Prediction of Heart Rate Abnormalities Using Data Mining Techniques

MSc Research Project

Data Analytics

Jhanavi Govinde Gowda

Student ID: X18128998

School of Computing

National College of Ireland

Supervisor: Dr. Cristina Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jhanavi Govide Gowda
Student ID:	X18128998
Programme:	Data Analytics **Year:** 2019
Module:	Research Project
Supervisor:	Dr. Cristina Muntean
Submission Due Date:	12-Dec-2019
Project Title:	Prediction of Heart Beat Abnormalities Using Data Mining Techniques
Word Count:	6214 **Page Count:** 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jhanavi Govinde Gowda

Date: 12-Dec-2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Prediction of Heart Rate Arrhythmias Using Data Mining Techniques

Jhanavi Govinde Gowda

X18128998

Abstract

Heart rate arrhythmias have become one of the most popular heart issues in human beings at recent years .It has opened a way to many researchers to classify the heart beats issues in the medical field for detecting the arrhythmia signals at early stages which will help in reducing the death rates. The research focuses on classifying the arrhythmias in the ECG signal for the 5 different categories of signals in the MIT-BIH dataset. A lot of research has been done in the classification for arrhythmias signal using various data mining techniques. Logistic Regression and CNN algorithms were implemented to classify the heart beat signals and aims at providing better accuracy in predicting the signals. Both the implemented models have been compared in terms of accuracy, precision, recall and computational time. The evaluation results of the deep leaning model - CNN turned out to be high in accuracy of 99.09% when compared with the machine learning model. This research helps out in classifying the heartbeat rate arrhythmias at early stages for enhanced treatment.

Keywords: Arrhythmias, Wavelet Transformation, Logistic Regression, CNN

1 Introduction

Recent statistics shows that 116 million people in US are estimated to have hypertension and in every 3.7 minutes a person dies from stroke and 1 out of every 3 persons dies due to heart diseases according to the reports of American Health Association (AHA). Depending on different criteria's, heart arrhythmias are classified into different categories. The arrhythmias such as atrial-flutter, atrial-fibrillation, premature-contractions, sinus-bradycardia, and atrial-tachycardia will lead to stroke and makes the people to experience flutter in the throat and chest. To detect these variations at early stages, manual attempt to analyse the heart beat signals by recording the ECG signal is tedious. As the data is increasing day by day, ECG signals are varying continuously and makes difficult for the researchers to analyse. ECG signals are periodic in nature; periodically varying signals is shown in Figure 1. Lot of researches is being carried out in this field and various data mining algorithms have been suggested in order to classify the signals.

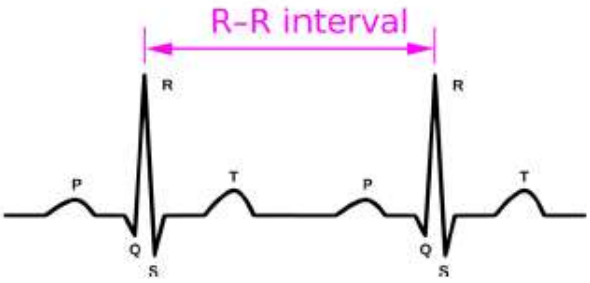

Figure 1: ECG waveform

The motivation behind this project is, the growing technology in the past 2-3 decades has increased the growth of risk factor in human health. The stressful and mechanical life, pollution has immensely affected human health, especially the issues related to heart. Cardio-vascular related diseases detected at later stages have accounted the major human death rate approximately 31% according to the World Health Organization (WHO). 82% of the people fall under middle and low economic classes making them difficult to afford the medical expenses. Cardiac arrhythmias, a serious heart condition found in the heart signal has acquired pathologies which increases and decreases the heart beat and causes sudden death. Early detection of variations in arrhythmias heart beat signals using data mining techniques will help in providing better treatment in a reduced medical expense.

The below research question helps in predicting the heart rate arrhythmias signals in human beings for better treatment when detected at early stages.

Can deep learning and machine learning techniques detect and classify the arrhythmias in heart beat signals to reduce the efforts in analysing the signal for enhanced treatments?

The project objective aims at classifying the heart beats for different input categories of heart beats like N, S, F, V and Q using machine learning and deep learning algorithms. The research helps to identify the problem easily and reduces the manual intervention efforts of doctors to analyse the ECG report. It also helps in giving early treatments to the patients who has variation in the heart beats signals.

The technical report is organized into different sections: Section 2 contains related work gives an insight to the various developments done in classifying the signals. Section 3 presents the methodology followed in the project and the project flow. Section 4 describes the design specifications and the architecture used in the project implementation. Section 5 explains the algorithm implementation. Section 6 evaluates the algorithms in terms of accuracy, recall and precision. Section 7 gives the discussion on the problems faced and Section 8 gives the conclusion of the project and specifies the future work for the project.

2 Related Work

Usually the research for classifying the ECG signals is done by considering the record of either a single patient (intra patient ECG data) or by collecting the records of many individual patients which is known as inter-patient ECG data. The researches have made one of the two choices for their research and in this regard, (Can Ye, Kumar, & Coimbra, 2012) categorized the work into, intra patient and inter patient data in order to classify the signal. (Martis, Acharya, Mandana, Ray, & Chakraborty, 2012) uses the record from the single patient which are only based on data labels of the heart beat signals and is classified into train and test data to provide positive results whereas (De Chazal, O'dwyer, & Reilly, 2004) adopted inter patient paradigms to build a generalized model by taking separate data to train and test which adopts to the practical situations. This proposed inter patient paradigm obtained an accuracy of 83% and sensitivity of 81% and produced more realistic results. On the other hand (De Chazal et al., 2004) proposed an hybrid model which is a patient specific, the global classifier's were built first and then local classifier's were used in order to train global classifier's. It obtained an accuracy and sensitivity of 97.4% and 94.4% respectively.

As the research progressed further, researches started building analytical ways of modelling and classifying the signals by using machine learning algorithms. (Alarsan & Younes, 2019) proposed the study of classification of heartbeat using Decision-Tree, XGBoost, and Random Forest. It uses spark – Scala tools since it simplify the algorithms and is preferred for huge data. The proposed algorithms were able to detect and classify the signal into 3 types: normal beats and abnormal beats such as Premature-Ventricular-Contraction (PVC) and Premature-Atrial-Contraction (PAC) by obtaining an accuracy of 96.75% and 97.98% for GDB and Random Forest respectively.

In this regard, (Díaz-Robles et al., 2008) have addressed the issue by adopting active learning in order to perform task adaptive and patient adaptive tasks. The SVM model has been built using MATLAB platform by producing an F-score of 99%. (Gar, Gladston, Men, & Eduardo, 2017) proposed the ECG representation using temporal vector cardiogram with complex network to extract the features and fine-tuned the classifier using SVM and particle swarm optimization algorithm to perform feature selection by producing a sensitivity of 87.3%. (Shi et al., 2019) classified the model by implementing XGBoost algorithm, which is one of the efficient machine learning algorithms to generalize the data, in order to bring regularity to control the complex model by following 3 steps: Pre-processing the data, Feature extracting and Hierarchical classification. Feature extraction is one of the important steps followed in order to classify the signal, it implements wrapper technique to extract the features and the features with smallest scores are removed. It includes extracting many different features like RR-intervals (Sannino & De Pietro, 2018) (Shi et al., 2019), statistical features, morphological features (De Chazal et al., 2004), vector cardiogram (Oh, Ng, Tan, & Acharya, 2018), higher order statistical features and interval features. In addition to that, some other features like, wavelet packet entropy (Li & Zhou, 2016), linear features (Martis et al., 2012) and abstract features are proposed.

The researchers conducted gives the insight to different machine learning algorithms but fails in handling huge data as the performance gradually decreases with increase in data, breaks down the problem into different parts and combine them at the end to give final result. These demerits in machine learning are well addressed by neural networks which give end to end results. Hence, the classification as taken a new turn by using neural networks which has an input and an output layer with many hidden layers which is in general called as deep learning techniques. It is one of the most efficient and powerful way of classifying complex data by increasing accuracy, sensitivity and precision. (Sannino & De Pietro, 2018) implemented DNN algorithm with 3 layers, firstly the data is analysed, features are extracted and the signals are classified. (Camacho, Collins, Powers, Costello, & Collins, 2018) The classification of signals include N (normal-beats), super-ventricular-beats (S), ventricular-beats (V) or fusion of normal beats and ventricular beats (F) by using a Google library called tensor flow framework which consists of 5, 10, 30, 50, 30, 10, 5 neurons and has seven layers which is hidden. The signal processing takes part in 4 steps: 1. De-noising using a low pass filter of order 12 with a cut off frequency of 35Hz and 2 median filters. 2. Peak detection is done using wavelet transformation which is there in MATLAB. 3. ECG signal is segmented into separate individual signal. 4. Feature extraction where the heartbeats are extracted like RR interval, global RR average interval and post interval. Then the algorithm is applied in order to compare accuracy, sensitivity and specificity with other most frequently used and popular algorithms like Naïve-Bayes, Ada-Boost, Random-Forest-Tree, SVM etc. DNN comparatively obtained an accuracy of 99%.

(Sellami & Hwang, 2019) used MIT-BIH data and implemented CNN model for the classification of the heartbeat. The model used 9 convolutional layers with each layer containing 64 size kernels and implements batch normalization technique and obtained an accuracy of 99.79%. The data is also implemented with the combination algorithms like CNN-RNN, CNN-LSTM, CNN-GRU. The Framework resulted in 83.4% accuracy applied using Keras library and Tensorflow-. (Hasan & Bhattacharjee, 2019) states that using CNN high recognition can be provided by applying 10 layers convolutional layers. (Sannino & De Pietro, 2018) improves the classification in intra patient paradigm as the model for individual patients equals to the number of patients. This results in redundancy and bad generalization, hence the author employed GRNN (Global-Recurrent-Neural-Network) (Wang et al., 2019), it is a higher version of RNN, generalizes the whole system and learns differences which is in different classes. The process follows 4 steps like System framework, construction of framework, Classification of models and the system implementation. The features like premature-or-Escape-Flag and morphological vector were extracted. The databases like SVDM, MIT-BIH and PTB were used with active learning algorithms to train the model. The model uses GRNN0, GRNN1, GRNN2, GRNN3 to evaluate the training data and testing data which is taken for different databases. It obtained an accuracy of 95.4% which is more accurate than other machine learning languages.

(Yildirim, Baloglu, Tan, Ciaccio, & Acharya, 2019) applied auto encoder algorithm, it is a new deep learning technique to decrease the size of the heart beats. It also employs CAE-LSTM (Long-Term Short-Term Memory) decreases the time duration for analysing the beats.

The CAE model has encoders and decoders which are compressed to a CAE model of 16 layers. The feature size at encoder part is compressed into 32*1 from 260*1 sampled signals and original signals are reconstructed using decoder to build LSTM network. Max pooling technique is applied to compress and in extracting the features of the signals, and they are normalized using batch normalization layer. It resulted in an accuracy of 99.42% showing it is highly efficient to classify the signals.

For more accuracy (Mathews, Kambhamettu, & Barner, 2018a) uses RBM (Restricted-Boltzman-Machine) and DBN (Deep-Belief-Network) to classify heart beats into 2 types which are VEB (Ventricular-Ectopic-Beats) and SVEB (Super-Ventricular-Ectopic-Beats). The processing of data is done in 3 phase, pre-processing, segmentation and feature extracting by using the popular data base MIT-BIH data base. 12 tap low pass filter, RLS (Recursive-Least-Square) and finite impulse response filter are used which is more efficient in order to remove the artefact's in the signal. Pre-processing includes down sampling the signal to 114Hz and extracting the features in two parts one having Twenty six features and the other one having Twenty two features. The model obtained an accuracy of 93.78% and 96.94% for SVEB and VEB signals at low sampling rate which is a better technique than ANN (Artificial-Neural-Network), QDA (Quadratic-classifier), LDA (Linear-Discriminant-Analysis) etc. Along with high accuracy it also reduces the training time and the CPU cycles.

(Dong, Wang, & Si, 2017) suggested a new technique by using deterministic learning algorithm to classify the signal to model and in order to rapidly recognize the signal. The process involved 2 steps: modelling the near signals and to classify the beats. It employs a unique feature called beat dynamics to classify the signals and instead of static features, it concentrates more on dynamic features. The system produced an overall accuracy of 97.78%, by using 5% training data. (Liu, Si, Wen, Zang, & Lang, 2016) gave a novel approach of using dictionary learning algorithm to classify signals. He employed k-medoid clusters, which is optimized by k-means++. The vector quantization features are applied with the proposed algorithm and the obtained results are compared with discrete wavelet transformation feature, Fourier transformation feature and sampling point feature.

(Dutta, Chatterjee, & Munshi, 2011) suggests ANN algorithm to classify the signal. It employs LVQ (Learning-Vector-Quantization) which is a part of supervised clustering to classify 3 types of signals like normal signals (N), Premature-Ventricular-Contraction (PVC) and Super-Ventricular-Ectopic Beats (SVEB). It also employs cross correlation technique in order to determine the similarity in the two signals. It uses MIT-BIH database and resulted in an accuracy of 95.24%. (Park & Kang, 2014) implements Pan-Tompkins algorithm to classify the signal. It extracts the features like QRS and P signals and also employs decision tree to classify the signal. In order to provide accurate result, the algorithm applies couple of mat functions like differentiators, integrators and cascades both low pass and high pass filter using MIT-BIH to provide 97.06% accuracy. (Ankışhan, 2019) uses SVM, CNN and MLR techniques for classification. The signal is smoothened and normalized with an accuracy of 98.09%, 98.56% and 96.88% respectively.

After the detailed study of the researches done in past, it can be concluded that a lot of classification models have been developed to classify the heart beat signals and in improving the accuracy. But the classification is only limited to two or three categories and few researches have limited there study by classifying the signal into normal and abnormal beats. Even though many deep learning algorithms give good accuracy, it uses complex architectures which require huge installations, making the model very complex and difficult to understand.

Hence this project aims at classifying the signals into 5 categories by implementing simple linear model and deep learning techniques to increase the accuracy and to study the approach followed in both the techniques and to give the comparison of the models based on computational time taken, accuracy, precision and recall and to decide the algorithm which best suits the data.

3 Research Methodology

The research project is followed with CRISP-DM methodology, it comprises of Business-Understanding, Data-Understanding, Data-Preparation, Modelling, Evaluation and Deployment has given in the Figure 2.

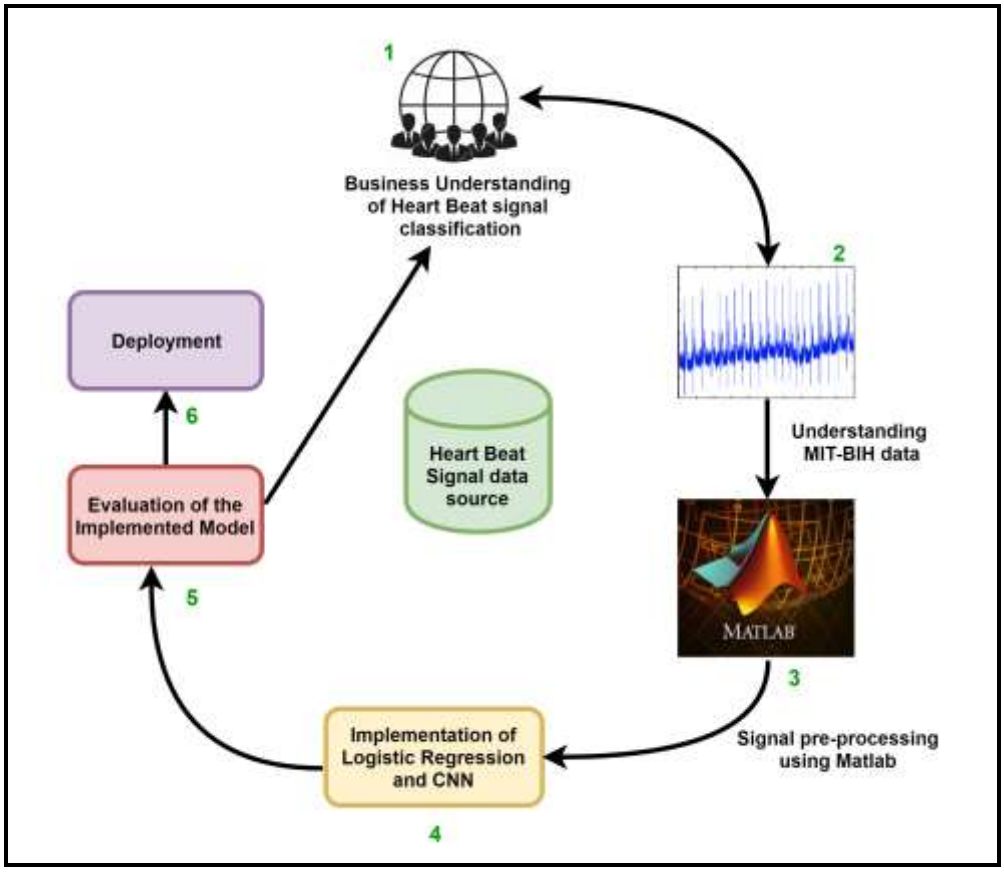

Figure 2: Process Flow diagram of the methodology implementation

3.1 Business understanding

The sudden increase in the heart disease rate has created the necessity for classification of heart beat. If the arrhythmias are detected at the early stages, it will help the doctor's in detecting the severity and analyze the patient's criticality and helps in giving the patients with proper treatment.

3.2 Data understanding

For this research, data was extracted from Massachusetts-Institute of Technology – Beth-Israel Hospital (MIT-BIH) at Boston from a publicly available website[1] which was studied from 1975 – 1979. The data contains the analysis of arrhythmias and cardiac dynamics. The dataset contains observations of 48 patients, in which 23 recordings has mixed population of in-patients and out-patients. The rest of the 25 recordings contain significant arrhythmias. Each sampling record was collected with a sampling frequency of 360 Hz and 11-bit resolution over 10mV range. The dataset contains 87554 rows which represent the number of samples and 187 columns represent the types of beats (N, S, V, F and Q). Figure 3 shows the sample of the dataset.

Types of Beats →

Number of Samples ↓

	0	1	2	3	4	5	6	7	8	9	...	177	178	179	180	181	182	183	184	185	186
0	1.000000	0.758264	0.111570	0.000000	0.080579	0.078512	0.066116	0.049587	0.047521	0.035124	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.908425	0.783883	0.531136	0.362637	0.366300	0.344322	0.333333	0.307692	0.296703	0.300366	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.730088	0.212389	0.000000	0.119469	0.101770	0.101770	0.110619	0.123894	0.115044	0.132743	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	1.000000	0.910417	0.681250	0.472917	0.229167	0.068750	0.000000	0.004167	0.014583	0.054167	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.570470	0.399329	0.238255	0.147651	0.000000	0.003356	0.040268	0.080537	0.070470	0.090604	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	1.000000	0.923664	0.656489	0.195929	0.111959	0.175573	0.122137	0.050891	0.035623	0.055980	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	1.000000	0.797260	0.320548	0.043836	0.049315	0.065753	0.030137	0.008219	0.005479	0.010959	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.909408	0.975610	0.533101	0.134146	0.066202	0.000000	0.010453	0.012195	0.031359	0.146341	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.927803	0.866359	0.299539	0.000000	0.231951	0.317972	0.274962	0.262673	0.270353	0.268817	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	1.000000	0.914230	0.473684	0.000000	0.064327	0.317739	0.405458	0.391813	0.382066	0.401559	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.997429	0.861183	0.365039	0.071979	0.082262	0.100257	0.074550	0.051414	0.051414	0.043702	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 3: Imported MIT-BIH Data

3.3 Signal Processing

The extracted data is processed in MATLAB using wavelet transformation to remove the power-line and base-line interference. The signal is de-noised in 5 levels by implementing DWT technique. The technique of using DWT helps in removing the high pass signals and provides the filtered output.

3.4 Implementation of the Models

Once the signals are pre-processed, the cleansed data was fed to the machine-learning and deep-learning algorithm for the implementation of the models to predict the arrhythmias in heart beats by classifying the different categories. During the implementation, the data processed out of the signal pre-processing was split into training and testing data in 70:30 ratio split. Then, Logistic Regression in machine learning and Convolution Neural Network

[1] https://physionet.org/content/mitdb/1.0.0/

in deep-learning models were implemented as part of the implementation of the research project is discussed in the next chapter.

3.5 Evaluation of the Models

Evaluation of the implemented models was done to evaluate the efficiency of the model. Some of the evaluation parameters were used to test the implemented model performance using accuracy, recall and precision, F1-measure and computational time. Confusion matrix was extracted from the implemented model to calculate the accuracy, precision and recall of the model as shown in the Figure 4.

Confusion Matrix		ACTUAL	
		YES	NO
PREDICTED	YES	True Positive	False Negative
	NO	True Negative	False Positive

Figure 4: Confusion-Matrix

3.6 Deployment

As part of deployment for the implemented models, a user interface using web integration will be deployed into cloud platform. This will be taken to future for classifying the heart beat arrhythmias using the web application.

4 Design Specification

Overall research project had a design following three-tier architecture. This three-tier architecture helps in establishing a link with the client to access the business logic layer and data interpretation layer at any point of time is shown in Figure 5.

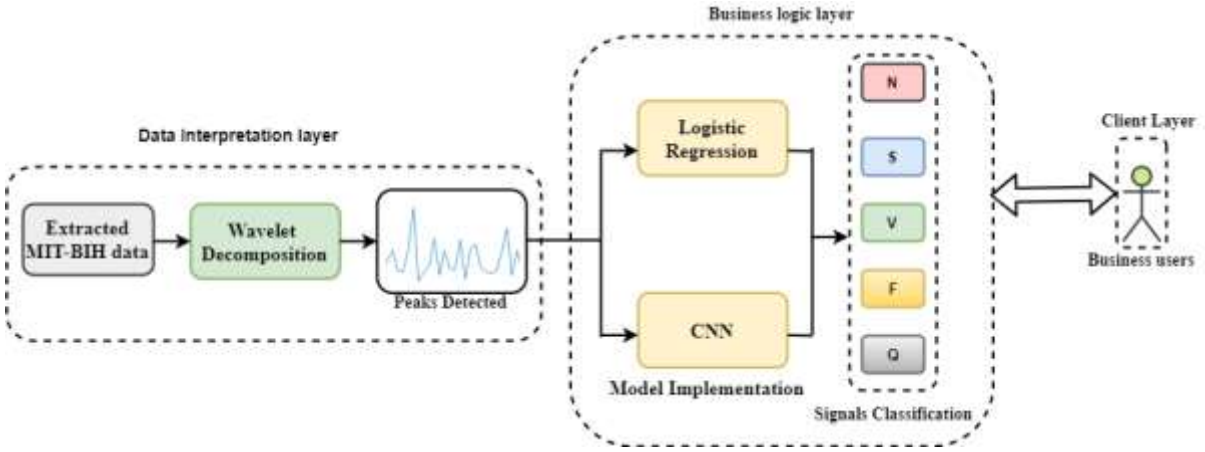

Figure 5: Design Flow

In the data interpretation layer, extracted data was stored in the local machine. The data was pre-processed initially. In the pre-processing phase, the data was removed with unwanted noise, power-line and base-line interference, artifacts and electrosurgical noise using MATLAB. Business Logic Layer covers the technical implementation of the model and

application of the business logics. Implementation of the Logistic Regression and Convolution Neural Network (CNN) algorithms was done in this layer. Implemented algorithms were evaluated in terms of accuracy, precision, recall and computational cost. Client layer addresses the Stakeholders. This model was approached by considering the Stakeholders (Doctors), where the output of the classification will help in reducing the manual intervention of analyzing the signal. Detection of arrhythmias will benefit the doctors by reducing the effort of manually analysing the report and checking for abnormalities and also helps the patients to get the best available treatment in early stages.

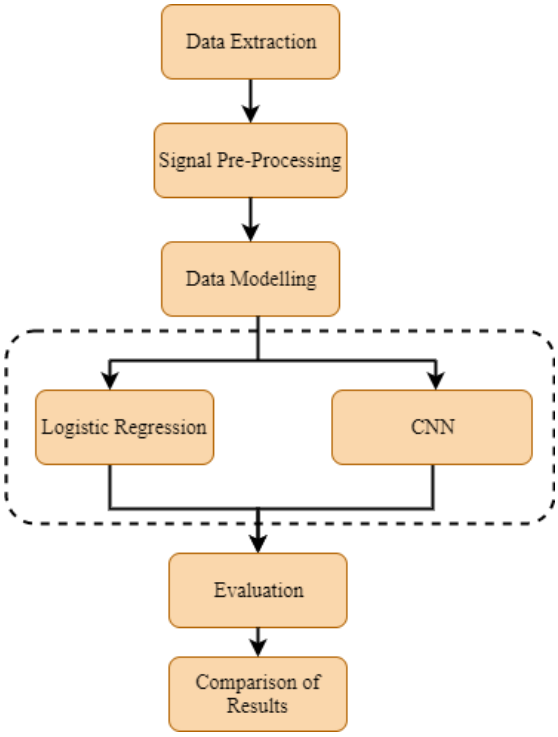

Figure 6: Process Flow diagram

The process flow of the project is given in Figure 6, the data is extracted from the database and it is pre-processed in MATLAB. The output obtained is implemented with two data mining models such as logistic regression and CNN. The implemented models are then evaluated and the results are compared and the best model is selected. The detailed explanation of the process flow is given in implementation part of the project.

5 Implementation of Heart Signal Classification

The research work mainly focuses on the classification of the heart beat signals for various categories. Signal pre-processing was carried out before implementation. The project has been implemented with machine-learning and deep-learning algorithms like Logistic Regression and CNN.

5.1 Signal Pre-Processing of the Data

The extracted MIT-BIH dataset was processed in MATLAB version R2019b (9.7.0.1190202) using wavelet analysis tool with Discrete-Wavelet-Transformation (DWT). The signal

contains distortions like baseline wanderings and power line interference. The distortions were caused due to respiration and patients movements. These distortions will interrupt the ECG signal and add noise to the signal. DWT was chosen as the signal shows slow oscillations and not localized with respect to time and frequency. Hence the most efficient way to address these abrupt changes is DWT. As it is finite, the signal was captured using stretched wavelets and abrupt changes were captured by compressed wavelet. According to the Nyquist criteria, implementation of the tree structure differentiates the signal into low pass and high pass filters, and discords half of the samples is shown in the Figure 7.

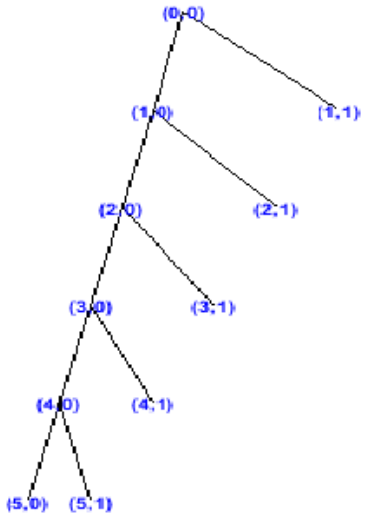

Figure 7: DWT Tree Filtering

Filters have the ability to cancel and reconstruct the sub-band. The next level sub-band will take low pass and iterates using the same technique to yield the narrower sub-band. The filter after de-noising was detected with R peaks and QRS complexions which is explained in below figures.

The original ECG signal with distortions at first stage as shown in Figure 8.

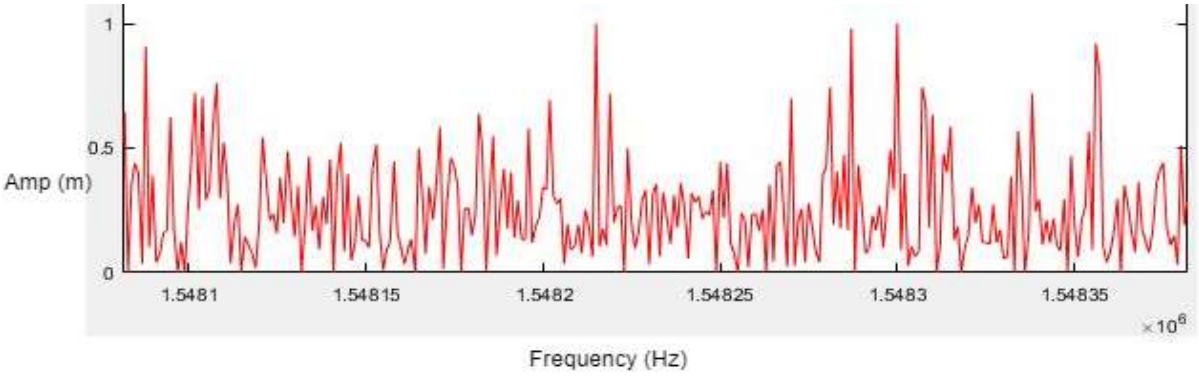

Figure 8: Original ECG signal at node (0, 0)

Signal obtained after the first level of filtering which removes second stage of high frequency signals as shown in Figure 9.

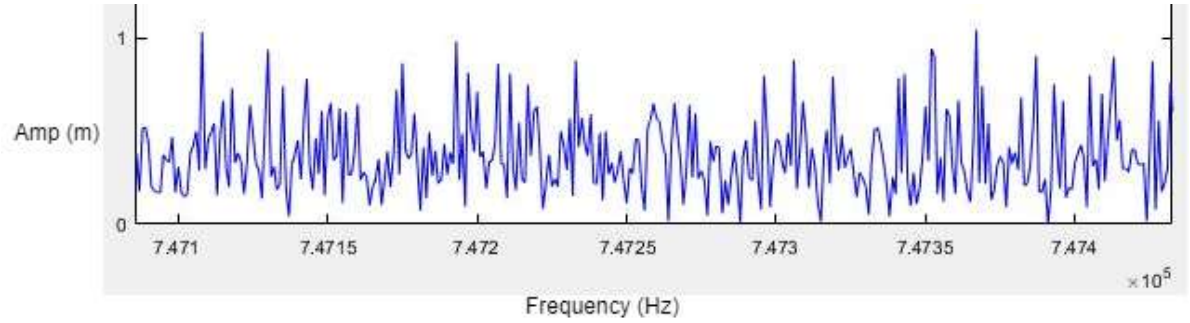

Figure 9: ECG signal at node (1, 0)

Signal obtained after the second level of filtering which removes third stage of high frequency signals as shown in Figure 10.

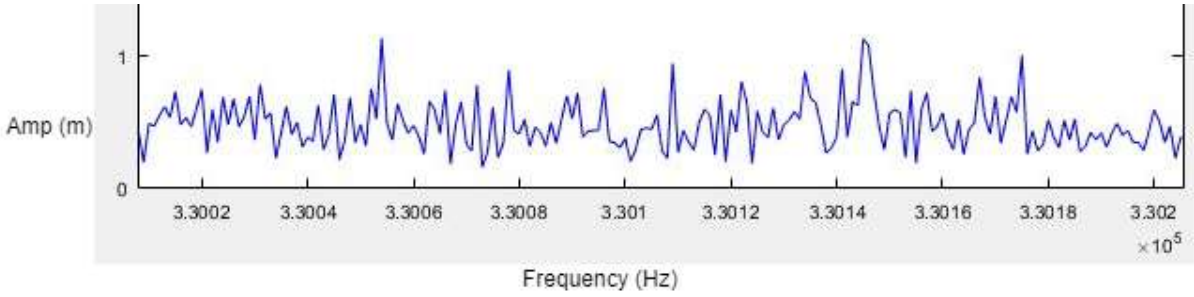

Figure 10: ECG signal at node (2, 0)

Signal obtained after the third level of filtering which removes fourth stage of high frequency signals as shown in Figure 11.

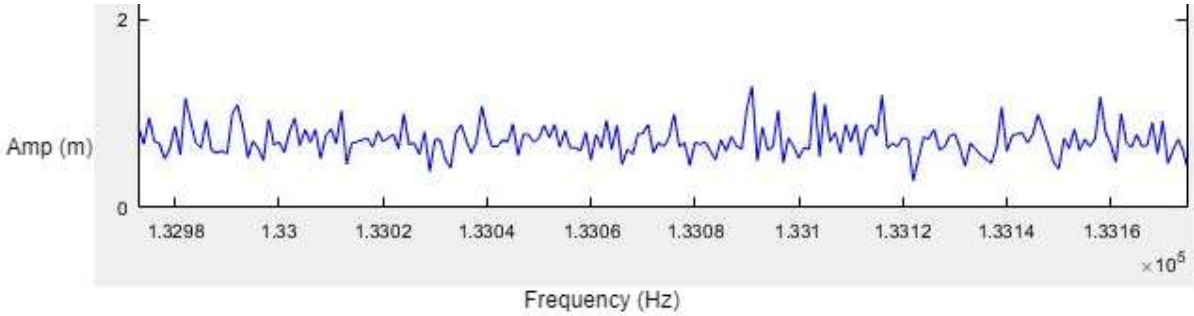

Figure 11: ECG signal at node (3, 0)

Signal obtained after the fourth level of filtering which removes fifth stage of high frequency signals as shown in Figure 12.

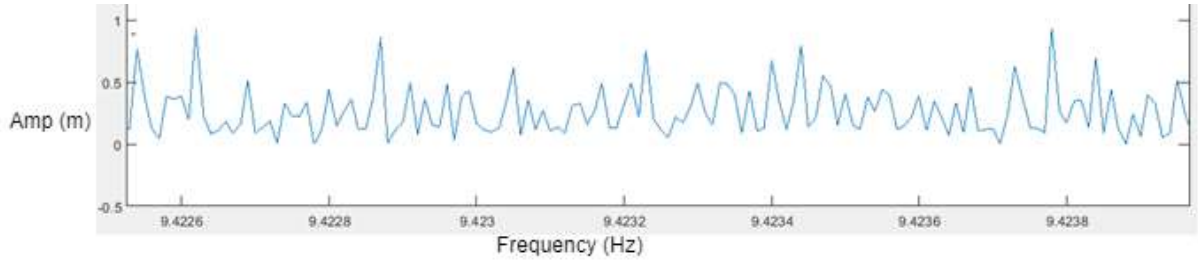

Figure 12: ECG signal at node (4, 0)

5.2 Logistic Regression

Among the machine learning models, logistic regression one of the most popular and simplest classification models falls in GLM (General Linear Model) framework. It helps in understanding how different variables affect the continuous variables. Sigmoid function uses cost function. Due to cost function, hypothesis of the model is limited to 0 and 1. The linear combination of input values using bias and coefficient values gives the output[2] of the dependent variable.

$$Y = \frac{e^{(bo+bix)}}{1+e^{(bo+bix)}}$$

Where, b_0 → intercept or bias term; b1 → co-efficient.

The train and test data was extracted from the MIT-BIH dataset. Extracted data has 80,000 samples per beat cycles of ECG. In order to fit into a unified timeframe each sample is normalized into a [0, 1] interval and padded with zeros. The graph and the spectrum of the dataset was plotted by specifying the range(5) and assigning plots_per_class to 10 and iterating them through each class, which divides the heart beat classes into 5 different categories. Each class represents a different state of heart is explained in Figure 13.

[2] https://machinelearningmastery.com/logistic-regression-for-machine-learning/

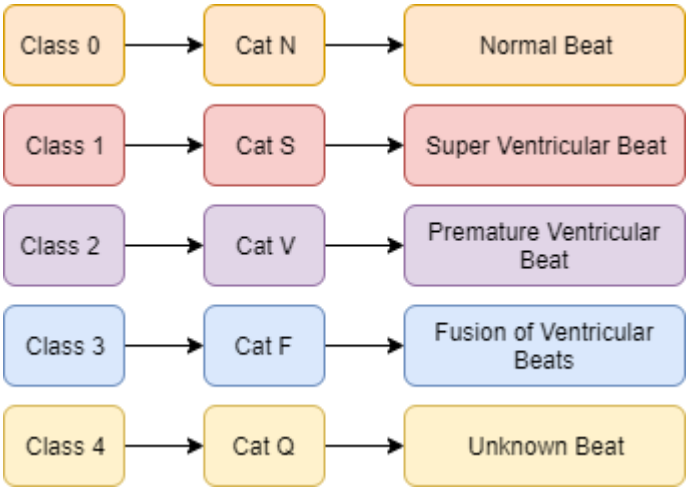

Figure 13: Different categories of heart beats (Mathews, Kambhamettu, & Barner, 2018b)

The model was implemented with Gaussian to smoothen the signal as smoothing helps in reducing the randomness of the signal and expose the signal for better understanding. Logistic Regression was applied on the extracted MIT-BIH data in classifying the model. The output of the model obtained is evaluated using accuracy, precision and recall. Logistic regression gives the classification of the model by classifying the signals into one of the category. But as the complexity increases, the approximation to the extended boundaries of multi class cases is difficult and hence the model needs to be introduced with non-linearity to increase accuracy and efficiency of the model. The model was then introduced with Fourier transformation to accommodate non-linearity to the system and used for simplifying the signal into smaller components.

The model imports kernel tricks such as RBF sampler. Kernels benefits the model by reducing the computational time, provides access to infinite dimensions of the feature space and gives similarities within the data instead of features.
The output of the Logistic Regression model is given by,

$$P\ (y = |\ x) = \frac{1}{1+\exp(-w^T x)}$$

As (x) is mapped to the implicit space of representation of kernel and the output of the formula gives the Kernel representation for Logistic Regression model.

$$P\ (y = |\ x) = \frac{1}{1+\exp(-\sum_{i=1}^{n} \alpha_i y_i k(x_i, x))}$$

The RBF Sampler kernel function benefits in modeling the data which cannot be linearly classified in the form of circles by providing smooth transitions. The RBF Sampler on two samples x, x` is given by,

$$K(x, x`) = \exp\left(-\frac{||x-x`||^2}{2\sigma^2}\right)$$

$||x - x`||^2$ Is the squared Euclidean-distance between the two vectors x, x`

The formulas mentioned above are taken from the website[3].

[3] https://en.wikipedia.org/wiki/Radial_basis_function_kernel

Gaussian mixture was imported for clustering purposes which groups the model into K number of groups with K representing the individual state of the machine. Again Logistic Regression was executed for better performance with high accuracy and precision along with confusion matrix was obtained.

5.3 Convolution Neural Networks (CNN)

CNN artificial neural network contains input, output and many hidden layers. The data vector for one dimensional vector input is given by x = ($x_1, x_2, x_3, x_4 \dots \dots \dots \dots x_n$) where $x_n \in R$ denotes the features.

The major libraries like Numpy, Pandas, Mathplot and Keras were installed for implementing CNN. Convolution 1D and Maxpooling 1D libraries were also imported from keras layers and models from keras.models. Training and testing data were loaded into the CNN. Both the training and testing data were concatenated, counts of unique values in the row were obtained, which indexes into 5 different categories like C0, C1, C2, C3 and C4 and flattened into row wise and graph is plotted on a (20,12) size window. The 5 different types of graph constitute to 5 different states (N, S, V, F and Q) of heartbeat with each state explaining different form of abnormality as listed in figure 11. The data is augmented to same level in order to train the model properly and the plot of different categories of the signal is shown in Figure 14.

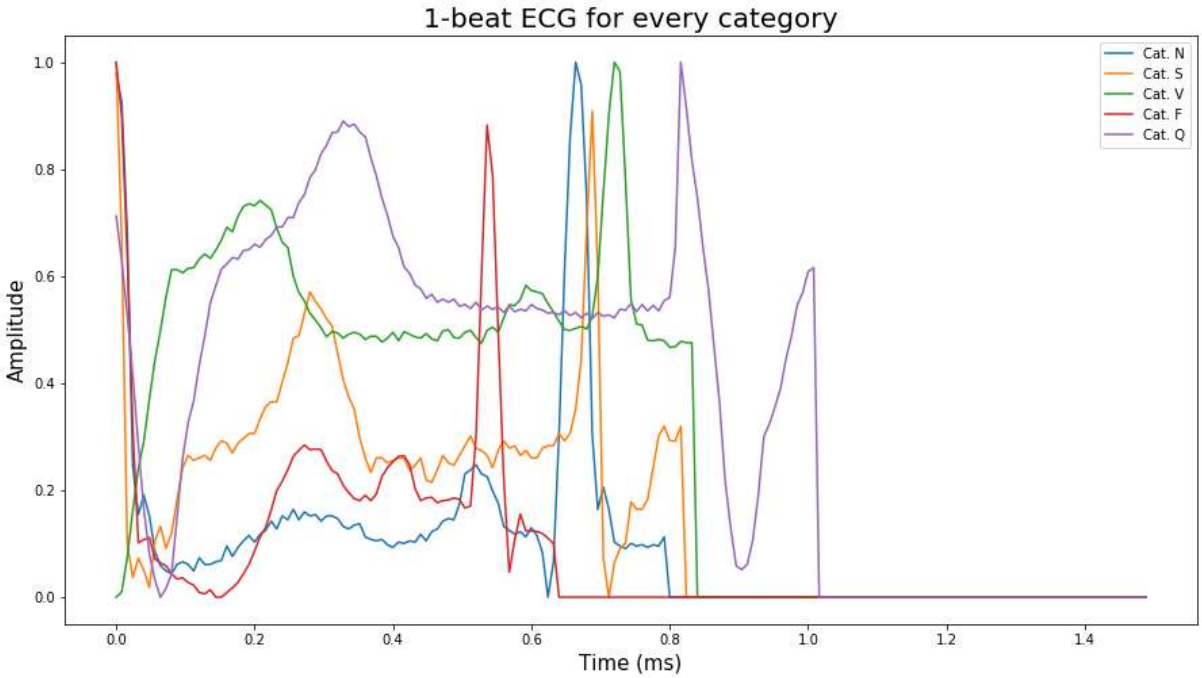

Figure 14: 1 Beat ECG classified into 5 categories

Fixing the batch size of 500, conv1D was implemented using 3 convolution layers. Each layer has filters of 32, with kernel size of 5*5 and strides of 1 were selected in order to jump from one item to another along the dimension of the array. The network was trained for 20 epochs and the activation layer was introduced to learn non linearity for separating the 5 classes as classes cannot be linearly separable. ReLU (Rectified-Linear-Units) activation

function was used since it is widely used and more effective in neural network architecture. After each convolutional layer, Max pooling technique was used to down sample the input and to reduce the dimensionality of the input by specifying a pool size of 5. The end to end architecture of the applied CNN model is given in Figure 15 and the output of the applied convolution layers with parameters are listed in Figure 16.

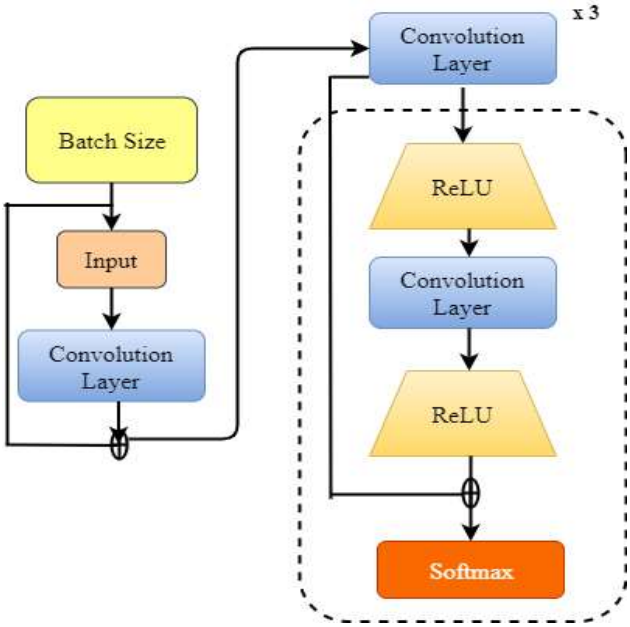

Figure 15: End to end architecture of the applied CNN model

As machine learning algorithms cannot convert the categorical data into vector numbers, the one-hot-encoding function was implemented in order to convert categorical data into vector numbers. Once the model was created, compilation of the model was performed using Adam optimizer. Adam Optimizer helps in computing efficiently as it uses less memory and well suited for large data. The model was finally trained using keras fit[4]() function, and was trained for 20 epochs with 109150 sample data.

[4] https://keras.io/scikit-learn-api/

Model: "model_1"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 187, 1)	0	
conv1d_1 (Conv1D)	(None, 183, 32)	192	input_1[0][0]
conv1d_2 (Conv1D)	(None, 183, 32)	5152	conv1d_1[0][0]
activation_1 (Activation)	(None, 183, 32)	0	conv1d_2[0][0]
conv1d_3 (Conv1D)	(None, 183, 32)	5152	activation_1[0][0]
add_1 (Add)	(None, 183, 32)	0	conv1d_3[0][0] conv1d_1[0][0]
activation_4 (Activation)	(None, 183, 32)	0	add_1[0][0]
max_pooling1d_2 (MaxPooling1D)	(None, 90, 32)	0	activation_4[0][0]
conv1d_6 (Conv1D)	(None, 90, 32)	5152	max_pooling1d_2[0][0]
activation_5 (Activation)	(None, 90, 32)	0	conv1d_6[0][0]
conv1d_7 (Conv1D)	(None, 90, 32)	5152	activation_5[0][0]
add_3 (Add)	(None, 90, 32)	0	conv1d_7[0][0] max_pooling1d_2[0][0]
activation_6 (Activation)	(None, 90, 32)	0	add_3[0][0]
max_pooling1d_3 (MaxPooling1D)	(None, 43, 32)	0	activation_6[0][0]
flatten_1 (Flatten)	(None, 1376)	0	max_pooling1d_3[0][0]
dense_1 (Dense)	(None, 32)	44064	flatten_1[0][0]
activation_7 (Activation)	(None, 32)	0	dense_1[0][0]
dense_2 (Dense)	(None, 32)	1056	activation_7[0][0]
dense_3 (Dense)	(None, 5)	165	dense_2[0][0]
softmax_1 (Softmax)	(None, 5)	0	dense_3[0][0]

Total params: 66,085
Trainable params: 66,085
Non-trainable params: 0

Figure 16: Details of the layers and the parameters for CNN

6 Evaluation

The results obtained by extensive study of both the models were highly accurate and shows good efficiency in classifying the heart beat signals. The performance matrix's such as accuracy[5], precision and recall was calculated for each model.

The formulas for precision and recall are taken from the website[6]

[5] https://developers.google.com/machine-learning/crash-course/classification/accuracy
[6] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

$$\text{Accuracy} = \frac{\textit{True Positive} + \textit{True Negative}}{\textit{Total Samples}}$$

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

$$\text{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Negative}}$$

Logistic regression model obtained an accuracy of 91.51% which is fair enough, with very less testing time of 33.04 seconds. The precision, recall, F1-score for each class and the confusion matrix has been presented in the Table 1.

Table 1: Evaluation Metrics Report for Logistic Regression

	precision	recall	f1-score	support
Cat N 0.0	0.98	0.89	0.93	18118
Cat S 1.0	0.31	0.65	0.42	556
Cat V 2.0	0.69	0.83	0.75	1448
Cat F 3.0	0.15	0.78	0.25	162
Cat Q 4.0	0.89	0.94	0.91	1608
accuracy			0.88	21892
macro avg	0.60	0.82	0.65	21892
weighted avg	0.93	0.88	0.90	21892

After introducing Non Linearity like Random Fourier feature into the model, the accuracy increased by 3.22% that is to 91.51% and the confusion matrix was obtained along with the classification report in Table 2.

Table 2: Evaluation Metrics Report after introducing Non Linearity

	precision	recall	f1-score	support
Cat N 0.0	0.98	0.92	0.95	18118
Cat S 1.0	0.40	0.71	0.51	556
Cat V 2.0	0.71	0.87	0.78	1448
Cat F 3.0	0.22	0.85	0.35	162
Cat Q 4.0	0.94	0.94	0.94	1608
accuracy			0.91	21892
macro avg	0.65	0.86	0.71	21892
weighted avg	0.94	0.91	0.92	21892

CNN model obtained an accuracy of 99.09% which shows that the model is highly significant to classify the signals and takes 5467 seconds computational time to train the data. The Evaluation report and the confusion matrix of CNN model is given in Figure 17 and Table 3.

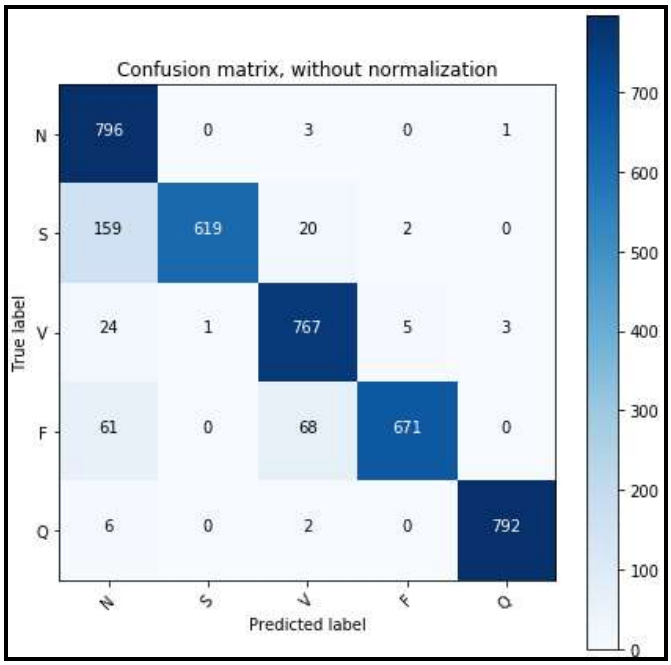

Figure 17: Confusion matrix for CNN model

Table 3: Evaluation Metrics Report of CNN

	precision	recall	f1-score
Cat N 0	0.76	0.99	0.86
Cat S 1	1.00	0.77	0.87
Cat V 2	0.89	0.96	0.92
Cat F 3	0.99	0.84	0.91
Cat Q 4	0.99	0.99	0.99

The evaluation matrix suh as precision, accuracy and recall for both the models is compared and graphs have been ploted.

The Figure 18 shows the precision graph for logistic regression and CNN models for all the 5 categories. The graph shows that precision for CNN is higher than the logistic regression.

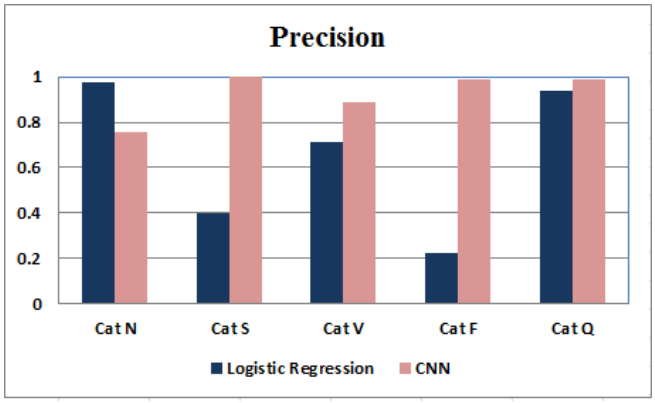

Figure 18: Precision comparison

The Figure 19 shows the accuracy graph for logistic regression and CNN models. The graph shows that accuracy for CNN is higher than the logistic regression.

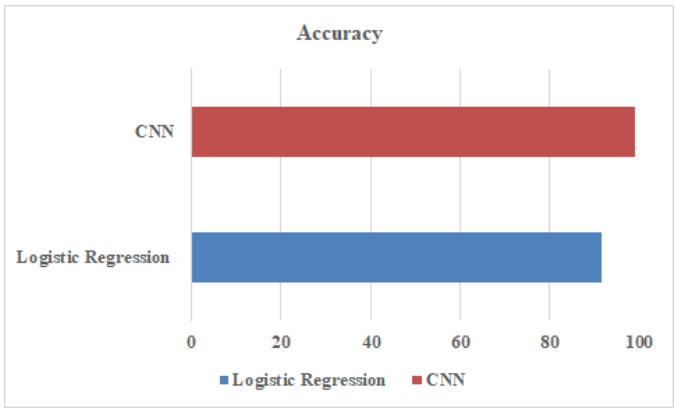

Figure 19: Accuracy comparison

The Figure 20 shows recall value for logistic regression and CNN models for all the 5 categories. The graph shows that recall value for CNN is higher than the logistic regression.

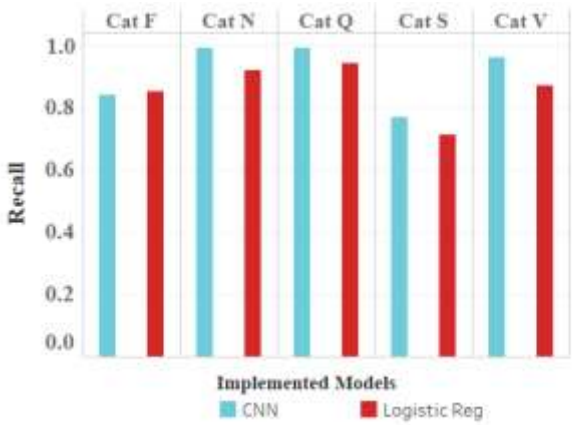

Figure 20: Recall value comparison

The results obtained after implementing both the models show that the accuracy obtained from CNN model is very high when compared to logistic regression. But CNN model takes more computational time. Since CNN model trains the data with high accuracy and efficiency by implementing many hidden layers, it can be concluded that CNN model fits the data well compared to logistic regression.

7 Discussion

Classification of the heart signal rates was main objective of the research project. The extracted data had different observations for the same patient as the recording of heart beat signals varied. The shape of the ECG recording slightly differs in shape as the heart's electrical activity travels in the direction which led to a positive deflection. There were few challenges in pre-processing data using MATLAB as the data had power-line and baseline interfaces. Pre-processing was done by de-noising the signals using wavelet transformation

tool with DWT approach. Another challenge in building the model using CNN algorithm was the model did not respond properly when it was executed for 70 epochs. Then the epoch was tuned to 20 for good performance and better accuracy. Computational cost was kept in consideration while developing model rather than a complex architecture. The results obtained after overcoming the challenges were better in predicting the heart signal rate and with higher accuracy.

8 Conclusion and Future Work

Cardiac arrhythmia is a kind of irregularity in the ECG cycle leads to major complications in heart and causes sudden death. Early detection of arrhythmia will help in resolving the problem has been addressed in the research project. Patient's heart beat signals were preprocessed using MATLAB for noise reduction. Classifying the heart beat signals into 5 different categories was carried out in the implementation using both machine-learning and deep-learning techniques. Implemented models were compared and studied in terms of computational time, accuracy and precision to evaluate the performance of one model over the other. CNN had higher accuracy and was cost effective in classifying the heart beat signals at early stages of arrhythmias.

Future work of the research can be carried by deploying the results of the implemented models in cloud platform. Further research can be conducted by collecting the data for more patients from the hospitals having CCU and implementing more machine-learning and deep-learning algorithms will help in understanding the other models performance and accuracy.

Acknowledgement

My sincere thanks to the project guide Dr. Cristina Muntean, who guided me consistently throughout the project and helped me to take the project in the right path. I also thank all my lectures and parents for their continuous support.

References

Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0244-x

Ankışhan, H. (2019). Estimation of heartbeat rate from speech recording with hybrid feature vector (HFV). *Biomedical Signal Processing and Control*, *49*, 483–492. https://doi.org/10.1016/J.BSPC.2019.01.015

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell*, *173*(7), 1581–1592. https://doi.org/10.1016/J.CELL.2018.05.015

Can Ye, Kumar, B. V. K. V., & Coimbra, M. T. (2012). Heartbeat Classification Using Morphological and Dynamic Features of ECG Signals. *IEEE Transactions on Biomedical Engineering*, *59*(10), 2930–2941. https://doi.org/10.1109/TBME.2012.2213253

De Chazal, P., O'dwyer, M., & Reilly, R. B. (2004). Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features. *IEEE TRANSACTIONS ON*

BIOMEDICAL ENGINEERING, *51*(7). https://doi.org/10.1109/TBME.2004.827359
Díaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., & Moncada-Herrera, J. A. (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, *42*(35), 8331–8340. https://doi.org/10.1016/J.ATMOSENV.2008.07.020
Dong, X., Wang, C., & Si, W. (2017). ECG beat classification via deterministic learning. *Neurocomputing*, *240*, 1–12. https://doi.org/10.1016/J.NEUCOM.2017.02.056
Dutta, S., Chatterjee, A., & Munshi, S. (2011). Identification of ECG beats from cross-spectrum information aided learning vector quantization. *Measurement*, *44*(10), 2020–2027. https://doi.org/10.1016/J.MEASUREMENT.2011.08.014
Gar, G., Gladston, M., Men, D., & Eduardo, L. (2017). *Inter-Patient ECG Heartbeat Classification with Temporal VCG Optimized by PSO*. (July), 1–11. https://doi.org/10.1038/s41598-017-09837-3
Hasan, N. I., & Bhattacharjee, A. (2019). Deep Learning Approach to Cardiovascular Disease Classification Employing Modified ECG Signal from Empirical Mode Decomposition. *Biomedical Signal Processing and Control*, *52*, 128–140. https://doi.org/10.1016/J.BSPC.2019.04.005
Li, T., & Zhou, M. (2016). ECG Classification Using Wavelet Packet Entropy and Random Forests. *Entropy*, *18*(8), 285. https://doi.org/10.3390/e18080285
Liu, T., Si, Y., Wen, D., Zang, M., & Lang, L. (2016). Dictionary learning for VQ feature extraction in ECG beats classification. *Expert Systems with Applications*, *53*, 129–137. https://doi.org/10.1016/J.ESWA.2016.01.031
Martis, R. J., Acharya, U. R., Mandana, K. M., Ray, A. K., & Chakraborty, C. (2012). Application of principal component analysis to ECG signals for automated diagnosis of cardiac health. *Expert Systems with Applications*, *39*(14), 11792–11800. https://doi.org/10.1016/J.ESWA.2012.04.072
Mathews, S. M., Kambhamettu, C., & Barner, K. E. (2018a). A novel application of deep learning for single-lead ECG classification. *Computers in Biology and Medicine*, *99*, 53–62. https://doi.org/10.1016/J.COMPBIOMED.2018.05.013
Mathews, S. M., Kambhamettu, C., & Barner, K. E. (2018b). A novel application of deep learning for single-lead ECG classification. *Computers in Biology and Medicine*, *99*, 53–62. https://doi.org/10.1016/J.COMPBIOMED.2018.05.013
Oh, S. L., Ng, E. Y. K., Tan, R. S., & Acharya, U. R. (2018). Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Computers in Biology and Medicine*, *102*, 278–287. https://doi.org/10.1016/J.COMPBIOMED.2018.06.002
Park, J., & Kang, K. (2014). PcHD: Personalized classification of heartbeat types using a decision tree. *Computers in Biology and Medicine*, *54*, 79–88. https://doi.org/10.1016/J.COMPBIOMED.2014.08.013
Sannino, G., & De Pietro, G. (2018). A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems*, *86*, 446–455. https://doi.org/10.1016/J.FUTURE.2018.03.057
Sellami, A., & Hwang, H. (2019). A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. *Expert Systems with Applications*, *122*, 75–84. https://doi.org/10.1016/J.ESWA.2018.12.037
Shi, H., Wang, H., Huang, Y., Zhao, L., Qin, C., & Liu, C. (2019). A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine*, *171*, 1–10. https://doi.org/10.1016/J.CMPB.2019.02.005
Wang, G., Zhang, C., Liu, Y., Yang, H., Fu, D., Wang, H., & Zhang, P. (2019). A global and

updatable ECG beat classification system based on recurrent neural networks and active learning. *Information Sciences*, *501*, 523–542. https://doi.org/10.1016/J.INS.2018.06.062

Yildirim, O., Baloglu, U. B., Tan, R.-S., Ciaccio, E. J., & Acharya, U. R. (2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. *Computer Methods and Programs in Biomedicine*, *176*, 121–133. https://doi.org/10.1016/J.CMPB.2019.05.004