UNIVERSITÀ DEGLI STUDI DI MILANO

DOCTORAL SCHOOL OF COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE


Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science
(XXXVII Cycle)

# AI-Driven Atrial Arrhythmia Detection: Development, Cross-Comparison and Uncertainty Quantification of Algorithms for Clinical Continuous ECGs

INF/01

ADVISORS:
Dr. Massimo Walter RIVOLTA
Prof. Fabio BADILINI
Prof. Roberto SASSI

DOCTORAL DISSERTATION OF:
**Md Moklesur RAHMAN**

DIRECTOR OF DOCTORAL PROGRAMME:
Prof. Roberto SASSI

Academic Year 2023/2024

# *Abstract*

**Background**: Atrial arrhythmias, particularly atrial fibrillation (AF), are prevalent cardiovascular disorders characterized by irregular heart rhythms originating from the atria, and affecting approximately 2-3% of the global population. These conditions are associated with increased risks of stroke, heart failure, and other severe complications. Traditional detection methods, primarily based on electrocardiograms (ECGs) analyzed by clinicians, are often time-consuming and prone to human error, especially when dealing with long-term monitoring (Holter recordings) and subtle, intermittent atrial arrhythmias. Recently, the development of artificial intelligence (AI)-based methods has garnered significant attention for automated AF detection from ECGs.

**Challenges**: Developing AI-based, particularly deep learning (DL), models to accurately detect atrial arrhythmias presents several significant challenges. First, extracting invariant representations across subjects of these arrhythmias is complex, necessitating high-quality annotated data and a substantial cohort of patients to ensure robust model training. Second, ECG datasets are typically imbalanced due to the scarcity of abnormal cases, complicating model training and evaluation, which can lead to bias and reduced performance in detecting rare arrhythmic events. Third, current methods are often validated on smaller patient populations of Holter recordings, which limits their clinical applicability and generalization, thereby restricting their effectiveness in diverse real-world settings. Finally, despite the promising performance of DL models in arrhythmia detection, their susceptibility to overfitting necessitates the exploration of uncertainty quantification to ensure safe integration into clinical practice.

**Objectives**: This thesis aims to design and develop DL models applied to ECG data for the automatic detection of AF from Holter recordings. Furthermore, it seeks to compare the performance of state-of-the-art models and commercial software solutions with the proposed model using a large, retrospective cohort of clinical data. Another key objective is to quantify the uncertainty in AF detection to assess the model's prediction confidence and improve its clinical reliability.

**Methods**: We obtained 1,346 Holter recordings from 1,346 distinct patients at Groupe Hospitalier Ambroise Paré in Paris, France, each with diverse

cardiac conditions. We developed a DL model for arrhythmia detection, focusing on residual attention models for comprehensive cross-comparisons with state-of-the-art models and two rule-based algorithms: (1) ABILE, a commercial software by AMPS LLC, New York, and (2) CBR, a research-based solution developed at the Center for Biological Research at UCSF, San Francisco. To enhance model performance, we systematically reviewed and applied various data augmentation techniques to improve the diversity and robustness of the training data. Furthermore, we investigated the impact of annotation errors (noisy labels) on model accuracy and implemented strategies to mitigate their effects. Additionally, we quantified the uncertainty in our DL model to assess prediction confidence and benchmarked 11 uncertainty quantification (UQ) methods for robust AF detection.

**Results**: The proposed DL model achieved 92.8% sensitivity and 91.5% specificity, outperforming state-of-the-art DL models. Moreover, when compared with the ABILE model, the proposed model achieved 95.1% sensitivity and 96.3% specificity, demonstrating superior specificity relative to ABILE's 48.9%, though with a slight reduction in sensitivity from ABILE's 98.4%. Additionally, in comparison with the CBR, which obtained 44.2% sensitivity and 99.9% specificity, the proposed model delivered a more balanced performance. Data augmentation techniques may improve the model's generalization and accuracy; however, in this context, they showed limited performance gains. Data augmentation techniques may improve model generalization and accuracy. However, in this context, it showed limited performance increase. The study of noisy labels provided valuable insights into model resilience. The model was found resilient up to 40% of a random change in label annotation. Finally, integrating UQ showed improved model's prediction confidence.

**Conclusions**: This research advances atrial arrhythmia detection through DL, offering potential improvements in clinical diagnostics and patient monitoring using Holter recordings. The methodologies and insights presented in this thesis lay a foundation for future research in cardiac arrhythmia detection using DL, addressing key challenges and enhancing the applicability of these models in clinical settings.

# *Acknowledgements*

Completing this PhD thesis has been a long and challenging journey, and I am deeply grateful to the many individuals and organizations who supported me throughout this journey.

First and foremost, I would like to express my sincere appreciation to my supervisors, Dr. Massimo W. Rivolta, Prof. Fabio Badilini, and Prof. Roberto Sassi. Your unwavering support, insightful guidance, and patience—especially during moments when my time-management faltered—were invaluable. Your kindness and understanding made me feel at home, even though I was thousands of miles away (7000 km), and eased the distance from everything familiar, reminding me that no matter how far I was, I was never truly alone. The warmth and sense of belonging you fostered will remain with me long after this journey, reminding me of the connection and support I was fortunate to experience.

To my colleagues and friends at the BiSP lab at the University of Milan, thank you for the stimulating discussions, collaboration, and camaraderie. A special thanks to Silvia Ibrahimi, whose support was invaluable throughout this journey. Our many conversations with lab mates, whether about science or life, helped me grow in ways I hadn't imagined, and for that, I am deeply grateful. I am also grateful for the lab's warm atmosphere, including the ever-dependable coffee machine, which played a subtle yet vital role in keeping me energized and focused.

I would like to extend my sincere thanks to *CardioCalm srl* for their financial support, which made this research possible. Your funding was crucial in enabling the progress of this work, and I am deeply appreciative.

Lastly, I would like to express my heartfelt gratitude to all the participants and collaborators involved in this research, especially the Cardiovascular Research Institute at the University of California, San Francisco, and Dr. Geoffrey Tison. Your directions were invaluable.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| **AF** | Atrial Fibrillation |
| **AFDB** | MIT-BIH Atrial Fibrillation Database |
| **AFL** | Atrial Flutter |
| **ANN** | Artificial Neural Network |
| **AT** | Atrial Tachycardia |
| **AUC** | Area Under Curve |
| **BE** | Batch Ensemble |
| **CNN** | Convolutional Neural Network |
| **CRN** | Convolutional Recurrent Network |
| **DA** | Data Augmentation |
| **DC** | Dice Coefficient |
| **DE** | Deep Ensemble |
| **DDCAEs** | Deep Denoising Convolutional Autoencoders |
| **DGM** | Deep Generative Models |
| **DL** | Deep Learning |
| **ECE** | Expected Calibration Error |
| **ECG** | Electrocardiogram |
| **EDL** | Evidential Deep Learning |
| **FLC** | Forward Loss Correction |
| **FID** | Fréchet Inception Distance |
| **FPR** | False Positive Rate |
| **FLOPs** | Floating-Point Operations |
| **GAN** | Generative Adversarial Networks |
| **GMM** | Gaussian Mixture Model |
| **GRU** | Gated Recurrent Unit |
| **ILR** | Implantable Loop Recorders |
| **iVOGN** | Improved Variational Online Gauss-Newton |
| **KNN** | K-Nearest Neighbour |
| **KMMD** | Kernel Maximum Mean Discrepancy |
| **KL** | Kullback–Leibler |
| **LBBB** | Left Bundle Branch Block |
| **LA** | Laplace Approximation |
| **LLLA** | Last Layer Laplace Approximation |
| **LTAF** | MIT-BIH Long Term Atrial Fibrillation Database |
| **MAP** | Maximum A Posteriori |
| **MAE** | Mean Absolute Error |

| | |
|---|---|
| **MC** | Monte Carlo |
| **MCD** | Monte Carlo Dropout |
| **MCM** | Markov Chain Model |
| **MSE** | Mean Square Error |
| **MFVI** | Mean Field Variational Inference |
| **MIT-BIH-AD** | MIT-BIH Arrhythmia Dataset |
| **MI** | Myocardial Infarction |
| **MMD** | Maximum Mean Discrepancy |
| **NLL** | Negative Log-Likelihood |
| **NSR** | Normal Sinus Rhythm |
| **PAC** | Premature Atrial Contraction |
| **PMSD** | Percent Mean Square Difference |
| **PE** | Packed Ensemble |
| **PRISMA** | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |
| **PPV** | Positive Predictive Value |
| **PVC** | Premature Ventricular Contraction |
| **RBBB** | Right Bundle Branch Block |
| **ResNet** | Residual Neural Network |
| **RNN** | Recurrent Neural Network |
| **RMSE** | Root Mean Squared Error |
| **RRI** | R-R Intervals |
| **RTA** | Residual Temporal Attention |
| **SE** | Snapshot Ensemble |
| **SGD** | Stochastic Gradient Descent |
| **SNR** | Signal-to-Noise Ratio |
| **SR** | Sinus Rhythm |
| **SWA** | Stochastic Weighted Average |
| **SVGD** | Stein Variational Gradient Descent |
| **SVM** | Support Vector Machine |
| **VAE** | Variational Auto Encoder |
| **VT** | Ventricular Tachycardia and Fibrillation |

*Dedicated to my parents*

# Chapter I

# Introduction

## 1.1 Motivation

Atrial arrhythmias, particularly atrial fibrillation (AF), represent a significant and growing challenge in cardiovascular disease due to their high prevalence and substantial impact on patient outcomes. AF is the most common sustained arrhythmia, affecting millions of individuals worldwide and contributing to increased risks of stroke, heart failure, and premature mortality [1]. As the global population ages and the prevalence of risk factors such as hypertension and diabetes rises, the burden of atrial arrhythmias is expected to escalate, necessitating advanced diagnostic and monitoring tools. Continuous electrocardiogram (ECG) is used as a critical technology for real-time assessment of cardiac electrical activity, offering the potential to detect and manage atrial arrhythmias more effectively. However, the performance of various algorithms designed for detecting these arrhythmias from continuous ECG data varies widely, influenced by factors such as the quality of the data, the specific algorithms used, and their validation across diverse patient populations. This variability underscores the urgent need for comprehensive cross-comparative studies to evaluate and enhance the performance of these algorithms, ensuring they are robust and generalizable across different clinical scenarios. This thesis aims to address this need by systematically comparing existing algorithms and developing new methodologies to improve their accuracy and reliability in detecting atrial arrhythmias from clinical continuous ECGs.

### Importance of Accurate Atrial Arrhythmia Detection

The accurate detection of atrial arrhythmias, particularly AF, is essential due to its substantial impact on patient outcomes and healthcare costs. The prevalence of AF is expected to double by 2060, driven by aging populations and the increasing prevalence of cardiovascular risk factors such as hypertension, diabetes, and obesity [1]. In the United States, an estimated 2.7 to 6.1 million people are currently living with AF, and this number is projected to rise to over 12 million by 2030 [2, 3].

One of the most critical outcomes associated with undiagnosed AF is stroke, with AF patients facing a fivefold increase in stroke risk compared to the general population [4]. Strokes caused by AF are often more severe, leading to greater morbidity, higher mortality rates, and longer hospitalizations. However, early detection through continuous ECG monitoring can dramatically reduce stroke risk. Studies have shown that timely initiation of anticoagulant therapy can reduce the risk of stroke by up to 64% in AF patients [5].

In addition to stroke prevention, early detection of atrial arrhythmias enables the prompt initiation of treatment strategies such as rate control, rhythm control, and anticoagulation, which are most effective when applied early in the disease course. Delays in diagnosis and treatment are associated with disease progression, including atrial remodeling, which increases the likelihood of persistent or permanent AF. Moreover, accurate and early arrhythmia detection can identify patients who are candidates for advanced interventions, such as catheter ablation, which has been shown to reduce AF recurrence and improve patient outcomes when performed early [6].

From a healthcare system perspective, the early detection and effective management of AF can lead to significant cost savings. AF-related strokes are associated with prolonged hospital stays, long-term disabilities, and considerable healthcare costs, with the average annual cost of treating an AF-related stroke exceeding $25,000 per patient in the United States [7]. By reducing the incidence of strokes and minimizing the need for repeated diagnostics, effective AF management helps to alleviate the economic burden on healthcare systems [8]. Continuous ECG monitoring plays a crucial role in achieving these outcomes, especially as the prevalence of AF continues to rise globally.

### The Need for Cross-Comparison of Algorithms

The development of algorithms for atrial arrhythmia detection has seen significant advancements in recent years, driven by the increasing availability of large-scale ECG datasets and improvements in computational technologies. Despite these advancements, the performance of various detection algorithms exhibits considerable variability, influenced by factors such as the quality and diversity of the datasets utilized, as well as the specific types of arrhythmias targeted. This variability underscores the necessity for rigorous cross-comparative studies to evaluate and benchmark the efficacy of different algorithms.

- Variability in Algorithm Performance: The effectiveness of atrial arrhythmia detection algorithms can vary significantly due to differences in the underlying data and methodological approaches. For instance, algorithms trained on specific types of ECG data, such as those from a particular demographic or clinical setting, may not perform as well when applied to diverse or heterogeneous populations. Studies have shown that algorithms that excel in detecting AF in one population may exhibit reduced accuracy or increased false positive rates when used in another population with different characteristics [9]. This variability highlights the importance of evaluating algorithms across a wide range of datasets to ensure their robustness and generalizability.

- Cross-Comparative Studies and Benchmarking: Cross-comparative studies are essential for identifying the strengths and weaknesses of different arrhythmia detection algorithms. By systematically comparing the performance of multiple algorithms on standardized datasets, researchers can assess various metrics, such as sensitivity, specificity, and predictive value, to determine which algorithms offer the most reliable detection capabilities. Such studies can also reveal how different algorithms handle specific challenges, such as noise, artifacts, or variations in ECG signal quality. For example, a comparative analysis may uncover that some algorithms are more adept at distinguishing between AF and other arrhythmias, while others may be better suited for detecting subtle or less frequent arrhythmias.

- Guiding Algorithm Development and Improvement: Insights gained

from cross-comparative studies can guide the refinement and development of more robust and generalizable algorithms. Understanding the limitations and performance characteristics of existing algorithms enables researchers to address specific weaknesses, such as improving algorithm sensitivity to rare arrhythmias or enhancing performance in noisy environments. Moreover, these studies can inform the design of future algorithms by highlighting best practices and identifying areas for innovation. For instance, integrating multiple data sources, such as combining ECG with patient demographics or clinical history, may enhance algorithm accuracy and utility.

- Implications for Clinical Practice and Research: Accurate and reliable arrhythmia detection is crucial for effective patient management and treatment. Algorithms that perform well across diverse populations and settings can lead to more timely and precise diagnoses, ultimately improving patient outcomes. Additionally, well-conducted cross-comparative studies contribute to the standardization of performance metrics and evaluation criteria, which can facilitate the adoption of new technologies in clinical practice. For researchers, these studies provide a benchmark for evaluating the impact of novel approaches and technologies, ensuring that advancements in the field are both scientifically rigorous and clinically relevant.

Overall, the cross-comparison of atrial arrhythmia detection algorithms is vital for advancing the field of cardiac monitoring. By evaluating and benchmarking different algorithms, researchers can enhance the development of more effective, generalizable, and clinically applicable detection methods. This approach not only improves algorithm performance but also supports better patient outcomes and drives innovation in cardiac healthcare technologies.

## 1.2 Background

### 1.2.1 Atrial Arrhythmias

Atrial arrhythmia refers to abnormal heart rhythms that originate in the upper chambers of the heart, called the atria. These arrhythmias cause the heart to beat irregularly or too quickly, which can reduce its ability to pump

blood effectively. To get a basic idea of the cardiac rhythms being classified in this work, a brief description of a few atrial arrhythmias follows.

- AF: AF is a common cardiac arrhythmia characterized by rapid and disorganized electrical impulses originating in the atria, the upper chambers of the heart. This irregular electrical activity causes the atria to quiver (fibrillate) instead of contracting effectively, leading to inefficient blood flow. AF increases the risk of stroke, thromboembolism, and heart failure due to the potential for blood to pool and clot within the atria. Clinically, AF presents with symptoms such as palpitations, fatigue, dyspnea, or may be asymptomatic. Management often involves anticoagulation therapy, rate or rhythm control medications, cardioversion, or catheter ablation. AF progresses through distinct clinical stages: silent AF $\rightarrow$ first detected $\rightarrow$ paroxysmal $\rightarrow$ persistent $\rightarrow$ long-standing persistent $\rightarrow$ permanent. Silent AF is asymptomatic and often undocumented, while paroxysmal AF lasts less than seven days. Persistent AF extends beyond seven days, long-standing persistent AF persists for over 12 months, and permanent AF occurs when the heart's rhythm cannot be restored to normal through drugs or ablation [10]. Importantly, more than one-third of patients initially diagnosed with paroxysmal AF are likely to progress to persistent AF within a decade [11].

- Atrial Tachycardia (AT): AT is characterized by an abnormally rapid heart rate, typically exceeding 100 beats per minute, originating from the atria [12]. Unlike AF, AT maintains a regular rhythm but with an elevated rate due to an abnormal electrical focus within the atrial tissue. This condition may present as paroxysmal (occurring in sudden bursts) or sustained, and it can result in symptoms such as palpitations, lightheadedness, and shortness of breath. Treatment may include pharmacologic intervention (antiarrhythmics or beta-blockers), catheter ablation, or lifestyle modification depending on severity and frequency.

- Atrial Flutter (AFL): AFL is a type of supraventricular tachycardia characterized by a rapid, but organized, electrical circuit within the atria, typically resulting in atrial rates of 240 to 400 beats per minute [12]. Although the atrial rhythm is regular, the rapid rate leads to suboptimal filling of the ventricles and reduced cardiac output. AF shares

many clinical features with AF, including an increased risk of stroke and thromboembolism. Symptoms include palpitations, dizziness, and fatigue. Management strategies often include anticoagulation, rate control medications, cardioversion, and catheter ablation.

- Premature Atrial Contractions (PACs): PACs are early electrical impulses that originate from ectopic foci in the atria, leading to premature heartbeats. PACs are typically benign and often asymptomatic, though individuals may experience palpitations or the sensation of a skipped heartbeat. PACs can be triggered by factors such as stress, caffeine, alcohol, or electrolyte imbalances. While usually not clinically significant, frequent PACs may warrant further evaluation to rule out underlying heart conditions. Treatment is generally unnecessary unless symptoms are severe, in which case beta-blockers or lifestyle adjustments may be recommended.

### 1.2.2 Ventricular Arrhythmias

Ventricular arrhythmias refer to abnormal heart rhythms originating in the ventricles, the lower chambers of the heart. These arrhythmias vary in severity, ranging from relatively benign conditions, such as premature ventricular contractions (PVCs), to potentially life-threatening disorders like ventricular tachycardia and fibrillation (VT). In ventricular arrhythmias, the electrical signals in the ventricles become disorganized or excessively rapid, which can severely impair the heart's ability to pump blood efficiently. Symptoms may include palpitations, dizziness, fainting, and in severe cases, cardiac arrest. Immediate treatment for dangerous ventricular arrhythmias often requires defibrillation or antiarrhythmic medications, while long-term management may involve an implantable cardioverter-defibrillator. Below, the two ventricular arrhythmias relevant to this thesis are described:

- PVCs: PVCs occur when the ventricles depolarize prematurely, resulting in an early heartbeat. This often leads to an irregular rhythm, which may manifest as palpitations or the sensation of a "skipped" beat. Although PVCs are typically benign and asymptomatic, frequent occurrences may be linked to underlying cardiac conditions, especially in individuals with structural heart disease. Common triggers for PVCs include stress, caffeine, alcohol, and electrolyte imbalances. While occasional PVCs usually do not require treatment, recurrent or

symptomatic cases may be managed with beta-blockers or catheter ablation.

- VT: Ventricular tachycardia is a life-threatening arrhythmia characterized by a rapid heart rate exceeding 100 beats per minute, originating in the ventricles. This accelerated rhythm can hinder the ventricles' ability to pump blood efficiently, leading to hypotension, syncope, or, in severe cases, progression to cardiac arrest. Immediate management of VT includes the use of antiarrhythmic drugs, electrical cardioversion, or defibrillation. Long-term prevention may require the implantation of an implantable cardioverter-defibrillator (ICD) to regulate the heart's rhythm and prevent future episodes.

### 1.2.3 Clinical Continuous ECG

Clinical continuous ECG is a vital diagnostic tool employed to monitor the heart's electrical activity over prolonged periods, typically in both inpatient and outpatient settings. In contrast to the standard 12-lead ECG, which provides only a brief snapshot of cardiac function, continuous ECG enables real-time, uninterrupted tracking. This continuous monitoring is particularly valuable in identifying transient cardiac events, such as arrhythmias which may not be detectable through short-duration tests.

Recent technological advancements have significantly enhanced the utility of continuous ECG, particularly through the incorporation of artificial intelligence (AI). AI algorithms have been shown to increase the accuracy of detecting abnormal cardiac rhythms and facilitate more efficient data interpretation. Furthermore, the growing prevalence of wearable devices equipped with ECG capabilities, such as smartwatches, has expanded the possibilities for long-term, patient-initiated cardiac monitoring. These innovations allow for the continuous assessment of heart function outside traditional clinical environments, thereby broadening the scope of cardiac care.

#### Types of Continuous ECG Monitoring

Continuous ECG monitoring encompasses several methods, each tailored to different clinical needs and durations of monitoring. These methods vary in terms of how they collect and transmit data, as well as the length of time

they can monitor the heart's electrical activity. The most common types are outlined below.

Holter Monitor: The Holter monitor is a portable, non-invasive device worn by patients for a period typically ranging from 24 to 48 hours. It continuously records the heart's electrical activity, allowing for the detection of transient arrhythmias or other abnormalities that may not be apparent during a standard ECG. During the monitoring period, patients are encouraged to engage in their normal daily activities, which provides a more comprehensive assessment of heart function across various conditions, such as rest, exercise, and sleep. Patients are often asked to keep a diary of symptoms (*e.g.*, palpitations, dizziness) to correlate with the recorded ECG data, enhancing diagnostic accuracy. The Holter monitor is particularly useful for diagnosing conditions like intermittent AF, bradycardia, tachycardia, or unexplained syncope. However, its relatively short monitoring period limits its ability to capture infrequent events.

Telemetry: Telemetry systems are primarily used in hospital settings to provide continuous, real-time monitoring of cardiac activity. Data from the patient's heart is wirelessly transmitted to a central monitoring station, where healthcare providers can observe heart rhythms and detect abnormalities instantaneously. Telemetry is often employed in critical care units or post-operative settings, where real-time monitoring is essential for detecting and responding to acute cardiac events such as myocardial infarctions or arrhythmias. Its immediate feedback system is crucial for patients with high-risk cardiac conditions who require constant observation.

Event Monitors: Event monitors are portable devices similar to Holter monitors but are typically used for longer periods, often up to 30 days. Unlike Holter monitors, event monitors do not record continuously but instead are activated either by the patient or automatically when abnormal heart rhythms are detected. This allows for prolonged monitoring, which is beneficial for patients who experience infrequent or unpredictable symptoms, such as occasional palpitations or fainting episodes. Event monitors can be external or implantable, and their ability to store episodic data makes them useful for diagnosing arrhythmias that occur sporadically.

Implantable Loop Recorders (ILR): ILRs are small, subcutaneously implanted devices that continuously monitor the heart's electrical activity

for extended periods, sometimes lasting up to several years. These devices are particularly useful for detecting rare, unexplained episodes of syncope, AF, or cryptogenic stroke, where more short-term monitoring options like Holter or event monitors might be insufficient. ILRs are programmed to automatically record abnormal heart rhythms and can be manually activated by the patient when symptoms occur. Due to their long-term monitoring capability and minimal interference with daily activities, ILRs provide a highly effective solution for diagnosing elusive or infrequent cardiac events.

### Clinical Applications of Continuous ECG

Continuous ECG monitoring plays a pivotal role in the diagnosis and management of various cardiac conditions. Its ability to provide uninterrupted, long-term data makes it invaluable for detecting intermittent or transient events that may be missed in conventional, short-duration monitoring. Key clinical applications include:

Arrhythmia Detection: Continuous ECG is crucial for diagnosing irregular heart rhythms, such as AF, ventricular tachycardia, and bradycardia. Many arrhythmias occur sporadically and may not be captured during a brief ECG. Long-term monitoring enhances the likelihood of detecting these intermittent events, thus facilitating more accurate diagnosis and timely intervention.

Ischemia Monitoring: Continuous ECG can assist in identifying silent or transient episodes of myocardial ischemia, which may not produce noticeable symptoms but are indicative of underlying coronary artery disease. Early detection of ischemia through continuous monitoring enables more prompt therapeutic interventions, potentially preventing more severe cardiovascular events.

Post-Surgical and Cardiac Care Monitoring: Continuous ECG is widely employed in the postoperative setting following cardiac surgeries or interventional procedures such as angioplasty. It allows for the real-time monitoring of heart function, providing early detection of complications such as arrhythmias or myocardial ischemia. This close surveillance is essential in the immediate recovery phase to ensure optimal patient outcomes.

Medication Monitoring: For patients undergoing treatment with medications that influence the heart's electrical activity, such as antiarrhythmic

drugs, continuous ECG monitoring is critical. It enables clinicians to assess the efficacy of the treatment, ensure that the medications are achieving the desired therapeutic effect, and identify any potential adverse effects, such as proarrhythmia or QT interval prolongation.

## 1.3 Research Problems

The rapid advancement of AI technologies has significantly impacted the field of cardiology, particularly in the detection of AF. However, the clinical applicability of these state-of-the-art AI methods is often constrained by several limitations. Most contemporary AI models for AF detection have been validated on relatively small-scale datasets that may not adequately represent the diverse array of clinical scenarios encountered in practice [13]. Additionally, many of these methods rely on 12-lead resting ECGs collected over short durations, which fail to capture the complexities associated with continuous ECG monitoring [14]. This limitation becomes particularly pronounced when dealing with Holter data, which involves long-term recordings and introduces challenges such as variations in signal quality and the need for effective detection of intermittent or irregular arrhythmias.

The clinical significance of accurate AF detection is profound, given the increasing prevalence of atrial arrhythmias and the emergence of new wearable ECG devices that offer continuous monitoring capabilities. These devices generate large volumes of continuous data, which underscores the necessity for robust and scalable AI-based approaches that can handle such data effectively. Despite their potential, current deep learning (DL)-based models face several critical challenges that need to be addressed to enhance their clinical utility:

- Data Quality and Quantity: High-quality, annotated ECG datasets, particularly those encompassing large and diverse patient populations, are scarce. The limited availability of such comprehensive datasets hampers the development and training of AI models that can generalize well across different demographics and clinical settings. Models trained on limited or non-representative data may exhibit poor performance when applied to broader or varied populations, thereby reducing their clinical relevance and effectiveness.

- Imbalanced Data: ECG datasets frequently suffer from significant class imbalances, where normal recordings vastly outnumber those with abnormal arrhythmias such as AFL or AF [15]. This imbalance poses a major challenge for training models, as they may become biased towards the majority class, leading to suboptimal performance on rare but clinically significant arrhythmias. Addressing this issue requires innovative approaches to dataset augmentation and algorithmic adjustments to ensure that models can accurately detect and classify less common arrhythmias.
- Invariant Representations: The variability and complexity inherent in ECG signals present a challenge for AI models in extracting invariant representations of atrial arrhythmias. Variations in signal quality, noise, and individual patient differences can affect the performance of detection algorithms [16]. Developing models that can effectively generalize across these variations is essential for achieving reliable and consistent results.
- Overconfidence in Predictions: DL models are known for their tendency to be overconfident in their predictions, which can be problematic, especially in clinical settings where accurate decision-making is crucial [17]. This overconfidence can lead to erroneous conclusions and inappropriate clinical actions if not properly managed. Ensuring that model predictions are well-calibrated and that output probabilities reflect true clinical uncertainty is vital for maintaining the trustworthiness and reliability of AI-based diagnostic tools.
- Mis-annotation (Noisy Labels): ECG data can sometimes be mis-annotated due to human error or variability in expert interpretation. Mis-labelled data can adversely affect model training and performance, leading to inaccurate predictions [18]. Addressing this issue involves developing methods for robust handling of noisy labels to ensure the reliability of training data.

## 1.4 Research Objectives

The main goal of this thesis is to advance the automatic detection of AF from continuous ECG data using AI, specifically DL techniques. To achieve this, the research focuses on several specific objectives aimed at improving the

accuracy, reliability, and practical applicability of AF detection algorithms. The detailed research objectives are as follows:

- Designing and Developing DL Models for AF Detection: This objective involves creating and refining AI techniques, particularly DL models, to enhance the automatic detection of AF from continuous Holter monitor recordings. The development process includes designing novel neural network architectures tailored to the unique characteristics of ECG signals, such as temporal dependencies.
- Evaluating and Comparing Model Performance Against State-of-the-Art Methods: A comprehensive performance comparison is conducted to benchmark the proposed model against existing state-of-the-art AI algorithms and clinical software. This evaluation utilizes a large retrospective cohort of clinical data to ensure a robust assessment across diverse patient populations and cardiac conditions. This objective aims to identify strengths and areas for improvement in the new models, providing insights into their potential for clinical integration.
- Quantifying and Addressing Uncertainty in Model Predictions: Given the critical importance of reliable and trustworthy predictions in clinical settings, this objective focuses on quantifying uncertainty associated with model predictions. Techniques such as uncertainty quantification and confidence interval estimation are employed to assess the reliability of the model outputs. This involves developing methods to measure prediction confidence and incorporating uncertainty estimates into the clinical decision-making process. By addressing prediction uncertainty, the research aims to enhance the trustworthiness of the AI models and support their integration into clinical workflows.
- Addressing Data Imbalance through Augmentation Techniques: ECG datasets often exhibit significant class imbalance, with a disproportionate number of normal recordings compared to those with atrial arrhythmias. This imbalance can lead to biased model performance, particularly in detecting less frequent arrhythmias. To address this issue, this objective focuses on implementing data augmentation techniques to balance the dataset. Strategies such as synthetic data generation are explored to improve model performance on underrepresented arrhythmia classes.

- Handling Noisy or Incorrect Labels: In practical settings, ECG data may contain noisy or incorrect labels due to various factors, including manual annotation errors or signal artifacts. This objective involves developing robust techniques to handle and mitigate the impact of label noise on model training. Methods such as noise-resistant learning algorithms are investigated to improve the models' resilience to inaccurate labels. By addressing label noise, the research aims to ensure more reliable model training and evaluation.

These research objectives collectively aim to advance the field of atrial arrhythmia detection by developing more effective and reliable AI-based techniques. By addressing challenges related to model performance, data imbalance, label noise, and representation learning, this thesis seeks to contribute to improved clinical outcomes and the integration of advanced cardiac monitoring technologies into routine practice.

## 1.5 Conclusion

In conclusion, the motivation for this thesis stems from the critical need to improve the detection of atrial arrhythmias from continuous ECGs. While significant progress has been made in the development of automated algorithms, challenges remain in terms of signal quality, variability in arrhythmias, patient diversity, computational efficiency, and interpretability. Cross-comparative studies are essential for evaluating existing algorithms and guiding the development of more robust solutions. Additionally, the potential for DL offer exciting opportunities to advance this field.

The implications of this research extend beyond the technical domain, with the potential to significantly impact clinical practice and patient care. By improving the accuracy and efficiency of atrial arrhythmia detection, this work aims to contribute to the early diagnosis and management of these conditions, ultimately reducing the burden of cardiovascular disease on patients and healthcare systems alike.

## 1.6 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter II: Design and Development of Deep Learning Model for AF Detection From Holter Recordings**
  This chapter outlines the process of data collection, preprocessing, model design, and evaluation for AF detection. It delves into the specific DL architectures used and provides a thorough explanation of the uncertainty quantification methods integrated into the model.

- **Chapter III: Systematic Review of Data Augmentation Techniques for ECG Signals in AI Applications**
  This chapter presents experimental findings, focusing on the performance of various data augmentation techniques for ECG signals.

- **Chapter IV: Uncertainty Quantification in DL Models for AF Detection**
  This chapter explores the results of uncertainty quantification techniques applied to DL models for AF detection. Performance comparisons among different uncertainty quantification methods are discussed, with a focus on clinical applicability. Additionally, an evidential DL model that incorporates evidence-based theory is used to quantify the uncertainty in AF and AFL detection.

- **Chapter V: The Impact of Label Noise on Deep Learning Models for Atrial Fibrillation Detection from Holter Recordings**
  This chapter details the experimental results concerning the impact of noisy labels on DL models for AF detection. The findings are compared to existing methods, and their relevance to clinical applications is thoroughly analyzed.

- **Chapter VI: Conclusions and Final Remarks**
  The final chapter encapsulates the primary contributions of this research to AI-based arrhythmia detection and offers recommendations for future studies in this evolving field.

# Chapter II

# Design and Development of Deep Learning Model for Atrial Fibrillation Detection From Holter Recordings

## 2.1 Introduction

Traditionally, AF detection from Holter recordings relies heavily on manual interpretation by trained clinicians. However, this process is labor-intensive, time-consuming, and prone to inter-observer variability [19]. Moreover, the increasing prevalence of AF necessitates scalable and efficient detection methods. In response to this challenge, automated systems have been used since the 1970s [20], which can automate AF detection, enhance diagnostic accuracy, and expedite patient care [21, 22].

The rapid advancement of AI has revolutionized medical diagnostics, offering promising performance for automated AF detection [23, 24, 25, 26]. Traditional ML techniques necessitate manual feature engineering, which is time-consuming and requires domain expertise. Additionally, these techniques may struggle to capture complex patterns and relationships within high-dimensional ECG signals. The dynamic nature of ECG data, influenced by factors like patient movement, environmental noise, demographics, and disease prevalence, poses challenges for traditional ML techniques.

DL emerged as a compelling solution to address the shortcomings of traditional ML models in AF detection [27, 14, 28]. DL models are typically

defined by their architectures, with convolutional neural networks (CNNs) and recurrent neural networks (RNNs) being prominent examples. These architectures, particularly CNNs and RNNs, demonstrate remarkable prowess in learning hierarchical representations and capturing temporal dependencies from raw data [29]. Nonetheless, DL models encounter challenges in effectively leveraging long-range dependencies within sequential data, which are prevalent in ECG signals. One notable advancement in DL architectures, the attention mechanism, presents a paradigm shift in addressing these challenges [30]. The integration of residual-attention mechanisms into DL models holds profound implications for AF detection from Holter recordings. By leveraging both residual connections and attention mechanisms, these models can effectively capture long-range dependencies and salient features within ECG signals, thus likely improving performance for AF detection.

Typically, many existing methods for AF detection from Holter recordings have been validated using a small number of patients *e.g.*, AFDB [31] and long-term AF dataset [32], limiting their applicability to real-world clinical settings. In contrast, our work aims to address this limitation by leveraging a new and clinically significant dataset comprising diverse patient populations and high-quality ECG recordings. Being very flexible in nature, DL models are prone to learning specific characteristics of the dataset used to train them, potentially resulting in a model that struggles to generalize in practice. By utilizing a larger and more representative dataset, we can enhance the generalizability and reliability of DL models. Particularly in the context of AF detection, the residual-attention DL model presents significant advantages over other DL-based methods when applied to Holter recordings. Firstly, the attention mechanism enables the model to focus on important segments of the ECG signal, enhancing interpretability and robustness [33, 34]. Secondly, the inclusion of residual connections facilitates the training of deeper networks, enabling a better capture of complex temporal dependencies in ECG segments [33, 34].

- We obtained a large retrospective dataset from Holter monitors and developed a residual-attention DL model specifically for detecting AF from these recordings.
- We conducted a comparative analysis of the performance between our proposed DL model and several leading state-of-the-art DL models.

- We conducted a comprehensive evaluation comparing the effectiveness of software solutions for detecting AF and AFL against the performance of our proposed DL model. To the best of our knowledge, this is the first study to perform such a comparison using Holter recordings.
- We examined the performance of our DL model across various demographic groups to assess its generalizability.

## 2.2 Related Works

This section provides a comprehensive overview of AI techniques utilized in the detection of AF. The review encompasses a range of methodologies, with a primary focus on ML techniques and advanced DL models. The DL approaches discussed are predominantly characterized by the use of CNNs, RNNs, hybrid models that combine both CNNs and RNNs, and attention-based mechanisms.

### 2.2.1 Machine Learning-Based Approaches

In the field of ML for AF detection, support vector machine (SVM) classifiers have garnered significant attention and application. Asgari *et al.* [24] utilized feature extraction across various frequency bands through stationary wavelet transform, coupled with an SVM for AF detection, achieving a sensitivity of 97.0% and specificity of 97.1% on the AFDB dataset. Similarly, Colloca *et al.* [35] extracted ten R-peak-related features from ECG signals and employed SVM for AF detection on both the AFDB and MIT-BIH NSR datasets, obtaining a sensitivity of 96.35% and a specificity of 98.91% respectively. Kumar *et al.* [36] utilized entropy-based features and applied a random forest (RF) classifier, achieving a sensitivity of 95.8% and specificity of 97.6%. Additionally, Czabanski *et al.* [25] computed various heart rate features—such as mean, median, and quartiles—which were subsequently used to train an SVM classifier for AF detection. These studies collectively highlight the widespread use and effectiveness of SVM classifiers in AF detection, demonstrating their robustness and adaptability across different methodologies and datasets.

Kennedy *et al.* [37] proposed an integrated approach combining RF and k-nearest neighbors (KNN) classifiers for AF detection based on R-R intervals

(RRI) in ECG signals. However, this RRI-focused method has limitations, notably its exclusion of additional relevant features such as P-wave analysis. Zabihi *et al.* [38] extracted features across multiple domains, including time, frequency, time–frequency, and phase space, and selected optimal features for classification using an RF classifier, reporting an overall performance score of 82.6% on the PhysioNet 2017 challenge dataset. Despite their efficacy, ML-based approaches generally require manual feature extraction and optimal feature selection, which can be cumbersome. These studies underscore the need for more comprehensive feature consideration in AF detection frameworks, highlighting the necessity for continued research to refine methodologies and enhance clinical applicability.

### 2.2.2 Deep Learning-Based Approaches

In AF detection, CNNs have been employed extensively. Xia *et al.* [39] utilized short-time Fourier transform and stationary wavelet transform to convert one-dimensional ECG signals into two-dimensional representations, subsequently applying CNNs to achieve sensitivity and specificity rates of 98.34% and 98.24%, respectively. Cai *et al.* [40] introduced a one-dimensional DenseNet for AF detection using 12-lead ECG recordings; however, the model's large scale presents practical limitations. Fan *et al.* [41] proposed a multi-scale CNN designed for single-lead ECG recordings, which demonstrated a sensitivity of 97.72% for 10-second segments, though its generalizability requires further validation on independent datasets. Shi *et al.* [42] developed a multiple-input deep learning model leveraging transfer learning and active learning to enhance classification performance, but the incorporation of hand-crafted features increased computational complexity. Tutuko *et al.* [43] proposed the CNN-based AFibNet model for two-class (normal vs. AF) and three-class (normal, AF, and Non-AF) detection tasks. Despite its focus on spatial features, the model overlooks temporal aspects and lacks interpretability. Prabhakararao *et al.* [44] introduced an ensemble method combining multiple CNN classifiers for multi-class arrhythmia classification, achieving average F1-scores of 84.5% and 88.3% on the PTBXL-2020 and Physionet-2017 datasets, respectively.

For RNN-based approaches, Maknickas *et al.* [45] utilized long short-term memory (LSTM) networks to classify ECG signals based on pre-computed QRS complex features. Sun *et al.* [46] developed a stacked LSTM network

for AF detection, addressing gradient issues to improve feature learning. Beak *et al.* [47] proposed a novel RNN algorithm for AF detection during sinus rhythm using 12-lead ECGs. Wang [48] combined CNNs with bidirectional GRU networks for AF and AFL classification, while Wang *et al.* [49] developed a model based on an 11-layer CNN and an improved Elman neural network, although requiring extensive and diverse training datasets. Ping *et al.* [15] proposed a hybrid model integrating an 8-layer CNN with a 1-layer LSTM, effectively managing long-term dependencies compared to traditional RNNs and multi-scale CNNs. Wang *et al.* [13] designed a DL model incorporating multi-scale convolution kernels and bidirectional GRU, demonstrating high accuracy on the AFDB and MIT-BIH arrhythmia datasets (MIT-BIH-AD). These RNN-based models, however, often fail to address the differential contributions of spatial and temporal features in ECG signals for effective AF detection.

Recent advancements in attention-based models include Jin *et al.* [50], who developed a twin attentional convolutional LSTM to extract multi-domain features from ECG signals, analyzing the impact of various input segments on prediction accuracy. Zhang *et al.* [51] proposed a dual-domain attention cascade network that utilizes channel-spatial and time series features to identify discriminative AF patterns, achieving accuracies of 99.49% and 99.28% in two-class and three-class tasks, respectively, on the CPSC-2018 dataset. Zhao *et al.* [52] integrated a temporal CNN with a self-attention mechanism to encode ECG heartbeat sequences, capturing both global and local features, and achieved a sensitivity of 91.85% on the MIT-BIH arrhythmia dataset. Li *et al.* [53] developed a model incorporating a self-complementary attention mechanism to extract both shallow and deep features from ECG signals, achieving AUC values of 99.79%, 95.51%, and 98.77% on the AFDB dataset, Physionet-2017, and CPSC-2018 datasets, respectively.

After reviewing the literature, it becomes clear that many existing methods for AF detection from Holter recordings have primarily been validated on relatively small patient cohorts, such as the MIT-BIH-AD dataset [54], AFDB [31], and the MIT-BIH long-term AF (LTAF) dataset [32]. While these datasets have been pivotal in advancing AF detection research, their limited sample size constrains the models' ability to generalize effectively across broader, more diverse populations. This issue raises significant concerns

about the robustness, clinical relevance, and scalability of these methods when applied in real-world settings, where patient variability is higher, and the detection environment is more complex. Expanding the size and diversity of the patient cohorts used for model training and validation remains a critical challenge for improving the clinical applicability and performance of AF detection techniques.

## 2.3 Dataset

In this chapter, we utilized three datasets to both develop and assess our DL model for detecting AF from Holter recordings. Among these, one dataset is private, while the remaining two—IRIDIA-AF and SHDB-AF—are publicly available. Below, we provide short descriptions of each dataset.

**Our Dataset:** The dataset consists of 661 Holter recordings obtained from 661 patients at Groupe Hospitalier Ambroise Paré in Paris, France. Each recording, with an average duration of approximately 23 hours, was captured using a Microport Spiderview Holter monitor. This system is a 2-lead setup with a sampling rate of 200 Hz and an amplitude resolution of $10\mu V$. The details about the dataset are described in [55]. The patients had an average age of around 60 years, and women represented about 39% of the recordings. Approximately 50% of the recordings (totaling 333) documented at least one episode of AF or AFL, with episode durations ranging from brief occurrences to the entire recording period (indicating chronic AF or AFL). The remaining recordings were predominantly in sinus rhythm (totaling 195), though they also included a notable number of PVCs (totaling 41), episodes of AT (totaling 61), and VT (totaling 31). The distribution of these conditions is depicted in Figure 2.1. The dataset shows that AF occupies 193,000 minutes, AFL accounts for 93,000 minutes, and AT covers 48,000 minutes, while normal sinus rhythm (NSR) encompasses 180,000 minutes. To ensure high data quality, the annotations were meticulously reviewed, with a minimum of 59 minutes per hour verified by a single cardiologist, thus minimizing noise and enhancing the accuracy of the analysis.

**IRIDIA-AF Dataset:** The dataset comprised 167 Holter records collected from 152 patients at an outpatient cardiology clinic located in Belgium [56]. The records were collected using a Microport Spiderview Holter recorder, which is the same device employed for our data collection (please refer

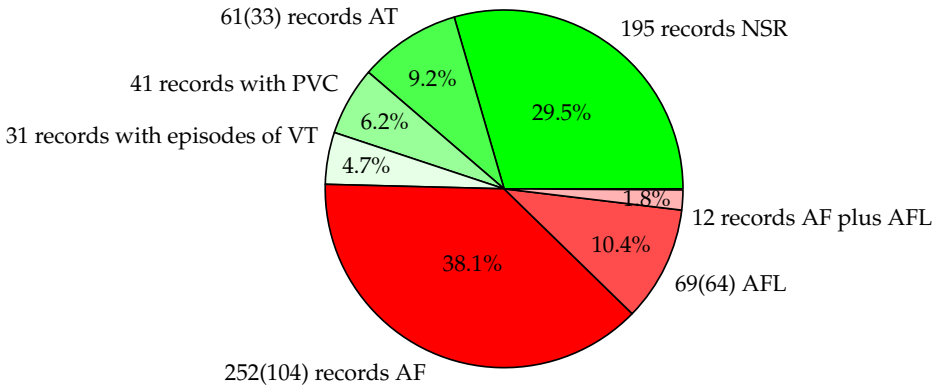



FIGURE 2.1: Distribution of patients by records in the dataset: Records with AF/AFL are shown in red, while records without AF/AFL are shown in green. The number of "chronic" records (*i.e.*, entire records under the labeled rhythm) is indicated in parentheses.

to the previous paragraph). Notably, records from patients with specific conditions, such as cardiac implantable electronic devices, persistent or permanent AF, or other cardiac diseases were excluded from the dataset. These exclusion criteria were implemented to ensure the homogeneity of the dataset and to focus the analysis on paroxysmal AF.

**SHDB-AF Dataset:** This dataset comprised ECG recordings from adult patients who underwent Holter monitoring between November 2019 and January 2022 in Japan [57]. The Holter monitors used were Fukuda devices, which recorded data from two leads—modified CC5 and NASA—at a sampling rate of 125 Hz. Each recording spans approximately 24 hours. In total, 147 Holter recordings were collected, from which a subset of 100 recordings, each corresponding to a unique patient, was selected. The data underwent preprocessing, including filtering with a zero-phase second-order infinite impulse response bandpass filter with a passband of 0.67–100 Hz to remove baseline wander and high-frequency noise. The recordings were then resampled to 200 Hz using an anti-aliasing filter. The patients exhibited a variety of cardiac conditions, including supraventricular arrhythmias such as AF, AFL, AT, and other supraventricular tachycardias, including Wolf-Parkinson-White syndrome and intranodal tachycardia.

**Preprocessing:** The dataset was collected from two distinct batches. The first batch (268 records) was used for training (250 records) and validation (18 records) and included several AT events. The second batch (393 records),

TABLE 2.1: Number of 10-s ECG segments (in thousand units) for each of the three datasets considered.

| Dataset | Training | | | | Validation | | | | Testing | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | | AF | | Non-AF | | AF | | Non-AF | AF | |
| | NSR | AT | AF | AFL | NSR | AT | AF | AFL | NSR | AF | AFL |
| Our dataset | 748 | 261 | 848 | 136 | 24 | 25 | 66 | 34 | 2324 | 498 | 124 |
| IRIDIA-AF | | | | – | | | | | 1872 | 536 | – |
| SHDB-AF | | | | – | | | | | 674 | 167 | 13 |

used exclusively for testing, did not contain AT episodes but included other challenging ventricular rhythms, such as ventricular tachycardia. Additionally, all 167 records from the IRIDIA-AF dataset and 100 records from the SHDB-AF dataset were exclusively used for testing purposes.

During the preprocessing phase, a third-order zero-phase Butterworth band-pass filter with cutoff frequencies of 0.5 Hz and 40 Hz was applied to mitigate baseline wander and reduce power line interference. Subsequently, each recording was segmented into 10-second windows without overlap. The number of 10-second segments for the training, validation, and testing sets is detailed in Table 2.1. In our classification scheme for binary classification, NSR and AT were collectively categorized as Non-AF, whereas AF and AFL were grouped under AF. AT was classified as non-AF due to its distinct characteristics in heart-rate stability, risk level, and treatment considerations compared to AF and AFL. Additionally, we explored a ternary classification scheme where NSR/AT remained "Non-AF", but AF and AFL were classified into separate categories.

## 2.4 DL Model

The DL architecture incorporated various layer types to perform both feature extraction and detection tasks. Figure 2.2 illustrates the schematic of the residual-temporal attention (RTA) DL model proposed in this study. This figure demonstrates the integration of an RTA block with a gated recurrent unit (GRU) layer. The addition of the GRU layer following the RTA block enhanced the model's ability to capture temporal dependencies and sequential patterns within 10-second ECG segments, thereby improving its accuracy in detecting AF [58]. The RTA block was utilized six times, with the initial number of kernels set to 32 and doubling every two iterations,

reaching a maximum of 128 kernels. This repetitive application of the RTA block enabled the model to extract hierarchical representations of the input signals effectively, which contributed to enhanced performance in AF detection. The RTA block itself consists of two components, which are detailed below.

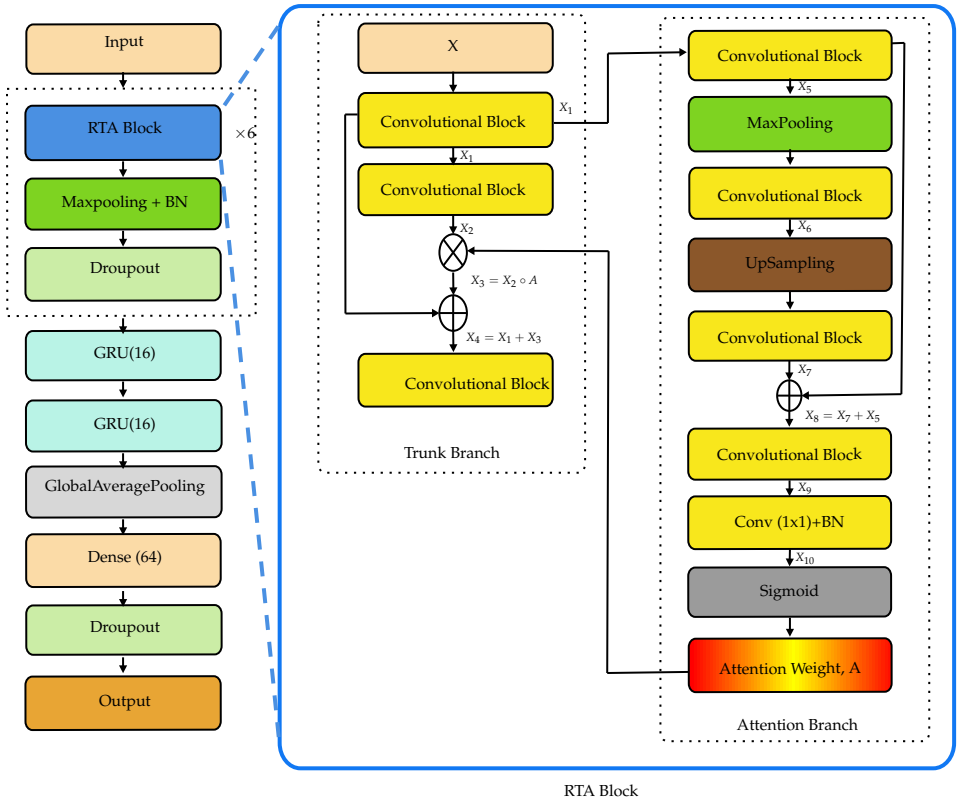



FIGURE 2.2: Diagram of the proposed DL architecture. Here, Conv and BN refer to the convolutional layer and batch normalization, respectively.

**Trunk Branch**: Assuming that $X$, a 10-second segment of ECG data, is provided as input to the trunk branch of the RTA block, the RTA block is repeated six times, with the output of each block serving as the input to the subsequent block. The input $X$ to the trunk branch passes through a convolutional block to generate a feature map $X_1$, which is then fed to the attention branch.

In the trunk branch, the feature map $X_1$ is passed through another convolutional block to generate $X_2$, which is element-wise multiplied with the attention map $A$ (obtained from the attention branch). The product

$X_3 = X_2 \odot A$ results in a refined feature map, which is then combined with $X_1$ using a residual connection. This final refined feature map, $X_3$, is passed through another convolutional block to generate the feature map $X_4$ before proceeding to the next layer of the main DL architecture.

**Attention branch:** The attention branch starts with the intermediate feature map $X_1$ from the trunk branch. This feature map is then passed through a convolutional block to generate the feature map $X_5$. To capture global features, the attention branch incorporates both down-sampling and up-sampling operations. Down-sampling is performed using max-pooling, while up-sampling is performed through the nearest-neighbour interpolation. These operations facilitate the extraction of features at different scales, thereby enhancing the model's ability to capture relevant information from the input data.

Following the down-sampling operation on $X_5$, a convolutional block is applied to expand the feature dimensions, yielding the feature map $X_6$. Subsequently, an up-sampling operation is performed, and the resulting feature map is fed into another convolutional block, which produces the feature map $X_7$. This feature map $X_7$ is then fused with the local feature map $X_5$ through a residual connection, resulting in the feature map $X_8$. The fused feature map $X_8$ is passed through an additional convolutional block to generate the refined feature map $X_9$, which further refines the attention map.

Finally, the feature map $X_9$ is processed by a convolutional layer with a $1 \times 1$ filter size and a sigmoid activation function. This operation produces the output feature map $X_{10}$, which contains the temporal attention weights, denoted as $A$. These attention weights are then used to adjust the importance of each feature within the trunk branch, enabling the model to prioritize the most relevant features of the input data, denoted as $X$.

**Implementation Details**: To optimize the model parameters, we employed the focal cross-entropy loss function as described by Lin *et al.* [59]. The focal loss is defined by the following equation:

$$\text{focal loss}(\mathbf{p}) = -\sum_{j=1}^{C} \alpha_j (1 - p_j)^{\gamma} \log(p_j) \tag{2.1}$$

where $C$ represents the number of classes (here, 2 and 3), $p_j$ denotes the

predicted probability for the $j$-th class, $\alpha_j$ is the class balancing parameter to address class imbalance, and $\gamma$ is the focusing parameter that down-weights easy samples. In our study, we set $\alpha_{\text{AF}} = 0.8$ and $\alpha_{\text{Non-AF}} = 0.2$, with $\gamma = 3$ for binary classification. For the three-class classification scenario, we assigned the alpha values as follows: $\alpha_{\text{Non-AF}} = 0.2$, $\alpha_{\text{AF}} = 0.3$, and $\alpha_{\text{AFL}} = 0.5$. The Adam optimizer was utilized with a learning rate of 0.001 to adjust the model parameters. The following hyperparameters were used during training: 50 epochs and a batch size of 128. To enhance training efficiency, a learning rate scheduler was implemented, which decreased the learning rate by 75% if no improvement was observed over six consecutive epochs. These hyperparameter values and the architecture of the DL model were selected based on their performance on the validation set.

## 2.5 Results and Discussions

The performance of the DL model was evaluated using key metrics including recall, false positive rate (FPR), and positive predictive value (PPV). Additionally, the area under the ROC curve (AUC) was computed to provide a comprehensive assessment of the model's overall performance in detecting AF. The AUC score was calculated using a one-vs-all approach. The recall, FPR, and PPV were calculated using the following formulas:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.2}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{2.4}$$

Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

### 2.5.1 Results for Non-AF and AF Classification

In Table 2.2, we present a comparative analysis of the performance of our proposed DL model against several state-of-the-art DL models on the test set

from our dataset. To evaluate the efficacy of our model relative to existing state-of-the-art DL models [27, 14, 28], we re-implemented these models from scratch, utilizing the same hyperparameters as those used in our proposed model. Notably, our DL model exhibits a marginal improvement in performance compared to the others. Specifically, in terms of recall, our model achieves scores of 0.915 for Non-AF and 0.928 for AF, surpassing the results obtained by Burke *et al.* [28] and Ribeiro *et al.* [14] for AF. Additionally, our model demonstrates superior performance in terms of AUC, with scores of 0.961 for Non-AF and 0.972 for AF, outperforming all other models listed, including those by Hannun *et al.* [27] and Ribeiro *et al.* [14]. The values highlighted in yellow represent the highest performance metrics observed, underscoring the superior accuracy and effectiveness of our proposed model in distinguishing between AF and Non-AF instances.

TABLE 2.2: Comparing the performance of the proposed DL model with other state-of-the-art DL models on our dataset. The best values are highlighted in bold.

| Method | Recall | | AUC |
|---|---|---|---|
| | Non-AF | AF | |
| Hannun *et al.* [27] | 0.913 | 0.870 | 0.949 |
| Ribeiro *et al.* [14] | **0.921** | 0.851 | 0.944 |
| Burke *et al.* [28] | 0.832 | 0.921 | 0.951 |
| Our model | 0.915 | **0.928** | **0.967** |

The efficacy of our proposed DL model for AF detection was further validated using two external datasets: the IRIDIA-AF and SHDB-AF datasets. For the IRIDIA-AF dataset, the model achieved a recall score of 0.942 for AF and 0.932 for Non-AF cases. On the SHDB-AF dataset, it attained a recall score of 0.925 for AF and 0.924 for Non-AF cases. Across all test sets, the model consistently achieved an AUC of over 0.960 for AF detection, demonstrating its strong ability to accurately distinguish between AF and Non-AF cases.

We further assessed the generalizability of our DL model across various demographic groups, with a particular emphasis on gender and age. The performance metrics are detailed in Table 2.3. In the gender-based analysis, the model exhibited recall rates for AF of 0.897 in males and 0.949 in females. The recall rates for Non-AF were 0.938 in males and 0.908 in females. Regarding age stratification ($\leq 60$ and $> 60$), the model demonstrated recall

TABLE 2.3: Performance of our proposed DL model across demographic groups on the test set of our dataset.

| Group | No of Records | Recall | |
|---|---|---|---|
| | | Non-AF | AF |
| Gender: Male | 258 | 0.938 | 0.897 |
| Gender: Female | 135 | 0.908 | 0.949 |
| Age $\leq 60$ | 164 | 0.904 | 0.934 |
| Age $> 60$ | 229 | 0.945 | 0.901 |
| VT | 31 | 0.832 | 0.973 |
| PVC | 41 | 0.934 | – |

rates for AF of 0.934 and 0.901, respectively, and recall rates for Non-AF of 0.904 and 0.945.

Additionally, we evaluated the performance of the model in two distinct patient cohorts: individuals with a cardiac condition of VT and those with PVC. In the VT cohort, the model achieved a recall rate for AF of 0.973 and a recall rate for Non-AF of 0.832. In the PVC cohort, the model attained a recall rate for Non-AF of 0.934. Notably, there were fewer than 18,000 10-second segments of AF in the VT group, and no 10-second segments of AF in the PVC group. Overall, these findings underscore the robustness and wide applicability of our DL model across diverse demographic groups, highlighting its potential for extensive clinical implementation.

### 2.5.2 Results for Non-AF, AF, and AFL Classification

In Table 2.4, we present a comparative analysis of the performance of our proposed DL model against several state-of-the-art DL models in the context of distinguishing between Non-AF, AF, and AFL cardiac conditions. Our model demonstrates superior performance across all evaluated metrics. Specifically, in terms of recall, our model achieved scores of 0.951 for Non-AF, 0.931 for AF, and 0.812 for AFL. These recall values surpass those of the other models, highlighting its enhanced sensitivity in identifying each condition. In terms of AUC, our model also outperforms the state-of-the-art models with scores of 0.981 for Non-AF, 0.973 for AF, and 0.942 for AFL. These AUC values reflect the model's overall effectiveness in distinguishing between the conditions, with our model achieving the highest scores across all three categories.

Comparatively, Hannun *et al.* [27] and Ribeiro *et al.* [14] exhibit strong performance, but our model consistently delivers higher recall and AUC values, particularly in the AF and AFL categories. Burke *et al.* [28] also shows competitive results, yet our model surpasses it in recall and AUC for Non-AF and AF.

TABLE 2.4: Comparing the performance of the proposed DL model with other state-of-the-art DL models for Non-AF, AF, and AFL on the test set of our dataset.

| **Model** | **Recall** | | | **AUC** | | |
|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Hannun *et al.* [27] | 0.925 | 0.889 | 0.718 | 0.967 | **0.936** | 0.935 |
| Ribeiro *et al.* [14] | 0.898 | 0.936 | 0.735 | 0.964 | 0.956 | 0.923 |
| Burke *et al.* [28] | 0.924 | 0.898 | 0.674 | 0.912 | 0.902 | 0.853 |
| Our Model | **0.951** | 0.931 | **0.812** | **0.981** | **0.973** | **0.942** |

### 2.5.3 Performance Comparison with Rule-Based Software

TABLE 2.5: Performance comparison of our model with rule-based software.

| **Metrics** | | **Our Model** | **ABILE** | **CBR** |
|---|---|---|---|---|
| Recall | Non-AF | 0.963 | 0.984 | **0.999** |
| | AF or AFL | **0.951** | 0.489 | 0.442 |
| | AF | 0.931 | **0.953** | 0.893 |
| | AFL | **0.798** | – | 0.068 |
| FPR | AF or AFL | 0.037 | 0.016 | **0.001** |
| | AF | 0.046 | 0.059 | **0.029** |
| | AFL | 0.016 | – | **0.0** |
| PPV | AF or AFL | 0.898 | 0.915 | **0.991** |
| | AF | 0.651 | 0.595 | **0.736** |
| | AFL | 0.909 | – | **1.00** |

Table 2.5 presents a comparative analysis of the performance of our proposed model against two rule-based software: (1) ABILE, a commercial software by AMPS LLC, New York [55], and (2) CBR, a research-based solution developed at the Center for Biological Research at UCSF, San Francisco [60]. The comparison is based on a subset of 193 records from a total of 393 records in the test set. Our model demonstrates a recall rate of 0.963 for

Non-AF cases, which is slightly lower than ABILE's recall rate of 0.984 and CBR's recall rate of 0.999. This suggests that while our model performs well in identifying Non-AF cases, it is somewhat less effective compared to CBR and slightly behind ABILE and CBR. In terms of recall for AF or AFL, our model achieves a rate of 0.951, markedly outperforming ABILE (0.489) and CBR (0.442). Additionally, our model exhibits a recall rate of 0.798 for AFL, which is substantially better than CBR's 0.068, although ABILE does not provide data for AFL.

Regarding FPR, our model has a rate of 0.037 for AF or AFL, which is higher than ABILE's 0.016 and CBR's 0.001, indicating that CBR is the most effective at minimizing false positives. For AF specifically, our model's FPR of 0.046 is lower than ABILE's 0.059 but higher than CBR's 0.029. Concerning PPV, our model's value of 0.898 for AF or AFL is lower than CBR's 0.991 and ABILE's 0.915. For AF, our model's PPV of 0.651 exceeds ABILE's 0.595 but falls short of CBR's 0.736. For AFL, our model achieves a PPV of 0.909, while CBR attains a perfect score of 1. Overall, our model excels in detecting AF and AFL with strong recall rates and demonstrates robust performance. However, CBR generally exhibits superior results in reducing false positives and achieving higher PPV, particularly for the Non-AF cases.

TABLE 2.6: Performance comparison of our model with rule-based software without chronic AFL case.

| **Metrics** | | **Our Model** | **ABILE** | **CBR** |
|---|---|---|---|---|
| Recall | Non-AF | 0.963 | 0.984 | **0.999** |
| | AF | 0.933 | **0.953** | 0.894 |
| FPR | AF or AFL | 0.037 | 0.016 | 0.001 |
| | AF | 0.022 | 0.015 | **0.001** |
| | AFL | 0.016 | – | **0.0** |
| PPV | AF or AFL | 0.750 | 0.874 | **0.986** |
| | AF | 0.827 | 0.879 | **0.986** |
| | AFL | 0.052 | – | **1.00** |

Additionally, we investigated the performance of our model after excluding the 24 chronic AFL records from the test set. The results are detailed in Table 2.6. In terms of recall, our model achieves a value of 0.963 for Non-AF cases,

which is slightly lower than ABILE's 0.984 and CBR's 0.999. This indicates that while our model performs effectively in identifying Non-AF cases, it is somewhat outperformed by CBR and slightly behind ABILE. For AF cases, our model's recall of 0.933 is comparable to ABILE's 0.953 and surpasses CBR's 0.894. This demonstrates that our model is competitive in detecting AF, with performance comparable to ABILE and superior to CBR.

Regarding the FPR, our model exhibits an FPR of 0.037 for AF or AFL, which is higher than ABILE's 0.016 and CBR's 0.001. This suggests that CBR and ABILE are more effective in minimizing false positives. Specifically for AF, our model's FPR is 0.022, slightly higher than ABILE's 0.015 and CBR's 0.001. For AFL, our model's FPR is 0.016, with CBR achieving a perfect score of 0.0, and ABILE not providing data for AFL.

In terms of PPV, our model has a value of 0.750 for AF or AFL, which is lower than ABILE's 0.874 and significantly less than CBR's 0.986. This indicates that CBR demonstrates the highest precision. For AF alone, our model's PPV is 0.827, lower than ABILE's 0.879 and CBR's 0.986. For AFL, our model's PPV is notably low at 0.052, while CBR achieves a perfect score of 1.

In addition, to make a fair comparison, we adjusted the thresholds of the DL model to match the recall values for Non-AF cases achieved by ABILE and CBR. The thresholds were set to 0.999 for CBR and 0.978 for ABILE. To determine these thresholds, we transformed the ternary classification into a binary one by combining the probability predictions of AF and AFL from the softmax output. Under this configuration, the recall values for AF were 0.862 for the threshold aligned with ABILE and 0.338 for the threshold aligned with CBR. When examining AFL detection within the same configuration, the recall values for AFL were 0.588 for the DL model, 0.0 for ABILE (as it does not detect AFL by design), and 0.068 for CBR.

Overall, our DL model demonstrates competitive performance, particularly in detecting AF and Non-AF cases. While ABILE and CBR outperform the DL model in reducing false positives and achieving higher PPV, especially for Non-AF cases, the DL model shows a clear advantage in AFL detection. Notably, ABILE does not provide results for AFL, and CBR's performance in this category is substantially lower. These findings position

our model as a strong alternative for scenarios prioritizing AFL detection, while maintaining robust performance in AF and Non-AF detection.

TABLE 2.7: Performance comparison of our model with rule-based software for patients with PACs.

| Metric | Our Model | ABILE | CBR |
|---|---|---|---|
| Recall (Non-AF) | 0.954 | 0.941 | **0.999** |

Finally, we evaluated the performance using an additional set of 685 Holter recordings from 685 patients. In this analysis, the recordings were categorized as Non-AF because they exclusively featured PACs as the cardiac condition. Table 2.7 presents a comparison of specificity for our proposed model with two software solutions, ABILE and CBR. Our model achieves a specificity of 0.954, indicating a high rate of identification of Non-AF cases, though CBR surpasses this with an impressive specificity of 0.999, reflecting its exceptional ability to correctly classify Non-AF cases. ABILE, with a specificity of 0.941.00, is slightly less effective than our model but still performs well in distinguishing Non-AF cases. Overall, while our model shows robust performance with high specificity, CBR excels in reducing false positives and achieving the highest specificity, whereas ABILE shows slightly lower specificity. While setting the thresholds to match the recall values of Non-AF cases for ABILE and CBR, our DL model achieved recall values of 0.980 and 0.992, respectively, for the 685 PACs patients.

## 2.6 Conclusion

Our DL model's effectiveness and reliability are demonstrated by its performance on a comprehensive and clinically relevant dataset, affirming its practical value. The model leverages a residual temporal attention mechanism to identify critical features in AF and AFL rhythms, leading to strong performance across our dataset and two external test sets. It outperforms three leading state-of-the-art DL models, showcasing its superior capabilities. The model's performance across diverse demographic groups, including various genders and age ranges, reflects its robustness and broad applicability. It consistently achieved high recall rates for both AF and Non-AF cases, and showed notable improvements for specific patient cohorts, such as those with VT and PVC, underscoring its versatility. In comparison with rule-based software, our model demonstrated competitive

results, particularly in detecting AF and AFL. It surpassed ABILE and CBR in recall rates for AF or AFL, highlighting its effectiveness in identifying these conditions. However, CBR and ABILE showed superior performance in minimizing false positives and achieving higher PPV, pointing to areas where our model could be further refined.

# Chapter III

# A Systematic Survey of Data Augmentation of ECG Signals for AI Applications

## 3.1 Introduction

DL models rely heavily on the availability of large, high-quality labeled datasets to achieve optimal performance. Insufficient or imbalanced data can result in degraded model accuracy, unstable training processes, and biased classification outcomes. Consequently, DL models typically require extensive, well-balanced datasets for reliable performance. However, acquiring such datasets presents significant challenges, especially in the domain of ECG, where abnormal cardiac events are rare, and the labeling of waveforms demands expertise from specialized cardiologists. The scarcity of annotated data is further compounded by the fact that only trained physicians can accurately interpret and label ECG recordings. To overcome these obstacles and enhance model robustness, data augmentation (DA) techniques are frequently employed. DA involves applying various transformations to the original data to generate additional, non-redundant training samples, thereby reducing the risk of overfitting and improving the model's ability to generalize effectively across a wider range of decision boundaries.

In the field of image recognition, DA has reached a mature phase, with leading CNN architectures consistently incorporating various DA strategies to improve performance. For example, residual networks (ResNet) utilize techniques such as color augmentation, scaling, and cropping [61], while

DenseNet incorporates methods like mirroring and translation [62], and Inception networks adopt mirroring and cropping techniques [63]. In contrast, these random transformations are not as suitable for ECG data because they can distort the clinical significance of key cardiac features, such as the relative amplitudes and durations of P waves, the QRS complex, T waves, and ST segments. For instance, applying time inversion to ECG signals would result in an unrealistic reversal of wave sequences, which is clinically unacceptable. While specific augmentation techniques like spectral modification might offer some benefits, most DA methods tailored for image data are often counterproductive for ECG signals. Randomly cropping or merging ECG segments, for example, could inadvertently alter a normal sinus rhythm into an abnormal arrhythmia, compromising diagnostic accuracy.

To the best of our knowledge, no comprehensive review focusing specifically on data DA techniques for AI applications in ECG signal analysis had been conducted prior to 2023. Considering the practical significance and the potential impact of these techniques in the development of ECG classification models, a review of DA methods employed in AI-based ECG classification is both timely and essential. In this study, we systematically examined relevant literature, identifying key features of the various methods. As a result of our analysis, we introduce, for the first time, a taxonomy of ECG DA techniques, as depicted in Figure 3.1. This taxonomy categorizes the techniques into two main groups: basic DA techniques and advanced DA techniques.

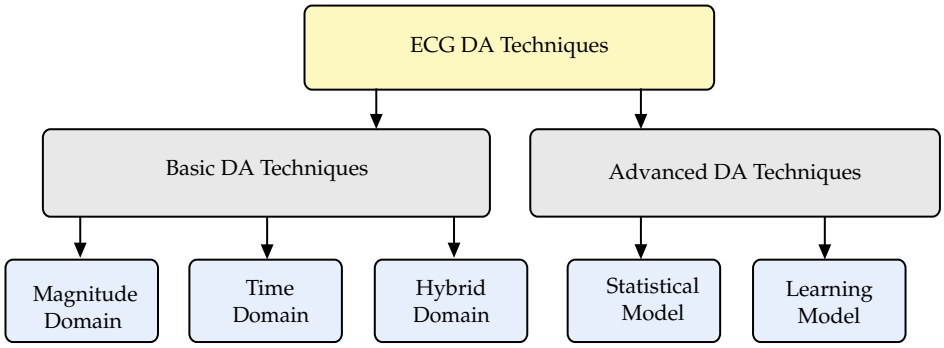



FIGURE 3.1: Taxonomy of ECG DA techniques.

Basic DA techniques involve random or structured transformations of ECG signals in the amplitude, time, or combined time-amplitude domains. In contrast, advanced DA methods utilize more sophisticated approaches,

including statistical and learning-based models, to better capture the underlying data distributions and generate novel patterns. Notable examples of statistical models used for ECG generation include the Gaussian mixture model (GMM) and Markov chain model (MCM), both of which have been employed to produce synthetic ECG samples [64, 65, 66]. More recent advancements in learning-based models, such as the variational autoencoder (VAE) [67] and generative adversarial networks (GANs) [68], offer even greater capability in generating realistic ECG data.

This chapter is structured as follows:

- A comprehensive review of contemporary techniques for ECG signal analysis employing DA methods.
- A detailed taxonomy and classification of ECG DA techniques, including their applications, datasets, and pertinent AI approaches.
- The implementation of DA techniques specifically for minority classes, such as AFL, and an evaluation of the effectiveness of these augmentations in enhancing the performance of DL classifiers.
- An in-depth discussion of existing research gaps and unresolved challenges within the field that warrant further investigation.

## 3.2 Method

### 3.2.1 Literature Search Strategy

A thorough literature search was performed across three major databases: IEEE Xplore, PubMed, and Web of Science. The search strategy encompassed a wide range of topics, such as different signal types (*e.g.*, ECG), AI methodologies, and various DA techniques. To ensure the relevance and validity of the results, the search was limited to peer-reviewed articles, conference papers, book chapters, and magazine articles published in English over the past decade, from January 1, 2013, to January 31, 2023. Details of the specific search terms and queries used in the search are presented in Table 3.1.

TABLE 3.1: List of search queries and the final query.

| Parameter | Search Query |
|---|---|
| Signal type (Q1) | "ECG" OR "electrocardiography" OR "electrocardiogram" OR "EKG" |
| AI technique (Q2) | "DNN" OR "deep learning" OR "neural network" OR "AI" OR "artificial intelligence" OR "machine learning" |
| DA technique (Q3) | "augmentation" OR "synthesis" OR "generation" |
| Specific technique (Q4) | "GAN" OR "generative adversarial network" OR "normalizing flow" OR "stable diffusion" |
| Final query | Q1 AND Q2 AND (Q3 OR Q4) |

### 3.2.2 Study Selection

To ensure a rigorous and systematic approach to article selection, we adhered to the guidelines established by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [69]. Initially, we used reference management software to remove duplicate entries. We then screened the remaining studies by evaluating their titles and abstracts. Following this initial screening, we performed a comprehensive review of the full texts to apply our inclusion and exclusion criteria. For transparency and clarity, a flowchart summarizing our selection process is presented in Figure 3.2. This approach allowed us to effectively refine our search and identify studies that were most relevant to our research objectives.

### 3.2.3 Results of the Research

Following the application of our search query and the established inclusion/exclusion criteria, we initially identified 625 articles. From this pool, 193 duplicates were removed using reference management software and through manual review. The remaining 432 articles were then screened based on their titles and abstracts, resulting in the selection of 350 papers for full-text evaluation, in accordance with the criteria detailed in Table 3.2. Ultimately, 119 papers met our criteria and were retained for further analysis. In the following, the key findings derived from the literature review were described.

In the following sections, we present the major outputs of the literature review.

TABLE 3.2: Inclusion and exclusion criteria for selecting papers

| Inclusion criteria | Exclusion criteria |
|---|---|
| Works published between January 1, 2013, and January 31, 2023 | Review papers and papers not written in English |
| Studies applying DA specifically to ECG signals | Studies not applying DA or lacking a clear description of DA and datasets |
| Papers providing a clear description of DA | Studies not considering ECG signals |
| Inclusion of AI techniques | Papers not reporting performance metrics |

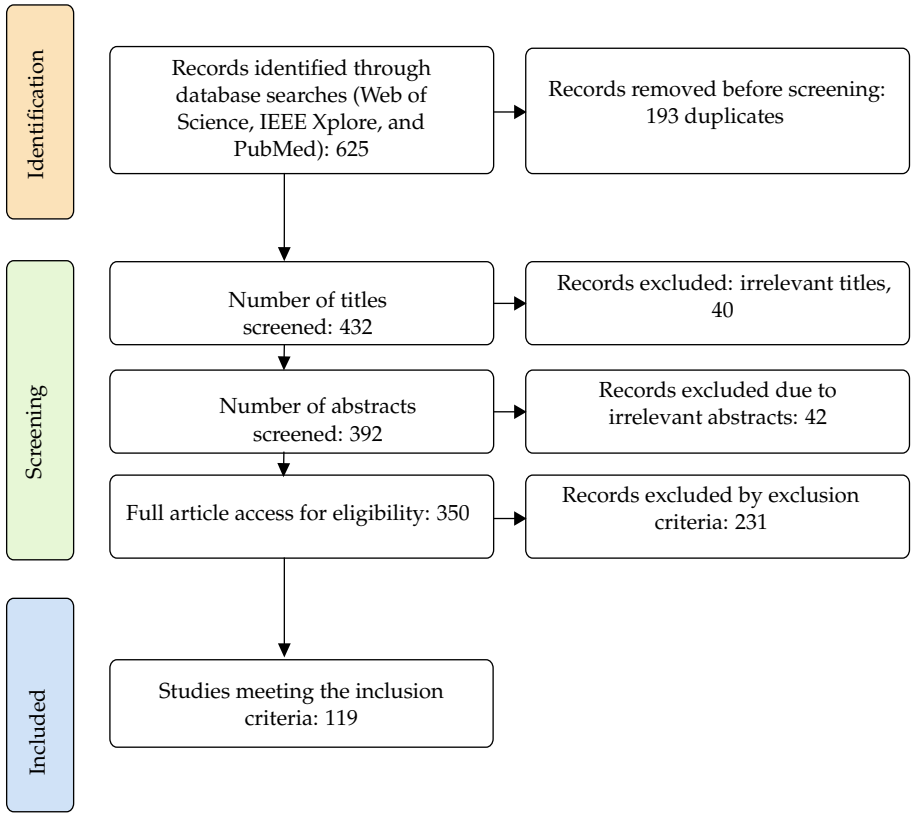



FIGURE 3.2: The search method for identifying relevant studies.

## 3.3 Applications and Datasets of ECG

### 3.3.1 Common Applications of ECG Analysis

AI has significantly enhanced ECG analysis, with automatic ECG interpretation being one of the most prevalent applications [70, 71, 67, 68]. Other critical applications include the localization and annotation of specific rhythms and beats, which are instrumental in detecting conditions such as myocardial infarction (MI) [66] and classifying fetal heart rate series [72]. Recent advances in biometric-based human identification also highlight the potential of ECG data for accurate personal recognition [73, 74]. Furthermore, ECG analysis has proven useful in detecting emotions and stress [75], pain [76], sleep apnea [77, 78], and even identifying COVID-19 infections [79, 80, 81, 82]. It is also employed in assessing signal quality [83, 84], among various other applications. This study encompasses all applications that investigate the use of AI techniques for ECG DA.

### 3.3.2 Datasets

The majority of the studies reviewed utilized a limited number of ECG datasets. The MIT-BIH-AD was employed in 46% of the studies, followed by PhysioNet-2017 at 13%, PTB at 7%, PhysioNet-2020 at 5%, and PhysioNet-2021 at 3%. The INCART, CPSC-2018, and PTB-XL datasets were each used in 2% of the studies. The specific characteristics of these datasets are described in detail below.

- MIT-BIH-AD: This dataset contains 48 ambulatory ECG records, each comprising two leads and spanning 30 minutes. The recordings were collected between 1975 and 1979 [54]. Each record is sampled at 360 Hz with an 11-bit resolution over a 10 mV range and was gathered from 47 individuals tested in the BIH Arrhythmia Laboratory. The dataset includes a variety of cardiac abnormalities, such as AF, atrial bigeminy, AFL, ventricular premature beats, right bundle branch block (RBBB), and left bundle branch block (LBBB).
- PhysioNet-2017: This dataset comprises 8,528 single-lead ECG records obtained from 3,658 individuals [85]. The data are sampled uniformly at 300 Hz and span durations of 9 to 61 seconds. It includes four distinct rhythm categories: normal, AF, noise, and other.

- INCART: The St. Petersburg INCART dataset consists of 75 records extracted from 32 Holter recordings, each spanning 24 hours. These recordings include patients diagnosed with various heart conditions such as coronary artery disease, ischemia, conduction abnormalities, and arrhythmia. In this dataset, 17 patients are male and 15 patients are female, and the mean age of the patients is 58. The data are sampled at 257 Hz, capturing subtle changes in heart function, and each record contains 12 standard leads.

- CPSC-2018: The China Physiological Signal Challenge 2018 dataset features 6,877 recordings of 12-lead ECG data from a diverse patient population [86]. The average age of the patients was about 89. The number of female patients was 4788. The recordings, collected from 11 hospitals, vary in length from 6 to 60 seconds and are sampled at 500 Hz. The dataset includes nine different types of cardiac abnormalities, including AF, LBBB, RBBB, normal, premature atrial contraction, premature ventricular contraction, intrinsic paroxysmal atrioventricular block, ST-segment depression, and ST-segment elevation.

- PTB: The PTB dataset consists of 549 ECG records, which include 15 leads (12 standard leads and 3 Frank leads) from 290 individuals [87]. The data were sampled at 1000 Hz with 16-bit resolution. Each individual has up to five records, allowing for longitudinal health assessments. Of the subjects, 216 were diagnosed with various heart diseases, including MI, cardiomyopathy/heart failure, bundle branch block, dysrhythmia, myocardial hypertrophy, valvular heart disease, and myocarditis. The remaining 52 individuals constitute a healthy control group, while the health status of 22 individuals remains unknown.

- PTB-XL: The PTB-XL dataset is a comprehensive collection of clinical ECGs, including 21,837 records from 18,885 patients [88]. The ECGs are 10 seconds in length and were recorded at two sampling rates, 100 Hz and 500 Hz, with 16-bit resolution. The dataset includes several ECG rhythms and abnormalities, such as normal, MI, conduction disturbances, and hypertrophy.

- PhysioNet-2021: The PhysioNet-2021 dataset features 12-lead ECG recordings from a large cohort of 6,877 patients with various cardiac

abnormalities [89]. The recordings were collected from six different hospital systems across four countries on three continents. The dataset is publicly available as training data, with over 88,000 ECGs provided. Some of the databases previously mentioned (*e.g.*, INCART, PTB, and PTB-XL) are included in PhysioNet-2021.

## 3.4 Basic Data Augmentation Methods

The concept of basic DA techniques for ECG signals is derived from random transformations commonly applied to images and time series, such as scaling, flipping, and noise addition. Broadly, basic DA methods for ECGs can be categorized into three types: time domain, magnitude domain, and hybrid domain transformations.

Time domain transformations modify the ECG signal along the time axis, shifting the data points to different time steps while preserving the sequence structure. In contrast, magnitude domain transformations retain the original time steps but alter the signal values (*e.g.*, in millivolts). Techniques such as scaling, noise addition, and dropping fall under this category, where only the amplitude of the signal is changed. Hybrid methods combine both time and magnitude transformations to manipulate the signal.

In general, basic DA generates a transformed pattern $\mathbf{x}^{'}$ by applying a random transformation function to the original pattern $\mathbf{x}$:

$$\mathbf{x}^{'} \leftarrow f(\mathbf{x}), \tag{3.1}$$

where $\mathbf{x}$ is represented as $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^\intercal$, with $N$ denoting the number of time steps in the original dataset. Each $\mathbf{x}_n$ corresponds to the ECG amplitude at time step $n$, collected across $L$ leads. For instance, a standard clinical ECG is often stored in an $N \times 12$ matrix (or its transpose).

Based on our review of the literature, we identified several works related to basic DA methods, summarized in Table 3.3. Below, we outline the most frequently used DA techniques applied in these studies:

- Noise Addition: Gaussian noise, denoted as $\mathbf{n}$, is applied to the ECG signal $\mathbf{x}$. This noise is characterized by a mean of 0 and a standard deviation of $\sigma$. The resulting transformation is expressed as the addition

of the noise to the original signal, *i.e.,* $\mathbf{x} + \mathbf{n}$ [90, 91, 92, 72, 93, 94, 95, 96, 97].

- Scaling: Each lead of the ECG signal is scaled by a random factor, which is sampled from a normal distribution [90, 91, 98, 92, 94, 74, 78]. Mathematically, this process can be represented by multiplying the ECG signal by a diagonal matrix, where each diagonal element corresponds to the scaling factor for a specific lead.
- Temporal Inversion: The ECG signal undergoes a reversal in the time domain, represented as $\mathbf{x}^{'} = [\mathbf{x}_N, \mathbf{x}_{N-1}, \ldots, \mathbf{x}_1]^\intercal$ [90, 92, 99, 93].
- Spatial Inversion: This method applies a spatial inversion by multiplying the amplitude of the ECG signal by $-1$, resulting in an inverted waveform, denoted as $-\mathbf{x}$ [90, 92, 93].
- Temporal-Spatial Inversion: A combination of temporal and amplitude inversion is applied, where the signal is first vertically inverted and then horizontally reversed, resulting in a temporal-spatial transformation [100, 99, 93].
- Permutation: The ECG signal is divided into multiple segments, which are then shuffled to randomly alter the temporal sequence before recombining [90, 92].
- Dropping: Random segments of the ECG signal are masked with a certain probability [90, 92, 96, 101].
- Cutout: Similar to dropping, cutout randomly zeros out portions of the signal. However, each portion has a fixed length [90, 102, 92, 72].
- Sine Wave Addition: A sine wave with randomly chosen frequency and amplitude is added to the ECG signal [90, 92].
- Square Pulse Addition: A square pulse, with randomly varying frequency and amplitude, is added to the ECG signal [90, 92].
- Time Warping: Segments of the ECG signal are randomly stretched or compressed along the time axis [90, 103].
- Baseline Wandering: Low-frequency sinusoidal signals are generated and added to simulate baseline wandering [104, 103].

- Lead Removal: A random lead is selected, and its values are set to zero [104].
- Lead Order Shuffling: The order of ECG leads, or a subset of them, is randomly rearranged [104, 91].
- High-Pass Filtering: A Butterworth filter with a fixed cutoff frequency (e.g., 0.5 Hz) is applied to remove baseline wander noise [104].
- Low-Pass Filtering: A Butterworth filter (e.g., cutoff frequency 47 Hz) is used to eliminate high-frequency noise [104]. This technique is also known as Gaussian blur when a one-dimensional Gaussian kernel is applied.
- Band-Pass Filtering: A Butterworth filter with low and high cutoff frequencies (e.g., 0.5 Hz and 47 Hz) is used to remove both baseline drift and high-frequency noise [104, 91, 97].
- Sigmoid Compression: This technique applies a sigmoidal activation function to compress the ECG signal [104].
- Powerline Noise Addition: Powerline interference (50 Hz or 60 Hz) is added to the signal to simulate environmental noise [103, 105, 98].
- EMG Noise Addition: Simulated electromyographic (EMG) noise, induced by muscle contractions, is added to the ECG signal [103, 105].
- Baseline Shift: Random direct current offsets are added to simulate baseline shifts caused by changes in electrode-skin impedance [103, 105, 98].
- Peak Alteration: This involves modifying the shape or duration of peaks, such as the QRS complex or T-wave [106, 74].
- Mixup: New signals are generated by linearly interpolating between two real signals with varying weights [107].

## 3.5 Advanced Data Augmentation Techniques

Conventional DA techniques often modify the characteristics of ECG signals in ways that introduce "noise" rather than generating meaningful new samples, which can negatively impact classification performance. For instance,

TABLE 3.3: Summary of basic DA methods for ECG classification using AI techniques.

| Type | Lead | Input | Classifier | Improvem. after DA | Dataset | Refs. |
|---|---|---|---|---|---|---|
| CA | 12 | ECG | CNN | 2.24% | Physionet-2020 | [108] |
| CA | 12 | ECG | CNN-LSTM | 3% | Physionet-2020 | [91] |
| CA | 12 | ECG | ResNet | −0.063–2.54% | CPSC-2018 | [104] |
| CA | 12 | ECG | CNN | – | Physionet-2020 | [102] |
| CA | 12 | ECG | CNN | – | Physionet-2020 | [109] |
| CA | 12 | ECG | ResNet | 1.4–3.5% | ICBEB and PTB-XL | [98] |
| CA | 1 | ECG | CNN | – | MIT-BIH AD | [110] |
| CA | 1 | Spectral | Residual Attention | 0.8% | MIT-BIH AD | [111] |
| CA | 12 | ECG | CNN | 7.73% | Physionet-2021 | [92] |
| CA | 1 | ECG | CNN | – | MIT-BIH-AD | [112] |
| CA | 12 | ECG | ResNet | 40% | INCART | [106] |
| CA | 2 | ECG | CNN | 2.3% | Physionet-2017 | [113] |
| CA | 1 | Spectral | CNN | – | MIT-BIH-AD | [114] |
| CA | 1 | ECG | CNN | 0.028% | MIT-BIH-AD | [115] |
| CA | 1 | Spectral | CNN | – | MIT-BIH-AD | [116] |
| CA | 12 | ECG | CNN | – | Physionet-2020 | [117] |
| CA | 8 | ECG | CNN | – | Private | [95] |
| CA | 12 | ECG | CNN | 1% | Physionet-2020 | [118] |
| CA | 12 | Spectral | CNN | 4.64% | PTB | [119] |
| CA | 1 | Spectral | CNN | – | Physionet-2017 | [120] |
| CA | 1 | ECG | CNN | 5% | MIT-BIH-AD | [99] |
| CA | 12 | ECG | CNN | – | Physionet-2021 | [97] |
| CA | 1 | ECG | BeatGAN | 0.28% | MIT-BIH-AD | [121] |
| CA | 1 | ECG | ResNet-LSTM | – | MIT-BIH-AD, AFDB and Physionet-2017 | [122] |
| CA | 1 | Spectral | Residual-Attention | – | MIT-BIH-AD and Supraventricular Arrhythmia | [123] |
| CA | 1 | Spectral | CNN | – | MIT-BIH-AD | [124] |
| CA | 1 | ECG | LSTM | 42% | Physionet-2017 | [125] |
| CA | 2 | ECG | CNN-RNN | – | Private | [126] |
| CA | 1 | ECG | CNN-LSTM | 3% | MIT-BIH-AD | [127] |
| CA | 1 | ECG | CNN-RNN | 1.91% | Physionet-2017 | [107] |
| CA | – | Spectral | CNN | – | MIT-BIH-AD and PTB | [128] |
| CA | 1 | ECG | CNN | – | Physionet-2017 | [96] |
| CA | 1 | ECG | CNN | – | Physionet-2017 | [129] |
| CA | 1 | ECG | CNN | – | Physionet-2017 | [101] |
| CA | 1 | ECG | ResNet-RNN | – | Physionet-2017 | [130] |
| CA | 12 | ECG | CNN | – | Physionet-2021 | [131] |
| CA | 1 | ECG | CNN | 0.62–5.61% | MIT-BIH-AD | [132] |
| CA | 1 | Spectral | Transformer | – | MIT-BIH-AD | [133] |
| Biometric | 1 | ECG | CNN | – | CYBHi and UofTDB | [134] |
| Biometric | 1 | ECG | CNN | 0.19% | PTB and LivDet2015 [135] | [136] |
| Biometric | 1 | ECG | CNN | 12% | Physionet-2018 | [74] |
| Frailty Identification | 1 | ECG | LSTM | 3.2% | Private | [94] |
| Sleep apnea | 1 | ECG | CNN | – | Private | [78] |
| Peak detection | 2 | ECG | CNN | 2.5% | MIT-BIH-NST | [137] |
| QA | 1 | ECG | CNN | 2% | Physionet-2017 | [138] |
| QA | 12 | Spectral | CNN | 2.91% | PhysioNet-2011 | [83] |
| QA | 2 | ECG | U-Net | – | QT [139] | [84] |

in [104], it was reported that certain basic DA techniques, such as horizontal and vertical flipping, adversely affected the accuracy of their classifier. To address the limitations of traditional DA methods, more sophisticated techniques have been developed as viable alternatives. Through a literature search, we identified several studies that explore advanced DA approaches, the key findings of which are summarized in Table 3.5. These techniques can be broadly classified into two categories: statistical generative models and learning-based models. A detailed description of each approach is provided in the following subsections.

### 3.5.1 Statistical Generative Models

Advanced ECG DA methods based on statistical generative models aim to model the underlying dynamics of ECG signals through statistical techniques. For example, Hatamian *et al.* [64] proposed a GMM to address class imbalance in AF detection. Their results showed that the GMM outperformed traditional oversampling methods for minority class augmentation. Similarly, Silva *et al.* [65] developed a cardiorespiratory signal synthesizer using conditional sampling from a multimodal stochastic system based on Gaussian copulas, integrated with a Monte Carlo (MC) approach. Zhu *et al.* [66] introduced an innovative DA technique that leveraged both probabilistic distribution and geometric information. Their method applied variations to the data distribution along the geodesic path in Wasserstein space, a mathematical framework that measures the distance between probability distributions. By analyzing cardiovascular features within ECG signals, they were able to account for geometric properties when generating augmented samples, which were subsequently fed into a multi-feature transformer model alongside real data. This approach yielded substantial performance gains, improving the AUC-ROC on the PTB-XL dataset by 6-17% compared to models trained on unaugmented data.

### 3.5.2 Learning-Based Models

In the realm of AI, DL-based generative models have emerged as powerful tools for producing diverse synthetic data samples that closely resemble real-world data. These models have garnered significant attention due to their ability to generate high-quality data, making them suitable for a variety of applications. While numerous generative models exist, only a

subset has been applied to ECG DA. In the following subsection, we focus on specific learning-based approaches that have been employed for ECG DA, with the goal of addressing the challenge of limited labeled data in AI-driven ECG applications.

### Embedding Space

ECG DA techniques should not only be capable of generating diverse samples but also adept at mimicking the characteristics of real ECG signals. It is hypothesized that applying transformations to encoded representations, rather than raw inputs, could result in more convincing synthetic data due to the unfolding of manifold structures in the feature space. For instance, Zhang *et al.* [100] employed fundamental DA techniques for representational learning within the embedding space. Their learning model consists of two primary modules: an encoder and a classifier. The encoder creates representations using a temporal-spatial reverse detection method, while the classifier executes the temporal-spatial reverse detection task during training. After training is complete, the encoder is transferred to the second stage, where the classifier applies the learned representations to various downstream tasks.

TABLE 3.5: Summary of advanced DA methods for ECG classification using AI techniques.

| Types | Lead | DA Methods | Input | Classifier | Improvem. after DA | Dataset | Refs. |
|---|---|---|---|---|---|---|---|
| CA | 1 | Style-transfer | ECG | CRN | 3% | Physionet-2017 & Private | [140] |
| CA | 2 | CGAN | ECG | CNN | 1.3–2.6% | MIT-BIH-AD & Physionet-2017 | [141] |
| CA | 12 | VAE | Spectral | CNN | 0-6% | Private | [67] |
| CA | 1 | GAN | ECG | CNN | 1% | MIT-BIH-AD | [142] |
| CA | 1 | GAN | ECG | CNN | 1.3% | MIT-BIH-AD | [68] |
| CA | 1 | Embed-ding space | ECG | CNN | – | Physionet-2017 | [100] |
| CA | 1 | GAN | Spectral | CBAM-ResNet | – | MIT-BIH-AD | [143] |
| CA | 12 | Embed-ding space | ECG | Self-supervised | – | Physionet-2021 | [144] |
| CA | 1 | GAN | Spectral | CNN | 3% | Physionet-2017 | [64] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [70] |
| CA | 1 | GAN | ECG | CNN | 5-37% | MIT-BIH-AD | [71] |

*Continued on next page*

Table 3.5 – *Continued from previous page*

| Types | Lead | DA Methods | Input | Classifier | Improvem. after DA | Dataset | Refs. |
|---|---|---|---|---|---|---|---|
| CA | 1 | GAN | ECG-PPG | CNN | – | BIDMC | [145] |
| CA | 1 | MC | ECG | CNN | – | MIT-BIH-AD | [65] |
| CA | 1 | Embed-ding space | ECG | CNN | 5.8% | ICENTIA11K [146] | [147] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [148] |
| CA | 1 | VAE | ECG | CNN-LSTM | 2% | MIT-BIH-AD | [149] |
| CA | 1 & 12 | BiLSTM-CNN & TimeGAN | ECG | CNN | – | MIT-BIH-AD & PTB | [150] |
| CA | 12 | GAN | ECG | ResNet | 5% | CPSC-2018 | [151] |
| CA | 1 | GAN | ECG | CRNN | 14% | Physionet-2017 | [152] |
| CA | 1 | GAN | ECG | Bi-LSTM | 1.9% | MIT-BIH-AD | [153] |
| CA | 1 | GAN | ECG | RF | 11% | MIT-BIH-AD | [154] |
| CA | 1 | GAN | ECG | LSTM | – | MIT-BIH-AD & MIT-BIH NSR | [155] |
| CA | 1 | GAN | ECG | CNN | 1.45% | MIT-BIH-AD | [156] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [157] |
| CA | 1 | GAN | ECG | CNN-LSTM | 2.65% | MIT-BIH-AD | [158] |
| CA | 1 & 12 | GAN | ECG | CNN | – | MIT-BIH-AD & PTB | [159] |
| CA | 1 | GAN | ECG | CNN | 0.24% | MIT-BIH-AD | [160] |
| CA | 2 | GAN | ECG | SVM | 32% | MIT-BIH-AD | [161] |
| CA | 1 | GAN | ECG | Bi-LSTM | 2–51% | MIT-BIH-AD | [162] |
| CA | 1 | VAE & GAN | ECG | CNN | 5% | MIT-BIH-AD | [163] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [164] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [165] |
| CA | 1 | GAN | ECG | LSTM | – | MIT-BIH-AD | [166] |
| CA | 1 | GAN | ECG | ResNet-BiLSTM-attention | – | MIT-BIH-AD | [167] |
| CA | 1 | AE | ECG | CNN | – | Physionet-2017 | [168] |
| CA | 1 | GAN | Spectral | CNN | – | MIT-BIH-AD | [169] |
| CA | 1 | GAN | ECG | Multi-head Attention | 5–10% | MIT-BIH-AD | [170] |
| CA | 1 | GAN | ECG | CNN | 20.5% | MIT-BIH-AD | [171] |
| CA | 1 | GAN | ECG | CNN | – | MIT-BIH-AD | [172] |
| CA | 1 | GAN | ECG | CNN | 32% | MIT-BIH-AD | [161] |
| CA | 1 | BiRNN | ECG | Ensemble Bagged Trees | – | MIT-BIH-AD | [173] |
| CA | 1 | GAN | ECG | CNN | 4.8–8.1% | Private | [174] |
| CA | 1 | GAN | ECG | LSTM | 4% | MIT-BIH-AD | [175] |
| CA | 1 | GMM | ECG | ResNet | 6.7% | MIT-BIH-AD | [176] |

*Continued on next page*

Table 3.5 – *Continued from previous page*

| Types | Lead | DA Methods | Input | Classifier | Improvem. after DA | Dataset | Refs. |
|---|---|---|---|---|---|---|---|
| CA | 12 | Embedding space | Spectral | Self-supervised | – | Private | [177] |
| CA | 1 | GAN | ECG | CNN | – | AHADB, VFDB, & CUDB | [178] |
| MI | 1 | Encoder-decoder | ECG | CNN | – | PTB | [179] |
| MI | 12 | Wasserstein Geodesic Perturbation | ECG | MFT | 6–17% | PTB-XL | [66] |
| MI | 1 | GAN | ECG | CNN | 4–6% | PTB | [180] |
| Fetal | 1 | GAN | ECG | CNN | 12% | CTU-UHB | [181] |
| Emotion | 1 | GAN | ECG | LSTM | 17% | CASE | [182] |
| Biometric | 1 | GAN | ECG | CNN | – | ECG-ID | [183] |
| Sleep-Apnea | 1 | GAN | ECG | CNN-LSTM | 1.78 | Apnea-ECG & MIT-BIH AD | [77] |
| Emotion | – | GAN | ECG | CNN | 5.64% | Private | [184] |
| MI | 12 | GAN | ECG | SVM | 0.75% | PTB | [185] |
| Emotion | 1 | GAN | ECG | SVM | – | DECAF | [75] |
| Pain intensity | 1 | DDCAE | ECG | NN | – | BioVid Heat Pain | [76] |

### 3.5.3 Deep Generative Models

Deep generative models (DGMs) have recently shown remarkable promise in generating realistic high-dimensional data, such as images, time series, and sequential data. In the context of ECG data, DGMs can be broadly categorized into two types: encoder-decoder networks and GANs. The following sections elaborate on these two DGMs.

**Variational Autoencoder:** The VAE is a powerful DL architecture that has significantly advanced unsupervised learning. At its core, the VAE consists of three essential components: an encoder, a decoder, and a loss function. The encoder and decoder are distinct neural networks, with the

encoder responsible for mapping high-dimensional inputs into a lower-dimensional latent space and the decoder tasked with reconstructing those inputs back into high-dimensional outputs. The loss function used in VAEs includes a negative log-likelihood term, augmented with a regularization term to ensure that the generated outputs closely resemble the original data. By sampling vectors from the latent space and decoding them, VAEs can generate novel patterns, making them highly effective for data synthesis and augmentation. In [67], the authors utilized a vector-quantized VAE (VQ-VAE) to augment training samples of spectral ECG images, reporting a 6% performance improvement over unaugmented data. Al Nazi *et al.* [149] applied a VAE model to enhance the diversity of ECG data, while Thiam *et al.* [76] employed deep denoising convolutional autoencoders (DDCAEs). Their method optimizes both the joint representation of input channels, generated by a multimodal DDCAE, and an additional neural network that is trained concurrently to perform classification.

**Generative Adversarial Networks:** GANs, introduced by Ian Goodfellow and his colleagues in 2014 [186], have become a standard approach for generating synthetic samples for training datasets. GANs rely on adversarial training to simultaneously optimize two neural networks: a generator and a discriminator. The generator creates samples that resemble those from the original data distribution, typically by sampling from a multivariate normal distribution and feeding it as input. The discriminator compares the generator's outputs with real samples, assigning a probability between 0 and 1 to determine whether a sample is synthetic or real.

In the context of ECG DA, several studies [142, 68, 70, 71, 148, 153, 155, 156, 157, 158, 159, 160, 162, 163, 164, 165, 166, 167, 170, 171, 172, 175, 77] utilized GANs to augment the minority class samples in the MIT-BIH arrhythmia dataset (MIT-BIH-AD). These augmented samples were subsequently used in DL models for ECG beat classification, yielding significant performance improvements, with gains ranging from 0.24% to 32% compared to unaugmented samples. Additionally, other studies such as [154] and [161] also implemented GANs for ECG DA but employed machine learning classifiers like RF and support SVM, respectively.

Zhou *et al.* [141] introduced a Conditional Generative Adversarial Network (CGAN) to generate diverse ECG signals aimed at enhancing the training efficiency of DL models. Their approach resulted in a performance

improvement of 1.3–2.6% across two datasets, namely MIT-BIH AD and PhysioNet-2017. In contrast to using raw ECG signals as input for GANs, some researchers have transformed ECG data into spectral images. For instance, Hatamian *et al.* [64] converted ECG signals into images using a logarithmic spectrogram.

Xiong *et al.* [140] developed an ECG generator comprising three main components: clinical ECG recordings, a mathematical model based on ordinary differential equations, and a 37-layer convolutional recurrent network (CRN) for style transfer. Initially, the mathematical model was employed to generate ECG waveforms that simulate an idealized heart rate and RR interval pacing, using parameters such as the mean and standard deviation of the heart rate. These synthetic ECG waveforms were then input into the neural network for style transfer. The authors reported that their approach improved AF detection accuracy by 3% when DA techniques were applied.

Similarly, Fangyu *et al.* [150] introduced an innovative method to enhance the accuracy of abnormal ECG signal detection. To mitigate the impact of imbalanced data on model performance, they implemented two DA techniques—BiLSTM-CNN and TimeGAN—to enrich the semantic representation of different ECG features. Additionally, they proposed a contrastive learning framework to ensure consistency in data representation across two channels. By maximizing the similarity between data representations and minimizing contrastive loss, they achieved more comprehensive data embeddings and correlations, resulting in a 3% performance improvement over a model that did not utilize contrastive learning.

Some researchers have exclusively utilized GANs for ECG synthesis. ECG synthesis holds the potential to advance our understanding of the underlying mechanisms of various cardiac conditions and foster the development of more precise diagnostic models. However, it is essential to rigorously validate the accuracy and reliability of models trained on synthetic ECG data before their application in clinical settings. Based on our review criteria, we identified several papers that employed generative methods solely for ECG synthesis; a summary of these methods is presented in Table 3.6. In these studies, various metrics were used to evaluate the performance of GAN models for ECG synthesis, with the choice of metrics depending on the specific objectives and characteristics of the generated ECG signals. Commonly used evaluation metrics include:

- Mean Squared Error (MSE) and Root MSE (RMSE): These metrics assess the average squared difference between generated and real ECG signals. Lower values indicate better performance.
- Signal-to-Noise Ratio (SNR): This metric evaluates the ratio of signal power to noise power in generated ECG signals. A higher SNR suggests better quality.
- Fréchet Inception Distance (FID): FID measures the distance between the distributions of generated and real ECG signals. Lower FID values indicate greater similarity between the two distributions.
- Maximum Mean Discrepancy (MMD): MMD evaluates the distance between two distributions by comparing the means of their feature representations in a reproducing Kernel Hilbert Space. A smaller MMD indicates that the two distributions are more similar.
- Dice Coefficient (DC): The Dice coefficient measures the similarity or overlap between two sets or binary masks. Values range from 0 to 1, with 0 indicating no overlap and 1 indicating a perfect match.
- Percent Mean Square Difference (PMSD): PMSD is calculated as the square of the difference between generated and real ECG values, divided by the average of the values, expressed as a percentage. Lower PMSD values reflect better performance.
- Kernel Maximum Mean Discrepancy (KMMD): KMMD, an extension of MMD, maps data to a high-dimensional space using a kernel function to evaluate similarity between data points. Lower KMMD values suggest greater similarity between generated and real data.

## 3.6 Implementation of DA Techniques to Enhance Atrial Flutter Performance

In our DL model, we observed relatively lower recall values for AFL cardiac conditions, as reported in Table 2.4. We hypothesized that this performance discrepancy could be attributed to the class imbalance problem. To address this issue, we implemented DA techniques aimed at improving the model's performance on AFL cases. Specifically, we employed two basic DA methods—lead order changing and Mixup (see Section 5.3.3)—along

TABLE 3.6: Summary of generative methods for ECG synthesis using AI techniques.

| Lead | Input | Method | Metric | Dataset | Refs. |
|---|---|---|---|---|---|
| 1 | ECG | GAN | MMD ($3.83 \times 10^{-3}$) | LUDB [187] | [188] |
| 1 | ECG | GAN | KMMD (5.53) | MIT-BIH AD | [189] |
| 1 | ECG | GAN | MSE (0.017–0.099) | PTB-XL | [190] |
| 1 | ECG | GAN | SNR (40.85 dB) | MIT-BIH AD | [191] |
| 1 | ECG | GAN | RMSE (0.126) | MIT-BIH AD | [192] |
| 1 | ECG | AE | MSE (0.2) | MIT-BIH AD | [193] |
| 1 | ECG | GAN | FID (4.77–17.19) | MIT-BIH AD | [194] |
| 2 | ECG | GAN | PMSD (7.21%) | – | [195] |
| 1 | ECG | BiLSTM-CNN GAN | RMSE (0.276) | – | [196] |
| 12 | ECG | U-Net generator | DC (0.868) | Private and INCART | [197] |
| 1 | ECG | GAN | RMSE (0.015–0.028) | MIT-BIH AD | [198] |
| 12 | ECG | Genetic Algorithm-NN | RMSE (44.9–90) $\mu$V | PTB | [199] |
| 12 | ECG | CycleGAN | MSE ([0.5–31]$\times 10^{-3}$) | Private | [200] |

with three advanced techniques to effectively double the AFL sample size to $2 \times 136,000$. The synthetic samples were then integrated with the real minority class samples to train our DL model. For more details on the model architecture and training process, refer to Section 2.4. The advanced DA methods used are described below:

- WaveGAN: This generative model takes a one-dimensional input vector of 2000 data points, sampled from a uniform distribution, and processes it through five deconvolution blocks to generate a signal with 2000 time steps across 2 ECG leads. Each deconvolution block consists of four layers: upsampling, padding, one-dimensional convolution, and ReLU activation. Both the generator and discriminator in this model are configured with 50 units. For further details, refer to [201].

- Pulse2Pulse GAN: Pulse2Pulse is a neural network architecture inspired by U-Net, designed specifically for generating ECG signals using one-dimensional convolutional layers. The generator receives an input of 2000 time steps across 2 leads, sampled from a uniform distribution. This input is processed through a series of five downsampling and five upsampling blocks. The upsampling process is similar to that in WaveGAN, while the downsampling blocks use one-dimensional convolutions followed by Leaky ReLU activation functions. Additional details can be found in [201].

- Diffusion Model: Diffusion models consist of two main processes:

the forward process and the backward process. The forward process progressively adds noise in a Markovian manner, corrupting the data incrementally. The backward process involves the model removing the added noise to reconstruct the original data distribution. Our implementation of this model is adapted to our specific use case. More information is available in [202].

### Performance for DL model after DA

Table 3.7 presents a comparative analysis of the performance of DA techniques applied to our DL model (see details in Section 2.2) across three diagnostic categories: Non-AF, AF, and AFL. The model without DA serves as the baseline, achieving Recall values of 0.951 for Non-AF, 0.931 for AF, and 0.812 for AFL, with corresponding AUC values of 0.981, 0.973, and 0.942.

Among the DA methods, Mixup demonstrates the most notable improvement in AFL detection, achieving the highest AUC for AFL (0.982). However, this gain in AFL performance is accompanied by a slight reduction in AF Recall, which decreases to 0.921. This trade-off reflects Mixup's focus on enhancing AFL detection, albeit at the cost of a marginal decline in AF performance.

In contrast, methods such as WaveGAN, Pulse2Pulse, and the Diffusion technique exhibit more balanced performance across all diagnostic categories but do not surpass Mixup in AFL detection. For instance, WaveGAN attains a Recall of 0.801 and an AUC of 0.952 for AFL, which, although respectable, falls short of Mixup's performance. However, WaveGAN maintains a higher Recall for AF (0.945), indicating that while these methods provide more consistent performance, they do not achieve the same level of efficacy in AFL detection as Mixup.

## 3.7 Discussion

Small-scale and imbalanced datasets present significant challenges for the application of AI-based models in cardiology. DA is a widely adopted solution to these issues, demonstrating success across various fields. However, applying DA to ECG signals poses unique challenges. ECG signals contain fine-grained information, such as the relative amplitudes of ECG

TABLE 3.7: Comparative performance of DA techniques applied to generate synthetic samples for the AFL cardiac condition and their impact on the performance of our DL model.

| Method | Recall | | | AUC | | |
|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Baseline | 0.951 | 0.931 | 0.812 | 0.981 | 0.973 | 0.942 |
| Lead-Shuffling | 0.921 | 0.950 | 0.791 | 0.981 | 0.972 | 0.941 |
| Mixup | 0.960 | 0.921 | 0.830 | 0.989 | 0.973 | 0.982 |
| WaveGAN | 0.963 | 0.945 | 0.801 | 0.988 | 0.970 | 0.952 |
| Pulse2Pulse | 0.954 | 0.952 | 0.798 | 0.980 | 0.972 | 0.963 |
| Diffusion | 0.951 | 0.961 | 0.792 | 0.982 | 0.973 | 0.960 |

waveforms (accurate to a few microvolts) and temporal relationships between data points (on the scale of milliseconds). These details are critical for AI classifiers. Synthetic ECG signals can be beneficial if they accurately capture this fine-grained information; otherwise, DA may degrade model performance. Another challenge in DA is universality—DA techniques are highly dependent on factors such as input type, input shape, the number of leads, and the hyperparameters of the AI or DL model. Furthermore, the same DA technique may have varying effectiveness depending on the type of ECG rhythm, enhancing performance in one scenario while deteriorating it in another.

Several DA methods have been developed to generate synthetic ECGs and improve AI model performance. These methods can be broadly categorized into basic and advanced techniques. Basic DA methods are typically easy to apply and computationally efficient, often producing promising results. However, certain techniques, such as time inversion, spatial inversion, permutation, and lead shuffling, should be avoided. Careful design is required for other basic methods, as inappropriate augmentation can produce non-physiological ECGs or signals that belong to different diagnostic categories. For example, scaling the QRS complex can simulate cardiac hypertrophy, while artificially prolonging the PR interval may mimic atrioventricular block. Similarly, in the context of MI, preserving the correct lead order is essential for accurate infarct localization.

In the field of advanced DA, researchers have extensively explored generative models such as GANs for synthesizing ECG data. Most research focuses

on generating ECG beats, particularly from datasets like MIT-BIH, rather than on rhythm-level augmentation. This emphasis limits the generalizability of DA techniques for applications that require rhythm generation. The lack of research on rhythm generation poses a challenge for determining the optimal DA techniques in applications where broader rhythmic patterns are of primary importance.

Moreover, the effectiveness of DA varies significantly depending on the technique, dataset, preprocessing, and specific application. It is often impossible to predict the best augmentation method for a given dataset without empirical testing. Advanced DA techniques, particularly generative models, offer the potential to generate high-quality synthetic data that closely match the statistical properties of real-world data. This makes them promising for enhancing the accuracy and robustness of AI-based models.

In Sections 3.4 and 3.5, we discussed methods for generating ECG signals. Several studies, however, implement DL models using spectral images as inputs (see Tables 3.3 and 3.5). In these approaches, ECG signals are transformed into spectral images, which are then used for classification instead of raw ECG signals [67, 114, 64, 79, 143, 111]. DA is applied directly to the spectral images, bypassing the ECG signal itself. While this method is motivated by the observation that ECG features often correlate with changes in frequency band energy, it neglects the role of the phase of sinusoidal components, potentially reducing model performance. Although spectral methods have shown promising results, interpretability and explainability are compromised, as changes in spectral images are not directly linked to ECG features recognized by cardiologists.

Our analysis also reveals several unresolved challenges. First, there is no consensus on the optimal ratio of real to synthetic ECG data for improving model performance and addressing overfitting. Some studies suggest that increasing the number of synthetic samples does not always lead to improved performance [140, 98, 185]. The ideal ratio depends on the specific application and must be empirically determined for each dataset. Further research is needed to explore the most effective balance between real and synthetic ECGs in AI models for different applications.

Second, quantifying the quality of synthetic ECG signals remains a significant challenge. There is no universally accepted method for measuring

the similarity between synthetic and real ECGs. Visual inspection is often employed to determine if a synthetic signal appears realistic, but this method requires domain expertise and lacks scalability. Most studies focus on evaluating the performance improvements of classifiers, rather than directly assessing the quality of synthetic ECGs. One potential solution is to extract key ECG features (*e.g.*, heart rate, QRS complex amplitude, peak-to-peak differences) from both real and synthetic signals, then apply distance metrics such as Wasserstein distance, Kullback-Leibler divergence, or Kolmogorov-Smirnov tests to quantify the similarity of distributions.

While DA often enhances model performance by enabling it to learn more robust features, it is not a universal solution that guarantees improvement across all tasks or diagnostic categories. Several factors must be considered when assessing the effectiveness of DA techniques, including the desired performance improvement, the size of the available training data, and the specific diagnostic objectives. As illustrated in Table 3.7, the application of DA techniques to our DL model (detailed in Section 2.2) produced mixed outcomes across three diagnostic categories: Non-AF, AF, and AFL. Among these techniques, Mixup significantly improved the model's ability to detect AFL, raising the AUC for AFL detection to 0.982. This result demonstrates DA's potential to address specific weaknesses in model performance. However, the improvement in AFL detection came at the expense of a slight decline in AF recall, which dropped from 0.931 to 0.921. This trade-off underscores an important point: enhancing performance in one diagnostic category may lead to a reduction in another. Thus, it is crucial to carefully evaluate how much performance improvement is needed and whether the trade-offs are acceptable. The size and balance of the training data also play a critical role in determining the impact of DA. For small or imbalanced datasets, DA can significantly improve model generalization by generating additional, diverse training samples. However, when working with large and varied datasets, the benefits of DA may diminish. In some cases, overly aggressive augmentation can even degrade performance by introducing noise or unrealistic variations into the training process.

Overall, further research is required to address these challenges. One promising direction is the combination of various DA techniques to expand datasets, such as applying adversarial learning for secondary augmentation on synthetic ECGs generated from basic DA techniques. This

approach could increase the variability of synthetic data. Moreover, integrating meta-learning with DA might provide insights into how DA affects the performance of AI models for ECG classification. While DA through adversarial learning is gaining popularity, improving the quality of synthetic ECGs remains essential. Ongoing development in this area is crucial for advancing the utility of DA in ECG analysis, particularly in terms of enhancing sample quality and evaluating performance across diverse datasets.

## 3.8 Conclusion

Collecting large-scale ECG datasets poses significant challenges due to limitations in patient availability, access to expert cardiologists, lengthy recording times, and operational complexities. DA is a promising strategy for addressing these issues, particularly in augmenting small-scale datasets and balancing minority classes to reduce overfitting and improve the performance of AI models. This study reviews the current research on DA techniques for ECG interpretation using AI. Overall, our findings suggest that DA generally enhances the performance of automatic ECG analysis. However, its success depends on application, ensuring that improvements in one diagnostic area do not come at the expense of another. In conclusion, this study provides practical insights from the literature, offering guidance for ECG research and assisting in modeling the inter-patient variability in ECG interpretation. It highlights the potential of DA to improve AI-driven ECG analysis while emphasizing the importance of tailoring DA methods to specific goals and challenges.

# Chapter IV

# Uncertainty Quantification of Deep Learning Model for Atrial Fibrillation Detection from Holter Recordings

## 4.1 Introduction

In recent years, DL models have shown promising success for AF detection [203, 27, 28]. However, concerns about the reliability and acceptance of these models in clinical practice persist. Variability in ECG signal characteristics—due to factors such as artifacts, noise, and the diversity of ECG patterns—further exacerbates these concerns. Additionally, DL models often demonstrate inconsistent performance on new, unseen data, eroding the trust of medical professionals and limiting their integration into clinical workflows.

To mitigate these issues, it is essential to provide clinicians with supplementary information, such as the confidence levels associated with model predictions, rather than simply presenting the outputs themselves. The increasing volume of ECG recordings requiring interpretation has further highlighted the need for efficient review processes, as manual examination of automated ECG analyses has become increasingly time-consuming and resource-intensive. By incorporating uncertainty information into model outputs, clinicians can focus their attention on cases where the model exhibits uncertainty, thereby optimizing the allocation of time and clinical

resources.

A potential solution to enhance the reliability of DL models is uncertainty quantification (UQ). UQ reflects the model's confidence in its predictions, such as distinguishing between AF and Non-AF rhythms. Two primary types of uncertainty arise in DL classifiers: data (aleatoric) uncertainty and model (epistemic) uncertainty [204]. Data uncertainty is caused by factors such as sensor noise, data collection errors, and labeling ambiguity, while model uncertainty stems from a lack of knowledge about the model's parameters—especially when the model is trained on limited or insufficient data, resulting in incomplete representations of underlying data patterns. Addressing these uncertainties is essential for developing robust and reliable DL models for AF detection. Clinically, uncertainty estimates can guide or automate label corrections, reject unreliable model outputs, and aid in detecting classification failures at the patient level.

Several UQ methods have been developed and applied in the context of AF detection [205, 206, 207, 208, 209, 210, 211]. One approach is through variational inference (VI), where the weights of a DL model are treated as random variables and their posterior distribution is approximated [205, 207, 212]. However, VI-based methods can present challenges related to scalability, both in terms of model architecture and data size. Another commonly used UQ approach involves ensemble methods, where multiple models are trained independently, and uncertainty is captured through averaging their predictions [207, 213].

The majority of prior research has focused on deep ensemble (DE) and Monte Carlo dropout (MCD) methods for UQ. However, these methods face limitations in terms of scalability, particularly when applied to large datasets or complex architectures [214]. While innovative UQ techniques have been proposed across various domains, their application in AF detection remains underexplored. As a result, there is a critical need for a comprehensive evaluation of UQ methods tailored specifically for AF detection across diverse datasets.

This study addresses this gap by conducting a comparative analysis of various UQ techniques on three public datasets, utilizing both internal and external validation sets. The key contributions of this study are as follows:

- We developed a benchmark for evaluating different UQ methods specifically designed for AF detection using Holter ECG recordings.
- We assessed these UQ techniques under real-world conditions by introducing random Gaussian noise into the data, thereby simulating real-world noise and variability.
- We analyzed the performance of the UQ methods across varying rejection thresholds, offering critical insights into their robustness and reliability in clinical settings.

In this chapter, we conducted an additional study focused on the evidential (EDL) model [215]. The EDL model utilizes a variational Dirichlet distribution to capture and quantify the uncertainty associated with predictions. By parameterizing a Dirichlet distribution over categorical output probabilities, the EDL model offers a principled measure of epistemic uncertainty, providing varying levels of confidence in its predictions.

EDL presents several distinct advantages over traditional UQ techniques such as MCD, DE, and VI. One of its main advantages is the ability to represent uncertainty directly through belief functions, facilitating a clear interpretation of confidence levels in predictions. Moreover, EDL supports soft classification, allowing for degrees of belief across multiple classes rather than making binary decisions, thereby enhancing decision-making in ambiguous scenarios.

The model effectively captures both aleatoric and epistemic uncertainty simultaneously, a feature often lacking in methods like MCD. Additionally, EDL can be trained end-to-end within deep learning architectures, streamlining integration and reducing computational costs by requiring only a single model, in contrast to the multiple models needed for DEs. Its predictions are well-calibrated, ensuring that confidence levels closely align with actual outcome probabilities.

Overall, EDL's flexibility, fewer hyperparameters, and enhanced interpretability position it as a powerful and versatile approach for uncertainty quantification across various applications. The objectives of this additional study are twofold: (i) to develop a DL model incorporating evidence-based theory for improved AF detection from Holter recordings, and (ii) to evaluate the advantages of the EDL model over traditional deterministic (softmax-based) DL models.

## 4.2 Related Works

In recent years, there has been growing interest in quantifying the uncertainty of DL models for AF detection. This section of the study reviews recent research, highlighting current approaches to addressing uncertainty in AF detection.

Belen *et al.* [205] employed a variational autoencoder DL model, integrating the Kullback-Leibler (KL) Divergence loss function, for AF detection using the AFDB dataset. To assess uncertainty, they iteratively fed the input data through the DL model and computed the standard deviation of the softmax probabilities. Vranken *et al.* [207] explored several UQ methods *e.g.*, MCD, VI, DE, and snapshot ensemble (SE) techniques. The efficacy of these methods in estimating uncertainties was assessed using rank-based metrics, calibration assessment, and out-of-distribution (OOD) detection. The findings revealed that the utilization of VI with Bayesian decomposition and ensemble methods with auxiliary output exhibited superior performance.

In [208], a weakly supervised learning approach was developed by incorporating the MCD approach to consider a limited amount of labeled data. The model achieved a classification performance with an F1-score ranging from 0.64 to 0.67 and an expected calibration error (ECE) ranging from 0.05 to 0.07. Aseeri *et al.* [206] developed a gated recurrent unit-based DL model trained using three types of datasets and estimated uncertainty using MCD and DE methods. They demonstrated that DE methods outperformed the MCD method. Elul *et al.* [209] conducted an extensive investigation into the integration of AI within clinical settings, focusing on the crucial role of uncertainty estimation in managing OOD instances and enabling multilabel diagnoses. Their approach involved the development of a DL model comprising 10 binary classifiers, each corresponding to distinct trained ECG abnormalities. This design facilitated the model's capacity to identify any combination of recognized rhythms and address unknown classes when the model generated negative predictions across all binary classifications. To gauge prediction confidence, they implemented the MCD method.

Zhang *et al.* [216] employed a Bayesian DL model with MCD for arrhythmia classification with a rejection option. They computed total uncertainty using an entropy-based decomposition of data and model uncertainty and

explored different uncertainty thresholds to improve classification performance by rejecting high-uncertainty instances. Jahmunah *et al.* [211] developed a Dirichlet distribution-based Densenet model with reverse KL divergence to compute predictive entropy for model uncertainty in a multi-class classification task. The authors argue that their approach is faster and computationally lightweight compared to previous uncertainty quantification methods. Additionally, they included noisy ECG in their analysis. Recently, Park *et al.* [210] proposed a self-attention-based LSTM-FCN DL architecture using a DE approach to quantify uncertainty. Their results achieved state-of-the-art performance, showing that epistemic uncertainty is reliable for classifying the six arrhythmia types.

## 4.3 Dataset and Preprocessing

In this study, three public ECG datasets are utilized to create the UQ benchmark: the IRIDIA-AF dataset, the MIT-BIH Long-Term Atrial Fibrillation (LTAF) dataset [32], and the AFDB dataset [217, 217]. The AFDB dataset is used exclusively for testing. The IRIDIA-AF is described in Sections 2.3. The LTAF and AFDB datasets are described below.

**LTAF**: This database contains 2-lead ECG signals from 84 patients of subjects with paroxysmal or sustained AF events with varying record durations but are typically 24 to 25 hours. The records are sampled at a frequency of 128 Hz. The rhythm annotations within the LTAF dataset are classified into two types: AF and N.

**AFDB**: This database contains 2-lead ECG signals from 23 patients sampled at a frequency of 250 Hz. The rhythm types within AFDB are classified into four types: AF, AFL, J (atrioventricular junctional rhythm), and N (sinus rhythm). In this study, the annotations of N are considered as "non-AF", while AF and AFL were merged as "AF".

To reduce baseline drift and powerline interference in the recordings, we used the same Butterworth filter configuration described in 2.3. A patient-wise partitioning technique is employed to ensure robust model development, validation, and testing, with the datasets split into training, validation, and testing sets in an 8:1:1 ratio. Each recording is segmented into non-overlapping 10-second windows. In this study, the NSR rhythm is categorized as "Non-AF", while AF and AFL rhythms are grouped together

as "AF". Table 4.1 provides a summary of the total number of 10-second AF and Non-AF segments across the three datasets.

TABLE 4.1: Number of 10-second segments of LTAF, IRIDIA-AF and AFDB datasets.

| Dataset | Training | | Validation | | Test | |
|---|---|---|---|---|---|---|
| | Non-AF | AF | Non-AF | AF | Non-AF | AF |
| LTAF | 193,470 | 168,707 | 34,410 | 18,291 | 14,991 | 26,839 |
| IRIDIA-AF | 1,150,662 | 353,063 | 331,214 | 94,759 | 390,131 | 88,521 |
| AFDB | – | – | – | - | 42,041 | 26,247 |

## 4.4 Model Architecture

We consider a DL model whose architecture consists of 18 layers. To manage the optimization of such a complex network, shortcut connections are incorporated, similar to the residual network architecture. The network comprises 8 residual blocks, each containing two convolutional layers. The number of residual blocks is selected by maximizing the accuracy on the validation set. These convolutional layers have a filter size of 3 and $32 \times 2^k$ filters, where $k$ is a hyper-parameter that starts at 0 and increments by 1 every two residual blocks. Additionally, every alternate residual block reduces the input size by a factor of 2 through subsampling.

To improve convergence and training stability, the ReLU activation function and batch normalization are applied after each convolutional layer. Furthermore, dropout with a probability of 0.3 is introduced to prevent overfitting. Subsequently, two dense layers comprising 128 and 64 neurons are employed. Each dense layer is followed by ReLU activation, batch normalization, and a dropout layer. Ultimately, a softmax activation function is utilized to generate a probability in AF detection. The model architecture is depicted in Figure 4.1. It is important to note that all UQ methods are employed within the same DL architecture. In the subsequent subsections, the most commonly used UQ methods are described in detail.

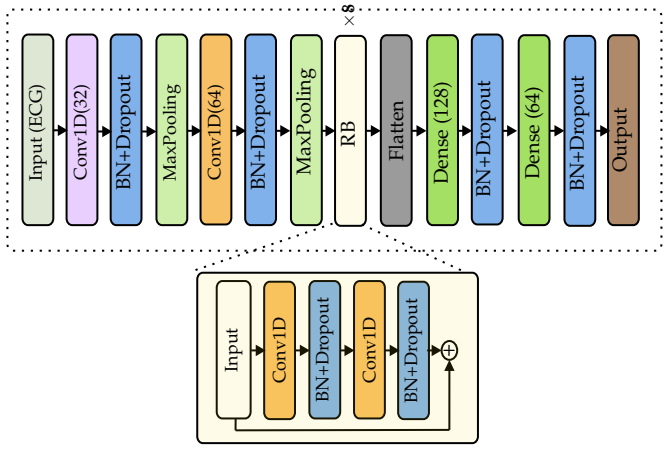



FIGURE 4.1: Diagram of the DL model. BN and RB stand for batch normalization and residual block, respectively.

## 4.5 Uncertainty Quantification Methods

Bayesian inference diverges from deterministic predictions by embracing a probabilistic approach. Instead of providing a single, definitive answer, it considers a range of possible values for model parameters, facilitating the incorporation of prior knowledge and the refinement of beliefs based on observed data.

To illustrate this concept, let us consider a training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ comprising $N$ instances and labels, considered sampled from the random variable $(X, Y) \sim P_{X,Y}$. For simplicity, let $\mathbf{x}_i \in \mathbb{R}^d$ denote a vector and $y_i$ a categorical variable. The input data $\mathbf{x}_i$ is fed into a neural network (NN) $\hat{y} = f_\theta(\mathbf{x}_i)$ with parameters $\theta$, yielding a classification output. This NN is conceived as a probabilistic model, where $f_\theta(\mathbf{x}_i) = P(Y \mid X = \mathbf{x}_i, \theta)$ and, differently from a deterministic approach, $\theta$ is considered a random variable as well. The posterior distribution of $\theta$ given the observed training set $\mathcal{D}$ can be used as a proxy for UQ using Bayesian inference.

In the context of Bayesian modeling, ensemble methods provide a means to quantify uncertainty by combining multiple models. The parameters of each model within the ensemble represent a distinct sample of the posterior distribution over the model parameters. Having at disposal different models, the ensemble prediction $\hat{y}^{\text{ensemble}}$ is obtained by aggregating individual

predictions from multiple models:

$$\hat{y}^{\text{ensemble}} = \frac{1}{M} \sum_{j=1}^{M} f_{\theta_j}(\mathbf{x}_i), \tag{4.1}$$

where $j = 1, 2, \ldots, M$ denotes the distinct models in the ensemble and $\hat{y}^{\text{ensemble}}$ indicates the probabilities for the label to predict. For all methods investigated in our study, we utilize an ensemble comprising $M = 4$ models, each parameterized differently to capture a diverse array of hypotheses from the DL model. This approach is implemented for the UQ methods, details of which are provided in the following subsections.

### 4.5.1 Monte Carlo Dropout

MCD is a powerful technique within the realm of DL that extends the traditional dropout regularization method [218]. In standard dropout, random units are dropped during training to prevent overfitting and encourage model robustness. MCD takes this concept further by employing dropout not only during training but also during the inference phase of a model. Instead of obtaining a single deterministic prediction, the model is run multiple times with dropout enabled, generating a distribution of predictions. The final prediction is then derived from the mean of these sampled predictions.

### 4.5.2 Ensemble Method with Different Initializations (DE)

In this approach, we leverage the fact that the typical strategy for training a NN is to initialize its weights randomly and then adjusting them through back-progatation. Here, we trained $M = 4$ different NNs with four different intializations and obtained $\hat{y}^{\text{ensemble}}$ as the average of these four output probabilities [219].

### 4.5.3 Snapshot Ensemble

SE creates multiple models through the training of a DL model using distinct snapshots of its parameters, obtained at various epochs during the training process [213]. These individual snapshots encapsulate the configuration of the model at different epochs, thereby offering diverse vantage points on the data manifold. The predictions derived from these varied models within

the SE framework possibly serve not only to enhance predictive accuracy but also to furnish a more robust estimation of uncertainty belonging to the predictions of the models. In this study, with $M = 4$, we took a snapshot at every 20 epochs to develop the SE model.

### 4.5.4 Batch Ensemble

Unlike traditional ensembles that combine predictions from independently trained models, batch ensemble (BE) utilizes ensemble members that share the same weights during training [220]. BE builds up an ensemble from a single base network (shared among ensemble members) and a set of layer-specific weight matrices unique to each member.

At each layer, the weight of each ensemble member is generated from the Hadamard product between a weight matrix shared among all ensemble members, called "slow weights" and a Rank-1 matrix that varies among all members, called "fast weights". Formally, let $W_{\text{share}} \in \mathbb{R}^{u \times v}$ be the slow weights in an NN layer with input dimension $u$ and output dimension $v$. Each member $m$ from an ensemble of size $M$ owns a fast weight matrix $W_m \in \mathbb{R}^{u \times v}$. $W_m$ is a Rank-1 matrix computed from a tuple of trainable vectors $r_m \in \mathbb{R}^u$ and $s_m \in \mathbb{R}^v$, with $W_m = r_m s_m^\top$. BE generates from them a family of ensemble weights by $\overline{W_m} = W_{\text{share}} \odot W_m$, where $\odot$ denotes the Hadamard product. Each $\overline{W_m}$ member of the ensemble is essentially a Rank-1 perturbation of the shared weights $W_{\text{share}}$. We implemented BE on all convolutional and dense layers in the NN. The loss function was Binary Cross Entropy averaged across the ensemble members.

### 4.5.5 Packed Ensemble

The utilization of ensemble methods is widely recognized for its advantages. However, a significant drawback is the considerable increase in both training time and memory usage during inference, which scales linearly with the number of models employed. To address these challenges, Olivier *et al.* [221] introduced the pack-ensembled (PE) method. This approach leverages grouped convolutions to significantly expedite the training and inference computations of ensembles. Grouped convolutions offer computational advantages by reducing the size of the subnetworks. Group convolutions can be extended to dense layers as well and the number of groups was set to $M = 4$. Here, PE was used for all convolutional and dense layers.

### 4.5.6 Mean Field Variational Inference

Mean field variational inference (MFVI) is a technique used in the Bayesian framework to approximate complex posterior distributions [222]. The goal of MFVI is to approximate the true posterior distribution by parameterizing it with a simpler, factorized distribution, known as the mean field distribution. The true posterior is often difficult to compute analytically due to its complexity. MFVI seeks to approximate this distribution by a factorized parametric distribution that factorizes over the individual parameters:

$$q(\theta;\omega_1,\cdots,\omega_K) = \prod_{i=1}^{K} q_i(\theta_i;\omega_i), \tag{4.2}$$

where $\theta_i$ represents the $i$-the parameter of the model, $\omega_i$ the parameters of the $i$-th $q_i(\theta_i;\omega_i)$ distribution, $q(\theta;\omega_1,\cdots,\omega_K)$ represents the complete variational distribution, and $K$ is the total number of parameters. Each distribution $q_i(\theta_i;\omega_i)$ was set as normal distribution over the variable $\theta_i$. The mean field approximation implies that the parameters are assumed to be independent given the mean field distribution. The objective is to find the mean field parameters $\omega_i$ that minimize the Kullback-Leibler (KL) divergence between the true posterior and the mean-field approximation.

Minimizing this divergence is equivalent to maximizing the Evidence Lower Bound (ELBO), which is defined as:

$$\text{ELBO} = \mathbb{E}_{q(\theta;\omega_1,\cdots,\omega_K)}[\log p(Y|X,\theta) - \log q(\theta;\omega_1,\cdots,\omega_K)] \tag{4.3}$$

which is a tractable objective function that can be optimized using various optimization algorithms, such as stochastic gradient descent (SGD).

MFVI was implemented in the first and last layers of our NN.

### 4.5.7 Rank-1 MFVI

Rank-1 MFVI method aims to approximate complex probability distributions by introducing a simplified, tractable family of distributions [223]. This approach merged the key ideas from BE and MFVI by constructing a posterior distribution over the parameters of the rank-1 matrices $rs^\intercal$. Similar to MFVI, we used the normal distribution for each of these parameters. Please, notice that in this case there are no four members, but only one.

Rank-1 MFVI was used for all convolutional and dense layers. The ELBO was maximized for this method too.

### 4.5.8 Stochastic Weighting Average Gaussian

Stochastic weighted average (SWA) centers around a learning rate schedule within SGD, and considers the weights of the models it encounters at consecutive epochs [224]. In this method, the weights obtained after each epoch, denoted as $\theta^{(e)}$, contribute to a running average, *i.e.*, the SWA solution, after $T$ epochs: $\theta_{\text{SWA}} = \frac{1}{T}\sum_{e=1}^{T}\theta^{(e)}$.

Maddox *et al.* [225] extends this method to estimate Gaussian posteriors for model parameters, by also estimating a covariance matrix for the parameters, using a low-rank plus diagonal posterior approximation. The diagonal part is obtained by keeping a running average of the second uncentered moment of each parameter, and then at the end of the training calculating:

$$\Sigma_{\text{diag}} = \text{diag}\left(\frac{1}{T}\sum_{e=1}^{T}{\theta^{(e)}}^2 - \theta_{\text{SWA}}^2\right) \tag{4.4}$$

while the diagonal part is approximated by keeping a matrix $GG^\top$ with columns $G_e = (\theta^{(e)} - \hat{\theta}^{(e)})$, $\hat{\theta}^{(e)}$ standing for the running estimate of the parameters' mean obtained from the first $e$ epochs. The rank of the approximation is restricted by retaining last $L$ vectors of the $G_e$ vectors and dropping the previous, with $L$ being a hyperparameter of the model, as follows

$$\begin{aligned}\Sigma_{\text{low-rank}} &\approx \frac{1}{L-1}GG^\top \\ &= \frac{1}{L-1}\sum_{e=T-L+1}^{T}(\theta^{(e)} - \hat{\theta}^{(e)})(\theta^{(e)} - \hat{\theta}^{(e)})^\top\end{aligned} \tag{4.5}$$

The overall posterior approximation is given by:

$$\theta_{\text{SWAG}}|\mathcal{D} \sim \mathcal{N}\left(\theta_{\text{SWA}}, \frac{1}{2}(\Sigma_{\text{diag}} + \Sigma_{\text{low-rank}})\right) \tag{4.6}$$

Once the posterior distributions are approximated, the model is used at test time by sampling from these approximations. Specifically, we used $T = 80$, $L = 4$ and a learning rate schedule dropping the learning rate by 25% every 20 epochs. After training, we draw $M = 4$ samples from

the approximated posterior distributions and compute the average of the predicted distributions from these samples.

### 4.5.9 Improved Variational Online Gauss-Newton

Improved variational online Gauss-Newton (iVOGN) introduces an enhanced Bayesian learning algorithm tailored to address positive-definite constraints within the learning process [226]. This method is part of the variational inference domain where the posterior distribution is approximated by a simpler one $q(\theta|\omega)$, where $\omega$ are the parameters. However, the parameters $\omega$ most often requires to satisfy constraints. For example, when a multivariate Gaussian variable is used for such approximation, the covariance matrix must be positive-definite. In this study, we assume the approximate distribution $q(\theta|\omega)$ to be a multivariate Gaussian distribution where $\omega$ representing the average and the covariance matrix for the multivariate variable $\theta$. The method ensures the covariance matrix to be positive-definite. Here, all model parameters were considered for this approximation.

### 4.5.10 Stein Variational Gradient Descent

Stein variational gradient descent (SVGD) is a gradient-based sampling algorithm for approximate inference [227]. Briefly, let consider the posterior distribution $p(\theta|\mathcal{D})$, SVGD finds a set of $n$ particles $\{z_i\}_{i=1}^n$ to approximate the posterior $p$. Each particle $z$ is a vector containing the model parameters. Particles' "positions" are updated by the following expression:

$$z_i \leftarrow z_i + \epsilon \frac{1}{n} \sum_{j=1}^{n} \left[ k(z_j, z_i) \nabla_{z_j} \log p(z_j|\mathcal{D}) + \nabla_{z_j} k(z_j, z_i) \right], \tag{4.7}$$

for all $i = 1, \ldots, n$, where $\epsilon$ is a step-size, and $k(z, z')$ is any positive definite kernel specified by the users, such as the radial basis function kernel $k(z, z') = \exp\left(-\frac{1}{h}\|z - z'\|_2^2\right)$, which can be thought of as encoding some similarity measure between different particles $z$. In this update, the term that contains the gradient of $\log p(z|\mathcal{D})$ drives the particles towards the high probability regions of $p(\theta|\mathcal{D})$, while the term with $\nabla_{z_j} k(z_j, z_i)$ acts as a repulsive force to push $z_i$ away from $z_j$ to avoid the particles from collapsing together. All model parameters were sampled using this technique, and SVGD's hyper-parameters were $n = 10$, $h = 10$ and $\epsilon = 0.001$.

### 4.5.11 Last Layer Laplace Approximation (LLLA)

Laplace approximation (LA) is derived through a second-order Taylor expansion centered around the mode of a distribution [228]. The mode can be determined using conventional gradient-based methods or, as in our case, substituted with a local optimum found with gradient descent. Specifically, this is achieved by approximating the log posterior over the weights of a NN given a dataset $\mathcal{D}$ around the Maximum A Posteriori (MAP) estimate $\theta_{MAP}$. Mathematically, this can be represented by the following expression.

$$\log p(\theta|\mathcal{D}) \approx \log p(\theta_{\text{MAP}}|\mathcal{D}) - \frac{1}{2}(\theta - \theta_{\text{MAP}})^{\top} \bar{H} (\theta - \theta_{\text{MAP}}), \tag{4.8}$$

where $\theta$ are the model parameters, and $\bar{H}$ denotes the Hessian of the negative log posterior. The absence of the first-order term is due to the expansion around a maximum ($\theta_{\text{MAP}}$), where the gradient is zero. Upon exponentiating this equation, it becomes evident that the right-hand side adopts a Gaussian functional form for $\theta$, leading to the approximation of a normal distribution through integration. The posterior over the weights is approximated as:

$$\theta|\mathcal{D} \sim \mathcal{N}(\theta_{MAP}, \bar{H}^{-1}). \tag{4.9}$$

In our study, we implement LA only for the last layer of our DL model.

### 4.5.12 Evidential Deep Learning

EDL combines DL with uncertainty quantification using evidence theory, specifically Dempster-Shafer Theory [215, 229]. Unlike softmax-based DL models, which produce point predictions, EDL models aim to explicitly model epistemic uncertainty in predictions in a principled way. In EDL, a set of belief masses $b_k \geq 0$ is assigned to each class $k \in [1, K]$, representing the potential class labels for a given input. These belief masses, along with an overall uncertainty mass $u \geq 0$ (corresponding to an "I don't know" category), must satisfy the constraint: $u + \sum_{k=1}^{K} b_k = 1$. The belief mass $b_k$ for class $k$ and the uncertainty mass $u$ are defined as follows:

$$b_k = \frac{e_k}{S}, \quad u = \frac{K}{S}, \quad S = K + \sum_{k=1}^{K} e_k, \tag{4.10}$$

where $e_k$ represents the evidence supporting the assignment of the input to class $k$. This formulation shows that as the evidence $e_k$ for a class increases, the associated uncertainty $u$ decreases. Conversely, in the absence of any evidence, $u = 1$ reflects complete uncertainty. This framework relates to the Dirichlet distribution through the concentration parameter $\alpha_k = e_k + 1$ for class $k$. The Dirichlet distribution, serving as the conjugate prior to the categorical distribution, enables sampling of probability assignments across all possible classes, denoted as $p \sim \text{Dir}(p \mid \alpha)$ and $\hat{y} \sim \text{Cat}(y \mid p)$. The expected probability for class $k$ is computed as the mean of its corresponding Dirichlet distribution: $\hat{p}_k = \alpha_k / S$.

**Evidential Loss Function**

Let $f(x \mid \Theta)$ represent the evidence vector $\mathbf{e} \in \mathbb{R}^K$ predicted by the DL model for an input $x$, with $\Theta$ denoting the model parameters. The parameters of the corresponding Dirichlet distribution are defined as $\boldsymbol{\alpha} = f(x \mid \Theta) + 1$.

Given an observation $x_i$, let $y_i$ be the one-hot encoded vector representing the true class, where $y_{ij} = 1$ for the true class $j$ and $y_{ik} = 0$ for all $k \neq j$. The Dirichlet distribution $\text{Dir}(\mathbf{p}_i \mid \alpha_i)$ is used to model the uncertainty over the class probabilities $\mathbf{p}_i$. The Dirichlet distribution acts as a prior for the multinomial likelihood of the class labels.

To train the model, we minimize the following evidential loss function for each instance $i$, which penalizes incorrect predictions while accounting for uncertainty:

$$\begin{aligned} L_{\text{ev},i}(\Theta) &= \int \|\mathbf{y}_i - \mathbf{p}_i\|^2 \frac{1}{\text{B}(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{(\alpha_{ij}-1)} \, d\mathbf{p}_i \\ &= \sum_{j=1}^{K} (y_{ij} - \hat{p}_{ij})^2 + \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{S_i + 1}, \end{aligned} \tag{4.11}$$

where $\hat{p}_{ij} = \alpha_{ij} / S_i$ with $S_i = \sum_{j=1}^{K} \alpha_{ij}$ is the total evidence for observation $i$, and $\text{B}(\alpha_i)$ represents the multivariate Beta function. The term $\|\mathbf{y}_i - \mathbf{p}_i\|^2$ corresponds to the squared error between the ground-truth label and the predicted class probabilities, and the second term regularizes the variance of the predicted probabilities. Additionally, to prevent overconfidence and encourage appropriate uncertainty, a regularization term is introduced

by penalizing the divergence between the predicted Dirichlet distribution $\mathrm{Dir}(\mathbf{p}_i \mid \alpha_i)$ and the uniform Dirichlet prior $\mathrm{Dir}(\mathbf{p}_i \mid \mathbf{1})$. The total loss function is thus formulated as:

$$L(\Theta) = \sum_{i=1}^{N} L_{\mathrm{ev},i} + \lambda_t \sum_{i=1}^{N} \mathrm{KL}\left(\mathrm{Dir}(\mathbf{p}_i \mid \tilde{\alpha_i}) \| \mathrm{Dir}(\mathbf{p}_i \mid \mathbf{1})\right), \tag{4.12}$$

where $\lambda_t$ is an annealing coefficient that controls the impact of the regularization term during training. It is defined as $\lambda_t = \min\left(1, \frac{t \cdot \text{Batch size}}{10 \cdot N}\right)$ where $t$ is the current training step, and $N$ represents the total number of training samples. The $\mathrm{KL}(\cdot \| \cdot)$ represents the Kullback-Leibler divergence between two Dirichlet distributions. The adjusted Dirichlet parameters $\tilde{\alpha}_i$ are computed to update the prior belief about the class probabilities by incorporating the true label, effectively removing non-informative evidence and focusing on the true class. They are computed as: $\tilde{\alpha}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \alpha_i$. Equation 4.12 enables the model to simultaneously minimize the prediction error and maintain calibrated uncertainty, particularly in cases where the available evidence is limited.

### 4.5.13 Training Details

We used an NVIDIA A100 80GB GPU to run all methods within the PyTorch environment. For training the DL model, we employed the Adam optimizer with a learning rate of 0.001 and a batch size of 128. The training process was capped at a maximum of 100 epochs, with early stopping applied to prevent overfitting: training was halted if the validation loss did not improve for 6 consecutive epochs. Given the imbalance in the dataset, we addressed this issue by incorporating focal loss during training, with parameters set to $\alpha = 0.1$ and $\gamma = 2$.

## 4.6 Results

### 4.6.1 Evaluation Metrics

In this section, we present the results of our benchmark study on uncertainty estimation in DL models for AF detection. We evaluate the performance of the models using key metrics, including sensitivity, specificity, ECE, AUC and negative log-likelihood (NLL).

Sensitivity and specificity are fundamental metrics for assessing the performance of AF detection from the DL model. Sensitivity measures the ability of the model to correctly identify positive cases (AF), while specificity gauges the model's accuracy in identifying negative cases (Non-AF).

ECE is a measure of how well the predicted probabilities align with the actual outcomes. It measures the difference between the average predicted probability and the actual observed frequency of events across various confidence intervals. Lower ECE values indicate better calibration. Formally, ECE is computed as follows:

$$\begin{aligned} \mathrm{ECE} &= \sum_{z=1}^{Z} \frac{|B_z|}{N} \left|\mathrm{acc}(B_z) - \mathrm{conf}(B_z)\right| \\ \mathrm{acc}(B_z) &= \frac{1}{|B_z|} \sum_{i \in B_z} \mathbb{I}(\hat{y}_i = y_i) \\ \mathrm{conf}(B_z) &= \frac{1}{|B_z|} \sum_{i \in B_z} \hat{p}_i \end{aligned} \tag{4.13}$$

where $\mathbb{I}(\hat{y}_i = y_i)$ denotes the indicator function, which equals 1 if the predicted label $\hat{y}_i$ matches the true label $y_i$ for the $i$-th sample, and 0 otherwise. $B_z$ is the set of samples whose confidence predicted by the model (*i.e.*, model's output probability) is in the interval $[z-1, z)/Z$ where $Z$ represents the total number of bins. $N$ is the total number of instances across all bins, $\mathrm{acc}(B_z)$ denotes the accuracy of the $z$-th bin, and $\mathrm{conf}(B_z)$ refers to the average confidence score of the samples in the $z$-th bin. $i \in B_z$ indicates a subset of instances that have similar confidence scores and are grouped together in the same bin. We set the total number of bins $Z = 10$.

NLL is a measure of how well model's predicted probabilities match the true distribution of the data. Lower NLL values suggest better alignment. Mathematically, the NLL is formulated as:

$$\mathrm{NLL} = -\frac{1}{N} \sum_{i=1}^{N} \log\left(p_{i,y_i}\right), \tag{4.14}$$

where $p_{i,y_i}$ denotes the predicted probability for instance $i$ and the correct class $y_i$.

TABLE 4.2: Performance for different UQ methods on the test set of IRIDIA-AF dataset.

| Model | Sensitivity | Specificity | AUC | ECE | NLL |
|---|---|---|---|---|---|
| MCD | 0.936 | 0.937 | 0.962 | 0.054 | 0.241 |
| DE | 0.935 | 0.936 | 0.961 | 0.054 | 0.251 |
| SE | 0.937 | 0.937 | 0.962 | 0.054 | 0.230 |
| BE | 0.934 | 0.948 | 0.973 | 0.053 | 0.622 |
| PE | 0.937 | 0.942 | **0.974** | **0.046** | 0.312 |
| SGVD | 0.840 | 0.850 | 0.911 | 0.089 | 0.441 |
| MFVI | 0.833 | 0.841 | 0.892 | 0.082 | 0.412 |
| MFVI (rank-1) | 0.830 | 0.840 | 0.901 | 0.091 | 0.433 |
| SWAG | 0.929 | **0.955** | 0.971 | 0.056 | 0.201 |
| iVOGN | **0.943** | 0.880 | 0.951 | 0.115 | 0.272 |
| LLLA | 0.935 | 0.935 | 0.961 | 0.069 | 0.085 |

### 4.6.2 Comparative Performance for Different UQ Methods

Table 4.2 presents a comparative performance of UQ methods applied to the test set of IRIDIA-AF dataset. Methods such as PE, BE, and SWAG stand out for their robust performance, characterized by high sensitivity, specificity, AUC, and well-calibration (low ECE). Additionally, these methods demonstrate competitive performance in terms of NLL, indicating their ability to provide accurate probabilistic predictions. In contrast, SGVD and MFVI exhibit comparatively weaker performance across these metrics, indicating more uncertainty, poorer calibration, and less accurate probabilistic predictions.

Additionally, Table 4.3 presents the performance of each UQ method on the LTAF dataset. Similar to the results on the IRIDIA-AF dataset, PE and BE exhibit high AUC scores and low ECE and NLL values, underscoring their superior discriminative ability and calibration. In contrast, models like SGVD and iVOGN show lower performance and higher uncertainty.

Table 4.4 presents the performance various UQ methods trained on the LTAF dataset and tested on the AFDB (external) dataset. In this case, the MCD, DE, SE, SWAG, and iVOGN methods demonstrate competitive sensitivities, specificities, and AUC scores. MCD provides a low NLL of 0.08, while DE, SE, SWAG, and iVOGN exhibit higher NLLs greater than 0.11. The PE and BE methods demonstrate high sensitivities of 0.959 and 0.872, respectively, along with competitive AUC scores of 0.94 and 0.88, indicating

TABLE 4.3: Performacne for UQ methods on the test set of LTAF dataset.

| **Model** | **Sensitivity** | **Specificity** | **AUC** | **ECE** | **NLL** |
|---|---|---|---|---|---|
| MCD | 0.826 | 0.978 | 0.960 | 0.095 | 0.092 |
| DE | 0.834 | 0.978 | 0.960 | 0.092 | 0.123 |
| SE | 0.915 | 0.978 | 0.970 | 0.061 | 0.232 |
| BE | 0.951 | 0.988 | 0.991 | 0.020 | 0.121 |
| PE | **0.996** | **0.994** | **0.992** | **0.007** | **0.021** |
| SWAG | 0.941 | 0.948 | 0.960 | 0.051 | 0.201 |
| MFVI | 0.851 | 0.931 | 0.971 | 0.087 | 0.232 |
| MFVI (rank-1) | 0.885 | 0.941 | 0.971 | 0.067 | 0.210 |
| SGVD | 0.802 | 0.931 | 0.920 | 0.121 | 0.191 |
| iVOGN | 0.791 | 0.930 | 0.920 | 0.112 | 0.173 |
| LLLA | 0.873 | 0.986 | 0.992 | 0.013 | 0.061 |

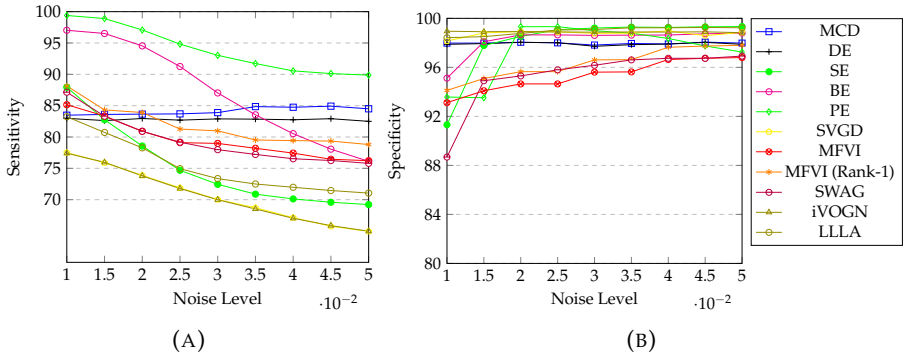
TABLE 4.4: Performance of different UQ methods on the entire AFDB dataset.

| **Model** | **Sensitivity** | **Specificity** | **AUC** | **ECE** | **NLL** |
|---|---|---|---|---|---|
| MCD | 0.887 | 0.602 | 0.861 | 0.150 | 0.080 |
| DE | 0.897 | 0.601 | 0.871 | 0.131 | 0.110 |
| SE | 0.894 | 0.648 | 0.851 | **0.081** | 0.141 |
| BE | 0.872 | 0.632 | 0.881 | 0.160 | **0.050** |
| PE | **0.959** | 0.578 | **0.941** | 0.133 | 0.071 |
| SWAG | 0.836 | 0.693 | 0.862 | 0.101 | 0.142 |
| MFVI | 0.827 | 0.622 | 0.861 | 0.112 | 0.091 |
| MFVI (rank-1) | 0.857 | 0.602 | 0.871 | 0.142 | 0.101 |
| SGVD | 0.826 | 0.703 | 0.870 | 0.099 | 0.152 |
| iVOGN | 0.940 | 0.621 | 0.861 | 0.110 | 0.181 |
| LLLA | 0.893 | **0.752** | 0.891 | 0.159 | 0.341 |

their effectiveness in identifying true positive cases. Despite lower specificities of 0.578 and 0.632, PE and BE maintain relatively low NLL values of 0.07 and 0.05, suggesting robust probabilistic predictions. In contrast, SGVD and MFVI (rank-1) show varying performance, with SGVD demonstrating a higher specificity of 0.703 but a relatively lower sensitivity of 0.826 compared to PE. These models also exhibit moderate to high NLL values, indicating potential challenges in probabilistic modeling for AF detection on the AFDB dataset. Finally, the LLLA method exhibits balanced sensitivity of 0.893 and specificity of 0.752, achieving an AUC of 0.89. However, its higher NLL of 0.34 indicates significant challenges in probabilistic predictions despite its overall balanced performance.

### 4.6.3 External Validation

Our study underscores the critical importance of selecting UQ methods that maintain consistent performance across external test sets, particularly those trained on one dataset and tested on another. In this case, we trained a DL model using the LTAF dataset and evaluated its performance on the AFDB dataset, with results detailed in Table 4.4. Notably, the sensitivity dropped from approximately 90% during internal validation to a range of 57.74% to 75.28% during external testing. These findings are consistent with existing literature, such as the work by Seo et al. [230], which demonstrates that models trained on data from a specific source may not generalize well to external datasets, underscoring the adage "one-size-does-not-fit-all".




(A) (B)

FIGURE 4.2: Comparative performance of different UQ methods under random noise addition. (A) Sensitivity across models with random noise addition to the ECG signal. (B) Specificity across models with random noise addition to the ECG signal.

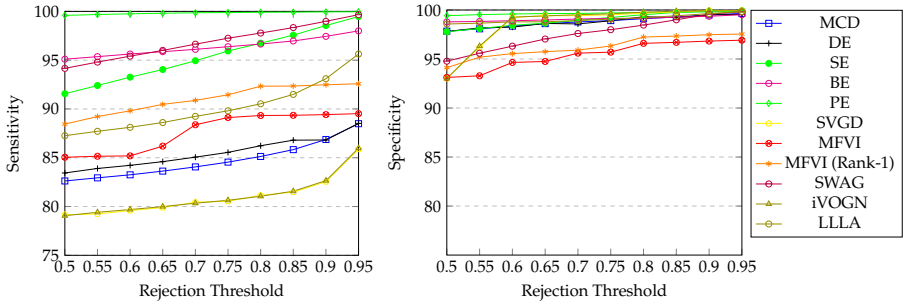
### 4.6.4 Impact of Random Noise Addition

By evaluating UQ methods under noisy conditions, our study provides insights into their reliability in real-world environments where data may be corrupted. To achieve this, we added a Gaussian noise to the ECG signal during inference, with standard deviations ranging from 0.01 to 0.055 mV with step of 0.005 mV, ensuring a comprehensive assessment of the UQ methods' performance under varying noise intensities. The sensitivity analysis, illustrated in Figure 4.2a, portrays the performance of various UQ methods across distinct noise levels on the test set of the LTAF dataset. Notably, the PE model outperforms its counterparts in this scenario. Despite the consistent performance of the MCD and DE methods, their computational demands and inference times are notably higher compared to the PE model, as shown in Table 4.5. In Figure 4.2b, the SE, BE, and PE models exhibit consistent performance across varying noise levels. Interestingly, under specific noise levels, a trade-off is observed between the specificity and sensitivity of UQ methods. While specificity diminishes, sensitivity increases, indicating a major impact of noise on the performance of these methods.

### 4.6.5 Classification with a Rejection Thresholds

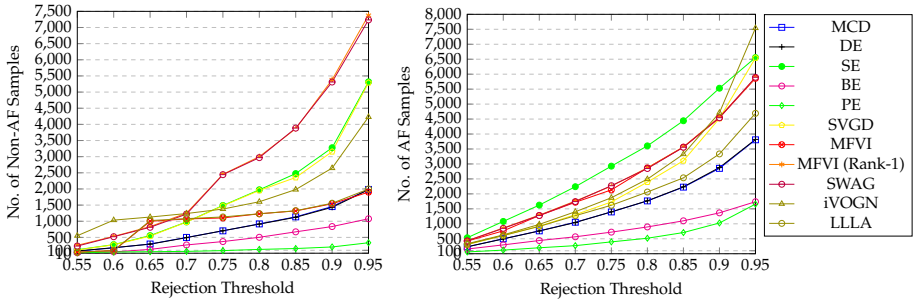
Classification with a rejection threshold, also known as reject inference, may enhance conventional classification models by enabling them to discard predictions when uncertainty is high. This approach is particularly beneficial in handling inputs that present classification challenges, as making low-confidence predictions could lead to errors.

In this study, we implement a decision threshold mechanism to assign class labels based on predicted probabilities. By varying the threshold from 0.55 to 0.95 with the interval of 0.05, we let the model to reject predictions when confidence is below the threshold. Our findings show that increasing the rejection threshold enhances the classifier's performance. Figure 4.3a demonstrates the sensitivity of various UQ methods across different rejection thresholds, revealing that higher thresholds improve overall performance with minimal impact on PE methods. Similarly, Figure 4.3b shows a corresponding trend in specificity across different UQ methods, paralleling the sensitivity results.




(A) (B)

FIGURE 4.3: Comparative performance of different UQ methods under rejection thresholds. (A) Sensitivity (B) Specificity




(A) (B)

FIGURE 4.4: Number of samples are discarded for different UQ methods under different rejection thresholds. (A) No. of Non-AF samples. (B) No. of AF samples. The total number of Non-AF and AF samples are 14,991 and 26,839, respectively, in the test set of the LTAF dataset.

In Figure 4.4, the rejection rate for each method pertains to the number of AF and Non-AF instances discarded at various thresholds. As the rejection threshold increases, the model becomes more confident in discarding instances, leading to higher rejection rates. Methods such as SE, SVGD, MFVI (Rank-1), SWAG, and iVOGN exhibit high rejection rates for AF cases, indicating a strong sensitivity to uncertainty and a more aggressive approach. This conservatism reduces false positives but risks over-rejecting true AF cases, potentially missing some genuine AF instances. In contrast, methods like MCD, DE, and MFVI show a more balanced rejection rate, increasing steadily with higher thresholds. These methods provide a moderate trade-off, aiming to reject Non-AF instances while minimizing the risk of misclassifying true AF cases as Non-AF. On the other hand, BE, PE, and LLLA demonstrate a more conservative strategy, rejecting fewer AF cases compared to other methods. This cautious approach suggests a focus on reducing false negatives, ensuring that fewer true AF cases are incorrectly classified as Non-AF. Thus, selecting the appropriate method involves balancing the rejection of Non-AF instances with the risk of missing true AF cases, tailored to the specific needs and priorities of the application.

### 4.6.6 Efficiency of UQ Methods

In addition to assessing performance metrics, our study explores the practical efficiency of various UQ methods by evaluating trainable parameters, inference time, and floating-point operations (FLOPs). Understanding the computational efficiency of these methods is crucial for their integration into clinical workflows, where resource constraints are prevalent.

Table 4.5 provides an evaluation of UQ techniques for efficiency on the LTAF dataset. Among these methods, MCD and DE demonstrate moderate inference times, each equipped with one million parameters and 5056 billion FLOPs. Notably, the SE stands out due to its expedited inference time, marginally reduced parameter count (0.99 million), and significantly lower FLOPs (1,686 billion), making it particularly suitable for efficiency-oriented applications.

The MFVI (rank-1) exhibits an increase in both parameters and FLOPs. While SGVD, iVOGN, and LLLA show competitive inference times, their performance metrics—including sensitivity, specificity, AUC, ECE, and

NLL—do not meet high standards. Conversely, PE excels in efficiency, boasting an exceptionally fast inference time of $3.36 \times 10^{-5}$ seconds, a modest 0.07 million parameters, and minimal 130 billion FLOPs. This unparalleled combination of speed and simplicity positions PE as an attractive choice for scenarios with stringent computational resource constraints.

Similarly, BE achieves an optimal balance, demonstrating a reasonable inference time of $9.48 \times 10^{-5}$ seconds, 0.26 million parameters, and 421 billion FLOPs. This notable performance, particularly with respect to both speed and model complexity, highlights the advantages of PE and BE over other methods in our comparative analysis. Their efficiency makes them prominent candidates in scenarios where resource optimization is of paramount importance.

TABLE 4.5: Average efficiency of UQ methods for a 10-s segment of the test set of LTAF dataset

| Model | Inference Time (s) | Parameters ($10^6$) | FLOPs ($10^9$) |
|---|---|---|---|
| MCD | $1.33 \times 10^{-4}$ | 1.00 | 5056 |
| DE | $1.33 \times 10^{-4}$ | 1.00 | 5056 |
| SE | $8.00 \times 10^{-5}$ | 0.99 | 1686 |
| BE | $9.48 \times 10^{-5}$ | 0.26 | 421 |
| PE | $3.36 \times 10^{-5}$ | 0.07 | 130 |
| SWAG | $3.20 \times 10^{-5}$ | 0.25 | 421 |
| MFVI | $3.17 \times 10^{-5}$ | 0.25 | 1264 |
| MFVI (Rank-1) | $1.89 \times 10^{-4}$ | 0.52 | 842 |
| SGVD | $3.23 \times 10^{-5}$ | 0.25 | 421 |
| iVOGN | $3.23 \times 10^{-5}$ | 0.25 | 421 |
| LLLA | $3.17 \times 10^{-5}$ | 0.25 | 1264 |

Overall, our comprehensive analysis of diverse UQ methods in AF detection highlights the multifaceted considerations essential for their effective application in clinical settings. By evaluating these methods through external validation, robustness under noise, classification with rejection options, and computational efficiency, we provide a holistic view of their strengths and limitations. Notably, while methods like PE and BE demonstrate superior efficiency and performance consistency, their integration into clinical workflows must be carefully balanced with the specific requirements of sensitivity, specificity, and computational resources.

### 4.6.7 Results for EDL

We trained and evaluated the EDL model, considering the architecture of the RTA model and the datasets used in our study. For more detailed information about the RTA model and the datasets, please refer to Section 2.2 and Section 2.3, respectively. Table 4.6 presents a comparative analysis of the performance of both our RTA model and the EDL model across key metrics. Both models demonstrate strong performance in the Non-AF and AF classes, achieving identical recall scores of 0.98, indicating their ability to accurately detect Non-AF and AF cases. They also exhibit excellent discriminative power for Non-AF and AF, with nearly identical AUC scores of 0.982 for Non-AF and 0.977 for AF.

TABLE 4.6: Performance of softmax-based DL and EDL model for AF detection.

| **Model** | **Recall** | | | **AUC** | | | **ECE** |
|---|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL | |
| Softmax-DL (RTA Model) | 0.951 | 0.931 | 0.812 | 0.981 | 0.973 | 0.942 | 0.16 |
| EDL | 0.953 | 0.934 | 0.838 | 0.982 | 0.977 | 0.972 | 0.09 |

Table 4.6 presents a comparative analysis of the performance of both our RTA model and the EDL model across key metrics. Both models demonstrate strong performance in the Non-AF and AF classes, achieving identical recall scores of approximately 0.95 for Non-AF and 0.93 for AF, indicating their ability to accurately detect Non-AF and AF cases. They also exhibit excellent discriminative power for Non-AF and AF, with nearly identical AUC scores of 0.982 for Non-AF and 0.977 for AF.

A slight difference emerges in the AFL class, where the EDL model outperforms the softmax-DL model, achieving a recall of 0.838 compared to 0.812. Additionally, the EDL model shows superior AUC performance for AFL, with a score of 0.972 compared to 0.942 for the softmax-DL model, highlighting the EDL model's enhanced ability to detect and differentiate AFL cases.

Furthermore, the EDL model demonstrates better calibration, as indicated by its lower ECE of 0.09, compared to the softmax-DL model's ECE of 0.16. This suggests that the EDL model's predicted probabilities are more accurately aligned with the true outcomes.

The results of this study highlight the effectiveness of the EDL model for AF detection from Holter ECG recordings, addressing key limitations of the softmax-based DL model. By explicitly modeling uncertainty, this framework offers significant advantages over the softmax-based DL model, particularly in clinical settings where decision-making under uncertainty is critical.

## 4.7 Discussion

This study conducts a comprehensive analysis of different UQ methods, focusing specifically on their application in AF detection. Acknowledging the pivotal role of uncertainty awareness in clinical decision-making, our objective is to identify the most suitable UQ methods for this purpose.

PE and BE consistently performed well across the IRIDIA-AF, LTAF, and AFDB datasets, demonstrating high sensitivity, specificity, and AUC values. These methods also provided low ECE and competitive NLL scores, reflecting their robustness and calibration abilities. This aligns with findings from previous studies which suggest that methods incorporating probabilistic approaches or ensembles, such as Bayesian methods and perturbation-based techniques, can offer enhanced reliability [231, 221].

SGVD and MFVI showed comparatively weaker performance, especially in terms of sensitivity and calibration. These methods faced challenges in aligning predicted probabilities with true outcomes, resulting in higher ECE and NLL values. The less effective performance of SGVD and MFVI is consistent with literature indicating that methods relying on variational inference or optimization techniques might struggle with calibration and probabilistic accuracy [232, 233].

The drop in performance when models trained on the LTAF dataset were evaluated on the AFDB dataset underscores a common issue in ML and medical diagnostics: model generalizability. The significant decrease in sensitivity from internal to external validation highlights the challenge of achieving robust performance across different datasets. This finding emphasizes the need for models to be trained on diverse datasets to improve their generalizability and reduce the risk of overfitting to specific data characteristics [234].

Our noise robustness analysis demonstrated that methods like PE maintained superior performance compared to others when subjected to varying levels of noise. This finding supports the use of PE in real-world applications where data corruption is a common concern. Conversely, methods like MCD and DE, while robust, exhibited higher computational demands, which may limit their practical applicability in scenarios with limited resources [220]. This is consistent with studies suggesting that while certain methods provide robustness, they may come at the cost of increased computational complexity.

Implementing rejection thresholds improved performance metrics, particularly sensitivity and specificity [235, 236]. This approach allows for high-confidence predictions while avoiding uncertain cases, thereby potentially reducing error rates. However, it also presents a trade-off between rejecting too many instances and missing true positives. The conservative strategies of methods like BE and PE, which reject fewer AF cases, indicate a focus on minimizing false negatives: a critical aspect in medical diagnostics where missing a true positive can have serious consequences.

In terms of computational efficiency, PE and BE emerged as the most balanced methods, offering both high performance and low computational overhead. PE, in particular, stood out for its minimal inference time and parameter count, making it highly suitable for scenarios with stringent resource constraints. This is particularly relevant given the practical constraints of deploying UQ methods in clinical settings where real-time performance is crucial.

Regarding the additional study, the EDL model presents substantial advantages for the detection of AF from Holter ECG recordings. In comparison to traditional softmax-based models, the EDL model demonstrates superior recall, enhanced discriminative power, and improved calibration. By providing well-calibrated confidence estimates, the EDL model facilitates more reliable clinical decision-making, thereby minimizing the occurrence of both false positives and false negatives. These advancements position the EDL model as a promising tool for accurate AF detection and diagnosis. Future research will focus on validating these findings within larger patient populations and investigating additional clinical applications, particularly in the context of out-of-distribution datasets.

## 4.8 Conclusion

In this study, we made a contribution to the field of AF detection through a comprehensive investigation of UQ methods. Firstly, we examine 11 distinct UQ techniques specifically tailored for AF detection using Holter recording data, thereby expanding the analytical toolkit available to healthcare professionals and researchers. Secondly, we conduct a rigorous evaluation of these UQ methods by introducing random Gaussian noise into the data, to evaluate the impact on UQ methods. This approach not only assesses the robustness of the methods but also underscores their practical applicability in noisy environments. Thirdly, by analyzing the performance of the UQ techniques across various rejection thresholds, we provide valuable insights into their reliability and robustness. Lastly, we used the EDL model to quantify uncertainty and compare its performance with that of our RTA model. Overall, these detailed assessments aid in understanding the strengths and limitations of each method, facilitating more informed decision-making in clinical settings.

# Chapter V

# The Impact of Label Noise on Deep Learning Models for Atrial Fibrillation Detection from Holter Recordings

## 5.1 Introduction

The effectiveness of DL models in detecting AF is largely dependent on the availability of large, accurately labeled datasets. However, the manual labeling of ECG data is a labor-intensive process that is prone to human error, even among experts. The ambiguous characteristics of the signals, along with their overlap with other arrhythmias, further complicate the annotation process, often leading to label noise [237, 238]. Additionally, Holter recordings, which capture long-term cardiac activity, introduce further variability that challenges accurate labeling.

A common strategy to mitigate these challenges involves employing multiple annotators to review the same samples in order to establish a consensus label. While this approach effectively reduces subjective bias, it is resource-intensive and impractical for large-scale training datasets, typically being reserved for test sets [238].

DL models are particularly vulnerable to overfitting when trained on noisy

data, which can result in significant performance degradation [239]. Research has demonstrated that even small amounts of label noise can adversely affect the performance of DL models [240]. While certain architectures, such as ResNet, exhibit some resilience to label noise [241], the development of innovative algorithms capable of tolerating and managing label noise remains essential.

These challenges underscore the necessity for noise-robust DL architectures that can accommodate label imperfections while preserving high diagnostic accuracy. In this study, we simulate the effects of both random and class-dependent label noise in the training data to replicate real-world annotation errors arising from human fatigue and bias. The core contributions of this study are as follows:

- Analyzing the impact of different levels of label noise on the performance of DL models for AF detection.
- Comparing the effectiveness of various techniques to identify those that are most resilient to label noise.
- Evaluating the performance of noisy label handling techniques using two external datasets for AF detection from Holter recordings.

## 5.2 Related Work

Various methods have been developed to tackle label noise, including semi-supervised learning approaches, where models leverage both labeled and unlabeled data, selectively utilizing the labeled data [242], and techniques designed to identify and filter noisy labels from the training set [243]. However, discarding noisy data can risk losing critical information, particularly in the medical field, where mislabeled instances may still contain essential patterns. Label noise can propagate through models, leading to erroneous clinical predictions. Despite its significant consequences, research addressing label noise in biomedical data classification—especially AF detection—remains sparse.

Zhang *et al.* [244] highlighted the vulnerability of DL models to label noise, noting that at high noise levels, these models often memorize incorrect labels, a phenomenon known as "overfitting to noise". This issue arises due to the capacity of DL models to fit both correct and incorrect data.

Nonetheless, the authors also pointed out that regularization techniques, such as dropout and weight decay, can help alleviate some negative effects of label noise. Similarly, Pasolli *et al.* [245] employed a genetic algorithm for optimal subset selection, aiming to remove data outside the optimal subset. While effective, their approach only managed label noise up to 20%. Another study by Li *et al.* [246] applied multiple machine learning classifiers to detect mislabeled samples, achieving comparable accuracy to clean datasets when label noise remained below 20%. However, these studies primarily focus on identifying noisy labels rather than exploring how noise impacts model performance.

In contrast, Liu *et al.* [247] proposed a data-cleaning method that combines a bootstrapped hard loss function to improve classification accuracy by minimizing the influence of mislabeled data. They experimented with tuning parameters for the loss function and identified an optimal epoch for data cleaning. Tested on the MIT-BIH-AD using a 1-D CNN, their method effectively mitigated the impact of noisy labels, even at noise levels reaching 50%.

Accurate labeling is critical for training reliable DL models for AF detection. However, label noise remains a substantial challenge, particularly in medical applications. While prior research has primarily focused on beat classification in the MIT-BIH-AD, the effects of noisy labels on rhythm classification—critical for AF detection—are less explored. This gap underscores the necessity for developing robust methods tailored to managing noisy labels in AF detection, especially for classification tasks involving 10-second ECG segments.

## 5.3 Methodology

### 5.3.1 Model

To evaluate the effect of label noise on the performance of our DL model, we utilized the RTA-DL model, designed to process 10-second ECG segments using two leads as input. This model was selected based on previous research, which conducted a comprehensive comparison of state-of-the-art DL models for AF detection. For detailed information on the architecture and development of the RTA-DL model, refer to Chapter II. In this study,

we used the same hyperparameters as used in the original RTA model to ensure consistency in our analysis.

### 5.3.2 Artificial Label Noise

This study investigates the impact of two common forms of label noise that often complicate manual annotation and affect classification performance.

The first type is **random label noise**, which simulates errors arising from factors like annotator fatigue or external distractions. This noise uniformly affects all classes, reflecting the conditions under which the annotator worked rather than the data's inherent properties. To replicate random label noise, we randomly selected a specified percentage of labels and changed them to other classes. For example, we converted "Non-AF" labels to "AF/AFL", "AF" labels to "Non-AF/AFL", and "AFL" labels to "Non-AF/AF". In binary classification, we inverted "Non-AF" labels to "AF" and vice versa. To assess the influence of varying noise levels, we introduced increments of noise at 10%, 20%, 30%, 40%, 50%, and 60% of the labels, thereby creating multiple variations of the training dataset with increasing noise levels. For binary classification, we limited the maximum noise level to 40%. Notably, the test set remained unaltered, ensuring that all modifications were confined to the training set. Although we cannot definitively confirm that the error rate on the test set is 0%, we reasonably assume a very low error rate due to the high level of curation and quality control applied to this dataset.

The second type, **class-dependent noise**, reflects the biases that can occur during human annotation, where certain classes may be misclassified more frequently due to subjective influences. To model this bias, we systematically flipped labels within a specific class while leaving the other class unchanged. We then repeated this process for the opposite class to determine if the impact varied based on the class subjected to noise. Similar to the random noise approach, we examined noise levels up to 60% for the three-class classification problem and 40% for the binary classification scenario, simulating varying degrees of class-specific mislabeling to evaluate its effect on overall model performance.

At each designated noise level, we trained the DL model for a fixed number of epochs. After each epoch, we evaluated the model on a separate validation set with clean labels, selecting the best-performing model based on

validation accuracy for subsequent testing. In addition to our DL model trained with focal cross-entropy loss, we implemented several techniques to mitigate the effects of noisy labels. The following sections provide a comprehensive description of each methodology used in this study.

### 5.3.3 Techniques for Noisy Label Handling

There are numerous techniques available for handling noisy labels in DL. In this study, we focus on several methods that are explained in more detail below.

#### Label-Smoothing

Label smoothing is a regularization technique designed to mitigate overconfidence in deep learning models by modifying the one-hot encoded target distributions [248]. Rather than assigning a probability of 1 to the correct class and 0 to all other classes, label smoothing distributes a small amount of probability mass across the incorrect classes. Formally, for a true class label $y$ for an instance, the one-hot encoded target vector is represented as:

$$\mathbf{y} = [0, 0, \ldots, 1, \ldots, 0, 0] \tag{5.1}$$

With label smoothing, this target distribution is adjusted to:

$$\mathbf{y}_{\text{smooth}} = \left[\frac{\epsilon}{K-1}, \ldots, 1-\epsilon, \ldots, \frac{\epsilon}{K-1}\right] \tag{5.2}$$

where $\epsilon$ is a small constant (typically between 0.1 and 0.2) and $K$ is the number of classes. This adjustment prevents the model from becoming excessively confident in its predictions, thereby reducing the likelihood of overfitting, especially in the presence of noisy labels. In our study, we set $\epsilon$=0.2.

#### Bootstrap

We implemented a loss correction technique based on the method proposed by Reed *et al.* [249]. This approach modifies the cross-entropy loss function to account for noisy labels by incorporating the model's predicted probabilities. Specifically, we employed a bootstrapping technique that combines the original labels with the model's predictions to reduce the impact of noisy

labels over time. In this method, the loss function is adjusted as follows:

$$L_{\text{boot}} = -\sum_{i=1}^{N} [\lambda\, y_i \log p_i + (1-\lambda)\, \hat{y}_i \log p_i] \tag{5.3}$$

where $N$ is the number of samples, $p_i$ is the predicted probability for the true class of sample $i$, $\hat{y}_i$ is the model's prediction for sample $i$ after the initial training phase, $\lambda$ is a weight parameter that controls the balance between the original label and the model's prediction. The parameter $\lambda$ typically ranges from 0 to 1. When $\lambda$ is close to 1, the loss function heavily relies on the original labels, whereas a smaller $\lambda$ places more emphasis on the model's predictions. This technique iteratively refines the model by progressively correcting the influence of noisy labels, possibly leading to improved robustness and accuracy in the presence of label noise. In this study, we set $\lambda = 0.8$.

**Huber Loss**

Unlike the traditional cross-entropy loss, which can be sensitive to outliers and noisy labels, Huber loss provides a balance between MSE and mean absolute error (MAE) and is designed to be less sensitive to label noise by reducing the influence of incorrect labels during training [250, 251].

$$L_\delta(y, \hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & \text{if } |y-\hat{y}| \le \delta, \\ \delta \cdot \left(|y-\hat{y}| - \frac{1}{2}\delta\right) & \text{otherwise,} \end{cases} \tag{5.4}$$

where $\delta$ is a threshold parameter that determines the transition between quadratic and linear behavior. For small errors (*i.e.*, when the absolute difference between the true label and the predicted value is less than $\delta$), Huber loss behaves like mean squared error (quadratic loss). However, for larger errors (*i.e.*, when the error exceeds $\delta$), it behaves like mean absolute error (linear loss). This combination provides the benefit of smooth gradients for small errors, while limiting the impact of larger errors, which are often the result of noisy labels or outliers. In our study, we set $\delta = 0.8$.

Even though the Huber loss may seem appropriate to regression tasks only, in scenarios where labels may not accurately reflect the true class (noisy labels), it reduces the effect of large discrepancies, thus preventing the model from being overly influenced by potentially incorrect labels.

### Mixup

Mixup is a DA strategy that generates new training examples by interpolating both the input data and their corresponding labels, thereby smoothing decision boundaries and mitigating overfitting [252].

Given two randomly selected training samples $(\mathbf{x_i}, y_i)$ and $(\mathbf{x_j}, y_j)$, Mixup constructs a synthetic training sample $(\tilde{x}, \tilde{y})$ using a convex combination of the two original samples. This is formalized as:

$$\tilde{x} = \lambda \mathbf{x_i} + (1 - \lambda)\mathbf{x_j} \tag{5.5}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{5.6}$$

where $\lambda \in [0, 1]$ is a mixing coefficient sampled from a Beta distribution $Beta(\alpha, \alpha)$, with $\alpha$ as a hyperparameter controlling the degree of interpolation. The label $\tilde{y}$ is similarly a linear interpolation of the original labels, resulting in a soft label for the new synthetic example. In this study, we set $\alpha = 0.2$.

### Forward Loss Correction (FLC)

Forward Loss Correction (FLC) is a technique designed to modify the standard loss function to account for label noise, thereby enhancing the robustness of the model. This adjustment aligns the training process with the underlying true distribution of the labels, as demonstrated by Patrini et al. (2017) [253].

Let $\check{y} \in \{1, \ldots, C\}$ represent the observed, potentially noisy label, where $C$ denotes the number of classes. The relationship between the true label and the noisy label is modeled through a *noise transition matrix* $T \in \mathbb{R}^{C \times C}$. The entry $T_{ij}$ of the matrix represents the probability of observing label $\check{y} = j$ given that the true label is $y = i$:

$$T_{ij} = P(\check{y} = j \mid y = i) \tag{5.7}$$

Under this formulation, the distribution of noisy labels $\check{y}$ can be expressed as a transformation of the true label distribution using the noise transition matrix $T$:

$$P(\check{y} \mid \mathbf{x}) = T^{\top} P(y \mid \mathbf{x}), \tag{5.8}$$

where $P(y \mid x)$ denotes the true class probability given input $x$. This relationship allows us to represent the noisy label distribution as a product of the true label distribution and the noise transition matrix $T$.

In this study, the noise transition matrix $T$ is defined as follows:

$$T = \begin{bmatrix} 0.9 & 0.1 & 0.1 \\ 0.1 & 0.9 & 0.1 \\ 0.1 & 0.1 & 0.9 \end{bmatrix} \tag{5.9}$$

In Equation 5.9, the diagonal entries indicate a probability of 0.9 for correctly identifying the true class, while the off-diagonal entries indicate a probability of 0.1 for misclassification among the other classes. This configuration assumes that the model is generally reliable but also acknowledges the presence of label noise.

To correct the influence of noisy labels during training, we adjust the standard loss function using the noise transition matrix $T$. Let $L(\theta)$ denote the loss function (e.g., cross-entropy loss) used to train DL model. The corrected predicted probabilities $\hat{y}$ are calculated as follows:

$$\hat{y} = y_{\text{pred}} T, \tag{5.10}$$

where $y_{\text{pred}} \in \mathbb{R}^{N \times C}$ represents the predicted probabilities from the model for each class $c$. The multiplication effectively adjusts the predicted probabilities by incorporating the noise characteristics captured in $T$. Notably, this calculation can also be expressed equivalently in terms of the label distribution:

$$P(\check{y} \mid \mathbf{x}) = \hat{y}, \tag{5.11}$$

indicating that the corrected predicted probabilities and the distribution of noisy labels are the same. The modified loss function incorporating FLC is then expressed as:

$$L_{\text{FLC}}(\theta) = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \tag{5.12}$$

where $y_c$ represents the true class labels, and $\hat{y}_c$ represents the adjusted predictions after accounting for noise.

**Knowledge Distillation**

Knowledge distillation (KD) is a model compression and training technique that transfers the knowledge from a large, complex model (the "teacher") to a smaller, more efficient model (the "student"). KD has shown effectiveness in improving robustness against noisy labels by leveraging soft predictions from the teacher model, which smooth out noisy or incorrect labels [254].

In KD, the teacher model is first trained on the dataset, including noisy or mislabeled examples. Once trained, the teacher generates a set of soft predictions, which represent class probabilities rather than hard labels. These soft labels are used to train the student model, which minimizes the following objective:

$$L_{KD} = (1 - \alpha) \cdot L_{CE}(y, z_s) + \alpha \cdot \tau^2 \cdot L_{KL}(z_t/\tau, z_s/\tau) \tag{5.13}$$

where $L_{CE}$ is the cross-entropy loss between the $y$ and the student model's output $z_s$, and $L_{KL}$ is the Kullback-Leibler divergence between the teacher's softened output $z_t$ and the student's output $z_s$. The parameter $\tau$ (temperature) controls the smoothness of the teacher's probabilities, while $\alpha$ balances the contribution of the true labels and the distilled knowledge. In this study, we used our DL model as the teacher model. For the student model, we removed the last RTA block from our DL model. We set $\alpha = 0.5$ and $\tau = 2$.

## 5.4 Results

We assess the performance of various techniques under two types of label noise: random label noise and class-dependent label noise, applied to both binary and ternary classification tasks. The following subsections provide a detailed evaluation of the results.

### 5.4.1 Performance for Non-AF, AF and AFL Under Different Noise Levels and Techniques

Table 5.1 presents the classification performance of various methods for detecting three classes—Non-AF, AF, and AFL—on our test set under different levels of random label noise. The noise levels range from 0% to 60%, with the performance metrics displayed for each method across these increments.

At 0% noise, our DL model achieved high accuracy across all classes, particularly with Non-AF (95.1%) and AF (93.1%). As the noise level increased to 10%, 20%, and beyond, the performance of most methods varied, indicating the models' sensitivity to label noise. For instance, Label Smoothing performed well at low noise levels, particularly with AF, where it reached an accuracy of 96.0% at 0% noise and maintained competitive performance up to 20% noise. However, its effectiveness diminished at higher noise levels, particularly for AFL, dropping to 62.1% accuracy at 20% noise.

In contrast, the Mixup method consistently demonstrated resilience against increasing noise levels, maintaining strong performance across all classes, especially at higher noise levels, where it achieved an accuracy of 83.4% for AFL at 60% noise. The Bootstrap method exhibited a similar trend, particularly excelling in the Non-AF and AF categories, where it achieved an accuracy of 96.1% at 0% noise and remained robust across varying noise levels.

Overall, the results indicate that while all methods experienced a decline in performance as noise levels increased, some methods, such as Mixup and Bootstrap, were more effective at mitigating the impacts of label noise compared to others.

TABLE 5.1: Recall for Non-AF, AF, and AFL on our test set with varying levels of random label noise (NL=0% to NL=60%).

(a) NL=0% to NL=30%

| Methods | NL=0% | | | NL=10% | | | NL=20% | | | NL=30% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Our DL Model | 0.951 | 0.931 | 0.812 | 0.945 | 0.954 | 0.785 | 0.954 | 0.956 | 0.792 | 0.902 | 0.932 | 0.783 |
| Huber Loss | 0.931 | 0.961 | 0.781 | 0.939 | 0.945 | 0.785 | 0.947 | 0.939 | 0.784 | 0.932 | 0.941 | 0.780 |
| Label Smoothing | 0.941 | 0.960 | 0.621 | 0.970 | 0.941 | 0.642 | 0.922 | 0.950 | 0.707 | 0.971 | 0.920 | 0.721 |
| Mixup | 0.960 | 0.921 | 0.831 | 0.951 | 0.929 | 0.820 | 0.950 | 0.923 | 0.821 | 0.952 | 0.941 | 0.801 |
| Bootstrap | 0.941 | 0.961 | 0.741 | 0.931 | 0.948 | 0.781 | 0.925 | 0.935 | 0.781 | 0.948 | 0.956 | 0.800 |
| FLC | 0.931 | 0.960 | 0.820 | 0.930 | 0.950 | 0.791 | 0.941 | 0.951 | 0.781 | 0.940 | 0.950 | 0.741 |
| KD | 0.932 | 0.952 | 0.788 | 0.944 | 0.951 | 0.783 | 0.942 | 0.956 | 0.790 | 0.942 | 0.933 | 0.781 |

(b) NL=40% to NL=60%

| Methods | NL=40% | | | NL=50% | | | NL=60% | | |
|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Our DL Model | 0.951 | 0.944 | 0.803 | 0.915 | 0.924 | 0.821 | 0.912 | 0.920 | 0.868 |
| Huber Loss | 0.932 | 0.944 | 0.763 | 0.925 | 0.913 | 0.804 | 0.941 | 0.901 | 0.869 |
| Label Smoothing | 0.841 | 0.951 | 0.702 | 0.930 | 0.912 | 0.735 | 0.929 | 0.901 | 0.743 |
| Mixup | 0.942 | 0.944 | 0.823 | 0.931 | 0.944 | 0.819 | 0.931 | 0.941 | 0.834 |
| Bootstrap | 0.931 | 0.941 | 0.744 | 0.921 | 0.902 | 0.860 | 0.834 | 0.881 | 0.921 |
| FLC | 0.941 | 0.952 | 0.741 | 0.931 | 0.940 | 0.734 | 0.931 | 0.942 | 0.742 |
| KD | 0.924 | 0.931 | 0.779 | 0.945 | 0.922 | 0.780 | 0.941 | 0.932 | 0.767 |

Table 5.2 presents the classification performance of various methods under different levels of class-dependent label noise ranging from 0% to 60%. Each method's performance is measured using accuracy scores for each class, indicating how well the model predicts the correct labels amidst increasing noise levels. Our DL model shows strong performance across all noise levels, with the highest accuracy of 0.951 for Non-AF at 0% noise and a drop to 0.871 at 60% noise. Other techniques, such as Huber Loss and Label Smoothing, also demonstrate competitive accuracy, particularly at lower noise levels. For instance, Label Smoothing achieves its best accuracy of 0.971 for Non-AF at 10% noise but experiences a decline at higher noise levels, reflecting the challenges posed by label noise. The Mixup technique and Bootstrap methods show varied performance, with the Mixup technique maintaining relatively high accuracy across most noise levels for AF and AFL classes. FLC and KD also yield comparable results, particularly in the Non-AF and AF classes, indicating that these methods can effectively mitigate the impact of label noise to some extent. Overall, the table highlights how different methodologies influence classification performance in the presence of label noise, providing valuable insights into their robustness and effectiveness in practical scenarios.

TABLE 5.2: Recall for Non-AF, AF, and AFL on our test set with varying levels of class-dependent label noise (NL=0% to NL=60%).

(a) NL=0% to NL=30%

| Methods | NL=0% | | | NL=10% | | | NL=20% | | | NL=30% | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Our DL Model | 0.951 | 0.931 | 0.812 | 0.945 | 0.954 | 0.785 | 0.954 | 0.956 | 0.792 | 0.902 | 0.932 | 0.783 |
| Huber Loss | 0.931 | 0.961 | 0.781 | 0.935 | 0.954 | 0.785 | 0.934 | 0.945 | 0.791 | 0.912 | 0.922 | 0.784 |
| Label Smoothing | 0.941 | 0.960 | 0.621 | 0.971 | 0.941 | 0.645 | 0.922 | 0.950 | 0.687 | 0.983 | 0.910 | 0.731 |
| Mixup | 0.960 | 0.921 | 0.831 | 0.951 | 0.921 | 0.780 | 0.951 | 0.942 | 0.741 | 0.950 | 0.941 | 0.780 |
| Bootstrap | 0.941 | 0.961 | 0.741 | 0.941 | 0.957 | 0.750 | 0.945 | 0.931 | 0.756 | 0.948 | 0.956 | 0.752 |
| FLC | 0.931 | 0.960 | 0.820 | 0.931 | 0.951 | 0.791 | 0.940 | 0.951 | 0.780 | 0.941 | 0.952 | 0.741 |
| KD | 0.932 | 0.952 | 0.788 | 0.941 | 0.942 | 0.791 | 0.942 | 0.951 | 0.783 | 0.941 | 0.951 | 0.739 |

(b) NL=40% to NL=60%

| Methods | NL=40% | | | NL=50% | | | NL=60% | | |
|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | AFL | Non-AF | AF | AFL | Non-AF | AF | AFL |
| Our DL Model | 0.921 | 0.934 | 0.803 | 0.925 | 0.923 | 0.801 | 0.921 | 0.923 | 0.871 |
| Huber Loss | 0.911 | 0.924 | 0.801 | 0.922 | 0.932 | 0.801 | 0.891 | 0.892 | 0.860 |
| Label Smoothing | 0.840 | 0.951 | 0.701 | 0.935 | 0.925 | 0.735 | 0.934 | 0.921 | 0.740 |
| Mixup | 0.943 | 0.945 | 0.763 | 0.933 | 0.954 | 0.789 | 0.920 | 0.951 | 0.841 |
| Bootstrap | 0.931 | 0.942 | 0.745 | 0.961 | 0.901 | 0.780 | 0.961 | 0.891 | 0.781 |
| FLC | 0.942 | 0.951 | 0.740 | 0.921 | 0.953 | 0.723 | 0.951 | 0.962 | 0.708 |
| KD | 0.941 | 0.952 | 0.732 | 0.920 | 0.951 | 0.721 | 0.939 | 0.943 | 0.721 |

### 5.4.2 Performance for Non-AF and AF

Table 5.3 presents the classification performance of different methods for detecting Non-AF and AF across different levels of random label noise, ranging from 0% to 40%. At 0% noise, Mixup achieved the highest accuracy for Non-AF (0.951), but its performance for AF (0.895) was lower compared to other methods like Label-Smoothing (0.929). As label noise increased, the performance of most methods showed some degradation, particularly at 40% noise. For instance, Mixup and Bootstrap saw drops in AF classification accuracy at higher noise levels (0.87 and 0.87, respectively). In contrast, Label-Smoothing maintained relatively stable performance, achieving strong results even at 40% noise (Non-AF: 0.921, AF: 0.930), indicating its robustness to label noise. Overall, while all methods are affected by increasing noise, Label-Smoothing and Huber Loss demonstrate better resilience, particularly in the AF classification task.

TABLE 5.3: Recall for Non-AF and AF on our test set with varying levels of random label noise.

| Methods | NL=0% | | NL=10% | | NL=20% | | NL=30% | | NL=40% | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | Non-AF | AF | Non-AF | AF | Non-AF | AF | Non-AF | AF |
| Our DL Model | 0.915 | 0.928 | 0.921 | 0.914 | 0.913 | 0.925 | 0.918 | 0.924 | 0.949 | 0.901 |
| Huber Loss | 0.925 | 0.922 | 0.919 | 0.904 | 0.909 | 0.915 | 0.920 | 0.912 | 0.909 | 0.911 |
| Label-Smoothing | 0.925 | 0.929 | 0.924 | 0.904 | 0.912 | 0.914 | 0.920 | 0.932 | 0.921 | 0.930 |
| Mixup | 0.951 | 0.895 | 0.932 | 0.914 | 0.913 | 0.905 | 0.918 | 0.891 | 0.901 | 0.871 |
| Bootstrap | 0.931 | 0.895 | 0.916 | 0.923 | 0.913 | 0.915 | 0.929 | 0.895 | 0.932 | 0.870 |
| FLC | 0.932 | 0.898 | 0.920 | 0.921 | 0.911 | 0.908 | 0.918 | 0.895 | 0.909 | 0.901 |
| KD | 0.912 | 0.918 | 0.911 | 0.921 | 0.913 | 0.918 | 0.914 | 0.910 | 0.917 | 0.901 |

The results presented in Table 5.4 compare the classification performance of different methods under varying levels of class-dependent label noise, specifically for distinguishing between Non-AF and AF. At 0% noise, the methods generally exhibit strong performance, with most achieving accuracy levels above 0.92 for both Non-AF and AF classifications. Label-Smoothing and Huber Loss yield the highest performance at this noise-free level. As the noise level increases, most methods maintain stable performance up to 20% noise. However, at higher noise levels (30% and 40%), performance begins to degrade more noticeably, particularly for methods like Mixup and Bootstrap, which show a significant drop in AF classification accuracy at 40% noise. Huber Loss and Label-Smoothing demonstrate greater resilience to noise, maintaining relatively consistent performance even at higher noise levels. Conversely, FLC and Mixup techniques suffer from more pronounced decreases in AF classification accuracy as noise

increases. These results indicate that while some methods are robust to moderate noise, their efficacy varies significantly under severe class-dependent label noise conditions.

TABLE 5.4: Recall for Non-AF and AF on our test set with varying levels of class-dependent label noise

| Methods | 0% | | 10% | | 20% | | 30% | | 40% | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Non-AF | AF | Non-AF | AF | Non-AF | AF | Non-AF | AF | Non-AF | AF |
| Our DL Model | 0.915 | 0.928 | 0.939 | 0.923 | 0.930 | 0.934 | 0.942 | 0.921 | 0.912 | 0.932 |
| Huber Loss | 0.939 | 0.921 | 0.940 | 0.920 | 0.932 | 0.934 | 0.942 | 0.921 | 0.941 | 0.921 |
| Label-Smoothing | 0.944 | 0.922 | 0.938 | 0.923 | 0.929 | 0.934 | 0.942 | 0.921 | 0.921 | 0.934 |
| Mixup | 0.940 | 0.932 | 0.933 | 0.929 | 0.935 | 0.914 | 0.932 | 0.920 | 0.901 | 0.871 |
| Bootstrap | 0.921 | 0.933 | 0.923 | 0.920 | 0.915 | 0.933 | 0.912 | 0.901 | 0.911 | 0.871 |
| FLC | 0.921 | 0.921 | 0.913 | 0.903 | 0.915 | 0.914 | 0.922 | 0.911 | 0.921 | 0.901 |
| KD | 0.935 | 0.932 | 0.913 | 0.922 | 0.915 | 0.914 | 0.922 | 0.904 | 0.941 | 0.911 |

### 5.4.3 Performance on External Test Sets

We assess the performance of each method at two distinct noise levels: 0% and 40% label noise, using two external datasets, IRIDIA-AF and SHDB-AF, as test sets for classifying Non-AF and AF cases. The model was trained on our dataset, and the resulting performance metrics for each method are summarized in the table below.

TABLE 5.5: Performance on the IRIDIA-AF dataset for Non-AF and AF classification. Values in parentheses indicate results with class-dependent label noise.

| **Methods** | NL=0% | | NL=40% | |
|---|---|---|---|---|
| | Recall | | Recall | |
| | Non-AF | AF | Non-AF | AF |
| Our DL Model | 0.942 | 0.932 | 0.951 (0.941) | 0.882 (0.922) |
| Huber Loss | 0.911 | 0.932 | 0.972 (0.941) | 0.843 (0.921) |
| Label-Smoothing | 0.912 | 0.923 | 0.921 (0.931) | 0.853 (0.911) |
| Mixup | 0.961 | 0.920 | 0.912 (0.952) | 0.903 (0.886) |
| Bootstrap | 0.912 | 0.922 | 0.909 (0.911) | 0.903 (0.921) |
| FLC | 0.902 | 0.922 | 0.912 (0.932) | 0.902 (0.908) |
| KD | 0.932 | 0.921 | 0.943 (0.932) | 0.90 (0.922) |

Table 5.5 presents the recall performance of various methods for classifying Non-AF and AF on the IRIDIA-AF dataset, evaluated at two noise levels: 0% and 40% label noise. For the noise-free scenario (0% label noise), Mixup achieves the highest recall for Non-AF (0.961), while our DL model exhibits

strong performance for AF classification (0.932). At 40% label noise, the results vary, with Mixup maintaining robust recall for AF (0.903), while our DL model achieves the highest Non-AF recall (0.951). Notably, under class-dependent label noise (values in parentheses), most methods show slight performance degradation, though Mixup and our DL model demonstrate resilience, particularly for Non-AF classification. Among noise-handling techniques, Huber Loss and FLC perform comparably well, though they show a decline in AF recall. Overall, Mixup and our DL model emerge as the most effective across both noise levels.

TABLE 5.6: Performance on the SHDB-AF dataset for Non-AF and AF Classification. Values in parentheses indicate results with class-independent label noise.

| **Methods** | NL=0% | | NL=40% | |
|---|---|---|---|---|
| | Recall | | Recall | |
| | Non-AF | AF | Non-AF | AF |
| Our DL Model | 0.925 | 0.924 | 0.961 (0.962) | 0.813 (0.913) |
| Huber Loss | 0.946 | 0.945 | 0.982 (0.940) | 0.861 (0.941) |
| Label-Smoothing | 0.943 | 0.953 | 0.953 (0.952) | 0.851 (0.939) |
| Mixup | 0.951 | 0.922 | 0.931 (0.982) | 0.921 (0.892) |
| Bootstrap | 0.931 | 0.921 | 0.935 (0.942) | 0.861 (0.941) |
| FLC | 0.925 | 0.901 | 0.922 (0.932) | 0.871 (0.927) |
| KD | 0.935 | 0.934 | 0.931 (0.953) | 0.891 (0.931) |

Table 5.6 presents the performance of various methods on the SHDB-AF dataset for Non-AF and AF classification. At 0% noise, all methods demonstrate strong recall performance for both Non-AF and AF classes, with Huber Loss and Mixup achieving the highest recalls. At 40% noise, there is a noticeable decline in recall for most methods, especially for the AF class, reflecting the challenge of handling label noise. However, Huber Loss, Mixup, and KD show relatively better robustness, particularly with class-independent label noise (indicated in parentheses). Notably, our DL model shows a drop in AF recall under 40% noise but performs better when class-independent noise is applied. Mixup and Huber Loss exhibit the most consistent performance across noise levels, suggesting their effectiveness in mitigating the impact of label noise on classification tasks.

## 5.5 Discussion

In this study, we observed that the performance of DL models for AF detection did not significantly degrade, even under high levels of label noise as high as 40%. This result may seem counterintuitive, as it is commonly expected that DL models suffer from substantial performance loss when trained on noisy labels. However, this phenomenon can be rationalized through several theoretical and empirical insights from existing literature.

First, DL models, particularly CNNs and architectures like ResNet, have been shown to possess a certain degree of robustness to label noise. Rolnick *et al.* [241] demonstrated that DL models can tolerate surprisingly high levels of label noise, particularly in cases where the noise is random or symmetric across classes. In such scenarios, the model learns to generalize the underlying patterns in the data while disregarding mislabeled instances, especially when trained with sufficient data. This could explain why the model's performance in AF detection remains relatively stable, even at high noise levels. Li *et al.* [246] used five different ML classifiers to improve the reliability of the mislabeled samples identification. The authors reported that if the label noise level is not higher than 20%, the classification accuracy can be improved to the same level as there is no mislabeled sample in the training set.

Additionally, it has been observed that the early training phase of DL models predominantly focuses on learning clean and easy-to-classify samples before memorizing noisy or mislabeled data. Arpit *et al.* [255] provide evidence that DL models, during their initial epochs, tend to learn meaningful representations of the data, which enables them to build a strong predictive model, even in the presence of noise. This phenomenon, known as the "memorization effect", allows models to maintain competitive performance in noisy settings, provided that training is stopped before extensive overfitting to mislabeled data occurs.

Moreover, the class imbalance and domain characteristics in the AF detection task may further contribute to the model's resilience to noise. AF detection from ECG signals is a relatively well-defined task with distinctive features, such as irregular heart rhythms, that may be less prone to confusion with normal or other arrhythmic signals, even under noisy conditions. This domain-specific feature distinctiveness could reduce the negative

impact of noisy labels.

Furthermore, advanced techniques for mitigating the effects of label noise, such as robust loss functions and early stopping, were likely beneficial. For instance, Zhang and Sabuncu [256] showed that robust loss functions like focal loss or Huber loss can make DL models less sensitive to mislabeled data. These methods down-weight the influence of potentially noisy or hard-to-classify samples, thereby improving model robustness in high-noise environments. In our study, the use of focal cross-entropy loss may have played a critical role in preventing significant performance degradation under noisy conditions.

Lastly, the evaluation of the model on external datasets (*e.g.*, IRIDIA-AF and SHDB-AF) also suggests that noisy labels during training did not drastically impact the generalization ability of the model. This is consistent with the findings of Frenay and Verleysen [239], who argue that DL models can maintain robust generalization in real-world noisy data, as long as the noise does not dominate the entire dataset.

In summary, the observed stability in performance under high label noise can be attributed to several factors, including the inherent noise robustness of DL architectures, the memorization effect during early training, domain-specific feature clarity in AF detection, and the use of advanced noise-handling techniques. These insights are supported by various studies, reinforcing the notion that, under certain conditions, DL models can exhibit considerable resilience to label noise without suffering substantial performance degradation.

## 5.6 Conclusion

Label noise poses a significant challenge for deploying DL models within the healthcare system, especially in sensitive tasks like AF detection from ECG signals, where the risk of classifiers learning from mislabeled data is heightened. An ideal model should effectively differentiate between representative patterns and label noise. This study demonstrated that DL models, particularly those tailored for noisy label handling, outperform traditional methods, showing less performance degradation in the presence of substantial label noise. Notably, annotation bias, particularly against

minority classes, proved more detrimental than random noise. This highlights the importance of selecting robust models that can adapt to real-world scenarios where label noise is prevalent. Overall, this research provides a systematic evaluation of how various types of label noise affect model performance, offering valuable insights into model selection strategies to mitigate the challenges posed by noisy label data.

# Chapter VI

# Conclusions and Final Remarks

## 6.1 Conclusions

Accurate detection of AF is of paramount clinical significance, especially given the rising prevalence of atrial arrhythmias and the increasing adoption of wearable ECG devices that offer continuous monitoring. These devices generate vast amounts of data, highlighting the urgent need for robust and scalable AI-based approaches. This thesis has explored various methodologies to enhance the automatic detection of atrial arrhythmias within clinical continuous ECG signals. Recognizing the critical role of precise AF detection in patient care, we investigated multiple strategies aimed at developing a reliable AI-based system for the detection of atrial arrhythmias from clinical continuous ECG data.

### 6.1.1 Design and Development of Deep Learning Model for Atrial Fibrillation Detection From Holter Recordings

This study utilized a comprehensive dataset of 1,346 retrospective Holter recordings, representing one of the largest collections available for the analysis of atrial arrhythmias across a wide spectrum of cardiac conditions. Through this dataset, we successfully developed and validated a residual attention-based DL model, which demonstrated superior performance in detecting AF and AFL compared to existing state-of-the-art DL models and two rule-based software, namely ABILE and CBR. The model exhibited high recall rates across diverse demographic groups, including patients with VT and PVC, underscoring its robustness and versatility.

Our model achieved a specificity rate of 0.963, which, while slightly lower than ABILE's 0.984 and CBR's 0.999, indicates its effectiveness in correctly identifying Non-AF cases. Conversely, the sensitivity for AF/AFL reached 0.951, substantially outperforming both ABILE (0.489) and CBR (0.442). These findings highlight the model's capability in accurately predicting AF events, which is crucial for clinical decision-making.

In terms of FPR, our model recorded a value of 0.037 for AF or AFL, which is higher than ABILE's 0.016 and CBR's 0.001. Specifically for AF, the FPR was measured at 0.046, which is lower than ABILE's 0.059 but higher than CBR's 0.029. The PPV for AF or AFL was determined to be 0.898, which, although lower than CBR's 0.991, was slightly superior to ABILE's 0.915. For AF, the PPV stood at 0.651, exceeding ABILE's 0.595 yet falling short of CBR's 0.736. Notably, for AFL, our model achieved a PPV of 0.909, while CBR attained a perfect score of 1.00.

Furthermore, we assessed the performance of our model using an additional dataset of 685 Holter recordings categorized as Non-AF, based solely on the presence of PACs. Our model demonstrated a specificity of 95.4%, reflecting strong accuracy in identifying Non-AF cases, although CBR maintained a lead with a specificity of 99.0%. ABILE followed closely with a specificity of 94.1%, which was slightly below that of our model.

In summary, while our model exhibits robust performance characterized by high specificity and a moderate FPR, it also identifies critical areas for enhancement, particularly in minimizing false positives and improving PPV. The findings of this study contribute valuable insights to the ongoing discourse in AF detection and provide a foundation for future research aimed at refining predictive models for better clinical outcomes.

### 6.1.2 A Systematic Survey of Data Augmentation of ECG Signals for AI Applications

This study systematically reviewed current research on DA techniques for AI-based ECG analysis. The findings indicate that DA can significantly improve the performance of automated ECG systems; however, its effectiveness is highly dependent on the specific application.

Our experiments revealed that while DA often enhances model performance by facilitating the learning of more robust features, it is not a one-size-fits-all

solution. The success of DA varies across different tasks and diagnostic categories. Several key factors must be considered when evaluating the effectiveness of DA techniques, including the extent of desired performance improvement, the size of the available training data, and the specific diagnostic objectives. Therefore, it is essential to assess the required performance enhancement and determine whether the trade-offs associated with DA are acceptable.

Additionally, the size and balance of the training dataset are critical in influencing the impact of DA. For small or imbalanced datasets, DA can substantially improve model generalization by generating diverse training samples. Conversely, in the context of large and heterogeneous datasets, the benefits of DA may diminish. In some instances, excessive augmentation can lead to performance degradation by introducing noise or unrealistic variations into the training process.

### 6.1.3 Uncertainty Quantification of Deep Learning Model for Atrial Fibrillation Detection from Holter Recordings

While the effectiveness of DL has shown considerable promise, the susceptibility of DL models to overfitting underscores the need for robust UQ methods to facilitate safe integration into clinical practice. Despite the availability of various UQ approaches for DL models, there is a significant lack of comprehensive evaluations that systematically compare these techniques within a scalable framework. This gap is particularly evident when employing metrics that elucidate the trade-offs between performance and computational cost in the context of AF detection.

In this study, we assess eleven distinct UQ techniques using Holter recording data. To evaluate their robustness, we introduce random Gaussian noise and examine the performance of each method in noisy environments. Additionally, we explore the effectiveness of UQ across varying rejection thresholds, thereby providing valuable insights into the reliability of these techniques.

Moreover, in a related investigation, we incorporate evidence theory into our DL model, referred to as the RTA model, to quantify calibration error. Our findings indicate that the evidential DL model outperforms traditional

softmax-based DL models. This study elucidates the strengths and limitations of each UQ method, thereby supporting more informed clinical decision-making and contributing to the development of more robust tools for AF detection.

Overall, this research enhances the understanding of UQ methods in AF detection and paves the way for the development of more accurate and reliable diagnostic tools.

### 6.1.4 The Impact of Label Noise on Deep Learning Models for Atrial Fibrillation Detection from Holter Recordings

Label noise is a significant challenge in healthcare, particularly in AF detection from ECG signals, where mislabeled data can mislead classifiers and compromise diagnostic accuracy. In such scenarios, a robust DL model must be capable of distinguishing true patterns from noisy labels to ensure reliable performance. This thesis demonstrates that DL models specifically designed to handle label noise show remarkable resilience, exhibiting minimal performance degradation even in the presence of substantial noise. Notably, the model maintains stability when up to 60% of labels are altered across three cardiac conditions (non-AF, AF, and AFL) and up to 40% for two conditions (non-AF and AF). These findings underscore the critical role of DL models that can manage random label noise while also addressing systematic labeling biases. This research offers a thorough evaluation of the impact of label noise on model performance, providing valuable insights into strategies for mitigating the effects of noisy data in AF detection from Holter recordings.

## 6.2 Final Remarks

The findings of this thesis contribute to the advancement of AI-driven atrial arrhythmia detection from Holter recordings. The research highlights the critical importance of robust DL model design, data augmentation, uncertainty quantification, and effective label noise management in achieving accurate and reliable AF detection. The implications of this research extend beyond mere algorithmic improvements; they pave the way for enhanced patient care through early detection and timely intervention, which are crucial for reducing AF-related complications.

However, it is essential to acknowledge some limitations inherent in our DL-based AF detection from Holter recordings. While our model effectively predicts 10-second segments of ECG data, this design raises concerns in the context of continuous monitoring, particularly regarding the management of false alarms. In a continuous monitoring environment, even a small number of false alarms can lead to unnecessary anxiety for patients and clinicians, as well as potentially inappropriate medical responses. This raises a critical issue: how can we effectively manage false alarms in real-time monitoring scenarios? Addressing this question remains an area for further exploration. One potential approach could involve implementing a decision rule to merge predictions of consecutive 10-s segments detected as AF, allowing the system to differentiate between isolated anomalies and sustained arrhythmias.

Overall, the findings of this thesis not only demonstrate the current capabilities of AI in atrial arrhythmia detection but also establish a strong foundation for future research and model refinement, ultimately enhancing the reliability, robustness, and clinical applicability of AI systems in atrial arrhythmia detection.

# Publications

## Journal Articles

- Rahman, Md Moklesur, Massimo Walter Rivolta, Fabio Badilini, and Roberto Sassi. 2023. "A Systematic Survey of Data Augmentation of ECG Signals for AI Applications" *Sensors* 23, no. 11: 5237.
- Gavidia, Marino, Hongling Zhu, Arthur N. Montanari, Jesús Fuentes, Cheng Cheng, Sergio Dubner, Martin Chames et al. "Early warning of atrial fibrillation using deep learning". *Patterns* 5, no. 6 (2024).
- Rahman, Moklesur, Massimo Walter Rivolta, Pierre Maison-Blanche, Fabio Badilini, and Roberto Sassi. "Residual-attention deep learning model for atrial fibrillation detection from Holter recordings". *Journal of Electrocardiology* 84 (2024): 12.

## Conference Proceedings

- Rahman, Md Moklesur, Massimo Walter Rivolta, Pierre Maison-Blanche, Fabio Badilini, and Roberto Sassi. "Quantifying Uncertainty of a Deep Learning Model for Atrial Fibrillation Detection from ECG Signals". *In Computing in Cardiology (CinC)*, vol. 50, pp. 1-4. IEEE, 2023.
- Rahman, Md Moklesur, Massimo Walter Rivolta, Fabio Badilini, and Roberto Sassi. "Evidential Deep Learning Model for Atrial Fibrillation Detection from Holter Recordings". *In Computing in Cardiology (CinC)*, 2024

# Bibliography

[1] Sumeet S Chugh, Rasmus Havmoeller, Kumar Narayanan, David Singh, Michiel Rienstra, Emelia J Benjamin, Richard F Gillum, Young-Hoon Kim, John H McAnulty Jr, Zhi-Jie Zheng, et al. Worldwide epidemiology of atrial fibrillation: a global burden of disease 2010 study. *Circulation*, 129(8):837–847, 2014.

[2] Centers for Disease Control and Prevention (CDC). Atrial fibrillation fact sheet, 2020. Retrieved from https://www.cdc.gov/heartdisease/atrial_fibrillation.htm.

[3] Susan Colilla, Ann Crow, William Petkun, Daniel E Singer, Teresa Simon, and Xianchen Liu. Estimates of current and future incidence and prevalence of atrial fibrillation in the us adult population. *The American journal of cardiology*, 112(8):1142–1147, 2013.

[4] Philip A Wolf, Robert D Abbott, and William B Kannel. Atrial fibrillation as an independent risk factor for stroke: the framingham study. *stroke*, 22(8):983–988, 1991.

[5] Robert G Hart, Lesly A Pearce, and Maria I Aguilar. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. *Annals of internal medicine*, 146(12):857–867, 2007.

[6] Hugh Calkins, Gerhard Hindricks, Riccardo Cappato, Young-Hoon Kim, Eduardo B Saad, Luis Aguinaga, Joseph G Akar, Vinay Badhwar, Josep Brugada, John Camm, et al. 2017 HRS/EHRA/ECAS/APHRS/SOLAECE expert consensus statement on catheter and surgical ablation of atrial fibrillation. *Ep Europace*, 20(1):e1–e160, 2018.

[7] Michael H Kim, Stephen S Johnston, Bong-Chul Chu, Mehul R Dalal, and Kathy L Schulman. Estimation of total incremental health care

costs in patients with atrial fibrillation in the united states. *Circulation: Cardiovascular Quality and Outcomes*, 4(3):313–320, 2011.

[8] Hooman Kamel, Peter M Okin, Mitchell SV Elkind, and Costantino Iadecola. Atrial fibrillation and mechanisms of stroke: time for a new model. *Stroke*, 47(3):895–900, 2016.

[9] Ronald J Prineas, Elsayed Z Soliman, George Howard, Virginia J Howard, Mary Cushman, Zhu-Ming Zhang, and Claudia S Moy. The sensitivity of the method used to detect atrial fibrillation in population studies affects group-specific prevalence estimates: ethnic and regional distribution of atrial fibrillation in the REGARDS study. *Journal of epidemiology*, 19(4):177–181, 2009.

[10] Gerhard Hindricks, Tatjana Potpara, Nikolaos Dagres, Elena Arbelo, Jeroen J Bax, Carina Blomström-Lundqvist, Giuseppe Boriani, Manuel Castella, Gheorghe-Andrei Dan, Polychronis E Dilaveris, et al. 2020 esc guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (EACTS) the task force for the diagnosis and management of atrial fibrillation of the european society of cardiology (ESC) developed with the special contribution of the european heart rhythm association (EHRA) of the ESC. *European heart journal*, 42(5):373–498, 2021.

[11] Andrei D Margulescu and Lluis Mont. Persistent atrial fibrillation vs paroxysmal atrial fibrillation: differences in management. *Expert review of cardiovascular therapy*, 15(8):601–618, 2017.

[12] Craig T January, L Samuel Wann, Hugh Calkins, Lin Y Chen, Joaquin E Cigarroa, Joseph C Cleveland Jr, Patrick T Ellinor, Michael D Ezekowitz, Michael E Field, Karen L Furie, et al. 2019 AHA/ACC/HRS focused update of the 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society in collaboration with the society of thoracic surgeons. *Circulation*, 140(2):e125–e151, 2019.

[13] Tan Wang and Yan Qin. A novel multi-scale convolutional network with attention-based bidirectional gated recurrent unit for atrial fibrillation discrimination. *Biocybernetics and Biomedical Engineering*, 41(2):445–455, 2021.

[14] Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature communications*, 11(1):1760, 2020.

[15] Georgios Petmezas, Kostas Haris, Leandros Stefanopoulos, Vassilis Kilintzis, Andreas Tzavelis, John A Rogers, Aggelos K Katsaggelos, and Nicos Maglaveras. Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets. *Biomedical Signal Processing and Control*, 63:102194, 2021.

[16] Monika Butkuvienė, Andrius Petrėnas, Andrius Sološenko, Alba Martín-Yebra, Vaidotas Marozas, and Leif Sörnmo. Considerations on performance evaluation of atrial fibrillation detectors. *IEEE Transactions on Biomedical Engineering*, 68(11):3250–3260, 2021.

[17] Thomas Grote and Philipp Berens. Uncertainty, evidence, and the integration of machine learning into medical practice. In *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, volume 48, pages 84–97, 2023.

[18] Görkem Algan and Ilkay Ulusoy. Label noise types and their effects on deep learning. *arXiv:2003.10471*, 2020.

[19] Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

[20] Will Gersch, David M Eddy, and Eugene Dong Jr. Cardiac arrhythmia classification: A heart-beat interval-Markov chain approach. *Computers and Biomedical Research*, 3(4):385–392, 1970.

[21] HUBERT V PIPBERGER and JEROME CORNFIELD. What ecg computer program to choose for clinical application: the need for consumer protection. *Circulation*, 47(5):918–920, 1973.

[22] Spencer Z Rosero, Valentina Kutyifa, Brian Olshansky, and Wojciech Zareba. Ambulatory ECG monitoring in atrial fibrillation management. *Progress in cardiovascular diseases*, 56(2):143–152, 2013.

[23] Felix K Wegner, Lucas Plagwitz, Florian Doldi, Christian Ellermann, Kevin Willy, Julian Wolfes, Sarah Sandmann, Julian Varghese, and Lars Eckardt. Machine learning in the detection and management of atrial fibrillation. *Clinical Research in Cardiology*, 111(9):1010–1017, 2022.

[24] Shadnaz Asgari, Alireza Mehrnia, and Maryam Moussavi. Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine. *Computers in biology and medicine*, 60:132–142, 2015.

[25] Robert Czabanski, Krzysztof Horoba, Janusz Wrobel, Adam Matonia, Radek Martinek, Tomasz Kupka, Michal Jezewski, Radana Kahankova, Janusz Jezewski, and Jacek M Leski. Detection of atrial fibrillation episodes in long-term heart rhythm signals using a support vector machine. *Sensors*, 20(3):765, 2020.

[26] Nuryani Nuryani, Bambang Harjito, Iwan Yahya, and Anik Lestari. Atrial fibrillation detection using support vector machine. In *Proceedings of the Joint International Conference on Electric Vehicular Technology and Industrial, Mechanical, Electrical and Chemical Engineering (ICEVT & IMECE)*, pages 215–218, 2015.

[27] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, 2019.

[28] David Burke, James Carey, Peter Doggart, and Alan Kennedy. Novel AI algorithm improves the automated detection of atrial arrhythmias from the Apple Watch. *Heart Rhythm*, 20(5):S613–S614, 2023.

[29] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] George Moody. A new method for detecting atrial fibrillation using RR intervals. *Proceedings of the Computing in Cardiology*, 10:227–230, 1983.

[32] Simona Petrutiu, Alan V Sahakian, and Steven Swiryn. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace*, 9(7):466–470, 2007.

[33] Yifan Sun, Jingyan Shen, Yunfan Jiang, Zhaohui Huang, Minsheng Hao, and Xuegong Zhang. MMA-RNN: A multi-level multi-task attention-based recurrent neural network for discrimination and localization of atrial fibrillation. *Biomedical Signal Processing and Control*, 89:105747, 2024.

[34] Xiang Lu, Xingrui Wang, Wanying Zhang, Anhao Wen, and Yande Ren. An end-to-end model for ECG signals classification based on residual attention network. *Biomedical Signal Processing and Control*, 80:104369, 2023.

[35] Roberta Colloca, Alistair EW Johnson, Luca Mainardi, and Gari D Clifford. A support vector machine approach for reliable detection of atrial fibrillation events. In *Computing in Cardiology 2013*, pages 1047–1050, 2013.

[36] Mohit Kumar, Ram Bilas Pachori, and U Rajendra Acharya. Automated diagnosis of atrial fibrillation ECG signals using entropy features extracted from flexible analytic wavelet transform. *Biocybernetics and Biomedical Engineering*, 38(3):564–573, 2018.

[37] Alan Kennedy, Dewar D Finlay, Daniel Guldenring, Raymond R Bond, Kieran Moran, and James McLaughlin. Automated detection of atrial fibrillation using RR intervals and multivariate-based classification. *Journal of electrocardiology*, 49(6):871–876, 2016.

[38] Morteza Zabihi, Ali Bahrami Rad, Aggelos K Katsaggelos, Serkan Kiranyaz, Susanna Narkilahti, and Moncef Gabbouj. Detection of

atrial fibrillation in ECG hand-held devices using a random forest classifier. In *2017 Computing in Cardiology (CinC)*, pages 1–4, 2017.

[39] Yong Xia, Naren Wulan, Kuanquan Wang, and Henggui Zhang. Detecting atrial fibrillation by deep convolutional neural networks. *Computers in biology and medicine*, 93:84–92, 2018.

[40] Wenjuan Cai, Yundai Chen, Jun Guo, Baoshi Han, Yajun Shi, Lei Ji, Jinliang Wang, Guanglei Zhang, and Jianwen Luo. Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. *Computers in biology and medicine*, 116:103378, 2020.

[41] Xiaomao Fan, Qihang Yao, Yunpeng Cai, Fen Miao, Fangmin Sun, and Ye Li. Multiscaled fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings. *IEEE journal of biomedical and health informatics*, 22(6):1744–1753, 2018.

[42] Haotian Shi, Haoren Wang, Chengjin Qin, Liqun Zhao, and Chengliang Liu. An incremental learning system for atrial fibrillation detection based on transfer learning and active learning. *Computer methods and programs in biomedicine*, 187:105219, 2020.

[43] Bambang Tutuko, Siti Nurmaini, Alexander Edo Tondas, Muhammad Naufal Rachmatullah, Annisa Darmawahyuni, Ria Esafri, Firdaus Firdaus, and Ade Iriani Sapitri. AFibNet: an implementation of atrial fibrillation detection with convolutional neural network. *BMC Medical Informatics and Decision Making*, 21:1–17, 2021.

[44] Eedara Prabhakararao and Samarendra Dandapat. Multi-scale convolutional neural network ensemble for multi-class arrhythmia classification. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3802–3812, 2021.

[45] Vykintas Maknickas and Algirdas Maknickas. Atrial fibrillation classification using QRS complex features and LSTM. In *Computing in Cardiology (CinC)*, pages 1–4, 2017.

[46] Le Sun, Yukang Wang, Jinyuan He, Haoyuan Li, Dandan Peng, and Yilin Wang. A stacked LSTM for atrial fibrillation prediction based on multivariate ECGs. *Health information science and systems*, 8:1–7, 2020.

[47] Yong-Soo Baek, Sang-Chul Lee, Wonik Choi, and Dae-Hyeok Kim. A new deep learning algorithm of 12-lead electrocardiogram for identifying atrial fibrillation during sinus rhythm. *Scientific reports*, 11(1):12818, 2021.

[48] Jibin Wang. An intelligent computer-aided approach for atrial fibrillation and atrial flutter signals classification using modified bidirectional LSTM network. *Information Sciences*, 574:320–332, 2021.

[49] Jibin Wang. Automated detection of atrial fibrillation and atrial flutter in ECG signals based on convolutional and improved Elman neural network. *Knowledge-Based Systems*, 193:105446, 2020.

[50] Yanrui Jin, Chengjin Qin, Yixiang Huang, Wenyi Zhao, and Chengliang Liu. Multi-domain modeling of atrial fibrillation detection with twin attentional convolutional long short-term memory neural networks. *Knowledge-Based Systems*, 193:105460, 2020.

[51] Peng Zhang, Chenbin Ma, Fan Song, Yangyang Sun, Youdan Feng, Yufang He, Tianyi Zhang, and Guanglei Zhang. D2AFNet: A dual-domain attention cascade network for accurate and interpretable atrial fibrillation detection. *Biomedical Signal Processing and Control*, 82:104615, 2023.

[52] Yuxuan Zhao, Jiadong Ren, Bing Zhang, Jinxiao Wu, and Yongqiang Lyu. An explainable attention-based TCN heartbeats classification model for arrhythmia detection. *Biomedical Signal Processing and Control*, 80:104337, 2023.

[53] Yongjian Li, Liting Zhang, Lin Zhu, Lei Liu, Baokun Han, Yatao Zhang, and Shoushui Wei. Diagnosis of atrial fibrillation using self-complementary attentional convolutional neural network. *Computer Methods and Programs in Biomedicine*, 238:107565, 2023.

[54] G.B. Moody and R.G. Mark. The mit-bih arrhythmia database on cd-rom and software for use with it. In *Computers in Cardiology*, pages 185–188, 1990.

[55] Martino Vaglio, Pierre Maison-Blanche, Gianfranco Toninelli, Lamberto Isola, Francesca Ferrari, and Fabio Badilini. CER-S, an ECG platform for the management of continuous ECG recordings and

databases. In *2022 Computing in Cardiology (CinC)*, volume 498, pages 1–4, 2022.

[56] Cédric Gilon, Jean-Marie Grégoire, Marianne Mathieu, Stéphane Carlier, and Hugues Bersini. IRIDIA-AF, a large paroxysmal atrial fibrillation long-term electrocardiogram monitoring database. *Scientific data*, 10(1):714, 2023.

[57] Kenta Tsutsui, Shany Biton Brimer, Noam Ben-Moshe, Jean Marc Sellal, Julien Oster, Hitoshi Mori, Yoshifumi Ikeda, Takahide Arai, Shintaro Nakano, Ritsushi Kato, et al. SHDB-AF: a japanese holter ECG database of atrial fibrillation. *arXiv:2406.16974*, 2024.

[58] Qianxi Zhao, Liu Yang, and Nengchao Lyu. A driver stress detection model via data augmentation based on deep convolutional recurrent neural network. *Expert Systems with Applications*, 238:122056, 2024.

[59] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[60] Priya A Prasad, Jonas L Isaksen, Yumiko Abe-Jones, Jessica K Zègre-Hemsey, Claire E Sommargren, Salah S Al-Zaiti, Mary G Carey, Fabio Badilini, David Mortara, Jørgen K Kanters, et al. Ventricular tachycardia and in-hospital mortality in the intensive care unit. *Heart Rhythm O2*, 4(11):715–722, 2023.

[61] Sarah Ali Abdelaziz Ismael, Ammar Mohammed, and Hesham Hefny. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. *Artificial intelligence in medicine*, 102:101779, 2020.

[62] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[63] Mehdi Habibzadeh, Mahboobeh Jannesari, Zahra Rezaei, Hossein Baharvand, and Mehdi Totonchi. Automatic white blood cell classification using pre-trained deep learning models: Resnet and inception. In *International conference on machine vision*, volume 10696, pages 274–281, 2018.

[64] Faezeh Nejati Hatamian, Nishant Ravikumar, Sulaiman Vesal, Felix P Kemeth, Matthias Struck, and Andreas Maier. The effect of data augmentation on classification of atrial fibrillation in short single-lead ECG signals using deep neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1264–1268, 2020.

[65] Diogo Silva, Steffen Leonhardt, and Christoph Hoog Antink. Copula-based data augmentation on a deep learning architecture for cardiac sensor fusion. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2521–2532, 2020.

[66] Jiacheng Zhu, Jielin Qiu, Zhuolin Yang, Douglas Weber, Michael A Rosenberg, Emerson Liu, Bo Li, and Ding Zhao. GeoECG: Data augmentation via wasserstein geodesic perturbation for robust electrocardiogram prediction. *arXiv:2208.01220*, 2022.

[67] Han Liu, Zhengbo Zhao, Xiao Chen, Rong Yu, and Qiang She. Using the VQ-VAE to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records. *Computer Methods and Programs in Biomedicine*, 196:105639, 2020.

[68] Shuai Ma, Jianfeng Cui, Chin-Ling Chen, Xuhui Chen, and Ying Ma. An effective data enhancement method for classification of ECG arrhythmia. *Measurement*, page 111978, 2022.

[69] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Philip J Devereaux, Jos Kleijnen, and David Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10):e1–e34, 2009.

[70] Tomer Golany and Kira Radinsky. PGANs: generative adversarial networks for ECG synthesis to improve patient-specific deep ECG classification. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 557–564, 2019.

[71] Tomer Golany, Kira Radinsky, and Daniel Freedman. SimGANs: Simulator-based generative adversarial networks for ecg synthesis to improve deep ECG classification. In *International Conference on Machine Learning*, pages 3597–3606, 2020.

[72] Arash Shokouhmand and Negar Tavassolian. Fetal electrocardiogram extraction using dual-path source separation of single-channel non-invasive abdominal recordings. *IEEE Transactions on Biomedical Engineering*, 2022.

[73] Ruggero Donida Labati, Enrique Muñoz, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Deep-ECG: Convolutional neural networks for ECG biometric recognition. *Pattern Recognition Letters*, 126:78–85, Sep 2019.

[74] Alex Barros, Paulo Resque, João Almeida, Renato Mota, Helder Oliveira, Denis Rosário, and Eduardo Cerqueira. Data improvement model based on ECG biometric for user authentication and identification. *Sensors*, 20(10):2920, 2020.

[75] Genlang Chen, Yi Zhu, Zhiqing Hong, and Zhen Yang. EmotionalGAN: Generating ecg to enhance emotion state classification. In *International Conference on Artificial Intelligence and Computer Science*, pages 309–313, 2019.

[76] Patrick Thiam, Hans A Kestler, and Friedhelm Schwenker. Multimodal deep denoising convolutional autoencoders for pain intensity classification based on physiological signals. In *ICPRAM*, pages 289–296, 2020.

[77] Pandu Wicaksono, Samuel Philip, Islam Nur Alam, and Sani M Isa. Dealing with imbalanced sleep apnea data using DCGAN. *Traitement du Signal*, 39(5), 2022.

[78] Dorien Huysmans, Ivan Castro, Pascal Borzée, Aakash Patel, Tom Torfs, Bertien Buyse, Dries Testelmans, Sabine Van Huffel, and Carolina Varon. Capacitively-coupled ECG and respiration for sleep–wake prediction and risk detection in sleep apnea patients. *Sensors*, 21(19):6409, 2021.

[79] Nebras Sobahi, Abdulkadir Sengur, Ru-San Tan, and U Rajendra Acharya. Attention-based 3D CNN with residual connections for efficient ECG-based COVID-19 detection. *Computers in Biology and Medicine*, 143:105335, 2022.

[80] Ismail Shahin, Ali Bou Nassif, and Mohamed Bader Alsabek. COVID-19 electrocardiograms classification using CNN models. In *International Conference on Developments in eSystems Engineering*, pages 448–452, 2021.

[81] Talha Anwar and Seemab Zakir. Effect of image augmentation on ECG image classification using deep learning. In *International Conference on Artificial Intelligence*, pages 182–186, 2021.

[82] Mahmoud M Bassiouni, Islam Hegazy, Nouhad Rizk, El-Sayed A El-Dahshan, and Abdelbadeeh M Salem. Automated detection of covid-19 using deep learning approaches with paper-based ECG reports. *Circuits, Systems, and Signal Processing*, 41(10):5535–5577, 2022.

[83] Guoyang Liu, Xiao Han, Lan Tian, Weidong Zhou, and Hui Liu. ECG quality assessment based on hand-crafted statistics and deep-learned s-transform spectrogram features. *Computer Methods and Programs in Biomedicine*, 208:106269, 2021.

[84] Guillermo Jimenez-Perez, Alejandro Alcaine, and Oscar Camara. Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks. *Scientific reports*, 11(1):1–11, 2021.

[85] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. AF classification from a short single lead ECG recording: The physionet/computing in cardiology challenge 2017. In *Computing in Cardiology (CinC)*, pages 1–4, 2017.

[86] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[87] Ralf Bousseljot, Dieter Kreiseler, and Allard Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedizinische Technik*, 40:317–318, 1995.

[88] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter.

Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020.

[89] Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit J Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, et al. Will two do? varying dimensions in electrocardiography: the physionet/computing in cardiology challenge 2021. In *Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[90] Naoki Nonaka and Jun Seita. RandECG: Data augmentation for deep neural network based ecg classification. In *Advances in Artificial Intelligence: Selected Papers from the Annual Conference of Japanese Society of Artificial Intelligence*, pages 178–189, 2022.

[91] Hosein Hasani, Adeleh Bitarafan, and Mahdieh Soleymani Baghshah. Classification of 12-lead ECG signals with adversarial multi-source domain generalization. In *Computing in Cardiology*, pages 1–4, 2020.

[92] Naoki Nonaka and Jun Seita. Electrocardiogram classification by modified EfficientNet with data augmentation. In *Computing in Cardiology*, pages 1–4, 2020.

[93] Gengyuan Guo, Pengzhi Gao, Xiangwei Zheng, and Cun Ji. Multimodal emotion recognition using CNN-SVM with data augmentation. In *International Conference on Bioinformatics and Biomedicine*, pages 3008–3014. IEEE, 2022.

[94] Maryam Eskandari, Saman Parvaneh, Hossein Ehsani, Mindy Fain, and Nima Toosizadeh. Frailty identification using heart rate dynamics: A deep learning approach. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3409–3417, 2022.

[95] Xiaolong Xu, Haoyan Xu, Liying Wang, Yuanyuan Zhang, and Fu Xiao. Hygeia: A multilabel deep learning-based classification method for imbalanced electrocardiogram data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[96] Johnson Loh, Jianan Wen, and Tobias Gemmeke. Low-cost DNN hardware accelerator for wearable, high-quality cardiac arrythmia detection. In *International Conference on Application-specific Systems, Architectures and Processors*, pages 213–216, 2020.

[97] Yamin Liu, Hanshuang Xie, Qineng Cao, Jiayi Yan, Fan Wu, Huaiyu Zhu, and Yun Pan. Multi-label classification of multi-lead ECG based on deep 1D convolutional neural networks with residual and attention mechanism. In *Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[98] Jingna Qiu, Maximilian P Oppelt, Michael Nissen, Lars Anneken, Katharina Breininger, and Bjoern Eskofier. Improving deep learning-based cardiac abnormality detection in 12-lead ECG with data augmentation. In *International Conference of the Engineering in Medicine & Biology Society*, pages 945–949, 2022.

[99] Garrett I Cayce, Arthur C Depoian, Colleen P Bailey, and Parthasarathy Guturu. Improved neural network arrhythmia classification through integrated data augmentation. In *2022 IEEE MetroCon*, pages 1–3, 2022.

[100] Wenrui Zhang, Shijia Geng, and Shenda Hong. A simple self-supervised ECG representation learning method via manipulated temporal–spatial reverse detection. *Biomedical Signal Processing and Control*, 79:104194, 2023.

[101] Martin Zihlmann, Dmytro Perekrestenko, and Michael Tschannen. Convolutional recurrent neural networks for electrocardiogram classification. In *Computing in Cardiology (CinC)*, pages 1–4, 2017.

[102] Ran Duan, Xiaodong He, and Zhuoran Ouyang. MADNN: a multi-scale attention deep neural network for arrhythmia classification. In *Computing in Cardiology*, pages 1–4, 2020.

[103] Temesgen Mehari and Nils Strodthoff. Self-supervised representation learning from 12-lead ECG data. *Computers in Biology and Medicine*, 141:105114, 2022.

[104] Junmo An, Richard E Gregg, and Soheil Borhani. Effective data augmentation, filters, and automation techniques for automatic 12-lead ECG classification using deep residual neural networks. In *International Conference of the Engineering in Medicine & Biology Society*, pages 1283–1287, 2022.

[105] Gary M Friesen, Thomas C Jannett, Manal Afify Jadallah, Stanford L Yates, Stephen R Quint, and H Troy Nagle. A comparison of the noise

sensitivity of nine QRS detection algorithms. *IEEE Transactions on biomedical engineering*, 37(1):85–98, 1990.

[106] Edmund Do, Jack Boynton, Byung Suk Lee, and Daniel Lustgarten. Data augmentation for 12-lead ECG beat classification. *SN Computer Science*, 3(1):1–17, 2022.

[107] Mou Wang, Sylwan Rahardja, Pasi Fränti, and Susanto Rahardja. Single-lead ECG recordings modeling for end-to-end recognition of atrial fibrillation with dual-path RNN. *Biomedical Signal Processing and Control*, 79:104067, 2023.

[108] Halla Sigurthorsdottir, Jérôme Van Zaen, Ricard Delgado-Gonzalo, and Mathieu Lemay. ECG classification with a convolutional recurrent neural network. In *Computing in Cardiology*, pages 1–4, 2020.

[109] Maximilian P Oppelt, Maximilian Riehl, Felix P Kemeth, and Jan Steffan. Combining scatter transform and deep neural networks for multilabel electrocardiogram signal classification. In *Computing in Cardiology*, pages 1–4, 2020.

[110] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396, 2017.

[111] Pengyao Xu, Hui Liu, Xiaoyun Xie, Shuwang Zhou, Minglei Shu, and Yinglong Wang. Interpatient ECG arrhythmia detection by residual attention CNN. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[112] Tanvir Mahmud, Shaikh Anowarul Fattah, and Mohammad Saquib. Deeparrnet: An efficient deep CNN architecture for automatic arrhythmia detection and classification from denoised ECG beats. *IEEE Access*, 8:104788–104800, 2020.

[113] Zhaocheng Yu, Junxin Chen, Yu Liu, Yongyong Chen, Tingting Wang, Robert Nowak, and Zhihan Lv. DDCNN: A deep learning model for AF detection from a single-lead short ECG signal. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[114] Hadaate Ullah, Md Belal Bin Heyat, Hussain AlSalman, Haider Mohammed Khan, Faijan Akhtar, Abdu Gumaei, Aaman Mehdi, Abdullah Y Muaad, Md Sajjatul Islam, Arif Ali, et al. An effective and lightweight deep electrocardiography arrhythmia recognition model using novel special and native structural regularization techniques on cardiac signal. *Journal of Healthcare Engineering*, 2022, 2022.

[115] Shenghua Liu, Bin Zhou, Quan Ding, Bryan Hooi, Zheng bo Zhang, Huawei Shen, and Xueqi Cheng. Time series anomaly detection with adversarial reconstruction networks. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[116] D Sangeetha, S Selvi, and M Siva Anandha Ram. A CNN based similarity learning for cardiac arrhythmia prediction. In *International Conference on Advanced Computing*, pages 244–248. IEEE, 2019.

[117] Sebastian D Goodfellow, Dmitrii Shubin, Robert W Greer, Sujay Nagaraj, Carson McLean, Will Dixon, Andrew J Goodwin, Azadeh Assadi, Anusha Jegatheeswaran, Peter C Laussen, et al. Rhythm classification of 12-lead ECGs using deep neural networks and class-activation maps for improved explainability. In *2020 Computing in Cardiology*, pages 1–4, 2020.

[118] Lucas Weber, Maksym Gaiduk, Wilhelm Daniel Scherz, and Ralf Seepold. Cardiac abnormality detection in 12-lead ECGs with deep convolutional neural networks using data augmentation. In *2020 Computing in Cardiology*, pages 1–4, 2020.

[119] P Natesan, E Gothai, et al. Classification of multi-lead ECG signals to predict myocardial infarction using cnn. In *International Conference on Computing Methodologies and Communication*, pages 1029–1033, 2020.

[120] Mohammed Tali Almalchy, Sarmad Monadel Sabree ALGayar, and Nirvana Popescu. Atrial fibrillation automatic diagnosis based on ECG signal using pretrained deep convolution neural network and svm multiclass model. In *International Conference on Communications*, pages 197–202, 2020.

[121] Bin Zhou, Shenghua Liu, Bryan Hooi, Xueqi Cheng, and Jing Ye. BeatGAN: Anomalous rhythm detection using adversarially generated time series. In *IJCAI*, volume 2019, pages 4433–4439, 2019.

[122] Yun Kwan Kim, Minji Lee, Hee Seok Song, and Seong-Whan Lee. Automatic cardiac arrhythmia classification using residual network combined with long short-term memory. *IEEE Transactions on Instrumentation and Measurement*, 71:1–17, 2022.

[123] Hongfu Xie, Hui Liu, Shuwang Zhou, Tianlei Gao, and Minglei Shu. A lightweight 2-D CNN model with dual attention mechanism for heartbeat classification. *Applied Intelligence*, pages 1–16, 2022.

[124] Kogilavani Shanmugavadivel, VE Sathishkumar, M Sandeep Kumar, V Maheshwari, J Prabhu, and Shaikh Muhammad Allayear. Investigation of applying machine learning and hyperparameter tuned deep learning approaches for arrhythmia detection in ECG images. *Computational & Mathematical Methods in Medicine*, 2022.

[125] Ping Cao, Xinyi Li, Kedong Mao, Fei Lu, Gangmin Ning, Luping Fang, and Qing Pan. A novel data augmentation method to enhance deep neural networks for detection of atrial fibrillation. *Biomedical Signal Processing and Control*, 56:101675, 2020.

[126] Dengao Li, Xuemei Li, Jumin Zhao, and Xiaohong Bai. Automatic staging model of heart failure based on deep learning. *Biomedical Signal Processing and Control*, 52:77–83, 2019.

[127] Jinyuan He, Jia Rong, Le Sun, Hua Wang, and Yanchun Zhang. An advanced two-step DNN-based framework for arrhythmia detection. In *Advances in Knowledge Discovery and Data Mining: Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II 24*, pages 422–434, 2020.

[128] Anita Pal, Ranjeet Srivastva, and Yogendra Narain Singh. CardioNet: an efficient ecg arrhythmia classification system using transfer learning. *Big Data Research*, 26:100271, 2021.

[129] Deepankar Nankani and Rashmi Dutta Baruah. An end-to-end framework for automatic detection of atrial fibrillation using deep residual learning. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 690–695, 2019.

[130] Yuxi Zhou, Shenda Hong, Junyuan Shang, Meng Wu, Qingyun Wang, Hongyan Li, and Junqing Xie. K-margin-based residual-convolution-recurrent neural network for atrial fibrillation detection.

*arXiv:1908.06857*, 2019.

[131] Hyeongrok Han, Seongjae Park, Seonwoo Min, Hyun-Soo Choi, Eunji Kim, Hyunki Kim, Sangha Park, Jinkook Kim, Junsang Park, Junho An, Kwanglo Lee, Wonsun Jeong, Sangil Chon, Kwonwoo Ha, Myungkyu Han, and Sungroh Yoon. Towards high generalization performance on electrocardiogram classification. In *Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[132] Nabil Sabor, Garas Gendy, Hazem Mohammed, Guoxing Wang, and Yong Lian. Robust arrhythmia classification based on QRS detection and a compact 1d-cnn for wearable ecg devices. *IEEE Journal of Biomedical and Health Informatics*, 26(12):5918–5929, 2022.

[133] Pingping Bing, Yang Liu, Wei Liu, Jun Zhou, and Lemei Zhu. Electrocardiogram classification using TSST-based spectrogram and ConViT. *Frontiers in Cardiovascular Medicine*, 9:983543–983543, 2022.

[134] Eduardo José da Silva Luz, Gladston J. P. Moreira, Luiz S. Oliveira, William Robson Schwartz, and David Menotti. Learning deep off-the-person heart biometrics representations. *IEEE Transactions on Information Forensics and Security*, 13(5):1258–1270, 2018.

[135] Valerio Mura, Giulia Orrù, Roberto Casula, Alessandra Sibiriu, Giulia Loi, Pierluigi Tuveri, Luca Ghiani, and Gian Luca Marcialis. Livdet 2017 fingerprint liveness detection competition 2017. In *International Conference on Biometrics*, pages 297–302, 2018.

[136] Mohamed Hammad and Kuanquan Wang. Parallel score fusion of ECG and fingerprint for human authentication based on convolution neural network. *Computers & Security*, 81:107–122, 2019.

[137] Donghwan Yun, Hyung-Chul Lee, Chul-Woo Jung, Soonil Kwon, So-Ryoung Lee, Kwangsoo Kim, Yon Su Kim, and Seung Seok Han. Robust r-peak detection in an electrocardiogram with stationary wavelet transformation and separable convolution. *Scientific Reports*, 12(1):19638, 2022.

[138] Álvaro Huerta, Arturo Martínez-Rodrigo, José J Rieta, and Raúl Alcaraz. ECG quality assessment via deep learning and data augmentation. In *Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[139] Pablo Laguna, Roger G Mark, A Goldberg, and George B Moody. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ecg. In *Computers in cardiology 1997*, pages 673–676, 1997.

[140] Zhaohan Xiong, Martin K Stiles, Anne M Gillis, and Jichao Zhao. Enhancing the detection of atrial fibrillation from wearable sensors with neural style transfer and convolutional recurrent networks. *Computers in Biology and Medicine*, 146:105551, 2022.

[141] Xue Zhou, Xin Zhu, Keijiro Nakamura, and Mahito Noro. Electrocardiogram quality assessment with a generalized deep learning model assisted by conditional generative adversarial networks. *Life*, 11(10):1013, 2021.

[142] Haixu Yang, Jihong Liu, Lvheng Zhang, Yan Li, and Henggui Zhang. Proegan-MS: A progressive growing generative adversarial networks for electrocardiogram generation. *IEEE Access*, 9:52089–52100, 2021.

[143] Ke Ma, A Zhan Chang'an, and Feng Yang. Multi-classification of arrhythmias using resnet with CBAM on CWGAN-GP augmented ECG gramian angular summation field. *Biomedical Signal Processing and Control*, 77:103684, 2022.

[144] Jangwon Suh, Jimyeong Kim, Eunjung Lee, Jaeill Kim, Duhun Hwang, Jungwon Park, Junghoon Lee, Jaeseung Park, Seo-Yoon Moon, Yeonsu Kim, et al. Learning ECG representations for multi-label classification of cardiac abnormalities. In *Computing in Cardiology (CinC)*, volume 48, pages 1–4, 2021.

[145] Pritam Sarkar and Ali Etemad. CardioGAN: Attentive generative adversarial network with dual discriminators for synthesis of ECG from PPG. In *AAAI Conference on Artificial Intelligence*, volume 35, pages 488–496, 2021.

[146] Shawn Tan, Guillaume Androz, Ahmad Chamseddine, Pierre Fecteau, Aaron Courville, Yoshua Bengio, and Joseph Paul Cohen. Icentia11k: An unsupervised representation learning dataset for arrhythmia subtype discovery. *arXiv:1910.09570*, 2019.

[147] Karen Fonseca, Sergio Osorio, Jeyson Castillo, and Carlos Fajardo. Contrastive learning for atrial fibrillation detection in challenging

scenarios. In *European Signal Processing Conference*, pages 1218–1222, 2022.

[148] Edmond Adib, Fatemeh Afghah, and John J Prevost. Arrhythmia classification using CGAN-augmented ecg signals. In *International Conference on Bioinformatics and Biomedicine*, pages 1865–1872, 2022.

[149] Zabir Al Nazi, Ananna Biswas, Md Abu Rayhan, and Tasnim Azad Abir. Classification of ECG signals by dot residual LSTM network with data augmentation for anomaly detection. In *International Conference on Computer and Information Technology*, pages 1–5, 2019.

[150] Fangyu Li, Hui Chang, Min Jiang, and Yihuan Su. A contrastive learning framework for ECG anomaly detection. In *International Conference on Intelligent Computing and Signal Processing*, pages 673–677, 2022.

[151] Pu Wang, Borui Hou, Siyu Shao, and Ruqiang Yan. Ecg arrhythmias detection using auxiliary classifier generative adversarial network and residual network. *IEEE Access*, 7:100910–100922, 2019.

[152] Rohan Banerjee and Avik Ghose. Synthesis of realistic ECG waveforms using a composite generative adversarial network for classification of atrial fibrillation. In *European Signal Processing Conference*, pages 1145–1149, 2021.

[153] Yi Xia, Yangyang Xu, Peng Chen, Jun Zhang, and Yongliang Zhang. Generative adversarial network with transformer generator for boosting ecg classification. *Biomedical Signal Processing and Control*, 80:104276, 2023.

[154] Han Sun, Fan Zhang, and Yunxiang Zhang. An LSTM and GAN based ECG abnormal signal generator. In *Advances in Artificial Intelligence and Applied Cognitive Computing: Proceedings from ICAI'20 and ACC'20*, pages 743–755, 2021.

[155] Eoin Brophy, Maarten De Vos, Geraldine Boylan, and Tomás Ward. Multivariate generative adversarial networks and their loss functions for synthesis of multichannel ecgs. *IEEE Access*, 9:158936–158945, 2021.

[156] Abdelrahman M. Shaker, Manal Tantawi, Howida A. Shedeed, and Mohamed F. Tolba. Generalization of convolutional neural networks for ECG classification using generative adversarial networks. *IEEE Access*, 8:35592–35605, 2020.

[157] Khondker Fariha Hossain, Sharif Amit Kamran, Alireza Tavakkoli, Lei Pan, Xingjun Ma, Sutharshan Rajasegarar, and Chandan Karmaker. ECG-Adv-GAN: Detecting ECG adversarial examples with conditional generative adversarial networks. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 50–56, 2021.

[158] Jian Liu, Xiaodong Xia, Xiang Peng, Jiao Hui, and Chunyang Han. Research on ecg signal classification based on data enhancement of generative adversarial network. In *Artificial Intelligence and Security: International Conference, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part I*, pages 405–419, 2022.

[159] Adyasha Rath, Debahuti Mishra, Ganapati Panda, and Suresh Chandra Satapathy. Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomedical Signal Processing and Control*, 68:102820, 2021.

[160] Haiyan Wang, Yanjie Zhou, Bing Zhou, Xiangdong Niu, Hua Zhang, and Zongmin Wang. Interactive ECG annotation: An artificial intelligence method for smart ECG manipulation. *Information Sciences*, 581:42–59, 2021.

[161] Xiaoyu Wang, Bingchu Chen, Ming Zeng, Yuli Wang, Hui Liu, Ruixia Liu, Lan Tian, and Xiaoshan Lu. An ECG signal denoising method using conditional generative adversarial net. *IEEE Journal of Biomedical and Health Informatics*, 26(7):2929–2940, 2022.

[162] Yilin Wang, Le Sun, and Sudha Subramani. CAB: classifying arrhythmias based on imbalanced sensor data. *KSII Transactions on Internet and Information Systems*, 15(7):2304–2320, 2021.

[163] Chaofan Du, Peter Xiaoping Liu, and Minhua Zheng. Classification of imbalanced electrocardiosignal data using convolutional neural network. *Computer Methods and Programs in Biomedicine*, 214:106483, 2022.

[164] Md Shofiqul Islam, Md Nahidul Islam, Noramiza Hashim, Mamunur Rashid, Bifta Sama Bari, and Fahmid Al Farid. New hybrid deep learning approach using BiGRU-BiLSTM and multilayered dilated CNN to detect arrhythmia. *IEEE Access*, 10:58081–58096, 2022.

[165] Yan He, Bin Fu, Jian Yu, Renfa Li, and Rucheng Jiang. Efficient learning of healthcare data from IoT devices by edge convolution neural networks. *Applied Sciences*, 10(24):8934, 2020.

[166] Tomer Golany, Gal Lavee, Shai Tejman Yarden, and Kira Radinsky. Improving ECG classification using generative adversarial networks. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13280–13285, 2020.

[167] Shuai Ma, Jianfeng Cui, Weidong Xiao, and Lijuan Liu. Deep learning-based data augmentation and model fusion for automatic arrhythmia identification and classification algorithms. *Computational Intelligence and Neuroscience*, 2022, 2022.

[168] Valeriia Guryanova. Online augmentation for quality improvement of neural networks for classification of single-channel electrocardiograms. In *Analysis of Images, Social Networks and Texts: International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers 8*, pages 37–49, 2020.

[169] Dong-Hoon Shin, Roy C. Park, and Kyungyong Chung. Decision boundary-based anomaly detection model using improved AnoGAN from ECG data. *IEEE Access*, 8:108664–108674, 2020.

[170] Taki Hasan Rafi and Young Woong Ko. HeartNet: self multihead attention mechanism via convolutional network with adversarial data synthesis for ECG-based arrhythmia classification. *IEEE Access*, 10:100501–100512, 2022.

[171] Haixu Yang, Jihong Liu, Lvheng Zhang, Yan Li, and Henggui Zhang. ProEGAN-MS: A progressive growing generative adversarial networks for electrocardiogram generation. *IEEE Access*, 9:52089–52100, 2021.

[172] Tomer Golany, Daniel Freedman, and Kira Radinsky. ECG ODE-GAN: Learning ordinary differential equations of ecg dynamics via generative adversarial learning. In *AAAI Conference on Artificial Intelligence*,

volume 35, pages 134–141, 2021.

[173] Andres Hernandez-Matamoros, Hamido Fujita, and Hector Perez-Meana. A novel approach to create synthetic biomedical signals using BiRNN. *Information Sciences*, 541:218–241, 2020.

[174] Zhenge Jia, Feng Hong, Lichuan Ping, Yiyu Shi, and Jingtong Hu. Enabling on-device model personalization for ventricular arrhythmias detection by generative adversarial networks. In *ACM/IEEE Design Automation Conference (DAC)*, pages 163–168, 2021.

[175] Le Sun, Yilin Wang, Zhiguo Qu, and Neal N. Xiong. BeatClass: A sustainable ECG classification system in IoT-based ehealth. *IEEE Internet of Things Journal*, 9(10):7178–7195, 2022.

[176] Barbara Mukami Maweu, Rittika Shamsuddin, Sagnik Dakshit, and Balakrishnan Prabhakaran. Generating healthcare time series data for improving diagnostic accuracy of deep neural networks. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15, 2021.

[177] Han Liu, Zhenbo Zhao, and Qiang She. Self-supervised ECG pre-training. *Biomedical Signal Processing and Control*, 70:103010, 2021.

[178] Kamana Dahal and Mohd Hasan Ali. A hybrid GAN-based dl approach for the automatic detection of shockable rhythms in AED for solving imbalanced data problems. *Electronics*, 12(1):13, 2022.

[179] Yu Deng, Zhongquan Gao, Songhua Xu, Pengyu Ren, Yang Wen, Ying Mao, and Zongfang Li. ST-Net: Synthetic ECG tracings for diagnosing various cardiovascular diseases. *Biomedical Signal Processing and Control*, 61:101997, 2020.

[180] Wenqiang Li, Yuk Ming Tang, Kai Ming Yu, and Suet To. SLC-GAN: An automated myocardial infarction detection model based on generative adversarial networks and convolutional neural networks with single-lead electrocardiogram synthesis. *Information Sciences*, 589:738–750, 2022.

[181] Yefei Zhang, Zhidong Zhao, Yanjun Deng, and Xiaohong Zhang. FHRGAN: Generative adversarial networks for synthetic fetal heart rate signal generation in low-resource settings. *Information Sciences*, 594:136–150, 2022.

[182] Andrei Furdui, Tianyi Zhang, Marcel Worring, Pablo Cesar, and Abdallah El Ali. AC-WGAN-GP: Augmenting ECG and GSR signals using conditional generative models for arousal classification. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing and ACM International Symposium on Wearable Computers*, pages 21–22, 2021.

[183] Amit Garg and Nima Karimian. ECG biometric spoofing using adversarial machine learning. In *International Conference on Consumer Electronics*, pages 1–5, 2021.

[184] Jiayuan Hu and Yong Li. Electrocardiograph based emotion recognition via WGAN-GP data enhancement and improved CNN. In *Intelligent Robotics and Applications: International Conference, ICIRA 2022, Harbin, China, August 1–3, 2022, Proceedings, Part I*, pages 155–164, 2022.

[185] Munawara Saiyara Munia, Mehrdad Nourani, and Sammy Houari. Biosignal oversampling using wasserstein generative adversarial network. In *International Conference on Healthcare Informatics*, pages 1–7, 2020.

[186] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[187] Alena I Kalyakulina, Igor I Yusipov, Viktor A Moskalenko, Alexander V Nikolskiy, Konstantin A Kosonogov, Grigory V Osipov, Nikolai Yu Zolotykh, and Mikhail V Ivanchenko. LUDB: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8:186181–186190, 2020.

[188] VV Kuznetsov, VA Moskalenko, DV Gribanov, and Nikolai Yu Zolotykh. Interpretable feature generation in ECG using a variational autoencoder. *Frontiers in genetics*, 12:638191, 2021.

[189] Fei Ye, Fei Zhu, Yuchen Fu, and Bairong Shen. ECG generation with sequence generative adversarial nets optimized by policy gradient. *IEEE Access*, 7:159369–159378, 2019.

[190] Hyo-Chang Seo, Gi-Won Yoon, Segyeong Joo, and Gi-Byoung Nam. Multiple electrocardiogram generator with single-lead electrocardiogram. *Computer Methods and Programs in Biomedicine*, 221:106858, 2022.

[191] Bingxin Xu, Ruixia Liu, Minglei Shu, Xiaoyi Shang, and Yinglong Wang. An ECG denoising method based on the generative adversarial residual network. *Computational and Mathematical Methods in Medicine*, 2021:1–23, 2021.

[192] Debapriya Hazra and Yung-Cheol Byun. SynSigGAN: Generative adversarial networks for synthetic biomedical signal generation. *Biology*, 9(12):441, 2020.

[193] Reza Soleimani and Edgar Lobaton. Enhancing inference on physiological and kinematic periodic signals via phase-based interpretability and multi-task learning. *Information*, 13(7):326, 2022.

[194] Deepankar Nankani and Rashmi Dutta Baruah. Investigating deep convolution conditional GANs for electrocardiogram generation. In *International Joint Conference on Neural Networks*, pages 1–8, 2020.

[195] JeeEun Lee, KyeongTaek Oh, Byeongnam Kim, and Sun K Yoo. Synthesis of electrocardiogram V-lead signals from limb-lead measurement using R-peak aligned generative adversarial network. *IEEE journal of biomedical and health informatics*, 24(5):1265–1275, 2019.

[196] Fei Zhu, Fei Ye, Yuchen Fu, Quan Liu, and Bairong Shen. Electrocardiogram generation with a bidirectional LSTM-CNN generative adversarial network. *Scientific reports*, 9(1):1–11, 2019.

[197] Shaobin Huang, Peng Wang, and Rongsheng Li. Noise ECG generation method based on generative adversarial network. *Biomedical Signal Processing and Control*, 81:104444, 2023.

[198] Pratik Singh and Gayadhar Pradhan. A new ECG denoising framework using generative adversarial network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(2):759–764, 2021.

[199] Fangjian Chen, Yun Pan, Ke Li, Kwang-Ting Cheng, and Ruohong Huan. Standard 12-lead ECG synthesis using a GA optimized BP neural network. In *International Conference on Advanced Computational Intelligence*, pages 289–293, 2015.

[200] Mohamed Amine Abdelmadjid and Mounir Boukadoum. Neural network-based signal translation with application to the ECG. In *IEEE Interregional NEWCAS Conference*, pages 542–546, 2022.

[201] Vajira Thambawita, Jonas L Isaksen, Steven A Hicks, Jonas Ghouse, Gustav Ahlberg, Allan Linneberg, Niels Grarup, Christina Ellervik, Morten Salling Olesen, Torben Hansen, et al. DeepFake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine. *Scientific reports*, 11(1):21896, 2021.

[202] Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ECG generation with structured state space models. *Computers in biology and medicine*, 163:107115, 2023.

[203] Md Moklesur Rahman, Massimo Walter Rivolta, Fabio Badilini, and Roberto Sassi. A systematic survey of data augmentation of ECG signals for AI applications. *Sensors*, 23(11):5237, 2023.

[204] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[205] James Belen, Sajad Mousavi, Alireza Shamsoshoara, and Fatemeh Afghah. An uncertainty estimation framework for risk assessment in deep learning-based AFib classification. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*, pages 960–964, 2020.

[206] Ahmad O Aseeri. Uncertainty-aware deep learning-based cardiac arrhythmias classification model of electrocardiogram signals. *Computers*, 10(6):82, 2021.

[207] Jeroen F Vranken, Rutger R van de Leur, Deepak K Gupta, Luis E Juarez Orozco, Rutger J Hassink, Pim van der Harst, Pieter A Doevendans, Sadaf Gulshad, and René van Es. Uncertainty estimation for deep learning-based automated analysis of 12-lead electrocardiograms. *European Heart Journal-Digital Health*, 2(3):401–415, 2021.

[208] Brian Chen, Golara Javadi, Alexander Hamilton, Stephanie Sibley, Philip Laird, Purang Abolmaesumi, David Maslove, and Parvin Mousavi. Quantifying deep neural network uncertainty for atrial fibrillation detection with limited labels. *Scientific Reports*, 12(1):20140, 2022.

[209] Yonatan Elul, Aviv A Rosenberg, Assaf Schuster, Alex M Bronstein, and Yael Yaniv. Meeting the unmet needs of clinicians from AI systems showcased for cardiology with deep-learning–based ECG analysis. *Proceedings of the National Academy of Sciences*, 118(24):e2020620118, 2021.

[210] JaeYeon Park, Kichang Lee, Noseong Park, Seng Chan You, and Jeong-Gil Ko. Self-attention LSTM-FCN model for arrhythmia classification and uncertainty assessment. *Artificial Intelligence in Medicine*, 142:102570, 2023.

[211] V Jahmunah, EYK Ng, Ru-San Tan, Shu Lih Oh, and U Rajendra Acharya. Uncertainty quantification in DenseNet model using myocardial infarction ECG signals. *Computer Methods and Programs in Biomedicine*, 229:107308, 2023.

[212] Md Moklesur Rahman, Massimo Walter Rivolta, Fabio Badilini, and Roberto Sassi. Quantifying uncertainty of a deep learning model for atrial fibrillation detection from ECG signals. In *2023 Computing in Cardiology*, volume 50, pages 1–4, 2023.

[213] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get M for free. *arXiv:1704.00109*, 2017.

[214] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

[215] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *NeurIPS*, 31, 2018.

[216] Wenrui Zhang, Xinxin Di, Guodong Wei, Shijia Geng, Zhaoji Fu, and Shenda Hong. A deep bayesian neural network for cardiac arrhythmia classification with rejection from ECG recordings. *arXiv:2203.00512*, 2022.

[217] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit,

and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[218] Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in neural information processing systems*, 26, 2013.

[219] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[220] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020.

[221] Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed-ensembles for efficient uncertainty estimation. *arXiv:2210.09184*, 2022.

[222] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[223] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. In *International conference on machine learning*, pages 2782–2792, 2020.

[224] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv:1803.05407*, 2018.

[225] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. *Advances in neural information processing systems*, 32, 2019.

[226] Wu Lin, Mark Schmidt, and Mohammad Emtiyaz Khan. Handling the positive-definite constraint in the Bayesian learning rule. In *International conference on machine learning*, pages 6116–6126, 2020.

[227] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[228] Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6, 2018.

[229] Glenn Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1:330–331, 1992.

[230] Hyo-Chang Seo, Seok Oh, Hyunbin Kim, and Segyeong Joo. ECG data dependency for atrial fibrillation detection based on residual networks. *Scientific Reports*, 11(1):18256, 2021.

[231] Xiongjie Chen, Yunpeng Li, and Yongxin Yang. Batch-Ensemble stochastic neural networks for out-of-distribution detection. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[232] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330, 2017.

[233] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804, 2018.

[234] Luiz Vasconcelos, Bryan Perez Martinez, Madeline Kent, Sardar Ansari, Hamid Ghanbari, and Ivan Nenadic. Multi-center atrial fibrillation electrocardiogram (ecg) classification using fourier space convolutional neural networks (fd-cnn) and transfer learning. *Journal of electrocardiology*, 81:201–206, 2023.

[235] Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim GJ Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud De Kroon, and Yarin Gal. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv:1912.10481*, 2019.

[236] Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Eli Gibson, RS Vishwanath, Abishek Balachandran, James M Balter, Yue Cao, Ramandeep Singh, et al. Quantifying and leveraging predictive uncertainty for medical image assessment. *Medical Image Analysis*, 68:101855, 2021.

[237] Andrejs Fedjajevs, Willemijn Groenendaal, Carlos Agell, and Evelien Hermeling. Platform for analysis and labeling of medical time series. *Sensors*, 20(24):7302, 2020.

[238] Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *International conference on machine learning*, pages 1789–1799, 2020.

[239] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[240] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, Claire Marais Sicre, and Gérard Dedieu. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing*, 9(2):173, 2017.

[241] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv:1705.10694*, 2017.

[242] Yifan Ding, Liqiang Wang, Deliang Fan, and Boqing Gong. A semi-supervised two-stage approach to learning from noisy labels. In *IEEE Winter conference on applications of computer vision (WACV)*, pages 1215–1224, 2018.

[243] Qingrui Jia, Xuhong Li, Lei Yu, Jiang Bian, Penghao Zhao, Shupeng Li, Haoyi Xiong, and Dejing Dou. Learning from training dynamics: Identifying mislabeled data beyond manually designed features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8041–8049, 2023.

[244] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[245] Edoardo Pasolli and Farid Melgani. Genetic algorithm-based method for mitigating label noise issue in ECG signal classification. *Biomedical Signal Processing and Control*, 19:130–136, 2015.

[246] Yaoguang Li and Wei Cui. Identifying the mislabeled training samples of ecg signals using machine learning. *Biomedical signal processing and control*, 47:168–176, 2019.

[247] Xinwen Liu, Huan Wang, and Zongjin Li. An approach for deep learning in ECG classification tasks in the presence of noisy labels. In *International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 369–372, 2021.

[248] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[249] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv:1412.6596*, 2014.

[250] Boyan Gao, Henry Gouk, and Timothy M Hospedales. Searching for robustness: Loss learning for noisy classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6670–6679, 2021.

[251] S Balasundaram and Subhash Chandra Prasad. Robust twin support vector regression based on Huber loss function. *Neural Computing and Applications*, 32(15):11285–11309, 2020.

[252] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv:1710.09412*, 2017.

[253] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[254] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015.

[255] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization

in deep networks. In *International conference on machine learning*, pages 233–242, 2017.

[256] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.