lisa Soulard for editorial assistance.

Patient Consent Statement
The authors confirm that patient consent is not applicable to this article.

Funding Sources
The study was supported by the Canadian Institutes of Health Research grant # 178158.

Disclosures
The authors report no conflict of interest.