ECG data analysis and heart disease prediction using machine learning algorithms

Sushmita Roy Tithi, Afifa Aktar, Fahimul Aleem, Amitabha Chakrabarty

***Abstract*— In the modern world, there have been some revolutionary advancement in the field of medical science and research and this is no different for electrocardiogram. Electrocardiogram (also abbreviated as ECG) illustrates the electrical activity of one's heart over a time period. Over the years, number of people suffering from heart disease have increased to some extent.Therefore, in our research, we aim to design a model using supervised machine learning that can find anomalies in one's ECG report by analyzing it. We have applied six supervised machine learning algorithms to distinguish between normal and abnormal ECG. In addition,we used them to predict the chances of a patient suffering from a certain heart disease. We divided our data set into two parts. 75% data in one group for training the model and rest 25% data in another group for testing. To avoid any kind of anomalies or repetitions, Cross Validation and Random Train-Test Split was used to obtain an answer as accurate as possible. We have compared the results with each other for a better understanding.**

***Index Terms*— ECG, Machine learning, Logistic regression, Decision tree, Nearest neighbour, Naïve Bayes, Support Vector Machine, Arificial Neural Network, Right bundle branch block, Myocardial infarction, Sinus tachycardia, Sinus bradycardia, Coronary artery disease, Abnormal ECG.**

I. INTRODUCTION

In this era, the 21st century, almost every other person is dependent on technology. This remarkable development of technology has made life better in many possible ways. Number of people suffering from heart disease is increasing day by day mostly because of the unhealthy lifestyle. So data acquisition methods have been developed for the past decades to accurately sense, collect, record, and analyze the patient's' physical condition [9]. ECG provides us with series of sinus rhythm which defines the condition of heart. ECG is useful for detecting certain types of conditions and it is the most common test for screening heart diseases for its low cost [1]. Furthermore, ECG pattern recognition is often useful as an early warning system for heart diseases. Very little change in any section of the ECG graph can result in different kind of diseases. During medical emergencies, like in ER or ICU, where time is of the essence, it would be more advantageous to find out what is ailing the patient for immediate treatment [1]. At times, there is high chance to miss out any abnormality in the ECG report as the change in the ECG wave shape is hardly noticeable. With the recent development in Machine Learning algorithms, the scope of performing in different sectors and concluding with better accuracy and optimized performance has increased [1]. Medical science has also improved over time. Considering all these factors, we planned about finding anomalies in Heart Rate of ECG reports and figure out which algorithm gives better and reliable result for some particular heart diseases. We distinguished between normal and abnormal ECG using different machine learning algorithms. Afterwards, we predicted few diseases depending on the availability of the data for a particular disease. The model that we trained to predict the outcome for different diseases and classify between normal and abnormal ECG can also be used to predict outcome for other diseases that we did not work for. In addition, we figured out which algorithm gives the best result in predicting the diseases we worked with. We analyzed the results that we predicted by using our model and calculated the accuracy level of different algorithms for our selected diseases.

Department of CSE, BRAC University, Dhaka, Bangladesh,

II. LITERATURE REVIEW

For a long time researchers have been working on identifying and predicting different diseases using machine learning. Exploration supervised learning, unsupervised learning and reinforcement learning, which is better for machine learning are analyzed in [2], [3] and [4]. There have been numerous works in the field of classification and involving mostly neural networks, Markov chain models and support vector machines (SVMs) [5]. In [6], a comparison between three different machine learning algorithms were discussed. It has been done by many to improve past algorithms or create new one for machine learning. In [2], a new algorithm VF15 was developed to classify arrhythmia. In [8], a definite review of preprocessing strategies, ECG databases, highlight extraction methods, classifiers and execution measures are displayed whereas in [7], a survey on the best machine learning approach on reading ECG is given. Machine learning in medical science is an unmistakable research related on osmosis of present day innovations: programming, PC and data advancements [9]. There is even research on defining the ECG wave from other wave or mixed signal. For example Independent component analysis (ICA) is connected on the blended signs and the isolated signs are recreated utilizing wavelet remaking and correlating the results demonstrate that Lifting Wavelet Transformation and FASTICA algorithm creates the best SNR estimation of 11.39 for maternal and 10.10 for fetal Electro Cardio Gram signals [10]. In [11], the purpose was to make an algorithm with elevated amounts of precision and less dimensions of false cautions by classifying the hearts electrical signals as demonstrative of ischemia or not. ECG early warning system comes to light for meeting public health, medical informatics of health services and

978-1-7281-0297-9/19/$31.00 ©2019 IEEE

information sending or enhancing through the internet and related technologies [9]. For example, ROC (Receiver operating characteristics) diagrams are helpful for sorting out classifiers and observing their execution, moreover they are normally utilized in making medical decision, and lately have been utilized progressively in machine learning and information mining research[12]. Ultimately, there are many research work involving machine learning and medical science, however we choose to work with ECG or the functionality of the human heart. During the time most common defect in human body primarily related with heart.

III. DATA AND METHODS

A. Dataset

We worked with the data set we found in the UCI Machine Learning Repository (UCI) database [13]. The data set or the input was a .data file. In this .data file the data are given in rows where each rows are the instances and the columns are the attributes. However, the attribute name of each column is not given in the .data file. The names of the attributes are given for each column in a different file. It is a pre-processed data which contains 452 instances and 279 attributes of which 206 are linear valued and the rest are nominal where 16 types of cardiac arrhythmia are classified.Among these, the first one being the normal ECG and class distribution of this data set is really unjust as three of the class's (11, 12, 13) instances does not exist and Class01 (normal) is most repeated [2]. These data are taken by using 10 electrodes. There are repetitive attributes for 12 different leads. The class distributions are Normal, Ischemic changes (Coronary Artery Disease), Old Anterior Myocardial Infarction, Old Inferior Myocardial Infarction, Sinus tachycardia, Sinus bradycardia, Ventricular Premature Contraction (PVC), Supra ventricular Premature Contraction, Left bundle branch block, Right bundle branch block, First-degree atrioventricular block (AV block), Second-degree atrioventricular block (AV block), Third-degree atrioventricular block (AV block), Left ventricle hypertrophy, Atrial Fibrillation or Flutter, and Others[13]. Attributes of this data set are Age, Sex, Height, Weight, QRS duration: Average of QRS duration in msec., P-R interval: Average duration between onset of P and Q waves in msec., Q-T interval: Average duration between onset of Q and offset of T waves in msec., T interval: Average duration of T wave in msec., P interval: Average duration of P wave in msec., Vector angles in degrees on front plane of:, QRS , T , P, QRST , J, Heart rate: Number of heart beats per minute which are the first fifteen columns [2][13]. Additionally attributes from channel DI, channel DII, channel DIII, channel AVR, channel AVL, channel AVF, channel V1, channel V2, channel V3, channel V4, channel V5, channel V6 are Average width in msec. of Q wave, R wave, S wave, R' wave, small peak just after R, S' wave, and Number of intrinsic deflections, Existence of ragged R wave, Existence of diphasic derivation of R wave, Existence of ragged P wave, Existence of diphasic derivation of P wave, Existence of ragged T wave, Existence of diphasic derivation of T wave, Amplitude of JJ wave, Q wave, R wave, S wave, R' wave, S' wave, P wave, T wave and QRSA , Sum of areas of all segments divided by 10 (Area= width * height / 2), QRSTA = QRSA + 0.5 * width of T wave * 0.1 * height of T wave [2][13]. Around 0.33 percent of data of the dataset is missing [2].

B. Data Refining and Categories

The dataset we found from UCI database, the columns are attributes and the rows are instances. It has some data missing for different attributes, as well instances. There are 10 instances where data is missing. That is why, we deleted them from the dataset before using it. The respective rows are 5, 66, 91, 200, 213, 238, 242, 360, 372 and 412. Moreover, the maximum data is missing in column 12, 14 and 199 which are respectively vector angles in degree P, vector angles in degree J and amplitude of channel AVR, T wave. Additionally 17 columns have 0.0 values. They are 20:(of channel DI) S' wave, 68:(of channel AVL) S' wave , 70:(of channel AVL) Existence of ragged R wave, 84:(of channel AVF) Existence of ragged P wave, 132:(of channel V4) Existence of ragged P wave, 133:(of channel V4) Existence of diphasic derivation of P wave, 140:(of channel V5) S' wave, 142:(of channel V5) Existence of ragged R wave, 144:(of channel V5) Existence of ragged P wave, 146:(of channel V5) Existence of ragged T wave, 152:(of channel V6) S' wave, 157:(of channel V6) Existence of diphasic derivation of P wave, 158:(of channel V6) Existence of ragged T wave, 165:(of channel DI) S' wave, 205:(of channel AVL) S' wave, 265:(of channel V5) S' wave, and 275:(of channel V6) S' wave. These columns were deleted just for predicting abnormal and normal ECG.

After this, we considered all data into two categories, normal and abnormal. Considering all ECGs of heart diseases as abnormal and normal ECGs are normal where normal is 1 and abnormal 0. Although for specific disease prediction the complete dataset is used without those 10 instances or rows mentioned before.

C. Algorithms

We have implemented six algorithms separately to analyze our data. They are as follows: Logistic regression (LR), Decision tree (DT), Nearest neighbour (NN), Naïve Bayes (NB), Support Vector Machine (SVM) and Arificial Neural Network (ANN) [14][15][16][17]. In Table I and II, the abbreviated form of these algorithms are used. Beside this, the abbreviated forms are also used in V as per need.

IV. EXPERIMENTAL SET UP

We divided the dataset to train and test our classifier with different algorithms that we have mentioned in III C. We also divided some sample data from the main dataset only to predict the disease of that particular data using our classifiers. At the beginning, we removed all the rows of the data set which had missing values, there were 10 such rows. Therefore our instances came down to 442 from 452. After that we did three different works here.

ISBN: 978-1-7281-0297-9 PART: CFP19TEO-ART

First, we extracted x and y where we consider x as the attributes and y as their respective classes of diseases. We normalized and scaled those data and train them through 6 different algorithm using cross validation. Subsequently, we did our second experiment. We considered all diseases in y are abnormal ECG valued 0 and normal ECG valued 1. We removed the columns where all values are '0'and columns with many missing data. We again scaled and normalized the data, trained them through 6 different algorithms using both random train test split and cross validation. Furthermore, we created 5 different data sets according to 5 different diseases. The attributes we took for each data set are related to those diseases. After that, we again trained those data set by the 6 different algorithms using cross validation, making 30 classifiers. Additionally, we used the sample data as input to test and the chances of a heart disease to occur.

V. RESULTS AND DISCUSSIONS

A. Results for normal and abnormal ECG

Fig 1 shows the cross validation score for predicting all 15 diseases without the removal of any missing data and Fig 2 and 3 shows the cross validation and random train-test split result for normal and abnormal ECG classified by the six algorithms mentioned in III C. The missing data are removed where needed. In both cases, the results are shown for both scaled and not scaled.

1) Cross Validation outcome including missing data: Fig 1 shows the cross validation score for all 15 diseases predicted by each algorithm for both with and without scaling. All the outcomes are below 70 percent. This is because it includes all the missing data for few attributes. Moreover, the data overlaps with each other while they are being calculated.

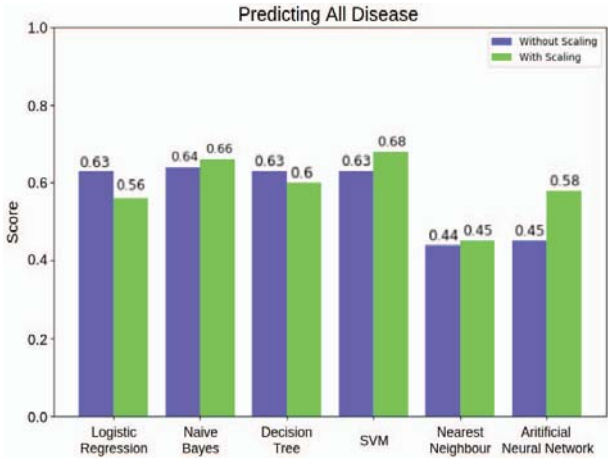


Fig. 1: Cross Validation Score of predicting all 15 diseases

2) Cross Validation outcome for normal and abnormal ECG: Fig 2 shows the cross validation score for normal and abnormal ECG for both with and without scaling. From this result, we can say that SVM gives the best score when scaling is used and Decision tree gives the best result when it is not scaled.

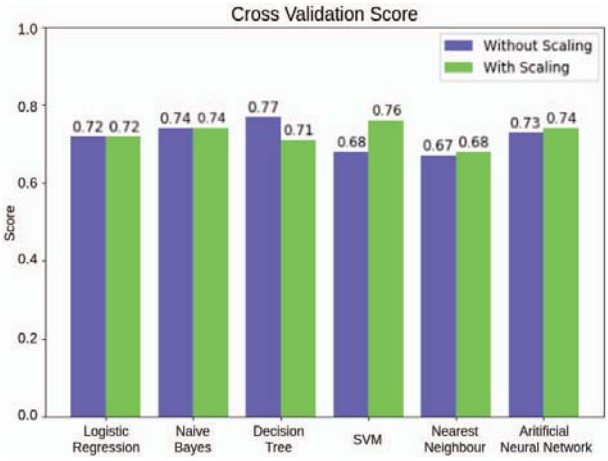


Fig. 2: Cross Validation Score for predicting normal and abnormal ECG

3) Random Train-Test Split outcome for normal and abnormal ECG: Fig 3 shows the random train-test split score for normal and abnormal ECG for both with and without scaling. From this result, we can say that SVM gives the best score when scaling is used and Decision tree gives the best result when it is not scaled.

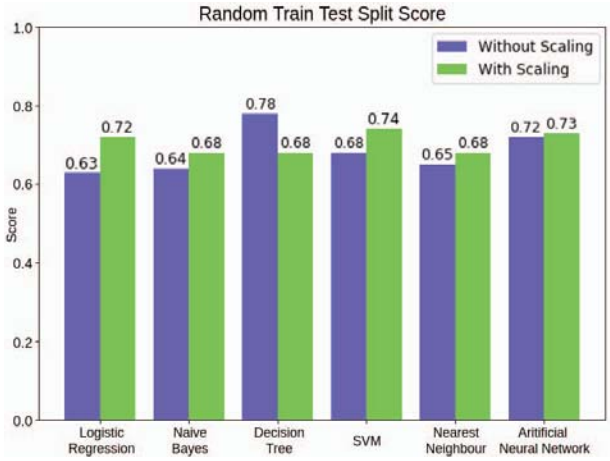


Fig. 3: Random Train-Test Split Score for predicting normal and abnormal ECG

B. Results for individual disease and algorithm

We chose 5 diseases to work with depending on the availability of the data for each disease. The diseases we worked with are Coronary artery disease(CAD), Myocardial infarction(MI), Sinus tachycardia(ST), Sinus bradycardia(SB) and Right bundle branch block(RBBB). In Table I, II and III, the abbreviated form of these diseases are used. Beside this, the abbreviated forms are also used in V as per need.

1) Cross Validation score: Here is the summary of the cross validation(CV) accuracy score in table I that we found by classifying each individual disease with different algorithms.

ISBN: 978-1-7281-0297-9

PART: CFP19TEO-ART

TABLE I: Cross Validation score for individual disease and algorithm

	LR	NN	DT	NB	SVM
CAD	0.86	0.76	0.87	0.94	0.86
MI	0.89	0.92	0.96	0.92	0.94
ST	0.95	0.64	0.95	0.95	0.95
SB	0.9	0.64	0.95	0.9	0.91
RBBB	0.85	0.8	0.88	0.9	0.92

Naive Bayes gives the a score above 0.90 for all the diseases. We got the lowest accuracy score from Nearest Neighbor algorithm which is 0.64 for both ST and SB. For Coronary Artery disease, NB gives the highest score of 0.94 and NN gives the lowest score which is 0.76. For Myocardia Infarction, DT gives the highest score of 0.96 and LR gives the lowest score which is 0.89. For Sinus Tachycardy, all the algo gives a score of 0.95 except NN which gives a score of 0.64. For Sinus Bradycardy, DT gives a highest score 0.95 and NN gives the lowest score which is 0.64. For RBBB, SVM gives the highest score 0.92 and NN gives the lowest score which is 0.8. Overall, Right Bundle Branch Block disease's scores is pretty consistent.

A line graph generated from the table I is shown in fig 4.

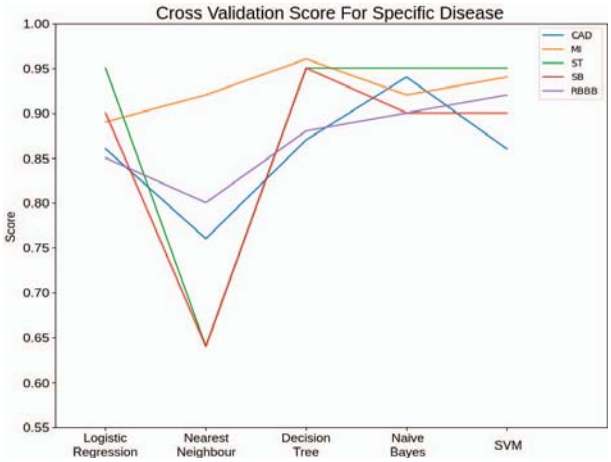


Fig. 4: Cross Validation Score for specific diseases

2) Random Train-Test Split score: Here is the summary of the Random Train-Test accuracy score in table II that we found by classifying each individual disease with different algorithms.

TABLE II: Random Train-Test Split score for individual disease and algorithm

	LR	NN	DT	NB	SVM
CAD	0.9	0.89	0.87	0.83	0.89
MI	1	0.96	0.99	0.91	0.99
ST	0.94	0.7	0.97	0.94	0.94
SB	0.91	0.69	0.94	0.91	0.96
RBBB	0.96	0.86	0.91	0.81	0.93

Logistic Regression gives the a score above 0.90 for all the diseases. We got the lowest accuracy score from Nearest Neighbor algorithm which is 0.67 for SB. For Coronary Artery disease, LR gives the highest score of 0.9 and NB gives the lowest score which is 0.83. For Myocardia Infarction, LR gives the highest score, 1 and NB gives the lowest score that is 0.91. For Sinus Tachycardy, DT gives the highest score, 0.97 and NN which gives a score of 0.7 which is the lowest. For Sinus Bradycardy, SVM gives a highest score 0.96 and NN gives the lowest score which is 0.69. For RBBB, LR gives the highest score 0.96 and NB gives the lowest score which is 0.81. Myocardial Infarction has got the best accuracy score for all the algorithms. Overall, all the algorithms gave relatively very good result for all diseases.

A line graph generated from the table II is shown in fig 5.

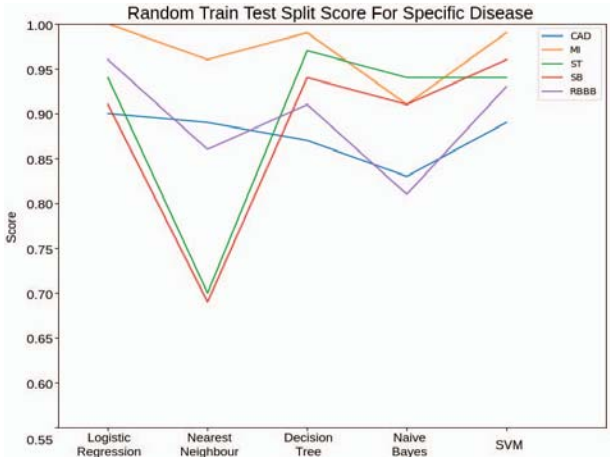


Fig. 5: Random Train-Test Split score for specific diseases

3) Cross Validation Score for ANN: Table III shows the Artificial Neural Network accuracy score of different neuron count. We used 3 hidden layers and every layers has the same number of neurons as given in table III. The range of the neuron is between 50 to 400. The iteration count is 5000. We can deduce that every disease has the accuracy score between 0.85 to 0.93.

TABLE III: CV score of ANN of individual disease vs Neuron count

	Neuron Count							
	50	100	150	200	250	300	350	400
CAD	0.854	0.861	0.861	0.861	0.861	0.861	0.865	0.865
MI	0.941	0.937	0.912	0.908	0.897	0.901	0.904	0.926
ST	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
SB	0.903	0.907	0.907	0.907	0.907	0.907	0.907	0.907
RBBB	0.931	0.925	0.88	0.907	0.928	0.915	0.918	0.914

A line graph generated from the table III is shown in fig 6.

ISBN: 978-1-7281-0297-9

PART: CFP19TEO-ART

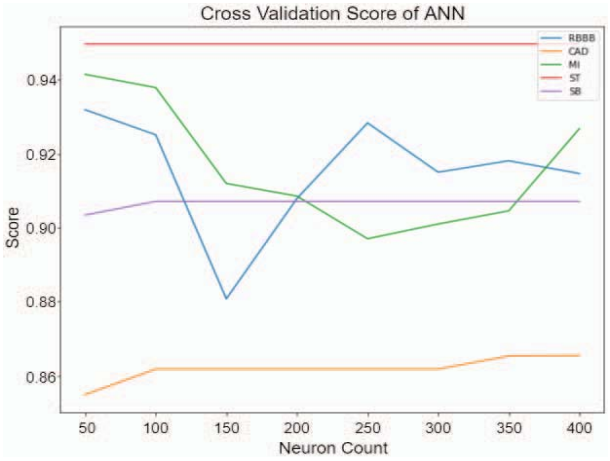
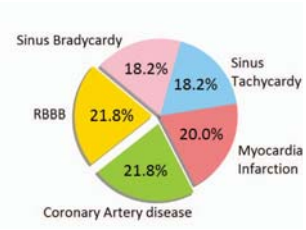


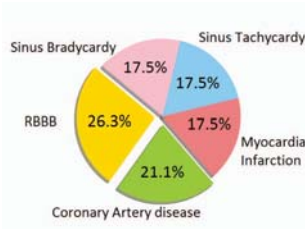
Fig. 6: Cross Validation Score of ANN for specific diseases

C. Prediction of disease with an input data set

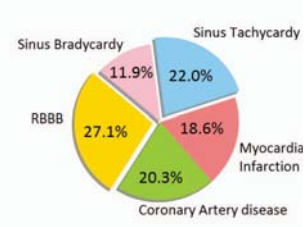
An unknown data set is used as an input for all the six algorithms to figure out the possible disease for this input. The result for each algorithm is shown in a different pie-chart below respectively. Each pie-chart shows the best two possible results.



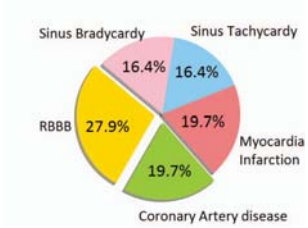
(a) Logistic Regression



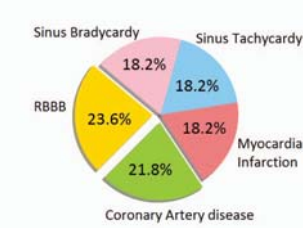
(b) Decision Tree



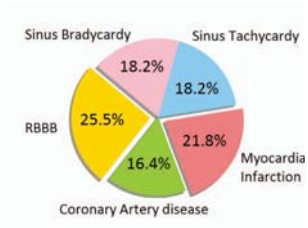
(c) Nearest Neighbour



(d) Naive Bayes



(e) SVM



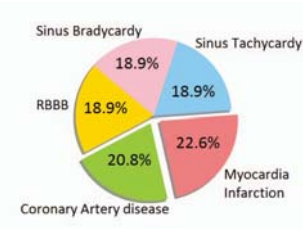
(f) Artificial Neural Network

Fig. 7: Prediction for all diseases using first input

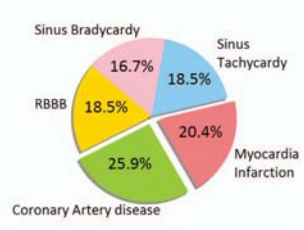
Considering the result we found from the pie charts in fig 7, this can be concluded that the input data set is a sufferer from Right Bundle Branch Block since all the algorithms give a maximum score for RBBB. However, in real life, when we tried to figure out the real disease, we found out that this input data was indeed a data set of RBBB (the person whose data we used was a sufferer of RBBB).

D. Prediction of disease with a second input dataset

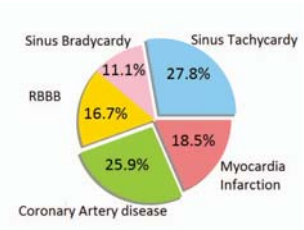
Another unknown data set is used as an input for all the six algorithms to figure out the possible disease for this input. The result for each algorithm is shown below in a different pie-chart below respectively. Each pie-chart shows the best two possible results similar to the previous sub section.



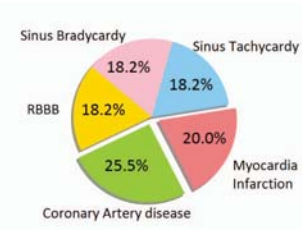
(a) Logistic Regression



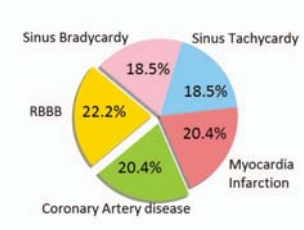
(b) Decision Tree



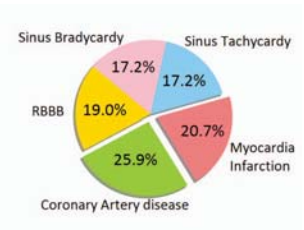
(c) Nearest Neighbour



(d) Naive Bayes



(e) SVM



(f) Artificial Neural Network

Fig. 8: Prediction for all diseases using second input

Considering the result we found from the charts in fig 8 this can be concluded that the input data set is a sufferer from Coronary Artery Block since all the algorithms gave the highest score for CAD. However, in real life, when we tried to figure out the real disease, we found out that this input data was indeed a data set of CAD (the person whose data we used was a sufferer of CAD).

E. Result Analysis

Therefore, by examining the results given in table IV, we can see that for predicting CAD it is best to use Naive Bayes Classifier since it has the highest score of 94% accuracy among all the algorithms. For predicting Myocardial Infarction, Decision Tree Classifier worked the best, the score being 96%. For Sinus Tachycardia all the algorithm has

ISBN: 978-1-7281-0297-9

PART: CFP19TEO-ART

a score of 95% except Nearest Neighbour. Decision Tree Classifier also worked well for Sinus Bradycardia with a score of 95%. And Lastly for Right Bundle Branch Block, Logistic Regression Classifier scored the highest, 96%.

TABLE IV: Best algorithm to use for individual disease and their score

Disease Name	Best Algorithm	Score
Right Bundle Branch Block	Logistic Regression	96%
Myocardial Infarction	Decision Tree	96%
Sinus Tachycardia	All except NN	95%
Sinus Bradycardia	Decision Tree	95%
CAD	Naive Bayes	94%

F. Comparison

In order to compare our models, we have used the same Cross Validation setting which was used by the Bilkent University [2]. We used 10-fold cross validation technique where we have divided the data set into 10 subsets. We used one set as testing data and rest of the 9 sets as training set. This process was repeated 10 times for each subset being the test set once. Then the average of the 10 results were taken in account.

In the experiment done by the Bilkent University, they used a algorithm called VFI5. The accuracy they found was only by using this algorithm which was 62%. They found an accuracy of 68% using genetic algorithm and VFI5. Their goal was to classify every disease. In contrast to their experiment, we further did 2 more experiments. Firstly we divided the data set into two groups. One group containing all the normal arrhythmia instances and other containing all the abnormal arrhythmia instances. We considered all the disease instance into one class and ran 6 different algorithms. The lowest accuracy that we found out is from Nearest Neighbour which is 68% and the highest is from Support Vector Machine which is 78%. On the other hand, when we isolated a single disease with the normal arrhythmia, our accuracy increases drastically. We got an accuracy result between 85% to 92% by classifying individual disease with different algorithms.

The cross validation accuracy scores were found automatically by our algorithm, however we had to use the equation 1 to calculate the accuracy score for in case of random train-test split whenever needed.

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

VI. CONCLUSIONS

The main aim of our research was to distinguish between normal and abnormal ECG by means of machine learning. In addition we figured out which algorithm works best for predicting a particular disease. We tried collecting data set from different mediums a couple of times, but the availability of dataset relating ECG is rare. Therefore, we used the data set we found in UCI Machine Learning Repository database to complete this. We faced a lot of trouble in some particular parts while doing our research, however we are happy that we managed to overcome them and reach at this level. While doing the research we figured out that the accuracy score for each algorithm increased when we did it for a particular disease rather than considering normal and abnormal ECG where abnormal ECG included all the 15 diseases. From our research, we found out that Logistic regression is the best algorithm to be used for Right Bundle Branch Block. Decision Tree gives the best result for Myocardial Infarction. All the algorithm except Nearest Neighbour can be used in case of Sinus Tachycardia. Decision tree is the best algorithm for Sinus Bradycardia. Lastly, Naive Bayes gives the best score for Coronary Artery Disease.

REFERENCES

[1] Neophytou, N., (2012). ECG event detection & recognition using time-frequency analysis (Doctoral dissertation).

[2] Guvenir, H. A., Acar, B., Demiroz, G., and Cekin, A. (1997). Supervised machine learningalgorithm for arrhythmia analysis.Computers in cardiology, pages 433–436.

[3] Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. InThe elementsof statistical learning, pages 485–585. Springer.

[4] Mahajan, S. (2014). Reinforcement learning: A review from a machine learning perspec-tive.International Journal, 4(8).

[5] Soman, T. and Bobbie, P. O. (2005). Classification of arrhythmia using machine learningtechniques.WSEAS Transactions on computers, 4(6):548–552.

[6] Grzymala-Busse, J. W. (1993). Selected algorithms of machine learning from examples.Fundam. Inform., 18:193–207.

[7] Roopa, C. and Harish, B. (2017). A survey on various machine learning approaches forecg analysis.Int J Comput Appl, 163(9).

[8] Jambukia, S. H., Dabhi, V. K., and Prajapati, H. B. (2015). Classification of ecg signalsusing machine learning techniques: A survey. InComputer Engineering and Applications(ICACEA), 2015 International Conference on Advances in, pages 714–721. IEEE.

[9] Asfaqul Islam, A. (2015). A real-time ecg warning system on myocardial infarction,hyperkalemia and atrioventricular block.

[10] Immanuel, J. J. R., Prabhu, V., Christopheraj, V. J., Sugumar, D., and Vanathi, P. (2012).Separation of maternal and fetal ecg signals from the mixed source signal using fastica.Procedia Engineering, 30:356–363.

[11] Zimmerman, M. W. (2004).Classification of ECG ST Events as Ischemic or Non-IschemicUsing Reconstructed Phase Spaces. PhD thesis, Marquette University.

[12] Fawcett, T. (2006). An introduction to roc analysis.Pattern recognition letters,27(8):861–874.

[13] Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel,M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python.Journal of Machine Learning Research, 12:2825–2830.

[15] Quinlan, J. R. (1986). Induction of decision trees.Machine learning, 1(1):81–106.

[16] Raschka, S. (2014). Naive bayes and text classification i-introduction and theory.arXiv preprint arXiv:1410.5329.

[17] Mitchell, T. M. et al. (1997). Machine learning. 1997.Burr Ridge, IL: McGraw Hill,45(37):870–877.