University of Essex
School of Computer Science and EE

A DISSERTATION SUBMITTED FOR THE DEGREE OF
MASTERS IN ARTIFICIAL INTELLIGENCE

Automatic Diagnosis of Heart Diseases Using Artificial Intelligence

Ahmad Raza

2101194

Supervisor: Dr. Cunjin Luo

August 30, 2022
Colchester Campus, Essex

Abstract

Among many other non-infectious diseases, heart diseases are a major cause of death all over the World. Early diagnosis of the abnormalities in the heart can help in curing the diseases and reducing the death toll. A device known as an "electrocardiogram" (ECG) is widely used for the identification of the patterns and rhythms in the heart. This device contains multiple electrodes that are placed over different parts of the body. In return, they detect the electrical signals in the heart. It does so by recording and analyzing the contractions in both chambers of the heart. Results from the ECG device are then manually analyzed by the physicians to detect different arrhythmic diseases of the heart, but it comes with a lot of problems. It is a time taking process that requires a lot of human resources and experts. Due to the complex nature of ECG signals, chances of human error are high which can result in a disease going undetected and can become a cause of death. Keeping in view these problems, an automatic detection system is required for heart diseases. Researchers have spent the past few years finding a solution to the problem at hand. They used various machine learning classification models trained on expert features as well as deep learning models which themselves are capable of extracting the robust features from the ECG and performing classification based on them. Despite the efforts, they couldn't succeed up to an extent where their research could be implemented on a clinical level. This was due to the limited homogeneous and origin-oriented data sets available, use of single-lead ECGs and performing single-label classification. In 2020, physio net introduced a massive 12 lead data set collected from various hospitals all over the world. Using this data set, we attempted to increase the accuracy achieved by the previous works by proposing an ensemble model based on various baseline residual neural network models to classify the 27 heart diseases. By assigning weights to the predictions from the base models, our model achieved an accuracy of 94.57 percent beating the best-ranked model in the competition. We also used a novel approach of assigning weights to the labels in our project to deal with the data imbalance.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

A significant ratio of people are diagnosed with various heart diseases all over the world on daily basis and this graph keep on growing with every passing hour. [16] Heart diseases also referred to as "Cardio-Vascular" diseases are caused due to multiple behaviors and factors related to health such as smoking, irregular sleep patterns, imbalanced diet, unhealthy lifestyle, uncontrolled blood pressure, bad cholesterol, and glucose levels in the blood. [17] Heart diseases are a known cause of taking millions of lives annually grabbing the first place in the list of most dangerous non-infectious as well as infectious diseases combined. [43] According to one of the studies carried out by the World Health Organization(WHO), half of the casualties in developed parts of the World including States occur due to Cardio Vascular diseases. This number adds up to approximately 12 million a year. [40]

Heart diseases occur when the beating of the heart is either too slow or too fast as compared to the normal heart beat rate, [32] which is referred to as "sinus rhythm". [18] Such abnormal patterns of the heart are known as "heart arrhythmia" and can be categorized into many types. [6] The most common category among abnormal heart rhythms is atrial fibrillation which alone is responsible for heart failure and mortality rates in the major population of elderly people. [22]

1.1 Background

Morbidity and mortality rates can be controlled if the heart arrhythmias are diagnosed and treated at early stages. [1] Electrocardiogram (ECG) is the most widely used medical equipment to record the rhythms and electrical activities of the heart which make it a popular diagnostic tool for the detection of heart diseases. [9] It records and analyzes the contractions of both (upper and lower) chambers of the heart to detect arrhythmic diseases. Along with heart arrhythmia, it is also capable of detecting abnormalities in the heart that are not arrhythmic by nature such as heart attacks and enlargement or compression of heart chambers. [14] ECG consists of electrodes that are placed on the human body to trace and record the electrical signals produced by the heart. In the case of one Lead ECG, two to three electrodes are placed on the body at places that are already determined whereas 12 lead ECG is performed using ten electrodes that are positioned at various locations of the body. One lead ECG assists in the monitoring of the heart at a basic level and can be used for non-clinical purposes such as in self-monitoring devices for determining the heartbeat. Whereas, 12 Lead ECG provides a full picture of the electrical activity of the heart and plays an important role in diagnosing fatal heart diseases. [10] Among 12 leads, six leads are known as "limb leads" whereas the remaining six leads are known as "chest leads". They are named according to their positioning on the body. Lead I, lead II, lead III, lead aVR, aVF, and aVL are categorized as limb leads whereas leads v1, v2, v3, v4, v5, and v6 are known as chest leads. Figure 1.3

An electrocardiogram records the electrical activity of the heart and gives output in the form of waves. ECG waves constitute different components. Figure 1.1 The points in the ECG waves are named P wave, Q wave, R wave, S wave and, T wave. These points in return form different intervals and segments such as Q R S complex, P R segment, P R interval, S T segment, T P segment, Q T interval, and R R interval. A p-wave is a result of depolarization of the atria, whereas Q R S complex is formed when ventricles are depolarized and finally re-polarization of the ventricles generates T-wave. [27] These wave components prove very beneficial while extracting different features from the ECG wave in diagnosing diseases. The component's start points, endpoints, width, length, and height act as significant features and play their role in arrhythmia detection. [29]

Different ECG features have been visualized in Table 1.1 for better understanding. Their normal values in terms of the time interval are given in milliseconds. P and T Waves have

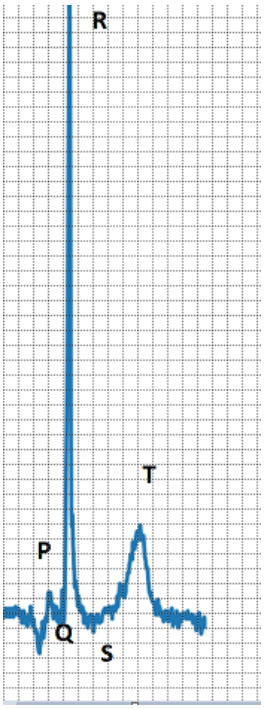

Figure 1.1: ECG Wave points and intervals.

a normal value of 80 and 160 respectively whereas the Q R S complex's normal interval is between 80 to 100. P R and S T segments have a maximum value of 120 whereas the P R interval's value can go up to 200 starting from 120. ST and QT intervals are considered to be normal at values 320 and 420 respectively. [44] If the values of these components (features) are different from the ones given in the table for Sinus Rhythm (Normal Heart Condition), then it's a red flag signaling the presence of an abnormal heart rhythm. Therefore, abnormality in these values can be used in the detection of different heart diseases.

1.2 Motivation

Once the results are obtained from the ECG, physicians examine those results to analyze the heart condition based on their experience and practice. Due to multiple reasons, this manual diagnosis done by humans is becoming a great concern. Due to the exponential increase in the number of heart patients day by day, its becoming problematic and almost impossible for cardiologists to analyze the Electrocardiogram (ECG) reports on time. These delays can worsen the patient's condition and may cost them their life. Moreover, Electrocardiograms are complex by nature, therefore chances of human error are very high. A single fatal

ECG components(features)	Time Interval(ms)
P Wave	80
T Wave	160
Q R S Complex	80 to 100
P R Segment	50 to 120
ST Segment	80 to 120
P R Interval	120 to 200
ST Interval	320
QT Interval	420

Table 1.1: ECG features w.r.t time intervals

disease that goes undetected can become a cause of death. [5] Lastly, analyzing hours of electrocardiograms for multiple patients and that too repeatedly requires hospitals to hire more resources and manpower, therefore clustering their funds in one place and forcing them to ignore other domains.

To address the concerns regarding the manual investigation of the ECG results, an automatic diagnosis system is required that can classify and diagnose multiple heart diseases based on the features extracted from the electrocardiogram. A lot of work has been carried out in this regard during the last decade where people have mainly used two different approaches to classify various heart diseases. In the first approach, they extracted features manually from the ECG waves also known as "expert features" and applied traditional machine learning models to these extracted features for classifying various heart arrhythmia. This method works well in the case of smaller data sets. In the second method, deep neural networks are used to automatically extract the features from the given data and detect diseases based on those "deep features". [39] Unlike the first approach, this method is well suited for a comparatively larger data set. The second approach produce better results in most cases as it uses deep neural networks, and due to their excellent learning capabilities, they are capable of extracting more robust features as compared to the expert features. [13]

1.3 Problem Statement

Despite a lot of efforts to automate ECG analysis, it has not been possible to implement it on a practical level in hospitals and clinics. Among many factors that have contributed to the failure of practical implementation of automated ECG classification, the main reason is that the data sets being used in these studies came from a single source and origin. Therefore, these small and origin-oriented data sets over fitted and failed to produce good results on the unseen data sets. Moreover, models were trained on a limited number of heart diseases, ignoring a significant number of abnormalities. As a result, the scope was only limited to a few diseases. In reality, ECG analysis is a much more complex process, therefore these attempts are not the actual representation of the complexity of ECG examination. [1] Also, a lot of those works were only focused on predicting single-labeled diseases, but in reality, one ECG can produce multi-labeled results. I.e. predict multiple heart abnormalities for a single person. Lastly, these days hospitals perform 12 lead ECG analyses but most of the data sets used in those studies used one lead ECG results.

In recent developments, a lot of ECG data sets collected from different parts of the World have been published publicly as a part of various competitions encouraging people to participate and submit their findings and evaluations regarding the data sets provided. In 2017, Physio Net Challenge published a single lead data set comprising a total of 12,186 recordings ranging from 6 to 61 seconds. 8,528 out of those were dedicated for the training purpose and 3,658 were kept hidden for testing. This challenge aimed at classifying only four heart conditions including normal heart functioning. [7] Later in 2018, China Physiological Signal Challenge (CPSC) released data of 9,458 patients from eleven different hospitals consisting of a total of 9,831 unique electrocardiograms with time interval lengths varying from 7 to 60 minutes. Physiological Signal Challenge provided 9 labels for the classification of this data set out of which 8 were cardiac diseases and the remaining one was labeled as normal heart rhythm. [24] Another well-known data set called MIT-BIH containing single lead ECG data has also been used widely to classify normal heart conditions from abnormal ones. [2]

All these data sets had limitations in one or the other way. For instance, the MIT-BIH arrhythmia data set performed only single-labeled binary classification for the 1 lead ECG data. Physio Net Challenge 2017 also used 1 lead ECG data to classify Atrial Fibrillation

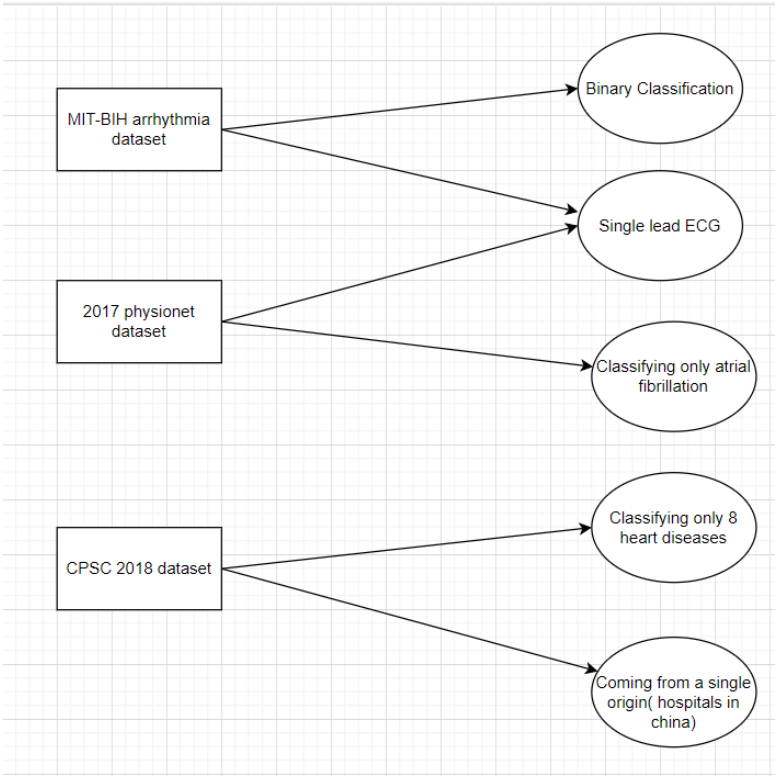

Figure 1.2: Data Sets and their Limitations.

mainly along with three other heart conditions. China Physiological Signal Challenge 2018 data set performed multi-labeled classification for 9 heart rhythms but the data came from the same origin. All of these shortcomings are visualized in Figure 1.2.

2020 has proved to be the year of breakthrough when Physio Net finally released multiple data sets containing a large collection of 43,101 twelve Lead ECG recordings as training data from four different countries. It is a multi-labeled data set aiming to classify 27 heart conditions where one ECG can predict more than one type of rhythm. [13] Physio Net 2020 data set contains 6 data sets for training among which Physikalisch-Technische Bundesanstalt (PTB-XL) data set contains 21,837 ECG recordings, Georgia data set contains 10,344 ECG recordings, China Physiological Signal Challenge data set contains 6877 ECG recordings, China Physiological Signal Challenge Extra data set contains 3453 ECG recordings, Physikalisch-Technische Bundesanstalt (PTB) data set contains 516 ECG recordings, St. Petersburg data set contains 74 ECG recordings. [1]

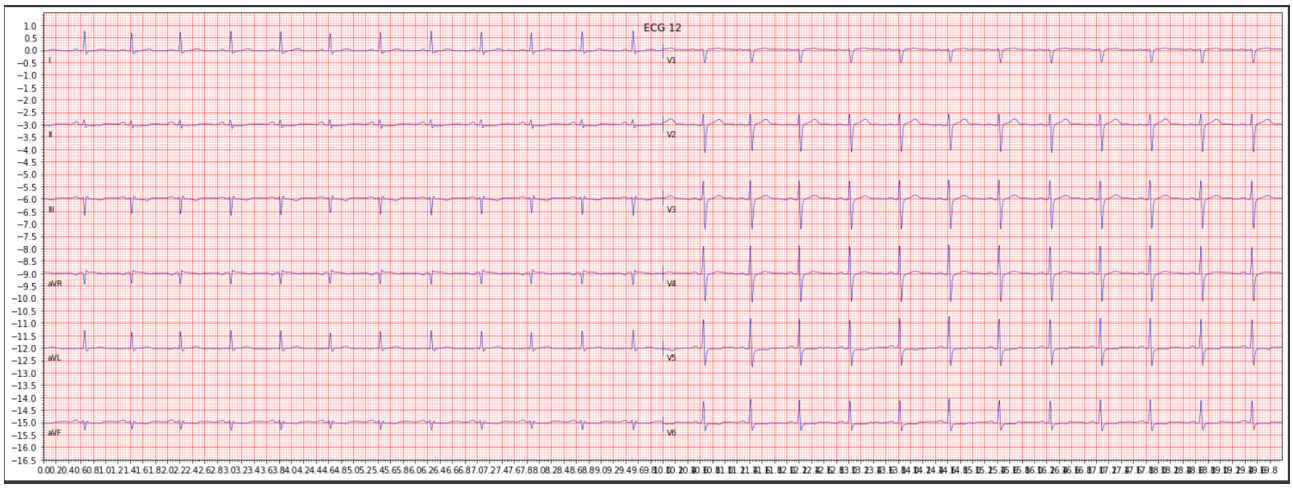

Figure 1.3: 12 Lead ECG.

1.4 Project Scope

In this paper, we are proposing a new ensemble approach attempting to perform multi-labeled classification on the Physio Net 2020 ECG data set and obtaining better results than that of already existing models. We are attempting to overcome the shortcomings of the previous works as we are using a heterogeneous data set coming from multiple sources. Also, we are classifying 27 diseases while previous works only classified a maximum of 9 heart conditions. As a baseline, we will train a Residual Neural Network with 50 layers on 12 Lead data. Then we will train the same model using only 6 chest leads data. For the third baseline model, a Residual Neural Network will be trained on 6 limb leads data. Finally, we will ensemble all the three models by assigning them different weights to see if the accuracy improves for different combinations of weights. Accuracy and AUC-ROC will be used as evaluation metrics as the data set is highly imbalanced and accuracy alone is not enough to justify the prediction results. Finally, we will compare our proposed model's results with the results in [21]. To further validate the results, they will be compared with the official results of the competition.

CHAPTER 2

Literature Review

This section discuss the work carried out by various researchers in the past few years.

2.1 Review on the Data Sets used for ECG Classification

[37] used MIT-BIH data set for the classification of heart diseases. Authors applied high as well as low Butter worth band pass filter for the removal of noise from the data. They extracted the desired information from the ECG signals at a frequency range of 0.5 to 40Hz. As the ECG signals for different ECG recordings vary in length, therefore a variable approach was used to collect segments of useful information from the electrocardiograms. Variable segmentation helped in bringing all the ECG signals to a fixed length of 300 samples. To deal with the imbalance in the MIT-BIH data set, the authors used generative adversarial networks (GAN) technique. This technique used two components. I.e Generator and discriminator. The former was used to generate new samples of data from the original data, and the latter was tuned on the generated data to distinguish real samples from fake ones. For classification, they used two distinct approaches based on the Convolutional Neural Networks. The first approach simply feeds the CNN with the ECG data and outputs the classification results. The second approach was based on a two-step hierarchy. Initially, it classifies the heartbeats and then as a final step, diseases are classified based on the results from the first step. They received an overall accuracy of 98.3 percent and 98 percent from both models respectively. Although a considerable percentage of accuracy was obtained in this paper, there are still

a few drawbacks to be discussed. For instance, the GAN method used in this paper is an unsupervised technique that can generate distorted samples resulting in outliers. Moreover, this paper has used a single lead ECG data therefore it cannot be used for clinical purposes as 12 Lead ECGs are more detailed and comprehensive. Also, single labels were predicted in this study, whereas in reality, one ECG reading can be classified into multiple diseases.

Authors in [15] also used MIT-BIH data set to classify heart arrhythmias. To pre-process the ECGs of variable lengths, they cropped them to a certain length, and for the samples below that threshold, they padded zeroes. For the training process, this paper used a Convolutional Neural Network (CNN). 1-Dimensional signals were fed to the network for training. Overall, their model achieved an accuracy of 93.4 percent. This approach also has certain drawbacks. First of all, they didn't deal with the data imbalance, resulting in the over-fitting of the trained model. Like [37], multi-labeled classification was not performed, therefore their model doesn't generalize well on other multi-labeled data sets, and is highly impractical for clinical purposes.

In [45], authors classified four heart conditions. I.e. Atrial fibrillation (AF), noisy recordings, other diseases, and normal heart conditions using the 2017 Physio Net Competition data set collected using a single ECG lead. This data set is one step ahead of the MIT-BIH data set as it was only providing binary labels for the classification whereas this data set has four labels for the classification. The data set used in this paper has a fixed sampling rate but ECG signal length varies for different samples. Like MIT-BIH data set, classes in this data set are also imbalanced. The authors split the training data to validate their model results as the test data was not made public by the competition. Instead of using traditional machine learning models, the authors of this paper used a 21-layer Neural Network that is capable of automatically extracting the features and performing classification based on those features. Their model was made up of 16 Convolutional layers, 3 and 2 recurrent and full layers respectively. The proposed model gave an accuracy of 87 percent on the validation data set but failed to produce the same results on the hidden test data set. This happened due to the poor generalization of the model as training data was limited. Therefore, a large and heterogeneous data set is needed for better generalization and to avoid over-fitting.

A 2018 China Physiological Signal Challenge data set was used by the authors in [25]. This data set contains 6,877 12 lead ECG recordings for classification. Unlike the Physio Net 2017 data set which only contained atrial fibrillation and "others" diseases for classification,

this data set contains labels for the actual diseases instead of marking them as "others". A total of nine labels are included in the data set out of which eight are diseases and one is a normal heart condition. This paper used deep features, expert features and a combination of both to train the proposed Convolutional Neural Network of 17 layers. The proposed model performed exceptionally well on the combination of both types of features. Performance on deep features alone was also acceptable but expert features didn't produce good results when passed solely to the model. The reason for the bad performance of expert features is that they didn't filter the noise from the data for the fear of losing important information. A lot of shortcomings in the MIT-BIH and Physio Net 2017 data set were eliminated with the introduction of the 2018 China Physiological Signal Challenge data set, but still, a few things are to be considered such as data set coming from only one origin (China) could result in over-fitting and fail to generalize on ECG data containing different demographics and coming from other parts of the World.

A new multi-labeled data set was introduced by physio net in 2020 to address the limitations of MIT-BIH, China Physiological Signal Challenge 2020 and Physio Net 2017 data sets. Authors in [47] attempted to classify 27 diseases using an ensemble model based on SE-Residual Network while using cross-entropy as a loss function. Their model managed to generalize well on the unseen data and produced a 68 percent metric score on the validation data set.

2.2 Traditional Machine Learning Methods applied in ECG Classification

[4] used MIT-BIH arrhythmia to perform binary classification, identifying the normal heartbeat from the abnormal. The authors used various filters to remove noise from the data followed by peak detection. Once the peaks were detected, various features were extracted based on these peaks. Based on the features extracted, various machine learning models were trained among which Naive Bayes produced the best results with an accuracy of 99.7 percent. Despite achieving excellent results, they could have applied a more advanced machine learning technique to extract the deep features as they are more robust in nature as compared to the hand extracted features fed to the machine learning model for the classification of ECG signals.

Authors of [26] took the challenge of distinguishing Atrial fibrillation rhythms from non-atrial fibrillation rhythms. For that purpose, they used a well-known machine learning approach known as Random Forest Classifier. As a pre-processing step, they applied the band-pass filter to remove noise from the data and performed down-sampling on the original ECG signals to decrease the computing time while features extraction. In second step, the Q R S complex was detected, and R R intervals were extracted using Pan-Tompkins. Following feature extraction, feature selection was performed and redundant features were removed to reduce the feature space. Finally, these features were passed to a Random Forest Classifier. To avoid the problem of over-fitting, bootstrapping was performed. This paper claims to achieve an F1 score of 0.78 on the testing data. Although this paper ensured that over-fitting doesn't take place by using bootstrapping but as mentioned earlier, for large data sets like ECG, deep learning approaches tend to produce more good results.

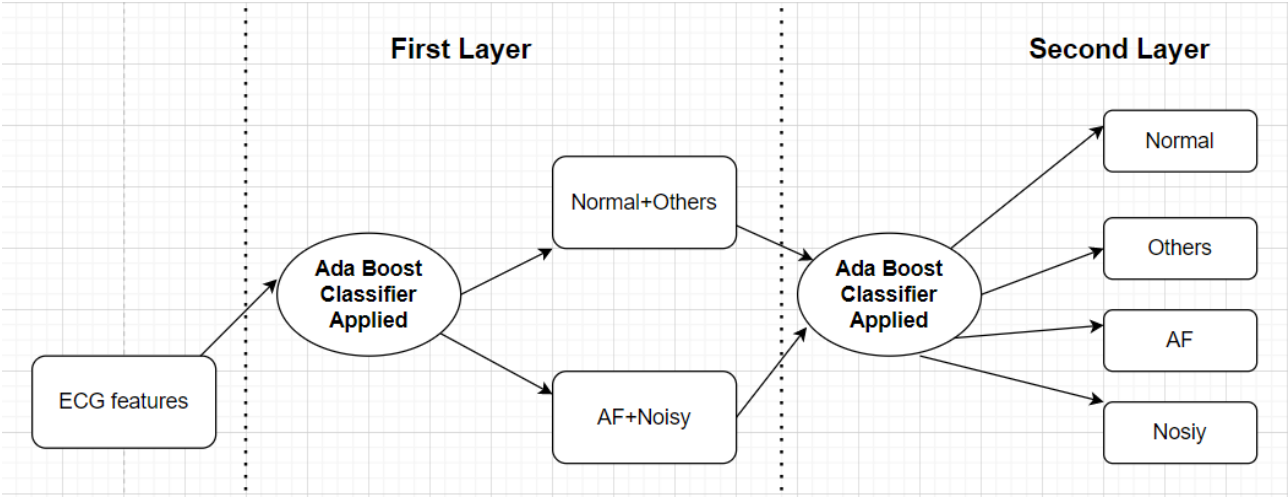

Figure 2.1: Multi-layered Ada Boost Classifier.

Physio Net 2017 Challenge data set was utilized by the authors of [8]. Unlike [45], which used deep learning for feature extraction and for classification, this paper used machine learning methods. A spectrogram was used to remove the noise from the data. Later on, features were extracted from the filtered data and finally feature selection was performed before the classification process. Ada Boost Classifier with two layers was used for classification. In the first layer, classes were classified into pairs of two, followed by the second layer which further classified those pairs to predict individual diseases. Figure 2.1 Although they managed to receive the F1 scores of 91, 80, 77, and 83 for the four classes respectively and grabbed the first spot in the competition, some of the samples from the 'normal' class

were predicted as 'other' by their classifier. One of the reasons it may have happened is to the absence of actual information about the diseases.

2.3 Deep learning methods applied in ECG Classification

[46] used the MIT-BIH data set to predict heart arrhythmias. For data denoising, they used the Daubechies wavelet 6 filter also known as "db6". The authors in this article used the Convolutional Neural Networks with cross entropy as a loss function. Their CNN model comprised 9 layers including convolutional, sub-sampling, fully connected and softmax layers. Stratification was applied while splitting the data set into training and testing to ensure the same level of data imbalance in both sets. Training data was further split into two for validation purposes. Like [15], they also passed one-dimensional signals to their network. Their model managed to obtain an accuracy of 99.3 percent on the test set but their way of splitting the test data was not very convincing as there is a possibility that duplicates exist in the data set, therefore this splitting approach can result in over-fitting the model and as a result produce bad results on the unseen data.

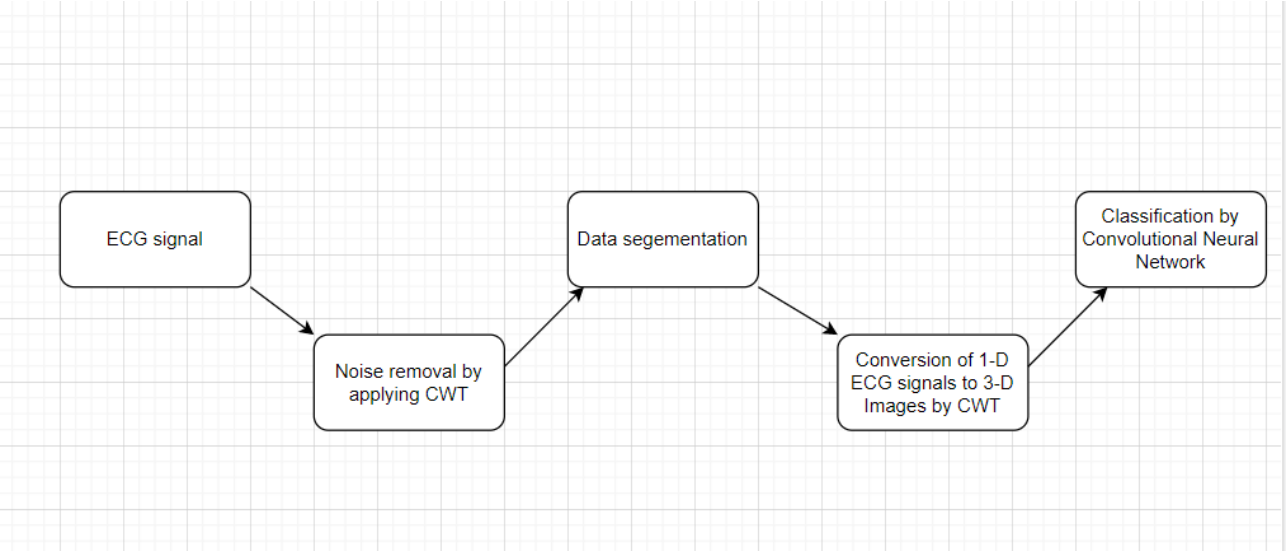

Figure 2.2: Continuous Wavelet Transformation on one-dimensional ECG data.

In another research, computer scientists used Continuous Wavelet transformation for removing the noise from the data and then creating data segments where each segment constitutes 1.2 seconds. As a post-processing step, they again applied Continuous Wavelet Transformation (CWT) and converted one-dimensional signals into three-dimensional pat-

terns. Finally, 3-dimensional images were passed to a Convolutional Neural Network (CNN) for classification. Figure 2.2 The overall accuracy of 99.23 percent was achieved while testing. In this paper, they used an under-sampling technique to address the issue of imbalanced data. The problem with this technique is that it can result in the loss of very important data samples as data is randomly under-sampled without any defined criteria. [11]

[21] performed multi-labeled classification on the China Physiological Signal Challenge data set. For that purpose, the noise was filtered from the data using a low pass band filter of level 8. To encounter the problem of varying signal length, a very unique approach was used that had not been used earlier. While truncating the signals to a fixed length, a block framing technique was used that detects the overlapping frames and places them together to ensure the smooth continuation of the signals. A random under-sampling technique was used based on the single unique labels as well as the combination of multiple labels to get rid of the data imbalance. For classification, a Residual Neural Network with 13 layers was used. The proposed model obtained an F1 score of 90.8 and an AUC-RUC score of 97.8. The data balancing technique used in this paper can result in losing a lot of important data samples as the majority class is being down-sampled to achieve the balance.

2.4 Improved (Ensemble) Artificial Intelligence Techniques applied in ECG Classification

[38] proposed a 20-layer model based on the combination of CNN and Bi-directional Long short-term memory (LSTM). It uses a level two Daubechies wavelet filter to remove noise from the data. To handle the data imbalance, SMOTE over-sampling technique was used. This technique creates new data samples from the actual data for the minority class using K-nearest neighbour. The original data set was divided into three parts with validation having 90 percent of the data, and training and testing parts having 10 percent each. The model was trained to predict five different types of heart rhythms. When tested on the test data, the proposed model achieved an overall accuracy of 98.71 percent. Although this paper proposed a new ensemble approach for classification but the technique they used for balancing the data comes with a few drawbacks. SMOTE technique creates new samples of the minority class based on the neighbouring samples of that particular class. Therefore, in case of high imbalance, data is highly skewed with the minority class dispersed in the space.

This can result in mixing both classes and creating redundant samples. [19]

[31] used a combination of machine learning and deep learning to classify the four labels in the 2017 Physio Net competition. Data provided in this competition is noisy due to multiple reasons such as muscle movements, electrodes not plugged properly and malfunctions in the ECG recording devices. Q R S complex was detected to extract features from the ECG signals. Two machine learning approaches were used in parallel for classification purposes. For the first approach, 43 extracted features were passed to the Bagged Tree Ensemble model for classification. For the second approach, ECG signals were filtered using nine different filters and this filtered data was passed to a model which was a combination of deep and shallow neural networks. Results from the second approach were used if they reached a certain threshold of accuracy otherwise outputs from the Bagged Ensemble model were used. This ensemble approach obtained an overall F1 score of 0.83 on the test data set. The work in this paper was ranked number two in the competition's official results.

In [5], a Convolutional Neural Network was built using five convolutional layers attached to a bi-directional Gated Recurrent Unit. This paper used a very unique technique to classify the data from CPSC 2018. The data set was stratified and divided into ten folds, followed by a cross-validation process run multiple times. Repeated cross-validations produced multiple trained models out of which the top ten models with the best results were selected. To further investigate the ECG data, 12 one-lead ECG data sets were produced from the 12-lead ECG data and again ten-fold cross-validation was performed. This resulted in 120 trained models. In the final stage, all the 130 trained models were ensemble to produce the best classification results. This model won the competition based on performance measures.

[36] used this physio net 2020 data set to perform multi-labeled classification. For that purpose, they introduced an ensemble approach. During pre-processing, the data was re-sampled at 4000Hz and ECG signals exceeding the fixed threshold of 4096 were split into batches to feed the ECG signals of a constant length to the Network. Multiple convolutional layers and residual blocks were used to construct the model. The model was trained on 200 epochs and the weights during the training were optimized using the Adam optimizer. The proposed Residual Network was trained seven times on the processed data with different randomly initialized weights every time. The results from the individual models were combined to produce the ensemble results. The motivation behind this approach was to address the problem of local minima. The ensemble approach resulted in better generalization

as compared to the individual models. The proposed model achieved a metric score of 13.2 percent on the test data and was ranked 28 in the competition.

In [30], authors used machine learning along with deep learning to perform classification. For pre-processing, the ECG recordings were down-sampled to 500Hz as these recordings came from different sources and therefore had different frequencies. With a bandwidth of 3 to 45 Hz, FIR band-pass filter was applied for removing noise from the ECG data. Using a fixed threshold of 15 seconds, zeros were padded to the signals if their length was less than the fixed threshold. Initially, 300 features were extracted from the lead II and a random forest classifier was applied to select the best performing features. It gave top 20 features as expert features. These expert features along with the demo-graphical (age and sex) features were fed to the deep neural network. Authors excluded the 81 unscored diseases from the list of labels and used only 27 scored diseases as labels for classification. The iterative stratification technique was applied while splitting the data into training and testing. The benefit of using this technique is that it ensures that the level of imbalance in the original data set is also maintained in the data folds based on individual labels as well as the combinations of various labels. The proposed model took 88 hours to train on the provided data set and produced an overall competition score of 58.7 percent on the hidden test data set.

[42] trained a modified version of Residual Neural Networks on the physio net competition data set using global skip connections and custom loss function. After performing all the necessary pre-processing steps, data was passed to a Residual Model with one-dimensional filters. At the end of each epoch, weights were updated to train a new model. All these models were ensemble using boot-strapping for better generalization. A total of 24 diseases were classified in this paper. The model was able to obtain a score of 69.9 percent on the validation data set and 0.202 on the test data set.

2.5 Proposed Ensemble Approach for the Classification of ECG

In this paper, we are attempting to perform multi-label classification on the physio Net 2020 competition data. We are building three baseline models using all 12 Leads ECG data, chest leads data, and limb leads data. As pre-processing steps, we are truncating all ECG signals to a fixed length of 5000. To deal with data imbalance, we are assigning weights to the

labels that are inversely proportional to the number of samples of respective labels in the data set. To split data into testing and training data, the iterative split approach is used to ensure the same ratio of data imbalance in the split data sets as present in the original data set. For noise removal, the butter low-pass band filter of level 2 is used with a sampling frequency of 257 Hz and a signal frequency of maximum of 30. Demo-graphical features are encoded before passing them to the Residual Neural Network. Labels are also one-hot encoded to create an array of size 27 for the labels containing 1's and 0's. To train the baseline models, we are using a Residual Neural Network of 50 layers with skip connections and binary cross-entropy loss function. The reason for using RNNs in our project is that they deal well with the problem of vanishing gradient in the models with a higher number of layers as compared to other Deep Neural Networks. As a final step, results from these models are ensemble by assigning weights to the models and checking which combination of weights gives the best performance on the test data. The reason for exploring the performance of different models on individual predictions by assigning weights is that various individual leads or combination of leads can predict certain diseases more efficiently as compared to other leads. Further details of the implementation are discussed in next sections.

CHAPTER 3

Methodology

3.1 Framework

For this project, we have used Tensor Flow and Keras frameworks. The reason we used them is that we are using a deep learning approach for the classification, and they are the most popular frameworks used in deep learning as compared to other frameworks such as MLTK, py-torch, etc. [35] Unlike tensor flow which is a fully-fledged framework, Keras is just a wrapper around other frameworks. Keras's main role is to provide a convenient and easy-to-use Application Program Interface (API) for tensor flow and several other frameworks. Figure 3.1

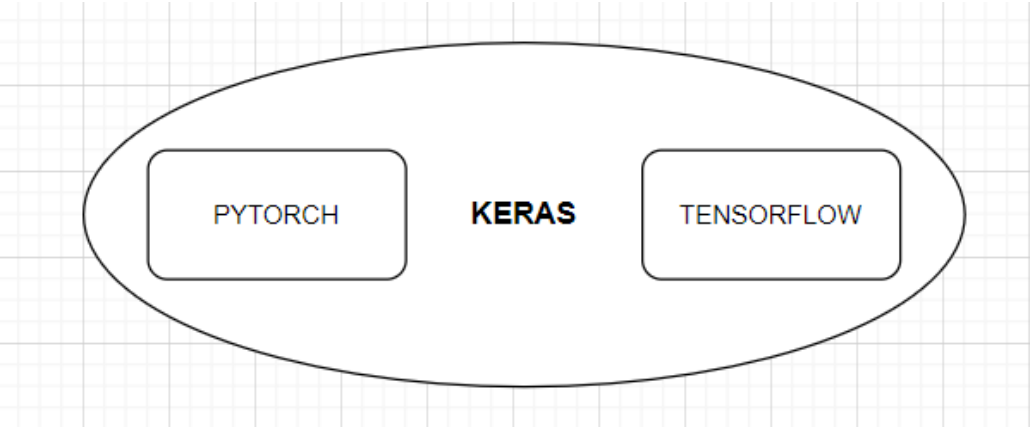

Figure 3.1: Keras Application User Interface for various frame works

3.2 Data Set

Physio Net 2020 data set is a collection of 43,101 ECG recordings of multiple patients coming from various hospitals in four different Countries. Each ECG recording consists of MATLAB and a WFDB file. MATLAB file contains the data for ECG signals whereas the WFDB file provides details about the patient in text format. It provides information about the patient ID, the number of leads used in the ECG recording, sampling frequency, number of data samples present in the ECG signal, date and time of the ECG recording, age and gender of the patient, medical history, pescription details, symptoms, individual details of the ECG signals from every lead and finally the diagnostic labels. [1]

```
1  E04001 12 500 5000
2  E04001.mat 16x1+24 1000.0(0)/mV 16 0 -107 -18586 0 I
3  E04001.mat 16x1+24 1000.0(0)/mV 16 0 19 24627 0 II
4  E04001.mat 16x1+24 1000.0(0)/mV 16 0 126 -21675 0 III
5  E04001.mat 16x1+24 1000.0(0)/mV 16 0 43 29670 0 aVR
6  E04001.mat 16x1+24 1000.0(0)/mV 16 0 -117 1621 0 aVL
7  E04001.mat 16x1+24 1000.0(0)/mV 16 0 73 -31320 0 aVF
8  E04001.mat 16x1+24 1000.0(0)/mV 16 0 -19 7724 0 V1
9  E04001.mat 16x1+24 1000.0(0)/mV 16 0 -87 -29422 0 V2
10 E04001.mat 16x1+24 1000.0(0)/mV 16 0 -9 4798 0 V3
11 E04001.mat 16x1+24 1000.0(0)/mV 16 0 -58 -14355 0 V4
12 E04001.mat 16x1+24 1000.0(0)/mV 16 0 9 3548 0 V5
13 E04001.mat 16x1+24 1000.0(0)/mV 16 0 -68 -11919 0 V6
14 # Age: 51
15 # Sex: Female
16 # Dx: 111975006,164930006,270492004
17 # Rx: Unknown
18 # Hx: Unknown
```

Figure 3.2: Header file containing patient details.

Figure 3.2 contains the header file for ECG recording E04001 where the first line indicates that it's a 12-Lead ECG recorded at a frequency of 500 Hz with data samples of 5000. The next 12 lines contains the details of all ECG leads including their names, offset, checksum and amplitude resolution. The remaining lines in the file indicate that the patient is a 51 years old female diagnosed with three heart diseases (Dx). Her pescription details (RX) and history (Hx) are Unknown.

As the ECG data in Physio Net Competition came from multiple sources, therefore it was

sampled at variable frequencies and has a different signal and time lengths. Six data sets introduced in the Competition are summarized in Table 3.1.

3.2.1 China Physiological Signal Challenge (CPSC) Data Set

CPSC data set was collected from various hospitals in China. ECG recordings in this data set were sampled at a frequency of 500 Hz and lasted between 6 to 60 seconds. This data set has a total of 6,877 recordings where 53.8 percent (3,699) belong to the male population and the rest of the 46.2 percent (3,187) were recorded from the female patients. [1]

3.2.2 China Physiological Signal Challenge Extra (CPSC-Extra) Data Set

Like the CPSC data set, this data set also originates from China. It has the same sampling frequency as of CPSC data set. The average duration for the ECG recordings in the data set is approximately 15.9 seconds. The gender distribution in the CPSC-Extra data set is 53.4 percent (1,843) males and 46.6 percent (1,610) females. CPSC and CPSC-Extra data sets combined contain data of 9,458 patients and a total of 10,330 ECG recordings. [1]

Data Set	Origin	Length(seconds)	Frequency(Hz)	Recordings	Male	Female
CPSC	China	6 to 60	500	6,877	3,699	3,178
CPSC-Extra	China	6 to 60	500	6,453	1,843	1,610
INCART	Russia	1800	257	74	40	34
PTB	Germany	10	500	516	377	139
PTB-XL	Germany	10	500	21,837	11,379	10,458
Georgia	USA	10	500	10,344	5,551	4,793

Table 3.1: Description of Data Sets

3.2.3 INCART Data Set

This data set was collected from the second largest city in Russia known as "St. Petersburg". It contains 74 ECG recordings for 32 patients sampled at a frequency of 257 Hz. The recordings are approximately 1800 seconds long. 54.1 percent (40) recordings were collected from males whereas 45.9 percent (34) recordings belong to females. Table 3.1

3.2.4 Physikalisch Technische Bundesanstalt (PTB) Data Set

Physikalisch Technische Bundesanstalt Data Set is named after the National Metrology Institute of Germany. The data set has ECG recordings lasting for a maximum of ten seconds. PTB data set sampled at 500 Hz frequency has a total of 516 ECG recordings with a ratio of 73.1 percent (377) recordings belonging to males and the rest of 26.7 percent (139) collected from females. [1]

3.2.5 Physikalisch Technische Bundesanstalt Extra Large (PTB-XL) Data Set

This data set was also collected from Germany but as the name indicates it's a comparatively larger data set than the PTB data set. As it comes from the same source, therefore it has same the ECG recordings length and frequency as the PTB data set. Out of the total 21,837 recordings, more than 50 percent (11,379) recordings were collected from male patients while the remaining 10,458 belong to female patients. PTB and PTB-XL data sets as a whole contain 22,353 recordings from 19,175 patients. Table 3.1

3.2.6 Georgia Data Set

The last data set used in this research comes from the city of Georgia in the United States of America. 10,344 ECG recordings were sampled at a rate of 500 Hz frequency from 7,871 patients. Among these recordings, 53.9 percent (5,551) were from male patients and 46.1 percent (4,793) were from female patients. Table 3.1

3.3 Class Imbalance

Figure 3.3 shows that the data used in this project is highly imbalanced. One of the widely used approaches to deal with data imbalance is the modification of the actual data set. In this modification technique, the data is re-sampled by down-sampling the majority class or up-sampling the minority class but these approaches come with problems. In down-sampling, the instances from the majority class are deleted to achieve balance which can result in losing important information whereas, during over-sampling, instances of the minority class are

simply copied randomly to increase the sample size of the minority class. This can result in over-fitting and can increase computational costs. [23] Another popular approach used to balance the data set is the SMOTE over-sampling technique. It uses the K-nearest neighbour to generate new instances for the minority class. The drawback to this approach is that it can mix both classes due to the minority class dispersed in the space. As a result of highly skewed data, redundant samples are created for the minority class. [19] To deal with the data imbalance in our project, we have used a very unique and novel approach that simply assigns weights to the classes while training. These weights are inversely proportional to the frequency of the classes in the data set where the minority class has the highest weight and majority class has the the lowest weight assigned to them. By using this technique, we have avoided the problem of increased computational costs, over-fitting and loss of important information.

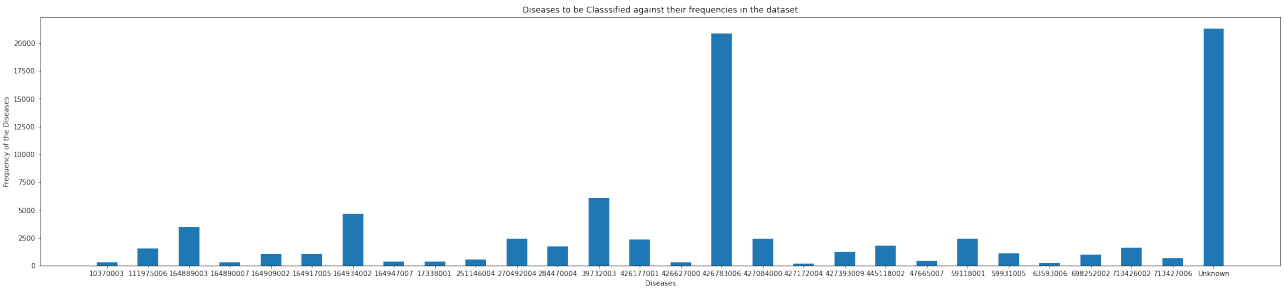

Figure 3.3: Class Frequencies in the Data Set.

3.4 Noise Removal

Figure 3.4 indicates that the data collected from the ECG contains a lot of noise. This noise can be due to multiple reasons such as muscle movements, distortion in the ECG gadgets and electrodes not being placed properly on the body while recording the ECG. Noisy data can result in a false diagnosis of the disease as it can change the morphology of the ECG signals creating a hindrance in extracting good quality features. Various wavelet, deep learning and sparsity-based techniques can be used to remove the noise from the ECG signals. In this paper, we have used the butter worth filter of order 2 with a low-passing technique. Butterworth low-pass filter only allows the signals that have a value less than the cut-off frequency whereas signals greater than the cut-off are weakened resulting in reducing the

noise in the signals.

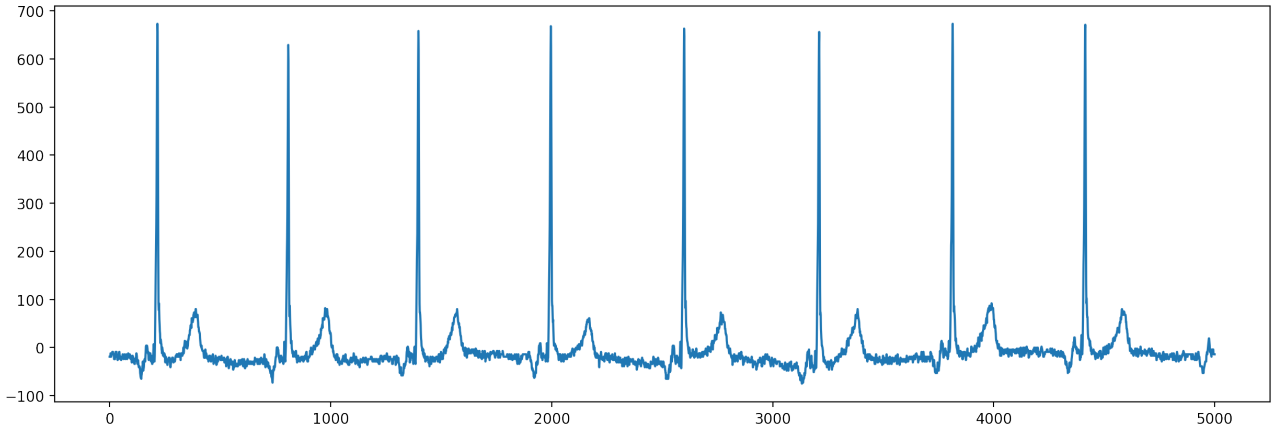

Figure 3.4: Noisy signal from ECG.

3.5 Features Extraction

Among the two popular approaches used for feature extraction in classification problems, we have used the deep learning approach. Unlike the machine learning approach where we extract the features manually by the peak detection from the Q R S complex, this approach is capable of extracting the features automatically from the given data. The reason for using the deep learning approach is because we have a large data set of 43,101 ECG recordings, and therefore due to its good learning capability, the deep learning model will be able to extract more robust features in our case as compared to hand-crafted human expert features. [13]

3.6 Evaluation Metrics

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + FalsePositives + TrueNegatives + FalseNegatives} \quad (3.6.1)$$

To evaluate our models, we are using two different metrics. i.e. Classification and Area Under the Curve Receiver Operator Characteristic (AUC-ROC). Classification refers to the ratio of the correct predictions to the total predictions by the model. [3] Sum of the correct predictions is obtained by adding the number of True positives and true negatives whereas the sum of the overall predictions is obtained by adding the true positives, true negatives, false positives and false negatives. (Eq. 3.6.1) True positives are the labels that are positive in

the actual and were also predicted as positive by the model whereas false positives are the examples that are negative but predicted as positive. True negatives are the instances that are negative in both predictions as well as actual data set whereas false negatives are the positive examples classified as negative. Figure 3.5

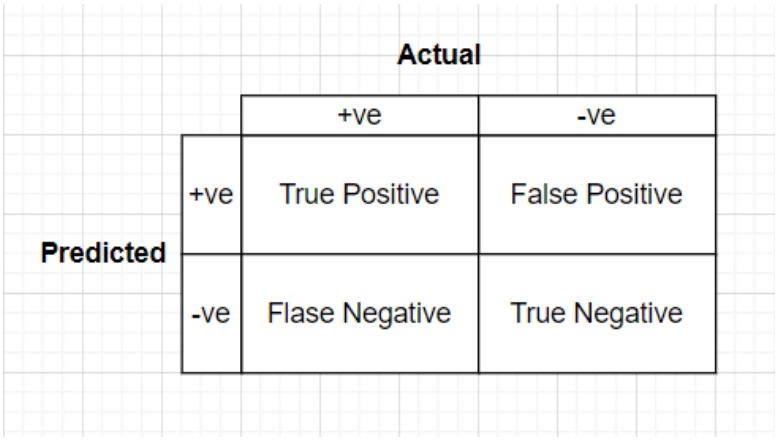

Figure 3.5: Confusion Matrix.

In our case, data is highly imbalanced therefore accuracy alone cannot be used to evaluate the model as classification models trained on such data are biased towards the majority class, ignoring the minority class while training. As a result, they over-fit the majority class. [20] Classification techniques are accuracy-driven and try to reduce the combined loss, therefore they assume that the cost of mispredictions for all classes is equal but in actuality, it's not true. For instance, in the case of heart disease, one disease can be more fatal than the other one. [41] For better generalization, we have used additional evaluation criteria known as "AUC-ROC". ROC is a curve plotted between True Positive Rate and False positive rate whereas AUC refers to the area under that curve. The true positive rate refers to the ratio of correctly predicted positive instances (eq. 3.6.2) and the false positive rate refers to the ratio of incorrectly identified negative classes (eq. 3.6.3). The value of AUC is between 0 and 1 where a value closer to 1 indicates that the model is very much capable of distinguishing between the different classes. [28]

$$TruePositiveRate = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{3.6.2}$$

$$FlasePositiveRate = \frac{FalsePositives}{FalsePositives + TrueNegatives} \tag{3.6.3}$$

CHAPTER 4

Model Design

4.1 Skip Connections in Residual Neural Networks

Despite the good learning ability of deep neural networks, while updating the weights during the back-propagation in a network having a large number of layers, the problem of vanishing gradient arises. This happens due to the reason that weights are updated in every layer during back-propagation by performing multiplication operations. Due to the repeated multiplication operations, weights in the starting layers become insignificant. [12] Residual Neural Networks are deep learning models with the exception that they solve the problem of the vanishing gradient by adding a skip connection layer to the model. They do so by adding the input of residual block to the loss function directly. Figure 4.1.

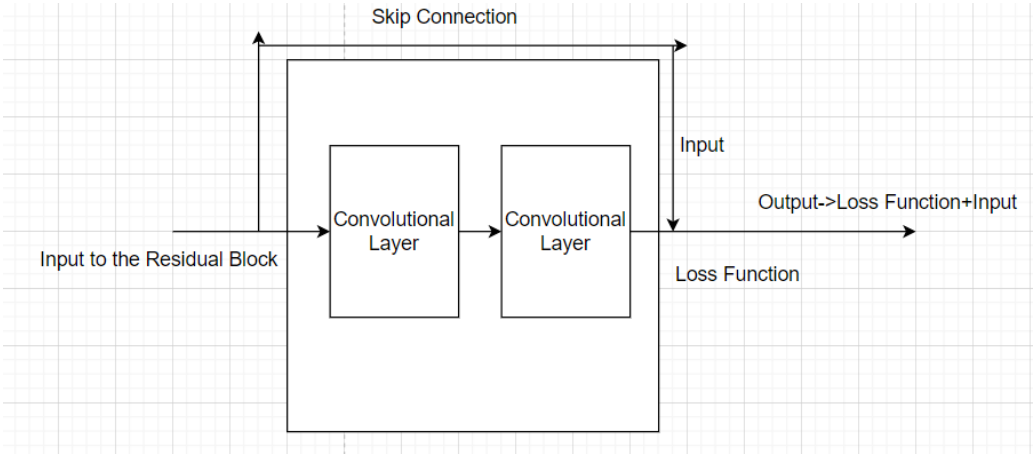

Figure 4.1: Skip Connection Block in Residual Neural Networks.

4.2 Residual Neural Network with 50 layers (RESNET-50)

For our project, we are using a residual neural network with 50 layers to train our baseline models. RESNET-50 is built up of two types of blocks. I.e. Identity block and convolutional block. Both blocks contains three convolutional layers and a skip connection. The only difference is that the convolutional block is used when the size of the input to the block is not same as the size of the output of the block. To make the size of the input and output the same in the convolutional blocks before adding them, we add a convolutional layer of kernel size 1 and padding equals "valid" to the skip connection. On the other hand, the identity block is used when the size of the input and output are the same. [34] Difference in the both blocks is visualized in Figure 4.2

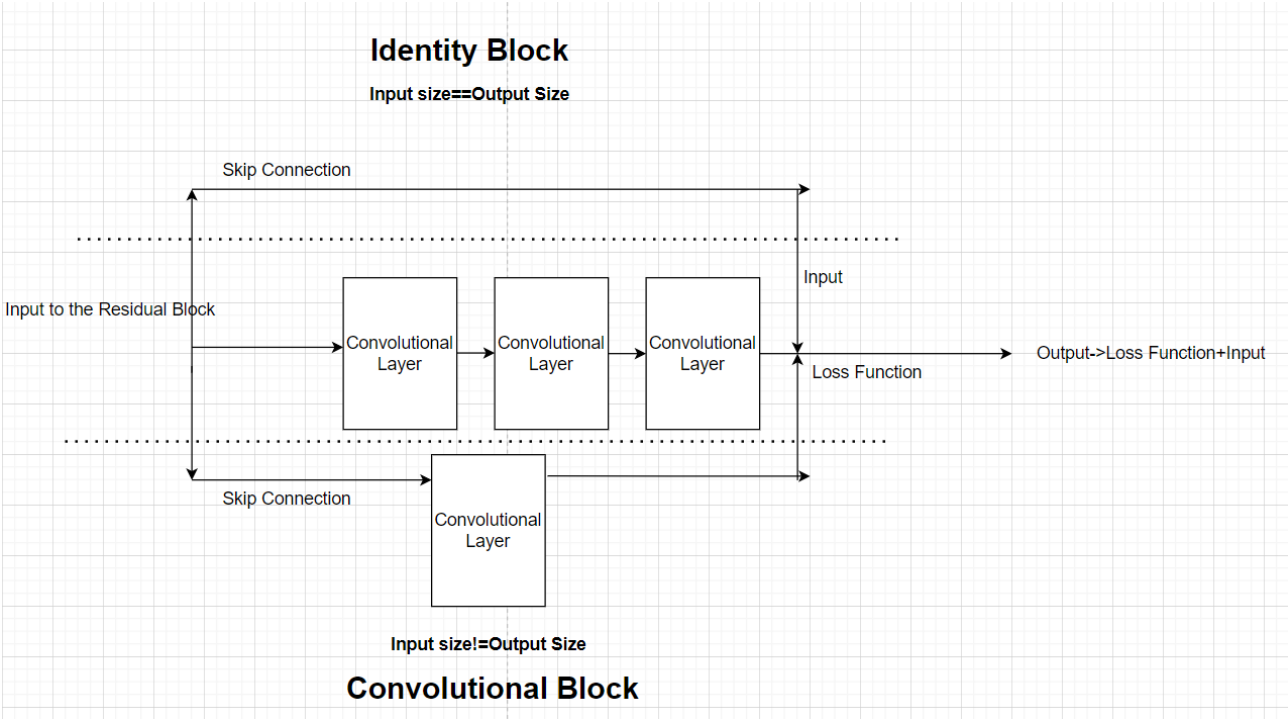

Figure 4.2: Convolutional block vs Identity block.

In the identity block, we perform batch normalization followed by the "Relu" activation function for the first two convolutional layers which converts the negative values to "zero". Whereas for the last layer, we don't apply the "Relu" function as we have to add the layer's output to the input of the block using skip connection. After adding the input and loss function, we apply the "Relu" activation function. Convolutional block also works in the same way. The only difference is that before adding the input to the output of the third layer,

the input size is changed using a convolutional layer in the skip connection. Once input and output sizes are equal, they are added and "Relu" activation is applied to the results. The benefit of using "Relu" activation is that it doesn't affect the convolutional layers while increasing the non-linearity of the network. [33]

After receiving input, zero-padding is performed to avoid the loss of information as while applying filters, dimension reduction takes place. After that input is passed to the first layer of the model which is a convolutional layer. Here, we perform batch normalization and reduce the dimension using max pooling. The first layer is followed by a combination of one convolutional block and two identity blocks. The three layers in these blocks contain filters of size 64, 64 and 256 respectively. In the third step, we have one convolutional block followed by three identity blocks. The layers of blocks in this step have filters of sizes 128, 128 and 512. A combination of one convolutional block and five identity blocks is followed whose layers have filters 256, 256 and 1024. The last combination of these blocks contains one convolutional block followed by two identity blocks with the three layers having filters 512, 512 and 2048. Once the output of the last identity block is received, average pooling is performed for the down-sampling of the connections. The last layer in the model is a fully connected layer also known as the "Dense Layer". It changes the dimensions from the output of the previous layer to the dimensions required in the final output. Figure 4.3 shows that the RESNET model developed has a total of 50 layers with each block having a total of 3 layers.

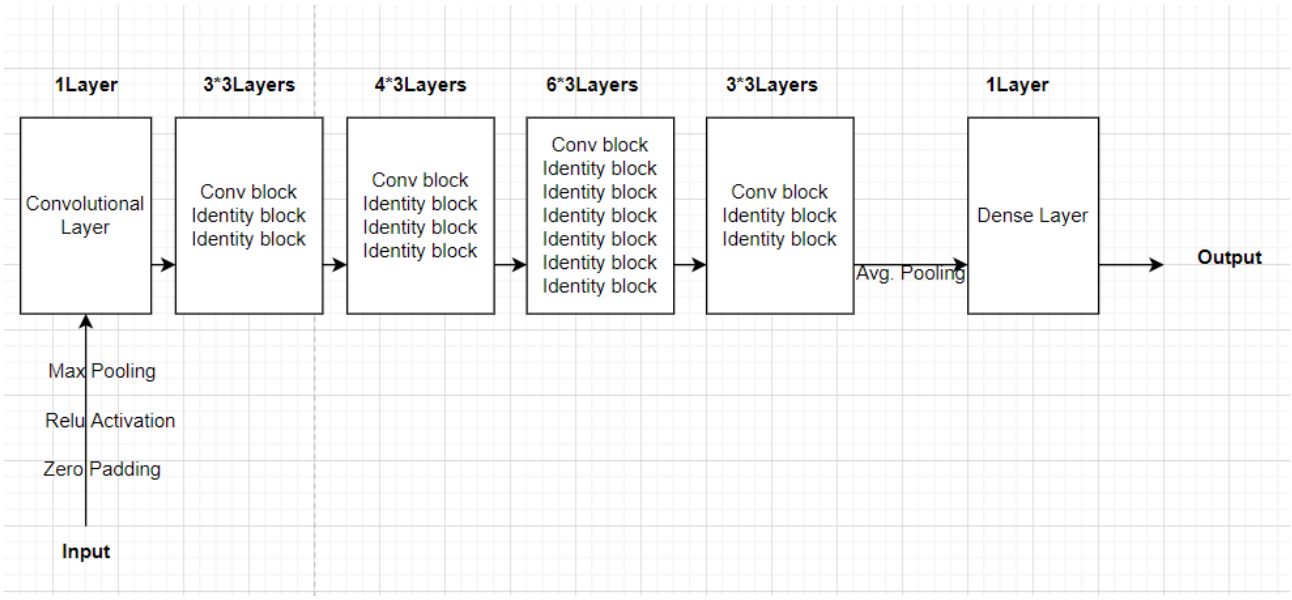

Figure 4.3: RESNET-50 Model.

In residual neural network, the output dimensions for a layer are calculated as shown in eq. 4.2.1:

$$Outputdimensions = (\frac{Inputdimension + 2(padding) - kernelsize}{strides} + 1), filters \tag{4.2.1}$$

CHAPTER 5

Implementation

As a first step during implementation, loadmat() function from scipy.io was used to read the MATLAB files containing the ECG signals data. ECG signals had variable lengths (1969 unique lengths), therefore they were truncated to the most frequently existing length of 5000 (Figure 5.1), bringing them to a constant length to feed to our model. Header files were also read and the diagnostic labels, sex, age, pescription, history, and surgery details of the patients were saved to a list. On analyzing the data, it was concluded that the surgery, history and pescription details for all the patients are unknown. Demo-graphical features (Sex and Age) were encoded. For the sex feature, males were assigned the value '1', females with value '0' and the 'NAN' values were replaced with integer '2'. The missing values in the age list were replaced with the most occurring value of age.

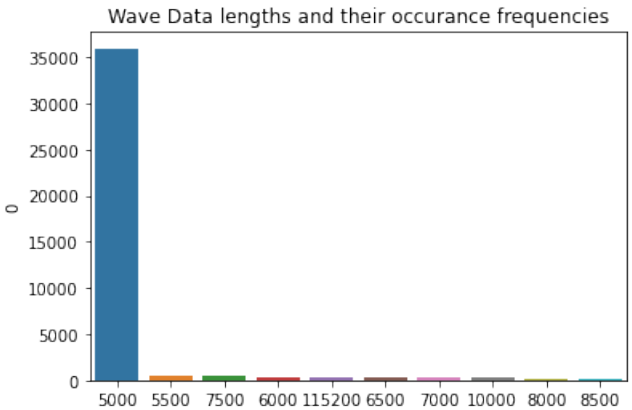

Figure 5.1: ECG signals lengths.

Butter low-pass level 2 filter with a signal frequency of 30 and a sampling frequency of 257 Hz was applied to remove the noise from the ECG signals. ECG signal after removing the noise is shown in Figure 5.2. ECG signals and demo-graphical features were combined to pass to our model.

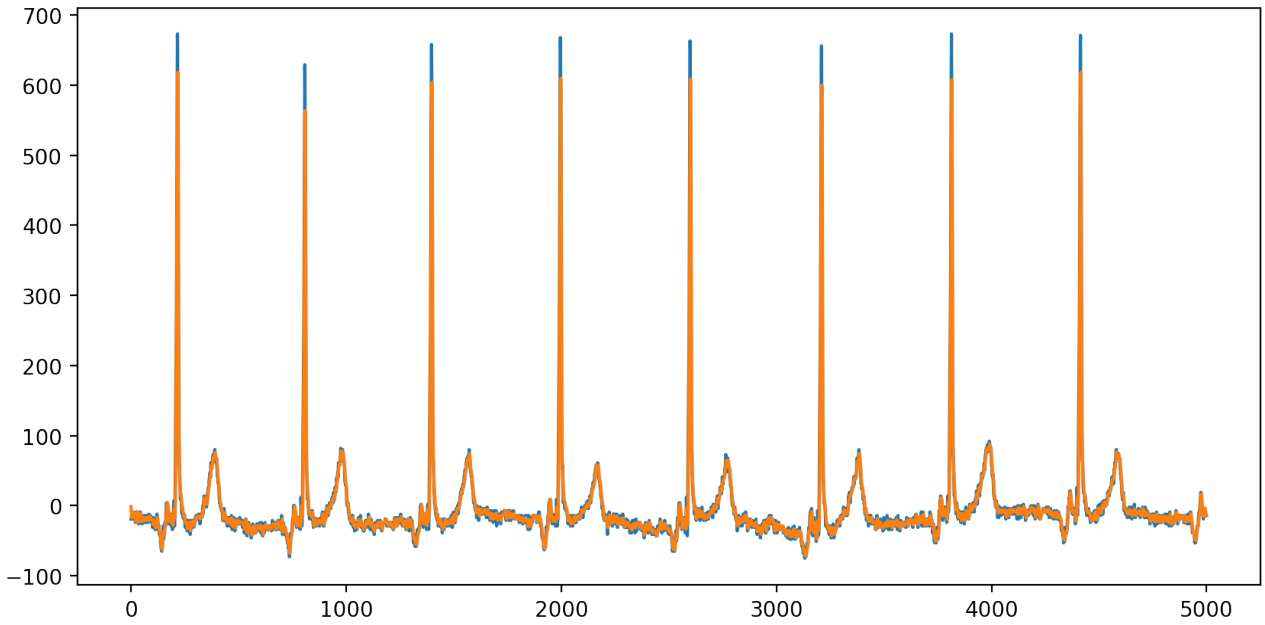

Figure 5.2: ECG signals after removing the noise.

Using the split() function, the diagnostic labels were comma-separated for the ECG recordings having multiple diagnoses and converted to type int. Labels not present in the list of 27 scored diseases were replaced with zero. Finally, the labels were one-hot encoded making it a multi-label multi-class classification problem. Data were split into test and train using an iterative split with 20 percent data going to the test set and 80 percent to training. To avoid the problem of over-fitting, weights were assigned to the labels in the training data set in such a way that the majority class gets the minimum weights during training while the minority class has the maximum weights. The frequency of diseases being classified in our project along with the weights assigned to them in the training data is shown in table 5.1.

For the RESNET-50 model, the binary cross entropy was used as a loss function as our list of multiple labels contains the binary values (0 and 1). While training, validation split was kept at 15 percent whereas 85 percent of the data was kept for training. To avoid the problem of local minima and for a more efficient gradient descent, the Adam optimizer was used and the learning rate was initialized at 0.01.

Disease	Frequency	Weights
10370003	299	72.61087866108787
111975006	1513	14.4496253122398
164889003	3475	6.242446043165468
164890007	314	69.13944223107569
164909002	1041	20.83313325330132
164917005	1013	22.24871794871795
164934002	4673	4.763656327202855
164947007	340	63.10545454545454
17338001	365	59.43150684931507
251146004	556	38.997752808988764
270492004	2394	9.062140992167102
284470004	1729	12.548083875632683
39732003	6086	3.5729874408070827
426177001	2359	9.196608373078961
426627000	288	76.1140350877193
426783006	20846	1.0405948312046531
427084000	2402	9.05741127348643
427172004	188	115.69333333333333
427393009	1240	17.493951612903224
445118002	1806	12.264310954063605
47665007	427	51.34319526627219
59118001	2402	9.02913631633715
59931005	112	19.498876404494382
63593006	215	103.91616766467065
698252002	997	22.16347381864623
713426002	1611	13.838915470494419
713427006	683	33.56673114119923
Unscored	21291	1.018845769975929

Table 5.1: Frequency and weights of diseases.

Validation loss was kept as an early stopping criterion with a patience value of 2 to avoid over-fitting. While training the model, training data along with the weights were passed to the model. The model was trained on the training data for 50 epochs in batches of size 75. The first baseline model was passed with an input of shape 12x5000 where 12 indicates the number of leads while 5000 are the data points in every lead. This model stopped training after seven epochs when the validation loss didn't improve and the fifth epoch was saved as the best model. It took approximately three hours to train the first baseline model. For the second model, ECG data for only limb leads were passed with the input having a shape of 6x5000 where 6 is the number of limb leads. Model 2 trained for six epochs with the fourth epoch saved as the best model, and took approximately 2.7 hours to train. The third baseline model was only trained on the six chest leads having an input shape of 6x5000 and it took 2 hours to train. The baseline models were trained with a total of 18,330,844 parameters (18,277,724 trainable and 53,120 non-trainable parameters). After training the three baseline models and saving them in ".h5" format, they were again loaded to make predictions on the test data. Predictions for all the models on test data were passed to an ensemble function which assigned weights ranging between (0 and 1) to the models predictions and iterated through all the possible combinations of weights for the models to output the combination giving the highest accuracy results in the end. Implementation flow is visualized in figure 5.3 for better understanding of the Project.

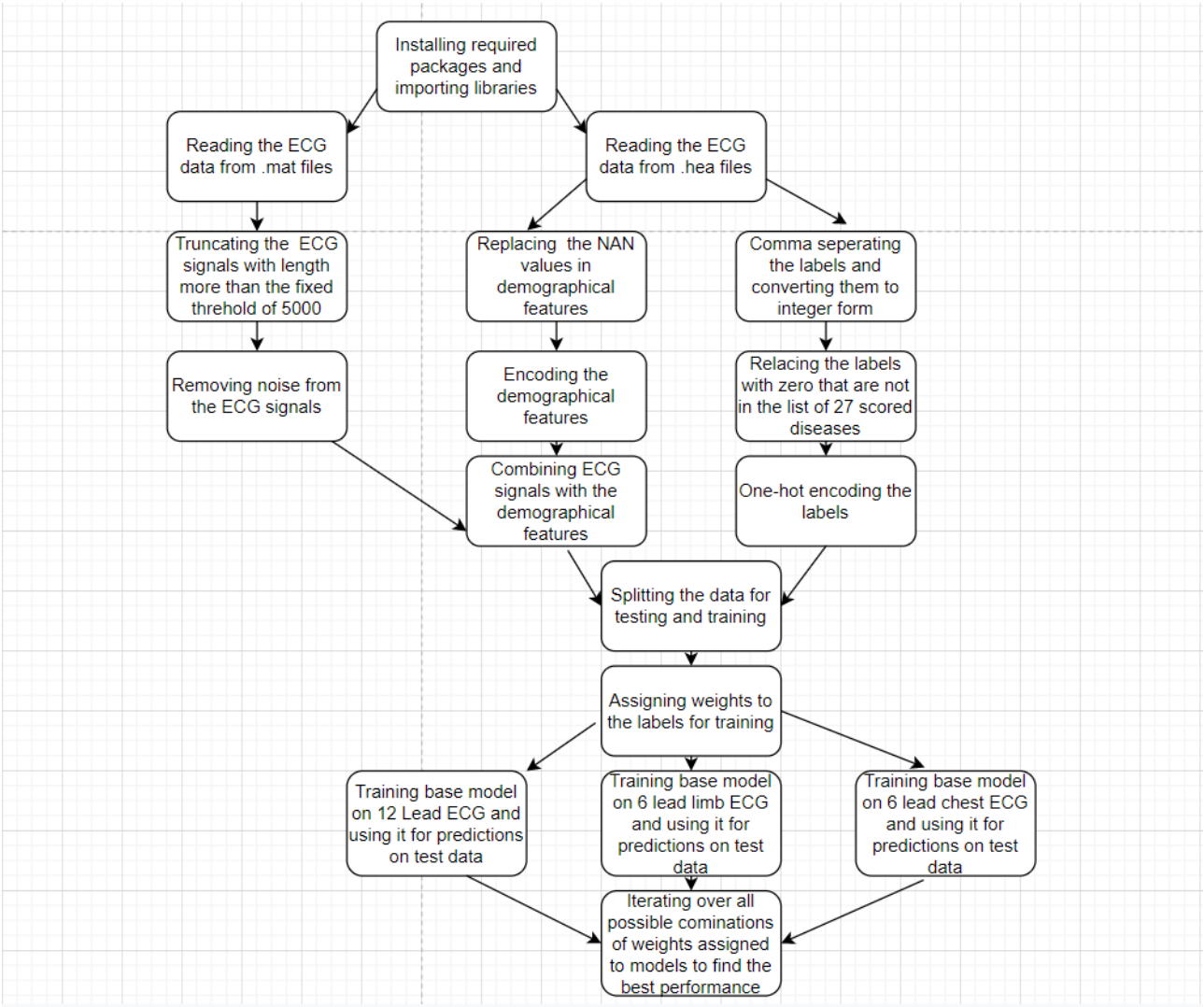

Figure 5.3: Implementation Flow.

CHAPTER 6

Results

6.1 Findings

This section discusses the results obtained from the proposed models and compares them with the results obtained by the best ranked-model in the 2020 Physio Net competition. These results are available on the official website of the competition. We also compare our results with the results obtained by [21].

Model	Training Accuracy	Validation Accuracy	Training AUC	Validation AUC
12-Lead	0.9341	0.9270	0.8052	0.8438
Limb-Leads	0.9341	0.9267	0.8159	0.8608
Chest-Leads	0.9341	0.9269	0.8081	0.8591

Table 6.1: Comparison of training and validation results for baseline models

Model	Validation Accuracy	Validation AUC
12-Lead proposed model	0.9270	0.8438
Limb-Leads proposed model	0.9267	0.8608
Chest-Leads proposed model	0.9269	0.8591
Physionet-2020 best model	0.279	0.893

Table 6.2: Comparison of Results with the competition results.

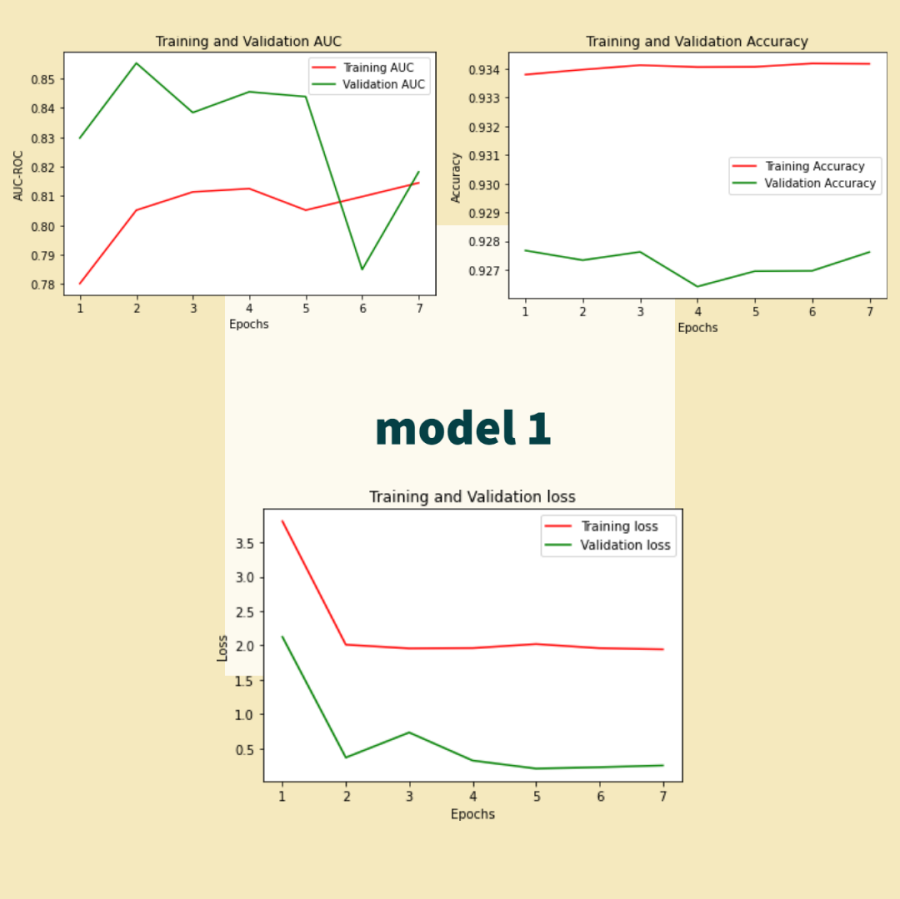

Figure 6.1: Results for the model trained on 12-lead ECG data.

12-Lead Model	Limb-Lead Model	Chest-Lead Model	Ensemble Model
0.9294311932853228	0.929320530500366	0.9290267399703834	0.94567356740345

Table 6.3: Accuracy on the test set.

6.1.1 Findings for 12 Lead ECG Model

For the ECG model trained on 12 leads data, validation loss has a falling curve from epoch one till five, but after the fifth epoch, there is no improvement in the validation loss (Figure 6.1), therefore we stop the training to avoid over-fitting. On observing the graph further, it can be seen that the training loss has a value higher than the validation loss. It doesn't necessarily mean that the model is under-fitting. It happens because we are training the model in batches, and while doing that the training loss is measured after every batch whereas the validation loss is measured at the end of the epoch. During the initial batches, the value of the loss is considerably higher and it drops in the following batches drastically. But at the end of the

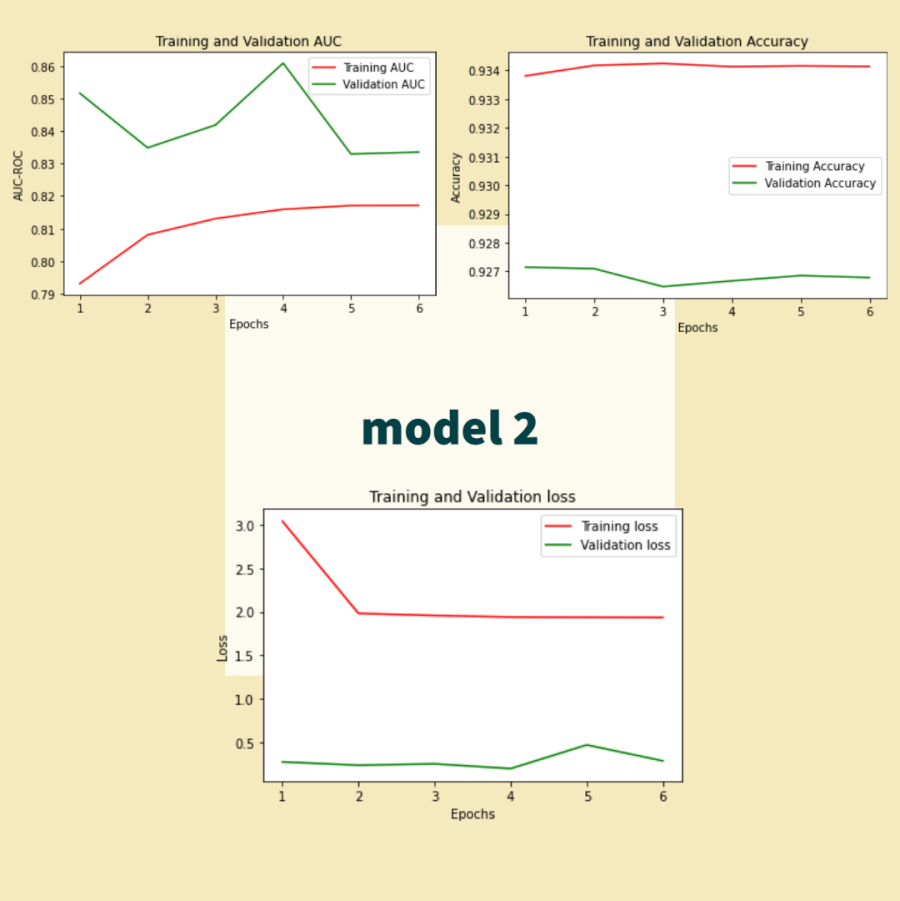

Figure 6.2: Results for the model trained on 6-lead limb data.

epoch, training loss is averaged for all the batches. Therefore, it has a value higher than the validation loss. On observing the graph for the AUC, it can be seen that the validation AUC at the final (5th) epoch is having value slightly higher than the training AUC. It means the model generalized well on the data. In case of the accuracy, the values for the validation and training differ by some decimal points. Table 6.1 contains the final training and validation scores for the baseline models. The model trained on 12 leads has the highest validation accuracy than the other models. The accuracy scores on the test data can be seen in table 6.3. When tested on the test data set, the 12 lead model gave the highest accuracy of 0.9294 percent as compared to the other two baseline models.

Table 6.2 shows the comparison of our results with the results of the best performing model in the 2020 physio net competition. As seen in the table, model trained on 12 lead ECG data has a much higher accuracy score than the best-performing model of the competition. Although we could not beat their ROC score, the ROC score for our 12 lead model is very close to the best performing model of the competition with only a difference of five percent.

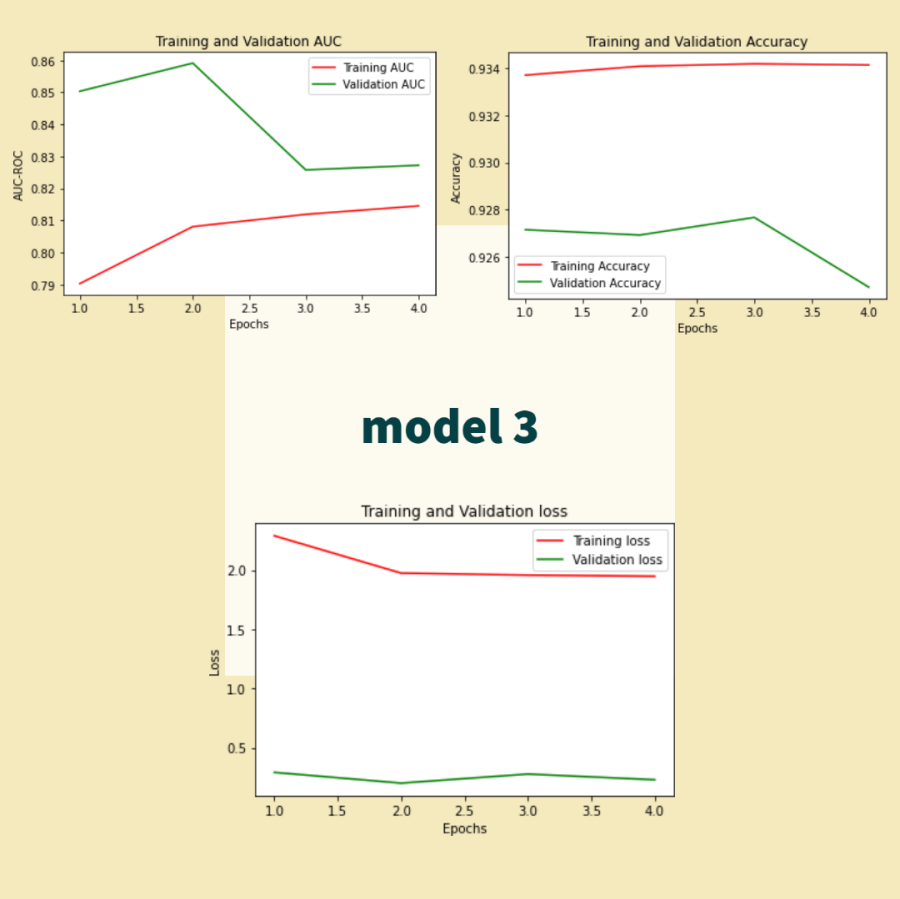

Figure 6.3: Results for the model trained on 6-lead chest data.

6.1.2 Findings for Limb-Leads Model

Figure 6.2 shows the training graph for the model trained on the limb leads. Validation loss didn't decrease after the fourth epoch, therefore we stopped the training and saved the fourth epoch as our final epoch. Training loss for this model is also a little bit higher than the validation loss for the same reason explained in the case of the first model. Validation AUC is slightly higher than the training AUC which indicates the good generalization abilities of the model. Training and validation AUC in this case differ by a small margin of 0.6. This model gave us an overall validation accuracy of 0.9267. For the AUC metric, the statistics are very interesting as the model trained on the limb leads only gives us the highest training as well as validation Area Under the Curve score of 0.8159 and 0.8608 respectively as compared to the model trained on all 12 leads. (Table 6.1) This supports our argument that some diseases can be diagnosed well only using chest or limb leads instead of using all 12 leads. Table 6.3 shows that the limb-lead model gave an accuracy of 0.9293 percent on the test data.

On comparing the results of model 2 with the 2020 physio net competition, it was noticed that the model trained on the limb leads outperformed the competition's best performing model with respect to the accuracy metric. AUC score obtained by this model was also very close to the AUC score of the competition by a difference of only 3 percent. Table 6.2

6.1.3 Findings for Chest-Leads Model

Training epochs of the third baseline model are visualized in figure 6.3. This model was only trained for two epochs as it didn't show any improvement after that and if continued could lead to the over-fitting of the model. Validation loss is lower than the training loss, therefore we can rule out the possibility of over-fitting. AUC graph indicates good generalization abilities whereas the accuracy graph rules out the possibility of under-fitting also. Training and validation Accuracy scores for the discussed model are 0.9341 and 0.9267 respectively whereas AUC scores for the training and validation data are 0.8081 and 0.8591 respectively. (Table 6.1) This model gave an overall accuracy of 0.9290 percent on the test data. Table 6.3

When compared to the 2020 physio net competition results, chest lead model was also able to beat the highest accuracy score obtained in the competition whereas AUC score differed by approximately three percent. Table 6.2 We also compared our results with the results in [21], but we were unable to beat their AUC score. One reason for that could be that the data set used by us had a very huge imbalance ratio between the majority and minority class with the majority class having 21,291 instances against the minority class of a population of 188, while they used a data set with a comparatively smaller imbalance ratio.

6.1.4 Ensemble Model Findings

The results from the ensemble approach further validate our argument that combination of certain leads can predict some diseases better than the other leads. While testing the ensemble model on the test data, the overall accuracy score improved from the previous highest score of 0.93 achieved by 12-lead model to a score of approximately 0.95 with an improvement of two percent. Table 6.3

6.2 Drawbacks

Although our proposed model is capable of generalizing well on the new ECG data due to its training on a massive range of multiple data sets but its performance could be affected due to the huge ratio of imbalance in the data set. As the data set contains realistic data coming from the hospitals, therefore we can't expect it to be a balanced data set because some diseases are very frequent in every age group whereas some diseases are very rare depending on the age. Therefore, we need to propose a more convincing technique to deal with the imbalance in ECG data sets. Moreover, the notes on the patient's medical history, medicine pescriptions and surgery details were marked at unknown in the data sets. Their presence in the data set could act as important features and help in better predictions.

6.3 Future Work

For future work, the individual effect of every ECG lead will be explored on every disease separately. This will help in further enhancing the performance of our proposed ensemble approach for ECG classification. Moreover, we plan to work on developing a more efficient and convincing technique to deal with the imbalanced data in ECGs.

CHAPTER 7

Conclusion

The novelty of our project is that we have performed multi-label classification using six data sets from multiple origins and hospitals. For that purpose, we proposed a model based on the ensemble approach which works by assigning weights to the predictions of multiple base models. We also explored the effect of chest and limb leads separately in the prediction of diseases. Our proposed model is capable of diagnosing 27 heart diseases from the ECG signals. Moreover, to deal with the data imbalance, we used a novel approach of assigning weights to the labels during the training process. Unlike the approaches used in the past, this approach prevents the under-fitting or over-fitting of the model, and creation of the redundant data. Lastly, we were able to beat the accuracy score achieved by the best-performing model in the Physio Net 2020 competition.

Bibliography

[1] E. A. P. Alday, A. Gu, A. J. Shah, C. Robichaux, A.-K. I. Wong, C. Liu, F. Liu, A. B. Rad, A. Elola, S. Seyedi, et al. Classification of 12-lead ecgs: the physionet/computing in cardiology challenge 2020. *Physiological measurement*, 41(12):124003, 2020.

[2] Z. F. M. Apandi, R. Ikeura, and S. Hayakawa. Arrhythmia detection using mit-bih dataset: A review. In *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)*, pages 1–5. IEEE, 2018.

[3] M. Bekkar, H. K. Djemaa, and T. A. Alitouche. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10), 2013.

[4] S. Celin and K. Vasanth. Ecg signal classification using various machine learning techniques. *Journal of medical systems*, 42(12):1–11, 2018.

[5] T.-M. Chen, C.-H. Huang, E. S. Shih, Y.-F. Hu, and M.-J. Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *Iscience*, 23(3):100886, 2020.

[6] C. C. Cheung, A. D. Krahn, and J. G. Andrade. The emerging role of wearable technologies in detection of arrhythmia. *Canadian Journal of Cardiology*, 34(8):1083–1087, 2018.

[7] G. D. Clifford, C. Liu, B. Moody, H. L. Li-wei, I. Silva, Q. Li, A. Johnson, and R. G. Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[8] S. Datta, C. Puri, A. Mukherjee, R. Banerjee, A. D. Choudhury, R. Singh, A. Ukil, S. Bandyopadhyay, A. Pal, and S. Khandelwal. Identifying normal, af and other abnormal ecg rhythms using a cascaded binary classifier. In *2017 Computing in cardiology (cinc)*, pages 1–4. IEEE, 2017.

[9] D. B. Geselowitz. On the theory of the electrocardiogram. *Proceedings of the IEEE*, 77(6):857–876, 1989.

[10] J. W. Grier. How to use 1-lead ecg recorders to obtain 12-lead resting ecgs and exercise (" stress") ecgs. *Department of Biological Sciences: printed from website http://www. ndsu. edu/pubweb/rvgrier*, 2008.

[11] R. He, K. Wang, N. Zhao, Y. Liu, Y. Yuan, Q. Li, and H. Zhang. Automatic detection of atrial fibrillation based on continuous wavelet transform and 2d convolutional neural networks. *Frontiers in physiology*, 9:1206, 2018.

[12] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[13] S. Hong, W. Zhang, C. Sun, Y. Zhou, and H. Li. Practical lessons on 12-lead ecg classification: Meta-analysis of methods from physionet/computing in cardiology challenge 2020. *Frontiers in Physiology*, page 2505, 2022.

[14] J. D. Howell. Early perceptions of the electrocardiogram: from arrhythmia to infarction. *Bulletin of the History of Medicine*, 58(1):83–98, 1984.

[15] M. Kachuee, S. Fazeli, and M. Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE international conference on healthcare informatics (ICHI)*, pages 443–444. IEEE, 2018.

[16] T. Karayılan and Ö. Kılıç. Prediction of heart disease using neural network. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 719–723. IEEE, 2017.

[17] R. U. Khan, T. Hussain, H. Quddus, A. Haider, A. Adnan, and Z. Mehmood. An intelligent real-time heart diseases diagnosis algorithm. In *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–6. IEEE, 2019.

[18] C. Lafuente-Lafuente, L. Valembois, J.-F. Bergmann, and J. Belmin. Antiarrhythmics for maintaining sinus rhythm after cardioversion of atrial fibrillation. *Cochrane Database of Systematic Reviews*, (3), 2015.

[19] K. Li, W. Zhang, Q. Lu, and X. Fang. An improved smote imbalanced data classification method based on support degree. In *2014 international conference on identification, information and knowledge in the internet of things*, pages 34–38. IEEE, 2014.

[20] Z. Li, K. Kamnitsas, and B. Glocker. Analyzing overfitting under class imbalance in neural networks for image segmentation. *IEEE transactions on medical imaging*, 40(3):1065–1077, 2020.

[21] Z. Li and H. Zhang. Automatic detection for multi-labeled cardiac arrhythmia based on frame blocking preprocessing and residual networks. *Frontiers in cardiovascular medicine*, 8:616585, 2021.

[22] G. Y. Lip and H.-F. Tse. Management of atrial fibrillation. *The Lancet*, 370(9587):604–618, 2007.

[23] A. Y.-c. Liu. *The effect of oversampling and undersampling on classifying imbalanced text datasets*. PhD thesis, Citeseer, 2004.

[24] F. Liu, C. Liu, L. Zhao, X. Zhang, X. Wu, X. Xu, Y. Liu, C. Ma, S. Wei, Z. He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

[25] Z. Liu, X. Meng, J. Cui, Z. Huang, and J. Wu. Automatic identification of abnormalities in 12-lead ecgs using expert features and convolutional neural networks. In *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*, pages 163–167. IEEE, 2018.

[26] R. Mahajan, R. Kamaleswaran, J. A. Howe, and O. Akbilgic. Cardiac rhythm classification from a short single lead ecg recording via random forest. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[27] R. A. Massumi, R. K. Sarin, A. A. Tawakkol, J. C. Rios, and H. Jackson. Time sequence of right and left atrial depolarization as a guide to the origin of the p waves. *The American Journal of Cardiology*, 24(1):28–36, 1969.

[28] S. Narkhede. Understanding auc-roc curve. *Towards Data Science*, 26(1):220–227, 2018.

[29] N. Naseer and H. Nazeer. Classification of normal and abnormal ecg signals based on their pqrst intervals. In *2017 International Conference on Mechanical, System and Control Engineering (ICMSC)*, pages 388–391. IEEE, 2017.

[30] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, and J. Rubin. A wide and deep transformer neural network for 12-lead ecg classification. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[31] F. Plesinger, P. Nejedly, I. Viscor, J. Halamek, and P. Jurak. Parallel use of a convolutional neural network and bagged tree ensemble for the classification of holter ecg. *Physiological measurement*, 39(9):094002, 2018.

[32] L. Politano, A. Palladino, G. Nigro, M. Scutifero, and V. Cozza. Usefulness of heart rate variability as a predictor of sudden cardiac death in muscular dystrophies. *Acta Myologica: Myopathies and Cardiomyopathies: Official Journal of the Mediterranean Society of Myology*, 27:114–122, 2008.

[33] S. Pramanik and H. A. B. Dahlan. Age estimation using shortcut identity connection of resnet50 based on convolutional neural network. In *2021 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 1–7. IEEE, 2021.

[34] L.-D. Quach, N. P. Quoc, N. H. Thi, D. C. Tran, and M. F. Hassan. Using surf to improve resnet-50 model for poultry disease recognition algorithm. In *2020 International Conference on Computational Intelligence (ICCI)*, pages 317–321, 2020.

[35] K. Ramasubramanian and A. Singh. Deep learning using keras and tensorflow. In *Machine Learning Using R*, pages 667–688. Springer, 2019.

[36] A. H. Ribeiro, D. Gedon, D. M. Teixeira, M. H. Ribeiro, A. L. P. Ribeiro, T. B. Schön, and W. Meira. Automatic 12-lead ecg classification using a convolutional network ensemble. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[37] A. M. Shaker, M. Tantawi, H. A. Shedeed, and M. F. Tolba. Generalization of convolutional neural networks for ecg classification using generative adversarial networks. *IEEE Access*, 8:35592–35605, 2020.

[38] A. Shoughi and M. B. Dowlatshahi. A practical system based on cnn-blstm network for accurate classification of ecg heartbeats of mit-bih imbalanced dataset. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–6. IEEE, 2021.

[39] S. Śmigiel, K. Pałczyński, and D. Ledziński. Ecg signal classification using deep learning techniques based on the ptb-xl dataset. *Entropy*, 23(9):1121, 2021.

[40] J. Soni, U. Ansari, D. Sharma, and S. Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, 17(8):43–48, 2011.

[41] C. Veni. On the classification of imbalanced data sets. 2018.

[42] T. Vicar, J. Hejc, P. Novotna, M. Ronzhina, and O. Janousek. Ecg abnormalities recognition using convolutional network with global skip connections and custom loss function. In *2020 Computing in Cardiology*, pages 1–4. IEEE, 2020.

[43] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, F. N. Delling, et al. Heart disease and stroke statisticsâ2021 update: a report from the american heart association. *Circulation*, 143(8):e254–e743, 2021.

[44] C. Wijaya, M. Harahap, M. Turnip, A. Turnip, et al. Abnormalities state detection from p-wave, qrs complex, and t-wave in noisy ecg. In *Journal of Physics: Conference Series*, volume 1230, page 012015. IOP Publishing, 2019.

[45] Z. Xiong, M. P. Nash, E. Cheng, V. V. Fedorov, M. K. Stiles, and J. Zhao. Ecg signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network. *Physiological measurement*, 39(9):094006, 2018.

[46] X. Xu and H. Liu. Ecg heartbeat classification using convolutional neural networks. *IEEE Access*, 8:8614–8619, 2020.

[47] Z. Zhu, X. Lan, T. Zhao, Y. Guo, P. Kojodjojo, Z. Xu, Z. Liu, S. Liu, H. Wang, X. Sun, et al. Identification of 27 abnormalities from multi-lead ecg signals: An ensembled se_resnet framework with sign loss function. *Physiological Measurement*, 42(6):065008, 2021.

APPENDIX A

Data and Code Availability

A.1 Source Code

The source code of the project is available at:

https://cseegit.essex.ac.uk/21-22-ce901-su/21-22_CE901-SU_raza_ahmad

The project file contains the following three folders:

- Labels" folder containing the SNOMED mappings for the scored and "unscored diagnostic labels
- "Models" folder containing the trained models
- "Notebooks" folder containing the python notebooks for the main code and the supporting functions

A.2 Data Availability

Data used in the project is available at:

https://physionet.org/content/ecg-arrhythmia/1.0.0/

APPENDIX B

Project break-down Structure

The whole project was broken down into three tasks which are further broken down into smaller tasks: Figure B.1

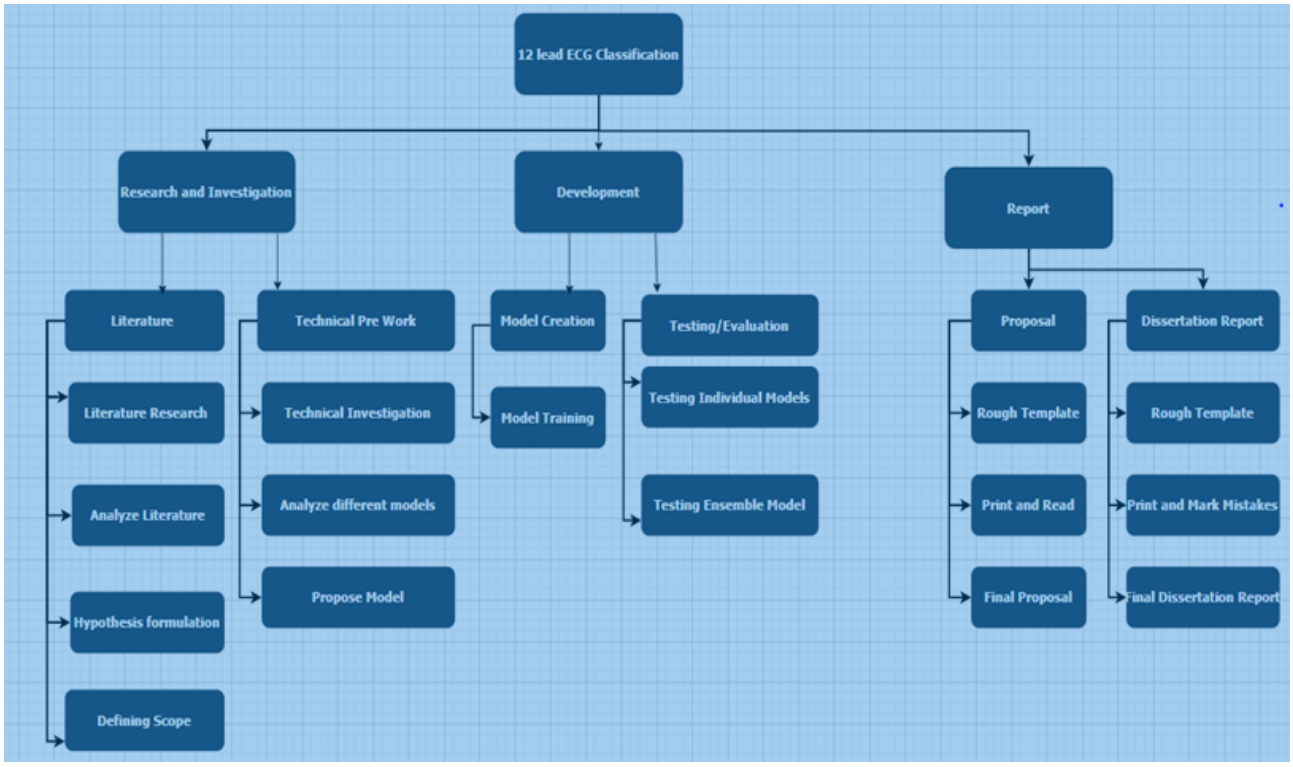

Figure B.1: Project break-down Structure.

APPENDIX C

Resources Utilized

C.1 Software Resources

As the project was based on a very large data set, therefore two main resources were used to compile the project. Initially, the subscription for the Google Colab Pro+ version was purchased but its computation power was not enough to run the project. Therefore, an additional resource (GPU system) was utilized, which was provided by the Supervisor Dr. Luo Cunjin.

C.2 Educational Resources

The major source of the information was the regular weekly meetings held with the supervisor. Other resources were utilized from the library at the University of Essex, Google Scholar and IEEE.