ADDIS ABABA UNIVERSITY
ADDIS ABABA INSTITUTE OF TECHNOLOGY (AAiT)
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

EXPLAINABLE RHYTHM-BASED HEART DISEASE DETECTION FROM ECG SIGNALS

BY
DEREJE DEGEFFA

ADVISOR
Dr. FITSUM ASSAMNEW

A thesis submitted to the School of Electrical and Computer Engineering in partial fulfillment of the requirements for the Degree of Master of Science in Computer Engineering

Jun, 2023
ADDIS ABABA, ETHIOPIA

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

The undersigned have examined the thesis titled:

EXPLAINABLE RHYTHM-BASED HEART DISEASE DETECTION FROM ECG SIGNALS

BY

DEREJE DEGEFFA

Approval by Boards of Examiners

Dr. Bisrat Derebssa	_____	_____
Dean, SECE, AAiT	Date	Signature
Dr. Fitsum Assamnew	_____	_____
Advisor	Date	Signature
Dr. Bisrat Derebssa	_____	_____
Internal Examiner	Date	Signature
Dr. Surafel Lemma	_____	_____
External Examiner	Date	Signature

Declaration

I, Dereje Degeffa Demissie, declare that this thesis is my original work. All sources of information in this study have been appropriately acknowledged. I further confirm that this thesis has not been submitted either in part or in full for any other requirements to any other learning institution.

Declared By:

_____

Student's Name and Signature

Approved By:

_____

Advisor's Name and Signature

Jun 2023

Acknowledgments

First and foremost, I thank **God** for the guidance, wisdom, and strength that He graciously granted me throughout my academic journey. Without His divine intervention, none of this would have been possible.

My sincere gratitude goes out to Dr. Fitsum Assamnew, my adviser, for his significant support and direction throughout my research. His insightful advice, constructive comments, and expert supervision were instrumental in shaping the direction and quality of my work. It is his passion that gave me strength, and his kindness inspires me to work hard. I am again grateful and thankful to him.

I extend my sincere gratitude to the School of Electrical and Computer Engineering at AAiT, AAU for providing me with excellent education and training. Their teachings and insights have profoundly impacted the way I approach research.

I would also like to thank everyone who generously shared their time and expertise for their invaluable comments and support in this study. Henok Banti (MD), Koricho Simie (MD), Mr. Degaga Wolde, Tariku Fikadu (MD), Mr. Yafet Philipos, and Muhammed Edris (PhD) have given me their invaluable comments and support that have enhanced the work of this research.

Finally, My family and friends also deserve my gratitude for standing by my side and offering their love, understanding, and motivation during challenging times.

Abstract

Healthcare decision support systems must function with confidence, trust, and a functional understanding. Many researches have been done to automate the identification and classification of cardiovascular conditions from Electrocardiogram (ECG) signals. One such area of research is the use of Deep Learning (DL) for classification of ECG signals. However, DL models do not provide the information on why they reached their final decision. This makes it difficult to trust their output in a medical environment. In order to resolve the trust issue, there is research being done to explain the decision the DL model has arrived at. Some approaches have been used to improve the interpretability of DL models, using the Shapley Value SHAP technique. However, SHAP's explanation happens to be computationally expensive.

In this research, we develop a deep learning model that can detect five rhythm-based heart diseases that incorporate explainability. We employ visual explainers: Grad-CAM and Grad-CAM++; as an explainability framework. These explainers are relatively lightweight and can be executed quickly on a standard CPU or GPU. Our model was trained using 12-leads ECG signals from the PTB-XL large dataset. We used 3,229 ECG records to train the model, 404 ECG records to validate it, and 403 ECG records to test it. Our model was effective, with a classification evaluation accuracy of 0.96 and an F1 of 0.88. In order to evaluate the explainability, we gave ten randomly selected outputs to two domain experts. The two experts agreed with at least 80% of the explanations given to them. In the explanations that were not completely accepted by the experts, many of the leads out of the 12 were correctly explained. Showing that the use of visual explainability like Grad-CAM++ could be useful in the diagnosis process of heart diseases. The outcomes of this evaluation suggest that our model output is, on average, on the ten sample cases, 80% correct and consistent with the evaluation of the two experts.

Keywords: Heart Disease, Rhythm-based, Explainability, ECG, Grad-CAM, Grad-CAM++

Table of Contents

List of Figures

List of Tables

List of Acronyms

Chapter 1

Introduction

Manual interpretation of the Electrocardiogram (ECG) signals takes a lot of time and expertise. The demand for diagnostic testing is increasing, which is encouraging the creation of new approaches to computerized interpretation of the ECG. Most importantly, precise detection of the type of heart problem is crucial for treatment planning to reduce the chance of serious outcomes. Using computer-aided diagnosis (CAD) systems that give explanations of the outputs can improve the facility to obtain outputs that are more accurate and reliable. In terms of addressing societal challenges, the inspiration for this work lies in automating the ECG signal analysis study for health care and assisting cardiologists in the diagnosis process.

1.1 Background

Cardiovascular Diseases (CVDs) are a class of diseases that affect the heart and its blood vessels [1]. CVD cause immense health and economic problems [2] and are also the leading cause of death in the world [3]. From 12.3 million (25.8%) mortalities in 1990 to 17.9 million (32.1%) deaths in 2015, the number of deaths worldwide due to CVD has increased. By 2030, it is projected that there would be 22 million fatalities worldwide if nothing is done [1, 4].

Therefore, an automated technique for identifying cardiac diseases is required to help doctors make a precise identification of different CVD. One of the most often used diagnostic tools for determining the existence of various cardiovascular diseases is the ECG. In order to automate the detection of cardiac diseases using ECG signals, many methods have been proposed in the literature [4, 5, 6]. However, many of them lack explanation for the classification.

1.1.1 Medical Background

Overview of Heart Anatomy

The heart, a muscular organ that is electrically stimulated, pumps blood throughout the body's organs and tissues. In this section, the tissue composition and anatomical components of the heart are described.

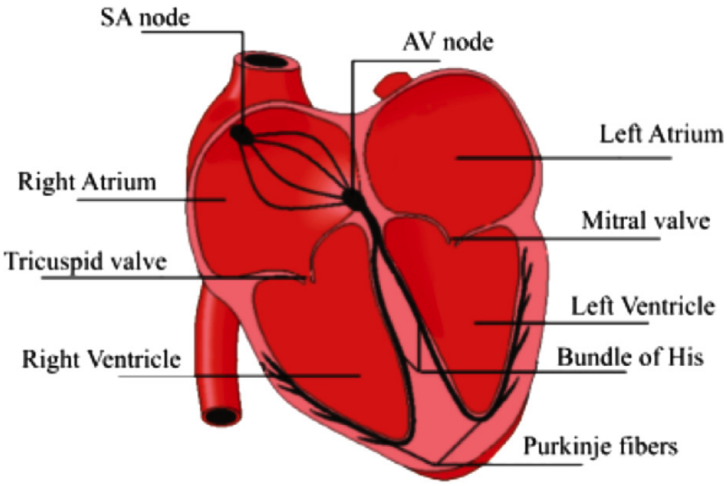


Figure 1.1: Heart Anatomy

Source: from article by A. Cheffer et al. [7]

Anatomically, the mammalian heart has two atria and two ventricles, each with two chambers, for a total of four chambers, as shown in Figure 1.1. In order to keep the heart beating properly, different parts of the organ work together. The Sinoatrial Node (SA), Atrioventricular Node (AV), and His-Purkinje Complex (HP) are the three important parts. A network made up of these components helps with heartbeat control.

A specific conduction system in the heart called the His-Purkinje complex is in charge of conveying electrical impulses from the AV node to the ventricles, enabling synchronized ventricle contraction and efficient blood pumping. The His-Purkinje complex is made up of the bundle of His, a group of specialized cardiac muscle fibers that runs through the interventricular septum from the AV node in the atrium and separates into left and right bundle branches that travel down the left and right sides of the septum, respectively. Purkinje fibers are a variety of small fibers that grow from bundle branches and extend throughout the ventricles, allowing the ventricular muscle to contract rapidly.

The normal functioning of the heart's electrical conduction system depends on the His-Purkinje complex. Defects in this system can cause a variety of heart rhythm disorders. A natural pacemaker, the SA node stimulates the atria by propagating as a wave. It initiates a pulse when it gets to the AV node, which then activates the His bundle and Purkinje fibers. The fibers spread the stimulus to the myocardial cells, resulting in the contraction of the ventricles [7].

1.1.2 Electrocardiogram (ECG)

ECG machine is a medical device that tracks the electrical activity and rhythm of the heart. The sensors affixed to the skin detect the electrical pulses generated by the heart organ each time the heart beats, and the device displays the voltage versus time electrical activity of the heart [8].

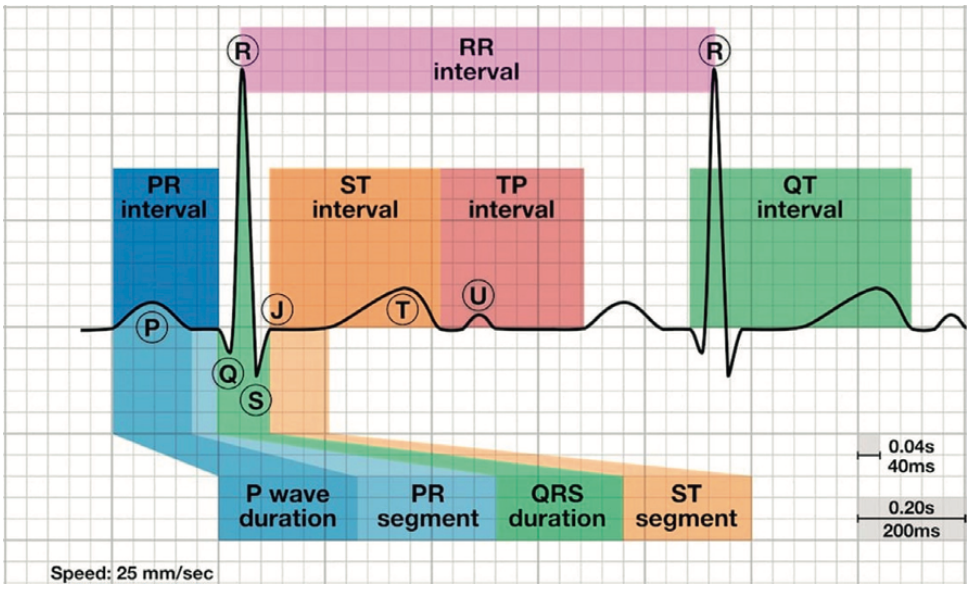


Figure 1.2: ECG intervals, waves, and segments

Source: from article by D. W. Feyisa et al. [8]

ECG devices are available in various styles and lead arrangements to meet different needs. Single-lead ECG machines are appropriate for personal usage because they are normally smaller and more portable. On the other hand, standard ECG machines use 12 leads, enabling a more complete examination of the electrical activity of the heart.

There are 12 directions on a 12-lead ECG where an electrical signal or heart impulse can be measured. As shown in Figure 1.2, the ECG signal produced by each lead has waves, intervals, segments, and one complex [8]. Each representing a specific electrical event. The impulse produced by the SA node is represented by the P wave. Ventricular contraction produces the QRS complex. Ventricular repolarization is reflected in the T wave. Heart rate variability, illustrated by the RR interval [7].

One of the most crucial phases in the diagnosis of heart diseases is the analysis of ECGsignals. Nearly all ECG analysis techniques require an understanding of the placement and shape of various waveforms (P-QRS-T) in ECG records in order to achieve good diagnosis accuracy. For instance, P-wave absence is one of the significant and clinically helpful aspects for the diagnosis of Atrial Fibrillation (AFIB), which is one of the most prevalent cardiac arrhythmias in the elderly population [7, 9].

It has been established that a healthy heart should always have a sinus rhythm, which may be identified by ventricular rates between 60 and 100 beats per minute and a consistent rate step between 0.120 and 0.20 seconds. It has been established that a Sinus Bradycardia (SBRAD) is a regular rhythm in which the ventricular rate is between 40 and 60 beats per minute with a consistent RR interval. Another was referred to as Sinus Tachycardia (STACH), a regular cardiac rhythm in which the ventricular step is quicker between 100 and 160 beats per minute. However, Sinus Arrhythmia (SARRH) has been described as an irregular rhythm with ventricular steps that typically vary from 60 to 100 [10].

1.2 Problem Statement

The main problem of deep learning models is their lack of explainability. They are complicated ”black boxes,” making it challenging to understand how they arrived at their decisions. If deep learning models are used for automatic detection of rhythmic disease of the heart from ECG signals, their decision needs to be supported by an explanation of which region of the signal was used to arrive at the decision for reliable clinical evaluation. As a result, there is a strong demand that DL models be supported by an understandable explanation to explain a specific decision without lowering classification accuracy. In this work, we will develop a DL model for the classification of rhythm-based heart disease. Then solve the black-box nature of DL by using a visual explanation method called gradient-weighted class activation mapping to weigh the regions of the ECG signals that are more important for the classification of rhythm-based heart disease. Then we will be able to answer our research question.

RQ: What is the effect of applying visual explainability using Grad-Cam and Grad-Cam++ on the reliability of the classification of rhythm-based heart disease?

1.3 Objective of the Thesis

1.3.1 General Objective

The overall objective of this study is to develop a rhythm-based heart disease classification model for 12-lead ECG signals and explain the classification using visual explanations.

1.3.2 Specific Objectives

Specific objectives include:

- To develop a model that can be used to classify heart disease from ECG signals.
- To evaluate the efficiency of the models' classification using the most common evaluation metrics
- To explain why the model made its decision using Grad-Cam and Grad-Cam++.
- To evaluate explainability using domain experts evaluations.

1.4 Contribution

In this study, we made significant contributions towards the improvement and trustworthiness of ECG signal classification using deep learning techniques. Following is a list of our contributions:

1. **Modification of the Resnet-18 model**:We adapted the Resnet-18 model architecture, which was originally designed for 2D image processing, to effectively handle 1D Electrocardiogram (ECG) signals that represent the electrical activity of the heart over time. The model was modified to take into account the unique features of ECG signals. To reduce the number of model parameters, prevent overfitting, and increase computing performance, we utilized Global Average Pooling (GAP). Each feature map of the GAP was concatenated in order to know the role of each lead in the classification of different classes.

2. **Increasing trustworthiness**: In medical applications such as ECG signal classification, obtaining the trustworthiness of predictions made by deep learning models is important. To address this, we used the Grad-CAM++ algorithm, which generates heatmaps to visualize the important regions of ECG signals. These heatmaps provided an explanation for the classification made by our model, helping to increase the explainability and clearness of the outputs.

3. **Model Explainability Evaluation**: To evaluate the performance of our DL model, we involved the aid of two domain experts. Both experts were given the same output of ten randomly selected ECG records in order to evaluate the model's classification and explainability. This evaluation was performed to validate, from the perspective of domain experts, how well our model classifies and explains ECG signals. It ensured that the classifications and explanations provided by our model were important and truthful.

1.5 Scope and Limitation

1.5.1 Scope

This study focused on developing a model that can classify only five rhythm-based heart disorders and explaining the classification that the model made and the reasons behind it.

1.5.2 Limitation

In this work, only five heart diseases related to rhythm problems were identified by our DL model. However, there are six more classes of rhythm-based heart diseases that we did not deal with.

Due to a shortage of experts, the outputs of our model were not fully validated by enough specialists; only ten samples of the outputs were evaluated and validated by two domain experts.

1.6 Methodology

The following procedures were followed in order to achieve the research objectives:

I. **Literature Review:** To understand the issue with our research work and potential solutions, a review of the literature on some of the related works about cardiac disorders was undertaken. We began our study by considering various deep learning methods used for classifying heart disease.

II. **Dataset:** The PTB-XL dataset, which consists of 21837 recordings from 18885 patients and is the largest publicly available 12-lead clinical ECG waveform dataset to date[11]. The dataset contains 71 distinct classes, of which 44 are diagnostic, 19 are form-related, and 12 are rhythm-related. The type of heart disease with abnormal rhythms is called rhythm-based heart disease. It includes: AFIB, STACH, SARRH, SBRAD, PACE, SVARR, BIGU, AFLT, SVTAC, PSVT, and TRIGU. We only considered five of them, involving AFIB, STACH, SARRH, SBRAD, and PACE.

III. **System Setup:** Data set selection is followed by the pre-processing of the collected data. To address the research question, the development of the ResNet-18 model comes next. And training has been provided for the model.

IV. **Evaluation Metrics:** Successful prediction models typically have accuracy and interpretability as their two defining characteristics [12]. To evaluate the performance of our model, standard evaluation metrics such as Accuracy, Precision, Recall and F1-Score are used. Additionally, visual explanations are also used.

1.7 Organization of the study

The article's remaining sections are organized as follows: Chapter two discusses the theoretical background and related works. This chapter presents a brief discussion on Residual Neural Network (ResNet), Class Activation Map (CAM), Gradient-Weighted Class Activation Mapping (Grad-CAM),Gradient-Weighted Class Activation Mapping++ (Grad-CAM++), and related works done in the Machine Learning (ML) and DL areas for ECG monitoring. The third chapter presents the proposed heart disease detection and explainable framework. The suggested model, together with its training algorithm and hyperparameters, are described in this chapter. The outputs from the experiments conducted using the explainable framework found in Grad-CAM and Grad-CAM++ are presented in Chapter 4, along with some sample outputs for evaluation and discussion on the outputs for performance comparison. Using generally accepted quantitative indicators, the proposed method for detecting heart disease is shown. Chapter 5 concludes the thesis and makes recommendations for future study.

Chapter 2

Background

In this chapter, we will consider the fundamentals of deep learning networks. And then focused on the specific architecture of Residual Neural Networks (ResNets). In addition to ResNets, we will also explore gradient-based visual explanations, which are used as a way to provide insight into the workings of deep neural networks. We also look into related works.

2.1 Deep learning

Deep Learning (DL) methodologies, especially reproductive models, have recently recognized a lot of attention in the field of medical image analysis [13]. It is a subsection of machine learning that employs multiple-layered artificial neural networks to learn complex data representations. DL has changed the field of artificial intelligence and made it possible for essential developments to take place in fields like speech and image recognition, robotics, and natural language processing. ResNets are a specific kind of deep neural network architecture that was formed to address the problem of vanishing gradients that can arise in extremely deep networks.

2.1.1 Residual Neural Network (ResNet)

Residual Neural Network (ResNet), a convolutional neural network that is very popular for its application in computer vision. It was the winner of the ILSVR-2015 classification competition and is yet another extremely deep model for neural network training that uses identity mapping for short connections [14]. ResNets have recently been shown to significantly improve the functionality of neural networks [15].It has been demonstrated that CNN-based architectures, which are frequently used in Deep Learning (DL), are the best at learning representations and resolving challenging computer vision and general artificial intelligence problems, such as object localization, image classification, image captioning, semantic segmentation, and visual question answering [16].

Even while Convolutional Neural Networks (CNNs) are successful at classification, training them is more challenging due to vanishing or exploding gradients. For instance, if the networks are sufficiently deep, they might not be able to learn basic operations like identity functions since a very small gradient prevents the updating of the layer weights. Whenever we stack additional layers on top of just the basics, let's say Alexnet, what happens is that a problem known as the vanishing gradient problem tends to occur. Occasionally, a neuron dies while being trained, and depending on how it is activated, it may never recover. Basically, we use the gradient to try to go back to the network and update its weights. And one of the challenges is that when we stack all these different layers on top of each other, the gradient becomes very small, and essentially, the network performance becomes extremely poor, and we are not able to train the deep convolutional neural network. The network's training of its deeper layers is exponentially slowed down by this issue [17].

The overall idea is to use a skip connection. Basically, the input is fed into the layer, and at the same time, a skip connection is used by directly passing the input to the next layer, bypassing the convolutional layer. This method helps overcome the vanishing gradient issues and permits the stacking of many layers. That is why the ResNets have been introduced as an effective solution to this problem.

ResNet also utilizes straightforward backpropagation to decrease processing costs and has a small parameter set, making it computationally effective. ResNets, a component of Convolutional Neural Networks (CNNs), add the input from the preceding layer to the output of the current layer. The network learns more easily and performs better as a result of this skip connection [15]. In residual learning, the residuals of the features train the network instead of the features themselves. ResNet is the most recent Deep CNN model that is based on the idea of residual learning. ResNet's architecture makes use of a shortcut connection that connecting the input of the n^{th} layer to some n + i layer. There are numerous residual building components that make up the ResNet architecture.

Let xi-1 serve as the input to the residual block and xi serve as the output from the block. We also consider that the output of various operations, such as convolution, batch normalization, and ReLU activation, is f(xi-1). Accordingly, we arrive at xi = f(x i-1) + x i-1. By employing this technique, it is feasible to propagate information from one layer to the next. In Figure 2.1, a fundamental ResNet building block is shown [18].

Why is it called residual?

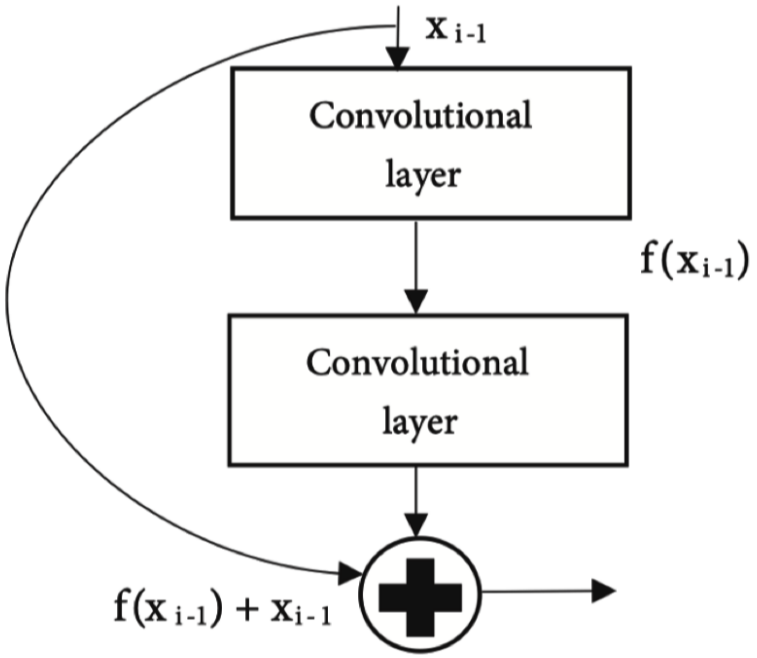


Figure 2.1: ResNet-18's structure: the fundamental building block of residual learning

Source: MD Rizvi et al. [18]

Each layer in conventional neural networks feeds information to the following layer. In a network with residual blocks, each layer feeds into the layer below it and then straight into levels that are two to three hops distant [19].

Let's consider a neural network block whose input is the value x. We want to get the true distribution H(x). Let's refer to the difference between this as:

R(x) = Output — Input = H(x) — x

By rearrangement, we obtain, H(x) = R(x) + x.

Generally, the residual block is making an effort to learn the actual output, H(x). We will find that the layers are actually attempting to learn the residual, R(x), because of the identity connection coming from x. Differently, this means that while the layers in a classical network are learning the true output (H(x)), those in a residual network are learning the residual (R(x)) [19]. As a result, Residual Block was named. In general, ResNet is a particular architecture for training deep neural networks. However, for a visual explanation of a trained neural network's decision, gradient-based localization is a well-known technique.

2.2 Gradient-based Localization for Visual Explanations

One of the common methods for understanding DL model predictions is to visualize the features that Convolutional Neural Networks (CNNs) have captured [20]. Over the last decade, CNN models have shown to be very successful at resolving difficult problems. However, due to the lack of knowledge regarding their fundamental workings, these deep models are viewed as **black box** approaches. Deep learning models that can be explained have gotten a lot of attention recently.

Explainability is a field that aims to increase the trust of users of Deep Learning (DL) models by opening their black-box and explaining how the models arrive at their predictions. Explainability algorithms help users by providing answers to questions such as Why does the model predict what it predicts [20]. Since explainability helps to know the models' internal workings and the reasons behind the models' particular decisions, explaining DL models has recently become increasingly accepted. For risk-sensitive applications like clinical decision assistance, this problem is extremely significant. One of the most popular techniques for visualizing and examining DL models is the creation of saliency maps, which highlight salient regions mostly linked to the model's decision-making [21]. The post-hoc explanation methods, such as Cam, Grad-cam, and Grad-cam++ explain the behavior of the model during the evaluation phase.

2.2.1 CAM

The goal of CAM is to explain how a model learns from data or why it performs badly on some tasks [16]. The foundation of CAMs is Global Average Pooling (GAP), in which the average value of each feature vector is obtained and immediately translated to a category label or an output node, which has been demonstrated to have significant localization capabilities. The final fully connected layer, which actually does the classification, receives the GAP operation's averaged feature maps from the last convolutional layer. The final classification score Y^c for a particular class c in a CNN with GAP can be expressed as a linear combination of its global average pooled last convolutional layer feature maps A^k [22]. Class activation maps are produced using a weighted sum of the feature maps from the last convolutional layer for each class.

$$Y^c = \sum_k W_k^c \sum_i \sum_j A_{ij}^k \tag{2.1}$$

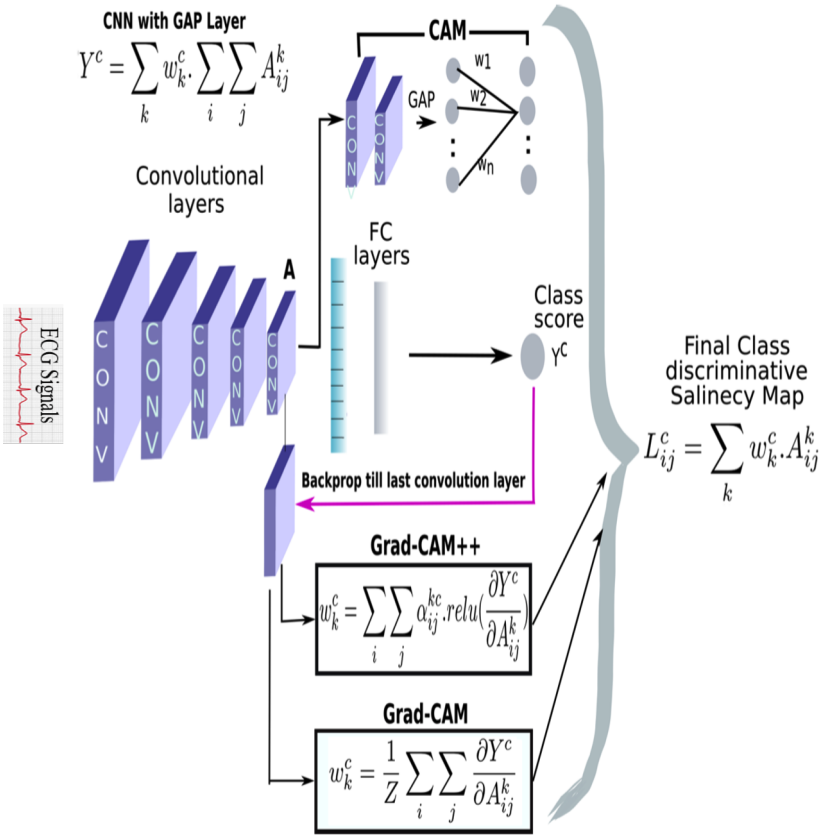


Figure 2.2: A summary of each of the three techniques: CAM, Grad-CAM, and Grad-CAM++, with their corresponding computation expressions.

Source: A. Chattopadhay et al. [22].

Each spatial position (i j) in the class-specific saliency map L^c is therefore determined as follows:

$$L_{ij}^c = \sum_k W_k^c A_{ij}^k \tag{2.2}$$

L_{ij}^{c} serves as a visual description of the class that the model predicts by directly correlating with the significance of a specific spatial location (i, j) for a specific class c. For each class c, CAM trained a linear classifier using the activation maps of the last convolutional layer to estimate the weights W_K^c. This limits its ability to explain CNNs with a GAP penultimate layer, though, and necessitates retraining several linear classifiers (one for each class) after the first model's training [22]. In Equation 2.2, W_k^c is the weight connected to feature map k and class c, and A_{ij}^k is the activation value of feature map k in the final convolutional layer at location (i, j). It makes sense that a high CAM value at position i, j would arise from the last convolutional layer's feature mappings having an average high activation value.

2.2.2 Grad-CAM

A modified form of CAM called Gradient-Weighted Class Activation Mapping (Grad-CAM) identifies the important ECG signal regions for concept prediction and takes into account both the gradients and the weights flowing into the last convolution layer [23]. In this manner, the layers that come before the last layer also contribute to the activation map. Grad-CAM visualized input regions with high resolution details that are crucial for predictions, increasing the transparency of CNN-based models. Grad-CAM can be used on any layer of the network, which is an advantage of employing gradients. However, the last one is particularly crucial for identifying the signal elements that make the biggest difference in the final prediction. Any module, not just a fully connected layer, may follow the layer used as the input for the prediction. Grad-CAM uses the parameters W_k^c, which are derived as follows and indicate the neuron significance weights:

$$W_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2.3}$$

where Z is a fixed number (number of pixels in the activation map). In Equation 2.3, $\frac{1}{Z} \sum_i \sum_j$ is used to represent the global average pooling and $\frac{\partial y^c}{\partial A_{ij}^k}$ denotes the back propagation gradients. The k^{th} feature map is represented by A^k in the gradient expression, and y^c is the class c score. Next, the following formula yields the Grad-CAM for a class c at point (x, y):

$$L_{Grad-Cam}^{c}{}^{(x,y)} = ReLUE\left(\sum_k W_k^c A^k(x,y)\right) \tag{2.4}$$

where the negative values were mapped to zero using the Rectified Linear Unit (ReLU) operator, which is an activation function if the input is positive, returns the value; else, it returns zero.We calculate the output of Grad-CAM for each class being analyzed, just like we did with CAM.

2.2.3 Grad-CAM++

A variant of Grad-CAM called Grad-CAM++ offers promising visual explanations for a specific CNN architecture across a variety of tasks that are human-interpretable [22] aims to more effectively localize multiple class instances. Grad-CAM++ uses a weighted average of the partial derivatives, which is different from Grad-CAM. Given a class c with a score of Y^c and the activation map A_{ij}^k calculated in the last convolutional layer, parameter α_{ij}^{kc} can be found as follows [24]:

$$\alpha_{ij}^{kc} = \frac{\frac{\alpha^2 y^c}{(\alpha A i j^k)^2}}{2\alpha^2 y^c \left(\alpha A_{ij}^k\right)^2 + \sum_a \sum_b A_{ab}^k \left(\frac{\alpha^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right)} \tag{2.5}$$

The definition of the parameter w_k^c, which has the same function as w_k^c in Grad-CAM, is as follows:

$$W_k^c = \sum_i \sum_j \alpha_{ij}^{kc} ReLUE\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{2.6}$$

Equation 2.6 brings us to

$$W_k^c = \sum_{ij} [\frac{\frac{\alpha^2 y^c}{(\alpha A i j^k)^2}}{2\alpha^2 y^c \left(\alpha A_{ij}^k\right)^2 + \sum_a \sum_b A_{ab}^k \left(\frac{\alpha^3 Y^c}{\left(\partial A_{ij}^k\right)^3}\right)}] ReLUE\left(\frac{\partial y^c}{\partial A_{ij}^k}\right) \tag{2.7}$$

It holds that, similar to other CAMs.

$$M_{Grad-Cam++}^{c\ (x,y)} = ReLUE\left(\sum_k w_k^c A^k(x,y)\right) \tag{2.8}$$

From Figure 2.2, we can sum it up that: Grad-cam estimates the weight by dividing Z, where Z is a constant (the number of pixels in the activation map). Grad-cam++ calculate the weight by a more sophisticated backpropagation. From equations 2.3 and 2.6, We can see that if α_{ij}^{kc} = 1/Z, Grad-CAM++ reduces to the formulation for grad-cam.Thus, Grad-Cam++ is a generalized formulation of Grad-Cam.

2.3 Hyperparameters for the deep learning model

The process of training a deep learning model needs iteratively adjusting its parameters to minimize the difference between the predicted outputs and the actual outputs. This optimization process aims to enhance the model's ability to accurately capture and generalize patterns in the data, ultimately improving its overall performance.

Activation Function

In deep neural networks, nonlinearity is introduced through the activation function [25]. It specifies a property of activated neurons that can be stored and mapped out by a non-linear function and used to address non-linear problems. The activation function is utilized to improve the neural network model's capacity for expression, which can give the neural network the appearance of artificial intelligence [26]. Some of the activation functions are:

ReLU (Rectified Linear Unit) function: Relu activation function is one of the major common neuron activation functions, and it is for deep network training. The vanishing gradient problem has been overcome by the function. For non-negative inputs, the ReLU is an identity function, while for negative inputs, it is a zero function [27].

Sigmoid function: The output of the sigmoid function, which may be understood as the likelihood that a given input belongs to one of two classes, is frequently employed in machine learning algorithms for tasks like binary classification.

Epochs

Epoch refers to a collection of samples that are passed through the training data set. In a single epoch, the algorithm traverses the whole dataset, updating the model in response to any faults it discovers, and then resets at the beginning of the dataset for the subsequent epoch. Increase the number of epochs until the testing accuracy starts to decline, even while the accuracy of the training cases is improving (overfitting). In order to calculate the weight update for each input sample, these values must be stored during an epoch, which is one trip over the training set. All contributions are added at the end of the epoch, and only the weights are updated with the composite value. In order to more closely follow the gradient, this approach adapts the weights with a cumulative weight update. The basic idea behind a training case is to send training samples to a neural network as input vectors.

Loss Function

The goal of testing and training various model parameters is to improve performance according to performance evaluation metrics while reducing the proportion of the loss function.

Learning Rate

The learning rate is a crucial training process parameter. A hyperparameter effectively and efficiently adjusts the step size during the training case, accelerating the training process. However, selecting the value of the learning rate hyperparameter is sensitive. The local minimum may be repeatedly overstepped if the chosen learning rate is too high, leading to oscillations and a very sluggish convergence to the lower error rate situation.If the chosen learning rate value is too small, the needed number of iterations may be more, which would lead to poor performance.

Optimizer

The goal of optimization is typically to reduce the loss function by updating the weights. Loss reduction in deep networks is accomplished by altering the network's hyperparameter configuration. The Loss function seems to point the gradient optimizer in the appropriate direction as it moves toward the global minimum. Several optimizers, such as Stochastic Gradient Descent (SGD), Root Mean Squared Propagation (RmsProp), and Adam optimizer, are utilized for gradient descent optimization techniques in the Deep Learning era. SGD, a stochastic estimate of gradient descent optimization that iteratively optimizes an objective function. It reduces the update time for handling massive amounts of data and removes a certain amount or number of computational redundancy.

Another optimization method is RmsProp, which adjusts the learning rate for each parameter. It limits oscillations in the vertical direction and splits the learning rate using the weighted running average of recent gradient magnitudes. One of the most important and effective optimization algorithms in the framework of the Deep Learning technique is Adam. It has benefits to need a little hyperparameter tuning for the learning rate. This optimizer also requires little memory for the network, making it simple to implement and computationally efficient.

Dropout

A dropout map is the most effective regularization method, and applied to increase the performance by disabling specific number of neurons in each layer randomly during training. In other ways, the dropout method can be used to reduce the impact of each layer's individual neuron number, which aids in the network's ability to generalize effectively and improves the accuracy of the outcome. A dropout network with the same size as the specific number of neurons in the previous layer is initialized randomly to mark the off or on state of the network structure's corresponding neuron at the beginning of each iterating train case.

2.4 Related works

The Electrocardiogram (ECG) was initially interpreted using computers in the 1950s, when an analog signal was converted to digital [28]. Because interpretative algorithms were added to portable ECG in the years that followed, and computer aided interpretations of the ECG became common. AI has recently been utilized in the study of ECG signal analysis. Numerous Machine Learning (ML) techniques, including in particular Deep Learning (DL) approaches, have shown good performance in detecting abnormal ECG waveforms and events, improving the detection accuracy of a number of heart-related disorders. Considering a one-dimensional (1D) data representation of the ECG signal and processing it using the same procedures as we would for a standard text is one workable data processing technique [3].

2.4.1 Models for detecting heart disease using deep learning

A Deep Learning (DL) model is used to represent the rules, which also incorporate any additional algorithm-specific information required to create a prediction, which describes what a machine learning algorithm has learned. An Artificial Neural Network (ANN) that can learn tasks from examples is known as a Deep Neural Network (DNN), which has a number of layers between the output and input layers. Although there are numerous different kinds of neural networks, neurons form a common element in all of them.

DNNs, which are further separated into long-term short-term memory (LSTM), recursive neural networks (RNNs), and convolutional neural networks (CNNs), are the foundation of DL models. Among these, CNNs are frequently employed in a variety of fields. The local receptive field permits for access to underlying features by neurons or processing units like directional edges or corners, which account for some of a CNN's complex structure's translation, scaling, and rotation in variance. This is one of the supreme important characteristics of a CNN. As a result, the CNN-based approach exhibits excellent performance in the classification of ECG signals. This is because of its high robustness and noise tolerance [3].

Convolutional neural networks (CNN)

The three steps of feature extraction, feature selection, and classification constitute the foundation of many algorithms. The alternative is to use machine learning techniques, for instance convolutional neural networks (CNN), which accept the raw (or pre-processed) information as input and do not require previously established features[6]. For signal analysis, natural language processing, signal analysis, and image classification, CNNs are a common class of DNNs. Since these networks can automatically extract hierarchical patterns from input using stacked trainable small filters or kernels, they require far less pre-processing than handcrafted features. [5] states that a CNN typically includes many convolutional layers, followed by layers for classification (such as fully connected layers), batch normalization, nonlinear activation, dropout, and pooling. SVMs, boosting classifier trees, and RNNs have all been used as replacements for fully connected layers in CNNs to summarize global features. Based on shared-weight designs and parallelization, CNNs can produce improved outputs and fast computation [29].

Recurrent Neural Network (RNN)

RNN structures are sequential, whereas CNN architectures are typically hierarchical[30]. A particular kind of neural network called an RNN is designed for modeling sequential data, such as time series, event sequences, and natural language. The outputs of the prior step are utilized as the current phase's input in an RNN. An RNN is able to remember information in sequential order by iteratively updating hidden states and memory. RNNs are an obvious choice for processing inputs of varying lengths and capturing temporal dependencies, especially for ECG data.

Convolutional recurrent neural network (CRNN)

A CRNN, as the name suggests, involves CNN (convolutional neural network) followed by the RNN (Recurrent neural networks).CRNNs utilize recurrent neural networks to summarize the extracted features over time and convolutional neural networks to extract local features [31]. Long ECG signals with variable sequence lengths and multichannel inputs are best handled with this design.

A 1D CNN or 2D CNN is utilized to extract local features from an ECG sequence. An RNN then summarizes local features along the time dimension to generate global features. To provide interpretable diagnosis, there is a need to incorporates a CRNN with a multi-level attention mechanism with beat-level, rhythm-level, and frequency-level expert features based on medical domain knowledge [29]. A CRNN model can perform better on the task for which it was trained when trained on a big dataset. In general because of how DL models work, a large dataset of good quality is required to fully understand the complex patterns and correlations present in the data.

2.4.2 Deep learning models' explainability

DL models are normally viewed as black boxes, despite the fact that they can perform at the state-of-the-art in many predictive tasks. Deep learning models frequently include many model parameters or complex model structures, making it challenging for a person to grasp their conclusions. As a result of the multi-layer nonlinear structure, deep learning models' decisions are not traceable by humans. Nevertheless, for many applications, understanding the behavior of the model when making predictions is just as important as their accuracy. Because medical specialists do not accept diagnoses without an explanation, this task is significantly more difficult in the medical field [29, 32].

A few researches have done with the goal of improving how interpretable deep learning approaches are for ECG data. In certain research, interpretable expert features that can be used for partial interpretation have been explicitly incorporated in deep learning models. On the basis of raw ECG data, other researchers have created salient maps using multi-level attention weights or attribute scores.There are two worthwhile research directions regarding interpretability.

The first is how to understand complex deep learning models using very straightforward models. For instance, one might first build a black-box deep learning model for a particular job, then build a separate explainable simple model that corresponds to the deep learning model's predictions, and finally interpret the prediction outputs using the simple model.

The second is a direct method for building an understandable deep model. For instance, since attention mechanisms can be more easily understood by humans, they can be added to hidden layers of deep model architecture instead of neuron connection concepts that are borrowed from tree-based models [12].

According to S.Somani et al. [33] a deep CNN trained from AliveCor ECG data, was deployed on a single-lead recorder system (KardiaBand, Apple Watch) to continuously monitor for AF in 24 patients. The model performed well, which was encouraging when compared to annotated reports from an insertable cardiac monitor (ICM) (episode sensitivity 97.5 percent and duration sensitivity 97.7 percent) on 24 patients, highlighting Deep Learning's usefulness in creating an inexpensive, non-invasive approach to AF surveillance and management. However, the single-lead record system cannot be used to identify many other types of heart disease.

D. Zhang et al. [34] built a deep neural network to classify heart arrhythmias automatically from 12-lead ECG data. Their model was able to perform multi-label classification of ECG signals into the following heart diseases: SNR, AF, IAVB, RBBB, LBBB, PAC, PVC, STD, and STE. An approach called 10-fold cross-validation was utilized for model training and evaluation. From the CPSC 2018 dataset, ten folds were generated at random. Eight of the available ten folds in each round were utilized for training, one for validation, and one for testing. The average F1 score and average area under the receiver operating characteristic curve (AUC) for their suggested model were both 0.813 and 0.970, respectively. They also applied the Shapley Additive Explanations (SHAP) method to understand the predictions made by the model at the patient and population levels. However, Post-hoc model explanation techniques based on SHAP are expensive due to their computational complexity [35]. The runtime is ultimately exponential in terms of the number of variables to analyze [36].

V. Jahmunah et al. [37] developed two deep learning models, DenseNet and CNN, to diagnose myocardial infarction (MI) based on ECG signals. Both models achieved good classification outputs with high accuracy rates of 98.9% and 98.5%, respectively, using a method called cross-validation to test their performance on data sets split into training and validation groups. In order to identify the exact ECG leads and wave portions that were most important in the models' predictions, they also employed a class activation mapping technique known as Grad-CAM. However, their model is limited to myocardial infarction (MI) detection.

H.Lee and M.Shin [38] used the PhysioNet/CinC Challenge 2017 dataset and proposed a novel image-based deep learning framework called BIT-CNN for automatic arrhythmia classification from short single-lead ECGs, receiving an overall F1 score of 81.75%. The proposed approach transforms variable-length 1D ECG signals into fixed-size 2D time-morphology representations and feeds them to the BIT-CNN model. The BIT-CNN model extracts a variety of more distinctive patterns from compressed sensing ECG morphology (CS-ECM). The proposed approach allows feature embedding vectors to provide interpretable time-morphology patterns focused on each step of the learning process. However, a single lead contains limited information compared to 12 leads.

2.5 Summary

ResNet, a kind of CNN, uses skip connections to overcome the vanishing gradient problem and simplify training for deeper layers. Its architecture is built on the concept of residual learning, which trains the model using the residuals of the features. The layers in a residual network are learning the residual. ResNet has a low parameter and is computationally effective, lowering processing costs by using simple backpropagation.

We considered the importance of building explainable DL models. One of the methods for getting explainability is to use visual explanations that highlight regions that are important for the model's decision-making. Some of the common methods for creating visual explanations involve the Class Activation Map (CAM) and its updated versions, Grad-CAM and Grad-CAM++. We also discussed the different training parameters for neural networks. Parameters are adjusted to reduce the difference between the expected and actual outputs of the model. The model becomes non-linear due to activation functions.

Epochs denote the full model iterations of the training dataset. Loss functions and optimizers are used to decrease the loss function and update the weights of the model. The training process is affected by the hyperparameters learning rate and optimizer, so selecting the suitable value is necessary. Dropout is a regulation method that arbitrarily disables some neurons in each layer during training, thus improving model performance.

In the section on related work, we cover some significant research on heart disease detection. The purpose of this study does not allow for a comprehensive presentation of all the suggested methods. This is because the most active area of research, automating heart disease detection from ECG data, has seen an increase in the number of articles published.

Chapter 3

Methodology

In this chapter, the proposed methodology for classification and visual explanation of ECG signals is discussed. The proposed methodology includes dataset selection, preprocessing of the data, model development and training, and an explainability framework. We have thoroughly discussed dataset selection, preprocessing, and the classification procedure used in this study.

3.1 Dataset

The PTB-XL dataset, the biggest 12-lead clinical ECG waveform dataset that is currently publicly available [11], is used for training, validating, and testing our proposed model. The waveform data in the PTB-XL dataset was gathered between 1989 and 1996, over a period of seven years. The dataset contains, 21837 clinical 12-lead ECG records of 18885 patients, each of which was recorded for 10 seconds. The dataset includes individuals in the complete age range of 0 to 95 years and is gender-balanced, with 52% men and 48% women. For all signals, the standard 12-lead configuration (lead I, lead II, lead III, lead aVL, lead aVR, lead aVF, lead V1-V6) with reference electrodes on the right arm is employed [11]. The dataset includes 71 different classes. The 71 classes consist of 44 diagnostic classes, 12 rhythm classes, and 19 form classes. The rhythm classes are given in Table 3.1.

Table 3.1: Overview of the Rhythm Statement.

	# Records	Description
SR	16782	sinus rhythm
AFIB	1514	atrial fibrillation
STACH	826	sinus tachycardia
SARRH	772	sinus arrhythmia
SBRAD	637	sinus bradycardia
PACE	296	normal functioning artificial pacemaker
SVARR	157	supraventricular arrhythmia
BIGU	82	bigeminal pattern (unknown origin, SV or Ventricular)
AFLT	73	atrial flutter
SVTAC	27	supraventricular tachycardia
PSVT	24	paroxysmal supraventricular tachycardia
TRIGU	20	trigeminal pattern (unknown origin, SV or Ventricular)

Source: [11]

3.2 Data preparation

The selection of suitable records among the huge open-source dataset is the most important stage in developing a deep learning model. We select the data from the PTB-XL 12-lead ECG waveform dataset. We only take into account the classes with sufficient records and those that are acceptable for our deep learning model, which is rhythm-related heart disease. We selected PTB-XL because it addressed the issues of a lack of a public data set and well-defined benchmarking protocols, both of which were obstacles to the creation of automatic ECG interpretation algorithms. PTB-XL is the largest publicly available clinical 12-lead ECG waveform data set to date [11]. The data is provided in CSV format, from which we selected the five rhythm-based ECG records. To prepare the acquired data for deep learning, we took the following steps, as shown in Figure 3.1.

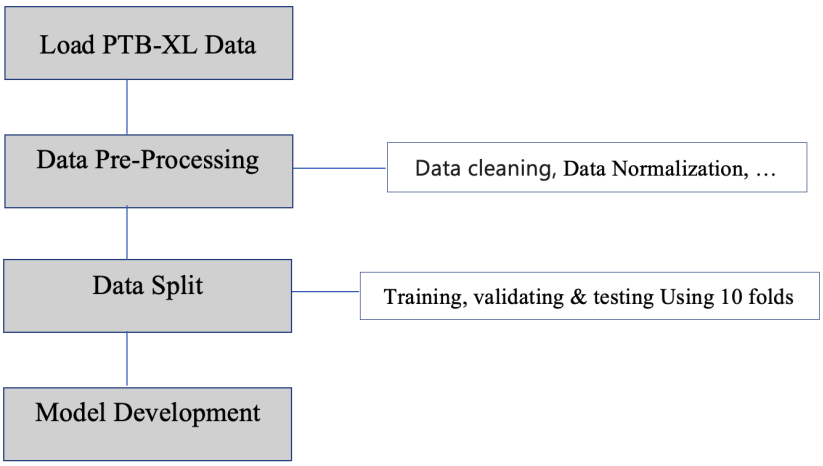


Figure 3.1: Methodology for preparing data.

3.2.1 Data pre-processing

Data preprocessing is a subset of data preparation. Before beginning to use the raw data in a deep learning (DL) environment, we arrange and format it as part of the data preprocessing procedure. This involves processes like **cleaning**, **normalizing**, converting, and integrating the data to make it more usable.

The PTB-XL dataset is divided into four major categories: diagnosis, form, and rhythm. Under the **data cleaning** practice, we filter the data and look for data that is rhythm-based and relevant to our needs. There were two separate sampling frequencies used in the PTB-XL 12-lead ECG waveform dataset. 100 Hz and 500 Hz. Every record will have 1000 discrete values for 100 Hz and 5000 discrete values for 500 Hz, provided that a 10-second recording of the signal was made. we utilized a sampling frequency of 100 Hz. We discovered records that are improper because some classes contain extremely few records that are insufficient to build a deep learning model, as can be seen from Table 3.1. We deleted inappropriate data from a dataset because inaccurate outputs would be produced if the data was not clean. We cleaned the data by removing classes from the rhythm category that were irrelevant to our task or topic. We removed insufficient records and experimented with relevant classes with at least 296 records, excluding Sinus Rhythm (SR), which is normal sinus rhythm. As a result, the five rhythms identified from the PTB-XL dataset were AFIB, STACH, SARRH, SBRAD, and PACE.

Atrial fibrillation(AFIB)

Atrial Fibrillation (AFIB) is the most prevalent clinically persistent cardiac arrhythmia condition, affecting 1% of the general population. A heart arrhythmia known as atrial fibrillation is highly related to other heart diseases that can raise the risk of heart failure and stroke [39]. Numerous risk factors, including hypertension, diabetes, coronary artery disease, obesity, cardiomyopathy, hyperthyroidism, and excessive alcohol use [39], increase the probability of developing it, all of which can reduce quality of life and raise mortality risk. AFIB significantly increases the risk of strokes, transient ischemic attacks (TIA), and heart failure and is strongly connected to other heart problems.

One of the most important areas in bioinformatics and the medical sciences is automatically classifying, particularly for the identification of atrial fibrillation. Currently, the diagnosis of AFIB is mostly determined by the presence of a few classic symptoms (such as breathing issues, chest pain, an irregular pulse, and so on) and the ECG's characteristics. However, it is challenging to correctly identify AFIB in the early stages, primarily because there are occasionally no evident symptoms when AFIB develops, and well-trained professional physicians are required to correctly analyze the feature information of the ECG [40]. The most typical arrhythmia identified in clinical practice is AFIB. It causes the atria to beat incoherently and inconsistently with the ventricles and is characterized by a rapid, irregular, and heterogeneous electrical activity of the heart due to inadequate atrial contractions.

AFIB can cause heart failure, subsequent **thromboembolic** events (the formation of blood clots), and even ischemic strokes in the general population. It is also directly linked to higher mortality rates. It primarily affects male patients, and its prevalence rises with advancing age [39, 40]. Its prevalence in clinical practice explains why many nations, notably those in the western world, allocate substantial resources to managing it. According to current studies, between 2050 and 2060, 6 to 12 million Americans and 18 million Europeans, respectively, would develop AFIB [39].

PACE

An electrical shock that causes the heart to contract is called a pace in an ECG. When someone has a slow or irregular heartbeat, it is utilized to get their ECG back to normal rhythm. A sharp spike can be seen on the patient's ECG trace when the pace is present, showing that the heart muscle has contracted due to electricity.

Sinus arrhythmia (SARRH)

Sinus Arrhythmia (SARRH) is defined by the variance of the R-R interval [41]. It is one typical example of a sinus rhythm deviation. Sinus arrhythmia commonly presents as a rate that is irregular and in which the R-R interval varies by greater than 0.12 seconds [42]. Furthermore, atrial activation coming from the sinus node is compatible with the pattern of P waves. Irregular vagus nerve initiation during breathing causes beat-to-beat variations in the resting heart rate [42]. Research have tried to show a higher frequency in people with underlying diabetes, obesity, and hypertension [43]. When a patient has sinus arrhythmia, symptoms are uncommon to appear. Shortness of breath, edema in the lower extremities, dyspnea with exertion, and peripheral neuropathy are examples of symptoms that, if they exist, are probably not caused by sinus arrhythmia but rather have an underlying cause. In older people, sinus arrhythmia seemed to be less common [44].

sinus tachycardia(STACH)

A clinical syndrome known as Sinus Tachycardia (STACH) is characterized by a consistently elevated resting heart rate (HR) and a further excessive increase with little physiological stimulation. According to consensus among experts, STACH is characterized by a resting sinus heart rate of 100 or more beats per minute, an average heart rate of more than 90 beats per minute during the entire day, and uncomfortable symptoms brought on by the persistent tachycardia. One percent of people in their middle years, primarily women, have STACH. Rarely do elderly individuals also have STACH. Palpitations, chest pain, exhaustion, shortness of breath, presyncope, and syncope are STACH symptoms [45].

Sinus bradycardia (SBRAD)

Sinus bradycardia can be a normal finding in healthy individuals, as it is responsible for the physiological slowing of heartbeats during sleep. However, it can also occur as a pathological response to other conditions [46].The symptomatic bradycardia that will lead to serious, potentially fatal consequences is what people are most concerned about.

On an Electrocardiogram (ECG), sinus bradycardia is characterized as bradycardia with a regular sinus rhythm. On an ECG, when the interval between each R wave is relatively constant, the heart rate is considered to be regular. In other words, the sinus node is firing impulses at a slower rate than usual. Normally, the sinus node beats between 60 and 100 times per minute. Heartbeats per minute that are less than 60 are referred to as bradycardia in people who are not well-trained athletes. Studies on the general population, however, frequently use a lower cutoff of 50 bpm [46, 47].

In the activity of **data normalization,** we made an effort to improve data values and organize data entries so that they appear uniformly structured through all fields and records, making it simpler to recognize, classify, and analyze information. In the data normalization, scaling numeric data, translating categorical data into numeric form, and encoding categorical labels are included. These strategies help to ensure smooth training for the model. Data normalization is an essential technique used in deep learning, as it helps improve the model's performance and accuracy. By preprocessing the data, we made it possible for DL models to operate with more precise and consistent data, which ultimately improves performance.

3.2.2 Split Dataset

In order to prepare the data for training, validation, and testing reasons, we utilized a method called "10-fold data splitting." This methodology involves splitting the available dataset into ten subgroups, or folds, with each subgroup being of equal size. In our specific case, we had approximately 4030 ECG signal records, which were divided into ten folds, each containing approximately 403 records. The assignment of records to the folds was done randomly during runtime, ensuring a fair distribution of data across the folds.

3.3 Model Development

After data preparation, which includes pre-processing the selected dataset, the next task is to develop and train the model. The diagram for the proposed methodology is shown in Figure 3.2. For the purpose of creating an explainable framework for the ECG classification task, we used 12 Resnet-18 with a bit of modification to show the role of each lead in the classification of the ECG signals.

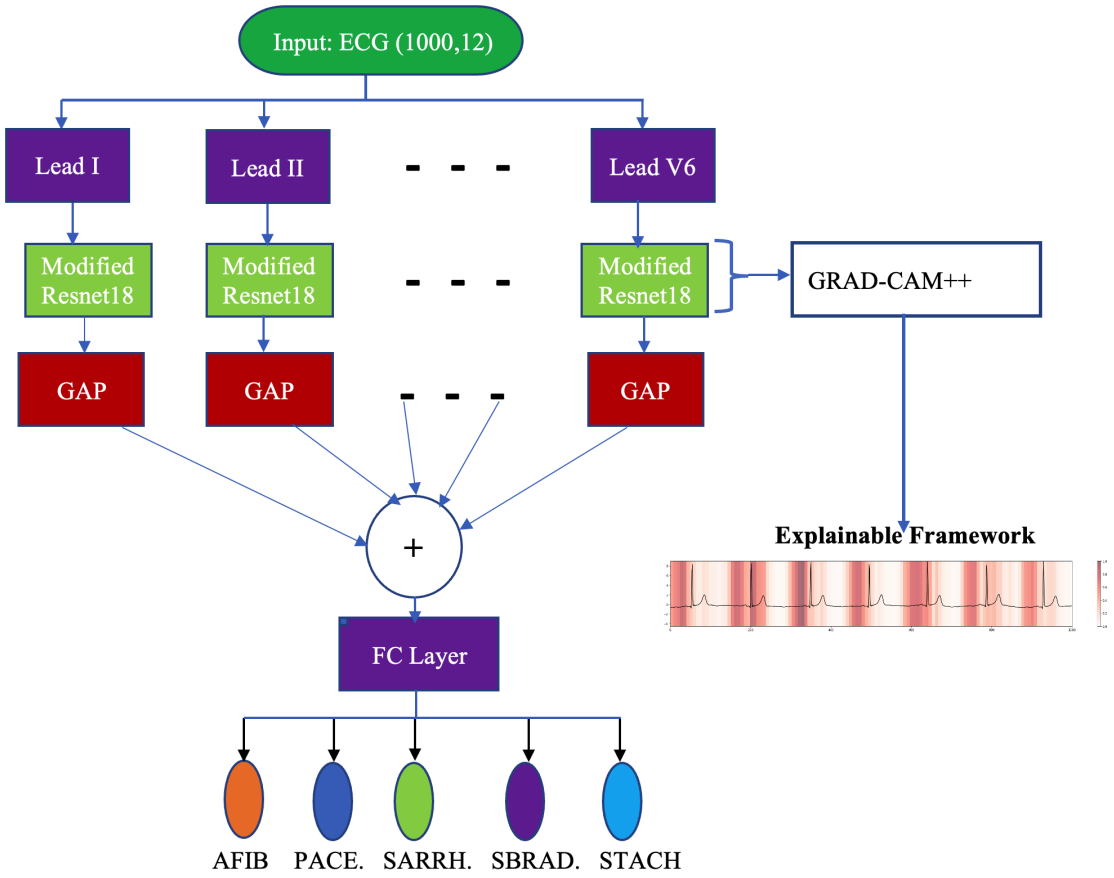


Figure 3.2: Diagram of our proposed method.

Based on the Residual Network architecture, we developed a deep learning model using the Resnet-18 convolutional neural network model for multi-class classification. The Resnet-18 architecture has been modified to work on 1D signals. The Resnet-18 architecture for 1D time series signals comprises several key components, including the identity blocks, convolutional blocks. We used the identity blocks and convolutional blocks that define a single block of the Resnet-18 for 1D time series data to implement the two types of residual blocks used in the ResNets architecture. The Resnet-18 1D function stacks multiple residual blocks to create a deep network suitable for processing time series ECG signal. The model takes an input of shape (1000, 12), where 1000 represents samples of a continuous signal from a sampling frequency of 100 Hz and a sampling time of 10 seconds, and 12 represents the number of channels.

The Resnet-18 architecture is a deep convolutional neural network with residual connections. The residual connections permit the gradients to flow directly through the network, facilitating the training of deeper networks. We build the Resnet-18 architecture for 1D signals using the Resnet-18 1D function, which employs the two residual blocks. We create the complete Resnet-18 1D CNN model through calling the Resnet-18 1D function multiple times to produce a sub-model for each input data channel (12 channels in this scenario). For each modified Resnet-18 model, we used GAP. This approach reduces the amount of network parameters, which prevents overfitting and boosts the model's computational efficiency. We concatenate the GAP of each feature map to know what contribution each lead makes to the classification of the class. The Fully Connected (FC) layer takes the concatenated feature vector generated by the GAP layer for classification.

In this research methodology, we also used the Grad-CAM++ algorithm as an explainable framework to generate heatmaps for visualizing the important regions of the ECG signals for classification by our model. The algorithm involves the following steps: (1) getting feature mappings from the final convolutional layer; (2) calculating the gradient of the predicted class with respect to the feature maps using backpropagation; (3) computing the weights of each feature map by averaging the gradient values; (4) determining a weighted sum of the feature maps to obtain a heatmap; (5) applying a ReLU activation function to the heatmap; and (6) normalizing the heatmap to enable visualization. The setup for hyperparameters and evaluation metrics is presented below.

3.3.1 Experimental parameters Setup

Achieving a good experimental setup for a DL model is essential to avoid under- or overfitting issues and find optimal outputs for the given problem. Key parameters are required to optimize the performance of a deep learning model; these include, those shown in Table 3.2. For the deep learning model to perform at its best, the hyperparameter variables that control the training process must be adjusted. To determine the optimal combination of hyperparameters that reduces loss or error metrics, hyperparameter selection, is an iterative procedure that requires trial and error. In general, there are common evaluation metrics used to evaluate the performance of deep learning models.

Table 3.2: Used parameters

Parameter Setup	
activation function	Relu
learning rate	0.001
optimizer	adam
dropout	0.2
loss function	categorical_crossentropy
number of epochs	60
batch size	32

3.3.2 Evaluation Metrics for the classification

Using evaluation metrics, the quality of deep learning models is evaluated. In any study, evaluating deep learning methods is a crucial task. A variety of assessment criteria, such as a confusion matrix, classification accuracy, the F1-score, and the model's explainability, can be used to evaluate a model. When we use the word accuracy, we typically refer to classification accuracy, which is the ratio of the number of correctly predicted events to the total number of input samples. A confusion matrix is a key tool for evaluating a deep learning (DL) model's performance. Using this metric, a model's capacity to predict the appropriate class for each sample of test data is evaluated. The confusion matrix includes true positive, false positive, true negative, and false negative values. These four metrics can be used to assess how accurately a DL model can categorize various classes in a given dataset.

Classification Accuracy

Classification accuracy in deep learning refers to how exactly a model can identify the correct class given a set of data. We employ it to measure our model's accuracy by looking

at the percentage of inputs in a given test set that it correctly classifies.

$$Accuracy = \frac{rightclassified * 100}{Allinputsamples} = \frac{(TP + TN)}{(TP + FP + TN + FN)} * 100 \tag{3.1}$$

Where; True Positives (TP) are a measure of how accurately a model identifies positive examples from test data. It indicates how well the model performs when tested on previously unseen data, classifying it as belonging to the class it was supposed to. False Positives (FP) are when the model incorrectly classifies a negative sample as being positive. That is, the model predicts that a given case belongs to the positive class, even though it actually belongs to the negative class. True Positives (TP) and True Negatives (TN) show values that are successfully classified, while False Positives (FP) and False Negatives (FN) represent instances that were misclassified.

Precision

Precision, also known as positive predictiveness, was used to measure the model's capacity to separate true positives (T p) from all of its positive predictions (T p + F p). It was useful to know the proportion of positive predictions from our model that were actually accurate.

$$Precision = \frac{TP}{(FP + TP)} \tag{3.2}$$

Recall

Recall is a metric that expresses how well a deep learning model can distinguish between significant data points among all other possible data points. We used this metric to know the proportion of correctly identified positive results in our model compared to all relevant cases in the relevant dataset. A high degree of recall means that the majority of important data points were accurately identified. Recall is also sometimes referred to as true positive rate sensitivity.

$$Recall = \frac{TP}{(FN + TP)} \tag{3.3}$$

F1−Score

The F1-Score, a statistical indicator, tells us how our model reliable. It is referred to as the harmonic mean of precision and recall. Good precision and great recall are both indicated by a high score. Using the following equation, the F1−Score of the detection technique is determined.

$$F1 - Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} \tag{3.4}$$

Confusion Matrix

To explain how effectively a classification of our model performs, we used a confusion matrix, which is a two-dimensional table containing real and predicted values.

Table 3.3: Confusion Matrix

		Actual	
		1	0
Predicted	1	True Positive	False Positive
	0	False Negative	True Negative

Using a set of test data for which the true values are known, we utilized it to measure how well a classification model performed. It allows the visualization of the performance of the model. The confusion matrix has four distinct categories: TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative) to enable visualization of the performance of an algorithm, as shown in Table 3.3.

Where TP: Actually True, which the model predicted as True.
TN: Actual False, which the model predicted as False.
FP: Actual False, which the model predicted as True.
FN: Actual True, which the model predicted as False.

By comparing appropriately classified instances to incorrectly classified samples, it can also be referred to as an error matrix and specifically gives information about the performance of classification tasks. Confusion Matrix shows outcomes not only in numerical form but also graphically, making them easier to analyze than other metrics like accuracy score. For these reasons, it is commonly used in ML and DL evaluation.

Evaluation of Model Explainability

Explainability is an emerging field, that emphasis on making models more transparent and explainable to humans. The explainability technique aims to identify features important for decision-making and helps to build trust and make sure a model is trustworthy. It has aspects of model evaluation and explanation approaches, helping to provide the information decision makers need while justifying the classification outputs of deep learning models.

Explainability helps people who may not possess technical skills for deep learning in their everyday lives but still be able to understand quality levels resulting from automated processes. Different needs and goals may require different evaluation metrics for evaluating explainability.

We used domain expertise feedback as an evaluation metric, as it was considered the most relevant to the specific explainability goals in our study.

Domain expertise evaluation: Our model's capability to generate explanations that are consistent with human knowledge and understanding is evaluated using feedback from domain experts.

3.4 Summary

The proposed classification approaches and the visual explanation of the ECG signals classification were covered in this chapter. For our proposed model, we used the PTB-XL dataset, the biggest clinical ECG waveform dataset that is publicly reachable. All ages and genders are included in the dataset, which consists of, 21837 clinical 12-lead ECG records from 18885 patients.

We selected only rhythm-based ECG records relevant to our needs and cleaned the data by deleting irrelevant classes and inadequate records. We also performed data normalization to improve the model's performance and accuracy. We used the 10-fold data splitting technique to arrange the data for training, validating, and testing, with around 4030 ECG signal records split into 10 folders of around 403 records each. We discussed the model development process in DL for ECG classification. We highlighted the importance of putting up the right hyperparameters to optimize the performance of the model and emphasizing the use of evaluation metrics such as classification accuracy, precision, recall, and F1-score to evaluate the quality of the DL model. In this chapter, we also presented the concept of explainability in DL model and its function in making models more trustworthy.

Chapter 4

Result and Discussion

This chapter describes the outcomes we found for the ECG signals classification task and the explainable framework. We have carefully covered the experimental procedures along with the classification outcome of the model. The model has been created and evaluated using test data for the classification of rhythmic cardiac disease. Precision, recall, F1-score, a confusion matrix, and a visual explanation make up the five performance evaluation measures that are used, and they all make it clear how effective the proposed model is. The research questions that were presented in the first chapter of this study were also addressed.

4.1 Experimentation Setup

The computer used has a configuration consisting of MacBook Pro, Processor: Apple M1, Total Number of Cores: 8, Memory: 16 GB, Storage: 1TB, and running macOS 12.3.1 (21E258) operating system. Python version 3.9.12 and its libraries are used as the main programming language. Keras, an open-source software library that provides a Python interface for artificial neural networks, is used to implement the model along with TensorFlow and Scikit-learn.

4.2 ECG classification results

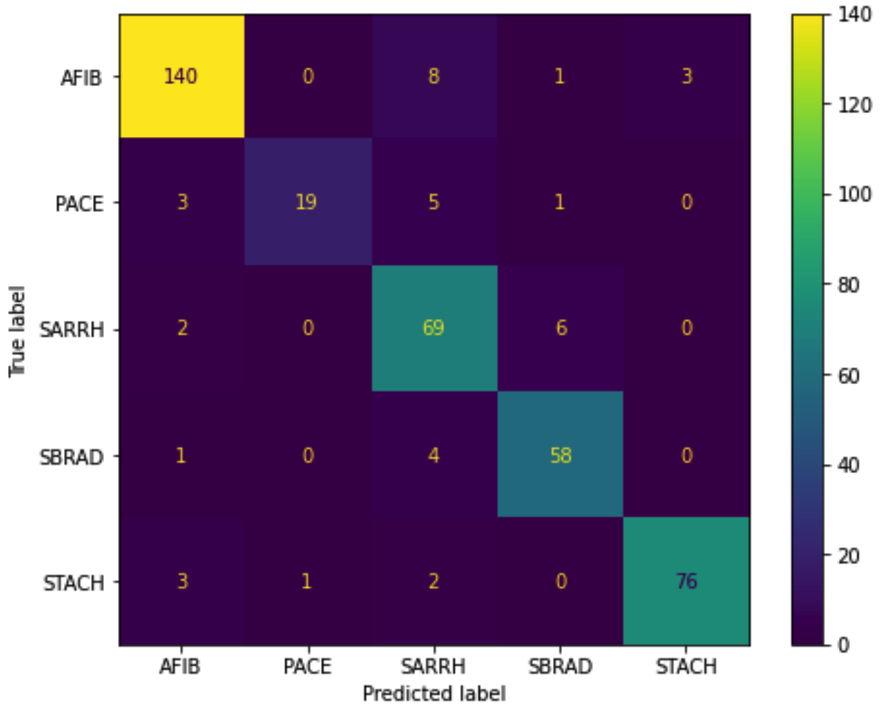
In Chapter 3, we have shown that a model was developed for the detection of rhythm-based heart diseases using the PTB-XL dataset. Our proposed model was used to classify heart disease from the ECG signals into five categories (AFIBs, PACEs, SARRHs, SBRADs, and STACHs) based on rhythm. We used 10-fold cross-validation for model training and evaluation. Following the selection of rhythm-related data from the PTB-XL dataset and randomly dividing the selected rhythm data into 10 folds, 8 of the 10 folds were used for training, 1-fold for validation, and 1-fold for testing. The test dataset was then used in order to get results. The reported results are the best results obtained on the test dataset during experimentation. The model was trained using catagorical-crossentropy as the loss function and the Adam optimizer as the optimization technique. The learning rate of 0.001, batch size of 32, and dropout rate of 0.2 were the hyperparameters of the model, and the maximum number of epochs was 60, at which we got good results. The results of our proposed approach are displayed in Table 4.1 for the evaluation metrics of accuracy, precision, recall, and F1-score for each class and the average results.

The model's predictions for each cardiac rhythm on the test dataset are provided in Table 4.1 along with their precision, recall, F1 score, and accuracy. Our model's overall average accuracy was greater than 0.96, and the average F1 score was 0.88. An average precision of 0.90 and an average recall of 0.87 were also attained by the proposed method. There are fewer false negatives when recall is good and fewer false positives when precision is good. The F1 score is the harmonic mean of precision and recall. In our work, when we compared at the class level, the following two classes showed our model performed better. AFIB was classified with an F1 score of 0.93, and STACH was classified with an F1 score of 0.94. We also noted that Normal Functioning Artificial Pacemaker (PACE)'s F1 score is as low as 0.79, which is because of the low recall value of PACE due to high False negative (FN) values in this class. From Table 4.2, we can see that there are 9 False Negative values, and and only 1 False positive out of 28 records.The high False negative (FN) values affected the result of Recall, and the low False positive value supported for good result in Precision. Since F1-Score is the harmonic mean of precision and recall,F1-Score of PACE is as lower as 0.79. However, the accuracy value of PACE is found to be high (0.98), which is due to high value of True Negative (TN), as can be seen from the formula for Accuracy, equation 3.1. The TN value for PACE can be seen from the Confusion matrix in Table 4.2 to be an extra-large value of 373.

Table 4.1: Classification Report for the five class labels.

Classes	Precision	Recall	F1-score	Accuracy	support
AFIB	0.94	0.92	0.93	0.95	152
PACE	0.95	0.68	0.79	0.98	28
SARRH	0.78	0.90	0.84	0.93	77
SBRAD	0.88	0.92	0.90	0.97	63
STACH	0.96	0.93	0.94	0.98	82
Average	0.90	0.87	0.88	0.96	402

Table 4.2: Confusion Matrix result for the five class labels



In Table 4.2, the rows represent the true labels of the samples, while the columns represent the predicted labels. Each cell in the matrix represents the number of samples that were classified as having a particular predicted label, given the true label of the sample. For example, the cell in row 1, column 1, shows that 140 samples were correctly classified as AFIB. The cell in row 1, column 3, shows that 8 samples that were actually classified as AFIB were incorrectly classified as SARRH. Similarly, the cell in row 4, column 3, shows that four samples that were actually classified as SBRAD were incorrectly classified as SARRH. We can use this confusion matrix to calculate various performance metrics of the classification model, such as F1 score, recall, accuracy, and precision.

The proposed model obtained a respectable F1-score of 0.88, which indicates the overall performance of the model.From Table 4.3 A sensitivity of 0.87, which is the capacity of our model to detect changes in input data and use it to evaluate how well it works even when dealing with unseen data, aids generalization and improves performance. AUC of 0.98 which measures the overall effectiveness of the classification.

Table 4.3: Performance of the proposed method in some metrics for the ECG classification task.

Evaluation Metric	Score Obtained
Hamming Loss	0.0398
AUC	0.98
Sensitivity	0.87

The loss and accuracy curves obtained when the model was being trained are shown in Figs. 4.1 and 4.2. These curves were determined on both the training and validation sets. The percentage of test points with correctly identified labels for each one is known as accuracy, which is the strongest metric. Attaining a value of 0.96, our proposed model demonstrated relevant accuracy. When accuracy is high, fewer false positives are produced. The proposed methodology also demonstrated respectable precision performance (0.90).

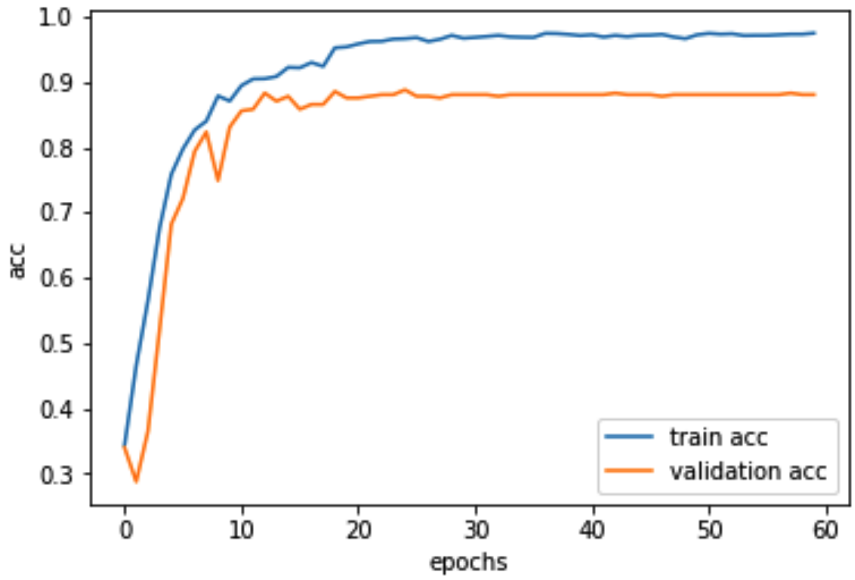


Figure 4.1: Accuracy curve of the model during training

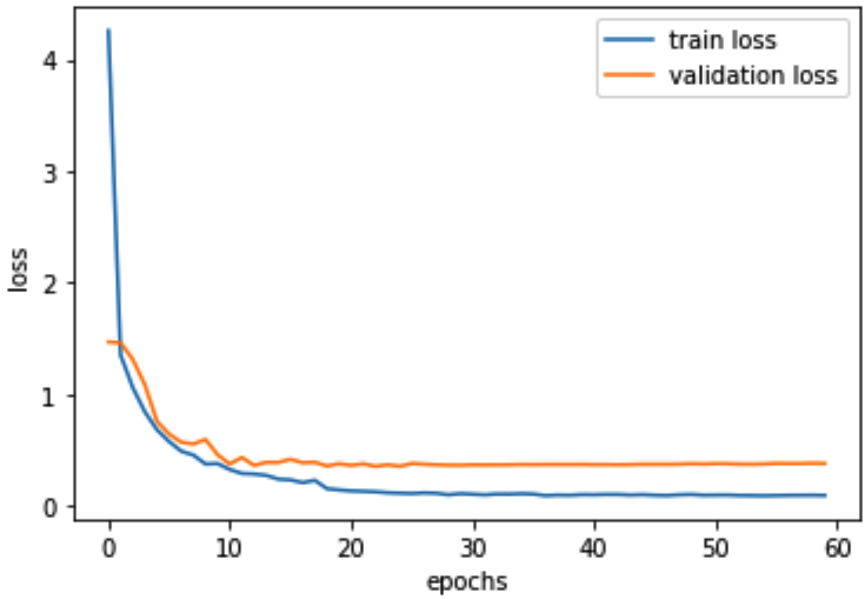


Figure 4.2: Loss curve of the model during training.

4.3 Model Explainability

For deep neural network applications in the real world, models' lack of explainability is a common challenge and limiting factor. In this work, to alleviate this problem, we used Grad-Cam and Grad-Cam++ visual explanation approaches to explain our model predictions.

4.3.1 Visual explanations for ECG classification using Grad-cam & Grad-cam++

Grad-CAM and Grad-CAM++ are methods for creating visual explanations that highlight the region of an ECG signal that are essential for a classification decision made by a deep neural network.

Grad-CAM creates visual explanation by calculating the gradients of a convolutional neural network's output with respect to the feature maps of the final convolutional layer. These gradients are then averaged, and the resulting map is utilized to weight the feature maps of the last convolutional layer. To create the final visual explanation, the weighted feature maps are then added together.

In addition to the first-order gradients present in Grad-CAM, Grad-CAM++ expands Grad-CAM by integrating second-order gradients. To produce a more precise visual explanation, it specifically computes the second-order gradients of the output with respect to the feature maps of the last convolutional layer and combines them with the first-order gradients. We consider a few samples to compare the two visual explainability measures: Grad-CAM and Grad-CAM++.

Even though, due to the computation of the second-order gradients, Grad-CAM++ require a higher runtime complexity of 10.52 sec/record compared to 6.38 sec/record for Grad-CAM as can be seen in Figures 4.3, 4.4, 4.5, and 4.6, Grad-CAM++ generally produces visual explanations that are more precise and focused than Grad-CAM; that is, Grad-Cam++ is better than Grad-Cam in visual explanations and important region localization.

Figure 4.3: Irregularly irregular ventricular rate: consecutive ventricular contractions are spaced at irregular intervals and with different shapes. When we compare visual explanations with Grad-CAM and Grad-CAM++, Grad-CAM++ shows the most crucial region of the ECG signals on lead v1 and the ECG signal on lead III in a dark color, indicating that it is the most important region of the ECG signal for the classification.

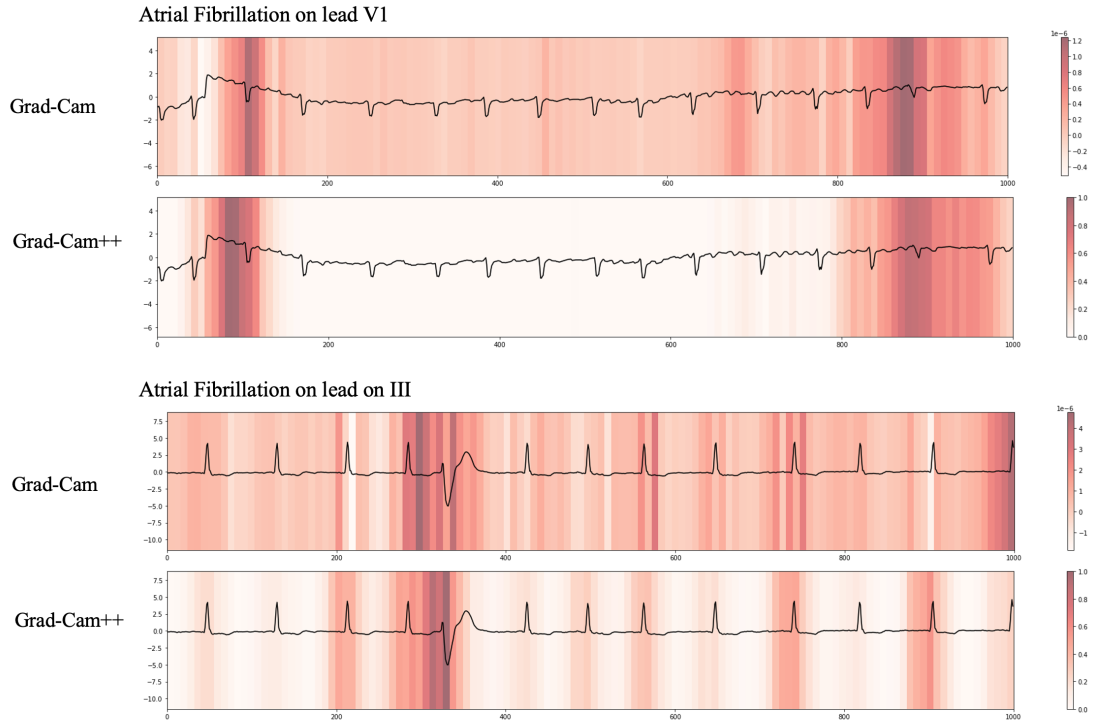


Figure 4.3: Grad-cam vs. Grad-cam++ visual explanations for AFIB on leads V1 & III

A heart rate that is higher than normal is referred to as sinus tachycardia, but it is still a sinus rhythm. Figure 4.4 shows this condition. From the figure, Grad-CAM++ is good at localizing the important region of the ECG signal by highlighting it with color.

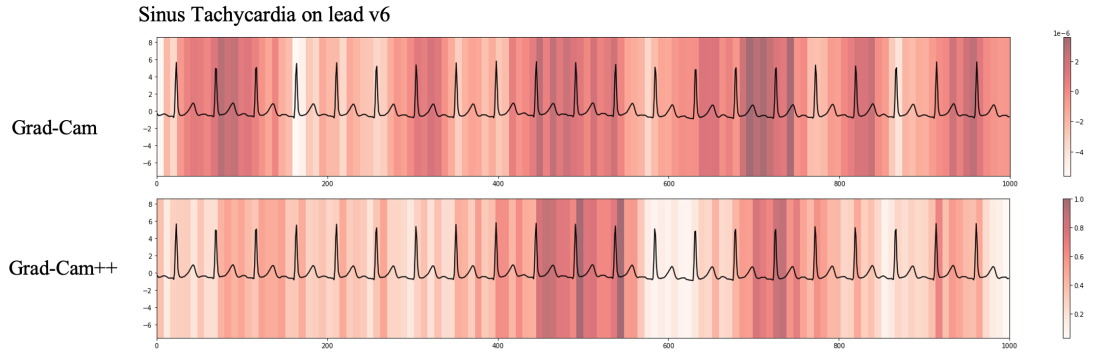


Figure 4.4: Grad-cam vs. Grad-cam++ visual explanations for STACH

There is a condition under which both Grad-CAM and Grad-CAM++ perform similarly. From equations 2.3 and 2.6, we have seen that if α_{ij}^{kc} = 1/Z, Grad-CAM++ reduces to the formulation for Grad-CAM. From Figure 4.5, it can be seen that both Grad-Cam and Grad-Cam++ have shown identical important region localization for a record classified as Sinus Arrhythmia on the lead V3.

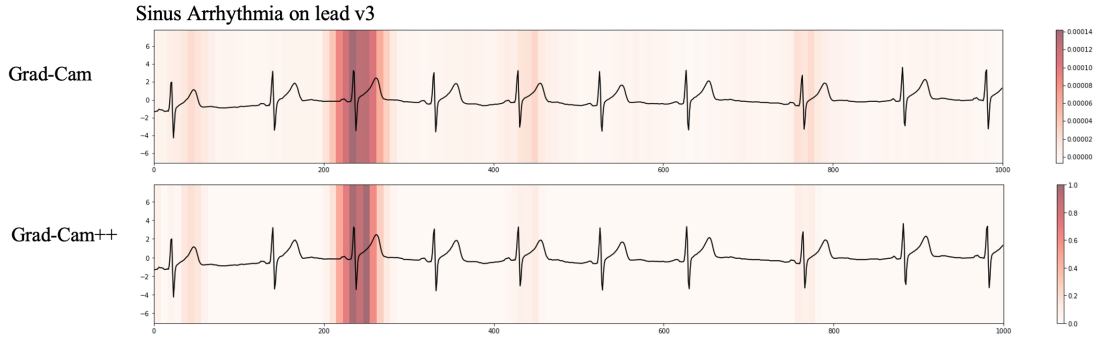


Figure 4.5: Grad-cam vs. Grad-cam++ visual explanations for SARRH

Figure 4.6 shows an output of both Grad-Cam and Grad-Cam++ for the record classified as SBRAD, with the interval step being slower than the normal sinus rhythm.The figure shows important region localization of SBRAD on the lead V5. However, Grad-Cam++ shows a narrower localization for the most important region for the classification. The narrower the important region localized, the more explainable the model becomes, as it precisely focuses on the most important regions for the classification.

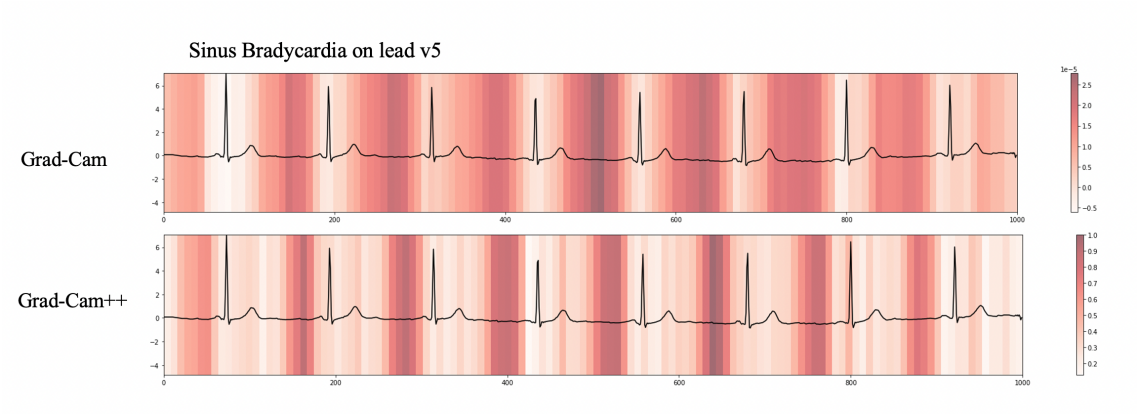


Figure 4.6: Grad-Cam vs. Gad-Cam++ visual explanations for SBRAD

We have seen that Grad-CAM++ outperformed Grad-CAM in the localization of important regions. By examining the Grad-CAM++ visualization, clinicians can gain insights into the features of the ECG signal that are most relevant to different classifications. For example, they may see that certain regions of the signal, such as the P-wave or R-R intervals, are particularly important for distinguishing between different types of heart rhythm problems. This can aid in the explanation of ECG signals and help clinicians make more informed diagnoses and treatment decisions.

4.3.2 Examples of Grad-cam++ visual explanations

The rate in normal sinus rhythm is generally regular, as shown in (Fig.4.7. Normal sinus rhythm). A sinus arrhythmia is a deviation from this normal sinus rhythm that typically establishes an irregular rate with a larger variation in the R-R interval. This implies that the heartbeat is irregular, which means that the period between each heartbeat changes over time. Additionally, P waves are typically monofom. This is clearly seen in (Fig 4.7.SARRH on Lead I, SARRH on Lead II, and SARRH on Lead V6) and highlighted with color.

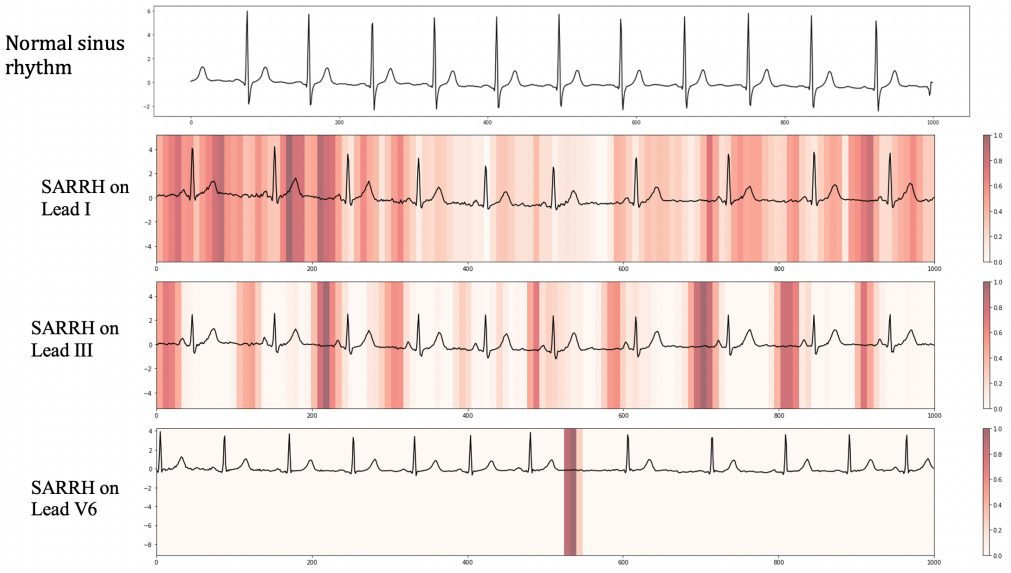


Figure 4.7: sinus arrhythmia on leads I, II, v6: irregular change in the R-R interval.

Sinus bradycardia is a medical disorder in which the heart beats more slowly than it should while at rest. This state is characterized by a regular but broader R-R interval. Because of the wider R-R intervals, when we count the number of R-R intervals in normal sinus rhythm (Fig.4.8: Normal sinus rhythm), it is 10 in number, but when we count the R-R intervals in the same figure (Fig.4.8 SBRAD on lead V4, SBRAD on lead V5), it is only seven. This means that the sinus bradycardia is slower than the normal range. Our model output, which identified sinus bradycardia, highlighted the problematic area in red. If the heart rate drops too low, it may cause decreased blood flow to the organs, which can lead to fainting, confusion, and even cardiac arrest. Clinicians may use medication to regulate the heart rate after looking at the model's output.

Figure 4.8: Sinus Bradycardia on leads V4, V5: A sinus rhythm slower than the normal range.

Sinus Tachycardia is a medical condition in which the heart beats faster than it should when a person is at rest. Normally, the heart rate is as shown in Fig. 4.9 (normal sinus rhythm), however on the same figure (Fig. 4.9: STACH on leads I, II) is characterized by a faster than normal heart rate, so that the R-R interval is very narrow and the number of R-R intervals is much higher compared to the normal sinus rhythm. Our model shows the abnormality by highlighting the region of the ECG signal with problems in red.

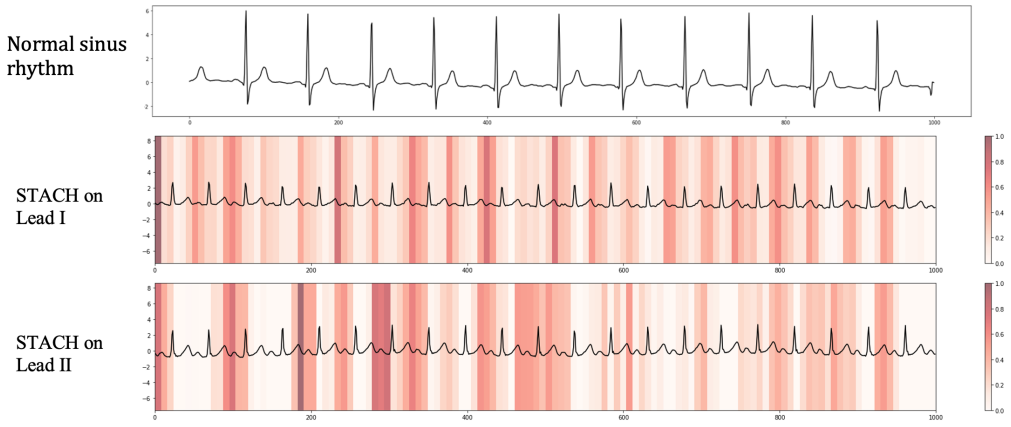


Figure 4.9: Sinus Tachycardia on lead I, II: A sinus rhythm that beats more quickly than usual.

When a person is in good health, their heart displays a predictable and regular wave pattern that consists of the P wave, QRS complex, and T wave.(Fig.4.10 Normal sinus rhythm) shows this pattern as it appears when the heart beats at a regular rate. However, various electrical patterns are shown in the same figure (Figure 4.10: AFIB on lead III, AFIB on lead on aVF, and AFIB on lead V1). In this condition, the impulses that control heartbeat become disordered, which causes disordered vibrations in the atria. As a result, a different and irregular wave pattern is seen that deviates from the characteristic sinus rhythm by fluttering rather than adequately contracting. Due to the absence of P waves and irregularly spread-out waves, it produces different forms than a normal sinus rhythm. Our model was able to identify this irregularity by highlighting the region in a red color to distinguish it from the typical rhythm.

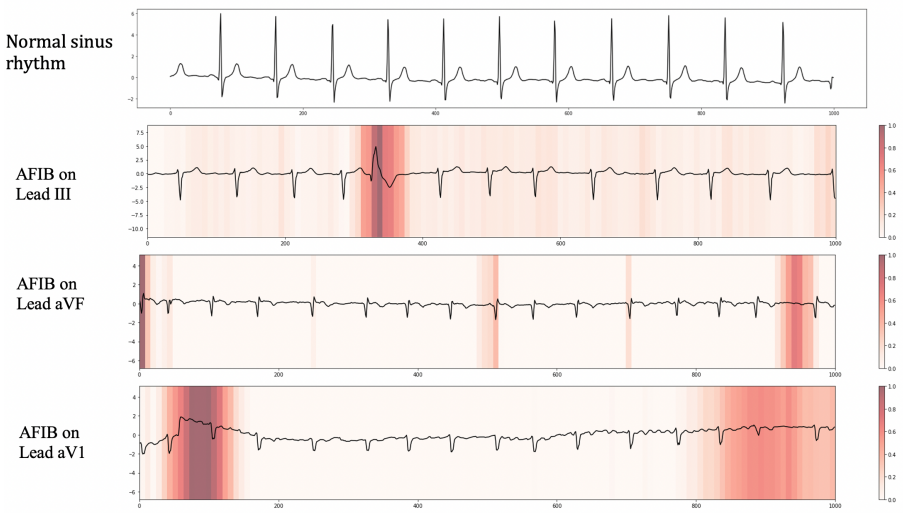


Figure 4.10: AFIB on lead III, aVF, V1: Irregularly irregular heartbeat rate, no visible P waves.

4.4 Discussion

In this work, a classification of rhythm-based heart disease has been carried out using the proposed model. From the result, we can observe that our model has achieved a good performance in classifying rhythm-based heart disease (AFIB, PACE, SARRH, ABRAD, and ATACH). Furthermore, we were able to reason out why our model made the classifications with a visual explanation using Grad-cam++.

By doing this experiment, we were able to answer the research question stated in Chapter one.

> RQ: What is the effect of applying visual explainability using Grad-CAM and Grad-CAM++ on the reliability of the classification of rhythm-based heart disease?

According to the findings of our research, the proposed model is capable of acceptably classifying heart disease with a rhythmic base. In particular, our model achieved an overall classification accuracy of above 93% for all the class labels contained in the unseen dataset.

These outcomes indicate that the proposed model performs well in classifying and distinguishing between different types of cardiac diseases based on their characteristic rhythms. Our model's accuracy and the explanation for its classification suggest that it can be used as a reliable tool to support the identification of people with these diseases. Overall, our results provide important insights into the effectiveness of deep learning approaches in the field of healthcare when the explainability framework is included in the work flow of deep learning.

Understanding and using the model's decision-making process establishes the explainability component of our model. Visual explanations, which highlight the significant regions or aspects that contribute to the model's output, are one method for achieving this.

The standard visual explanation methods we used include Grad-cam and Grad-cam++. In literature, it is stated that Grad-CAM++ outperforms Grad-CAM in localizing multiple class instances within a single image [22]. In our work, we applied Grad-Cam and Grad-Cam++ to a 1D ECG signals to localize the most important region used for the classification. It was found that Grad-CAM++ produced better visual explanations and more precise localization of important regions of the ECG signals when results from using Grad-cam and Grad-CAM++ for visual explanations were compared. In other words, Grad-CAM++ outperformed Grad-cam in its ability to pinpoint the most crucial areas of the input ECG signals that influenced the model's decision-making. In general, the use of visual explanations can enhance the transparency and explainability of deep learning models and offer useful insights into how they make decisions.

In order to evaluate a model's output with domain experts, we randomly chose 10 records from the output of our model and gave them to two medical specialists. We gave 10 records to each doctor for their feedback. Each expert looking at 10 records means that the doctor is required to analyze the ECG of 10 patients, each of which has 12 leads. It takes several hours to read the ECG result of a patient, and it requires careful investigation of each lead. That is why we are limited to only 10 records to receive domain experts comments.

According to the reports from the two specialists, First-Expert offered positive feedback on nine out of ten records that the model's output matched, indicating that for nine out of the ten records that First-Expert was given, the model's output matched his evaluation. However, First-Expert disagreed with the categorization and explanation of one particular record, PACE.

Table 4.4: comments from two domain experts on our model's results

Records	Condition	Expert's Comment		Agree, Disagree or Equivocal	
		Expert-One	Expert-Two	Expert-One	Expert-Two
1	**Sinus Arrythmia (SARRH)**	Comments: Sinus arrhythmia, baseline artifact poor R wave progression /late R waves transition, possibly counter clockwise heart rotation	Sinus Arrhythmia: EQUIVOCAL: No strong evidence, threshold for irregularity needs to be a bit higher	**Agree**	**Equivocal**
2	**Sinus Bradycardia (SBRAD)**	Comments: Sinus bradycardia, likely lead misplacement/reversal, baseline artifact	Sinus Bradycardia: AGREE	**Agree**	**Agree**
3	**Atrial Fibrillation (AFIB)**	Comments: atrial fibrillation with fast ventricular response	Atrial Fibrillation: AGREE	**Agree**	**Agree**
4	**Normal functioning artificial pacemaker (PACE)**	Comments: atrial fibrillation; impossible to see pacemaker spikes (if present) because of artifacts	Normal functioning artificial pacemaker (PACE): EQUIVOCAL (no visible clue for pacing but could be a paced rhythm like in bipolar VVI pacemakers): Irregularity due to premature ventricular contraction	**Disagree**	**Equivocal**
5	**Sinus Bradycardia (SBRAD)**	Comments: Sinus bradycardia, baseline artifact, benign early repolarization	Sinus Bradycardia (SBRAD): AGREE	**Agree**	**Agree**
6	**sinus tachycardia (STACH)**	Comments: sinus tachycardia, PVCs (premature ventricular complexes)	Sinus tachycardia (STACH): AGREE (with single PVC on second QRS)	**Agree**	**Agree**
7	**Atrial Fibrillation (AFIB)**	Comments: multifocal atrial fibrillation	Atrial Fibrillation: AGREE (with fast ventricular response)	**Agree**	**Agree**
8	**Sinus Arrythmia (SARRH)**	Comments: baseline artifacts and baseline wander. Sinus arrhythmia	Sinus Arrythmia (SARRH): AGREE but might still be normal if slight irregularities are ignored (same as above)	**Agree**	**Agree**
9	**Sinus Tachycardia (STACH)**	Comments: sinus tachycardia, aberrantly conducted PAC	Sinus Tachycardia: AGREE (with PVC)	**Agree**	**Agree**
10	**Atrial Fibrillation (AFIB)**	Comments: atrial fibrillation	Atrial Fibrillation: AGREE	**Agree**	**Agree**

Similarly, Second-Expert agreed with the model's classification for 8 records, which means that for 8 out of the 10 records provided to him, the model's output was in agreement with his evaluation. However, Second-Expert classified two of the recordings as equivocal, indicating that the model's classification and explanation of those two weren't precise or clear. PACE and SARRH were the names of the two records designated as equivocal.

Based on the results of those reports from the two domain experts given in Table 4.4, it can be said that the output of the model is mostly correct and consistent with the interpretations of the two doctors. The fact that both doctors had issues with the classification of the PACE record reflects the limited number of records used for training and testing in this class. Their issues also agreed with the smallness values of 0.68 on the recall scale and 0.79 on the F1-score found in the result for PACE, which are the smallest when compared to other classes recall values and F1-scores.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Our work aimed to develop an explainable model for classifying rhythm-based heart disease from 12-lead ECG signals using the PTB-XL dataset. We proposed an approach using the ResNet-18 deep learning model integrated with visual explainers (Grad-CAM and Grad-CAM++). We examined the impact of employing Grad-CAM and Grad-CAM++ visual explainers on the reliability of classifying rhythm-based cardiac disease. The findings demonstrate that our technique effectively identifies rhythm-based heart disease, while the visual explainability builds trust in the deep learning model's predictions. Our model was effective, with a classification accuracy of 96% and an F1 score of 0.88. We gave samples of our model's explanations to two domain experts, and the experts agreed with 80% of the explanations given to them. In general, our work addressed the need for explainable model in cardiac diagnostics and valuable insights were gained with regard to visual explainers.

5.2 Future Work

In future work, it is recommended to include all remaining rhythm types in the study, as the current research focused on only five out of eleven rhythm based heart diseases due to limited data availability. Gathering a larger range of ECG recordings from different datasets, would be beneficial for expanding the training dataset. This expansion has the potential to improve the explanation of the classification of all rhythm-based heart problems if the remaining six rhythms become available.

Collaboration with domain experts is important to ensuring the practical application of research results in the real world. Their expertise and experience can contribute to the successful implementation of the findings if they validate all the classifications made by the deep learning model.

References

[1] M. Ganeshkumar, V. Ravi, V. Sowmya, E. Gopalakrishnan, and K. Soman, "Explainable deep learning-based approach for multilabel classification of electrocardiogram," *IEEE Transactions on Engineering Management*, 2021.

[2] E. J. Benjamin, P. Muntner, A. Alonso, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, A. R. Chang, S. Cheng, S. R. Das, *et al.*, "Heart disease and stroke statistics—2019 update: a report from the american heart association," *Circulation*, vol. 139, no. 10, pp. e56–e528, 2019.

[3] E. Jing, H. Zhang, Z. Li, Y. Liu, Z. Ji, and I. Ganchev, "Ecg heartbeat classification based on an improved resnet-18 model," *Computational and Mathematical Methods in Medicine*, vol. 2021, 2021.

[4] E. Izci, M. A. Ozdemir, M. Degirmenci, and A. Akan, "Cardiac arrhythmia detection from 2d ecg images by using deep learning technique," in *2019 Medical Technologies Congress (TIPTEKNO)*, pp. 1–4, IEEE, 2019.

[5] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[6] M. Soliński, M. Lepek, A. Pater, K. Muter, P. Wiszniewski, D. Kokosińska, J. Salamon, and Z. Puzio, "12-lead ecg arrythmia classification using convolutional neural network for mutually non-exclusive classes," in *2020 Computing in Cardiology*, pp. 1–4, IEEE, 2020.

[7] A. Cheffer, M. A. Savi, T. L. Pereira, and A. S. de Paula, "Heart rhythm analysis using a nonlinear dynamics perspective," *Applied Mathematical Modelling*, vol. 96, pp. 152–176, 2021.

[8] 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, 2022.

[9] A. Peimankar and S. Puthusserypady, "Dens-ecg: A deep learning approach for ecg signal delineation," *Expert systems with applications*, vol. 165, p. 113911, 2021.

[10] A. Atangana and S. I. Araz, "Rhythmic behaviors of the human heart with piecewise derivative," *Math. Biosci. Eng*, vol. 19, pp. 3091–3109, 2022.

[11] 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, no. 1, pp. 1–15, 2020.

[12] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *Advances in neural information processing systems*, vol. 29, 2016.

[13] P. Celard, E. Iglesias, J. Sorribes-Fdez, R. Romero, A. S. Vieira, and L. Borrajo, "A survey on deep learning applied to medical images: from simple artificial neural networks to generative models," *Neural Computing and Applications*, vol. 35, no. 3, pp. 2291–2323, 2023.

[14] A. Jafar and L. Myungho, "Hyperparameter optimization for deep residual learning in image classification," in *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pp. 24–29, IEEE, 2020.

[15] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: a survey," *Applied Sciences*, vol. 12, no. 18, p. 8972, 2022.

[16] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, IEEE, 2020.

[17] M. S. Ebrahimi and H. K. Abadi, "Study of residual networks for image recognition," in *Intelligent Computing*, pp. 754–763, Springer, 2021.

[18] M. RIZVI, K. Deb, M. I. Khan, M. KOWSAR, M. SAKI, and T. KHANAM, "A comparative study on handwritten bangla character recognition," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 4, pp. 3195–3207, 2019.

[19] S. Sahoo, "Residual blocks," 2018.

[20] S. Sattarzadeh, M. Sudhakar, K. N. Plataniotis, J. Jang, Y. Jeong, and H. Kim, "Integrated grad-cam: Sensitivity-aware visual explanation of deep convolutional networks via integrated gradient-based scoring," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1775–1779, IEEE, 2021.

[21] Q. Zhang, L. Rao, and Y. Yang, "Group-cam: group score-weighted visual explanations for deep convolutional networks," *arXiv preprint arXiv:2103.13859*, 2021.

[22] 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)*, pp. 839–847, IEEE, 2018.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[24] N. O. Pinciroli Vago, F. Milani, P. Fraternali, and R. da Silva Torres, "Comparing cam algorithms for the identification of salient image features in iconography artwork analysis," *Journal of Imaging*, vol. 7, no. 7, p. 106, 2021.

[25] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic relu," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pp. 351–367, Springer, 2020.

[26] Y. Wang, Y. Li, Y. Song, and X. Rong, "The influence of the activation function in a convolution neural network model of facial expression recognition," *Applied Sciences*, vol. 10, no. 5, p. 1897, 2020.

[27] Y. Yu, K. Adu, N. Tashi, P. Anokye, X. Wang, and M. A. Ayidzoe, "Rmaf: Relu-memristor-like activation function for deep learning," *IEEE Access*, vol. 8, pp. 72727–72741, 2020.

[28] H. Smulyan, "The computerized ecg: friend and foe," *The American journal of medicine*, vol. 132, no. 2, pp. 153–160, 2019.

[29] S. Hong, Y. Zhou, J. Shang, C. Xiao, and J. Sun, "Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review," *Computers in Biology and Medicine*, vol. 122, p. 103801, 2020.

[30] W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of cnn and rnn for natural language processing," *arXiv preprint arXiv:1702.01923*, 2017.

[31] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 2392–2396, IEEE, 2017.

[32] Z. Zhang, Z. Li, and Z. Li, "An improved real-time r-wave detection efficient algorithm in exercise ecg signal analysis," *Journal of Healthcare Engineering*, vol. 2020, 2020.

[33] S. Somani, A. J. Russak, F. Richter, S. Zhao, A. Vaid, F. Chaudhry, J. K. De Freitas, N. Naik, R. Miotto, G. N. Nadkarni, *et al.*, "Deep learning and the electrocardiogram: review of the current state-of-the-art," *EP Europace*, vol. 23, no. 8, pp. 1179–1191, 2021.

[34] D. Zhang, S. Yang, X. Yuan, and P. Zhang, "Interpretable deep learning for automatic diagnosis of 12-lead electrocardiogram," *Iscience*, vol. 24, no. 4, p. 102373, 2021.

[35] 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, no. 1, p. 111, 2022.

[36] I. Kumar, C. Scheidegger, S. Venkatasubramanian, and S. Friedler, "Shapley residuals: Quantifying the limits of the shapley value for explanations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26598–26608, 2021.

[37] V. Jahmunah, E. Y. K. 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, 2022.

[38] H. Lee and M. Shin, "Learning explainable time-morphology patterns for automatic arrhythmia classification from short single-lead ecgs," *Sensors*, vol. 21, no. 13, p. 4331, 2021.

[39] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras, "Automated atrial fibrillation detection using a hybrid cnn-lstm network on imbalanced ecg datasets," *Biomedical Signal Processing and Control*, vol. 63, p. 102194, 2021.

[40] H. Dang, M. Sun, G. Zhang, X. Qi, X. Zhou, and Q. Chang, "A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals," *IEEE Access*, vol. 7, pp. 75577–75590, 2019.

[41] M. Wolf, G. Varigos, D. Hunt, and J. Sloman, "Sinus arrhythmia in acute myocardial infarction," *Medical Journal of Australia*, vol. 2, no. 2, pp. 52–53, 1978.

[42] D. S. shakya, “Sinus arrhythmia ecg - cardiac tamponade,” 2023.

[43] M. P. Soos and D. McComb, “Sinus arrhythmia,” in *StatPearls [Internet]*, StatPearls Publishing, 2021.

[44] M. P. S. D. McComb, “Ncbi books,” 2022.

[45] M. Ali, A. Q. Haji, A. Kichloo, B. P. Grubb, and K. Kanjwal, “Inappropriate sinus tachycardia: a review,” *Reviews in Cardiovascular Medicine*, vol. 22, no. 4, pp. 1331–1339, 2021.

[46] F. A. Alnajim, M. A. S. Alkhidhr, M. A. A. Alanazi, A. A. J. Bawazeer, A. I. Shahar, B. M. Alsharif, O. Q. K. Alanazi, O. Q. Darraj, N. M. Mohamed, A. A. M. Alawi, *et al.*, “An overview of diagnosis and management of bradycardia: Literature review.,” *Archives of Pharmacy Practice¦ Volume*, vol. 12, no. 1, 2021.

[47] F. M. Kusumoto, M. H. Schoenfeld, C. Barrett, J. R. Edgerton, K. A. Ellenbogen, M. R. Gold, N. F. Goldschlager, R. M. Hamilton, J. A. Joglar, R. J. Kim, *et al.*, “2018 acc/aha/hrs guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society,” *Journal of the American College of Cardiology*, vol. 74, no. 7, pp. e51–e156, 2019.