Addis Ababa University
College of Technology and Built Environment
School of Electrical and Computer Engineering

Interpretable Hybrid Multichannel Deep Learning for 12-Lead ECG-based Heart Disease Classification

By: Yehualashet Megersa Ayano

Supervisors:

1. Prof. Dr. Friedhelm Schwenker
2. Dr. Bisrat Derebssa Dufera
3. Dr. Taye Girma Debelee

A PhD dissertation submitted to the School of Electrical and Computer Engineering, College of Technology and Built Environment, Addis Ababa University in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Engineering.

21 February, 2025

Addis Ababa University

College of Technology and Built Environment (CTBE)

School of Electrical and Computer Engineering (SECE)

By: ***Yehualashet Megersa Ayano***

This is to certify that the dissertation prepared by Yehualashet Megersa Ayano titled, **"Interpretable Hybrid Multichannel Deep Learning for 12-Lead ECG-based Heart Disease Classification"** and submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Engineering complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved and signed by board of Examining Committee

	Name	Signature	Date
Dean, SECE	Dr. Sosina Gashaw	.....	.....
Supervisor	Prof. Dr. Friedhelm Schwenker	.....	.....
Co-supervisor	Dr. Bisrat Derebssa Dufera	.....	.....
Co-supervisor	Dr. Taye Girma Debelee	.....	.....
Internal Examiner	Dr. Fitsum Assamnew	.....	.....
External Examiner	Prof. Dr.-Ing. Achim Ibenthal	.....	.....

DECLARATION

I, the undersigned, declare that this dissertation titled **"Interpretable Hybrid Multichannel Deep Learning for 12-Lead ECG-based Heart Disease Classification"** was prepared by me, with the guidance of my supervisors. The work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted, in whole or in part, for any other degree or professional qualification. Parts of this work have been published in Journal of Diagnostics and Journal of IEEE Access.

By: Signature

Yehualashet Megersa Ayano

Witnessed by:

Name of student supervisor: Signature

Prof. Dr. Friedhelm Schwenker
Ulm University, Institute of Neural Information Processing

Name of student co-supervisor: Signature

Dr. Bisrat Derebssa Dufera
Addis Ababa University, CTBE, School of Electrical and Computer Engineering

---

Dr. Taye Girma Debelee Signature
Artificial Intelligence Institute, and Addis Ababa Science and Technology University

---

ABSTRACT

The electrocardiogram (ECG) is a noninvasive and affordable tool that offers valuable insights into heart activity from multiple perspectives. However, medical practitioners often face difficulties in diagnosing underlying heart conditions from ECG signals. To address these challenges and improve diagnostic accuracy, researchers have investigated the potential of deep learning (DL) techniques. Nevertheless, developing a robust and interpretable deep learning model that performs well across diverse ECG datasets remains a key research focus.

Thus, in this PhD research, an interpretable deep learning system is designed, incorporating preprocessing of ECG signal and post-hoc interpretability. The designed model is a multichannel hybrid deep learning architecture consisting of 12 blocks, each combining a one-dimensional (1D) convolutional neural network (CNN) with bidirectional long short-term memory (BiLSTM) networks. After the 12 blocks, the feature maps are concatenated and further processed by an attention mechanism and a two-dimensional (2D) CNN. All components, including the 1D CNN, BiLSTM layers, attention mechanism, and 2D CNN, are used as feature extraction backbones. Subsequently, fully connected (FC) layers are incorporated for classification. The model was independently trained and tested on three distinct 12-lead ECG datasets: (1) the PTB-XL dataset, using five super-diagnostic classes, (2) the CODE-15% dataset, encompassing six heart disease classes, and (3) the Chapman Arrhythmia datasets, which were analyzed using two configurations: seven reduced classes (Chapman-Reduced) and four merged classes (Chapman-Merged). The model achieved average test accuracy rates of 89.84%, 97.82%, 98.55%, and 98.80% for these datasets, respectively. The result indicates the model's effectiveness across different ECG datasets.

To understand how the model reached its classification result, we applied two post-hoc interpretability techniques: Gradient-weighted Class Activation Mapping plus (Grad-CAM++) and SHapley Additive exPlanations (SHAP). These techniques were used to visualize influential segments of the ECG signal, both at the instance level for specific samples and at the test set level for assessing the overall contributions of individual ECG leads. SHAP, with its theoretical grounding, ensures consistent feature attribution by capturing causal relationships within the ECG data. Meanwhile, Grad-CAM++, through causal localization, identifies regions of the ECG signals that influenced the model's decisions. The interpretability provided

from both techniques were cross-checked against heart disease manifestations in ECG signals using established cardiology literature, ensuring alignment with clinical patterns. The model's performance and the output interpretation techniques demonstrate that the proposed approach is a practicable tool for ECG-based heart disease diagnosis.

Keywords: Heart disease, 12-leads ECG, CNN-BiLSTM, deep learning, interpretability, Grad-CAM++, SHAP.

ACKNOWLEDGMENTS

First, I would like to express my heartfelt gratitude to my advisors, Prof. Dr. Friedhelm Schwenker, Dr. Bisrat Derebssa, and Dr. Taye Girma, for their unwavering support, inspiration, and invaluable feedback throughout my PhD study.

I would also like to express my sincere gratitude to Dr. Zelalem Chimdesa at Zewuditu Memorial Hospital, Dr. Tariku Fekadu, a specialist in internal medicine, and Dr. Yitagessu Getachew, a cardiologist at Yekatit 12 Hospital Medical College, for their invaluable guidance, support in pilot testing the model, and insightful discussions.

Additionally, I am grateful to the team of experts at the Ethiopian Artificial Intelligence Institute for their support and encouragement. In addition, I would like to extend my special thanks to the Electrical and Computer Engineering staff at the College of Technology and Built Environment for creating a conducive environment throughout my studies.

Finally, I would like to express my deepest gratitude for the unwavering love, support, and encouragement of my beloved wife, Rahel Gebresilassie, and my parents. I also cherish the joy and inspiration brought by my daughter, Heran Yehualashet.

Contents

List of Tables

List of Figures

LIST OF ACRONYMS

1D	One-Dimensional
2D	Two-Dimensional
1dAVb	First Degree Atrioventricular Block
AF	Atrial Flutter
AFIB	Atrial Fibrillation
API	Application Program Interface
AT	Atrial Tachycardia
AUC	Area Under the Curve
AUPRC	Area Under the Precision-Recall Curve
AVG	Average
AV	Atrioventricular
AVNRT	Atrioventricular Node Reentrant Tachycardia
AVRT	Atrioventricular Reentrant Tachycardia
Bi-GRU	Bi-directional Gated Recurrent Units
Bi-LSTM	Bidirectional Long Short Term Memory
BW	Baseline-wander
CAM	Class Activation Map
CCTA	Coronary Computed Tomography Angiogram
CD	Conduction Disturbance
CIE	Computerized Interpretation Of ECG
CNN	Convolutional Neural Network
CODE	Clinical Outcomes In Digital Electrocardiography
DCT	Discrete Cosine Transform
DL	Deep Learning
DNN	Deep Neural Network
ECG	Electrocardiogram
FC	Fully Connected
FCN	Fully Convolutional Network
FN	False Negative
FP	False Positive
GAP	Global Average Pooling
Grad-CAM++	Gradient-weighted Class Activation Map Plus Plus
GRU	Gated Recurrent Units
GSVT	Grouped Supraventricular Tachycardia
GUI	Graphical User Interface
HYP	Hypertrophy
Hz	Hertz
IDCT	Inverse Discrete Cosine Transform
IML	Interpretable Machine Learning

KNN	K-Nearest Neighbor
LBBB	Left Bundle Branch Block
LIME	Local Interpretable Model Agnostic Explanation
LRP	Layer-wise Relevance Propagation
LSTM	Long Short Term Memory
M	Millions
MI	Myocardial Infarction
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
NORM	Normal
PFI	Permutation Feature Importance
PLI	Power Line Interface
RBBB	Right Bundle Branch Block
ReLU	Rectifier Linear Unit
RNN	Recurrent Neural Network
SAAWR	Sinus Atrium to Atrial Wandering Rhythm
SA	Sinoatrial
SB	Sinus Bradycardia
SHAP	SHapley Additive exPlanations
SI	Sinus Irregularity
SR	Sinus Rhythm
ST	Sinus Tachycardia
STTC	ST/T Change
SVEB	Supra-Ventricular Ectopic Beats
SVT	Supraventricular Tachycardia
t-SNE	t-distributed Stochastic Neighbor-Embedding
TN	True Negative
TNMG	Telehealth Network Of Minas Gerais
TP	True Positive
VEB	Ventricular Ectopic Beats
WAVG	Weighted Average
WHO	World Health Organization
ZMH	Zewditu Memorial Hospital

LIST OF SYMBOLS

β_{ij}	Attention weight of the encoder state at position j for output position i
$\gamma[k]$	Scaling factor applied to the k^{th} frequency component in the Discrete Cosine Transform (DCT)
Ω	Asymptotic lower bound (algorithm complexity)
$\mathfrak{c}$	Channel (number of leads in 12-lead ECG)
α_{ij}^{kc}	Coefficient representing the contribution of the activation at spatial location (i, j) in the k^{th} feature map to the output class c
$\odot$	Hadamard product
$*_m$	Dilated convolution operator with dilation factor m
$\xi(\mathbf{x})$	The explanation for an instance feature vector $\mathbf{x}$ produced by LIME
$\neg$	Logical negation (NOT)
$\mathscr{L}$	Loss function
μ	Mean
$\mathscr{M}$	ML model
λ	Proximity penalty function in LIME
$\in$	Set membership operator
$\mathbb{R}$	Set of real numbers
$\cup$	Set union operator
ϕ	Shapley value
$\subseteq$	Subset operator
$\mathscr{D}$	Dataset
τ_l	The l^{th} permutation of the instances in $\mathscr{D}$
$\boldsymbol{\nu}_i$	Attention score vector at output position i, computed as the weighted sum of encoder states using attention weights β_{ij}
$a(\mathbf{s}_{i-1}, \mathbf{h}_j)$	Alignment scoring function that computes the relevance between $\mathbf{s}_{i-1}$ and $\mathbf{h}_j$
$\mathbf{A}$	Attention logits computed as $\mathbf{A} = \mathbf{Q}_r \cdot \mathbf{K}^T$, indicating similarity scores between query and key
$\mathbf{A}_r$	Reshaped attention logits, $\mathbf{A}_r \in \mathbb{R}^{s \times u \times u}$, representing the attention scores at each data point in sequence s
b	Bias term in the linear approximation of the model $\mathscr{M}$ for class c

$\mathbf{b}$	Learnable bias vector used in the attention computation
$\mathbf{b}_{n_f}$	Bias term added to the output feature map for filter n_f
c	Class category
$\hat{d}$	A set of $\hat{\mathbf{x}}$
d	A distance metric
e_{ij}	Attention score between the decoder state preceding output position i and the encoder hidden state at position j
$\mathbf{E}_{i,j,n_f}$	The output of the convolutional operation tensor value at spatial location (i, j) for the n_f^{th} filter
f_s	Sampling frequency
$f_{\max}$	Maximum frequency that can be represented (Nyquist frequency), equal to half the sampling rate: $f_{\max} = \frac{f_s}{2}$
f_{BW}	Low-pass cutoff frequency used to characterize baseline wander (BW)
F	2-dimensional feature matrix
$\mathbf{F}$	Convolutional tensor (feature map)
$\mathbf{g}_{f_{1\times N}}$	Discrete Cosine Transformed and baseline wander (BW) removed single-lead ECG signal represented as a vector, where $1 \times N$ indicates a single channel with N samples
$\mathbf{g}[k]$	Discrete Cosine Transformed representation of the input vector (time-domain signal) $\mathbf{x}$ in the frequency domain, the indices k correspond to the frequency axes
$h_{\mathbf{x}}$	Mapping function
$\mathbf{h}_j$	Encoder hidden state at input position j
$\mathbf{H}$	Output tensor from the BiLSTM, with shape $\mathbb{R}^{\mathfrak{c}\times s\times u}$, where $\mathfrak{c}$ is the number of channels (e.g., leads), s is the sequence length (number of data points), and u is the number of features (BiLSTM units)
$I\{\mathbf{x} \in R_m\}$	A binary identity function that gives 1 if $\mathbf{x}$ is in the R_m subset, or else it returns 0
$\mathbf{K}$	Key tensor for each data point in sequence s, $\mathbf{K} \in \mathbb{R}^{u\times\mathfrak{c}}$, preserving the channel and feature dimensions of $\mathbf{H}$
L_{ij}^c	Class-discriminative localization map value at spatial location (i, j), indicating the contribution of that location in the feature map to class c

l_r	Kernel size along the row dimension
$l_{\mathfrak{c}}$	Kernel size along the column dimension
$\mathscr{M}_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})$	Marginal value of $\mathscr{M}$ for the feature values present in S plus feature x_i
$\mathbf{m}_1, \mathbf{m}_2$	Binary masks where $\mathbf{m}_1$ selects features to retain, $\mathbf{m}_2$ selects features to replace
n_f	Number of filters in the convolutional kernel tensor $\mathbf{F}$, corresponding to the output feature map dimension
O	Class output
o_v	A constant value used to overwrite masked (irrelevant) features
$\text{PFI}(\mathscr{M}, j)$	The contribution of feature j to predicting y^c by quantifying the change in model $\mathscr{M}$'s performance when j is permuted
$\mathbf{Q}$	Query tensor computed as $\tanh(\mathbf{H}\cdot\mathbf{W}+\mathbf{b})$
$\mathbf{Q}_r$	Reshaped query tensor, $\mathbf{Q}_r \in \mathbb{R}^{u\times\mathfrak{c}}$, obtained at each data point in sequence s
R_m	The m^{th} subset of the input space in the decision tree
$R_i^{(l)}$	Relevance score of neuron i in layer l, obtained by aggregating its contributions to neurons in layer $l+1$
s	The output sequence length of the BiLSTM network
S	A subset from all features F
$\mathbf{s}_{i-1}$	Decoder hidden state immediately before producing the output at position i
t	Temporal instance or time step
$\tanh(\cdot)$	Hyperbolic tangent activation function, applied element-wise
u	Number of features for each element in the output sequence, equal to the number of units in the second BiLSTM layer
$\mathbf{v}[n]$	Intermediate output of the 1D dilated convolution at index n
w_k^c	Importance weight of the k^{th} convolutional feature map for class c in the class activation map
w_{ij}^{fc}	Weight connecting neuron i to neuron j in a fully connected (fc) layer
$\mathbf{w}[k]$	1D convolutional kernel represented as a vector, where k indexes the position within the kernel

$\mathbf{w}^T$	Transposed weight vector corresponding to class c in the linear approximation of the model $\mathscr{M}$
$\mathbf{w}^{\text{conv}}$	Kernel weight vector for a 1D CNN
$\mathbf{W}$	Learnable weight tensor used in the attention computation
$\mathbf{W}_{\text{attn}}$	Attention weights obtained by applying softmax to $\mathbf{A}_r$ along the last axis
$\mathbf{x}_k$	A vector in the k^{th} row of a data of size $N \times M$
x_i	A single element from the vector $\mathbf{x}_k$
X_f	BW filtered and inverse DCT-transformed 12-lead ECG Signal
$X_{f_{\text{lead}\times N}}$	BW-filtered and inverse DCT-transformed ECG signal, where lead $\times N$ denotes 12 leads and N samples per lead
$\hat{\mathbf{x}}$	Interpretable representation features that are sampled from the original feature space
$\mathbf{x}_S$	Input feature values in a set S
$\mathbf{x}_d$	Input ECG signal after baseline wander (BW) and powerline interference (PLI) noise removal
$x_{d_{\min}}, x_{d_{\max}}$	Minimum and maximum values of $\mathbf{x}_d$ used for normalization
$\mathbf{x}_{\text{norm}}, X_{\text{norm}}$	Normalized ECG signals: $\mathbf{x}_{\text{norm}}$ represents a single-lead ECG signal as a vector, and X_{norm} represents a 12-lead ECG signal as a matrix
y	Ground truth label
$\hat{y}$	Model predicted output
y^c	Class prediction score for class c in the output layer
$\mathbf{Y}_{\text{attn}}$	Output tensor from the attention layer
$\hat{\mathbf{z}}$	Binary perturbed instance in interpretable space used in LIME
$\mathbf{z}$	Real-valued input reconstructed from binary vector $\hat{\mathbf{z}}$ via mapping $h_{\mathbf{x}}$
Z	Set of perturbed samples generated around the original input in LIME, used to train the local surrogate model

CHAPTER 1

Introduction

1.1. Motivation

Heart disease is one of the deadliest health conditions, as it severely affects the heart's function and damages the blood vessels. According to a World Health Organization (WHO) report, in the year 2019, around 17.9 million cardiovascular disease-related deaths were registered. This accounts for 32% of all global mortality, and the highest among all non-communicable diseases [1]. In addition, more than three-fourths of all these mortalities occur in low and middle-income countries [1].

In heart diagnosis clinicians use different procedures and tools such as an electrocardiogram (ECG) [2], echocardiogram [3], coronary computed tomography angiogram (CCTA) [4], cardiac magnetic resonance imaging (MRI) [5], blood tests [6] and coronary angiograms [7]. Among the listed diagnosis techniques, ECG is a low-cost and non-invasive procedure that can easily be administered for diagnosing heart disease [2]. However, according to J. Higueras *et. al.* [8] physicians of all levels face a challenge in accurately picking the underlying heart disease by reading ECG waves. J. Higueras *et al.* [8], revealed that results from the study group of 195 physicians, 153 of whom are residents and 42 staff members, indicated that medical doctors' ability to read ECGs varies widely and reported accuracies are modest. Another study conducted by Amini *et. al.* [9], it is shown that ECG interpretation competency among 323 medical staff and students was 5.13 ± 2.25 for a maximum score of 10. Besides, the finding highlighted that 77.3%, 63.8%, and 62.2% of the participants could not identify normal, myocardial infarction (MI), and pathological Q-waves, respectively. In addition, Getachew *et al.* [10] conducted a cross-sectional study to evaluate the competency of

ECG interpretation among medical interns at Addis Ababa University and Haramaya University. The study reported that the ability of graduating medical students to identify ECG abnormalities, such as anteroseptal ST-segment elevation myocardial infarction (MI), atrial fibrillation (AFIB), and first-degree atrioventricular block (1dAVb) was 42.6%, 39.1%, and 32.1%, respectively. Their finding indicates the difficulty physicians at all levels face in diagnoses of heart diseases in reading and interpreting an ECG. The difficulty is mainly because of the variety of cardiac disease situations and the manifestation similarities of cardiac illness on an ECG.

To mitigate these challenges and aid physicians in the diagnosis of heart conditions, a computerized interpretation of ECG records (CIE) was introduced [11]. However, studies have shown significant inaccuracies of this method and limitations of computerized ECG interpretation [12]. Thus, despite attempts to improve the accuracies of automated ECG interpretation techniques, the final ECG interpretation still requires a physician re-read. Furthermore, the lack of an internationally accepted standard for computerized ECG interpretation poses a challenge for relying on CIE [11]. As a result, researchers have been examining the possibility of machine learning (ML) techniques in interpreting ECG signals for cardiac disease diagnosis.

1.2. Heart Anatomy Overview

The heart is a very critical organ responsible for pumping blood to all parts of the human body through blood vessels by using rhythmic contractions of cardiac muscle or myocardium. It consists of four chambers separated into left and right sides by the wall called the septum. Each side of the heart contains atria and ventricle [13]. The right atrium and ventricle, respectively, collect deoxygenated blood from the body via superior and inferior vena cava, and pump it into the lungs via right and left pulmonary arteries. Whereas, the left atrium and ventricle, respectively, collect oxygenated blood from the lungs via right and left pulmonary veins, and pump it throughout the body via the aorta, as shown in Figure 1.1 [13, 14].

The heart pumps blood in a synchronized manner under the control of the cardiac conduction system. The system consists of sinoatrial (SA) node, atrioventricular (AV) node, and conduction cells, as shown in Figure 1.2 [14]. The SA node initiates the cardiac cycle by

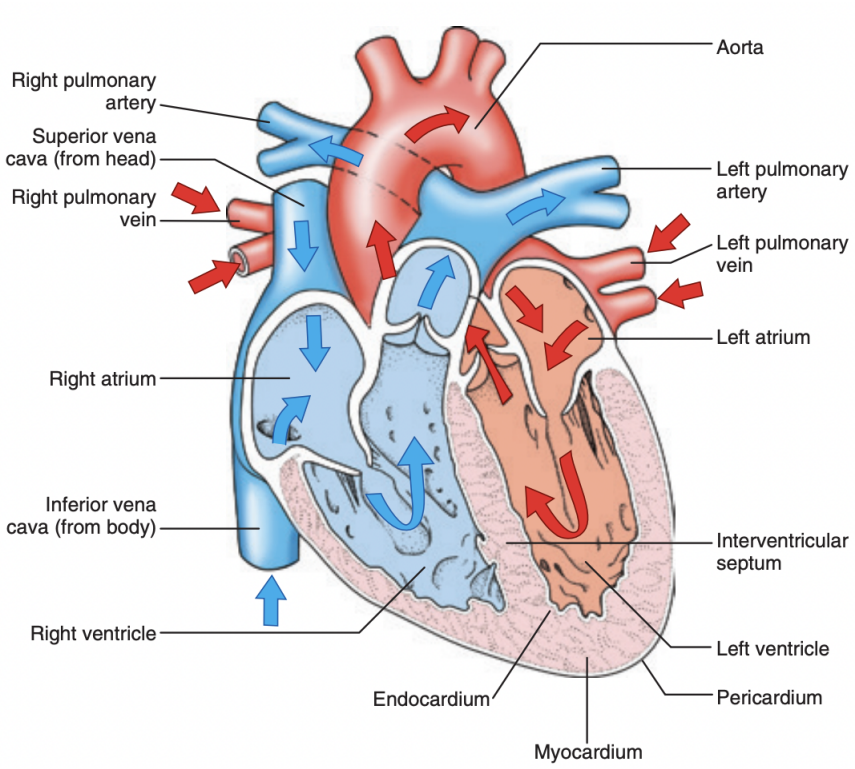


Figure 1.1: Sectional View of the Heart: the blood flow and physiology [14].

stimulating the contraction of cardiac muscle fibers [14]. The process is rhythmic and determines the heart rate, which is 70 to 80 times per minute for a normal adult. The characteristic rhythm of this SA node is called sinus rhythm. The impulse generated by the SA node passes to the AV node via junctional fibers of the cardiac conduction system. The AV node passes the cardiac impulse into the interventricular septum via the atrioventricular bundle. The atrioventricular (AV) bundle branches spread into enlarged Purkinje fibers and continue to the heart's apex curving around the ventricles. The Purkinje fibers have numerous small branches that become continuous with cardiac muscle fibers. The stimulation of the Purkinje fiber causes the ventricle walls to contract in twisting motion that forces blood into the aorta and pulmonary trunk [14]. So, the impulse initiated in the SA node passes through the AV node, atrioventricular (AV) bundle, right and left bundle branches, and finally ends in Purkinje fibers. This electrical activity of the heart is picked by electrodes placed on the skin and results in an ECG waveform.

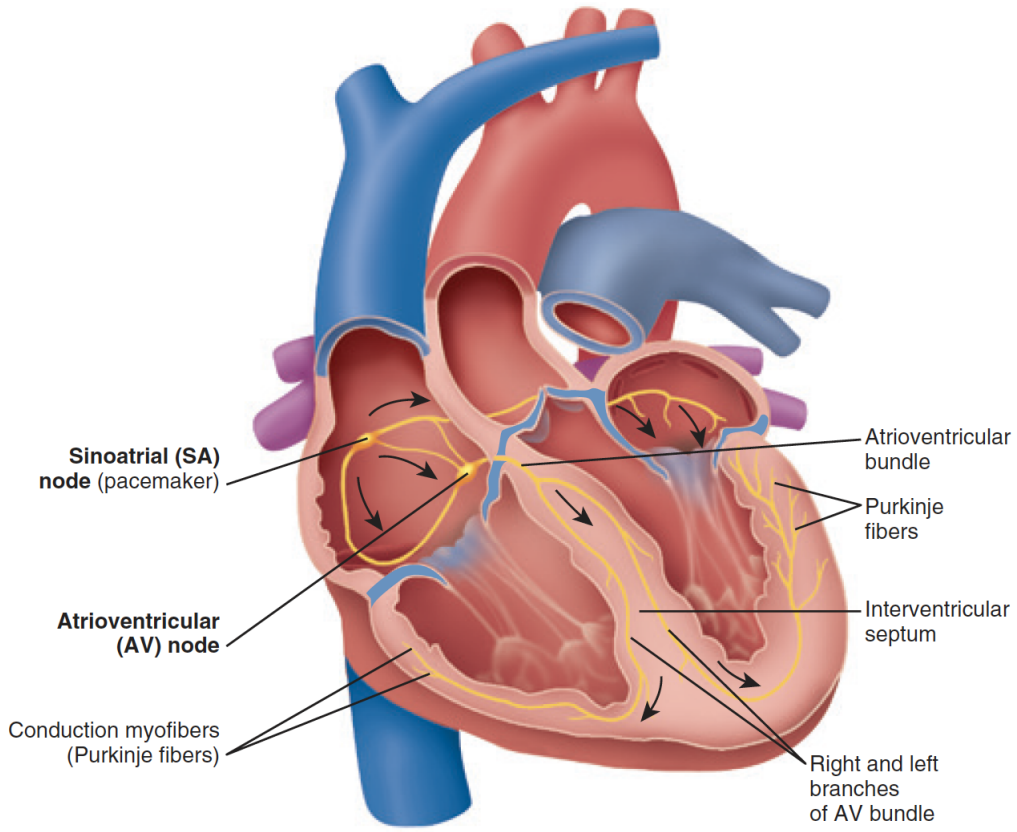


Figure 1.2: Sectional View of the Heart: the conduction system components [14]

1.3. ECG Signal

ECG machines are used to acquire electrical activities of the heart as observed from the electrodes attached to a patient's arms, legs, and chest, as shown in Figure 1.3. The electrical signals picked by these electrodes are associated with a 12-lead ECG machine that records the aggregate electrical activity of the heart from distinct angles over some time, commonly 12 seconds [15]. Besides, a 12-lead ECG is a standard technique in diagnosing various heart diseases [16]. Among the 12 leads, the three bipolar leads (I, II, and III) measure the potential differences between both arms, and one arm and the leg [14]. The six chest leads (V1 to V6) are unipolar. The remaining leads, aVR, aVL, and aVF, are augmented limb leads derived from standard limb leads (I, II, and III). The six chest leads (V1 to V6) view the heart's electrical activity in the horizontal plane, whereas the six limb leads (I, II, III, aVR, aVL, and aVF) view the heart's electrical activity in the vertical plane [2, 17]. A standard ECG record of a patient is shown in Figure 1.4.

A single cycle of an ECG contains a pattern of waves, as shown in Figure 1.5. When the sinoatrial (SA) node triggers an impulse, the atrial fibers depolarize to produce a potential

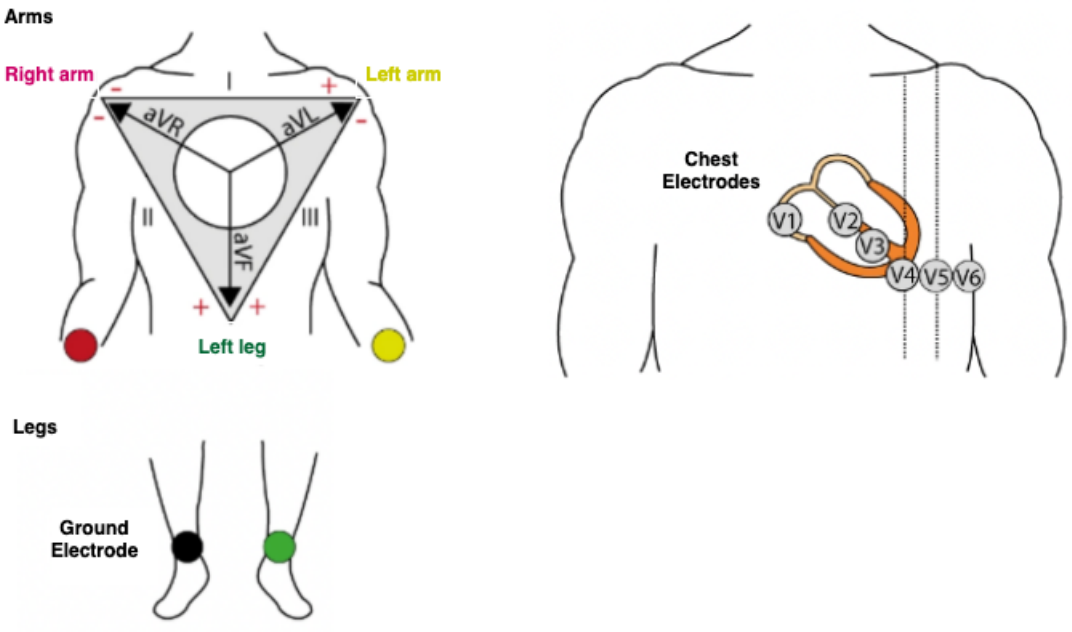


Figure 1.3: The placement of ECG electrodes on the chest, arms, and legs [18].

difference called a P-wave, leading to atrial contraction. In a normal ECG, as shown in Figure 1.5, a P-wave has a duration of about 0.08 seconds [14]. A P-wave is prominently seen in leads II and V1. Moreover, it leans inverted in the lead aVR and is upright in leads I and II, as shown in Figure 1.4.

After the atrial fiber depolarization, the impulse reaches the ventricular fibers and rapidly depolarizes them. Since the ventricular walls are thick, the depolarization results in more electrical changes; it is called the QRS-complex, which consists of Q, R, and S waves. The QRS-complex also lasts for about 0.08 seconds [14]. Then, as the ventricles repolarize, a T-wave is produced. The T-wave is about 0.16 seconds in a normal ECG. It can be seen from Figure 1.5 that the atrial repolarization is missing from the pattern due to atrial fiber repolarization at the same time as ventricular fiber depolarization [14].

As shown in Figure 1.5, the PR interval is the period between the P-wave and the QRS-complex. The PR interval indicates the impulse transmission times between the SA and atrioventricular (AV) nodes. It contains atrial depolarization, contraction, and depolarization waves via the conduction system. The ST segment, on the other hand, occurs during the depolarization of the ventricular myocardium, and it lasts about 0.22 seconds. The QT interval that lasts about 0.38 seconds is a period from the start of ventricular depolarization to repolarization [14]. The TP segment is an isoelectric region that indicates the absence of a substantial amount of potential difference in the ventricular myocardial cells. It is a rest-

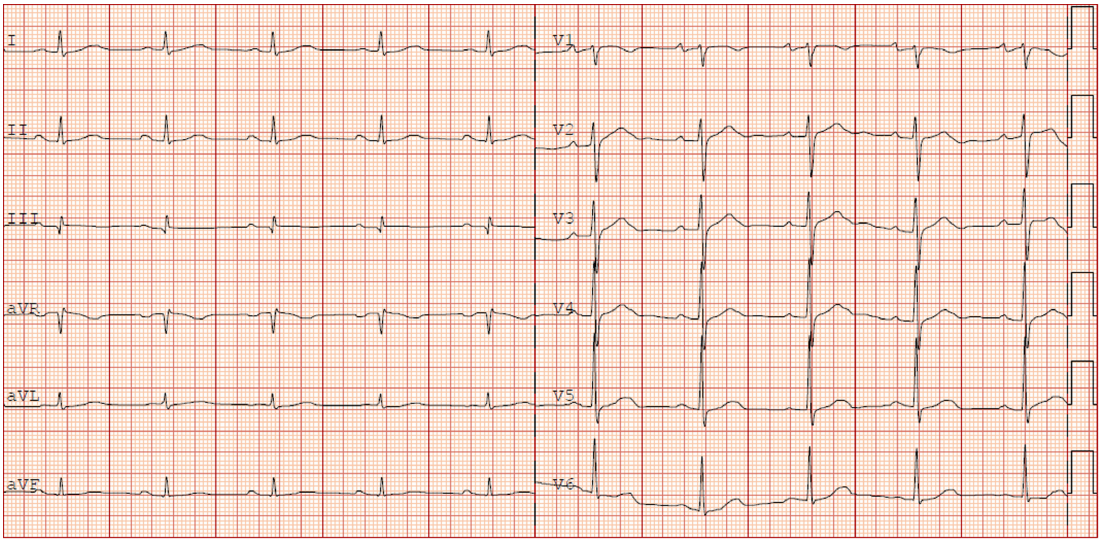


Figure 1.4: A standard 12-lead ECG of a single patient [17].

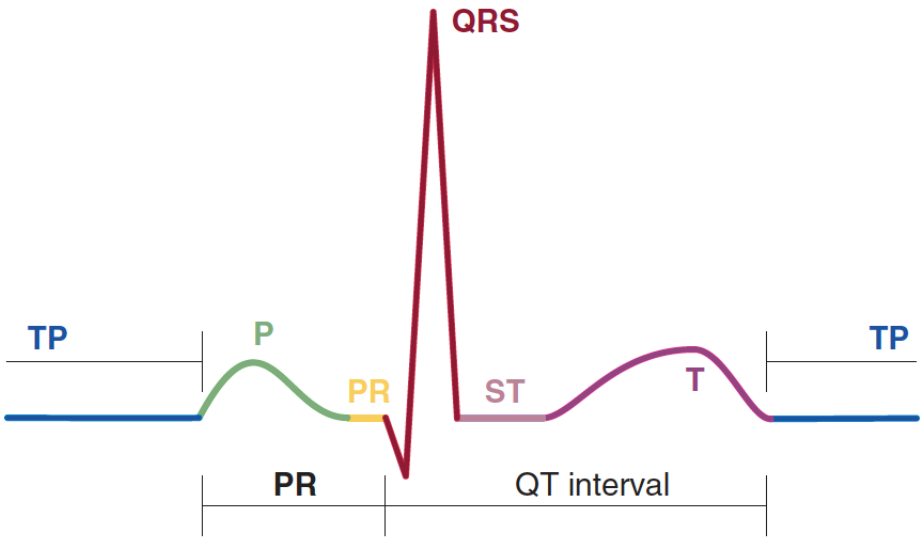


Figure 1.5: A single cardiac cycle of the ECG pattern [14].

ing state of the ventricular myocardial cell and covers a time from the end of repolarization to the onset of the next depolarization [19]. Any deviation from this normal cardiac cycle may indicate heart disease and conduction system problems. As shown in Figure 1.6, for instance, a QRS duration greater than 0.12 seconds, broad monophasic R waves in leads I, V5, and V6, and the absence of Q waves in leads V5 and V6 are indications of the left bundle branch block (LBBB) [2].

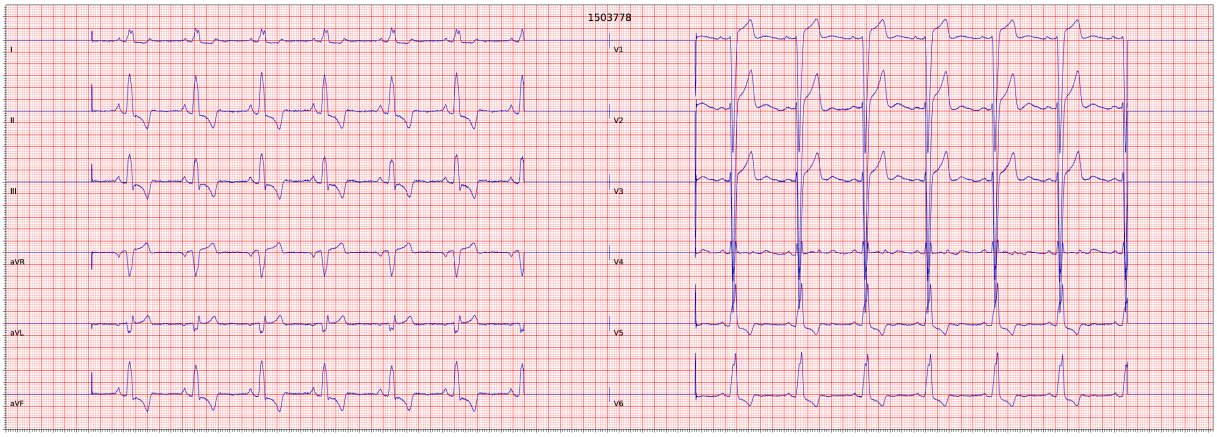


Figure 1.6: A 12-lead ECG of a patient with exam_id of 1503778 diagnosed for LBBB [20].

1.4. Machine Learning: In an ECG Signal Classification Prescriptive

Recently, several studies have examined the possibility of artificial intelligence (AI) techniques in interpreting an ECG in the diagnosis of cardiovascular diseases [20–30]. In addition, a review article written by Liu *et al.* [31] provided a detailed review of deep learning techniques used for ECG diagnosis. Majority of the literary studies focus on identifying small types of heart abnormalities from among several types of heart disease [20, 32]. Moreover, some of the literary studies only focus on normal and abnormal ECG signal classes from a single lead ECG signal [25, 28]. However, ML-based methods for heart disease detection and classification from an ECG signal have shown promising results and are an active area of research. Several studies have reported that ML-based ECG interpretation algorithms perform better at approximating human experts compared to existing computer-assisted interpretation (CIE) techniques [33].

However, the complexity and black-box nature of ML models have hindered medical practitioners from having confidence in the diagnostic results of these models [34]. To address this, ML model interpretation techniques have been introduced to explain how the model arrived at a particular classification decision [34]. These interpretation techniques enable human experts to trust the model's output, debug and troubleshoot the model, and avoid model bias [35]. However, the field is still developing, and researchers are concentrating on developing techniques that can explain the model's reasoning behind detecting or classifying abnormalities in healthcare settings [34] and other applications [35]. So, finding a better

interpretable heart disease classifier is an active research area [36].

1.5. Statement of the Problem

The electrocardiogram (ECG) is a widely used, non-invasive tool for assessing a heart condition. But, diagnosing underlying conditions from ECG signals remains challenging for medical practitioners. While machine learning models have shown promise in improving ECG interpretation, they are computationally complex and black-box. The computational complexity is mainly associated with the model's parameter size. Additionally, the black-box nature of deep learning models makes it difficult to interpret their classification output. This limitation hinders our understanding of how models arrive at their decisions. This situation makes it difficult for medical practitioners to trust the classification output of the model.

Therefore, this Ph.D. study aims to address the following research questions:

- What are the key limitations of existing 12-lead ECG-based heart disease diagnosis systems, and what strategies can be implemented to improve their accuracy and reliability?
- What types of noise commonly affect ECG signals, and what preprocessing techniques can be applied to enhance signal quality for deep learning models?
- Can a robust and optimized deep learning model be developed to achieve high generalizability across diverse ECG datasets?
- What IML methods can enhance the interpretability and clinical relevance of classification outputs generated by deep learning models for ECG signals?

1.6. Objective

1.6.1 General Objective

The main objective of this Ph.D. research is to design and implement a robust and interpretable deep learning algorithm that is capable of operating in real-life conditions for the automatic classification of heart diseases from a 12-lead ECG signal.

1.6.2 Specific Objectives

The general objective is associated with the following specific objectives:

- To perform a literature survey to identify gaps in developing robust deep learning (DL) systems for accurately classifying various heart diseases using real-world 12-lead ECG datasets.
- To collect, analyze, and preprocess 12-lead ECG datasets.
- To design and implement a robust deep learning-based ECG signal classification algorithm.
- To evaluate interpretable machine learning techniques for explaining classification outputs in a 12-lead ECG-based model.
- To evaluate the proposed model's computational complexity regarding trainable parameters, and ensure it remains lightweight for deployment.
- To evaluate the effectiveness and robustness of the proposed model using three different 12-lead ECG datasets.
- Finally, deploy the system and conduct pilot testing in hospital.

1.7. Significance of the Dissertation

The significance of developing a robust deep learning (DL)-based system for heart disease classification from 12-lead ECG signals is underscored by the current challenges clinicians face in accurately diagnosing heart conditions. As highlighted by studies, there is a considerable variance in physicians' ability to interpret ECG signals accurately. This underscores the need for advanced computational tools to support and enhance diagnostic accuracy.

A robust DL-based system has the potential to address these diagnostic challenges by identifying heart diseases from 12-lead ECG signals more effectively. Deep learning models can be trained to recognize complex patterns and anomalies in ECG data that might be difficult for human experts to easily identify. By integrating the system to existing medical practices, it can provide consistent, reliable, and objective classifications of heart diseases.

1.8. Contributions of the Dissertation

The main strengths and contributions of the study presented in this article include:

- The design and evaluation of a heart disease diagnosis system based on a 12-lead ECG signal that incorporates three components: preprocessing, design a hybrid DL model, and visual interpretation of the model's classification output.
- The classification performance of the proposed hybrid model was tested on three different ECG datasets and compared with state-of-the-art methods.
- The proposed hybrid model has shown robustness with high generalizability across the three datasets.
- Systematically reviewing and analyzing IML techniques for ECG-based heart disease classification.
- Investigated and showcased the visual interpretation of the model's classification output using Grad-CAM++ and SHAP techniques.

1.9. Publications

As a result of this research the following research outputs were published in peer-reviewed journals containing parts of the results.

- Ayano YM, Schwenker F, Dufera BD, Debelee TG. Interpretable Machine Learning Techniques in ECG-Based Heart Disease Classification: A Systematic Review. Diagnostics. 2023; 13(1):111. https://doi.org/10.3390/diagnostics13010111
- Y. M. Ayano, F. Schwenker, B. D. Dufera, T. G. Debelee and Y. G. Ejegu, "Interpretable Hybrid Multichannel Deep Learning Model for Heart Disease Classification Using 12-Lead ECG Signal," in IEEE Access, vol. 12, pp. 94055-94080, 2024, doi: 10.1109/ACCESS.2024.3421641.

1.10. Organization of the Dissertation

This dissertation is structured into six chapters, each addressing a critical aspect of the research. Chapter 1 introduces the study, outlining the problem statement, significance, and contributions of this dissertation. A comprehensive review of the relevant literature is provided in Chapter 2, establishing the foundation for the research. Chapter 3 details the research methodology, including an in-depth explanation of the proposed model and its overall workflow. The experimental results, accompanied by a thorough discussion of the findings, are presented in Chapter 4. Building upon the analysis in Chapter 4, Chapter 5 explores the interpretability of the model's classification outputs, emphasizing insights derived from the study. Finally, Chapter 6 concludes the dissertation by summarizing the research findings and proposing future directions.

CHAPTER 2

Literature Review

This chapter provides an overview of ML techniques proposed in the literature for an ECG-based heart disease diagnosis, starting with classical machine learning methods. We also examine the deep learning techniques proposed for complex and large 12-lead ECG datasets with large class labels. Additionally, this section discusses the importance of interpretability in ML models, emphasizing the need to understand the decision-making processes of these models to ensure evidence-based diagnoses. In addition, it discusses interpretability techniques proposed in the literature to improve the understanding of why an ML model classifies an ECG signal into a particular diagnostic class. Despite advances in ML methods, significant gaps remain in the literature, particularly regarding the integration of interpretability with deep learning models and the practical application of these techniques to diverse ECG datasets. These gaps will be highlighted and discussed.

2.1. Classical Machine Learning in ECG

Classical machine learning models, such as decision trees, random forests, support vector machines, and multi-layer perceptrons (MLP), rely on hand-crafted features for training [37]. They have demonstrated acceptable performance in binary classification tasks, such as distinguishing between normal and abnormal ECGs or detecting a single disease in heart disease classification [38]. These techniques heavily depend on the ECG data quality and the relevance of hand-crafted features. The reliance on manual feature engineering is time-consuming and often misses crucial clinical insights in complex and noise-prone signals such as ECG. Besides, classical machine learning models do not scale well to complex and large datasets with several class labels [39]. This claim has been proved by Bickmann

et al. [40], where XGBoost was trained on extracted features from the ECG signals. The classification result of the proposed model on one of the large ECG datasets, PTB-XL 12-lead ECG dataset [41], yielded a relatively modest classification accuracy of 70.90%. Apart from the challenge in feature extraction, XGBoost, and other classical ML techniques do not inherently capture temporal dependencies in time-series data. As a result, there is a decline in focus on using these techniques for heart disease classification from an ECG [40].

2.2. Deep Learning in ECG

Unlike classical ML methods that heavily rely on hand-crafted features, deep learning methods do not require manual feature extraction techniques. Deep learning methods such as convolutional neural networks (CNN), recurrent neural networks (RNN), Long Short-Term Memory (LSTM), Bi-Directional Long Short-Term Memory (Bi-LSTM), attention mechanisms, and hybrid models used in an ECG-based heart disease classification through mapping the ECG signal to its class-label in an end-to-end manner. This section discusses the results achieved by deep learning (DL) models in 12-lead ECG signal classification. It highlights the successes of these models while also discussing their limitations.

2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are deep neural networks designed to automatically and adaptively learn spatial hierarchies of features through back-propagation [42]. CNNs utilize multiple building blocks, including convolution layers, pooling layers, activation functions, batch normalization, and fully connected layers. The convolution and pooling layers perform feature extraction from the ECG signals, while activation functions introduce non-linearity, and batch normalization stabilizes and speeds up the training process. The fully connected (FC) layers then map the extracted features to the output class labels, facilitating accurate classification of various cardiac conditions [43].

Apart from various architectural design concepts in convolutional neural networks, 1D and 2D CNNs have been extensively experimented with in ECG signal analysis based on the data input format and the dimensions of filters/kernels [20, 28, 30, 43–55]. Many of these proposed techniques [48, 49, 51–56] were tested on the MIT-BIH arrhythmia database [57, 58], containing very few ECG data with beat- and rhythm-level annotations. However, these

annotations do not apply to clinical-level 12-lead ECG-based heart disease diagnosis [59]. Single-lead and beat-wise annotations are often unreliable for detecting most critical types of heart diseases, such as myocardial infarction (MI) [60]. Moreover, the beat-wise annotation process is costly and labor-intensive, making it less practical. There are limited approaches that utilize CNNs for classifying 12-lead ECGs that can be applied to the clinical level and various types of heart diseases beyond just rhythm disorders. The recently published research works [20,43,44,47,50,56], have used 12-lead ECG datasets, namely, PTB-XL [41], CODE-15% [61] and Chapman Arrhythmia dataset [62]. These datasets contain diverse cardiac conditions with patient-level annotations, enhancing their suitability for developing DL models with practical clinical applications. The following sections briefly present these approaches.

Pałczyński *et al.* [44] investigated few-shot learning methods for training a deep convolutional neural network to classify 2, 5, and 20 cardiac disease classes from the PTB-XL dataset. While the proposed network achieved high accuracy in classifying 2 classes, results showed lower accuracy for 5 and 20 class classifications, with maximum accuracies of 80.20% and 70.0%, respectively. Similarly, Sandra Śmigiel *et al.* [47] proposed a technique that uses features from a 5-layer 1D-CNN architecture and entropy features extracted from each channel of the 12-lead ECG signal. The model has achieved an accuracy of 89.2%, 76.5%, and 69.8%, with F1-scores of 89.1%, 68%, and 33.2% for classifying 2, 5, and 20 classes of heart diseases from PTB-XL dataset, respectively. On the other hand, Strodthoff *et al.* [50] evaluated multiple pre-trained DL algorithms, finding that ”resnet1dwang” achieved the highest performance, with an area under the curve (AUC) of 93.0% for classifying the 5 super-diagnostic classes. Similarly, Anand *et al.* [43] proposed a 2D-CNN for classifying 5 super-diagnostic classes of PTB-XL datasets. The proposed model has less number of parameters and achieved an accuracy of 89.73% and an AUC of 93.41%. In addition, Anand *et al.* [43] tested their proposed architecture using merged 4 classes of the Chapman Arrhythmia dataset through fine-tuning the initial hyperparameters and achieved an average accuracy of 95.8% and an AUC of 99.46%.

In another study, the performance of pre-trained CNN models in classifying heart diseases in the PTB-XL dataset was evaluated by Strodthoff *et al.* [50]. These models include 1D versions of Inception [63] (referred to as ”inception1d”), a modified ResNet [64] (”xre-

sent1d101"), and a modified Full convolutional Network (FCN) [65] ("fcn_wang"). The classification performance of these models, namely "inception1d", "xresent1d101", and "fcn_wang" on five (5) super diagnostic classes in terms of macro AUC are 92.10%, 92.8%, and 92.5%, respectively. Among the assessed pre-trained models, the modified version of ResNet (referred to as "resnet1dwang") achieved an AUC of 93.0% in the classification of five (5) super-diagnostic classes. Similarly, Narotamo *et al.* [56] compares various deep learning methods for ECG classification, focusing on different ECG signal representations (1D and 2D) and multimodal fusion approaches. They used the PTB-XL dataset and transformed the signals to 2D-image representation using Gramian angular field, recurrence plot, and Marko transition field to train AlexNet, VGG16, ResNet50, MobileNetV2, and AlexNetAttenention. The AlexNetAttention model, with a total of 46,335,880 trainable parameters, achieved the highest specificity of 81.98%, though it did not attain the best overall performance. The study also explores multimodal fusion strategies (early, late, and joint) fusion to combine 1D and 2D representations and improve classification performance. The late fusion resulted in the classification accuracy of 79.22%. However, the 1D approach with GRU achieved a better accuracy of 80.69%.

On the other hand, Ribeiro *et al.* [20] proposed a residual neural network-based model to classify six cardiac abnormalities, achieving an average F1-score of over 80%. The datasets used in this study were sourced from the Telehealth Network of Minas Gerais (TNMG) as part of the Clinical Outcomes in Digital Electrocardiography (CODE) study. The proposed model features over 6 million trainable parameters. This increases computational complexity, requiring significant processing power and memory. Besides, the complexity may result in a risk of overfitting, especially when trained on a dataset with a limited amount of data.

The above discussion shows the potential of CNN-based techniques in classifying cardiac conditions by capturing spatial features embedded in ECG signals. However, due to their fixed receptive field size, CNNs may struggle with processing lengthy sequences, potentially overlooking long-term dependencies crucial for accurate ECG interpretation [66]. Additionally, using 2D CNNs on transformed ECG signals introduces computational complexities and may not improve performance [56, 67, 68].

2.2.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of deep neural network suited for analyzing 1D signals that are sequential by nature, like an ECG. RNNs process the ECG data one step at a time, allowing the network to capture the temporal dependencies in the signal [69]. By maintaining an internal state, RNNs can capture information from previous time steps, effectively detecting patterns over time. However, to overcome their short-term memory, the vanishing and exploding gradient problem in basic RNNs, modified architectures such as long short-term memory (LSTM), bidirectional LSTM (Bi-LSTM), gated recurrent unit (GRU), and bidirectional GRU (Bi-GRU) have been proposed [69–71]. These architectures can model the sequential nature of ECG signals in the classification of heart diseases [56, 60, 72–77].

Kuila *et al.* [72] proposed a model that integrates RNN to classify an arrhythmia from MIT-BIH arrhythmia database [57, 58]. The MIT-BIH arrhythmia database, which includes recordings from two leads, is often used to evaluate ECG classification techniques in the context of wearable devices that typically operate with single-lead configurations. In addition, it covers only a limited range of arrhythmias and does not fully represent the broader spectrum of cardiac conditions encountered in clinical practice [87]. The method proposed by Kuila *et al.* [72] aids in handling sequential ECG data, capturing temporal dependencies critical for accurate arrhythmia detection with a classification accuracy of 96.41%.

On the other hand, Wang *et al.* [73] used a dual-path RNN to detect atrial fibrillation (AFIB) from a single-lead ECG signal with a classification accuracy of 84.5%. In another study, Boda *et al.* [76] compared a basic RNN, LSTM, and GRU in the classification of arrhythmia from MIT-BIH arrhythmia database [57, 58]. They focused on developing a framework for classifying the ECG beats into different categories, focusing on ventricular ectopic beats (VEB) and supra-ventricular ectopic beats (SVEB) on a patient-specific basis. Accordingly, the LSTM-based model achieved the highest performance with a sensitivity of 92.4% and 77.8% in classifying the VEB and SVEB, respectively. Similarly, Çınar *et al.* [77] evaluated an LSTM model in classifying MIT-BIH arrhythmia into normal, arrhythmia, and congestive heart failure, achieving an accuracy of 90.67%. In a related study, Hiriyannaiah *et al.* [71] focused on LSTM networks for ECG analysis and heartbeat classification from the MIT-BIH arrhythmia dataset. Their methodology involved using LSTM-based neural networks and

comparing their performance across different metrics. Among the evaluated LSTM architectures, the Bi-LSTM model outperformed the baselines across all metrics, demonstrating its superior capability in capturing dependencies and classifying heartbeats accurately.

Other studies have proposed RNN-based models for classifying heart diseases from 12-lead ECG signals. For example, Narotamo *et al.* [56] tested various modified RNN variants, including LSTM, Bi-LSTM, GRU, and Bi-GRU, on the PTB-XL dataset. Among these, the GRU-based model demonstrated the best performance, achieving an accuracy of 80.69% and a specificity of 81.04%. On the other hand, Zhang *et al.* [60] proposed Bi-LSTM network for detecting only myocardial infarction (MI) from 12-lead ECG signal of the PTB diagnostic ECG dataset [78]. The proposed method achieved a good performance with MI detection accuracy of 94.77%.

The RNN networks primarily focus on temporal dependencies and are not inherently equipped to handle spatial relationships between different leads in an ECG. So, hybridizing them with other networks like CNNs, can help in effectively analyzing both the spatial and temporal aspects of ECG data.

2.2.3 Hybrid Techniques

Hybrid approaches, which integrate different architectures or combine DL models, offer improved accuracy and enhanced interpretability. Besides, they are also better suited for limited-size datasets [79]. The hybridization allows the model to utilize the strengths of each technique; for instance, in CNN-LSTM hybrid architecture, the CNN architectures are good at extracting spatial features, and LSTM architectures are good at capturing sequential dependencies in a time series datasets like an ECG [80]. The benefits of hybrid techniques in ECG based heart disease classification have been demonstrated in the literature [62, 80–85].

Yang *et al.* [81] proposed a model that combines ResNet and GRU to classify the 12-lead PTB-XL ECG dataset. This hybridization enables the model to handle long-term dependencies and improved the super-diagnostic class classification accuracy to 88.4%. Similarly, Jing *et al.* [82] used the 12-lead PTB-XL dataset to evaluate their CNN-GRU hybridized model. By integrating convolutional and recurrent layers, this model captures both spatial and temporal features of ECG signals, leading to an average accuracy of 88.19% and macro F1-score 70.55%. Besides, the total number of trainable parameters of the model is 569,373,

which is a modest amount compared other architectures in this category.

Yildirim *et al.* [80] proposed a CNN-LSTM hybrid model with 2.9 million (M) parameters to classify reduced seven and merged four classes of arrhythmia from Chapman Arrhythmia 12-lead ECG dataset [62]. The model demonstrated an acceptable generalization ability to classify four classes of arrhythmias with a classification accuracy of 96.13%. Similarly, Lai *et al.* [83] proposed an architecture that integrates a Residual Convolutional Neural Network (ResCNN) with an LSTM layer to detect arrhythmias. This combination of ResCNN and LSTM networks provides complementary advantages, achieving efficient feature extraction and processing of ECG signals with variable lengths, with classification performance F1-score of 76.9% on the China Physiological Signal Challenge 2018 (CPSC 2018) dataset [86]. In another study, Xie *et al.* [84] proposed a model architecture that consists of multiple BranchNets, incorporating convolutional layers for spatial feature extraction, bidirectional Long Short-Term Memory (BiLSTM) layers for capturing temporal dependencies. The method achieved an F1-score of 81.02% and an accuracy of 74.84% on Computing in Cardiology Challenge 2020 (CinC2020) dataset, and an F1-score of 74.90% and an accuracy of 55.37% on Shandong Provincial Hospital (SPH) dataset [84]. Moreover, Chen *et al.* [85] proposed a model that combines 1-D ResNet-34 and an LSTM. The achitecture was evaluated on a 12-lead ECG dataset collected by the authors and achieved a classification accuracy of 81% in distinguishing 6 disease classes.

Table 2.1 summarizes the models proposed for 12-lead ECG datasets. Unlike heartbeat-level annotated datasets such as MIT-BIH, these models are designed for 12-lead ECG analysis in clinical settings. The results presented in Table 2.1 show the promise to incorporate DL models in diagnosing cardiac abnormalities. However, these methods have limitations. Firstly, some are highly complex with parameters counted in the millions [20, 50, 80]. Conversely, methods proposed in [44, 47] are lightweight architectures but exhibit lower performance levels. Secondly, these models are not well tested on different datasets and disease classes; as a result, it is not easy to conclude their applicability across a broader range of heart diseases after re-training. Thirdly, except [43], there is a limitation in integrating model output interpretation techniques, which are essential for providing evidence-based diagnoses and enhancing physician trust in using these models. So, these limitations motivate us to develop a robust interpretable model with minimal trainable parameters.

Table 2.1: A summary of 12-lead ECG based heart disease classification methods.

Author	Method	Dataset	Performance (%) [accuracy, macro F1-score]	Limitations
Ribeiro *et al.* [20]	DNN	CODE	[—, 80%]	• Interpretability methods were not incorporated. • The model has a parameter size exceeding 6 million.
Feyissa *et al.* [30]	1D-CNN	PTB-XL [super-diagnostics]	[89.70%, 72.0%]	• Heart disease classes with fewer than 20 samples were excluded, which may limit the model's ability to generalize to rare conditions. • Robustness was not evaluated across various datasets, limiting insights into its generalizability. • The study lacks interpretability methods.
Anand *et al.* [43]	2D-CNN	PTB-XL [super-diagnostics]	[89.75%, —]	• Requires extensive hyperparameter fine-tuning for different ECG datasets. • Interpretability analysis was limited to SHAP without test-set level interpretability.
		Chapman-Merged	[95.85%, 95.39%]	

Table 2.1: *Cont.*

Author	Method	Dataset	Performance (%) [accuracy, macro F1-score]	Limitations
Pałczyński *et al.* [44]	1D-CNN	PTB-XL [super-diagnostics]	[79.0%, 70.6%]	• Exhibits low performance. • The study lacks interpretability methods. • Robustness of the models was not tested on different datasets.
Demissie *et al.* [45]	ResNet-18	PTB-XL [a subset from rhythm category]	[96.0%, 88.0%]	• Only a subset of five ECG records from the rhythm category was used for model development. • Grad-CAM and Grad-CAM++ were used for instance-level interpretability analysis, but test-set level interpretability was not incorporated. • ResNet-18 has a large model size with millions of parameters.
Śmigiel *et al.* [47]	1D-CNN with entropy features	PTB-XL [super-diagnostics]	[73.0%, 60.0%]	• Interpretability methods were not incorporated. • The model's robustness was not assessed by evaluating its performance on different datasets. • Exhibits low performance.

Table 2.1: *Cont.*

Author	Method	Dataset	Performance (%) [accuracy, macro F1-score]	Limitations
Narotamo *et al.* [56]	1D-GRU	PTB-XL [super-diagnostics]	[80.69%, —]	• Interpretability methods were not incorporated. • The model's robustness was not assessed by evaluating its performance on different datasets. • Exhibits low performance.
Zhang *et al.* [60]	Bi-LSTM	PTB diagnostic ECG database	[94.77%, —]	• Only 369 MI and 79 normal ECG signals were used for model development. • Interpretability methods were not incorporated.
Yildirim *et al.* [80]	CNN-BiLSTM	Chapman-Reduced	[92.24%, 80.04%]	• The model's robustness was not evaluated across different datasets. • Interpretability methods were not incorporated.
		Chapman-Merged	[96.13%, 95.57%]	
Jing *et al.* [82]	CNN–GRU	PTB-XL [super-diagnostics]	[88.19%, 70.55%]	• The model's robustness was not assessed by evaluating its performance on different datasets. • Interpretability methods were not incorporated.

Table 2.1: *Cont.*

Author	Method	Dataset	Performance (%) [accuracy, macro F1-score]	Limitations
Yang *et al.* [81]	ResNet–GRU–Attention (ResGRu-Attention)	PTB-XL [super-diagnostics]	[88.04%, 75.90%]	• ResNet has a large model size with millions of parameters. • The model's robustness was not assessed by evaluating its performance on different datasets. • Interpretability methods were not incorporated.
Lai *et al.* [83]	ResCNN–LSTM	CPSC 2018 [rhythm and morphology abnormality]	[—, 76.9%]	• The model's robustness was not assessed by evaluating its performance on different datasets. • Interpretability methods were not incorporated.
Xie *et al.* [84]	BranchNets	CinC2020	[55.37%, 74.90%]	• Interpretability methods were not incorporated. • The model's performance is modest.
		SPH	[74.84%, 81.02%]	
Chen *et al.* [85]	1D ResNet-34 – LSTM	12-lead ECG dataset collected by authors	[81%, —%]	• The model's robustness was not assessed by evaluating its performance on different datasets. • Interpretability methods were not incorporated. • ResNet-34 has a large model size with millions of parameters.

2.3. Interpretable Machine Learning Techniques

The need to determine the rationale behind the output decisions of the ML models began in the 1970s [88]. However, considerable advancements in the field of IML have been attained in the last few years. Nevertheless, its conceptual foundation is still underdeveloped [89]. Currently, there is no well-established mathematical definition for the interpretability of ML models. It can also be called explainable artificial intelligence (XAI), and there is no well-agreed definition [90]. However, Murdoch et al. [91] defined the focus of an IML as "... the extraction of knowledge from an ML model concerning relationships either contained in data or learned by the model ...". According to their definition, knowledge is relevant if it provides insight for a particular audience in a given context. Based on the problems to be solved and users that use the output of an IML, this insight can be in the form of visual presentation, human-understandable languages, or mathematical equations.

IML techniques proposed in the literature for explaining black box ML models attempted to localize segments of an ECG signal that the ML used for output prediction. In addition, the performances of these IML techniques were not measured against ground truth, partially because of the unavailability of the annotated dataset and commonly agreed-on quantitative metrics. As far as we know, there are no publicly available ECG heart disease datasets that include the clinical descriptions necessary for categorizing the ECG tracings into their respective disease classes. This sub-section presents a comprehensive study of IML methods in the context of ECG-based heart disease diagnosis. It highlights the challenges in accurately interpreting ECGs and the need for understandable ML models. In addition, it discusses the use of class activation maps and their variants for localizing the segments and leads of an ECG signal that contributes most to the classification output, as well as feature relevance-based explanation techniques like SHAP. Besides, it outlines the advantages and limitations of various IML techniques. It also addresses the challenges in designing and evaluating these methods. Finally, it concludes by underlining the potential of these techniques to offer evidence-based diagnoses and bring trust in ML models.

2.3.1 Taxonomy of IML

Various explanation techniques have been proposed in the literature to explain the output and behavior of machine learning models. Based on discussions in the literature [90, 92–96], in this research, we propose a taxonomy for IML techniques as shown in Figure 2.1. Here, the classification of IML techniques is based on their interpretation result presentation, scope, model specificity of the method, and the complexity of the ML model. However, the IML technique can hold a place in more than one of the classes in taxonomy. In the following sections, we will provide a detailed explanation based on the taxonomy shown in Figure 2.1. In addition, the main concepts behind IML techniques and their usage for an ECG signal-based heart disease diagnosis are subsequently discussed.

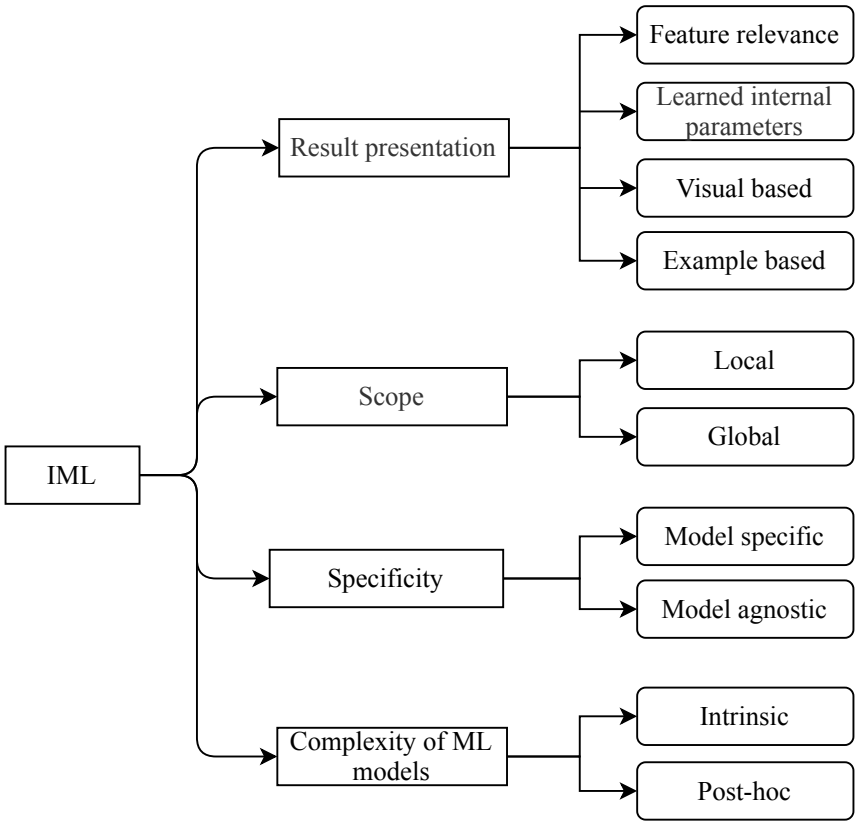


Figure 2.1: Taxonomy of machine learning interpretability.

2.3.2 Result Presentation in IML

In IML, results from interpretation methods can be presented in several ways to provide insightful information to the user. These methods include displaying feature relevance, the model's internal parameters, visual-based explanations, and example-based explanations.

2.3.2.1 Feature Relevance

Feature relevance-based model explanation is a technique used to interpret an ML model's output after the training process. This technique assigns a score to each feature, indicating its contribution to the prediction output of a trained model [90, 94]. Mathematically, this contribution can be quantified by analyzing the input/output behavior of the model. In feature relevance-based explanations, the contribution of each input feature, $\mathbf{x} = (x_1, \ldots, x_m)$, is measured to determine its impact on the model's output, $\mathscr{M}(x_1, \ldots, x_m)$. Several methods utilize feature relevance to explain AI models. This sub-section briefly discusses Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and permutation feature importance.

SHapley Additive exPlanations (SHAP) SHapley Additive exPlanations (SHAP) are derived from game theory, where SHapley values represent the marginal contribution of each player to the overall team effort. In the context of interpreting machine learning models, SHapley values indicate the contribution of each feature to the prediction or classification output of a given black-box model. To determine feature importance in model output prediction or classification, SHapley values can be calculated based on the complexity of the ML model. Various techniques for calculating SHapley values include linear SHAP, kernel SHAP, and deep SHAP [97, 98].

The linear SHAP method explains feature importance in linear ML models. Let $S \subseteq F$, where S is a subset of all features $F = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_M\}$, with $\mathbf{x}_k$ representing the feature in the k^{th} column of a dataset of size $N \times M$. The contribution of a feature x_i to the model output, $\mathscr{M}$, is assessed in two steps. First, the model is trained with the feature x_i included, resulting in a model represented as $\mathscr{M}_{S\cup\{i\}}$. Then, the model is retrained without the feature x_i, resulting in $\mathscr{M}_S$. To evaluate the contribution of feature x_i, the prediction from the two models are compared on the current input $\mathscr{M}_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})$ and $\mathscr{M}_S(\mathbf{x}_S)$. The SHapley value, ϕ_i, for the feature x_i is then calculated using Eqn. (2.1) [97]:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [\mathscr{M}_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - \mathscr{M}_S(\mathbf{x}_S)] \tag{2.1}$$

where $\mathbf{x}_S$ represent the input feature values in a set S, $\mathscr{M}_S(\mathbf{x}_S)$ represents the marginal value of $\mathscr{M}$ for the features present in S, and $\mathscr{M}_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}})$ denotes the marginal value of $\mathscr{M}$ for

the feature values present in S plus feature x_i. Thus, Eqn. (2.1) computes the disparity over all possible subsets $S \subseteq F \backslash \{i\}$ weighed by the number of features in the S from the total number of features, F.

While the interpretation derived from SHapley values for features can be understood and thoroughly tested in the context of models that used in ECG based heart disease classification [99–102], the SHapley technique still has limitations. The primary challenge is the significant computational burden of calculating SHapley values for all feature subsets, as the complexity increases exponentially [103].

Though the interpretation obtained from the SHapely values of the features can be comprehended and thoroughly tested for interpreting ECG-based ML models [99–102], the SHapley technique still has limitations. The major challenge is the computational burden associated with calculating SHapley values for all feature subsets where the computational complexity is exponential [103]. However, to mitigate these limitations, techniques, such as restricting the subset permutation using the causal relationship of features [104] and incorporating the constraint of correlations among feature values [105, 106] have been proposed. Moreover, to overcome the computational expensiveness of Eqn. (2.1), kernel SHAP [103], tree-SHAP [107], and SHAP with gradient explainer [108] have been introduced.

Local Interpretable Model-Agnostic Explanations (LIME) LIME, introduced by Ribeiro et al. [109], approximates complex non-linear ML models using a locally interpretable surrogate model. This approach helps explain which features contribute most to the output of the black-box ML model. However, LIME operates under the assumption that complex models can be approximated as linear on a local scale. Mathematically, this is captured by the penalty function $\lambda_{\mathbf{x}}(\mathbf{z})$, which measures the proximity between perturbed instances, $\mathbf{z} \in \mathbb{R}$ and the original instance feature $\mathbf{x}$. Thus, given $\mathscr{M}$, a black box ML model to be explained, and g being a surrogate model best approximates $\mathscr{M}$ among a class of potential interpretable models G, that is, $g \in G$. The explanation $\xi(\mathbf{x})$ for an instance feature vector $\mathbf{x}$ produced by LIME is obtained by minimizing the objective function $\mathscr{L}(\mathscr{M}, g, \lambda_{\mathbf{x}}) + \Omega(g)$, as given in Eqn. (2.2) [109]:

$$\xi(\mathbf{x}) = \underset{g \in G}{\operatorname{argmax}} \left(\mathscr{L}(\mathscr{M}, g, \lambda_{\mathbf{x}}) + \Omega(g) \right) \tag{2.2}$$

where $\mathscr{L}$ is a locality-aware loss function for measuring how g is unfaithful in closely resembling $\mathscr{M}$ in the locality defined by $\lambda_{\mathbf{x}}$ and $\Omega(g)$, a measure of g's complexity.

LIME uses a set of $\hat{d}$ interpretable representation features $\hat{\mathbf{x}} \in \{0,1\}^{\hat{d}}$ that are sampled from the original feature space of the data, $\mathbf{x} \in X$. By using binary vector represented perturbed instances $\hat{\mathbf{z}}$ around non-zero elements of $\hat{\mathbf{x}}$, a label for the explanation model, $\mathscr{M}(\mathbf{z})$, is obtained. The mapping of the binary vector representation of features to the original real-valued representation is performed via a mapping function $h_{\mathbf{x}}$, such that $h_{\mathbf{x}} : \hat{\mathbf{z}} \rightarrow \mathbf{z}$, that is, $\mathbf{z} = h_{\mathbf{x}}(\hat{\mathbf{z}})$. Thus, using this dataset, Z, of perturbed samples with their labels, that is, $\{(\hat{\mathbf{z}}, \mathscr{M}(\mathbf{z}))\}$, the locality-aware loss function, $\mathscr{L}$, is defined as Eqn. (2.3) [109]:

$$\mathscr{L}(\mathscr{M}, g, \lambda_{\mathbf{x}}) = \sum_{\mathbf{z},\hat{\mathbf{z}} \in Z} \lambda_{\mathbf{x}}(\mathbf{z})(\mathscr{M}(\mathbf{z}) - g(\hat{\mathbf{z}}))^2 \tag{2.3}$$

Few pieces of literature have attempted to show the applicability of LIME in interpreting ECG signal-based heart disease classification ML model outputs [110, 111]. The interpretations provided by LIME do not fully capture the essential manifestations of heart diseases in an ECG signal [112]. Moreover, they do not deliver a global explanation of the learned complex ML model over the entire spectrum of feature values. In addition, the random perturbations of feature instances left the LIME techniques to suffer from the instabilities that pose challenges in reproducing the explanations [113–116]. Furthermore, LIME can be manipulated to hide biases [117].

Permutation Feature Importance (PFI) PFI measures the change in the performance of the black box ML model while shuffling any given feature of the test dataset. Thus, PFI interprets the black box ML model by describing the contribution of a feature in the ML model's output accuracy [118]. Given a trained model $\mathscr{M}$, such that $\mathscr{M}(\mathbf{x}^{(i)}) \approx y^{(i)}$, where $\mathbf{x}^{(i)} = ({x_1}^{(i)}, {x_2}^{(i)}, \ldots, {x_j}^{(i)}, \ldots, {x_m}^{(i)})$ is feature vector of size m and $y^{(i)}$ is a target of the i^{th} instance. The PFI calculates the contribution of a given feature j in predicting $y^{(i)}$ as indicated in Eqn. (2.4) [119, 120]:

$$PFI(\mathscr{M}, j) = \frac{1}{nk} \sum_{i=1}^{n} \sum_{l=1}^{k} [\mathscr{L}[y^{(i)}, \mathscr{M}(\mathbf{x}_j^{(\tau_l)^{(i)}})] - \mathscr{L}[y^{(i)}, \mathscr{M}(\mathbf{x}^{(i)})]] \tag{2.4}$$

where τ_l is a random permutation vector of instances in a dataset, $\mathscr{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{n}$, with n instances for $l = 1, \ldots, k$ permutations. $\mathscr{L}$ is a loss function linking the model output $\mathscr{M}(\mathbf{x})$ to the target pair y. Thus, $\mathscr{L}[y^{(i)}, \mathscr{M}(\mathbf{x}_j^{(\tau_l)^{(i)}})]$ is the loss function linking the perturbed output of the model $\mathscr{M}(\mathbf{x}_j^{(\tau_l)^{(i)}}) = \mathscr{M}(x_1{}^{(i)}, x_2{}^{(i)}, \ldots, x_j{}^{(\tau_l)^{(i)}}, \ldots, x_M{}^{(i)})$ to the target $y^{(i)}$ with respect to the perturbed feature x_j and $\mathscr{L}[y^{(i)}, \mathscr{M}(\mathbf{x}^{(i)})]$ gives a baseline loss linking the baseline output of the model and $\mathscr{M}(\mathbf{x}^{(i)})$ to the target pair $y^{(i)}$ for the instance i.

PFI can give model-agnostic global insight into the black box model, $\mathscr{M}$. It considers the dependency between features while determining their importance, specifically when it is applied during model training. In addition, it avoids retraining a model with a different subset of features, which saves time and even prevents it from reaching a new model due to the retraining process. However, PFI needs a labeled ground truth of a given instance to calculate the feature importance. Similarly, when a dataset contains highly correlated features, the PFI may exhibit bias, leading to less important features being assigned the highest importance value [121]. Furthermore, permuting individual time steps, in a time series signal like ECG, disrupts the temporal sequence and ignores the inherent structure and strong temporal dependencies of the signals [120, 122].

2.3.2.2 Learned Internal Parameters of the Model

Explaining the internal learned parameters of a model is a common interpretability technique in classical machine learning. It is particularly useful in inherently transparent algorithms. For instance, in tree structures, the learned parameters include the features and splitting criteria [123]. This approach can aid in visualizing and interpreting the focus of the model in object classification by analyzing and adjusting intermediate filters in CNNs. However, it is less effective for complex signals such as ECGs, where the patterns are not as visually distinct as those in natural scenes [124].

Tree-based ML models, such as decision trees, random forests, XGBoost, and AdaBoost, operate by partitioning the dataset using criteria like Gini impurity, mean squared error, or information gain. These criteria are applied based on the feature values within the dataset to make optimal splits for better classification. Each splitting creates different subsets from the dataset of the final, intermediate, and first subsets, respectively, called leaf nodes, split nodes, and root nodes [93, 123, 125]. Mathematically, the predicted instance, $\hat{y}$, obtained

from the leaf node is represented in terms of feature **x**, as given in Eqn. (2.5) [125]:

$$\hat{y} = \sum_{m=1}^{M} \mu_m I\{\mathbf{x} \in R_m\} \tag{2.5}$$

where μ_m is the average value of all elements present in the subset (R_m), $I\{\mathbf{x} \in R_m\}$ is a binary identity function that gives 1 if **x** is in the R_m subset, or else it returns 0. As stated earlier, the criteria used to generate the R_m subsets can be the Gini impurity index, mean squared error, or information gain based on the problem and data type of the dataset.

In tree-based ML models, the learned parameters, including the splitting threshold values of a feature, the Gini impurity index value, and the number of data points of the model are explained more easily. However, as the tree depth increases, the interpretation becomes difficult, and the model becomes opaque. In addition, the interpretation of truthfulness is affected by the poor generalization properties of the tree models themselves, where most tree-based ensemble models lack stability, especially while modeling complex interactions among several features [93, 126–129].

2.3.2.3 Visual Explanation

One of the methods for interpreting the output of a black-box machine learning model is to identify influential segments or regions in the data. These segments or regions are highlighted based on their contribution to the model's decision [130]. Visual explanation-based result presentation techniques have been tested in interpreting black-box machine learning classifiers in an ECG signal-based heart disease diagnosis. Some of them include class activation map-based techniques [131–134], saliency maps [135, 136], layer-wise relevance propagation [137], occlusion maps [135], and attention maps [138–141]. Moreover, LIME [110], and SHAP [101, 142–144] are used to explain the decision of the ML techniques by visually representing the important regions of an ECG signal, which contributes most to the decision. A brief discussion of some of these methods is presented below.

Class Activation Maps The class activation map (CAM) technique introduced by Zhou et al. [145] provides a visual explanation by localizing the important regions in input data that play major roles in the decisions of ML models. In class activation, the descriptive regions of input data that an ML model used for classification are highlighted [146]. The

class activation map calculates the contribution of units (L_{ij}^{c}) in the last layer activation filter map (F_{ij}^{k}) of the convolutional layer for the class prediction score (y^{c}) of the output layer. The CAM technique proposed by Zhou et al. [145] used global average pooling (GAP) and fully connected layers (FC) to obtain L_{ij}^{c}. In [145], F_{ij}^{k} and y^{c} have a linear relationship as given in Eqn. (2.6).

$$y^{c} = \sum_{k} w_{k}^{c} \sum_{i} \sum_{j} F_{ij}^{k} \tag{2.6}$$

where w_{k}^{c} represents the importance weight for the k^{th} convolutional feature map in predicting class c, (i,j) are indices of the k^{th} feature map, c is the class category; and k is a filter index.

The main aim of CAM is to find the contribution of the last feature maps that satisfy $y^{c} = \sum_{i,j} L_{ij}^{c}$. Thus, the contribution of each unit in the last feature map, L_{ij}^{c}, can be obtained from Eqn. (2.6), as shown in Eqn. (2.7):

$$L_{ij}^{c} = ReLU\left(\sum_{k} w_{k}^{c} F_{ij}^{k}\right) \tag{2.7}$$

In a single-dimensional time series signal, such as an ECG signal, the class activation map for class c at the specific temporal instance t is as indicated in Eqn. (2.8):

$$L_{t}^{c} = ReLU\left(\sum_{k} w_{k}^{c} F_{t}^{k}\right) \tag{2.8}$$

where F_{t}^{k} is the activation of filter k in the last conventional layer at the temporal instance t, and L_{t}^{c} indicates the importance of the activation at the temporal location t leading to the categorization of a signal into class c.

CAM has been used for interpreting an ECG signal classification result of a convolutional neural network [147]. Accordingly, it allows the visualization of segments of an ECG signal that the classification model mainly uses in its decision. However, the linear layers vanish the non-linearity of deep classifiers. In addition, the integration of CAM changes the network architecture and needs retraining [148]. As a result, techniques Grad-CAM [131, 132, 149–157], and Grad-CAM++ [134, 158] have been proposed in the ECG signal-based heart disease classification.

Given z-number of units in the feature maps, Grad-CAM decides the importance weight, w_k^c, of each of the k^{th} feature map, F_{ij}^k, to the class score, y^c, as given Eqn. (2.9).

$$w_k^c = \frac{1}{z}\sum_i\sum_j \frac{\partial y^c}{\partial F_{ij}^k} \tag{2.9}$$

Grad-CAM generates coarse heatmaps to highlight influential regions of a signal, but it sometimes lacks precise localization. Moreover, it suffers from a gradient saturation problem that results in inaccurate localization of relevant segments and leads of the 12-lead ECG signal. In addition, the localization of the descriptive signal part is highly affected by small perturbations of the input signal. Furthermore, the explanation is noisy and contains discontinuities [148].

On the other hand, Grad-CAM++ generates more detailed heatmaps and provides better localization of the segments of a signal that contribute most to the model's classification output [159]. This improvement is due to Grad-CAM++ computes a weighted combination of higher-order partial derivatives of the class score with respect to the feature maps, as given in Eqn. (2.10).

$$w_k^c = \sum_i\sum_j \alpha_{ij}^{kc} \cdot ReLU\left(\frac{\partial y^c}{\partial F_{ij}^k}\right) \tag{2.10}$$

where α_{ij}^{kc} is a coefficient that represents the contribution of the activation at indices (i,j) of the k^{th} feature map to the output class c, and calculated as in Eqn. (2.11):

$$\alpha_{ij}^{kc} = \frac{\frac{\partial^2 y^c}{\partial (F_{ij}^k)^2}}{2\frac{\partial^2 y^c}{\partial (F_{ij}^k)^2} + \sum_a\sum_b F_{ab}^k \left(\frac{\partial^3 y^c}{\partial (F_{ij}^k)^3}\right)} \tag{2.11}$$

where $\sum_a\sum_b$ indicates that the summations over all positions (a,b) in the feature map. The coefficient, α_{ij}^{kc}, improves Grad-CAM++'s ability to generate accurate heatmaps by leveraging higher-order derivatives, that are, second and third-order derivatives of the class score, y^c, with respect to the activation, F_{ij}^k. This allows it to capture the sensitivity of the class score to changes in activations at various temporal and spatial locations.

Saliency Maps A feature saliency map highlights the regions of a signal that are most relevant for classifying the input signal into a given category. It shows which parts of the signal contribute most to the model's decision for a specific class. The saliency map can be built using gradients of the output, $y^c = \mathscr{M}(\mathbf{x})$, of an ML model, $\mathscr{M}$, with respect to the input, $\mathbf{x}$, for the class c, highlighting how sensitive the output is to changes in the input [135]. The idea is that the class score y^c can be approximated by using the first-order Taylor expansion as given in Eqn. (2.12):

$$y^c = \mathscr{M}(\mathbf{x}) \approx \mathbf{w}^T \mathbf{x} + b \tag{2.12}$$

where b is a scalar, and $\mathbf{w}$, as indicated in Eqn. (2.13), is the weight gradient vector that provides an explanation for the model classification outcome:

$$\mathbf{w} = \frac{\partial \mathscr{M}(\mathbf{x})}{\partial \mathbf{x}} \tag{2.13}$$

Among other techniques, the saliency map can be generated using guided backpropagation, where the gradient of each neuron is calculated, and those with the highest gradient values are activated to form a heatmap. Generating saliency maps using guided backpropagation can become computationally expensive for large CNN networks due to the need to compute and store gradients for every neuron across multiple layers [136]. The heatmap shows the most salient parts of the signal that contribute most to classifying the input $\mathbf{x}$ to class c.

Saliency maps were experimented with for explaining complex ML models in ECG signal-based heart disease diagnosis [135, 136, 160, 161]. Although the backpropagation gradient saliency map can visually enhance regions of the input signal that contribute the most to classification, it has certain limitations. At first, the backpropagation saliency suffers from a gradient saturation problem mainly because saliency maps depend on input sensitivity [162]. Next, the generated gradient heatmap often does not explain the direct relation to the classifier's decision [163]. More importantly, the saliency method is susceptible to small shifts in the input signal, and its explanation may not be reliable [164].

Layer-Wise Relevance Propagation (LRP) An LRP provides an explanation for the model output by decomposing it into relevance score (R_n). These scores quantify the contribution

of each input element x_n to the model's output prediction $y = \mathscr{M}(\mathbf{x})$, where the input sample is $\mathbf{x} = [x_1, \ldots, x_n, \ldots, x_N]$. Thus, an LRP explains the ML model's output by attributing relevant values to the essential components of the input by tracing back the trained model layer by layer, starting from the final output node [165]. This layer-by-layer relevance propagation holds the layer-wise conservation property, given that i and j are neurons at two consecutive layers of a neural network, l and $l+1$, respectively. The overall sum of the i^{th} neuron's relevance score sums to $R_i^{(l)}$, such that relevance conservation property is maintained according to Eqn. (2.14) [165]:

$$R_i^{(l)} = \sum_j R_{i \leftarrow j}^{(l,l+1)} \quad \textit{such that } i \textit{ contributes to } j \tag{2.14}$$

The propagation of relevant scores R_j of layer $l+1$ onto neurons of the l layer can be achieved using different types of rules. Moreover, different rules can be used at each layer of the network architecture [166]. One of the simplest rules is given in Eqn. (2.15) [165]:

$$R_i = \sum_j \frac{u_i w_{ij}^{\text{fc}}}{\sum_{0,i} u_i w_{ij}^{\text{fc}}} R_j \tag{2.15}$$

where u_i is an activation of the neuron i, w_{ij}^{fc} is the weight connecting neuron i to neuron j, and $\sum_{0,i}$ indicates the sum over all neurons j in the l layer. Moreover, the rule satisfies the basic properties in which deactivated neurons, neurons with no connection, and zero weight has no relevant value.

LRP has been used to interpret the output of deep learning models by generating heatmaps based on relevance scores, highlighting the regions of the input that contribute most to the prediction. LRP provides detailed and exact relevance scores for each input feature, offering fewer noises around the target class and the ability to highlight parts of the signal that negatively contribute to the output. However, it is complex to implement and interpret [166, 167]. Besides, the heatmap produced by LRP can still be noisy due to the initialization of the non-target class with a zero relevance value. In addition, LRP may not always produce intuitive or visually interpretable results compared to gradient-based methods. Moreover, it has a limitation in discriminating targets that produce identical heatmaps for different entities in an input signal [168]. Furthermore, selecting propagation rules is challenging, and obtaining the best parameters is trivial [169].

Occlusion Map The occlusion map is one of the attribution-based techniques where the model output is explained by changing part of the input data with different values [170]. The input can be altered on a specific location, for instance, in a time series signal such as an ECG with total t_h time points, the alteration can cover a certain time step durations (t_d). For a signal $\mathbf{x} = \{x_{t_1}, \ldots, x_{t_d}, \ldots, x_{t_h}\}$, the locally altered signal $(\hat{\mathbf{x}})$ can be obtained using Eqn. (2.16) [170]:

$$\hat{\mathbf{x}} = (\mathbf{x} \odot \mathbf{m}_1) + o_v \mathbf{m}_2 \tag{2.16}$$

where $\odot$ is Hadamard product [171], $\mathbf{m}_1$ and $\mathbf{m}_2$ are mask vectors that complement each other, that is, $\mathbf{m}_2 = \neg\mathbf{m}_1$ and o_v is the occluding value. The values for $\mathbf{m}_1$, $\mathbf{m}_2$, and o_v are determined based on the required modifications on $\mathbf{x}$.

The occlusion-based ML model's interpretation algorithms are simple to implement. Moreover, it can measure the marginal effects of each windowed region of the input signal given that the segments of the input are independent [172, 173]. In addition, the occlusion method is used to interpret the output of non-differentiable ML models, unlike gradient-based explanation techniques [135]. However, similar to other perturbation-based model output explanation methods, such as LIME and SHapley value maps, the computational complexity associated with the input occlusion is high [174, 175].

Attention Mechanisms Attention mechanisms are commonly used in time-series data due to their ability to overcome the limitations of traditional encoder-decoder models. They enhance the model's focus on the most relevant parts of the input sequence [139, 176]. The attention mechanism can be incorporated into machine learning networks, allowing the model to focus on specific regions of an input signal that contribute most to the output prediction [138, 139, 176–180]. This approach ensures that the contribution of each segment of a signal to the model's classification output is effectively captured by other IML techniques during inference [177].

The attention mechanism takes the encoder output (latent vector) as the input and performs three consecutive computations, which are alignment scoring (e_{ij}), computing attention weights, and attention score vector computation, as given in Eqn. (2.17), Eqn. (2.18), and

Eqn. (2.19) [181], respectively.

$$e_{ij} = a(\mathbf{s}_{i-1}, \mathbf{h}_j) \tag{2.17}$$

where a is an alignment model whose score e_{ij} measures how well the input around position j of the encoder's hidden state $\mathbf{h}_j$ matches the previous decoder hidden state $\mathbf{s}_{i-1}$ at position i just before emitting. Then, the attention weight score (β_{ij}) of each $\mathbf{h}_j$ is computed by applying an activation function, for instance, the softmax activation function, on the alignment score as shown in Eqn. (2.18).

$$\beta_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{ij})} \tag{2.18}$$

where T is the number of the encoder's hidden states. Finally, the attention score vector ($\boldsymbol{v}$), which is the output of the attention mechanism, is computed as a weighted sum of all encoder hidden states, as shown in Eqn. (2.19).

$$\boldsymbol{v}_i = \sum_{j=1}^{T} \beta_{ij} \mathbf{h}_j \tag{2.19}$$

Based on the techniques employed for generating attention scores, attention mechanisms are broadly classified into deterministic attention and stochastic attention [182]. In the case of a deterministic, attention scores are calculated as the weighted sum of all hidden states, whereas, in stochastic attention, attention scores are determined by selecting one of the hidden states, h_j.

The attention mechanism introduces the model's output interpretability scheme, in addition to improving the performance of the ML model's ECG signal-based heart disease classification [138–141,176]. However, attention mechanisms alone may not fully explain the model's behavior [183].

The attention mechanism introduces the model's output interpretability scheme, in addition to improving the performance of the ML model's ECG signal-based heart disease classification [138–141, 176]. However, attention mechanisms alone may not fully explain the model's behavior, especially considering the open question of whether query and key need to exist independently [183].

2.3.2.4 Example-Based Explanation

Example-based explanation techniques in ML help end-users understand a model's prediction for a specific instance by referencing similar examples from the training set [90, 184]. For instance, the Nearest Neighbors technique identifies and presents training examples closest to the input instance [185]. Additionally, counterfactual explanations show how the model's prediction would change if the input differed, comparing it to similar examples with different outcomes [186].

The concept in an example-based explanation technique is that if two data instances (X_i and X_j) are similar and the ML model's ($\mathscr{M}$) output for input data instance X_i is $y = \mathscr{M}(X_i)$, then the model output for a data instance X_j is also y.

Example-based ML output explanations include counterfactual [186, 187] and adversarial examples [188]. Moreover, inherently interpretable (transparent) shallow ML algorithms include the k-nearest neighbor (KNN) [94, 185] work based on an example-based approach. These techniques work through minimizing a loss function, commonly a distance metric between the instance to be explained $\mathbf{z}$ and its perturbed form $\hat{\mathbf{z}}$. In this method, the ML model's output is explained by finding the extent of perturbations on the input instance that brings changes to the outcome of the ML model. Formally, given an ML model $\mathscr{M} : Z \rightarrow Y$, a data instance $\mathbf{z} \in Z$ with model output $y = \mathscr{M}(\mathbf{z})$, and the desired model output target $\hat{y} \in Y \setminus \{y\}$, a counterfactual explanation solves the objective function, d, given in Eqn. (2.20) [187]:

$$\underset{\hat{\mathbf{z}} \in Z}{\text{minimize}}\ d(\hat{\mathbf{z}}, \mathbf{z}) \quad \text{such that} \quad \mathscr{M}(\hat{\mathbf{z}}) = \hat{y} \tag{2.20}$$

where d is any distance metric.

Example-based explanation techniques highlight part of an input instance or feature values changed to give the target class $\hat{y}$. In other words, the explanation gives the difference between $\mathbf{z}$ and $\hat{\mathbf{z}}$, such that $\mathscr{M}(\mathbf{z}) \neq \mathscr{M}(\hat{\mathbf{z}})$. In addition, an example-based explanation is easily implemented because of the objective function that can be easily optimized [189, 190]. However, there will be more than one example for a single sample instance that results in a lack of obtaining a unique explanation for a particular input instance. Moreover, several challenges need to be addressed, including limitations in visualizing results [190].

2.3.3 Scope of IML Techniques

IML models can also be classified as locally and globally scoped based on whether they explain predictions for a specific input instance or provide insight into the overall workings of the model, respectively. Local scoped interpretability methods aim to interpret the output of the DL model for a single instance. In contrast, globally scoped interpretable methods aim to explain the model's behavior across the dataset. They provide a comprehensive understanding of the whole logic of the model and the entire reasoning follows for all possible outcomes of the model [90, 91, 95].

Local model interpretation methods focus on answering "why is the ML model making a specific prediction?". Moreover, these methods can reveal the effects of a specific segment of input instances or feature values on the output of the model [90, 115]. Thus, these techniques help to understand the causal relations between specific input instances and their corresponding ML model outputs [95]. However, the explanation obtained from these techniques is valid only for a single input instance and does not generalize. In addition, the explanation result obtained from these techniques lacks stability. That means the explanation generated through consecutively running these techniques may result in a different outcome. Furthermore, the local surrogate model may spuriously approximate the complex ML models, i.e., the explanation outcome may have no real connection with the ML model [191, 192].

On the other hand, global model interpretation methods focus on answering "how is the ML model making a prediction?". These methods can try to understand how subsets of the model influence the model's decisions. Global interpretability can be achieved through training interpretable constraints together with the input data [95]. In addition, it can also be achieved by demonstrating the statistical contribution of each feature in the decision of the underlying black box model. Furthermore, the global explanation can also be obtained by capturing representation at the intermediate layers of complex DL models. Thus, these techniques help to understand the inner working mechanisms of ML models and increase the model's transparency [95]. However, globally scoped interpretation techniques often miss explaining a model output for specific input instances. However, different methods have been proposed in the literature for obtaining a global explanation of the black box model through aggregating local explanations [193].

2.3.4 Specificity of IML Techniques

Interpretability techniques in ML can also categorized into model-specific and model-agnostic methods based on their applicability across different models [90]. Model-specific techniques are designed to explain particular classes of models, but they may not always use internal parameters, as seen in convolutional neural networks (CNNs). Instead, these techniques often focus on the unique aspects of the model's architecture [95]. On the other hand, model-agnostic techniques provide explanations that are independent of the internal workings of any specific model. They relate the input data to the output predictions of a black box model without depending on internal parameters. This approach ensures that explanations can be provided for a broad range of models, regardless of their underlying structure [94].

Model-specific explanation techniques not only explain the model outputs based on the model characteristics but also help in improving the efficiency of the ML model by investigating the characteristics. Moreover, model-specific interpretation techniques have high translucency in which they can rely on more information to generate an explanation [90]. However, they are limited to a specific model and are less portable to explain other models. On the other hand, model-agnostic interpretable techniques are independent of the model to be explained and can be applied to any model [94]. However, due to the approximation and assumptions made in constructing model-agnostic interpretation techniques, their explanation results may become less accurate and even vulnerable to adversarial attacks [94, 117, 188]. In addition, it may be difficult to faithfully detail the explanation produced by model-agnostic methods, as to how they truly reflect the decision-making processes of the ML model [95]. Furthermore, the computational complexities of model-agnostic techniques, such as SHapley values, grow exponentially as the number of input features increases [192].

2.3.5 Complexity of ML Models

Based on the complexity of an ML model to be explained, the interpretability methods are categorized into intrinsic and post hoc. In intrinsic interpretability, the explanation is based on understanding how the ML model works. On the other hand, in post-hoc interpretability, the explanation is provided by extracting a piece of information from a trained complex black box ML model [90].

The intrinsic explanation methods used for ML models have simple architecture by design and provide self-explanatory results. However, these ML models cannot be used to solve complex problems and suffer a lot from capturing nonlinearity in the data. In the literature, methods have been proposed to mitigate the trade-off in reducing the model performance for interpretability. One of the methods is adding semantically meaningful constraints to complex models to improve interpretability without a significant loss in the performance [124].

The post hoc explanation methods are usually applied after the ML model is trained and provide an explanation without modifying the trained model. Moreover, the complex ML model can be approximated by surrogate models, such as decision trees and shallow neural networks. These surrogate models provide a global post hoc model-agnostic explanation by mimicking the complex ML model [194–196]. These techniques are much more flexible and can switch to explain different black box ML models. However, the post hoc methods compromise the fidelity of the explanation. In addition, they may fail to represent the behavior of the complex ML model [95].

As a summary, it is worth to note that, both globally and locally scoped IML techniques can be model specific or model agnostic and used for intrinsic model explanations or post-hoc explanations [95], as shown in Table 2.2.

2.4. Summary

The discussion above underscores the potential of incorporating ML/DL models for diagnosing cardiac abnormalities using 12-lead ECG signals. These models offer promising advancements in automating and enhancing the accuracy of ECG analysis. However, there are still challenges and areas that require further exploration to fully leverage their capabilities in clinical applications.

Firstly, some models are highly complex, with parameters reaching into the millions [20, 50, 80]. While these complex models often achieve high accuracy, their computational demands can be a drawback. In contrast, lightweight architectures, such as those proposed in [44, 47], offer reduced complexity. However, these simpler models tend to show lower performance levels compared to their more complex counterparts.

Secondly, these models have not been extensively tested on diverse datasets and disease

classes, limiting their generalizability. As a result, their performance on a broader range of heart conditions remains uncertain. Without thorough testing on various datasets, it is difficult to assess their true applicability. This raises concerns about their effectiveness after re-training for different cardiac abnormalities.

Thirdly, there is a significant limitation in the integration of model output interpretation techniques across most studies. These techniques are crucial for understanding how the model's decisions are made and identifying which segments of the ECG signal contributed most to the classification output. Without these insights, the models function as black boxes, providing limited understanding of how their decisions are made. This lack of interpretability impedes the ability to provide evidence-based diagnoses. As a result, it becomes challenging for physicians to trust and adopt these models in clinical practice.

These limitations motivate us to develop a robust interpretable model while minimizing the number of trainable parameters. This ensures that the interpretability method provides clear insights into the model's heart disease classification outputs. Additionally, reducing the number of trainable parameters enhances the model's efficiency and practicality, making it more suitable for real-world applications in clinical settings.

Table 2.2: Summary of commonly used techniques for ML interpretation in ECG-based heart disease classification.

Technique	**Scope**	**Specificity**	**Complexity**	**Result Presentation**
LIME [110, 111]	Local	Model-agnostic	Post hoc	• Highlights relevant regions of an ECG.
Feature importance (FI) [110]	Global	Model-agnostic	Post hoc	• Features that have meaningful clinical significance are identified based on their importance in the ML model's output classification.
SHAP [99–102, 110, 142–144]	Local/Global	Model-agnostic	Post hoc	• Rank the global feature importance of an ECG signal, provide a local explanation for the model's classification output, and highlight morphological segments of the ECG signal.
Attention mechanisms [138, 139, 176–180]	Local	Model-specific	Intrinsic	• Uses attention weights to highlight segments of the input signal.
Layer-wise relevance propagations (LRPs) [137, 165]	Local	Model-agnostic	Post hoc	• Highlights regions of the input signal to indicate the contribution of each region through the back-propagating relevance score from the ML model's final output.

Table 2.2: *Cont.*

Technique	**Scope**	**Specificity**	**Complexity**	**Result Presentation**
Occlusion Maps [135, 173]	Local	Model-agnostic	Post hoc	• Identify segments of an ECG signal by replacing parts of the signal and observing the change in the output.
Class-Activation Maps [131–134, 140, 146, 147, 149–157]	Local	Model-specific	Post-hoc	• It highlights segments of an ECG signal to show each region's contribution by concatenating feature maps through gradients from the output class to the final convolutional layer.
Saliency Maps [135, 136, 160, 161]	Local	Model-agnostic	Post hoc	• Suggests a segment of an ECG signal that contributes the most to classifying a particular input instance to an output class.
Learned internal parameters (LIPs) [127–129]	Global	Model-specific	Intrinsic	• Provide the internal parameters of the ML models. For instance, the splitting conditions of the tree structure are based on functional feature components and provide the final decision probabilities on the leaf nodes.
Example-based [185]	Local	Model-agnostic	Post hoc	• The explanation output consists of a raw and combined information about ECG signals that are nearest neighbors to the ML model's input ECG tracing.

CHAPTER 3

Research Methodology

3.1. Overview

This chapter details the research methodology of the dissertation. It covers the discussion of the research framework, the datasets, the pre-processing methods, the designed model, evaluation metrics, and interpretability methods used in the study. In this dissertation, we used an experimental design research approach because it focused on practical solutions and allowed us to use scientific results to develop solutions for complex real-world problems. This approach aligns with the goal of creating robust and relevant solutions for ECG-based heart disease diagnosis. By utilizing this methodology, the research systematically designs and experiments a robust and interpretable hybrid deep learning system for ECG analysis. This approach addresses challenges in ECG interpretation and ensures its practical applicability and effectiveness across diverse datasets.

3.2. The Research Workflow

The workflow for the ECG-based heart disease classification method presented in this study, as shown in Figure 3.1, begins with dataset selection followed by data pre-processing. The raw ECG signals are first denoised to remove artifacts, normalized to standardize the signal, and then split into multiple segments or channels to facilitate analysis. This pre-processing step ensures the data is clean and ready for input into the model, improving the efficiency.

The model employs a multichannel, hybrid deep learning architecture designed to handle the complexity of ECG signals. The system splits the 12-lead ECG input into 12 channels, pro-

cessing them individually with a series of 1D convolutional neural networks (1D-CNN) that include residual connections to maintain feature information along the depth of the network. These layers are followed by a bidirectional long short-term memory (BiLSTM) network to capture temporal dependencies, an attention mechanism to focus on critical segments, and a 2D-CNN layer to refine feature extraction. The final classification is done using fully connected layers. Bayesian optimization was used to fine-tune the model's hyperparameters for improved performance across various ECG datasets [197]. Then, the classification performance is analyzed using metrics, including accuracy, specificity, confusion matrices, precision, recall, F1-score, AUC, and AUPRC. These metrics provide a comprehensive evaluation of the model's capability to classify different heart conditions from ECG data.

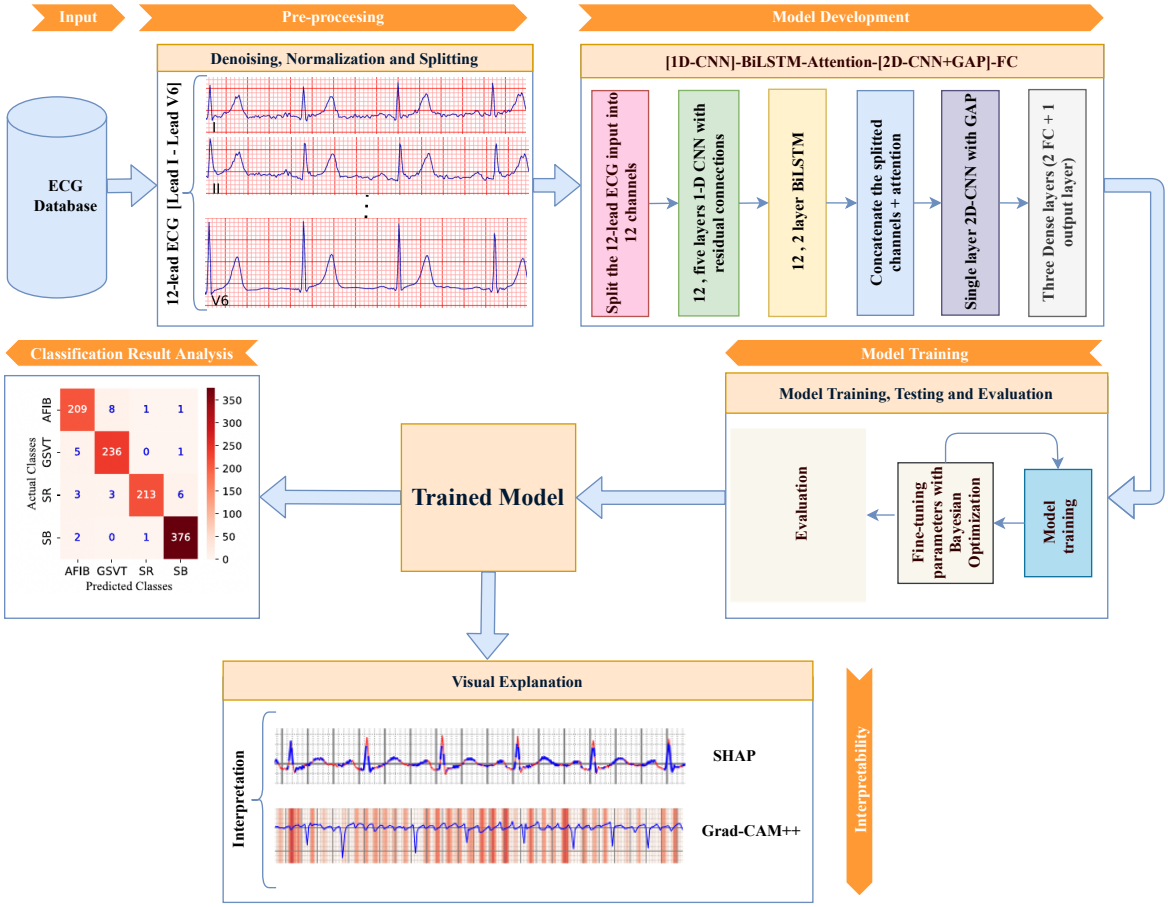


Figure 3.1: Workflow.

After the model was trained, its outputs were interpreted using post-hoc interpretability techniques. Specifically, Gradient-weighted Class Activation Mapping Plus Plus (Grad-CAM++) and SHapley Additive exPlanations (SHAP) were applied to visualize the most influential segments of the ECG signals driving the model's predictions. These techniques

were employed both at the instance level, analyzing individual samples, and at the test set level, assessing the contribution of individual ECG leads across the entire dataset.

3.3. Dataset

Among several cardiac ECG datasets, the PTB-XL [41], the CODE-15% [61], and the Chapman Arrhythmia [62] datasets are suitable for benchmarking and evaluating DL models. The datasets contain a wide variety of cardiac diseases. Besides, the public domain availability coupled with the 12-lead patient-level annotation makes these datasets ideal for developing DL models that have practical/clinical applications.

The PTB-XL dataset is a multi-class 12-lead ECG dataset comprising 21,837 samples, with each record 10 seconds long and recorded at a sampling frequency of 100 Hz. The dataset includes records of various cardiac conditions grouped into five super diagnostics classes: conduction disturbance (CD), hypertrophy (HYP), myocardial infarction (MI), normal (NORM), and ST/T change (STTC). The distribution of each class is depicted in Figure 3.2. A single ECG record may be associated with multiple disease classes. As shown in Figure 3.3, a total of 4079, 920, and 159 samples from the dataset were labeled into 2, 3, and 4 classes, respectively. In addition, 16272 samples were labeled in one of the five (5) classes. Besides, 407 cases do not belong to one of the five classes.

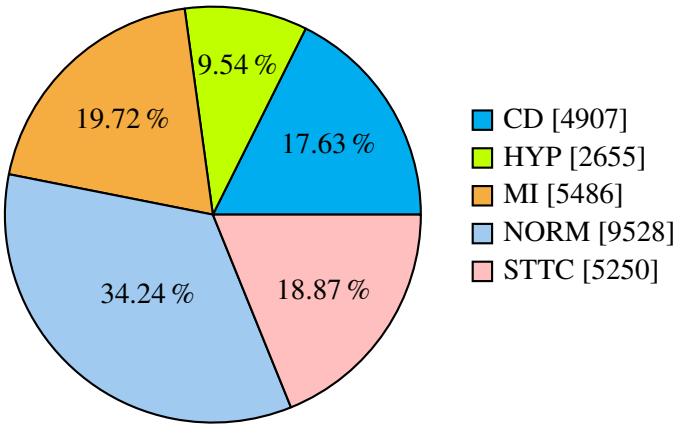


Figure 3.2: PTB-XL dataset's super-diagnostics class-wise distribution.

To make the comparison among ML techniques proposed by different researchers, Wagner *el al.* [36] have provided a stratified sampling-based partition for the super-diagnostic classes PTB-XL dataset as shown in Table 3.1.

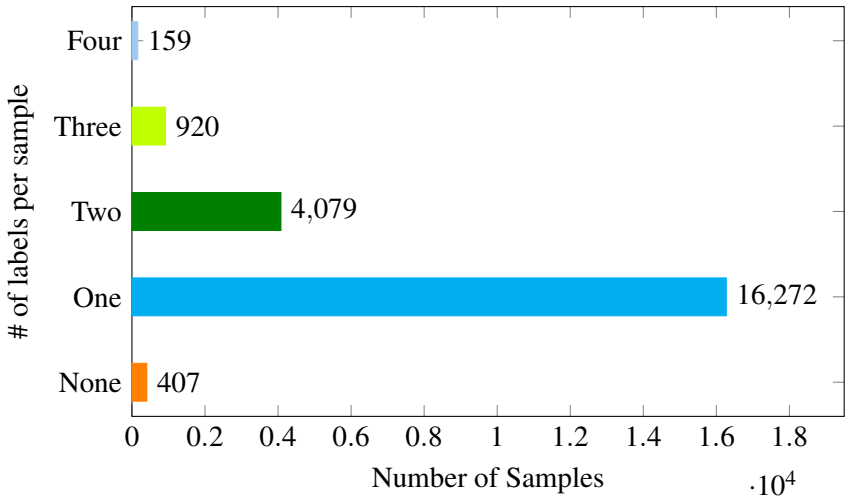


Figure 3.3: PTB-XL dataset's class label-wise distribution of records.

Table 3.1: Stratified sampling-based partition of super-diagnostic classes of PTB-XL dataset.

Classes	Partition									
	1	2	3	4	5	6	7	8	9	10
CD	481	487	487	494	496	496	479	492	497	498
HYP	263	264	264	261	265	270	265	269	271	263
MI	550	540	529	551	563	565	551	540	544	553
NORM	941	967	993	928	941	932	970	935	957	964
STTC	526	526	515	527	532	530	523	514	534	523
Total Count	2761	2784	2788	2761	2797	2793	2788	2750	2803	2801

Similarly, the CODE-15% dataset is a large multi-class 12-lead ECG dataset comprising 345,779 samples, each 10 seconds long and recorded at a sampling frequency of 400 Hz. The CODE-15% dataset was obtained through stratified sampling from the larger CODE dataset and is publicly available [61]. The dataset contains six (6) label annotated heart disease classes: first-degree AV block (1dAVb), right bundle branch block (RBBB), left bundle branch block (LBBB), sinus bradycardia (SB), sinus tachycardia (ST), and atrial fibrillation (AFIB). The distribution of each class is shown in Figure 3.4. Much like the PTB-XL dataset, the CODE-15% dataset is multi-class, with some records containing multiple class labels. As shown in Figure 3.5, among the six diseases, a total of 3486, 180, and 5 samples of the dataset contain 2, 3, and 4 disease labels, respectively. The majority of the records are normal, and cases that do not belong to one of these six classes. These records account for 89% of the total dataset, where the total count of normal ECG records is 134,657, and those that do not belong to one of six disease classes are 173,347.

In addition, the CODE-15% dataset authors provided a separate 827 ECG tracings, named CODE-test [198] that is annotated by specialist cardiologists, cardiology residents, emergency residents, and medical students. The annotation made by cardiologists is used as a gold standard to perform a performance comparison against ML models, medical students, and cardiology and emergency residents. The gold standard CODE test contains 146 samples representing the six disease classes, with 12 samples exhibiting dual disease cases. Specifically, it includes 28 instances of 1dAVb, 34 of RBBB, 30 of LBBB, 16 of SB, 13 of AFIB, and 37 of ST, as shown in Table 3.2. The remaining 82.35% of the samples do not belong to any of the six disease classes.

Table 3.2: The CODE-test dataset for six disease classes.

Class	1dAVb	RBBB	LBBB	SB	ST	AFIB
Count	28	34	30	16	37	13

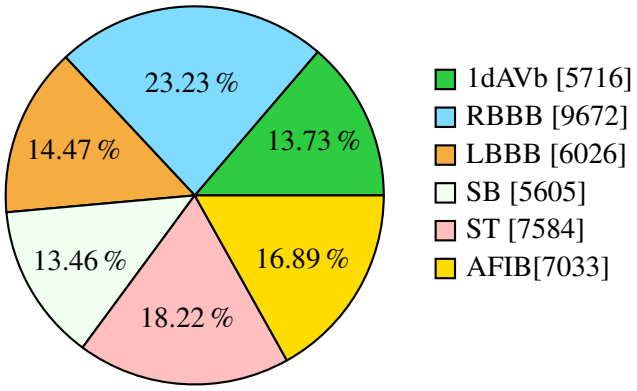


Figure 3.4: CODE-15% dataset's class-wise distribution.

On the other hand, the Chapman Arrhythmia dataset [62] is a 12-lead ECG record of 11 heart rhythms from 10,646 patients where each sample is 10 seconds long and sampled with a sampling frequency of 500 Hz. Among the dataset, Sinus Atrium to Atrial Wandering Rhythm (SAAWR), Atrioventricular Reentrant Tachycardia (AVRT), Atrioventricular Node Reentrant Tachycardia (AVNRT), and Atrial Tachycardia (AT) are in a few numbers with a total number of 7, 8, 16, and 121 samples, respectively. Researchers trained their proposed AI model using only the remaining seven (Chapman-Reduced) or the merged four (Chapman-Merged) rhythm classes. The seven disease classes are atrial flutter (AF), atrial fibrillation (AFIB), sinus irregularity (SI), sinus bradycardia (SB), sinus rhythm (SR), sinus

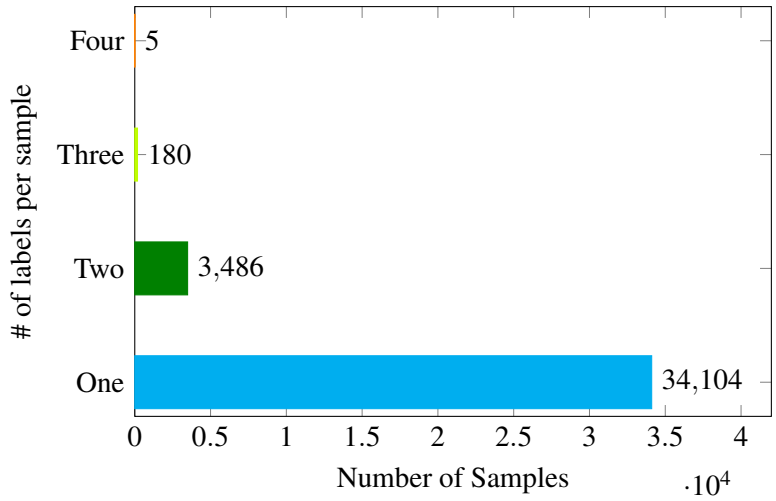
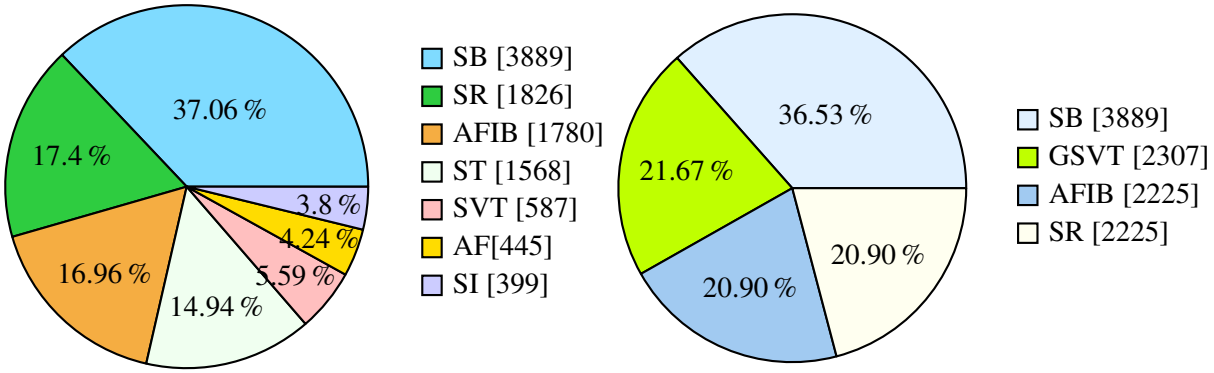


Figure 3.5: CODE-15% class label-wise distribution of records.

tachycardia (ST), and supraventricular tachycardia (SVT) As shown on Figure 3.6a, the reduction of the Chapman Arrhythmia dataset to seven classes is still imbalanced. As a result, Zheng *et al.* [199], dataset authors, suggested merging rare cases in the dataset to their upper-level arrhythmia types. The resulting four (4) arrhythmia disease groups, namely, SB, AFIB, GSVT, and SR, are shown in Figure 3.6b. SB contains only sinus bradycardia, AFIB contains – atrial fibrillation and atrial flutter, grouped supraventricular tachycardia (GSVT) contain – supraventricular tachycardia (SVT), atrial tachycardia (AT), atrioventricular node reentrant tachycardia (AVNRT), atrioventricular reentrant tachycardia (AVRT) and sinus atrium to atrial wandering rhythm (SAAWR), and SR contains – sinus rhythm and sinus irregularity.



(a) Seven rhythm classes

(b) Merged four rhythm classes

Figure 3.6: Chapman Arrhythmia dataset's seven and merged four rhythm class-wise distribution.

Additionally, a dataset from Zewditu Memorial Hospital (ZMH) containing ECG recordings of patients diagnosed with various heart conditions was used to pilot-test the trained model. For the pilot-testing phase, a subset of six disease classes have been selected from

this dataset: 1dAVb, RBBB, LBBB, SB, ST, and AFIB. The evaluation made by two cardiologists and one internal medicine specialist is used as a gold standard. The dataset contains 135 samples representing the six disease classes, with 19 samples exhibiting dual disease cases. Specifically, it includes 49 instances of 1dAVb, 17 of RBBB, 25 of LBBB, 44 of SB, 12 of AFIB, and 7 of ST, as shown in Table 4.8. These disease classes have been chosen because they align with the conditions in the CODE-15% public dataset, which has been used to train the model. This alignment allows for a controlled evaluation of the model's performance when applied to real-world data. The pilot testing aims to evaluate how effectively the model, which will be trained on the CODE-15% public dataset, can be generalized when applied to new ECG data from a different clinical setting. The dataset used in this study consists of 12-lead ECG signals with a duration of 10 seconds. Initially, the data was sampled at 1000 Hz, but to align with the model trained on the CODE-15% dataset, it was downsampled to 400 Hz. The data was downsampled to 400 Hz to ensure compatibility with the model trained on CODE-15% dataset.

The selected disease classes from ZMH, being clinically relevant and similar to those in the CODE-15% dataset, provide a valuable test case for evaluating the model's performance in real-world scenarios. This approach helps to get insight into the model's ability to handle diverse patient populations and varying ECG signals to test its robustness and potential for clinical application.

Table 3.3: The ZMH dataset for trained model evaluation.

Class	1dAVb	RBBB	LBBB	SB	ST	AFIB
Count	49	17	25	44	7	12

3.4. Data Preprocessing

Preprocessing ECG signals by removing artifacts and normalizing amplitudes is essential for ML/DL-based heart disease classification models. This process enhances the quality of the data. Improved data quality contributes to the stability and reliability of the model. ECG signals are often contaminated by internal and external (environmental) artifacts [200, 201]. External artifacts in the data include those originating from power line interference and electrode motion [200, 202]. Conversely, internal artifacts are primarily caused by the patient's

muscle tremors, shivering, and hiccups [200]. One of the significant artifacts that challenges ECG analysis, both for expert physicians and the ML/DL, is a baseline-wander (BW) [203]. The BW is a low-frequency noise, typically within the range of 0.15 to 0.8 Hz, that can cause deviations of segments or entire portions of an ECG signal from the isoelectric axis. Improper electrode placement, patient movement, and respiration during ECG acquisition commonly cause BW. The deviation of the ECG waves from their isoelectric axis, as illustrated in Figure 3.7, impedes the accurate diagnosis of the underlying cardiac condition.

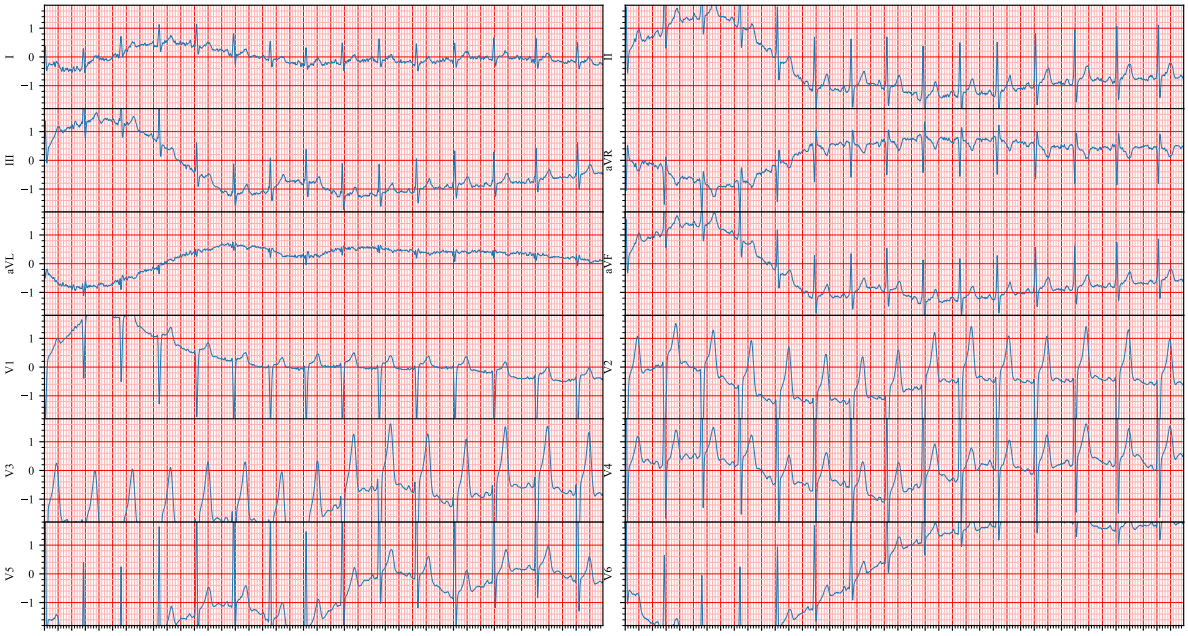


Figure 3.7: ECG record from CODE-15% Exam ID 821571.

In this study, we employed a discrete cosine transform (DCT) based technique depicted in Algorithm 1 to remove the baseline drift of ECG signals. DCT converts finite discrete sequences of an ECG wave as a sum of cosine functions, as indicated in Eqn. (3.1) [204,205]. Given a 12-lead ECG signal, $X \in \mathbb{R}^{12\times N}$, let $\mathbf{x} \in \mathbb{R}^{1\times N}$ represent a single lead among the 12-lead signals, the DCT operation on an $\mathbf{x}$ is given as:

$$\mathbf{g}[k] = \gamma[k] \sum_{i=0}^{N-1} \left(\cos \left[\frac{\pi k(2i+1)}{2N} \right] \mathbf{x}[i] \right) \tag{3.1}$$

where $\mathbf{g}[k]$ is frequency domain converted version of the time domain signal $\mathbf{x}$, i and k, respectively, are time and frequency domain axis indexes with N length, and the scaling

factor $\gamma[k]$ for both Eqn. (3.1) and (3.2) is defined as:

$$\gamma[k] = \begin{cases} \frac{1}{\sqrt{N}}, & \text{k=0} \\ \sqrt{\frac{2}{N}}, & \text{otherwise.} \end{cases}$$

Following the index-bound DCT concept proposed by Shin *et. al.* [206], frequency components below 0.8 Hz and above half of the sampling frequency (f_s) are masked as indicated on *line 8-11* of algorithm 1. Then, the DCT filtered signal ($\mathbf{g}_f$), converted back to time domain representation ($\mathbf{x}_f$) by inverse discrete cosine transform (IDCT) using Eqn. (3.2) [205]:

$$\mathbf{x}_f[i] = \sum_{k=0}^{N-1} \left(\gamma[k]\mathbf{g}_f[k] \cos\left[\frac{\pi i(2k+1)}{2N}\right]\right) \tag{3.2}$$

Algorithm 1 : Steps in DCT based BW removal.

1: **Data:** Input data ($X \in \mathbb{R}^{12\times N}$): 12-lead ECG signal with N-samples for each lead
2: **Parameter:** Low-pass frequency of BW $f_{BW} = 0.8$Hz
3: **Parameter:** High-pass frequency $f_{max} = fs/2$
4: **Result:** BW-removed ECG-signal ($X_f \in \mathbb{R}^{12\times N}$)
5: **for** $leads \leftarrow 0$ **to** 11 **do**
6: $\quad \mathbf{g}_{1\times N} \leftarrow \text{DCT}(X_{lead\times N})$
7: $\quad \text{freq}[k] \leftarrow (k \times f_s)/2N$ ▷ $0 \leq k \leq N-1$
8: $\quad mask \leftarrow (freq \geq f_{BW}) \wedge (freq \leq f_{max})$
9: $\quad dct\_filter \leftarrow [0] * N$
10: $\quad dct\_filter[mask] \leftarrow 1$
11: $\quad \mathbf{g}_{f_{1\times N}} \leftarrow \mathbf{g}_{1\times N} \cdot dct\_filter$
12: $\quad X_{f_{\text{lead}\times N}} \leftarrow \text{IDCT}(\mathbf{g}_{f_{1\times N}})$
13: **end for**

The technique mentioned above is implemented on a lead-by-lead basis for each lead to minimize the impact of BW. Figure 3.8b shows a sample result obtained after the BW artifact is removed from the raw ECG signal given on Figure 3.8a. On the other hand, the estimated BW in Figure 3.8a is the low-frequency BW artifact obtained by subtracting the BW compensated signal from the raw signal.

In addition to the BW artifact, the ECG signal is susceptible to electric power line interference (PLI) noise. Depending on the oscillating frequency of the electrical power supply, PLI noise occurs at 50/60 Hz with bandwidth $<$ 1Hz and often include harmonics [207]. As shown in Figure 3.9, PLI noise is manifested as a low-amplitude artifact with high fre-

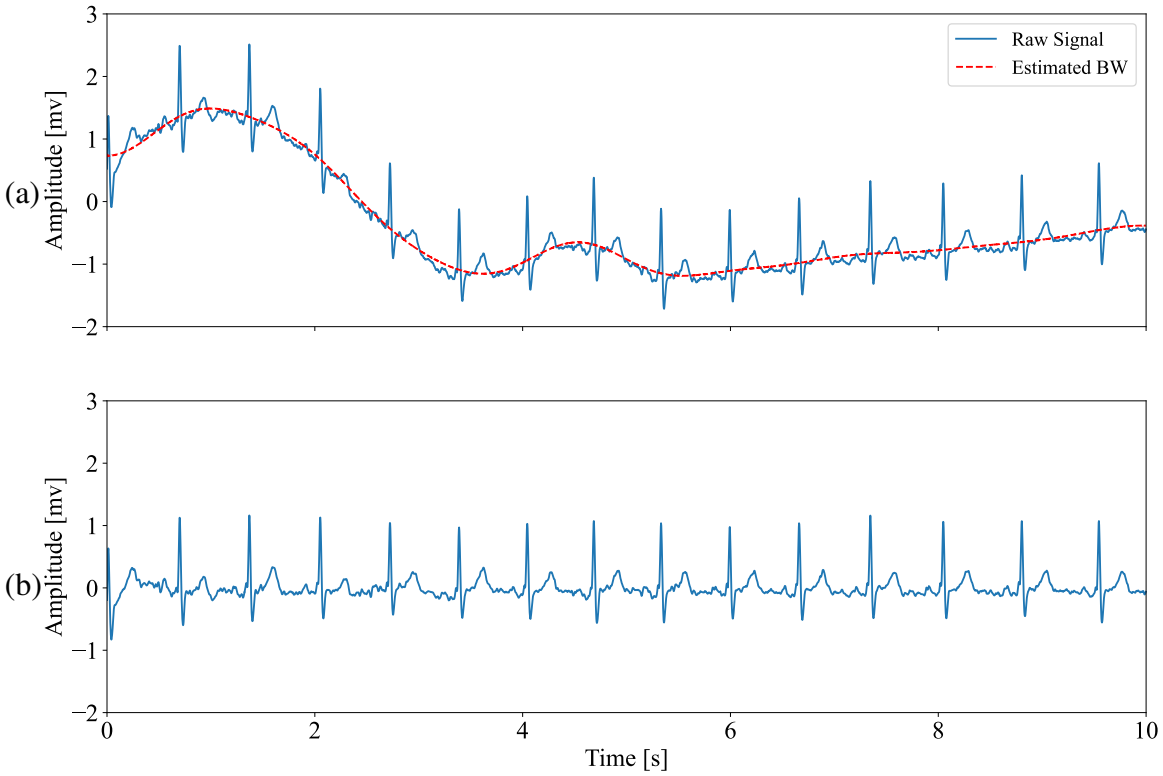


Figure 3.8: ECG record from CODE-15% Exam ID 821571: (a) raw Lead III ECG signal with estimated BW, (b) after cleaned from BW noise.

quency. It distorts the shape of an ECG signal and poses a difficulty for identifying the P-wave and T-wave in the analysis of cardiac conditions. PLI noise is compensated using a notch filter that removes the power line interference at 50/60 Hz and their harmonics [208]. Figure 3.10 shows PLI noise removal from lead III of an ECG sample given in Figure 3.9. A time-domain plot of the lead III wave after passing through the BW artifact cancellation module discussed above is shown in Figure 3.10a. Its corresponding frequency-magnitude distribution is depicted in Figure 3.10b and shows the signal contamination by PLI at 50 Hz and its harmonics 100 Hz. The effect of the PLI noise compensation using the notch filter is shown in Figure 3.10c. Furthermore, the frequency domain plot of the corresponding denoised signal in Figure 3.10d shows the elimination of the PLI noise together with its harmonics.

The third task in the preprocessing stage is normalizing the ECG signal. Normalization guarantees the amplitude of an ECG wave to have a similar range and helps to improve ML model stability against varying ECG signal amplitude across different records. In this paper, each lead of a denoised ECG signal, $\mathbf{x}_d$, is scaled into $[-1, 1]$ using min-max normalization using Eqn. (3.3) [158]:

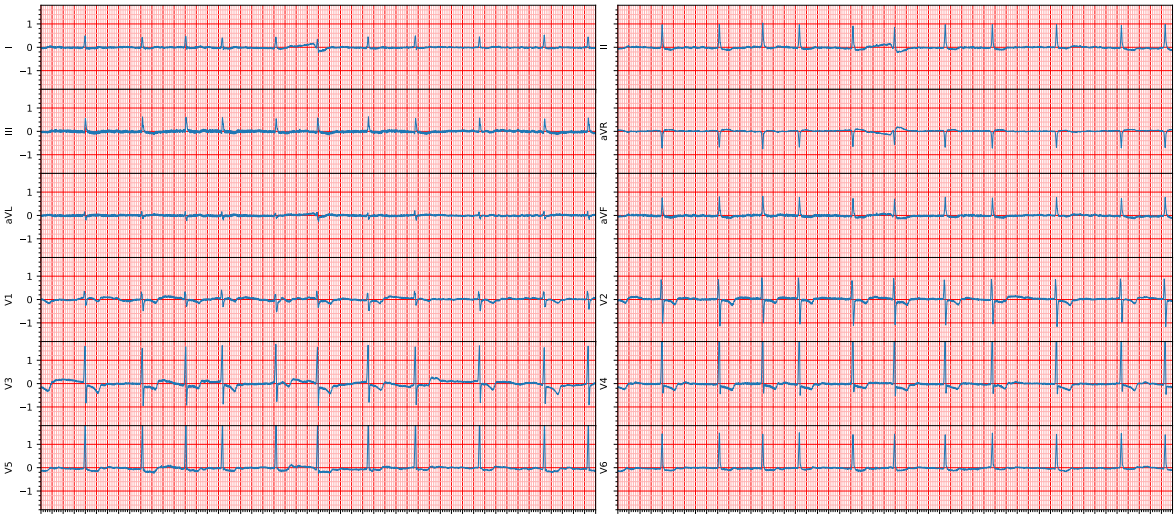


Figure 3.9: ECG record from Chapman Arrhythmia ID MUSE_20180112_071020_91000

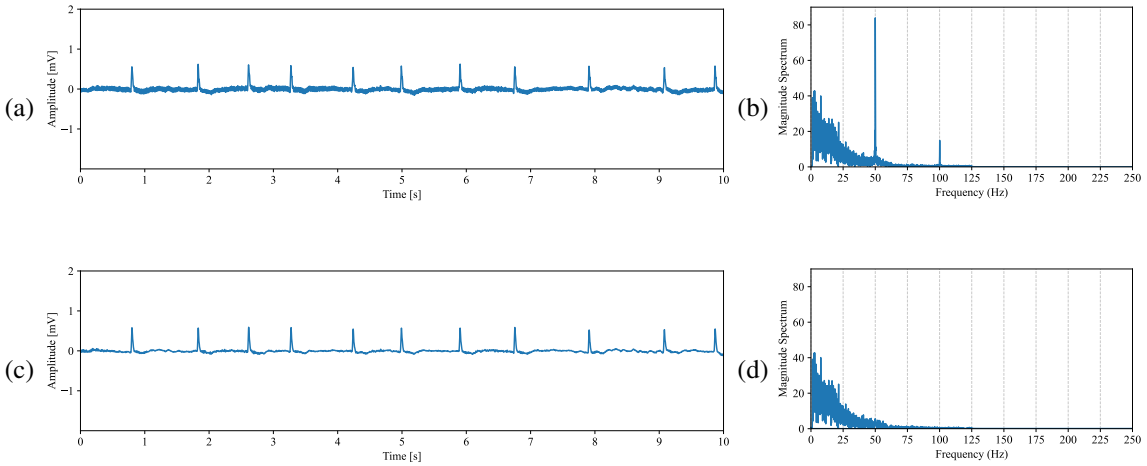


Figure 3.10: Filtering result of Lead III of Chapman Arrhythmia dataset shown in Figure 3.9: (a) PLI contaminated noisy signal, (b) the frequency magnitude distribution of (a), (c) filtered signal, and (d) the frequency magnitude distribution of (c).

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x}_d - x_{d_{min}}}{x_{d_{max}} - x_{d_{min}}}(x_{new_{max}} - x_{new_{min}}) + x_{new_{min}} \tag{3.3}$$

where $\mathbf{x}_{\text{norm}} \in \mathbb{R}^{1\times N}$ is a normalized signal, $x_{new_{min}} = -1$ and $x_{new_{max}} = 1$ are the new minimum and maximum values, respectively. Besides, $x_{d_{max}}$ and $x_{d_{min}}$ are the maximum and minimum values in $\mathbf{x}_d$, respectively. Then, the normalized 12-lead ECG signal which is obtained by concatenating $\mathbf{x}_{\text{norm}}$ is $X_{\text{norm}} \in \mathbb{R}^{12\times N}$.

In the preprocessing stage, BW artifact compensation, PLI noise removal, and normalization techniques have been successively used to improve ECG data quality.

3.5. Model

In this study, we designed a multichannel hybrid model that utilizes the 12 channels of a 12-lead ECG signal and includes post-hoc model output interpretability. Each lead provides a distinct 1D signal, contributing its share of information in identifying underlying heart diseases. The detailed architecture and configuration of the model is illustrated in Figure 3.11. To extract and capture features present in each of these 12 channels, we used 12 blocks of 1D CNN-BiLSTM hybrid networks and subsequent attention mechanism and 2D CNN blocks.

The first 1D CNN block, each having 24 filters organized into 4 groups, each group consisting of 6 filters, and dilation rates of 1, 2, 3, and 4 are applied to each group. The dilated convolution results of these four groups are then concatenated before batch normalization. This architecture enables the model to capture multi-scale temporal features from an ECG signal at different resolutions by increasing the receptive field [209]. Given a dilation factor m and a kernel weights $\mathbf{w}^{\text{conv}}$, the output $\mathbf{v}$ from a dilated convolutional operation on a normalized single lead signal $\mathbf{x}_{\text{norm}}$ is determined using Eqn.(3.4) [209]:

$$\mathbf{v}[n] = (\mathbf{x}_{\text{norm}} *_m \mathbf{w}^{\text{conv}})[n] = \sum_{k=0}^{k_s-1} \mathbf{w}^{\text{conv}}[k]\mathbf{x}_{\text{norm}}[n-mk] \tag{3.4}$$

where k_s is kernel size, $*_m$ refers to a dilated convolution, and n is in the interval $[0, N-1]$ is an index of a single lead ECG signal sample.

As shown in Figure 3.11, the proposed model has two blocks with a projection skip connections. These connections employ a 1D convolutional layer to transform the features from earlier layers before performing element-wise addition with the main path. This ensures dimensional compatibility and enables the model to effectively combine both low-level and high-level representations. Besides, the skip connections improve the gradient flow and model convergence that enhances the model performance [210, 211].

In the proposed model architecture, the 12 1D-CNN blocks are followed by two BiLSTM networks. The BiLSTM network has two stacked LSTM networks, forward LSTM and backward LSTM, used to capture different temporal characteristics within features at the end of 1D CNN blocks. The forward LSTM captures relationships within the extracted features by the 1D-CNN blocks, extending from the beginning to the end of the feature sequence.

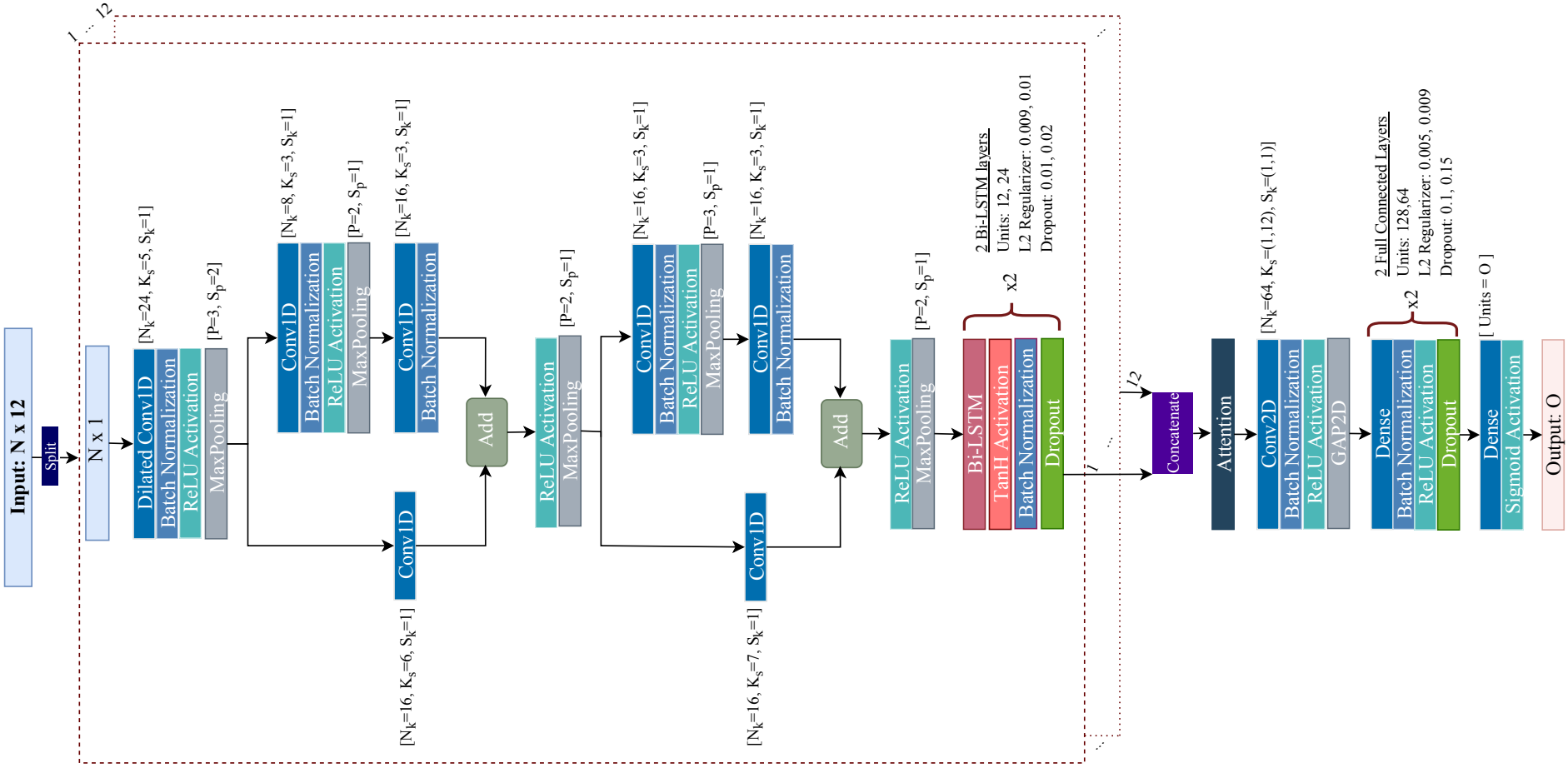


Figure 3.11: The Proposed Model Architecture.

Where N_k is the number of kernels/filters, K_s is the kernel size, S_k is the kernel stride, P is the pooling size, S_p is the pooling stride, $N \times 12$ denotes the input dimension of the 12-lead ECG signal, $N \times 1$ represents a single lead of the input ECG, and O is the output.

The backward LSTM, on the other hand, captures the temporal relationships within the features in the reverse order [212, 213].

Then, the outputs of all the 12 BiLSTM networks are concatenated to form a tensor, $\mathbf{H} \in \mathbb{R}^{\mathfrak{c} \times s \times u}$, where $\mathfrak{c}$ represents the 12 ECG channels, s is the output sequence length of the BiLSTM network which is in the time-step dimension and u is number of features for each element in the output sequence whose dimension equals to BiLSTM units of the second BiLSTM layer. However, all values of this tensor do not have equal contributions in the final classification task. So, we added a multiplicative attention mechanism in the proposed network to emphasize the tensor that contributes most to the classification. The attention layer is implemented by adapting the mechanisms proposed by Yao *et. al.* [214] and Vaswani *et. al.* [215]. Given weight tensor ($\mathbf{W} \in \mathbb{R}^{u \times u}$) and bias vector ($\mathbf{b} \in \mathbb{R}^{1 \times 1 \times u}$) in the attention network, the query tensor ($\mathbf{Q} \in \mathbb{R}^{\mathfrak{c} \times s \times u}$) for $\mathbf{H}$ is defined as in Eqn. (3.5):

$$\mathbf{Q} = \tanh(\mathbf{H} \cdot \mathbf{W} + \mathbf{b}) \tag{3.5}$$

Then, the query tensor obtained in Eqn. (3.5) is used to determine the attention logits as given in Eqn. (3.6):

$$\mathbf{A} = \mathbf{Q}_r \cdot \mathbf{K}^T \tag{3.6}$$

where $\mathbf{A}$ is the attention logits tensor that shows the similarity score between query and key tensors, $\mathbf{Q}_r \in \mathbb{R}^{u \times \mathfrak{c}}$ is obtained by reshaping the query tensor $\mathbf{Q}$, and $\mathbf{K} \in \mathbb{R}^{u \times \mathfrak{c}}$ is a key tensor obtained by keeping the feature and channel dimensions of $\mathbf{H}$. This is mainly because each element in the sequence is considered as an individual data point. The attention weight $\mathbf{W}_{attn} \in \mathbb{R}^{s \times u \times u}$ is then obtained by applying softmax on a reshaped attention logits tensor, that is, $\mathbf{A}_r \in \mathbb{R}^{s \times u \times u}$ along the last axis as given in Eqn. (3.7):

$$\mathbf{W}_{attn} = softmax(\mathbf{A}_r) \tag{3.7}$$

Finally, the output of the attention network, $\mathbf{Y}_{attn} \in \mathbb{R}^{\mathfrak{c} \times s \times u}$, is a attention weighted version of $\mathbf{H}$ as in Eqn. (3.8):

$$\mathbf{Y}_{attn} = (\mathbf{W}_{attn} \cdot \mathbf{H}^T)^T \tag{3.8}$$

where $\mathbf{H}^T \in \mathbb{R}^{s \times u \times \mathfrak{c}}$ and $\mathbf{Y}_{attn}$ is the attended input obtained through weighting the input, $\mathbf{H}$, by the attention weight, $\mathbf{W}_{attn}$.

The output of the attention layer is then fed to the 2D CNN network. Adding the 2D CNN in the network helps to capture spatial features in an ECG signal [200]. Besides, the 2D CNN layer with the global average pooling (GAP) activation makes the model interpretable through visualizing and localizing important features that contribute most to the classification [59, 216]. Given a tensor $\mathbf{F} \in \mathbb{R}^{l_r \times l_\mathfrak{c} \times u \times n_f}$ that incorporates n_f number of filters with kernel size $l_r \times l_\mathfrak{c}$ that transforms feature maps from u to n_f dimension. The output of the convolution operation, $\mathbf{E} \in \mathbb{R}^{(\mathfrak{c}-l_r+1)\times(s-l_\mathfrak{c}+1)\times n_f}$, is represented as in Eqn. (3.9):

$$\mathbf{E}_{i,j,n_f} = \sum_{u} \sum_{m=0}^{l_r-1} \sum_{n=0}^{l_\mathfrak{c}-1} (\mathbf{Y}_{\text{attn}}[i+m, j+n, u] \cdot \mathbf{F}[m,n,u,n_f]) + \mathbf{b}_{n_f} \tag{3.9}$$

where i and j are spatial coordinates of $\mathbf{E}$ with dimension $(\mathfrak{c}-l_r+1) \times (s-l_\mathfrak{c}+1) \times n_f$, $\mathbf{b}_{n_f}$ is a bias vector value that contributes to each feature map in $\mathbf{F}$. In the proposed architecture, $l_r = 1$ and $l_\mathfrak{c} = 12$ for a kernel size $k_s = (1, 12)$ and the number of filters $n_f = 64$. This convolution operation extracts features and preserves the information along the 12 leads ($\mathfrak{c}$).

The extracted features from previous layers are processed through a rectifier linear unit (ReLU) activation function and then subjected to 2D global average pooling (GAP). The GAP computes the average value of each feature map that preserves spatial information and reduces parameter size. Then, the resulting features are fed into three dense or fully connected (FC) layers. These layers contribute to the overall model's capacity to learn patterns in the ECG signals. The first and second dense layers have 128 and 64 neurons or units, respectively. The third (output) FC layer has a number of neurons equal to the class labels (O).

As shown in Figure 3.11, the proposed model integrates different activation functions and regularization methods across the architectures, including batch normalization, dropout, and $L2$ weight regularizers. The activation functions enable the model to approximate complex non-linear relationships in the dataset [217]. Similarly, the dropout and $L2$ weight regularizers can reduce overfitting and enable the model to be robust against variations in the input data [217]. The batch normalization is used to mitigate the internal covariate shifts,

vanishing and exploding gradients, and sensitivity of the network to initialization [209,217]. The batch normalization in each convolutional and FC layer, except the output FC layer, is followed by the ReLU activation function. The output FC layer, on the other hand, uses the sigmoid activation function [217]. Besides, a hyperbolic tangent (tanh) activation function [212] is used inside the BiLSTM output cells. The model hyperparameters including the number of kernels, kernel sizes, strides, pool sizes, and the number of units in both Bi-LSTM and dense layers were adjusted for better performance using empirical tuning based on related studies [43, 80, 218]. In addition, Bayesian optimization with expected improvement acquisition function [197] is used to fine-tune the dropout rate and weight regularization factors in both Bi-LSTM and FC layers.

3.6. Model's Classification Output Interpretability Methods

In an ECG-based heart disease classification, DL-based techniques have shown commendable performances [20, 43, 44, 47, 50, 80]. However, interpreting the classification output of DL models remains challenging due to their black-box nature [95]. As a result, various DL model output interpretation methods have been proposed in the literature to understand the rationale behind the classification outputs. In this study, we explored the interpretation results produced by two post-hoc model interpretability techniques: Grad-CAM++ and SHAP with gradient explainer.

SHAP is a feature attribution-based method that shows the contribution of each 12-lead of an ECG signal with their time segment in the final classification output. On the other hand, Grad-CAM++ is a gradient-based method that generates class-discriminative localization maps for highlighting the contribution of each time segment and lead of the ECG signal. It operates by computing gradients following the final convolutional layer of the CNN [59].

CHAPTER 4

Experimental Results and Discussion

4.1. Overview

One of the limitations of ML models is to replicate their best performance across different ECG datasets [43]. In this regard, we present experimental results of the model after being trained on the PTB-XL [41], the CODE-15% [61], and the Chapman Arrhythmia [62] datasets that are discussed on section 3.3. The aim is to effectively replicates the performance of the proposed model across three distinct ECG datasets. These datasets have different sampling frequencies, diverse cardiac diagnosis classes, and contamination by artifacts.

This chapter provides a comprehensive examination of the experimental setup, results, and insights into the performance of the proposed model for heart disease diagnosis using 12-lead ECG signals. The experimental setup section discusses on the hardware and frameworks used, as well as the model training details. Then, the metrics used for evaluating the model's performance are discussed. Then, experimental result analysis is presented to show the model's performance across the three heart disease datasets. Then, a model component analysis is presented to evaluate the contribution of individual components to the overall performance improvement. This analysis sheds light on the effectiveness of key design choices and their impact on model performance. Finally, the proposed model's performance is compared with state-of-the-art techniques applied to large 12-lead ECG datasets and a discussion is provided in demonstrating its competitive performance and potential advantages in clinical and real-world applications.

4.2. Experimental Setup

The proposed model was evaluated on three different ECG datasets separately. This approach provides insight into the data-specific characteristics of the model. Besides, it facilitates performance comparison against models trained on these datasets.

A total of four different experiments were conducted to evaluate the proposed model. The first two experiments were performed on the Chapman Arrhythmia dataset with seven (7) and merged four (4) heart disease classes. The third experiment studied the model performance on the PTB-XL dataset with 5 super-diagnostic classes. Finally, the fourth experiment was on the CODE-15% dataset. In these datasets, a single instance of an ECG signal may exhibit multiple diseases, making the classification a multi-label classification task. The proposed classification model was trained using a sigmoid cross-entropy loss function, optimized by the Adams optimizer, and an initial learning rate of 0.0005. The learning rate was halved when the validation loss did not improve for three consecutive epochs. In all experiments, the training was set to run for a maximum of a hundred epochs, with a batch size of 32. Besides, the training process stopped early if there was no improvement in the validation loss for seven consecutive epochs. For the CODE-15% dataset, the number of epochs maintained for learning rate reduction and early stopping were four and nine, respectively.

In all these experiments, the proposed model was trained and tested on a machine equipped with hardware specifications of Intel(R) Xeon(R) Gold 6278C CPU with 2.60 GHz clock rate, 64 GB RAM, and 16 GB Tesla T4 GPU. The proposed model was implemented using Keras API on TensorFlow framework version 2.11.0.

4.3. Performance Evaluation Metrics

The proposed model performance is evaluated using standard metrics commonly used in multi-labelled and multi-class classification tasks [219–221]. Given, $O \in \mathbb{R}^c$ a model classification output with c number of classes, the classification performance is usually evaluated using the individual class O_i in $\{O\}_{i=1}^{c}$. The evaluation metrics, $Accuracy_i$, $Specificity_i$, $Precision_i$, and $Recall_i$ are, respectively, defined in Eqn. (4.1), (4.2), (4.3), and (4.4) in terms true positive (TP_i), false negative (FN_i), true negative (TN_i), and false positive (FP_i). The true instances represent correct model predictions that correspond to the actual test data la-

bels, whereas false instances represent faulty model predictions that correspond to the true data labels.

Accuracy$_i$ measures the per-class classification effectiveness of the proposed model and computed as in Eqn. (4.1):

$$Accuracy_i = \frac{TP_i + TN_i}{TP_i + FN_i + FP_i + TN_i} \tag{4.1}$$

Specificity$_i$ measures the effectiveness of the proposed model in identifying cases that are negative in test data for disease class i, and defined as Eqn. (4.2):

$$Specificity_i = \frac{TN_i}{TN_i + FP_i} \tag{4.2}$$

Precision$_i$ measures the agreement of the positive cases detected by the proposed model with the actual positive cases in the test data for disease class i, and defined as Eqn. (4.3):

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{4.3}$$

Recall$_i$ measures the model's ability to capture all positive cases from the test data for disease class i. It is also known as sensitivity or true positive rate, which shows the model's performance in identifying true positive cases in the test data. It is defined as in Eqn. (4.4):

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{4.4}$$

F1-score$_i$ represents the harmonic mean of Precision$_i$ and Recall$_i$ for a model in identifying true positive instances in the test data for class i. It is defined as in Eqn. (4.5):

$$F1\text{-}score_i = 2 \cdot \frac{Recall_i \cdot Precision_i}{Recall_i + Precision_i} \tag{4.5}$$

Other metrics that measure model performance are the area under the receiver-operating characteristic curve (AUC) and the area under the precision-recall curve (AUPRC). The AUC measures the ability of a model to distinguish between classes by plotting the True Positive Rate (TPR) or $Recall_i$ against the False Positive Rate (FPR), which is $1 - Specificity_i$. A

higher AUC value indicates better class separation, where an AUC of 0 signifies a completely incorrect model, an AUC of 0.5 implies no ability to separate classes (equivalent to random guessing), and an AUC of 1 reflects perfect classification performance, with no overlap between classes. The AUC is particularly useful in imbalanced datasets, as it provides a balanced evaluation by focusing on both the sensitivity and specificity of the model. This makes it a robust metric for assessing a model's ability to identify both positive and negative cases accurately, irrespective of class proportions in the data.

The AUPRC, on the other hand, is more focused on evaluating the model's performance in identifying positive instances, especially in datasets where the positive class is rare. It measures the relationship between Precision (positive predictive value) and Recall (sensitivity) by emphasizing the need to minimize false positives while maximizing true positives. The value of AUPRC ranges from 0 to 1, with 1 indicating perfect performance where all positive instances are correctly identified without any false positives. Unlike AUC, AUPRC is particularly informative when dealing with imbalanced datasets, as it highlights the model's effectiveness in recognizing the minority class [222].

The overall classification performance of the model is measured in terms of macro-averaging, micro-averaging and weighted-averaging of the metrics discussed above. Macro-averaging treats all classes independently and computes the average value across all disease classes. On the other hand, micro-averaging computes the metrics globally on cumulative TP, FN, TN, and FP of all disease classes and it takes class imbalance into consideration [222]. Whereas, in weighted-averaging (wavg) measures the average of model performance metrics which is weighted by the TP of each disease class in the dataset. The macro precision (Precision_{macro}) and macro recall (Recall_{macro}) are determined using Eqn. (4.6) and Eqn. (4.7), respectively. Whereas, the micro precision (Precision_{micro}) and micro recall (Recall_{micro}) are determined using Eqn. (4.8) and Eqn. (4.9), respectively. Besides, the weighted precision (Precision_{wavg}) and weighted recall (Recall_{wavg}) are determined using Eqn. (4.10) and Eqn. (4.11), respectively, where N_i is the count of samples of class i in the test data. The micro, macro, and weighted averages F1-scores are computed based on their corresponding: micro, macro, and weighted averages of recall and precision, utilizing equivalent equation defined in Eqn. (4.5).

$$Precision_{macro} = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i+FP_i}}{c} \tag{4.6}$$

$$Recall_{macro} = \frac{\sum_{i=1}^{c} \frac{TP_i}{TP_i+FN_i}}{c} \tag{4.7}$$

$$Precision_{micro} = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} (TP_i + FP_i)} \tag{4.8}$$

$$Recall_{micro} = \frac{\sum_{i=1}^{c} TP_i}{\sum_{i=1}^{c} (TP_i + FN_i)} \tag{4.9}$$

$$Precision_{wavg} = \frac{\sum_{i=1}^{c} (Precision_i \cdot N_i)}{\sum_{i=1}^{c} N_i} \tag{4.10}$$

$$Recall_{wavg} = \frac{\sum_{i=1}^{c} (Recall_i \cdot N_i)}{\sum_{i=1}^{c} N_i} \tag{4.11}$$

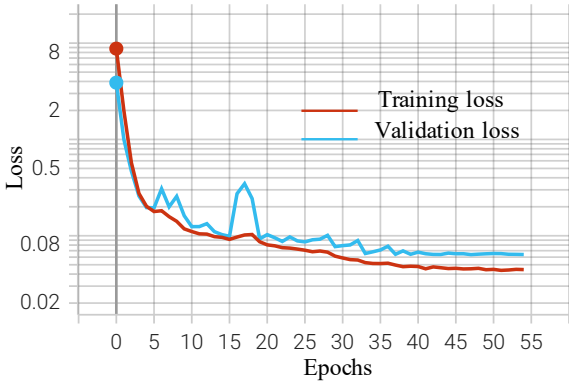
4.4. Result Analysis

4.4.1 Performance on Chapman Arrhythmia Dataset

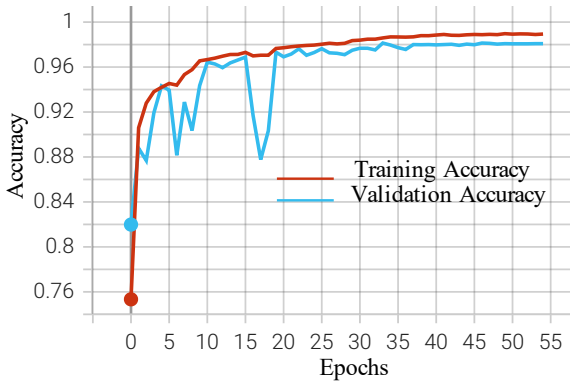
For the Chapman Arrhythmia dataset [62], the proposed model's classification performance was evaluated on reduced seven (Chapman-Reduced) and merged four (Chapman-Merged) disease classes. In both cases, the dataset was split with a ratio of 80%, 10%, and 10% into training, validation, and testing sets, similar to the strategy employed by Yildirim *et al.* [80].

Figure 4.1 shows the training and validation curves obtained from the model training process. These curves show that the proposed model optimizes its parameters after 20 epochs and convergence to the training and validation data. Then, the trained model performance was evaluated on an unseen test dataset. As depicted in Table 4.1, the model demonstrated a high generalization capability with an average classification accuracy of 98.55% and an average specificity of 99.20%. Nevertheless, the model struggles with AF and SI disease classes that

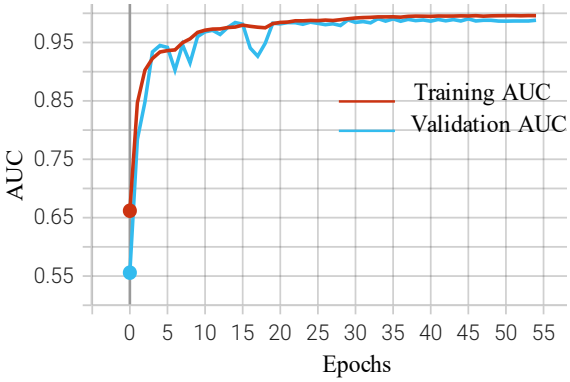
are few in number in datasets. However, as shown in Figure 4.2 the AUC curve indicates the proposed model's capability to distinguish between positive and negative instances of disease classes with an average AUC value of 99.58%. The confusion matrix, shown in Figure 4.3, also indicates the model's classification performance in capturing the true positives within each disease class.



(a) Loss curve



(b) Accuracy curve



(c) AUC curve

Figure 4.1: Training and validation curves of Chapman-Reduced Arrhythmia dataset: (a) loss curve, (b) accuracy curve, and (c) AUC curve.

In the seven (7) class Chapman-Reduced dataset, the model exhibits challenges in classifying AF and SI disease classes. The challenge is mainly due to the smaller number of samples and high similarity between AF and AFIB, as well as between SR and SI. To address the class imbalance and class similarity issues, the dataset's authors [62] recommended grouping the entire dataset into four classes: AFIB, GSVT, SB, and SR, as discussed in section 3.3. Then, we trained the model on this merged four (4) disease classes, and Figure 4.4a, Figure 4.4b, and Figure 4.4c depicts the training and validation curves. The curves indicate the

Table 4.1: The model's performance in (%) on test set of the Chapman-Reduced Arrhythmia dataset.

Classes	Accuracy	Specificity	Recall	Precision	F1-Score	AUC	AUPRC
AF	97.43	99.20	58.70	77.14	66.67	98.62	78.96
AFIB	97.72	98.17	95.48	91.35	93.37	99.66	98.29
SI	98.38	99.70	63.16	88.89	73.85	99.41	87.59
SB	99.52	99.40	99.74	98.97	99.35	99.97	99.95
SR	98.29	98.85	95.70	94.68	95.19	99.57	99.01
ST	99.14	99.55	97.04	97.62	97.33	99.96	99.80
SVT	99.33	99.50	96.15	90.91	93.46	99.85	97.36
Average	98.55	99.20					
Macro			86.57	91.37	88.46	99.58	94.42
Micro			94.58	95.22	94.90	99.78	99.01
Weighted			94.58	94.99	94.64	99.76	97.99

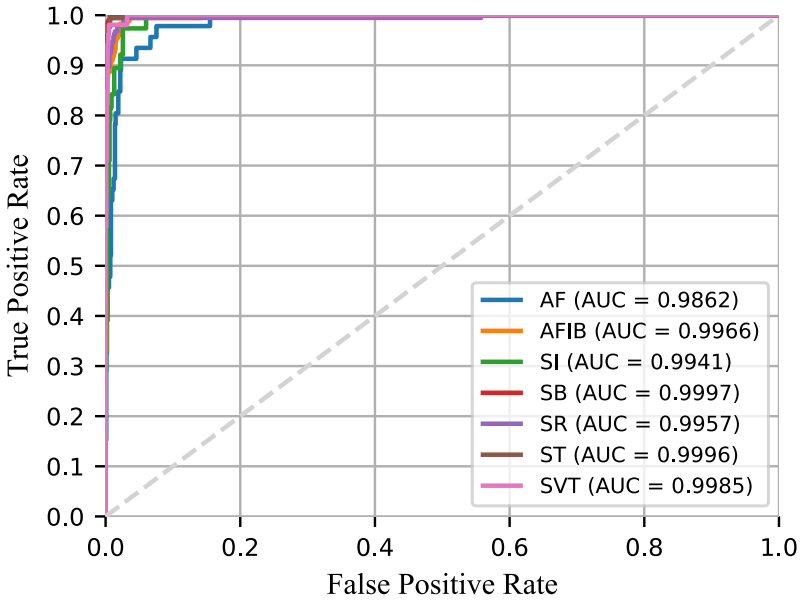


Figure 4.2: Area under receiver operating curve

model's capacity to fit the training and validation sets and showcase its convergence. Then, the trained model was tested on an unseen test set, and Figure 4.5 and 4.6 depicts its classification performance. As depicted on Table 4.2, the model achieved an average accuracy of 98.80%. The model demonstrates robust performance on the test set with a micro AUC of 99.77% and a macro AUC of 99.75%, indicating the high discriminating capability of the model across all classes. Additionally, the model achieves AUPRC with a micro AUPRC of 99.36% and a macro AUPRC of 99.18%, indicating the model's capability in identifying TP and minimizing FP across the test set. Also, these results are reflected in the AUC-ROC curve and the confusion matrix shown on Figure 4.5 and 4.6, respectively.

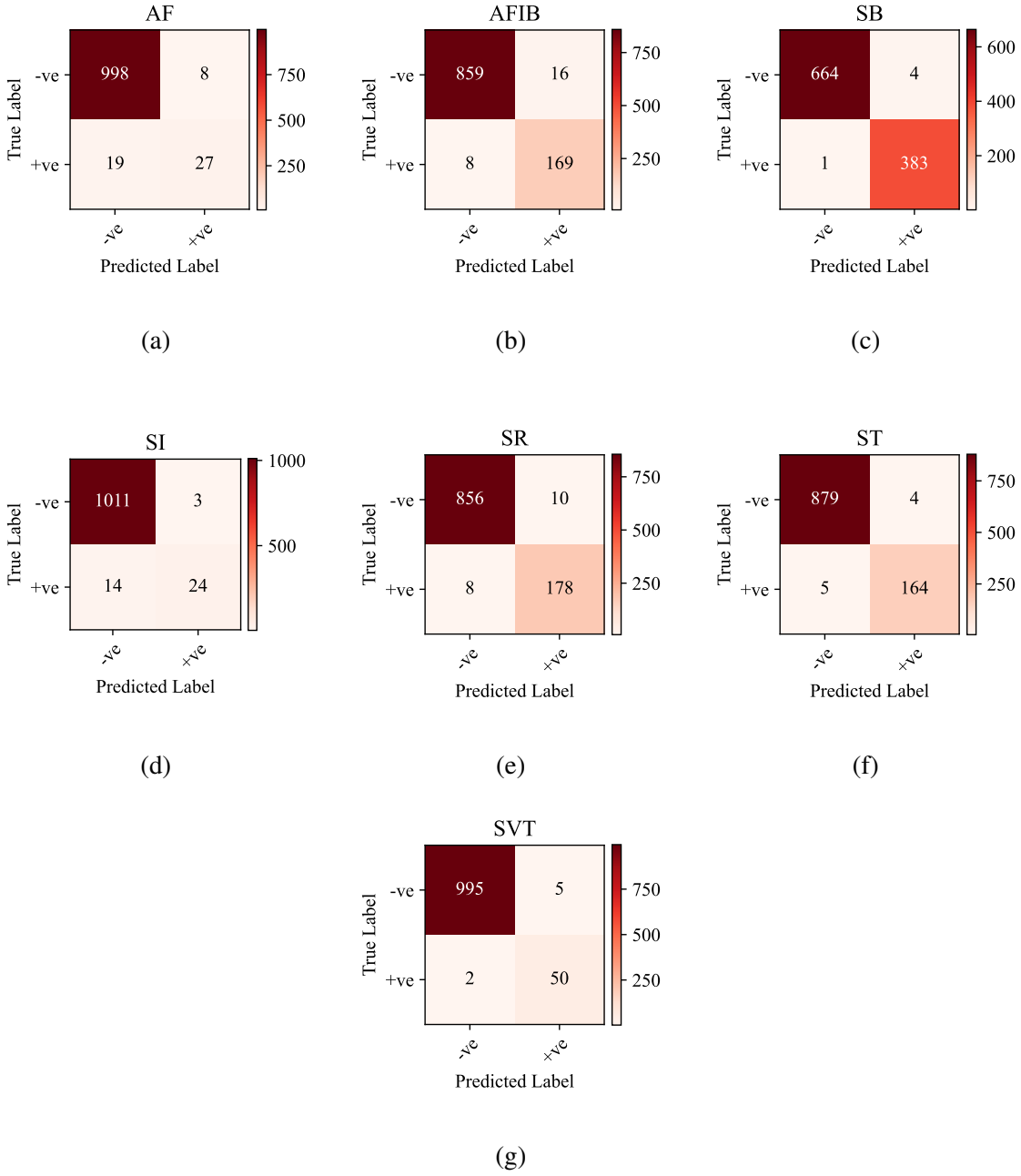
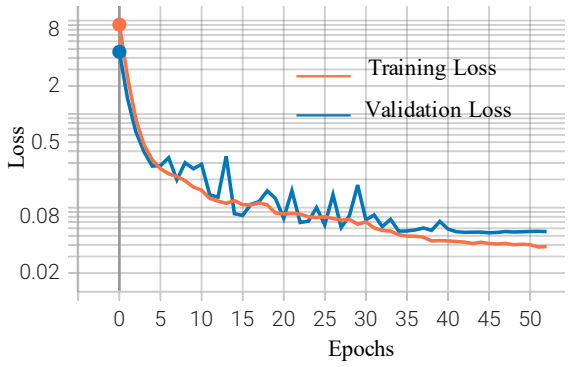


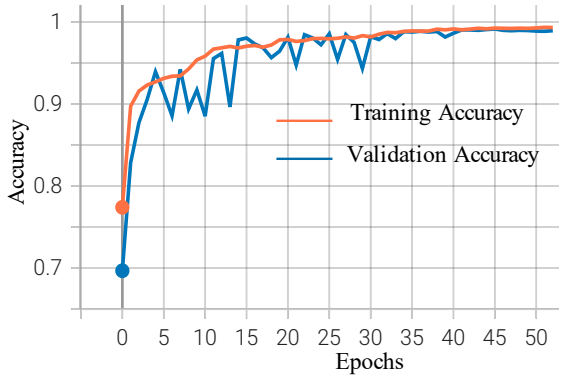
Figure 4.3: Confusion matrix of the test set of the Chapman-Reduced dataset: (a) AF, (b) AFIB, (c) SB, (d) SI, (e) SR, (f) ST, and (g) SVT

Furthermore, to demonstrate the effectiveness of the proposed model, t-distributed stochastic neighbor embedding (t-SNE) [223] visualization is performed on the output of the GAP2D layer which is before the fully connected layer, the test dataset before the input layer, and output layers for the Chapman-Reduced and Chapman-Mereged as shown on Figure 4.7 and Figure 4.8, respectively.

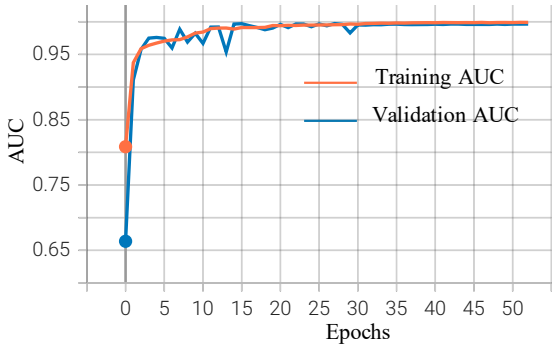
For the Chapman-Reduced, Figure 4.7a shows the test dataset complexity with an undefined



(a) Loss curve



(b) Accuracy curve



(c) AUC curve

Figure 4.4: Training and validation curves for Chapman-Merged Arrhythmia dataset: (a) loss curve, (b) accuracy curve, and (c) AUC curve.

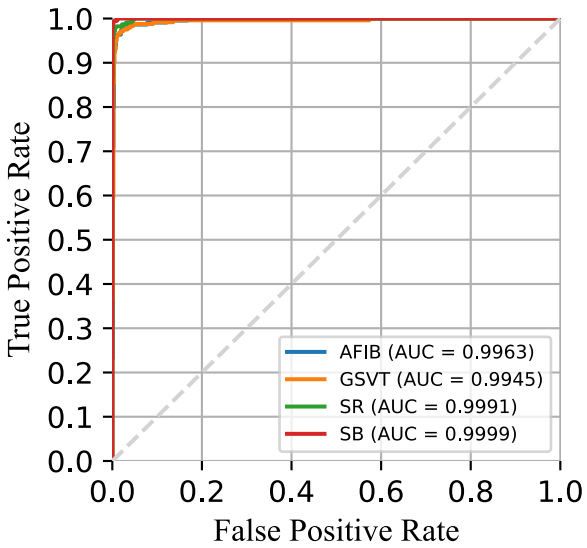


Figure 4.5: Area under receiver operating curve of Chapman-Merged Arrhythmia test set

cluster of data points. Then, on the output of the GAP2D layer of the proposed model, as shown in Figure 4.7b, the seven disease classes are grouped with clear boundaries. The t-SNE plot indicates that the feature extraction process has effectively captured important

Table 4.2: The model's performance in (%) on test set of the Chapman-Merged Arrhythmia dataset.

Classes	Accuracy	Specificity	Recall	Precision	F1-Score	AUC	AUPRC
AFIB	98.31	98.94	95.89	95.89	95.89	99.63	98.47
GSVT	98.31	98.91	96.28	96.28	96.28	99.45	98.57
SB	99.06	99.76	96.44	99.09	97.75	99.91	99.70
SR	99.53	99.56	99.47	99.21	99.34	99.99	99.98
Average	98.80	99.29					
Macro			97.02	97.62	97.32	99.75	99.18
Micro			97.37	97.83	97.60	99.77	99.36
Weighted			97.37	97.84	97.60	99.78	99.29

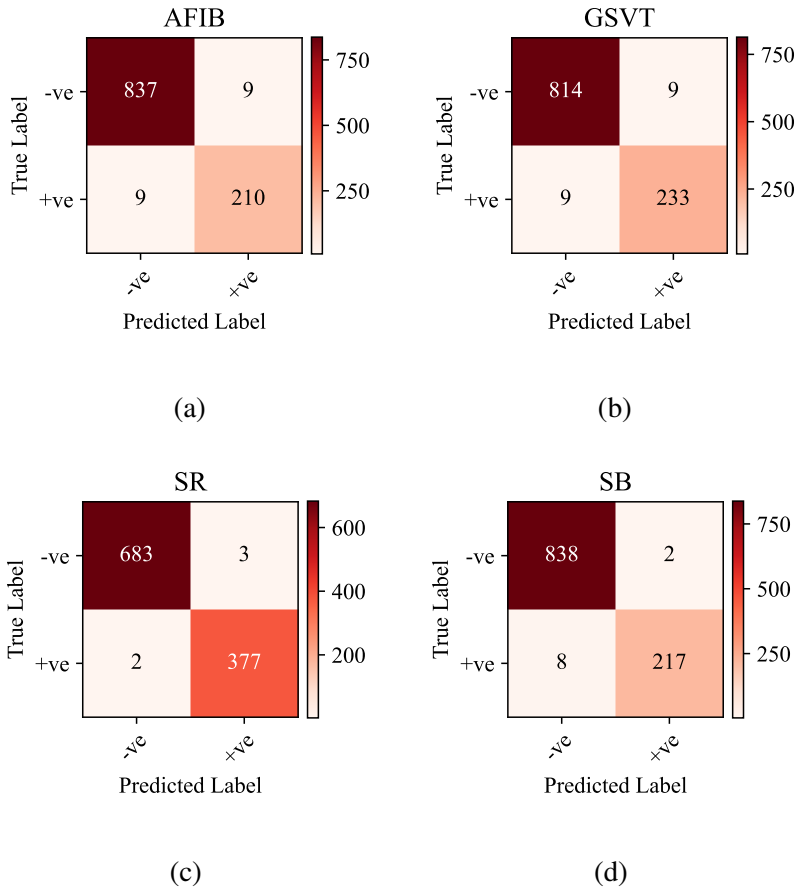


Figure 4.6: Confusion matrix of the test set of the Chapman-Merged arrhythmia dataset: (a) AFIB, (b) GSVT, (c) SR, and (d) SB

information from the raw ECG data. Finally, the t-SNE plot on Figure 4.7c shows class separation achieved by the model at the output layer. Most disease classes, such as SB, ST, and AFIB, form distinct clusters with minimal overlap, indicating that the model has learned distinguishable features for these classes. SR also shows good separability. Minor overlaps

exist, particularly between AF and AFIB, as well as SI and SR, which may suggest some feature similarity leading to potential misclassifications. The compact and isolated clustering of SVT implies it has unique, identifiable features but appears in a smaller number compared to SB, SR, ST, and AFIB. This visualization highlights the model's potential for effective disease classification, though further refinement could improve separability in overlapping classes.

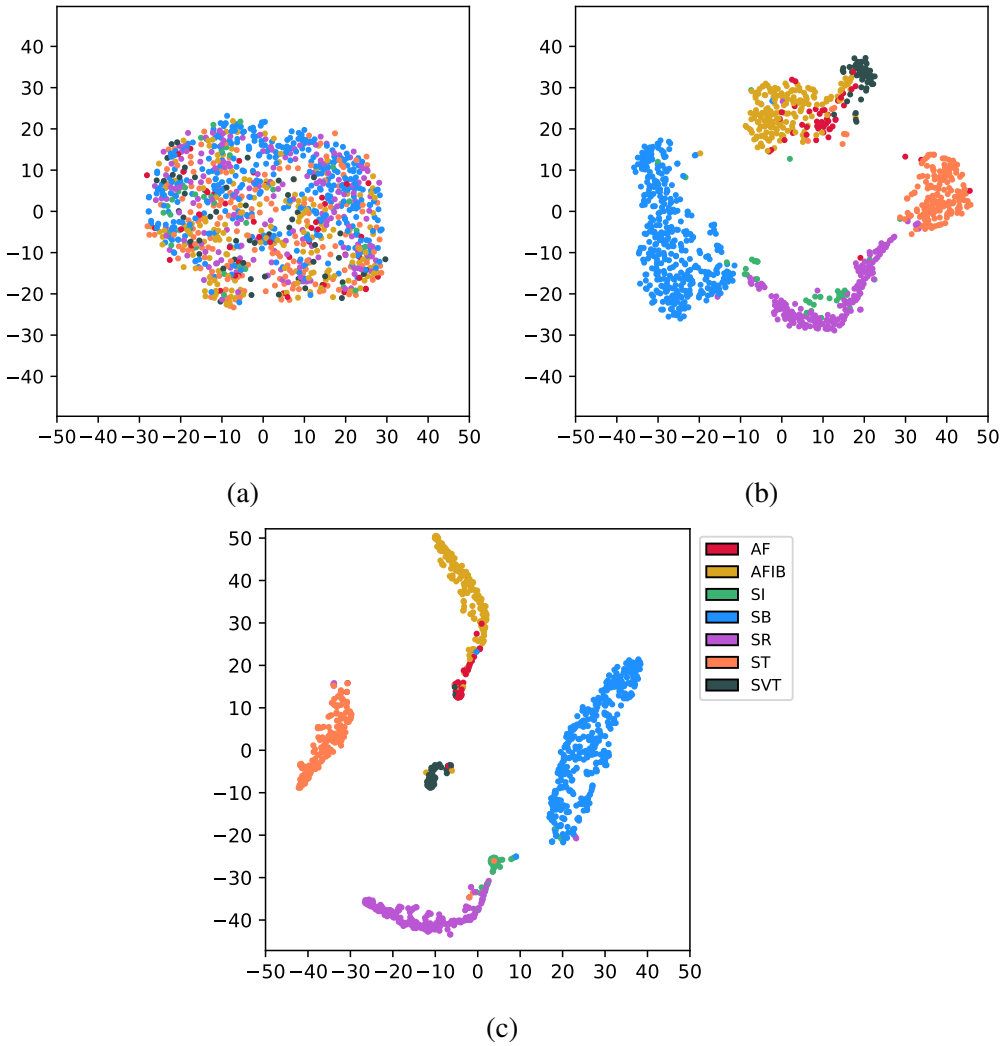


Figure 4.7: t-SNE visualization on Chapman-Reduced test dataset (a) test set ECG samples at the input layer, (b) GAP2D layer, and (c) Output layer of the proposed model.

Similarly, for the Chapman-Merged, Figure 4.8a shows the dataset complexity with an undefined cluster of data points. Then, on the output of the GAP2D layer of the proposed model, as shown in Figure 4.8b, the four disease classes are grouped with clear boundaries. The t-SNE plot indicates that the feature extraction process has effectively captured important information from the raw ECG data. Finally, The t-SNE plot on Figure 4.8c shows class

separation achieved by the model at the output layer. The four disease classes AFIB, GSVT, SB, and SR form distinct clusters, as shown in Figure 4.8c. These distinct clusters show that the model has generalized effectively in classifying these heart diseases.

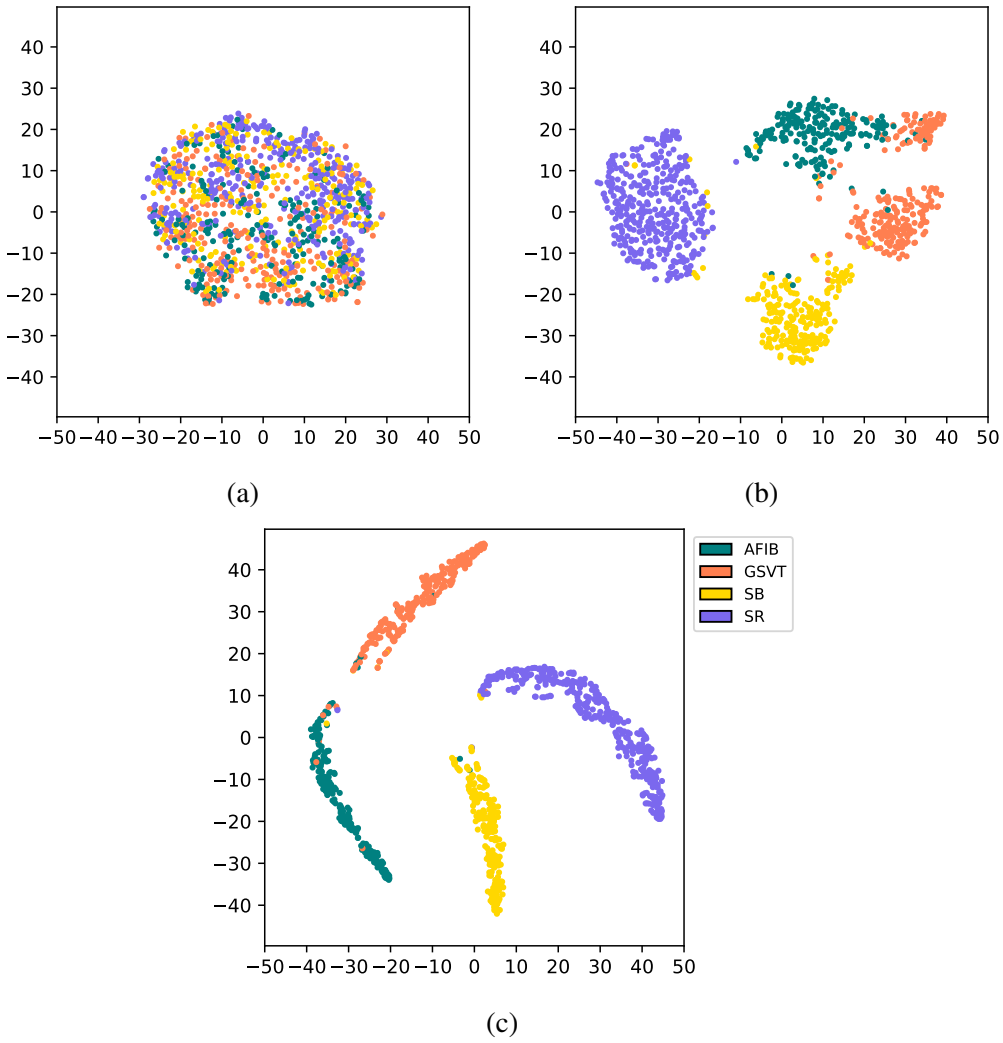


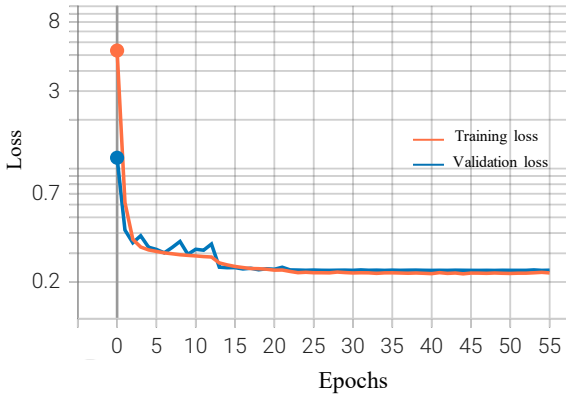
Figure 4.8: t-SNE visualization on Chapman-Merged test dataset (a) test set ECG samples at the input layer, (b) GAP2D layer, and (c) Output layer of the proposed model.

4.4.2 Performance on PTB-XL Dataset

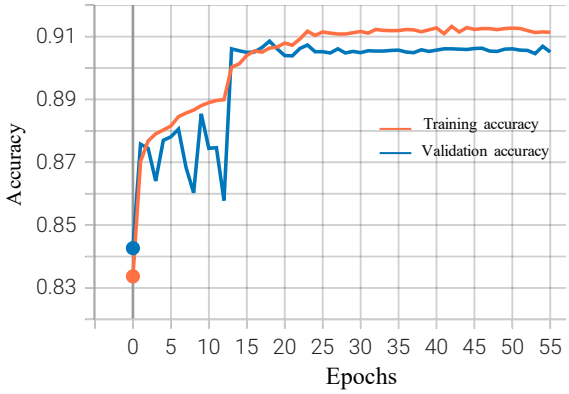
The PTB-XL dataset is divided into training, validation, and test sets based on stratified sampling techniques and provided by Wagner *et al.* [36]. Accordingly, Wagner *et al.* [36] recommended the first nine (9) partitions of the stratified samples for training and validation, whereas the last partition for testing as depicted on Table 3.1.

The training and validation curves derived from the model training procedure are displayed

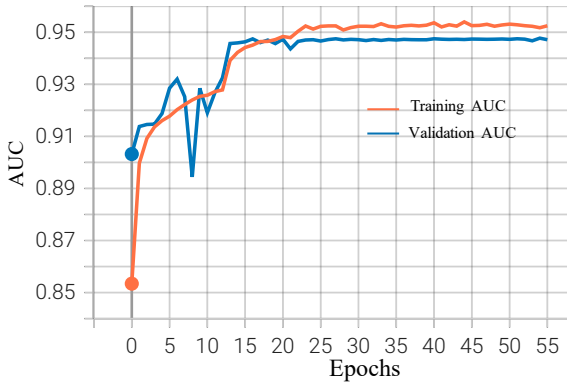
in Figure 4.9. These curves demonstrate that the model optimizes its parameters and converges to the training and validation sets in a few epochs. We assessed the trained model's performance on an unseen test dataset to determine the model's capacity for generalization. Table 4.3 shows that the model had an average test set classification accuracy of 89.84%, indicating a reasonably good performance. Nevertheless, the model struggles to generalize well for HYP class due to the small sample size of HYP in both the training and testing sets. However, the proposed model exhibits strong performance in other disease classes. For instance, the model achieved a recall and an F1-score of 91.08% and 87.23%, respectively, in detecting the NORM cases.



(a) Loss curve



(b) Accuracy curve



(c) AUC curve

Figure 4.9: Training and validation curves for PTB-XL dataset: (a) loss curve, (b) accuracy curve, and (c) AUC curve.

In addition, the AUC values shown in Table 4.3 indicates the model's ability to distinguish among the five disease classes with the micro and macro AUC of 94.51% and 93.51%, respectively. Besides, the confusion matrix depicted in Figure 4.10 shows that the proposed

model performs well in identifying instances that do not belong to a class of interest, that is, negative instances for each disease class. To further evaluate the effectiveness of the proposed model, t-SNE visualization was applied on three stages of the proposed model: the test dataset prior to entering the input layer, the GAP2D layer output which is before the fully connected layer, and the output layer of the proposed model on the PTB-XL dataset, as shown in Figure 4.11.

Table 4.3: The model's performance in (%) on test set of the super-diagnostics class of PTB-XL dataset.

Classes	Accuracy	Specificity	Recall	Precision	F1-Score	AUC	AUPRC
CD	90.20	95.43	72.29	82.19	76.92	92.80	85.59
HYP	92.01	97.16	53.99	72.08	61.74	90.90	68.92
MI	88.79	94.30	72.33	80.97	76.41	94.19	86.42
NORM	88.33	86.20	91.08	83.70	87.23	95.62	93.65
STTC	89.88	94.29	75.82	80.49	78.03	94.05	84.08
Average	89.84	93.48					
Macro			73.08	79.89	76.07	93.51	83.73
Micro			77.69	81.50	79.55	94.51	87.48
Weighted			76.87	81.83	78.75	94.10	86.68

The concentration of data points on the t-SNE plot of the test dataset, as shown in Figure 4.11a, suggests that the ECG data of the PTB-XL dataset is relatively homogeneous, with similarity across points and few variability.

The t-SNE plot after the GAP2D layer, as shown in Figure 4.11b, shows an intermix across most disease classes, with a better clear distinction observed for the CD and NORM disease classes. This pattern suggests that while the feature extraction backbone is generally effective, it excels in identifying patterns specific to CD and NORM. The intermixing among other classes highlights the inherent complexity of the dataset, indicating significant similarities in the features of these heart disease classes. This complexity challenges the model's ability to separate each class.

The t-SNE plot in Figure 4.11c shows distinct clustering for disease classes like CD and MI. While the wide dispersion of NORM points reflects flexibility and generalization capabilities, it also highlights the need for improved specificity. Additionally, the overlap between HYP and STTC disease classes contributes to misclassification. Given the limited variability among the ECG features of the disease classes, as shown in Figure 4.11a, the t-SNE plot

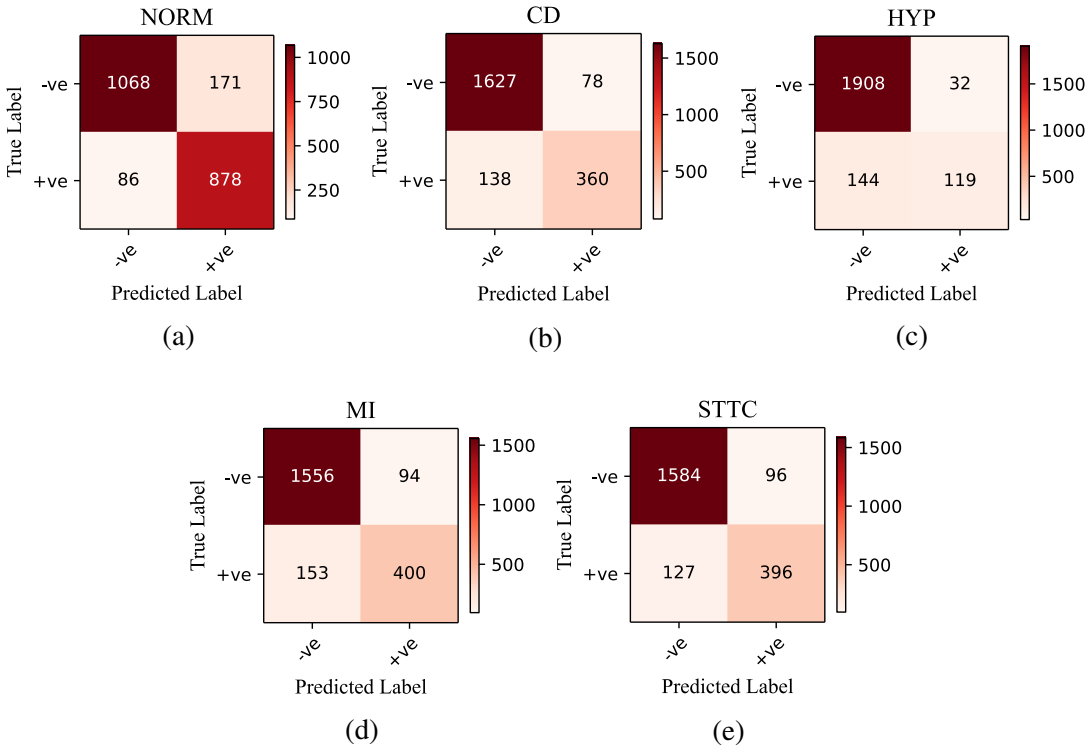


Figure 4.10: Confusion matrix of the test set of the PTB-XL dataset: (a) Normal, (b) Conduction Disturbance, (c) Hypertrophy, (d) Myocardial Infarction, and (e) ST/T Change.

of the proposed model at the output layer, as shown in Figure 4.11c, exhibits its potential in classifying complex conditions. This indicates that, despite inherent similarities in the ECG signals, the model can capture distinctive information between disease classes and its strength in managing complex cases.

4.4.3 Performance on CODE-15% Dataset

As discussed in section 3.3, the CODE-15% dataset comprises a total of 345,779 samples. Among these, 308,004 samples are categorized as not belonging to any of the specified six classes—1dAVb, RBBB, LBBB, SB, ST, and AFIB. In our experimental approach, the focus is on evaluating the proposed model's performance on a specific subset of the dataset. This subset is comprised of 37,775 samples, ensuring each sample contains at least one of the six disease classes. This approach allowed us to assess the model's performance exclusively on this subset of 37,775 samples, all of which include one or more of the specified disease classes. This approach addresses concerns related to data imbalance and computational resource constraints associated with training on the entire dataset

The 37,775 samples were split into training, validation, and test sets, respectively, with a

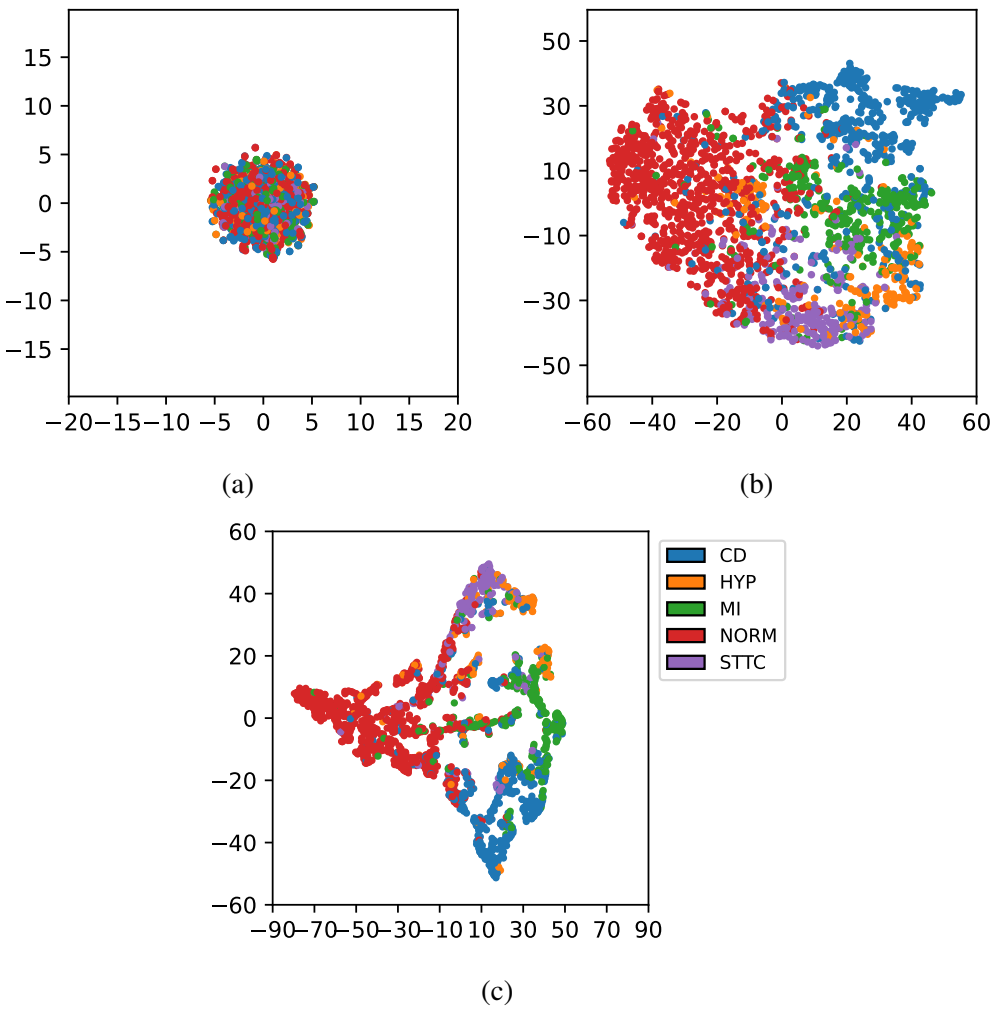
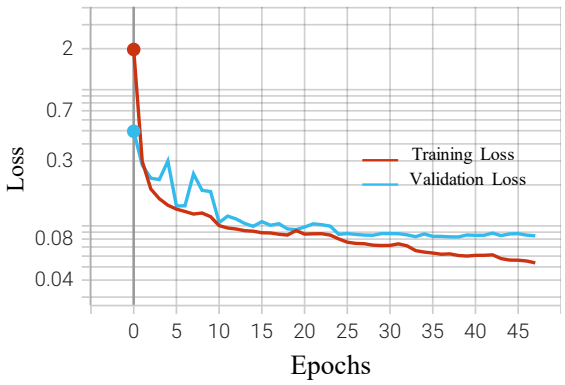


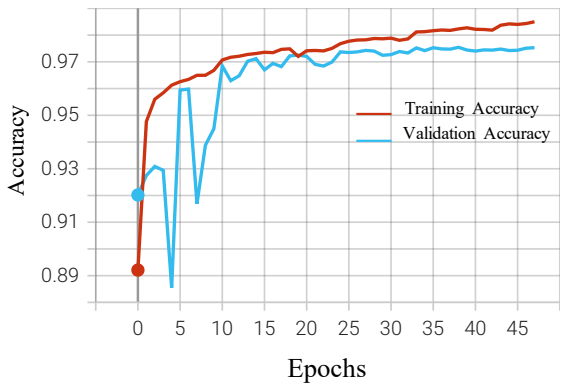
Figure 4.11: t-SNE visualization on PTB-XL test dataset (a) test-set ECG samples at the input layer, (b) GAP2D layer, and (c) Output layer of the proposed model.

ratio of 70%, 15%, and 15%. Figure 4.12 depicts the training and validation curves obtained from the model training. The loss curves in Figure 4.12a show that the training and validation loss is at 0.085 and 0.054, respectively, on the 47th epoch with early stopping. These small loss values, coupled with high accuracy and AUC values shown on curves of Figure 4.12b and 4.12c, indicate the model captures patterns from the data with acceptable discrimination capability between the six disease classes. Besides, the model's performance on the unseen test dataset is given in Table 4.4. The high recall and precision values show the model's capability to spot TP instances from all classes and minimize FP across the test dataset. Similarly, the macro, micro, and weighted F1-scores of 93.71%, 94.02%, and 94.01%, respectively, demonstrate the model's balanced performance in both precision and recall that is also clear in high values of AUPRC across the disease classes.

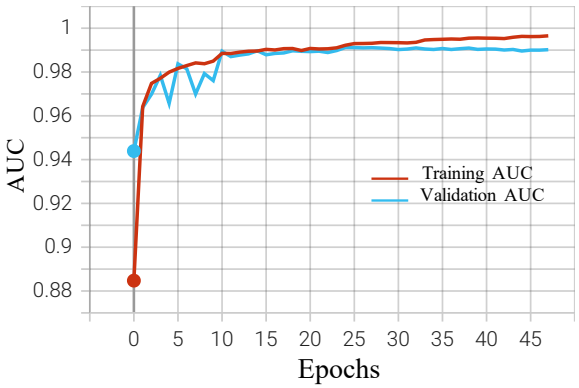
In addition, we evaluated the model performance using a gold standard of 146 samples from



(a) Loss Curve



(b) Accuracy curve



(c) AUC curve

Figure 4.12: Training and validation curves for CODE-15% dataset: (a) loss curve, (b) accuracy curve, and (c) AUC curve.

the CODE-test dataset, detailed in section 3.3 and summarized in Table 3.2. The classification performance is presented in terms of a confusion matrix as depicted in Figure 4.13. The results show that the model captures all true positives in SB and AF disease classes. Moreover, for LBBB and ST, the model does not classify negative samples as positive.

Furthermore, the t-SNE visualization of the GAP2D layer, located before the fully connected layer, along with the test dataset before the input layer and the output layer of the proposed model on the CODE-15% dataset, is presented in Figure 4.14. The colors represent different disease classes shown in the legend.

As shown in Figure 4.14a, the t-SNE distribution of the test set before the input layer reveals that the data points are intermixed, making it difficult to distinguish between the six disease classes. However, after the GAP2D layer, shown in Figure 4.14b, the data points for the disease classes transitioned from overlapping to more separable clusters. The t-SNE plot

Table 4.4: The model's performance in (%) on test set of the CODE15% dataset.

Classes	Accuracy	Specificity	Recall	Precision	F1-Score	AUC	AUPRC
1dAVb	96.85	98.41	88.17	90.86	89.50	99.01	96.23
RBBB	98.40	98.92	96.73	96.57	96.65	99.51	98.88
LBBB	98.13	99.26	92.39	96.04	94.18	99.40	97.78
SB	98.25	99.20	92.55	95.03	93.77	99.63	98.14
AFIB	98.31	99.28	94.68	97.24	95.94	99.62	99.11
ST	96.99	98.16	92.15	92.34	92.24	98.90	96.82
Average	97.82	98.87					
Macro			92.78	94.68	93.71	99.35	97.83
Micro			93.17	94.89	94.02	99.38	98.09
Weighted			93.17	94.89	94.01	99.35	97.94

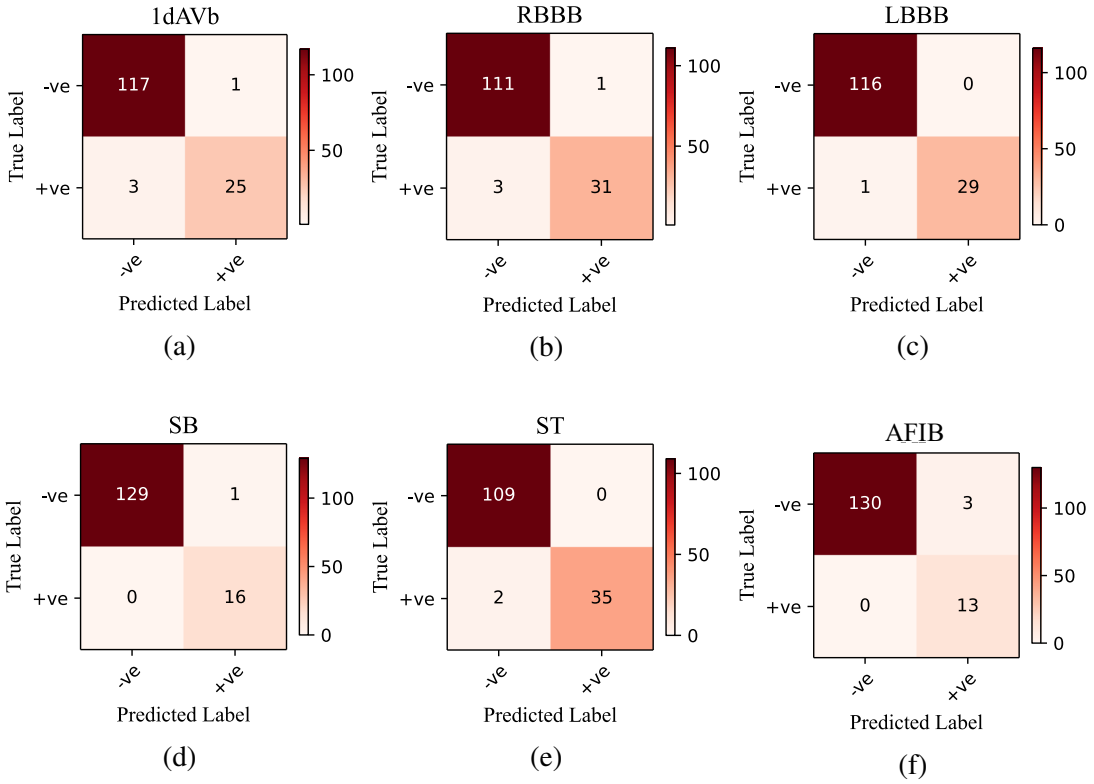


Figure 4.13: Confusion matrix of the CODE-test: (a) 1dAVb, (b) RBBB, (c) LBBB, (d) SB, (e) ST, and (f) AFIB.

of the model's output layer, as shown in Figure 4.14c, indicates that most disease classes, such as ST, AFIB, and LBBB, are well-separated. The t-SNE plot exhibits that the model has effectively learned to extract discriminative features and reduce inter-class overlap. The result shows the effectiveness of the proposed model in capturing important information from ECG signals and its strong classification performance.

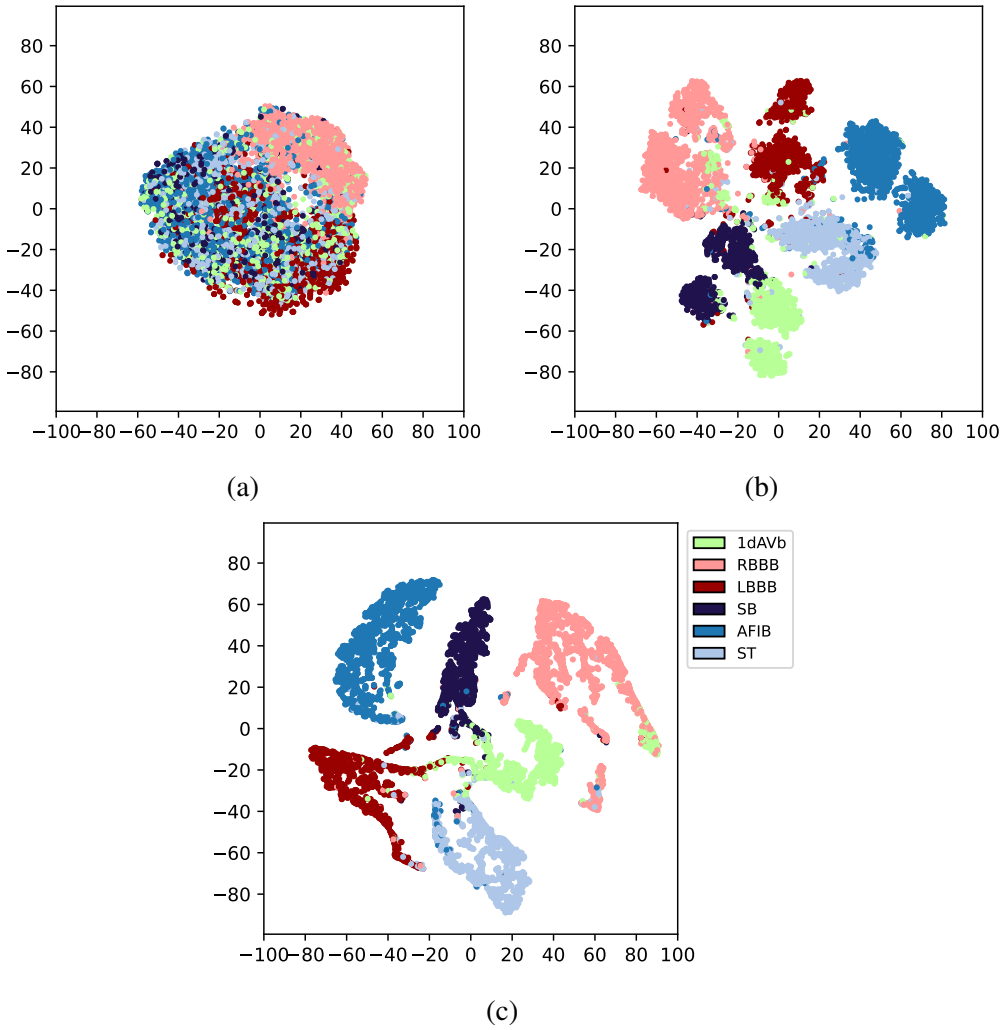


Figure 4.14: T-SNE visualization on CODE-15% test dataset (a) test-set ECG samples at the input layer, (b) GAP2D layer, and (c) Output layer of the proposed model.

4.5. Component Analysis and Ablation Test

This section elaborates the contribution of BiLSTM and attention networks in improving the proposed model generalization capability. As presented on Table 4.5, we trained a model without the BilLSTM-attention network and without the attention network on all three dataset. The performance metrics across three datasets: PTB-XL, Chapman Arrhythmia Merged and Reduced, and Code15% demonstrate how these architectural changes impact accuracy, recall, precision, and F1-score.

The results on the PTB-XL dataset show performance improvements across all metrics while moving from the configuration without BiLSTM-attention network to the proposed model. As shown in Table 4.5, the average accuracy improves from 87.24% to 89.84%, while macro recall shows a significant increase from 62.80% to 73.08%, and the F1-score improved from

68.26% to 76.07%. Adding the attention layer improved the macro F1-score, rising from 74.19% to 76.07%. However, precision remains relatively constant across the three configurations (without BiLSTM-attention, without attention, and the proposed model), with only slight improvements. This result indicates that adding BiLSTM-attention layers makes the proposed model better at detecting more true positives (higher recall) while slightly reducing false positives in the PTB-XL dataset.

Table 4.5: Ablation Study Results on test-set of each dataset

Dataset	Metric (%)	Without BiLSTM-attention	Without attention	Proposed
PTB-XL	Avg. Accuracy	87.24	89.14	89.84
	Macro Recall	62.80	71.13	73.08
	Macro Precision	77.20	78.17	79.89
	Macro F1-score	68.26	74.19	76.07
Parameters	Trainable	113,573	286,181	288,533
	Total	116,005	290,341	292,693
Chapman-Merged	Avg. Accuracy	95.07	98.38	98.80
	Macro Recall	87.76	96.12	97.02
	Macro Precision	91.01	96.83	97.62
	Macro F1-score	89.03	96.46	97.32
Parameters	Trainable	113,508	286,116	288,468
	Total	115,940	290,276	292,628
Chapman-Reduced	Avg. Accuracy	94.36	98.11	98.55
	Macro Recall	61.73	80.84	86.57
	Macro Precision	81.61	89.90	91.37
	Macro F1-score	63.38	83.96	88.46
Parameters	Trainable	113,703	286,311	288,663
	Total	116,135	290,471	292,823
Code15%	Avg. Accuracy	95.99	97.44	97.82
	Macro Recall	83.97	91.05	92.78
	Macro Precision	92.29	94.12	94.89
	Macro F1-score	87.73	92.55	93.71
Parameters	Trainable	113,638	286,246	288,468
	Total	116,070	290,406	292,628

The model performance results on the Chapman arrhythmia dataset, including both Chapman-Merged and Chapman-Reduced, show how class imbalance affects performance metrics. The Chapman-Merged dataset, with four merged classes and with less class imbalance, shows a significant improvement in both recall and F1-score, reaching a final macro F1-score of 97.32% as shown in Table 4.5. This improvement shows that the model benefits from a

more balanced dataset, and the BiLSTM and attention mechanisms help refine its ability to capture influential patterns, resulting in fewer false negatives and better generalization. In contrast, the Chapman-Reduced, with seven classes and a higher class imbalance, starts with a much lower recall and F1-score, reflecting the difficulty of classifying underrepresented classes. However, adding the BiLSTM and attention layers has significantly improved recall from 61.73% to 86.57% and F1-score from 63.38% to 88.46%. This improvement indicates the ability of these mechanisms to capture influential features in an ECG signal and help reduce the impact of class imbalance.

The performance metrics for the Code15% dataset also show the impact of different model configurations. Without the BiLSTM-attention mechanism, the model achieves an average accuracy of 95.99%. While this reflects a solid classification ability, the macro recall of 83.97% indicates that the model struggles to identify true positives. Despite a high macro precision of 92.29%, which suggests a low rate of false positives, the macro F1-score of 87.73% reveals a significant trade-off, as the recall is notably lower than the precision.

When BiLSTM layers are incorporated, the model's average accuracy improves to 97.44%, and macro recall rises to 91.05%, indicating enhanced performance in identifying true positives. The macro precision also slightly increases to 94.12%, reflecting higher reliability in predictions. Nonetheless, the proposed model, which integrates both BiLSTM and attention mechanisms, achieves the most significant improvements, attaining an average accuracy of 97.82% and a macro recall of 92.78%. This configuration captures more true positives and maintains a high macro precision of 94.89%. With a macro F1-score of 93.71%, the proposed model demonstrates a strong performance, effectively addressing the complexities within the dataset and showing the benefits of integrating both mechanisms for improved classification results.

Adding BiLSTM and attention mechanisms has increased the number of trainable parameters, as shown in Table 4.5 of various configurations of the ablation study. In the case of the PTB-XL dataset, the baseline model without BiLSTM-attention comprises 113,573 trainable parameters, reflecting its fundamental complexity for capturing essential patterns. Including the BiLSTM increases the total number of trainable parameters to 286,181. The proposed model, which integrates both BiLSTM and attention mechanisms, has slightly increased the trainable parameters to 288,533.

The observed variation in trainable parameters across datasets with the same architecture, as shown in Table 4.5, is primarily due to the differences in the number of output classes. Each dataset necessitates a specific number of neurons in the output layer: 5 neurons for PTB-XL, 4 for Chapman-Merged, 7 for Chapman-Reduced, and 6 for Code15%. This difference in output neurons directly influences the parameters of the model.

Across all datasets, substantial performance improvements are observed that show the effectiveness of increased model complexity. For instance, in the Chapman-Reduced seven-class dataset, the proposed model achieves a macro recall of 86.57%, improving from 61.73% without BiLSTM-attention, as shown in Table 4.5. The results suggest that the additional complexity introduced by the BiLSTM and attention mechanisms does not merely increase the number of parameters but significantly enhances the model's robustness and performance across diverse datasets. This improved performance comes from the capability of BiLSTM to capture the temporal dependencies in ECG signals, which addresses a limitation of 1D-CNNs that primarily focus on extracting local features.

4.6. Performance Comparison

Table 4.6 presents a performance comparison between our proposed model and existing state-of-the-art techniques. The results show that our model consistently performs well across all three datasets. Yildirim *et al.* [80] developed a model using CNN-LSTM blocks for the Chapman-Reduced seven and Chapman-Merged four classes of the Chapman Arrhythmia dataset. Their model achieved average accuracies of 92.24% and 96.13% for the reduced seven and merged four disease classes, respectively. In comparison, as presented in Table 4.6, our proposed model achieved higher average accuracies of 98.55% and 98.80%, respectively. In another study, Baygin *et al.* [224] applied a cascaded homomorphically irreducible tree model with maximum absolute pooling for multilevel feature generation to classify the Chapman Arrhythmia dataset into reduced seven and merged four classes. Their technique achieved an average classification accuracy of 92.95% for the reduced seven classes and 97.18% for the merged four classes, as shown in Table 4.6.In the case of reduced seven class classification, our proposed model showed a better classification accuracy, which is higher than the result obtained in [224] by 5.6%.

Similarly, Anand *et al.* [43] proposed a 2D-CNN model for heart disease classification,

evaluating its performance on five super-diagnostic classes of the PTB-XL dataset and the Chapman-Merged four classes Arrhythmia dataset. As shown in Table 4.6, their model achieved a macro AUC of 93.41% and an average accuracy of 89.73% on the PTB-XL dataset. When applied to the Chapman-Merged four classes using the same architecture and different fine-tuned hyperparameters, it attained an average accuracy of 95.8% and an AUC of 99.46%. However, they did not test their model on the Chapman-Reduced seven classes. By comparison, our proposed model achieved a slightly higher macro AUC and accuracy on the PTB-XL dataset. For the Chapman-Merged four classes, our model exceeded their accuracy by 2.70%, consistently outperforming across all metrics as highlighted in Table 4.6.

On the other hand, Śmigiel *et al.* [47] developed a 5-layer 1D-CNN model that incorporates entropy features separately, resulting in a computationally efficient but lower-performing model. Similarly, Pałczy'nski et al. [44] introduced a deep CNN trained in a few-shot learning approach for classifying the PTB-XL dataset. As detailed in Table 4.6, their model achieved an average accuracy of 79.00%, a specificity of 73.5%, and a macro F1-score of 70.60% on the five-class PTB-XL dataset. In contrast, our proposed technique demonstrated a better classification performance, achieving 89.84% in average accuracy, 93.48% in specificity, and 76.07% in macro F1-score.

Additionally, the proposed model performance was evaluated against cardiology residents, emergency medicine residents, and medical students using a subset of the CODE-test dataset. This subset, comprising 146 samples as detailed in Section 3.3 and summarized in Table 4.6, was annotated by cardiologists as the gold standard [198]. As shown in Table 4.6, our proposed model demonstrated performance comparable to that of the cardiology residents, emergency residents, and medical students. Ribeiro et al. [225] also presented a model trained on the entire CODE dataset, achieving good results when the sigmoid output threshold was adjusted to maximize the F1-score [20]. However, at a threshold value of 0.5, the classification outcome is shown in Table 4.7. Nevertheless, it resulted in no false positives across the disease classes.

The results presented in Table 4.7 highlight the potential of using ML models to support clinicians in diagnosing heart diseases from ECG signals. This is particularly significant given the challenges that physicians of varying experience levels face when reading and interpreting ECGs, as demonstrated by studies such as [8–10].

Table 4.6: Performance of the proposed model in (%) compared to prior works.

Metric (in %)	Chapman Arrhythmia Dataset [500 Hz]							PTB-XL Dataset [100 Hz]			
	Yildirim *et al.* [80]		Anand *et al.* [43]	Baygin *et al.* [224]		Proposed		Anand *et al.* [43]	Śmigiel *et al.* [47]	Pałczyński *et al.* [44]	Proposed
	Reduced	Merged	Merged	Reduced	Merged	Reduced	Merged				
Avg. Accuracy	92.24	96.13	95.85	92.95	97.18	98.55	98.80	89.73	76.50	79.00	89.84
Avg. Specificity	98.72	95.43	-	-	-	99.20	99.29	-	-	73.50	93.48
Macro Recall	80.15	95.43	95.34	80.98	96.77	86.57	97.02	-	66.20	70.60	73.08
Micro Recall	-	-	-	-	-	94.58	97.37	-	-	-	77.69
Wavg Recall	-	-	95.85	-	-	94.58	97.37	-	-	-	76.87
Macro Precision	80.31	95.78	95.44	90.17	97.07	91.37	97.62	-	71.40	-	79.89
Micro Precision	-	-	-	-	-	95.22	97.83	-	-	-	81.50
Wavg Precision	-	-	-	-	-	94.99	97.84	-	-	-	81.83
Macro F1-score	80.04	95.57	95.39	84.01	96.91	88.46	97.32		68.00	70.60	76.07
Micro F1-score	-	-	-	-	-	94.90	97.60	79.38	-	-	79.55
Wavg F1-score	-	-	95.84	-	-	94.64	97.60	-	-	-	78.75
Macro AUC	-	-	99.46	-	-	99.58	99.75	93.41	91.00	93.60	93.51
Micro AUC	-	-	-	-	-	99.78	99.77	-	-	-	94.51
Wavg AUC	-	-	-	-	-	99.76	99.78	-	-	-	94.10
Macro AUPRC	-	-	98.53	-	-	94.42	99.18	83.38	-	-	83.73
Micro AUPRC	-	-	-	-	-	99.01	99.36	-	-	-	87.48
Wavg AUPRC	-	-	-	-	-	97.99	99.29	-	-	-	86.68

Table 4.7: Confusion matrix for a subset of the CODE-test dataset

		Predicted Label									
		Ribeiro *et al.* [225]		Resident Cardiologist		Emergency Cardiologist		Medical Student		Proposed	
	True Label	Negative (-ve)	Positive (+ve)	Negative (-ve)	Positive (+ve)	Negative (-ve)	Positive (+ve)	Negative (-ve)	Positive (+ve)	Negative (-ve)	Positive (+ve)
1dAVb	Negative (-ve)	118	0	131	2	116	2	109	9	117	1
	Positive (+ve)	21	7	9	19	5	23	2	26	3	25
	Misclassified	21		11		7		11		4	
RBBB	Negative (-ve)	112	0	111	1	112	0	111	1	111	1
	Positive (+ve)	5	29	1	33	8	26	2	32	3	31
	Misclassified	5		2		8		3		4	
LBBB	Negative (-ve)	116	0	116	0	115	1	116	0	116	0
	Positive (+ve)	5	25	3	27	4	26	3	27	1	29
	Misclassified	5		3		5		3		1	
SB	Negative (-ve)	130	0	130	0	129	1	129	1	129	1
	Positive (+ve)	4	12	1	15	2	14	4	12	0	16
	Misclassified	4		11		3		5		1	
AFIB	Negative (-ve)	133	0	131	2	132	1	128	5	130	3
	Positive (+ve)	4	9	3	10	5	8	1	12	0	13
	Misclassified	4		5		6		6		3	
ST	Negative (-ve)	109	0	109	0	109	0	108	1	109	0
	Positive (+ve)	9	28	7	30	2	35	6	31	2	35
	Misclassified	9		7		2		7		2	

4.7. Pilot Test Results

The model, trained on the CODE-15% dataset, was deployed on the Ethiopian Artificial Intelligence Institute (EAII) data center for real-world evaluation. It was subsequently pilot-tested using the ZMH dataset described Section 3.3. The system's graphical user interface (GUI), shown in Figure B.1 in Appendix B, facilitates the analysis of ECG data, providing an intuitive platform for users to interact with the model's predictions and results.

The results of the pilot testing, shown in Table 4.8, demonstrate the strong performance of the model in classifying ECG signals into six clinically significant disease classes, as reflected in the overall metrics and class-wise evaluations.

Table 4.8: The proposed model's pilot-test performance in (%) on ZMH dataset.

Classes	Accuracy	Specificity	Recall	Precision	F1-Score	AUC	AUPRC
1dAVb	85.93	82.56	91.84	75.00	82.57	95.51	93.42
RBBB	98.52	100	88.24	100	93.75	99.30	97.34
LBBB	95.56	96.36	92.00	85.19	88.46	99.42	98.13
SB	91.11	100	72.73	100	84.21	98.95	98.10
AFIB	97.04	98.37	83.33	83.33	83.33	97.83	89.43
ST	98.52	100	71.43	100	83.33	95.65	78.84
Average	94.44	96.22					
Macro			83.26	90.59	85.94	97.78	92.55
Micro			84.42	86.09	85.25	97.63	93.03
Weighted			84.42	88.34	85.32	97.73	94.98

The model achieved an overall accuracy of 94.44%, demonstrating its dependability in distinguishing the disease classes. Furthermore, macro-averaged metrics indicate robust performance across all classes, with a macro recall of 83.26%, macro precision of 90.59%, macro F1-score of 85.94%, a macro AUPRC of 92.55%, and a macro AUC of 97.78%. The high macro AUC and macro AUPRC suggest acceptable discriminative ability and positive predictive power across all classes.

The model performed well across most classes, with the highest accuracy and specificity for RBBB and ST. However, there is room for improvement in Recall and Precision for some classes, such as 1dAVb and AFIB. Despite these areas of improvement, the model's overall performance, including high AUC and AUPRC scores, suggests it is well-suited for heart disease classification and can be considered for clinical applications.

CHAPTER 5

Interpretability Analysis

5.1. Overview

Interpreting the DL model's classification output is essential in providing evidence-based diagnosis. However, as briefed in Section 3.6, it is challenging to tell the rationale behind the model's decision in classifying an input instance into its diagnostic class. In our study, SHAP with gradient explainer [226] and Grad-CAM++ [159] attribution values were used to demonstrate the proposed model's classification output interpretability. The model's output result is explained at the test set and single instance level, as discussed in the following sections. To demonstrate the interpretability of the proposed model's classification output, we used the reduced seven-class Chapman Arrhythmia test set, as it presents a challenging scenario with its high number of classes and significant class imbalance. On the other hand, the single instance level explanation was provided by visualizing the attribution values from interpretation techniques along with a patient's raw ECG signal.

5.2. Test Set Level Interpretability

The test set level explanation is provided by computing the positive contribution of each lead in the classification of the entire test set. Using a similar technique employed by Agrawal *et al.* [227] and detailed in Algorithm 2, the cumulative attribution values for each lead across all patients within individual classes were calculated. This cumulative sum represents the contribution of each ECG lead for the diagnostic class. Then, these contributions were averaged for each diagnostic group. The result on Figure 5.1 and Figure 5.2, respectively,

show the contribution of each lead in the trained model's classification outputs using SHAP values and Grad-CAM++ attributions.

Algorithm 2 Steps in Test set level Analysis

Input: *GradCAM++* attribution or *SHAP* Values of Test data (**D**): Attr, Predicted classes (AF, AFIB, SI, SB, SR, ST, SVT): $NumClasses = 7$
$ClassIndices \leftarrow$ [AF, AFIB, SI, SB, SR, ST, SVT]
Result: Total and average (AVG) Contribution of each lead in classification
$a[i][j]$ ▷ Attribution of j^{th} lead for i^{th} class
$sum[i]$ ▷ Sum of attribution of all leads for i^{th} class
$r[i][j]$ ▷ Normalized attribution of j^{th} lead on i^{th} class
$ravg[j]$ ▷ Average attribution of j^{th} lead
for $i \leftarrow 0$ to $NumClasses - 1$ **do**
    $S \leftarrow$ Attr[`ClassIndices`$[i]$] ▷ Get SHAP values or GradCAM++ attributions for class i
    for each lead j **do**
        for each instance in test data, d **do**
            for each Timestamp or FeatureMap, τ **do** ▷ Timestamp in raw ECG signal
                $c[i][j] \leftarrow c[i][j] + S[d][\tau][j]$ ▷ for SHAP and FeatureMap from
            end for ▷ the last Conv2D layer for Grad-CAM++
        end for
    end for
end for
for each class i **do**
    for each lead j **do**
        $sum[i] \leftarrow sum[i] + c[i][j]$ ▷ Calculate the sum
    end for
end for
for each class i **do**
    for each lead j **do**
        $r[i][j] \leftarrow c[i][j]/sum[i]$ ▷ Normalize the attribution
    end for
end for
for each lead j **do**
    for $i \leftarrow 0$ to $NumClasses - 1$ **do**
        $avg[j] \leftarrow avg[j] + r[i][j]$ ▷ Accumulate the contribution of each lead
    end for
    $avg[j] \leftarrow avg[j]/NumClasses$ ▷ Compute the average contribution each lead
end for

The average (AVG) SHAP value across the seven diagnostic classes depicted in Figure 5.1 shows that, within the reduced Arrhythmia dataset test set, lead II, V1, I and V4 made more significant positive contributions to the model's classification output. Similarly, as shown in Figure 5.2, the average (AVG) Grad-CAM++ heatmap values of the 12-ECG leads across the diagnosis classes show that lead II, V4, I, and V1 are the influential leads in the model's

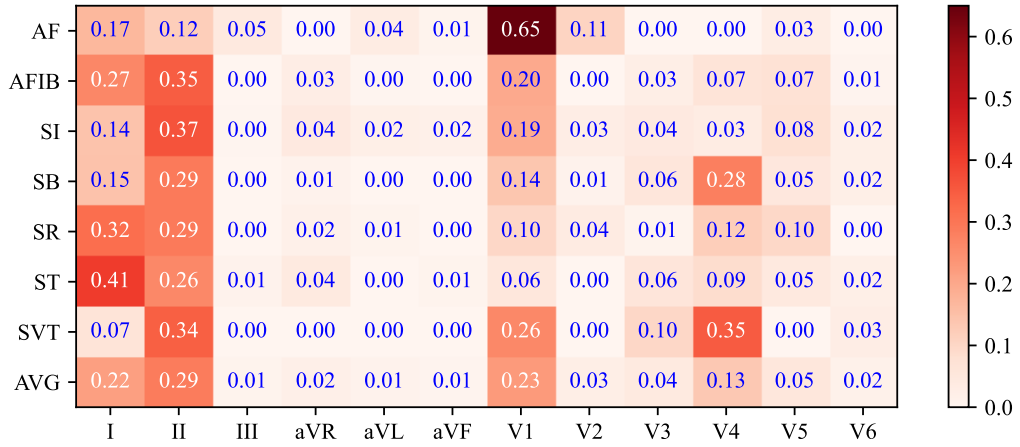


Figure 5.1: SHAP values based test set level interpretation through quantifying the contribution rate of each ECG leads to the diagnostic classes.

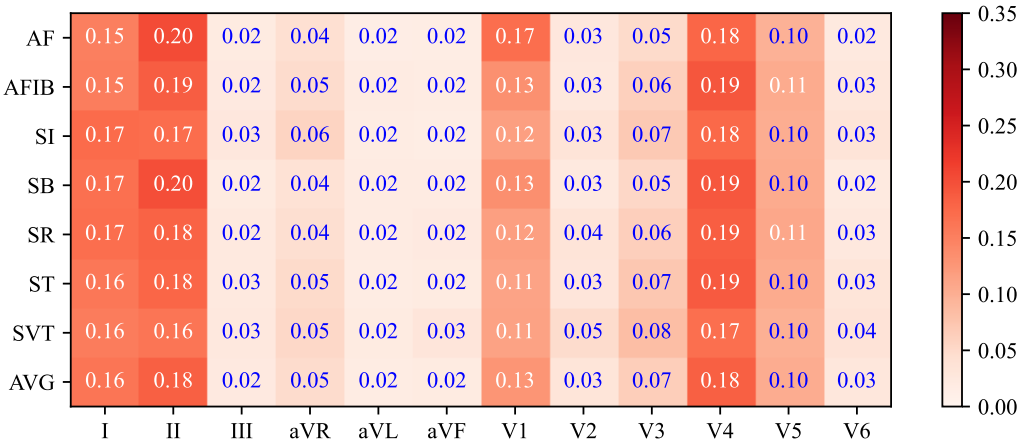


Figure 5.2: Grad-CAM++ heatmaps based test set level interpretation through quantifying the contribution rate of each ECG leads to the diagnostic classes.

classification output.

Analyzing the test set, as illustrated in Figure 5.1 and Figure 5.2, using SHAP values and Grad-CAM++ attributions, respectively, offers valuable insights by highlighting the most influential leads that contribute to predictions across the dataset. This method combines explanations from multiple ECG instances, uncovering consistent patterns regarding the significance of different leads in diagnosing various disease conditions within the dataset. These findings are essential for validating the model's reliability and helping clinicians prioritize specific leads for further investigation.

Another benefit of the result provided in Figure 5.1 and Figure 5.2 is identifying systematic errors or biases in the model, for instance, if the model relies on leads that might not be clinically relevant. This ensures that the model is both robust and clinically interpretable. How-

ever, test set-level analysis using cumulative attributions may obscure instance-specific variations. Additionally, because the ECG signal is inherently temporal, this approach may fail to pinpoint specific time intervals that contribute most significantly to the proposed model's classification output. So, augmenting test set-level interpretation with instance-specific explanations is crucial for optimizing interpretability.

5.3. Instance Level Interpretation

In the instance-level explanation, the SHAP method results in an output that matches in dimension to the raw ECG data. This output comprises SHAP values corresponding to each position within the input ECG signal. Then, a threshold value is selected to retain the top half of SHAP values from the lead that contributes most significantly to the model's classification. Additionally, for the remaining ECG leads, the retained SHAP values are those greater than or equal to the threshold. These values highlight the ECG segments in red, while the rest are shown in blue, as illustrated in Figure 5.3.

Figure 5.3a depicts an ECG signal diagnosed with Atrial Flutter (AF). As shown in Figure 5.3a, AF is manifested as saw-tooth-like atrial waves between consecutive QRS complexes [228] as shown in the orange marked region of Figure 5.3a. The saw-tooth atrial waves are more noticeable on lead V1 as depicted in Figure A.13. Besides, Figure 5.3b shows lead II of a subject with Sinus Tachycardia (ST). The regions of the plot with the red color indicate segments on an ECG signal that significantly contributed to the classification. ST is a cardiac condition with a faster rate resulting in shortened P-R and intervals between heartbeats (R-R intervals) while preserving a regular rhythm [200, 228]. In agreement with the clinical diagnosis criteria, the P-R intervals and the R-peaks are highlighted in red as depicted in Figure 5.3b. However, these segments of the ECG are not consistently highlighted across the time intervals of the signal. In addition, the clarity in highlighting the influential regions of an ECG is not sharp across all diagnosis classes.

Similarly, the SHAP-based interpretation provides a visual explanation comparable to clinical observations in the remaining diagnostic classes. In Supraventricular Tachycardia (SVT), the heart rate is high, and this rapid succession of cardiac cycles makes the P-wave often buried within the QRS complex, which makes it not visible on lead II. Besides, in lead V1, SVT makes the P-wave indistinguishable [200]. The red color in Figure A.1 of lead II and

V1 ECG segments highlights this manifestation of SVT. On the other hand, a heart rate below 60 beats per minute results in a condition called Sinus Bradycardia (SB). Despite the slower heart rate, ECG waveforms in SB maintain a regular rhythm and normal morphology [200, 228]. As shown in Figure A.7, the slow heart rate is evidenced by the prolonged isoelectric states between cardiac cycles in ECG leads. These isoelectric states in lead I, II, and V1 are highlighted in red.

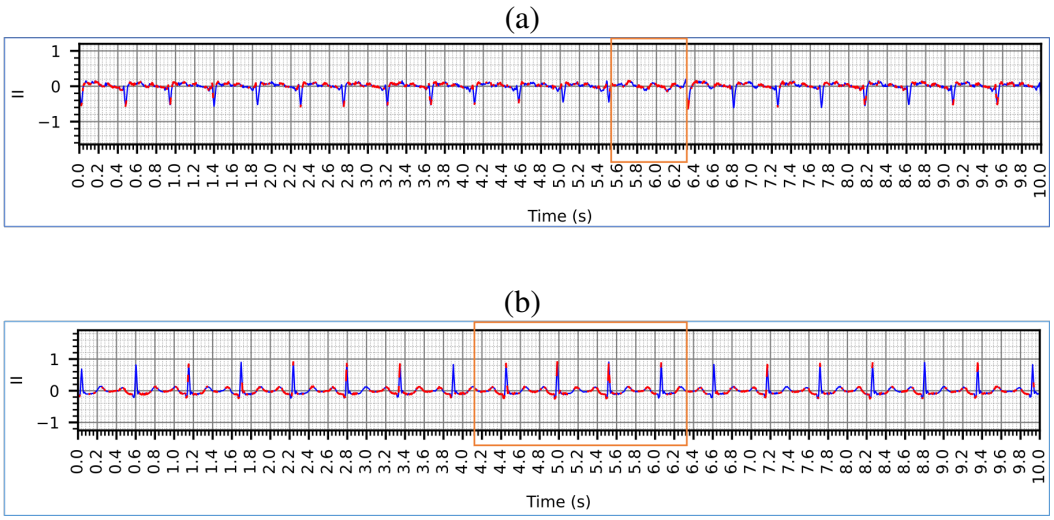


Figure 5.3: SHAP value based interpretation: (a) Lead II of a sample with Atrial Flutter (AF), (b) Lead II of a sample with Sinus Tachycardia (ST).

In Sinus Rhythm (SR), the heart rate is between 60 to 100 beats per minute [200]. In addition, every cardiac cycle exhibits a normal waveform morphology with a regular rhythm, the appearance of a P-wave before each QRS complex, and a P-R interval with normal range [200]. In this regard, Figure A.5 depicts that the SHAP values in lead I and V1 highlight the QRS complex in red as the influential segments of the ECG. Besides, the SHAP values in lead I highlight nearly the entire waveform covering the P-waves, QRS complexes, and P-R intervals. The deviation of the heart rhythm from the regular patterns observed in SR is known as Sinus Irregularity (SI) [200]. These irregularities may manifest as fluctuations in the waveform morphology, irregular R-R interval, and inconsistent patterns in the cardiac cycle [200]. On Figure A.9, the R-R interval irregularity is apparently visible; for instance, the time length between the 8^{th} and the 9^{th} beats R-R interval is 0.56 seconds, and between 9^{th} and the 10^{th} beats is 1.16 seconds. The SHAP values highlighted the peaks in red on lead I. On the other hand, Atrial Fibrillation (AFIB) is characterized by an irregularly

irregular ventricular rhythm and fibrillatory waves (f-waves), leading to an irregular undulation of the ECG baseline [200]. These f-waves are evident in lead V1 and highlighted in red, as shown in Figure A.11.

In Grad-CAM++, the interpretability at the instance level is achieved by highlighting specific segments in the ECG with a heatmap. Grad-CAM++ generates class activation maps (heatmaps) by calculating the gradients of the target class to the final convolutional layer's feature maps as given in Eqn. (2.10). These gradients are used to weigh the importance of different spatial locations in the feature maps, allowing the visualization of the leads and time segments of the ECG contributing to the model's classification decision. Calculating the gradient results in a tensor with the same shape as the feature map from the last convolutional layer, the Conv2D on Figure 3.11. After averaging along the channel axis, a new tensor is produced with a shape of [number of leads (12), sequence length in the last convolutional layer]. This new tensor was then transformed into a vector with a size of N using linear interpolation, maintaining the same sample size as the original ECG.

Figure 5.4 depicts a lead V1 of ECG signals diagnosed with AF and SB. The heatmap highlights segments of ECG that influence the model to classify the input into respective diagnosis classes. The heatmap on Figure 5.4a shows that the saw-tooth waves in AF cardiac conditions are more influential segments of an ECG. On the other hand, in Figure 5.4b, the prolonged isoelectric states between consecutive heartbeats on lead V1 of a subject with SB cardiac condition are emphasized with the warm red color.

On sample cases of the Chapman Arrhythmia dataset, as shown in the supplementary on Figure A.2, Figure A.4, A.6, A.8, A.10, A.12 and A.14, Grad-CAM++ technique has visually highlighted influential segments of an ECG that contribute significantly in the classification. The heatmaps on lead I and lead II in Figure A.2 illustrate the absence of a distinct P-wave, which is a clinical manifestation of SVT in addition to its high heartbeat rate. On the other hand, Figure A.6, the heatmap on lead I and II, highlights the presence of the P-waves, PR-intervals, and QRS-complexes in SR. The shortened PR-intervals on lead II of a subject with ST cardiac condition are emphasized with the warmest dark red color on Figure A.4. In the case of SI, the segments of the ECG with inconsistent patterns in cardiac cycles are highlighted as shown in leads I and II of Figure A.10. While in AFIB, the f-waves are evident in lead V1 and are highlighted with a warmer heatmap, as shown in Figure A.12. The

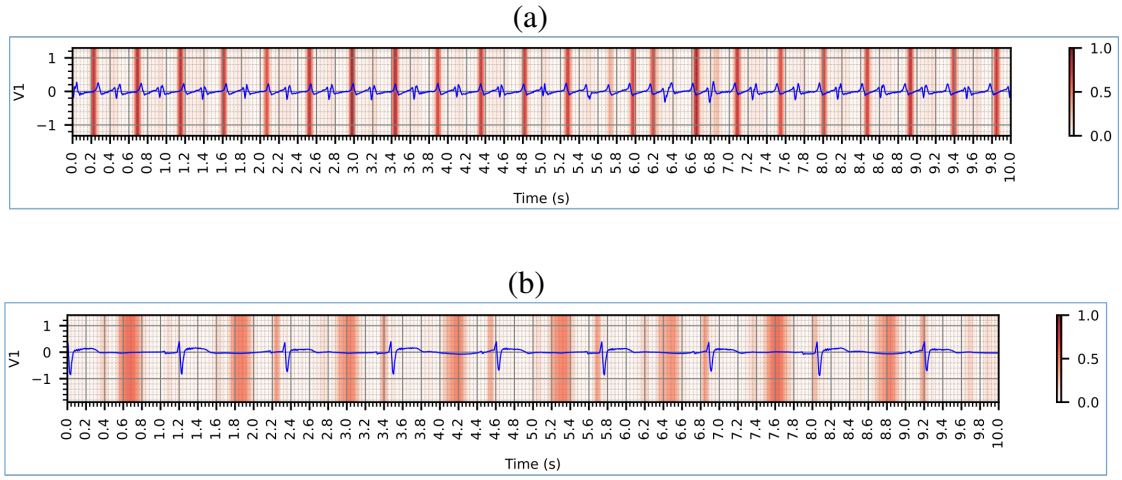


Figure 5.4: Grad-CAM++ based interpretation: (a) Lead V1 of a sample with Atrial Flutter (AF), (b) Lead V1 of a sample with Sinus Bradycardia (SB).

visualization result exhibits similarities with clinical findings of the cardiac diagnosis. This consistency indicates that the Grad-CAM++ technique can offer an intuitive visualization of influential segments of an ECG that the trained model used in its classification. However, the visualization techniques might not be easily understandable by medical practitioners. The challenge is mainly because of the difficulty correlating the heatmap visual explanation to existing clinical interpretations.

The above analysis shows the use case of SHAP value and GradCAM++ attribution in the proposed model classification output interpretation. Besides, the test set level interpretability analysis aids in determining which leads among the 12 lead ECG contribute significantly to the model in classifying the given dataset. The test set-level interpretability provides a global perspective by aggregating feature contributions across multiple instances, revealing consistent patterns in the proposed model decision-making process. This helps identify specific ECG leads that are influential for a particular classification in the given disease classes, enhancing trust and reliability in the model. In addition, it is used to ensure the proposed model classification output aligns with medical domain knowledge.

On the other hand, the instance-level interpretability offers insights into the temporal time interval on ECG leads that drive the proposed model's prediction for a single instance. By combining both levels of analysis, clinicians and researchers can validate DL models' general robustness. This dual model interpretability approach aids in observing where the model focuses on providing the classification output.

CHAPTER 6

Conclusion and Future Works

Heart disease diagnosis from ECG tracings poses significant challenges for physicians across various expertise levels. These challenges underscore the need for machine learning (ML) models to augment diagnostic capabilities. However, the black-box nature of these models and their limited performance metrics have hindered their adoption due to concerns about reliability and trustworthiness. Therefore, interpreting the output of black-box ML models is critical to earning the trust of clinicians and enhancing their practical utility. To address this, we examined the taxonomy of interpretable machine learning (IML) methods, categorizing them by result presentation approach, scope, specificity, and model complexity. Additionally, we evaluated these methods, highlighting their respective strengths and limitations.

We further explored advancements in integrating IML techniques into ECG-based heart disease diagnosis workflows and proposed an interpretable deep learning framework tailored for this task. The proposed model leverages 12-lead ECG data and combines multiple deep-learning components to enhance feature extraction and interpretability. The architecture integrates 12 blocks of 1D CNN-BiLSTM networks with an attention mechanism, followed by a 2D CNN as the feature extraction backbone and three fully connected layers for classification. The 1D CNN effectively captures local features, while the BiLSTM models long-term temporal dependencies inherent in ECG signals. This enables the identification of the most impactful ECG leads and the key temporal segments within these leads that contribute to the classification task. Moreover, the attention mechanism assigns greater weight to leads and temporal regions most relevant to the model's decision-making process, thereby improving interpretability. The integration of a 2D CNN further enhances the model by capturing spatial relationships in feature maps and the GAP layer reduces the overall parameter count and

enhances interpretability.

The performance of the proposed model was evaluated across three distinct ECG datasets and demonstrated better discrimination among diagnostic classes when compared to state-of-the-art methods. Specifically, the model achieved impressive test macro AUC values of 99.58%, 99.75%, 93.51%, and 99.35% on the Chapman-Reduced seven and Chapman-Reduced four classes of the Chapman Arrhythmia dataset, PTB-XL, and CODE-15% datasets, respectively. Additionally, the average test macro AUPRC values were 83.73%, 94.42%, 99.18%, 83.73%, and 97.83% for the same datasets. These results demonstrate the model's ability to perform well across different datasets, offering robust classification performance.

Further, Grad-CAM++ and SHAP techniques have been effectively employed to visualize the most influential leads and segments within the ECG signals that contribute to the model's classification decisions. Test set-level interpretability reveals global patterns by aggregating feature contributions, helping to identify which leads are most significant for classifying specific disease classes. Instance-level interpretability, on the other hand, provides insights into the temporal intervals within the ECG signals that drive the model's predictions for individual cases. By combining both test set and instance-level analyses, this dual approach ensures enhanced interpretability, aiding clinicians in understanding how the model makes its predictions.

Future studies will focus on quantifying and evaluating the performance of the visual interpretations generated by Grad-CAM++ and SHAP techniques. Additionally, we plan to explore the integration of language models to generate verbal descriptions of the model's output, further enhancing the interpretability of its predictions. These efforts aim to improve the understanding, reliability, and evidence-based diagnosis.

Appendix A: Supplementary Figures

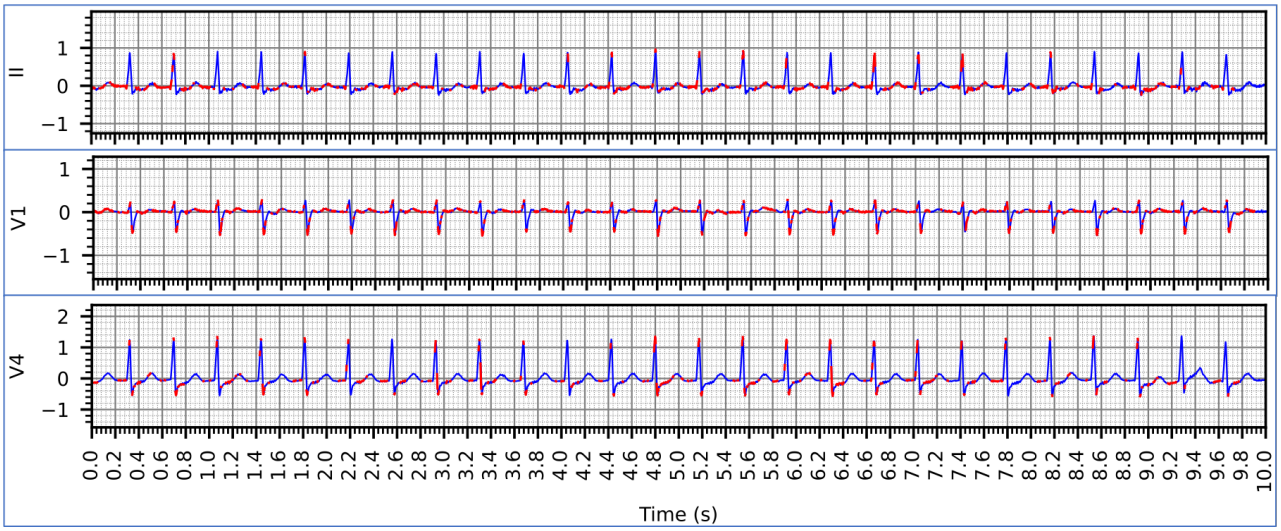


Figure A.1: Example of instance level interpretation for SVT using SHAP values.

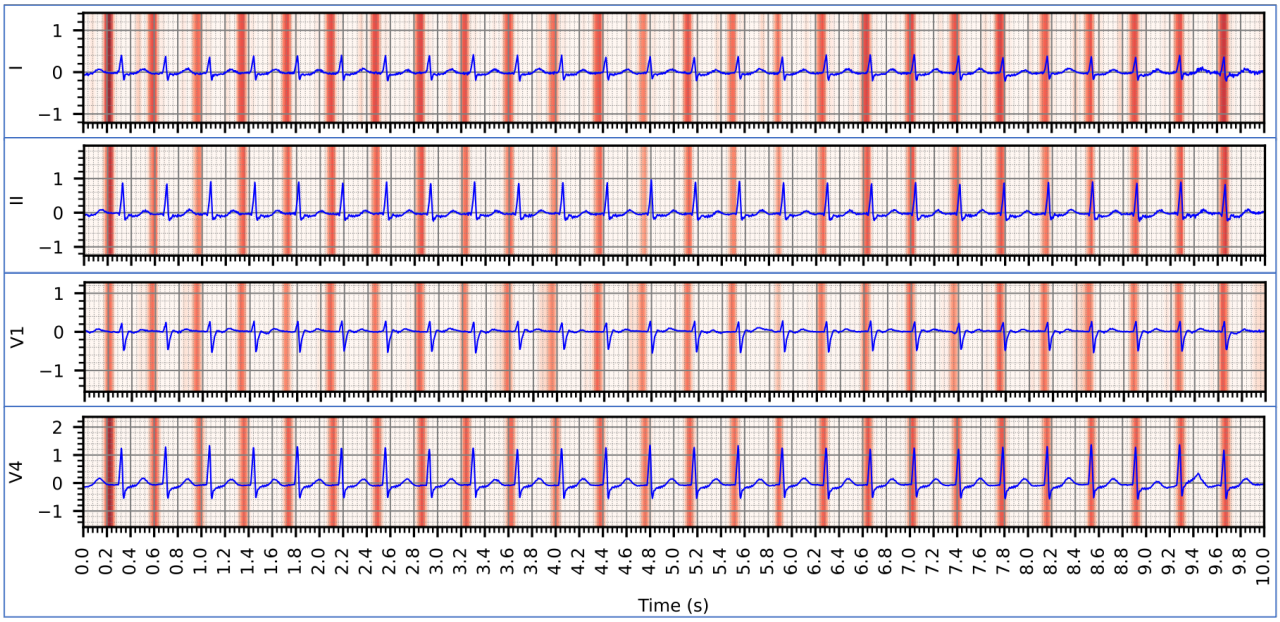


Figure A.2: Example of instance level interpretation for SVT using Grad-CAM++ heatmaps, the same ECG sample to Figure A.1.

Supraventricular Tachycardia (SVT) is characterized by an elevated heart rate, leading to shortened R-R intervals and a rapid succession of cardiac cycles. This condition often obscures the P-wave, making it difficult to distinguish, as it overlaps with the QRS complex or the T-wave, particularly in ECG leads II and V1 [200]. Figure A.1 presents a SHAP-based explanation, highlighting regions that contribute most to the model's prediction in red. Figure A.2 displays a Grad-CAM++ heatmap visualization. Both figures emphasize the hallmark characteristics of SVT, with leads II, V1, and V4 showing the shortened R-R intervals and obscured P waves.

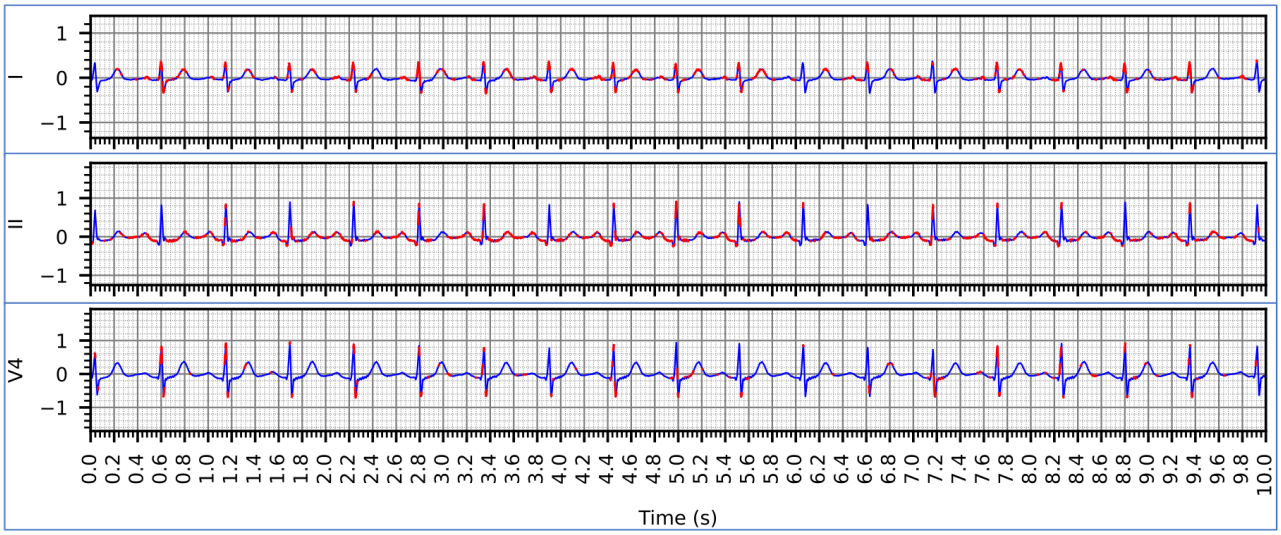


Figure A.3: Example of instance level interpretation for ST using SHAP values.

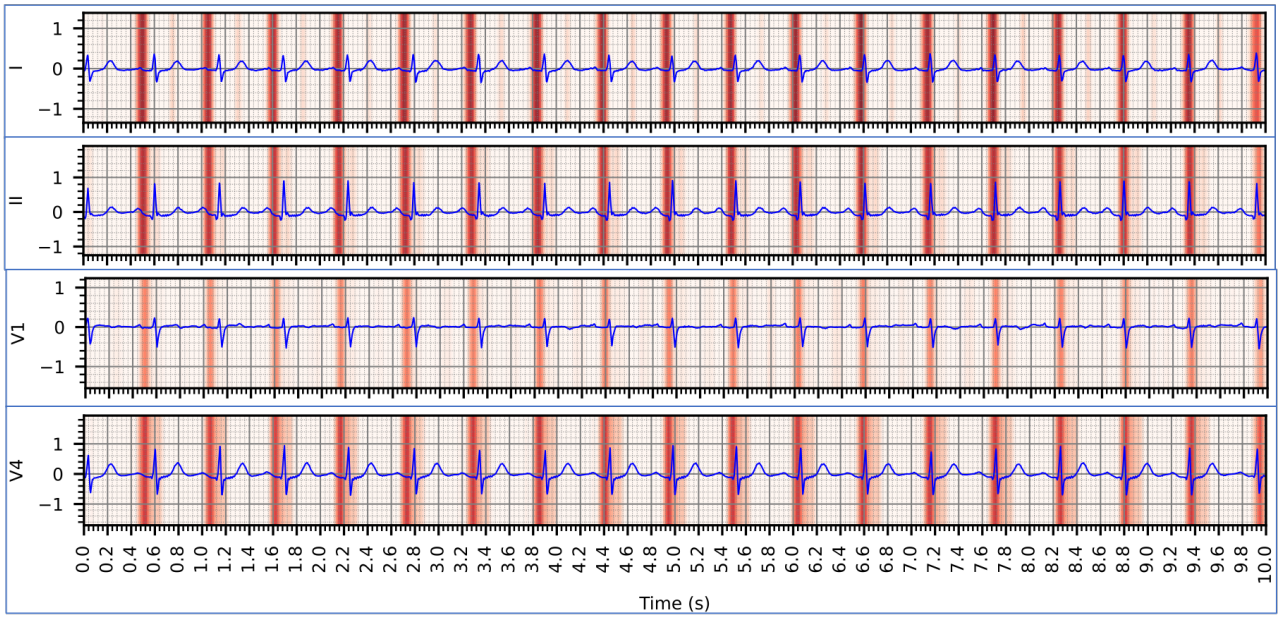


Figure A.4: Example of instance level interpretation for ST using Grad-CAM++ heatmaps, the same ECG sample to Figure A.3.

Sinus Tachycardia (ST) is characterized by an increased heart rate, leading to shortened P-R and R-R intervals while maintaining a regular rhythm [200, 228]. Figure A.3 and A.4, shows ECG readings from a subject diagnosed with Sinus Tachycardia. The red regions in Figure A.3 highlight the ECG segments that have the most significant impact on the model's classification. Consistent with clinical diagnostic criteria, the P-R intervals and R-peaks are marked in red, as seen in Figure A.3. Additionally, Figure A.4 emphasizes the shortened P-R intervals in lead I and II with dark red highlights.

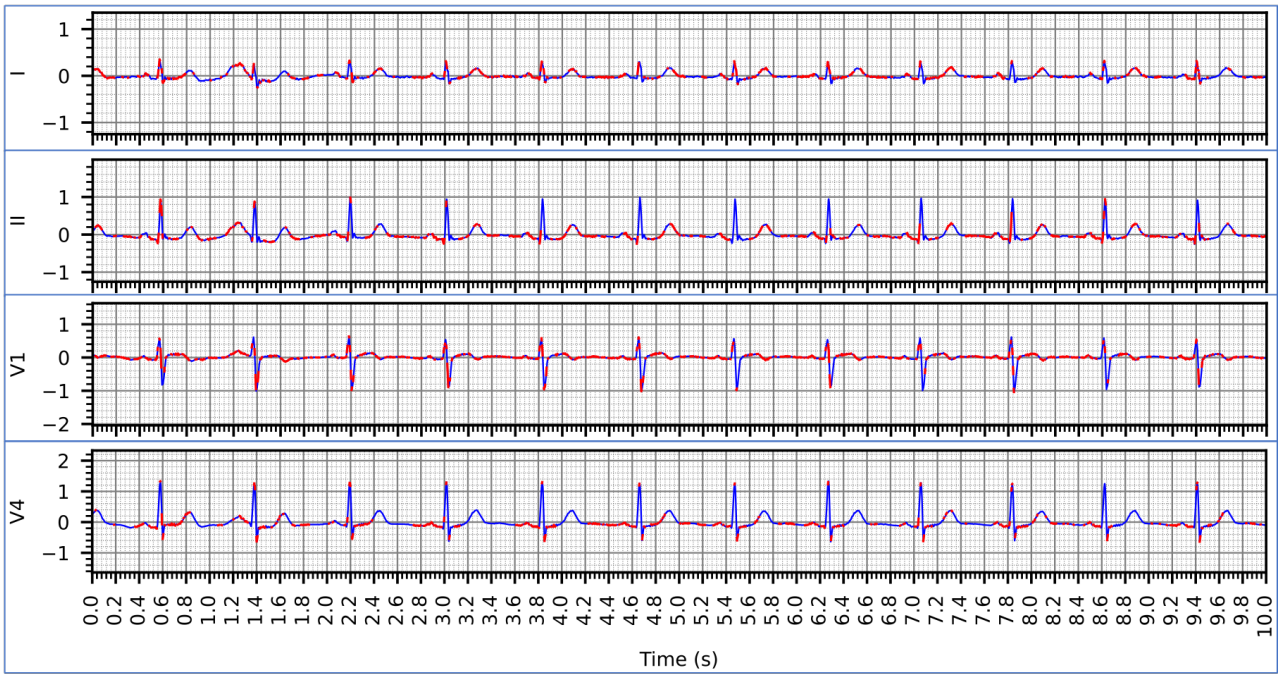


Figure A.5: Example of instance level interpretation for SR using SHAP values.

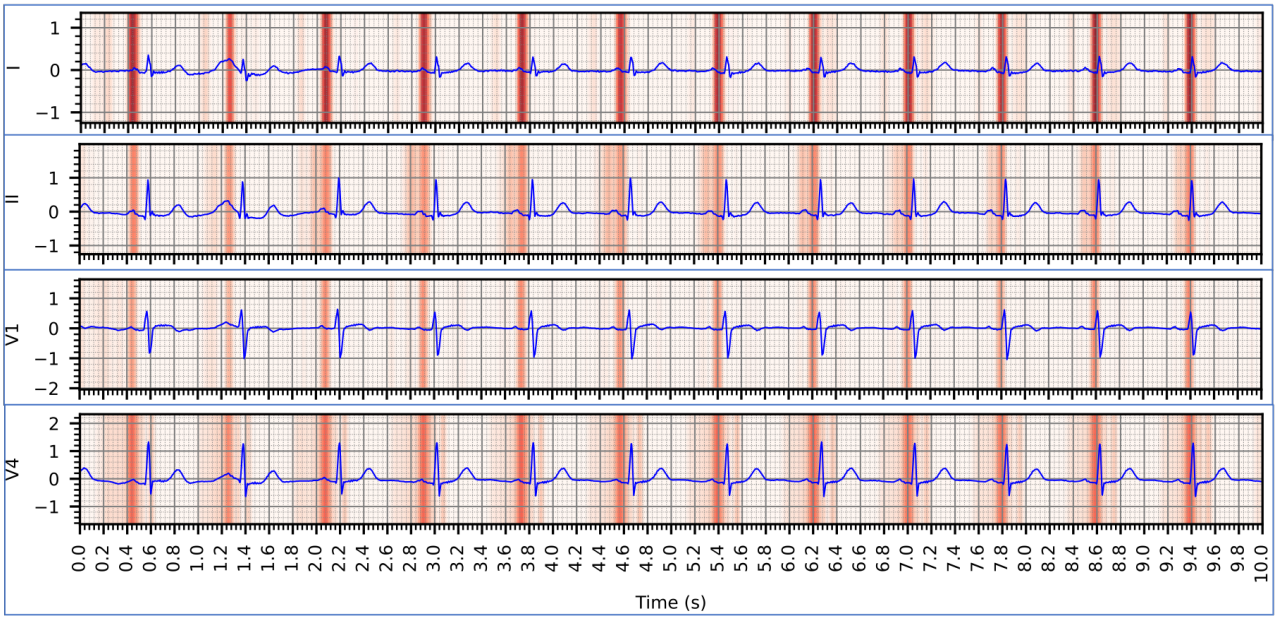


Figure A.6: Example of instance level interpretation for SR using Grad-CAM++ heatmaps, the same ECG sample to Figure A.5.

In Sinus Rhythm (SR), the heart rate ranges from 60 to 100 beats per minute [200]. Each cardiac cycle exhibits a normal waveform morphology with a regular rhythm, characterized by a visible P-wave preceding every QRS complex and a P-R interval within the normal range [200]. Figure A.5 presents SHAP-based explanations for leads I, II, V1, and V4. In particular, the QRS complexes in leads I and V1 are highlighted in red, indicating that the model identifies these regions that contribute most to its classification of the ECG signal as SR. In the lead I, the SHAP values emphasize nearly the entire waveform, including the P-waves, P-R intervals, and QRS complexes. Similarly, Figure A.6 displays a Grad-CAM++ heatmap, where leads I and II highlight the P-waves, P-R intervals, and QRS complexes, further confirming that the model uses these ECG segments to classify the ECG as SR.

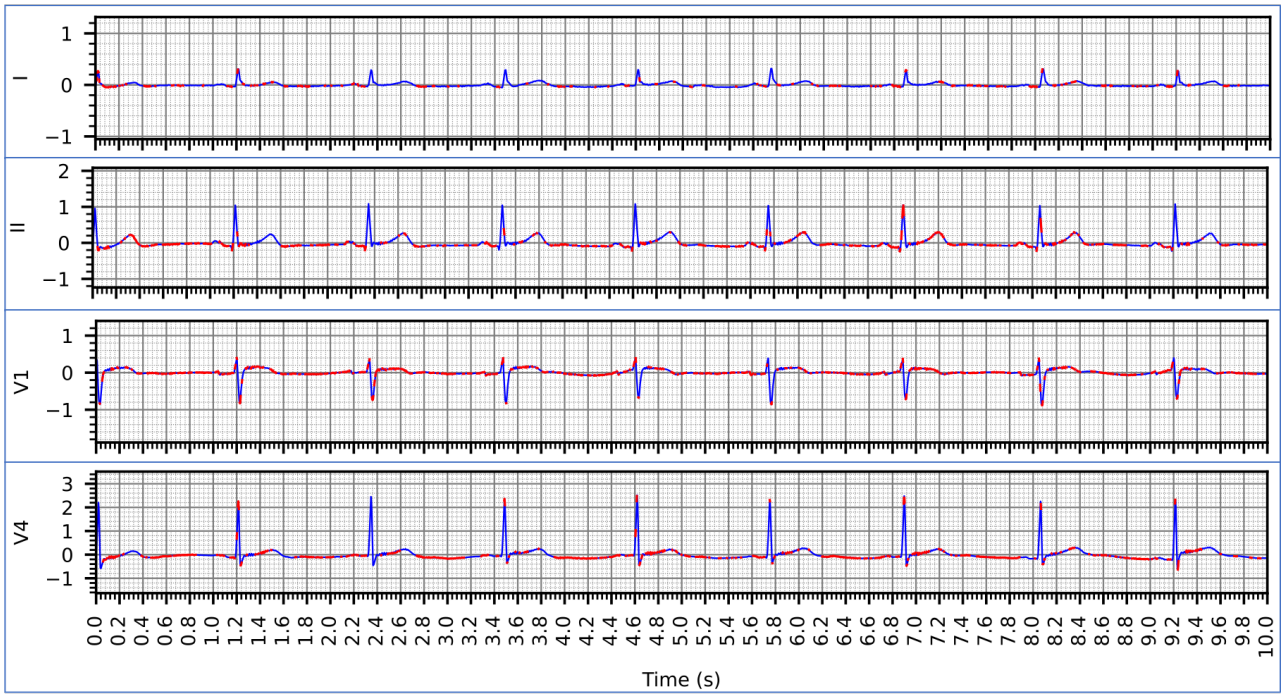


Figure A.7: Example of instance level interpretation for SB using SHAP values.

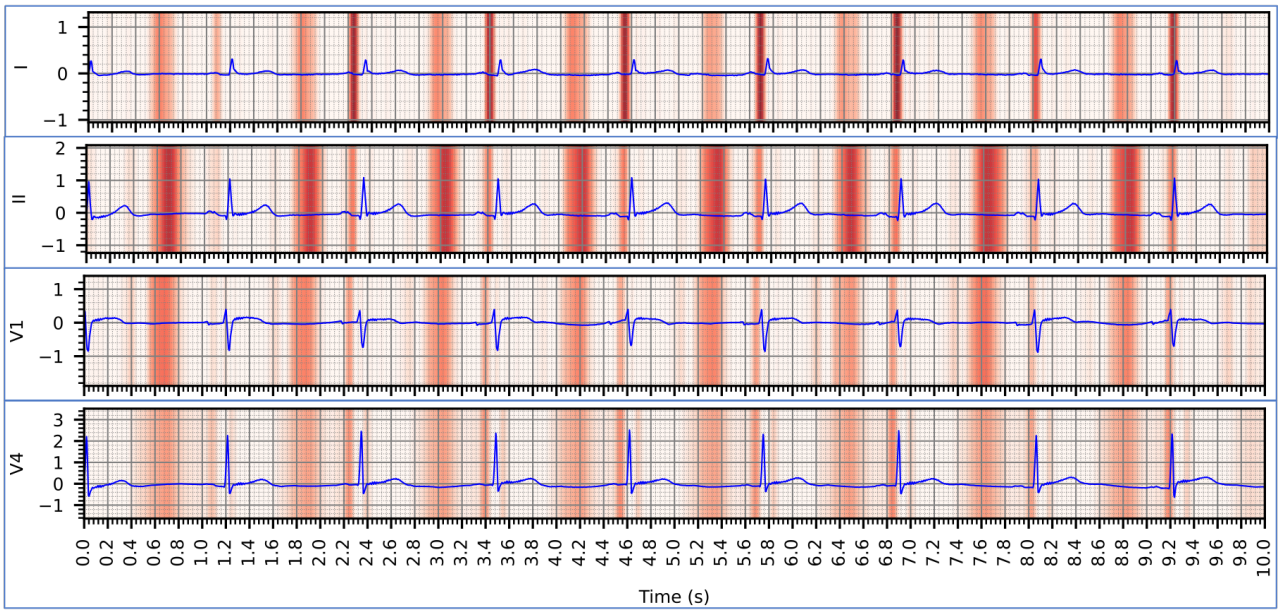


Figure A.8: Example of instance level interpretation for SB using Grad-CAM++ heatmaps, the same ECG sample to Figure A.7.

In Sinus Bradycardia (SB), the heart rate below 60 beats per minute [200, 228]. Despite the slower rate, ECG signals in SB typically preserve a regular rhythm and exhibit morphological features of a normal ECG. Figure A.7 displays SHAP-based explanations, where the prolonged isoelectric intervals between cardiac cycles, indicative of the slower heart rate, are prominently highlighted in red across leads I, II, V1, and V4. These regions are influential segments of the ECG signal that the model used to classify the ECG as SB. Similarly, Figure A.8 presents a Grad-CAM++ heatmap for lead I, II, V1, and V4. The prolonged isoelectric states between consecutive heartbeats are highlighted across all four leads, with the warmest red tones appearing on lead II, indicating the model's emphasis on these features for recognizing SB.

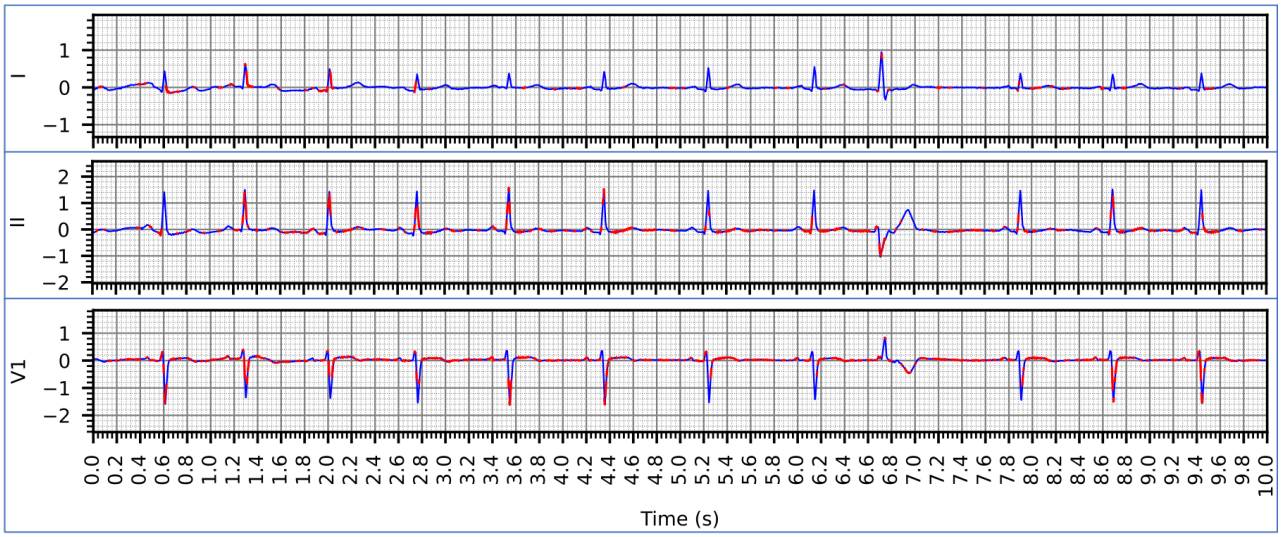


Figure A.9: Example of instance level interpretation for SI using SHAP values.

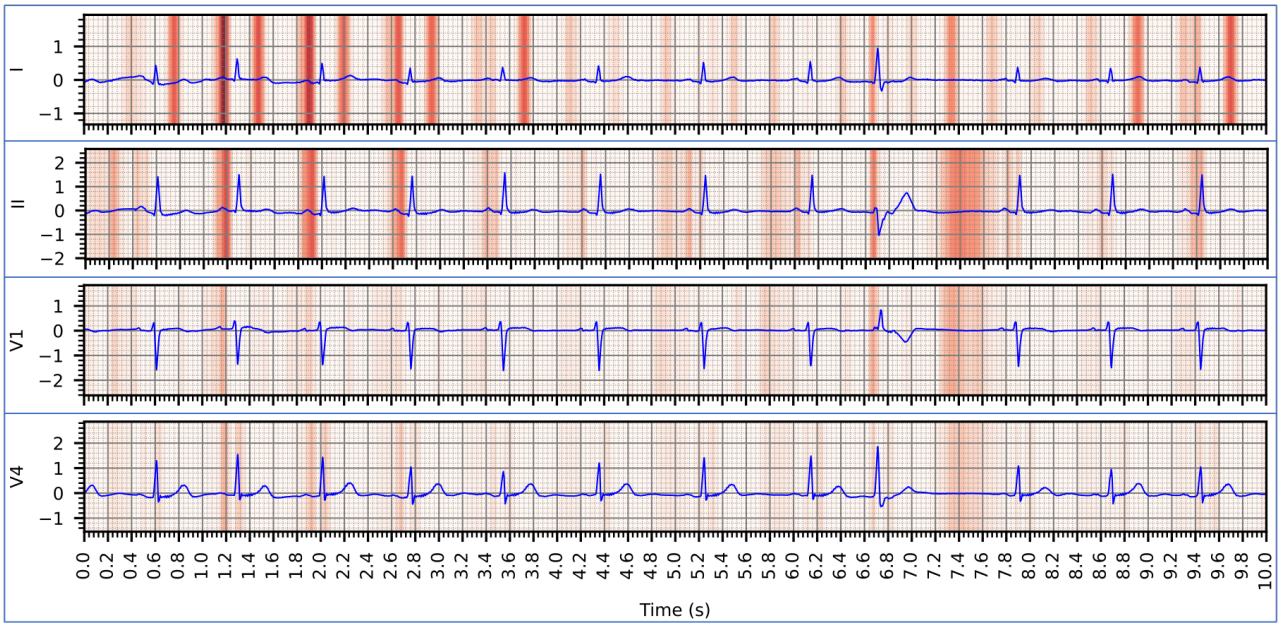


Figure A.10: Example of instance level interpretation for SI using Grad-CAM++ heatmaps, the same ECG sample to Figure A.9.

Sinus Irregularity (SI) is characterized by deviations from the regular rhythm characteristic of Sinus Rhythm (SR) [200]. These deviations often manifest as fluctuations in waveform morphology, irregular R-R intervals, and inconsistent cardiac cycle patterns [200]. Figure A.9 illustrates SHAP-based explanations that highlight this variability; for instance, the R-R interval between the 8^{th} and 9^{th} beats is relatively short, whereas the interval between the 9^{th} and 10^{th} beats is longer. Additionally, the QRS complex on lead II is inverted. In both instances, the R-peak of the 9^{th} beat is marked in red, indicating the model's attention to these irregularities in classifying the signal as SI. Similarly, Figure A.10 displays a Grad-CAM++ heatmap, where leads I and II emphasize segments of the ECG with inconsistent cardiac activity, demonstrating the model's focus on these features when identifying SI.

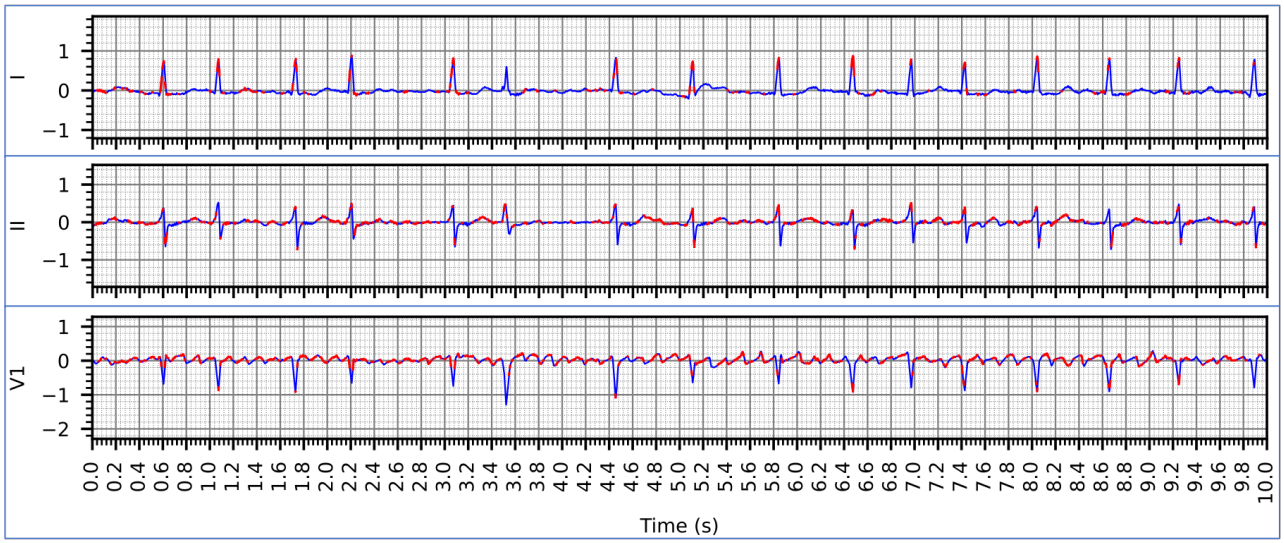


Figure A.11: Example of instance level interpretation for AFIB using SHAP values.

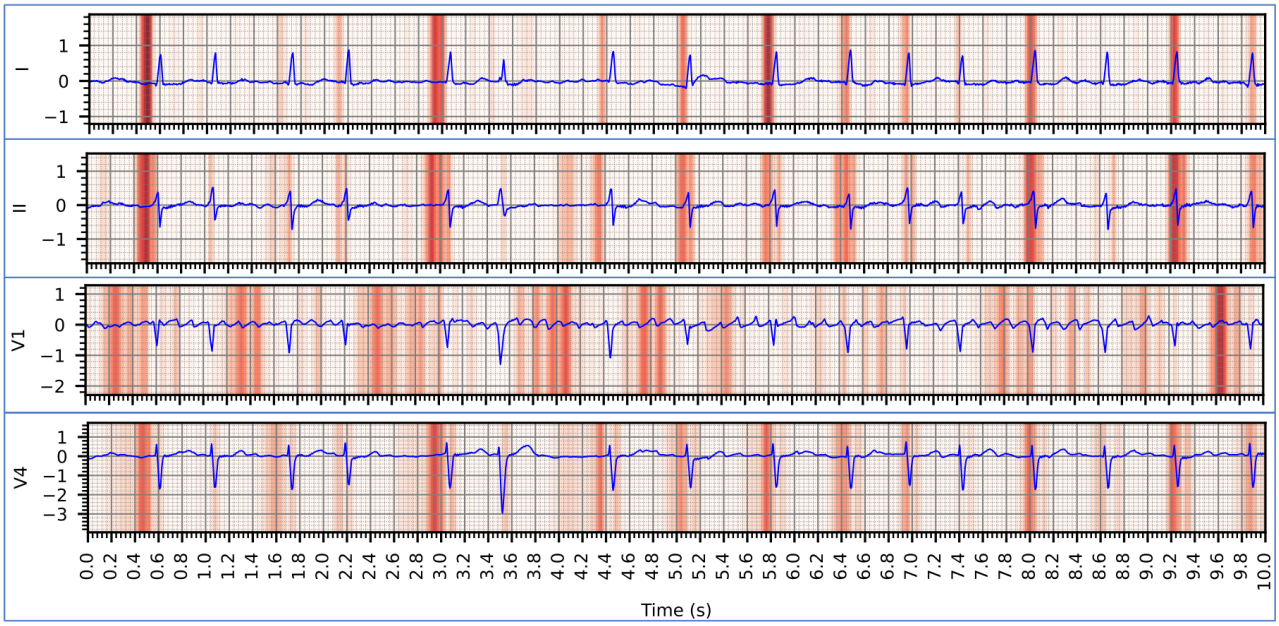


Figure A.12: Example of instance level interpretation for AFIB using Grad-CAM++ heatmaps, the same ECG sample to Figure A.11.

Atrial Fibrillation (AFIB) is characterized by an irregularly irregular ventricular rhythm and the presence of fibrillatory waves (f-waves) [200]. These f-waves are most prominent in lead V1 and are highlighted in red in Figure A.11, indicating their importance in the model's classification. Similarly, Figure A.12 shows a Grad-CAM++ heatmap, where lead V1 shows a warmer intensity, demonstrating the model's focus on the f-waves to identify AFIB.

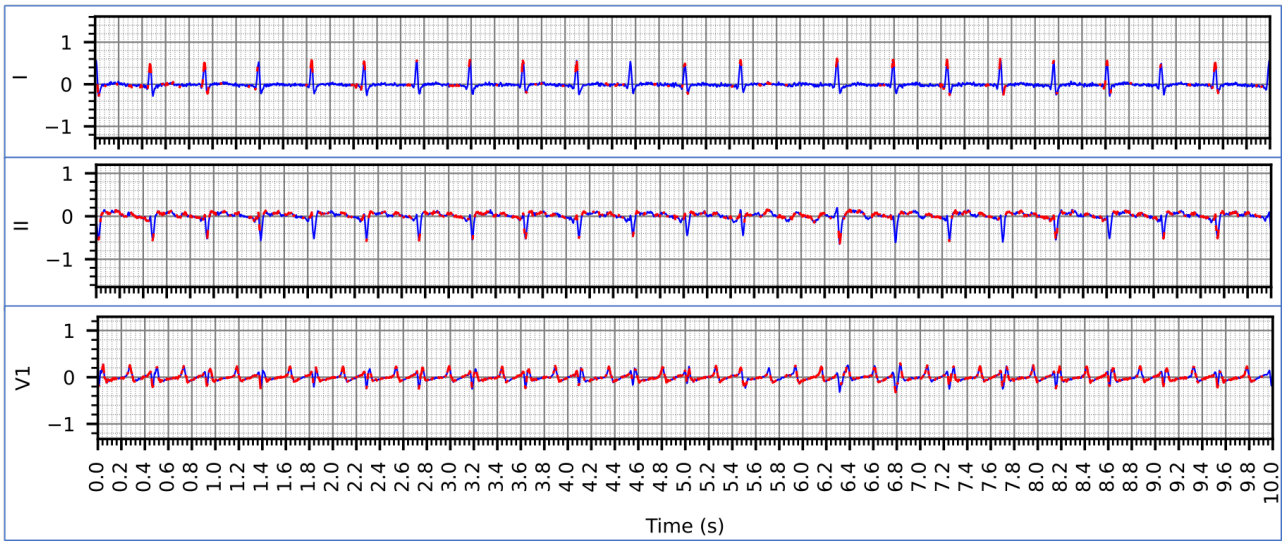


Figure A.13: Example of instance level interpretation for AF using SHAP values.

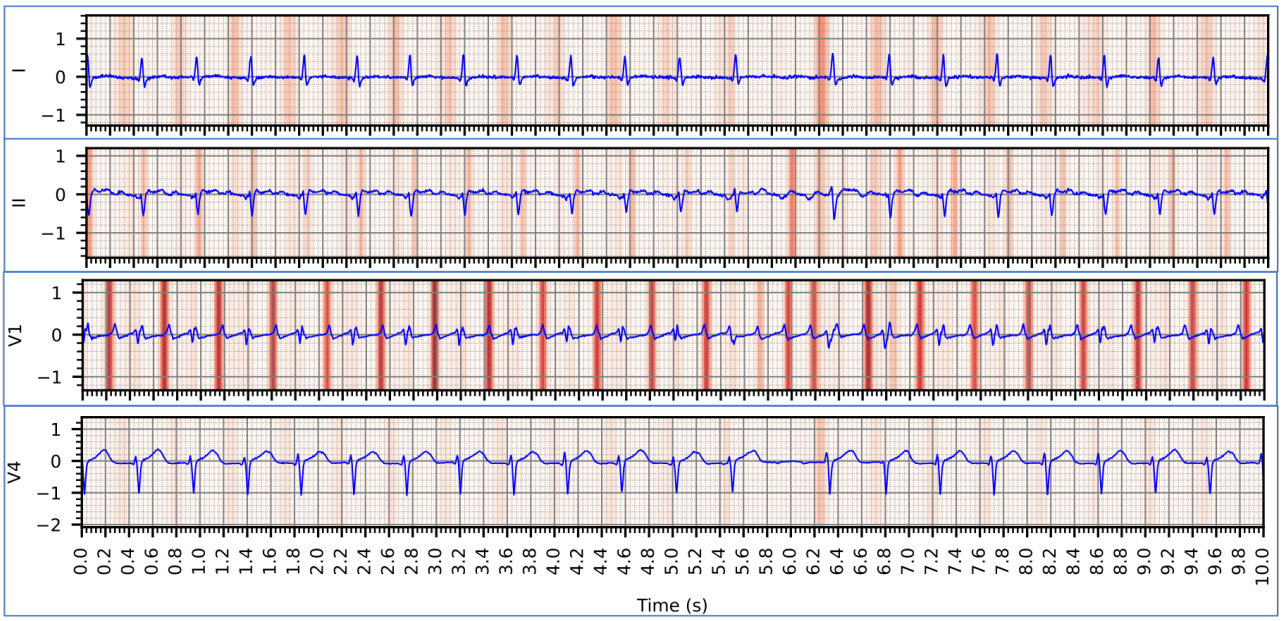


Figure A.14: Example of instance level interpretation for AF using Grad-CAM++ heatmaps, the same ECG sample to Figure A.13.

Atrial Flutter (AF) is characterized by saw-tooth-like atrial waves occurring between consecutive QRS complexes [228]. These atrial waves are prominent in lead V1. As illustrated in Figure A.13, the SHAP values highlight these features in red within lead V1. Similarly, the Grad-CAM++ heatmap in Figure A.14 indicates that the model emphasizes these saw-tooth waves in lead V1 in classifying the ECG as AF.

Appendix B: Deployed System

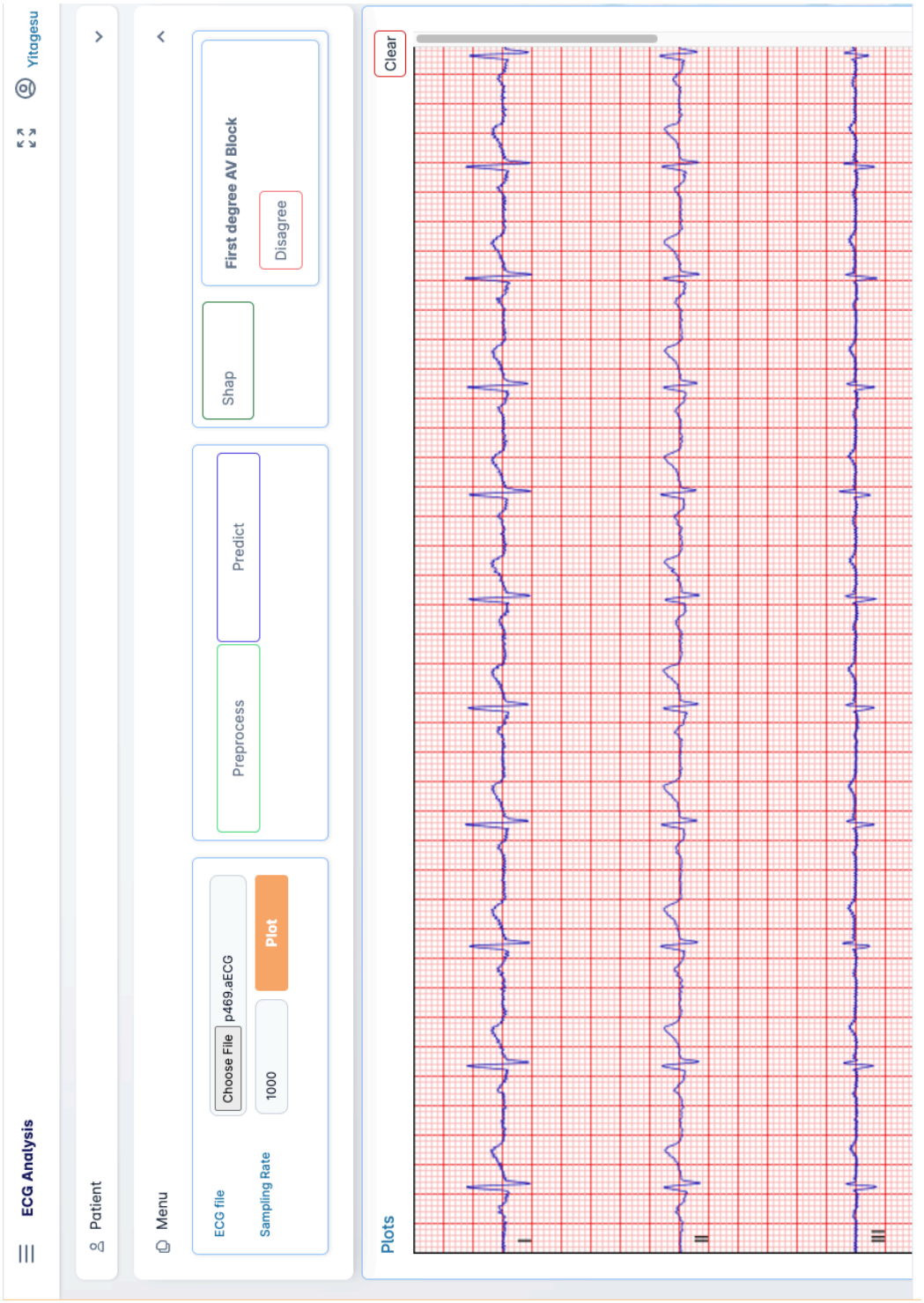


Figure B.1: GUI of ECG Analysis System.

Bibliography

[1] "Fact Sheet: Cardiovascular Diseases." https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2022-05-23.

[2] F. Morris, *ABC of Clinical Electrocardiography*. Malden, Mass. Oxford: Blackwell Pub, 2008.

[3] Yugandhar R. Manda and Krishna M. Baradhi, "Cardiac Catheterization Risks and Complications," *StatPearls Publishing*, January 2021.

[4] M. E. Jørgensen, C. Andersson, B. L. Nørgaard, J. Abdulla, J. B. Shreibati, C. Torp-Pedersen, G. H. Gislason, R. E. Shaw, and M. A. Hlatky, "Functional Testing or Coronary Computed Tomography Angiography in Patients With Stable Coronary Artery Disease," *Journal of the American College of Cardiology*, vol. 69, pp. 1761–1770, April 2017.

[5] I. S. Syed, J. F. Glockner, D. Feng, P. A. Araoz, M. W. Martinez, W. D. Edwards, M. A. Gertz, A. Dispenzieri, J. K. Oh, D. Bellavia, A. J. Tajik, and M. Grogan, "Role of Cardiac Magnetic Resonance Imaging in the Detection of Cardiac Amyloidosis," *JACC: Cardiovascular Imaging*, vol. 3, pp. 155–164, February 2010.

[6] J. Pannu, S. Poole, N. Shah, and N. H. Shah, "Assessing Screening Guidelines for Cardiovascular Disease Risk Factors using Routinely Collected Data.," *Scientific reports*, vol. 7, p. 6488, July 2017.

[7] Tammiraju Iragavarapu; T Radhakrishna; KJagadish Babu and R Sanghamitra, "Acute coronary syndrome in young - A tertiary care centre experience with reference to coronary angiogram," *Journal of the Practice of Cardiovascular Sciences*, vol. 5, no. 1, p. 18, 2019.

[8] J. Higueras, S. Gómez-Talavera, V. Cañadas, R. Bover, M.-L. P, J. C. Gómez-Polo, C. Olmos, C. Fernandez, J. Villacastín, and C. Macaya, "Expertise in Interpretation of 12-Lead Electrocardiograms of Staff and Residents Physician: Current Knowledge and Comparison between Two Different Teaching Methods," *Journal of Cardiology & Current Research*, vol. 5, February 2016.

[9] K. Amini, A. Mirzaei, M. Hosseini, H. Zandian, I. Azizpour, and Y. Haghi, "Assessment of electrocardiogram interpretation competency among healthcare professionals and students of ardabil university of medical sciences: a multidisciplinary study," *BMC Medical Education*, vol. 22, June 2022.

[10] M. Getachew, T. Beyene, and S. Kebede, "Electrocardiography interpretation competency of medical interns: Experience from two ethiopian medical schools," *Emergency Medicine International*, vol. 2020, pp. 1–6, May 2020.

[11] J. Schläpfer and H. J. Wellens, "Computer-Interpreted Electrocardiograms," *Journal of the American College of Cardiology*, vol. 70, pp. 1183–1192, August 2017.

[12] P. Martínez-Losas, J. Higueras, J. C. Gómez-Polo, P. Brabyn, J. M. F. Ferrer, V. Cañadas, and J. P. Villacastín, "The influence of computerized interpretation of an electrocardiogram reading," *The American Journal of Emergency Medicine*, vol. 34, pp. 2031–2032, October 2016.

[13] A. J. Weinhaus and K. P. Roberts, "Anatomy of the Human Heart," in *Handbook of Cardiac Anatomy, Physiology, and Devices*, pp. 59–85, Humana Press, 2009.

[14] J. Moini, *Anatomy and Physiology*, ch. 18: The Heart, pp. 449–471. Jones and Bartlett Learning, third ed., 2020.

[15] S. Dey, R. Pal, and S. Biswas, "Deep learning algorithms for efficient analysis of ECG signals to detect heart disorders," in *Biomedical Engineering*, IntechOpen, April 2022.

[16] L. Fabricius Ekenberg, D. E. Høfsten, S. M. Rasmussen, J. Mølgaard, P. Hasbak, H. B. D. Sørensen, C. S. Meyhoff, and E. K. Aasvang, "Wireless single-lead versus standard 12-lead ecg, for st-segment deviation during adenosine cardiac stress scintigraphy," *Sensors*, vol. 23, p. 2962, March 2023.

[17] J. Park, J. An, J. Kim, S. Jung, Y. Gil, Y. Jang, K. Lee, and I. young Oh, "Study on the use of standard 12-lead ECG data for rhythm-type ECG classification problems," *Computer Methods and Programs in Biomedicine*, p. 106521, November 2021.

[18] Dr. Araz Rawshani, "The ECG leads: electrodes, limb leads, chest (precordial) leads, 12-Lead ECG (EKG)." https://ecgwaves.com/topic/ekg-ecg-leads-electrodes-systems-limb-chest-precordial/. Accessed: 2022-06-16.

[19] P. M. Rautaharju, B. Surawicz, and L. S. Gettes, "AHA/ACCF/HRS Recommendations for the Standardization and Interpretation of the Electrocardiogram," *Circulation*, vol. 119, March 2009.

[20] A. H. Ribeiro, M. H. Ribeiro, G. M. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. S. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira, T. B. Schön, and A. L. P. Ribeiro, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications*, vol. 11, April 2020.

[21] K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Artificial intelligence-enhanced electrocardiography in cardiovascular disease management," *Nature Reviews Cardiology*, vol. 18, pp. 465–478, February 2021.

[22] M. Alfaras, M. C. Soriano, and S. Ortín, "A fast machine learning model for ECG-based heartbeat classification and arrhythmia detection," *Frontiers in Physics*, vol. 7, July 2019.

[23] A. H. Kashou, W.-Y. Ko, Z. I. Attia, M. S. Cohen, P. A. Friedman, and P. A. Noseworthy, "A comprehensive artificial intelligence–enabled electrocardiogram interpretation program," *Cardiovascular Digital Health Journal*, vol. 1, pp. 62–70, September 2020.

[24] M. Hammad, A. Maher, K. Wang, F. Jiang, and M. Amrani, "Detection of abnormal heart conditions based on characteristics of ECG signals," *Measurement*, vol. 125, pp. 634–644, September 2018.

[25] K. M. Aamir, M. Ramzan, S. Skinadar, H. U. Khan, U. Tariq, H. Lee, Y. Nam, and M. A. Khan, "Automatic heart disease detection by classification of ventricular arrhythmias on ECG using machine learning," *Computers, Materials & Continua*, vol. 71, pp. 17–33, October 2022.

[26] X. Zhang, K. Gu, S. Miao, X. Zhang, Y. Yin, C. Wan, Y. Yu, J. Hu, Z. Wang, T. Shan, S. Jing, W. Wang, Y. Ge, Y. Chen, J. Guo, and Y. Liu, "Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system," *Cardiovascular Diagnosis and Therapy*, vol. 10, pp. 227–235, April 2020.

[27] S. Śmigiel, K. Pałczyński, and D. Ledziński, "ECG signal classification using deep learning techniques based on the PTB-XL dataset," *Entropy*, vol. 23, p. 1121, August 2021.

[28] S. Ortín, M. C. Soriano, M. Alfaras, and C. R. Mirasso, "Automated real-time method for ventricular heartbeat classification," *Computer Methods and Programs in Biomedicine*, vol. 169, pp. 1–8, February 2019.

[29] J. Gao, H. Zhang, P. Lu, and Z. Wang, "An effective LSTM recurrent network to detect arrhythmia on imbalanced ECG dataset," *Journal of Healthcare Engineering*, vol. 2019, pp. 1–10, October 2019.

[30] D. W. Feyisa, T. G. Debelee, Y. M. Ayano, S. R. Kebede, and T. F. Assore, "Lightweight Multireceptive Field CNN for 12-Lead ECG Signal Classification," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–14, August 2022.

[31] X. Liu, H. Wang, Z. Li, and L. Qin, "Deep learning in ECG diagnosis: A review," *Knowledge-Based Systems*, vol. 227, p. 107187, September 2021.

[32] A. H. Khan, M. Hussain, and M. K. Malik, "Cardiac Disorder Classification by Electrocardiogram Sensing Using Deep Neural Network," *Complexity*, vol. 2021, pp. 1–8, March 2021.

[33] A. H. Kashou, S. K. Mulpuru, A. J. Deshmukh, W.-Y. Ko, Z. I. Attia, R. E. Carter, P. A. Friedman, and P. A. Noseworthy, "An artificial intelligence–enabled ECG algorithm for comprehensive ECG interpretation: Can it pass the 'turing test'?," *Cardiovascular Digital Health Journal*, vol. 2, pp. 164–170, June 2021.

[34] T. A. A. Abdullah, M. S. M. Zahid, and W. Ali, "A review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions," *Symmetry*, vol. 13, p. 2439, December 2021.

[35] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," *CoRR*, vol. abs/2006.11371, 2020.

[36] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F. I. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, May 2020.

[37] G. Quer, R. Arnaout, M. Henne, and R. Arnaout, "Machine learning and the future of cardiovascular care," *Journal of the American College of Cardiology*, vol. 77, pp. 300–313, January 2021.

[38] S. Dhyani, A. Kumar, and S. Choudhury, "Analysis of ecg-based arrhythmia detection system using machine learning," *MethodsX*, vol. 10, p. 102195, 2023.

[39] S. Śmigiel, "Ecg classification using orthogonal matching pursuit and machine learning," *Sensors*, vol. 22, p. 4960, June 2022.

[40] L. Bickmann, L. Plagwitz, and J. Varghese, *Benchmarking Approaches: Time Series Versus Feature-Based Machine Learning in ECG Analysis on the PTB-XL Dataset*. IOS Press, August 2024.

[41] P. Wagner, N. Strodthoff, R.-D. Bousseljot, W. Samek, and T. Schaeffter, "Ptb-xl, a large publicly available electrocardiography dataset," 2020.

[42] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, March 2021.

[43] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, "Explainable AI decision model for ECG data of cardiac disorders," *Biomedical Signal Processing and Control*, vol. 75, p. 103584, May 2022.

[44] K. Pałczyński, S. Śmigiel, D. Ledziński, and S. Bujnowski, "Study of the few-shot learning for ECG classification based on the PTB-XL dataset," *Sensors*, vol. 22, p. 904, January 2022.

[45] D. D. Demissie and F. A. Andargie, "Explainable rhythm-based heart disease detection from ecg signals," in *Pan-African Conference on Artificial Intelligence*, pp. 101–116, Springer Nature Switzerland, April 2024.

[46] F. Yang, X. Zhang, and Y. Zhu, "Pdnet: A convolutional neural network has potential to be deployed on small intelligent devices for arrhythmia diagnosis," *Computer Modeling in Engineering and Sciences*, vol. 125, no. 1, pp. 365–382, 2020.

[47] S. Śmigiel, K. Pałczyński, and D. Ledziński, "ECG signal classification using deep learning techniques based on the PTB-XL dataset," *Entropy*, vol. 23, p. 1121, August 2021.

[48] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ecg classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 664–675, March 2016.

[49] M. Sepahvand and F. Abdali-Mohammadi, "A novel method for reducing arrhythmia classification from 12-lead ecg signals to single-lead ecg with minimal loss of accuracy through teacher-student knowledge distillation," *Information Sciences*, vol. 593, pp. 64–77, May 2022.

[50] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1519–1528, May 2021.

[51] H. De Melo Ribeiro, A. Arnold, J. P. Howard, M. J. Shun-Shin, Y. Zhang, D. P. Francis, P. B. Lim, Z. Whinnett, and M. Zolgharni, "Ecg-based real-time arrhythmia monitoring using quantized deep neural networks: A feasibility study," *Computers in Biology and Medicine*, vol. 143, p. 105249, April 2022.

[52] J. Li, Y. Si, T. Xu, and S. Jiang, "Deep convolutional neural network based ecg classification system using information fusion and one-hot encoding techniques," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–10, December 2018.

[53] A. A. Ahmed, W. Ali, T. A. A. Abdullah, and S. J. Malebary, "Classifying cardiac arrhythmia from ecg signal using 1d cnn deep learning model," *Mathematics*, vol. 11, p. 562, January 2023.

[54] S. Nurmaini, A. Tondas, A. Darmawahyuni, M. N. Rachmatullah, R. Umi Partan, F. Firdaus, B. Tutuko, F. Pratiwi, A. H. Juliano, and R. Khoirani, "Robust detection of atrial fibrillation from short-term electrocardiogram using convolutional neural networks," *Future Generation Computer Systems*, vol. 113, pp. 304–317, December 2020.

[55] G. Sannino and G. De Pietro, "A deep learning approach for ecg-based heartbeat classification for arrhythmia detection," *Future Generation Computer Systems*, vol. 86, pp. 446–455, September 2018.

[56] H. Narotamo, M. Dias, R. Santos, A. V. Carreiro, H. Gamboa, and M. Silveira, "Deep learning for ecg classification: A comparative study of 1d and 2d representations and multimodal fusion approaches," *Biomedical Signal Processing and Control*, vol. 93, p. 106141, July 2024.

[57] G. B. Moody and R. G. Mark, "Mit-bih arrhythmia database," 1992.

[58] G. B. Moody and R. G. Mark, "Mit-bih atrial fibrillation database," 1992.

[59] Y. M. Ayano, F. Schwenker, B. D. Dufera, and T. G. Debelee, "Interpretable machine learning techniques in ECG-based heart disease classification: A systematic review," *Diagnostics*, vol. 13, p. 111, December 2022.

[60] Y. Zhang and J. Li, "Application of heartbeat-attention mechanism for detection of myocardial infarction using 12-lead ecg records," *Applied Sciences*, vol. 9, p. 3328, August 2019.

[61] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro, "Code dataset," November 2021.

[62] J. Zheng, C. Rakovski, Sidy Danioko, Jianming Zhang, H. Yao, and G. Hangyuan, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," 2019.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2016.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2016.

[65] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, May 2017.

[66] L. Sadouk, *CNN Approaches for Time Series Classification*. IntechOpen, November 2019.

[67] B. Król-Józaga, "Atrial fibrillation detection using convolutional neural networks on 2-dimensional representation of ecg signal," *Biomedical Signal Processing and Control*, vol. 74, p. 103470, April 2022.

[68] S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, and M. Gabbouj, "1-d convolutional neural networks for signal processing applications," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2019.

[69] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru," May 2023.

[70] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. B. Schön, "Beyond exploding and vanishing gradients: analysing rnn training using attractors and smoothness," June 2019.

[71] S. Hiriyannaiah, S. G M, K. M H M, and K. G. Srinivasa, "A comparative study and analysis of lstm deep neural networks for heartbeats classification," *Health and Technology*, vol. 11, pp. 663–671, April 2021.

[72] S. Kuila, N. Dhanda, and S. Joardar, "Ecg signal classification and arrhythmia detection using elm-rnn," *Multimedia Tools and Applications*, vol. 81, pp. 25233–25249, March 2022.

[73] M. Wang, S. Rahardja, P. Fränti, and S. Rahardja, "Single-lead ecg recordings modeling for end-to-end recognition of atrial fibrillation with dual-path rnn," *Biomedical Signal Processing and Control*, vol. 79, p. 104067, January 2023.

[74] S. Saadatnejad, M. Oveisi, and M. Hashemi, "Lstm-based ecg classification for continuous monitoring on personal wearable devices," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 515–523, February 2020.

[75] E. Prabhakararao and S. Dandapat, "Attentive rnn-based network to fuse 12-lead ecg and clinical features for improved myocardial infarction diagnosis," *IEEE Signal Processing Letters*, vol. 27, pp. 2029–2033, November 2020.

[76] S. Boda, M. Mahadevappa, and P. Kumar Dutta, "An automated patient-specific ecg beat classification using lstm-based recurrent neural networks," *Biomedical Signal Processing and Control*, vol. 84, p. 104756, July 2023.

[77] A. Çınar and S. A. Tuncer, "Classification of normal sinus rhythm, abnormal arrhythmia and congestive heart failure ecg signals using lstm and hybrid cnn-svm deep neural networks," *Computer Methods in Biomechanics and Biomedical Engineering*, vol. 24, pp. 203–214, September 2020.

[78] R.-D. Bousseljot, D. Kreiseler, and A. Schnabel, "The ptb diagnostic ecg database," September 2004.

[79] D. Salman, C. Direkoglu, M. Kusaf, and M. Fahrioglu, "Hybrid deep learning models for time series forecasting of solar power," *Neural Computing and Applications*, vol. 36, pp. 9095–9112, February 2024.

[80] O. Yildirim, M. Talo, E. J. Ciaccio, R. S. Tan, and U. R. Acharya, "Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ECG records," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105740, December 2020.

[81] Z. Yang, A. Jin, Y. Liu, W. Lv, and X. Zhu, *The Fusion Model of ResNet and GRU Based on Simplified Self-Attention for ECG Classification on PTB-XL Dataset*, pp. 87–103. Springer Nature Switzerland, May 2024.

[82] J. Jing, J. Zhang, A. Liu, M. Gao, R. Qian, and X. Chen, "Ecg-based multiclass arrhythmia classification using beat-level fusion network," *Journal of Healthcare Engineering*, vol. 2023, pp. 1–10, November 2023.

[83] C. Lai, S. Zhou, and N. A. Trayanova, "Optimal ecg-lead selection increases generalizability of deep learning on ecg abnormality classification," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, October 2021.

[84] X. Xie, H. Liu, D. Chen, M. Shu, and Y. Wang, "Multilabel 12-lead ecg classification based on leadwise grouping multibranch network," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, April 2022.

[85] Y.-J. Chen, C.-L. Liu, V. S. Tseng, Y.-F. Hu, and S.-A. Chen, "Large-scale classification of 12-lead ecg with deep learning," in *2019 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, May 2019.

[86] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, J. Li, and E. N. Yin Kwee, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics*, vol. 8, pp. 1368–1373, September 2018.

[87] H. Zhu, Y. Zhao, Y. Pan, H. Xie, F. Wu, and R. Huan, "Robust heartbeat classification for wearable single-lead ecg via extreme gradient boosting," *Sensors*, vol. 21, p. 5290, August 2021.

[88] Edwar d Hanc e Shortliffe, *Computer-Based Medical Consultations: Mycin*. Elsevier, 1976.

[89] D. S. Watson, "Conceptual challenges for interpretable machine learning," *Synthese*, vol. 200, March 2022.

[90] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning – a brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops*, pp. 417–431, Springer International Publishing, 2020.

[91] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 22071–22080, October 2019.

[92] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, p. 832, July 2019.

[93] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, June 2020.

[94] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, July 2021.

[95] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, pp. 68–77, December 2019.

[96] T. A. A. Abdullah, M. S. M. Zahid, and W. Ali, "A review of interpretable ML in healthcare: Taxonomy, applications, challenges, and future directions," *Symmetry*, vol. 13, p. 2439, December 2021.

[97] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 4768–4777, Curran Associates Inc., 2017.

[98] D. Rothman, *Hands-On Explainable AI (XAI) with Python*. Packt Publishing, July 2020.

[99] E. Angelaki, M. E. Marketou, G. D. Barmparis, A. Patrianakos, P. E. Vardas, F. Parthenakis, and G. P. Tsironis, "Detection of abnormal left ventricular geometry in patients without cardiovascular disease through machine learning: An ECG-based approach," *The Journal of Clinical Hypertension*, vol. 23, pp. 935–945, January 2021.

[100] R. Rouhi, M. Clausel, J. Oster, and F. Lauer, "An interpretable hand-crafted feature-based model for atrial fibrillation detection," *Frontiers in Physiology*, vol. 12, May 2021.

[101] A. Anand, T. Kadian, M. K. Shetty, and A. Gupta, "Explainable AI decision model for ECG data of cardiac disorders," *Biomedical Signal Processing and Control*, vol. 75, p. 103584, May 2022.

[102] L. Ibrahim, M. Mesinovic, K.-W. Yang, and M. A. Eid, "Explainable prediction of acute myocardial infarction using machine learning and shapley values," *IEEE Ac-*

cess, vol. 8, pp. 210410–210417, 2020.

[103] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to shapley values," *Artificial Intelligence*, vol. 298, p. 103502, September 2021.

[104] C. Frye, C. Rowat, and I. Feige, "Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[105] I. Basu and S. Maji, "Multicollinearity correction and combined feature effect in shapley values," in *Lecture Notes in Computer Science*, pp. 79–90, Springer International Publishing, 2022.

[106] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," June 2020.

[107] J. Yang, "Fast treeshap: Accelerating shap value computation for trees," September 2021.

[108] S. Sylvester, M. Sagehorn, T. Gruber, M. Atzmueller, and B. Schöne, "Shap value-based erp analysis (sherpa): Increasing the sensitivity of eeg signals with explainable ai methods," *Behavior Research Methods*, vol. 56, pp. 6067–6081, March 2024.

[109] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?"," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, August 2016.

[110] I. Neves, D. Folgado, S. Santos, M. Barandas, A. Campagner, L. Ronzio, F. Cabitza, and H. Gamboa, "Interpretable heartbeat classification using local model-agnostic explanations on ECGs," *Computers in Biology and Medicine*, vol. 133, p. 104393, June 2021.

[111] M. Bodini, M. W. Rivolta, and R. Sassi, "Interpretability analysis of machine learning algorithms in the detection of ST-elevation myocardial infarction," in *2020 Computing in Cardiology Conference (CinC)*, Computing in Cardiology, December 2020.

[112] J. Ojha, H. Haugerud, A. Yazidi, and P. G. Lind, "Exploring interpretable ai methods for ecg data classification," in *The Fifth Workshop on Intelligent Cross-Data Analysis and Retrieval*, ICMR '24, ACM, June 2024.

[113] Z. Zhou, G. Hooker, and F. Wang, "S-LIME: Stabilized-LIME for Model Explanation," in *https://doi.org/10.3390/make3030027*, KDD '21, (New York, NY, USA), p. 2429–2438, Association for Computing Machinery, 2021.

[114] G. Visani, E. Bagli, and F. Chesani, "Optilime: Optimized lime explanations for diagnostic computer algorithms," June 2020.

[115] M. R. Zafar and N. Khan, "Deterministic local interpretable model-agnostic explanations for stable explainability," *Machine Learning and Knowledge Extraction*, vol. 3, pp. 525–541, June 2021.

[116] S. M. Shankaranarayana and D. Runje, "ALIME: Autoencoder based approach for local interpretability," in *Intelligent Data Engineering and Automated Learning – IDEAL 2019*, pp. 454–463, Springer International Publishing, 2019.

[117] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ACM, February 2020.

[118] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of machine learning research : JMLR*, vol. 20, 2019.

[119] Q. Au, J. Herbinger, C. Stachl, B. Bischl, and G. Casalicchio, "Grouped feature importance and combined features effect plot," *Data Mining and Knowledge Discovery*, vol. 36, pp. 1401–1450, June 2022.

[120] A. Sood and M. Craven, "Feature importance explanations for temporal black-box models," February 2021.

[121] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," *Statistics and Computing*, vol. 31, October 2021.

[122] R. Hammachi, N. Messaoudi, and S. Belkacem, "Ecg beats classification with interpretability," in *2022 International Conference of Advanced Technology in Electronic and Electrical Engineering (ICATEEE)*, IEEE, November 2022.

[123] Y. Izza, A. Ignatiev, and J. Marques-Silva, "On explaining decision trees," October 2020.

[124] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, IEEE, June 2018.

[125] S. Masís, *Interpretable Machine Learning with Python*. Packt Publishing, March 2021.

[126] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Information Sciences*, vol. 572, pp. 522–542, September 2021.

[127] A. Rath, D. Mishra, and G. Panda, "Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique," *Frontiers in Big Data*, vol. 5, October 2022.

[128] W. Zhang, R. Li, S. Shen, J. Yao, Y. Peng, G. Chen, B. Zhou, and Z. Wang, "Interpretable detection and location of myocardial infarction based on ventricular fusion rule features," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–15, October 2021.

[129] F. Maturo and R. Verde, "Pooling random forest and functional data analysis for biomedical signals supervised classification: Theory and application to electrocardiogram data," *Statistics in Medicine*, vol. 41, pp. 2247–2275, February 2022.

[130] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, pp. 2674–2693, August 2019.

[131] M. Porumb, E. Iadanza, S. Massaro, and L. Pecchia, "A convolutional neural network approach to detect congestive heart failure," *Biomedical Signal Processing and Control*, vol. 55, p. 101597, January 2020.

[132] V. Jahmunah, E. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-CAM technique on ECG signals," *Computers in Biology and Medicine*, vol. 146, p. 105550, July 2022.

[133] S. A. Hicks, J. L. Isaksen, V. Thambawita, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, I. Strümke, C. Ellervik, M. S. Olesen, T. Hansen, C. Graff, N.-H. Holstein-Rathlou, P. Halvorsen, M. M. Maleckar, M. A. Riegler, and J. K. Kanters, "Explaining deep neural networks for knowledge discovery in electrocardiogram analysis," *Scientific Reports*, vol. 11, May 2021.

[134] R. Fang, C.-C. Lu, C.-T. Chuang, and W.-H. Chang, "A visually interpretable detection method combines 3-d ECG with a multi-VGG neural network for myocardial infarction identification," *Computer Methods and Programs in Biomedicine*, vol. 219, p. 106762, June 2022.

[135] M. Bodini, M. W. Rivolta, and R. Sassi, "Opening the black box: interpretability of machine learning algorithms in electrocardiography," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, October 2021.

[136] J. Bridge, L. Fu, W. Lin, Y. Xue, G. Y. H. Lip, and Y. Zheng, "Artificial intelligence to detect abnormal heart rhythm from scanned electrocardiogram tracings," *Journal of Arrhythmia*, vol. 38, pp. 425–431, March 2022.

[137] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1519–1528, May 2021.

[138] S. Mousavi, F. Afghah, and U. R. Acharya, "HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks," *Computers in Biology and Medicine*, vol. 127, p. 104057, December 2020.

[139] Y. Jin, J. Liu, Y. Liu, C. Qin, Z. Li, D. Xiao, L. Zhao, and C. Liu, "A novel interpretable method based on dual-level attentional deep neural network for actual multilabel arrhythmia detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022.

[140] H. Lee and M. Shin, "Learning explainable time-morphology patterns for automatic arrhythmia classification from short single-lead ECGs," *Sensors*, vol. 21, p. 4331, June 2021.

[141] L. Fu, B. Lu, B. Nie, Z. Peng, H. Liu, and X. Pi, "Hybrid network with attention mechanism for detection and location of myocardial infarction based on 12-lead electrocardiogram signals," *Sensors*, vol. 20, p. 1020, February 2020.

[142] N. L. Wickramasinghe and M. Athif, "Multi-label classification of reduced-lead ECGs using an interpretable deep convolutional neural network," *Physiological Measurement*, vol. 43, p. 064002, June 2022.

[143] D. Zhang, S. Yang, X. Yuan, and P. Zhang, "Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram," *iScience*, vol. 24, p. 102373, April 2021.

[144] M. Rashed-Al-Mahfuz, M. A. Moni, P. Lio', S. M. S. Islam, S. Berkovsky, M. Khushi, and J. M. W. Quinn, "Deep convolutional neural networks based ECG beats classification to diagnose cardiovascular conditions," *Biomedical Engineering Letters*, vol. 11, pp. 147–162, February 2021.

[145] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, June 2016.

[146] M. Goswami, B. Boecking, and A. Dubrawski, "Weak supervision for affordable modeling of electrocardiogram data.," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2021, pp. 536–545, 2021.

[147] S. D. Goodfellow, A. Goodwin, R. Greer, P. C. Laussen, M. Mazwi, and D. Eytan, "Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings," in *Proceedings of the 3rd Machine Learning for Healthcare Conference* (F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, eds.), vol. 85 of *Proceedings of Machine Learning Research*, pp. 83–101, PMLR, 17–18 August 2018.

[148] H. Jung and Y. Oh, "Towards better explanations of class activation mapping," February 2021.

[149] J. Wang, X. Qiao, C. Liu, X. Wang, Y. Liu, L. Yao, and H. Zhang, "Automated ECG classification using a non-local convolutional block attention module," *Computer Methods and Programs in Biomedicine*, vol. 203, p. 106006, May 2021.

[150] A. Raza, K. P. Tran, L. Koehl, and S. Li, "Designing ECG monitoring healthcare system with federated transfer learning and explainable AI," *Knowledge-Based Systems*, vol. 236, p. 107763, January 2022.

[151] G. M., V. Ravi, S. V, G. E.A, and S. K.P, "Explainable deep learning-based approach for multilabel classification of electrocardiogram," *IEEE Transactions on Engineering Management*, pp. 1–13, 2022.

[152] R. R. Lopes, H. Bleijendaal, L. A. Ramos, T. E. Verstraelen, A. S. Amin, A. A. Wilde, Y. M. Pinto, B. A. de Mol, and H. A. Marquering, "Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: An application to phospholamban p.arg14del mutation carriers," *Computers in Biology and Medicine*, vol. 131, p. 104262, April 2021.

[153] D. Li, H. Wu, J. Zhao, Y. Tao, and J. Fu, "Automatic classification system of arrhythmias using 12-lead ECGs with a deep neural network based on an attention mechanism," *Symmetry*, vol. 12, p. 1827, November 2020.

[154] Y. Cho, J. myoung Kwon, K.-H. Kim, J. R. Medina-Inojosa, K.-H. Jeon, S. Cho, S. Y. Lee, J. Park, and B.-H. Oh, "Artificial intelligence algorithm for detecting myocardial infarction using six-lead electrocardiography," *Scientific Reports*, vol. 10, November 2020.

[155] J. myoung Kwon, K.-H. Kim, K.-H. Jeon, S. Y. Lee, J. Park, and B.-H. Oh, "Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 28, October 2020.

[156] V. Sangha, B. J. Mortazavi, A. D. Haimovich, A. H. Ribeiro, C. A. Brandt, D. L. Jacoby, W. L. Schulz, H. M. Krumholz, A. L. P. Ribeiro, and R. Khera, "Automated multilabel diagnosis on electrocardiographic images and signals," *Nature Communications*, vol. 13, March 2022.

[157] J.-M. Kwon, S. Y. Lee, K.-H. Jeon, Y. Lee, K.-H. Kim, J. Park, B.-H. Oh, and M.-M. Lee, "Deep learning–based algorithm for detecting aortic stenosis using electrocardiography," *Journal of the American Heart Association*, vol. 9, April 2020.

[158] M. Jiang, Y. Qiu, W. Zhang, J. Zhang, Z. Wang, W. Ke, Y. Wu, and Z. Wang, "Visualization deep learning model for automatic arrhythmias classification," *Physiological Measurement*, vol. 43, p. 085003, August 2022.

[159] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, March 2018.

[160] J. myoung Kwon, K.-H. Kim, J. Medina-Inojosa, K.-H. Jeon, J. Park, and B.-H. Oh, "Artificial intelligence for early prediction of pulmonary hypertension using electrocardiography," *The Journal of Heart and Lung Transplantation*, vol. 39, pp. 805–814, August 2020.

[161] Y.-Y. Jo, J. myoung Kwon, K.-H. Jeon, Y.-H. Cho, J.-H. Shin, Y.-J. Lee, M.-S. Jung, J.-H. Ban, K.-H. Kim, S. Y. Lee, J. Park, and B.-H. Oh, "Detection and classification of arrhythmia using an explainable deep learning model," *Journal of Electrocardiology*, vol. 67, pp. 124–132, July 2021.

[162] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[163] E. Mohamed, K. Sirlantzis, and G. Howells, "A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation," *Displays*, vol. 73, p. 102239, July 2022.

[164] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un)reliability of saliency methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 267–280, Springer International Publishing, 2019.

[165] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209, Springer International Publishing, 2019.

[166] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, pp. 247–278, March 2021.

[167] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, February 2018.

[168] Y.-J. Jung, S.-H. Han, and H.-J. Choi, "Explaining CNN and RNN using selective layer-wise relevance propagation," *IEEE Access*, vol. 9, pp. 18670–18681, January 2021.

[169] X. Huang, S. Jamonnak, Y. Zhao, T. H. Wu, and W. Xu, "A Visual Designer of Layer-wise Relevance Propagation Models," *Computer Graphics Forum*, vol. 40, pp. 227–238, June 2021.

[170] M. Resta, A. Monreale, and D. Bacciu, "Occlusion-based explanations in deep recurrent models for biomedical signals," *Entropy*, vol. 23, p. 1064, August 2021.

[171] F. J. Caro-Lopera, V. Leiva, and N. Balakrishnan, "Connection between the hadamard and matrix products with an application to matrix-variate birnbaum–saunders distributions," *Journal of Multivariate Analysis*, vol. 104, pp. 126–139, February 2012.

[172] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018.

[173] H. Bleijendaal, L. A. Ramos, R. R. Lopes, T. E. Verstraelen, S. W. Baalman, M. D. O. Pool, F. V. Tjong, F. M. Melgarejo-Meseguer, F. J. Gimeno-Blanes, J. R. Gimeno-Blanes, A. S. Amin, M. M. Winter, H. A. Marquering, W. E. Kok, A. H. Zwinderman, A. A. Wilde, and Y. M. Pinto, "Computer versus cardiologist: Is a machine learning algorithm able to outperform an expert in diagnosing a phospholamban p.arg14del

mutation on the electrocardiogram?," *Heart Rhythm*, vol. 18, pp. 79–87, January 2021.

[174] M. Ivanovs, R. Kadikis, and K. Ozols, "Perturbation-based methods for explaining deep neural networks: A survey," *Pattern Recognition Letters*, vol. 150, pp. 228–234, October 2021.

[175] T. Dissanayake, T. Fernando, S. Denman, S. Sridharan, H. Ghaemmaghami, and C. Fookes, "A robust interpretable deep learning classifier for heart anomaly detection without segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2162–2171, June 2021.

[176] R. Li, X. Zhang, H. Dai, B. Zhou, and Z. Wang, "Interpretability analysis of heartbeat classification based on heartbeat activity's global sequence features and BiLSTM-attention neural network," *IEEE Access*, vol. 7, pp. 109870–109883, 2019.

[177] S. Hong, C. Xiao, T. Ma, H. Li, and J. Sun, "MINA: Multilevel knowledge-guided attention for modeling electrocardiography signals," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, August 2019.

[178] Q. Yao, R. Wang, X. Fan, J. Liu, and Y. Li, "Multi-class arrhythmia detection from 12-lead varied-length ECG using attention-based time-incremental convolutional neural network," *Information Fusion*, vol. 53, pp. 174–182, January 2020.

[179] Y. Elul, A. A. Rosenberg, A. Schuster, A. M. Bronstein, and Y. 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*, vol. 118, June 2021.

[180] S. S. Mousavi, F. Afghah, A. Razi, and U. R. Acharya, "Ecgnet: Learning where to attend for detection of atrial fibrillation with deep visual attention.," *IEEE-EMBS International Conference on Biomedical and Health Informatics. IEEE-EMBS International Conference on Biomedical and Health Informatics*, vol. 2019, May 2019.

[181] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations,*

ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (Y. Bengio and Y. LeCun, eds.), 2015.

[182] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," April 2022.

[183] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, September 2021.

[184] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, ACM, March 2019.

[185] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan, "Intuitively assessing ML model reliability through example-based explanations and editing model inputs," in *27th International Conference on Intelligent User Interfaces*, ACM, March 2022.

[186] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, April 2022.

[187] R. Mochaourab, A. Venkitaraman, I. Samsten, P. Papapetrou, and C. R. Rojas, "Post hoc explainability for time series classification: Toward a signal processing perspective," *IEEE Signal Processing Magazine*, vol. 39, pp. 119–129, July 2022.

[188] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, and R. Ranganath, "Deep learning models for electrocardiograms are susceptible to adversarial attack," *Nature Medicine*, vol. 26, pp. 360–363, March 2020.

[189] I. Karlsson, J. Rebane, P. Papapetrou, and A. Gionis, "Locally and globally explainable time series tweaking," *Knowledge and Information Systems*, vol. 62, pp. 1671–1700, August 2019.

[190] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: Challenges revisited," June 2021.

[191] A. Maratea and A. Ferone, "Pitfalls of local explainability in complex black-box models," *Proceedings of WILF 2021, the 13th International Workshop on Fuzzy Logic and Applications*, vol. 3074, December 2021.

[192] C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl, "General pitfalls of model-agnostic interpretation methods for machine learning models," in *xxAI - Beyond Explainable AI*, pp. 39–68, Springer International Publishing, April 2022.

[193] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "GLocalX - from local to global explanations of black box AI models," *Artificial Intelligence*, vol. 294, p. 103457, May 2021.

[194] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making*, vol. 19, July 2019.

[195] S. Marton, S. Lüdtke, and C. Bartelt, "Explanations for neural networks by neural networks," *Applied Sciences*, vol. 12, p. 980, January 2022.

[196] S. Jia, P. Lin, Z. Li, J. Zhang, and S. Liu, "Visualizing surrogate decision trees of convolutional neural networks," *Journal of Visualization*, vol. 23, pp. 141–156, November 2019.

[197] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, "Hyperparameter optimization for machine learning models based on bayesian optimizationb," *Journal of Electronic Science and Technology*, vol. 17, pp. 26–40, March 2019.

[198] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro, "Code-test: An annotated 12-lead ecg dataset," April 2020.

[199] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, and C. Rakovski, "A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients," *Scientific Data*, vol. 7, February 2020.

[200] D. G. S. D. D. Schocken, *Marriott's Practical Electrocardiography*. LWW, December 2021.

[201] H. Kestler, M. Haschka, W. Kratz, F. Schwenker, G. Palm, V. Hombach, and M. Hoher, "De-noising of high-resolution ecg signals by combining the discrete wavelet

transform with the wiener filter," in *Computers in Cardiology 1998. Vol. 25 (Cat. No. 98CH36292)*, pp. 233–236, IEEE, September 1998.

[202] X. kui Wan, H. Wu, F. Qiao, F. cong Li, Y. Li, Y. wen Yan, and J. xin Wei, "Electrocardiogram baseline wander suppression based on the combination of morphological and wavelet transformation based filtering," *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–7, March 2019.

[203] J. P. Allam, S. Samantray, and S. Ari, "Patient-specific ECG beat classification using EMD and deep learning-based technique," in *Advanced Methods in Biomedical Signal Processing and Analysis*, pp. 87–108, Elsevier, January 2023.

[204] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C–23, pp. 90–93, January 1974.

[205] Z. C. Oleiwi, D. Al-Shammary, M. Al-Asfoor, and A. Ibaida, "Light network high performance discrete cosine transform for digital images," *Visual Informatics*, vol. 5, pp. 41–50, June 2021.

[206] H. S. Shin, C. Lee, and M. Lee, "Ideal filtering approach on DCT domain for biomedical signals: Index blocked DCT filtering method (IB-DCTFM)," *Journal of Medical Systems*, vol. 34, pp. 741–753, April 2009.

[207] H. Liu, Y. Wang, D. Chen, X. Zhang, H. Li, L. Bian, M. Shu, and D. Chen, "A large-scale multi-label 12-lead electrocardiogram database with standardized diagnostic statements," June 2022.

[208] S. P. Heinrich, "Removing mains interference from the mfERG by applying a post-processing digital notch filter: for the good or the bad?," *Documenta Ophthalmologica*, vol. 144, pp. 31–39, November 2021.

[209] E. Fotiadou, R. J. G. van Sloun, J. O. E. H. van Laar, and R. Vullings, "A dilated inception CNN-LSTM network for fetal heart rate estimation," *Physiological Measurement*, vol. 42, p. 045007, April 2021.

[210] A. Choubineh, J. Chen, F. Coenen, and F. Ma, "A quantitative insight into the role of skip connections in deep neural networks of low complexity: A case study directed at

fluid flow modeling," *Journal of Computing and Information Science in Engineering*, vol. 23, July 2022.

[211] R. Yasrab, "SRNET: A shallow skip connection based convolutional neural network design for resolving singularities," *Journal of Computer Science and Technology*, vol. 34, pp. 924–938, July 2019.

[212] A. Peimankar and S. Puthusserypady, "DENS-ECG: A deep learning approach for ECG signal delineation," *Expert Systems with Applications*, vol. 165, p. 113911, March 2021.

[213] G. Ni, X. Zhang, X. Ni, X. Cheng, and X. Meng, "A WOA-CNN-BiLSTM-based multi-feature classification prediction model for smart grid financial markets," *Frontiers in Energy Research*, vol. 11, May 2023.

[214] X. Yao, X. Li, Q. Ye, Y. Huang, Q. Cheng, and G.-Q. Zhang, "A robust deep learning approach for automatic classification of seizures against non-seizures," *Biomedical Signal Processing and Control*, vol. 64, p. 102215, February 2021.

[215] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," June 2017.

[216] E. Mohamed, K. Sirlantzis, and G. Howells, "A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation," *Displays*, vol. 73, p. 102239, July 2022.

[217] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, March 2021.

[218] B. Abibullaev, I. Dolzhikova, and A. Zollanvari, "A brute-force CNN model selection for accurate classification of sensorimotor rhythms in BCIs," *IEEE Access*, vol. 8, pp. 101014–101023, May 2020.

[219] T. Schlosser, M. Friedrich, Trixy Meyer, and D. Kowerko, "A consolidated overview of evaluation and performance metrics for machine learning and computer vision," January 2024.

[220] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, pp. 427–437, July 2009.

[221] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-label confusion matrix," *IEEE Access*, vol. 10, pp. 19083–19095, February 2022.

[222] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *Journal of Big Data*, vol. 10, April 2023.

[223] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, December 2008.

[224] M. Baygin, T. Tuncer, S. Dogan, R.-S. Tan, and U. R. Acharya, "Automated arrhythmia detection with homeomorphically irreducible tree technique using more than 10,000 individual subject ecg records," *Information Sciences*, vol. 575, pp. 323–337, October 2021.

[225] A. H. Ribeiro, M. H. Ribeiro, G. M. Paixão, D. M. Oliveira, P. R. Gomes, J. A. Canazart, M. P. Ferreira, C. R. Andersson, P. W. Macfarlane, W. Meira Jr., T. B. Schön, and A. L. P. Ribeiro, "Pre-trained deep neural network models for ecg automatic abnormality detection," January 2020.

[226] R. Kitani and S. Iwata, "Verification of interpretability of phase-resolved partial discharge using a cnn with shap," *IEEE Access*, vol. 11, pp. 4752–4762, January 2023.

[227] A. Agrawal, A. Chauhan, M. K. Shetty, G. M. P, M. D. Gupta, and A. Gupta, "Ecg-icovidnet: Interpretable ai model to identify changes in the ecg signals of post-covid subjects," *Computers in Biology and Medicine*, vol. 146, p. 105540, July 2022.

[228] M. K. Das and D. P. Zipes, *Electrocardiography of Arrhythmias: A Comprehensive Review*. ELSEVIER, 2nd ed., March 2021.