Principles of Data Mining and Machine learning

Title: Implementation of machine learning models on Electrocardiogram (ECG) dataset for Arrhythmia Classification

Major: MSc Computer Science

Student ID: 2229453

Name: Etim Ubongabasi Ebong

Abstract

Machine learning is a vast field of study that aims with the use of mathematical algorithms to create models from correlated data through a process that often involves training and testing to produce the most optimum model. This optimum model can then be used to make predictions about any future results from the correlated data. Machine learning is divided into two broad types, supervised and unsupervised learning. With supervised learning, models are built using previous data with labels and predefined outcomes and used to predict future occurrences. I.e., the previous data is used for processes called training and testing where the goal is to produce an optimal model that can then be used to make predictions. With unsupervised learning however, data is not labeled and there are no predefined outcomes.

Keywords: Machine Learning, Linear regression, Model, Prediction

INTRODUCTION

Problem Identification

Cardiac arrythmia is defined as a medical condition in which the heart beat of a patient is too fast, too slow or irregular. It is a condition that is prevalent across all age groups of patients and is caused by the presence of underlying medical issues like high blood pressure, thyroid problems, an imblance of electrolytes in the body and bad lifestyle choices among other causes.Some symptoms of cardicac arrythmia include but are not limited to shortness of breath, dizzyness and palpitations i.e, an abnormality in the beat of the heart*[1].*

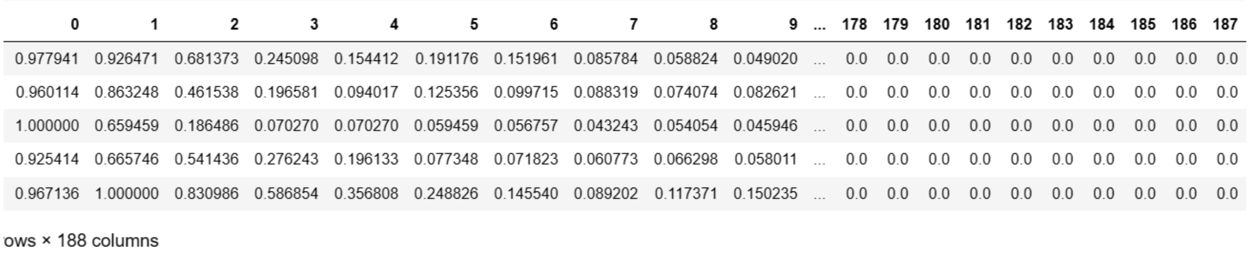
Ishcaemic heart disease is the leading cause of death worldwide according to the World Health Organization(WHO) [2]. This means that the accuracy in detection as well as early detection of heart disease is a is a very important step to be taken in the field of medicine in order to prevent more occurencies of death from heart disease. The use of machine learning models in detection and proper classification of different types of arrythmia is a much needed solution in this regard because it can ensure timely detection, improved accuracy and a higher level of convinience for medical practicioners and patients alike as it will save time when detecting heart disease which will inevitably lead to the prvention of deaths from heart disease.

The focus of this report is the documentation of the process of performing electrocardiogram (ECG) dataset analysis on an arrythmia dataset and the development of machine learning based classifiers for the purpose of classification of arrythmia heartbeat signals.

EXPLORATORY DATA ANALYSIS AND VISUALIZATION

Analysis and Visualization

The dataset used for this project was taken from the MIT-BIH Arrhythmia Kaggle dataset. [2]. It contains the records of forty-eight half-hour excerpts of two-channel ambulatory electrocardiogram (ECG) recordings, obtained from forty-seven subjects [3]. The dataset contains a training and testing set of 109446 total instances at a sampling frequency of 125Hz that divides arrythmias into five distinct categories.



0	1	2	3	4	5	6	7	8	9	...	178	179	180	181	182	183	184	185	186	187
0.977941	0.926471	0.681373	0.245098	0.154412	0.191176	0.151961	0.085784	0.058824	0.049020	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.960114	0.863248	0.461538	0.196581	0.094017	0.125356	0.099715	0.088319	0.074074	0.082621	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.000000	0.659459	0.186486	0.070270	0.070270	0.059459	0.056757	0.043243	0.054054	0.045946	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.925414	0.665746	0.541436	0.276243	0.196133	0.077348	0.071823	0.060773	0.066298	0.058011	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.967136	1.000000	0.830986	0.586854	0.356808	0.248826	0.145540	0.089202	0.117371	0.150235	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

ows × 188 columns

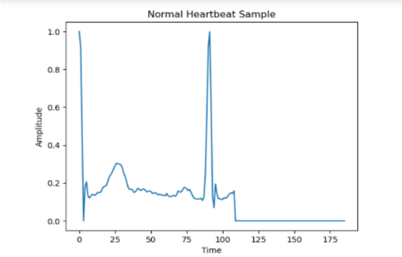
Figure 1 Merged dataframe containing train and test data for arrythmia classification

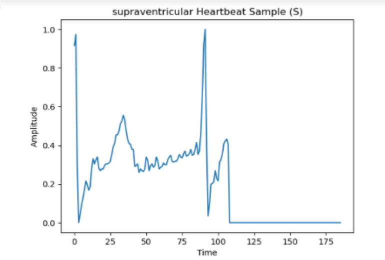
The aim of this project is to perform exploratory data analysis on this electrocardiogram (ECG) dataset and use machine learning classifiers including Random Forest Classifier, Multi Layer Perceptron and Support Vector Machines to accurately classify arrythmia efficiently. These machine learning methods will be implemented for use to classify different types of hearbeats into classes: ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4] where N is normal, S is supraventricular, V is ventricular, F is fusion and Q is unknown. The classes mentioned correspond to the 187th column of the dataframe above. Effectiveness of the methods will be determined by comparing carefully selected metrics including the accuracies, recall, confusion matrix and F-1 score. Oboservations will be made that suggest ways to futher improve the model performance. The dataset was merged and checked for null values to ensure that it did not contain any empty columns.

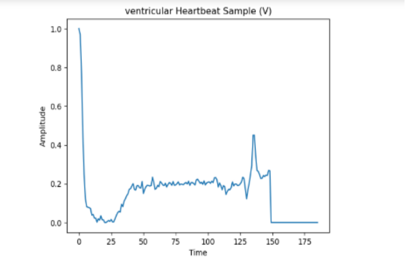
Data collection and Preprocessing

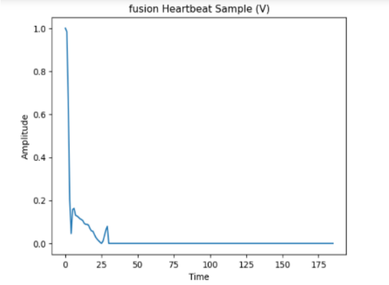
The prepeocessed and segmented dataset is is obtained from https://www.kaggle.com/datasets/shayanfazeli/heartbeat/data on Kaggle [3]. Each segment in the dataset corresponds to an individual heartbeat sample. These heartbeat samples are segmented into five arrythmia classes.

Below is a graphical representation of each invividual hearbeat from each of the classes in the order N,S,V,F,Q with the y axis signifying amplitude and the x axis the time in seconds.









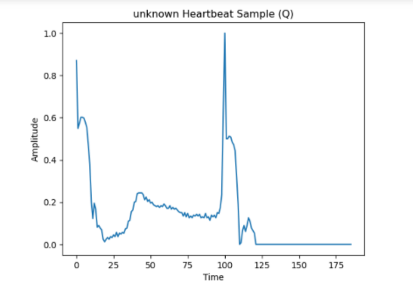
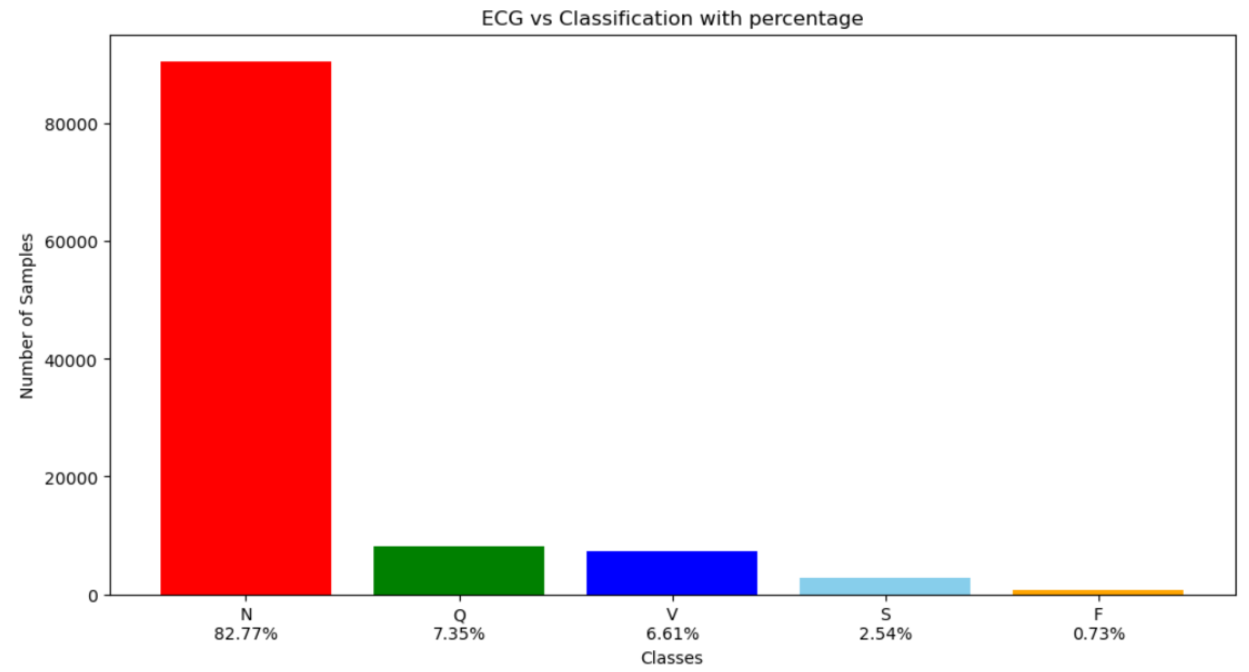


Table 1: Summary of MIT-BIH dataset classes[4]

Arrythmia class	Sample Amount (of 109446)	Sample percentage (%)
(N) Normal (0)	90589	82.77
(S) Supraventricular (1)	2779	2.54
(V) Ventricular (2)	7236	6.61
(F) Fusion (3)	803	0.73
(Q) Unknown (4)	8039	7.35

Table 1 shows a summary of the MIT-BIH arrythmia dataset using the classes feature. The majority of ECG heartbeat samples represented in the database is the normal samples with a total of 90589 samples accounting for more than 80% of all the samples in the dataset.



A look at the graph of the distribution of classes in the dataset shows that the dataset has a high amount of signals recorded as normal beats.

IMPLEMENTATION

The implementation of a machine learning approach for porperly classifying the MIT-BIH arrythmia dataset will employ conventional machine learning procedure in order to ensure that the dataset is ready for training and testing. This procedure includes the datacollection that has been done earlier, preprocessing of the dataset which includes merging the train and test set through concatenation, checking for columns with null and unique values, changing column names to prroperly match the description of the data in them, producing a correlation matrix that shows correlation coefficients between variables, seperating the inputs from the target variables and more.

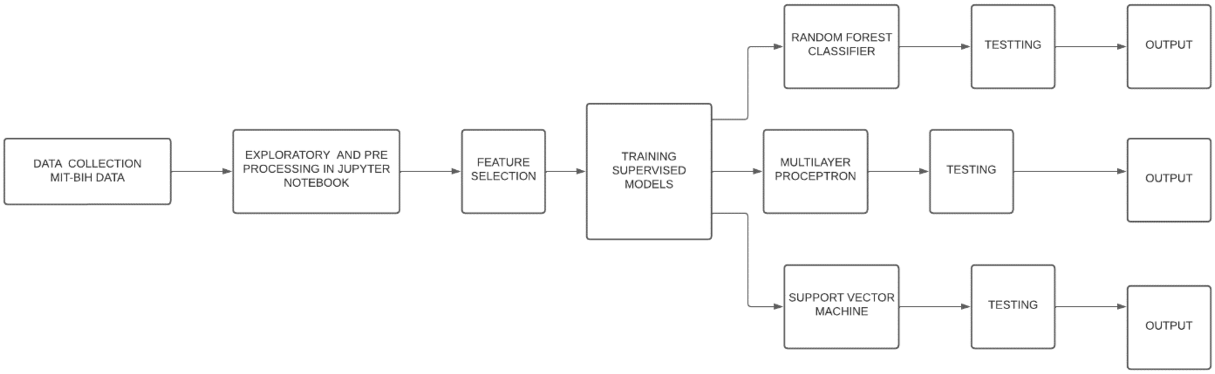


Figure 3 Visual representation of Machine learning pipeline.

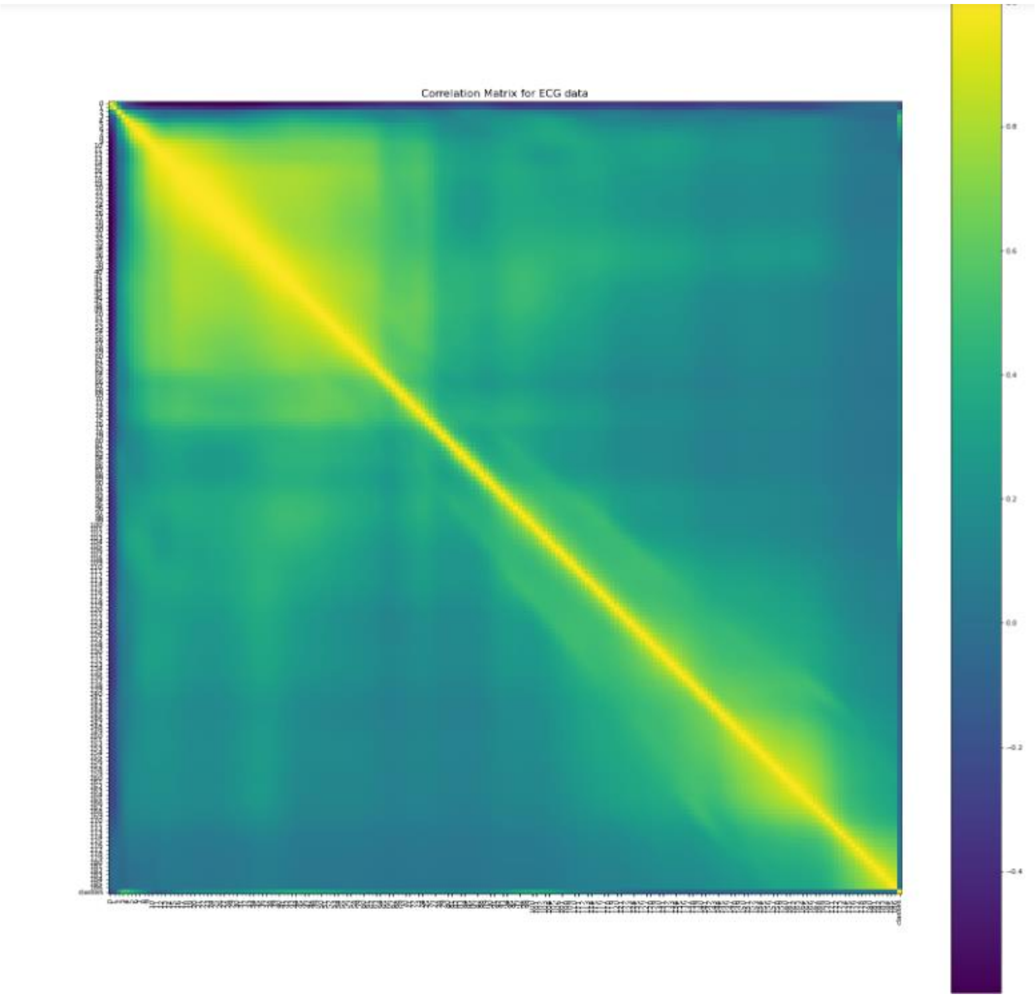


Figure 2 Correlation Matrix for MIT-BIH

Three machine learning classifier models have been developed to classify the different types of arrythmia, they include the Random Forest classifier, Multi layer Perceptron and Support Vector Machine.

Random Forest Classifier (RF)

Random Forest Classifier is known for its high level of adaptability as it can be used for both classification and regression problems. It has the ability to handle comlex datasets and mitigate overfitting(Sruthi E R)[5].

It works by usng the 'ensemble learning technique' where it makes use of multiple algoithms i.e., decision trees at random datapoints and using the average value of their predictions as the final result. The mathematical interpretation is thus:

Given training data $D = \{(\mathrm{X}_{i\,,}Y_i), i = 1 \ldots..n\}$ $where\ inputs\ are$ $(\chi_{i1}, \ldots., \chi_{ip})$ and targets are
y_i a random forest classifier generates a set of decision trees K where each tree k is built from a sample

D_k ,a sample drawn with replacement from the set D of size n. For predictions, $\hat{y}$ equals $\widehat{f(X)}$ for Random variables $(\chi_{i1}, \ldots, \chi_{ip})$ and y. Thus

$\hat{y} = \hat{f}\ (y_1(x), \ldots.\ y_k\ (x))$ where k is the ith decision tree (Adele Cutler)[6].

Multi Layer Perceptron (MLP):

The Multi Layer Perceptron is a type of deep artificial neural network that is comprised of three main layers. The input layer which receives inputs, hidden layer where all the processing of inputs is done using activation functions and output layers for dsplaying the predictions. The nodes in the input output and hidden layers are connected to each other through wieghts otherwise known as pearameters that are randomly chosen on the first iteration of training. Predicions are imporved throght a process called backpropagation, where the weights are updated to improve predictions (Prof. Alex Bronstein)[7]. It is often used for supervised regression and classification tasks, for image processing as well as for data where there is no linear relationship.

The mathematical illustration of a multi layer perceptron is as follows:

The output $\hat{y}$ of a MLP is given by the sum of weighted inputs. I.e., for input X ranging from $(x_i \ldots.\ x_n)$ the predictions $\hat{y} = \sigma(\sum_i^n (w_i . x_i))$.

Where σ is the activation function applied in the hidden layer i.e., sigmoid or relu functions, w_i $is\ the\ value\ of\ the\ ith\ weight\ and\ x_i\ is\ the\ value\ of\ the\ ith$ input.

Support Vector Machine (SVM):

Suport Vector Machines are a supervised machine learning cclassifier that is used for classification. It works by finding a hyperplane to classify data in an N-dimensional space.

The hyperplane is chosen in such a way that it maximizes the distance between the most extreme data points of each class. SVM handles classification of non linearly seperable data by using kernel functions which map the data unto a higher dimensional space. The mathematical illustration of Support Vector Machine is as follows:

Suppose we have a set of training points L where each input x_i has D attributes (is of dimensionality D) and is in one of two classes $y_i = +1$ or $y_i = -1$. The training data is of the form (x_i, y_i) where i ranges from 1 to L. $y_i \in (-1,\ +1), x \in R^D$. Assuming the data is linearly seperable we can draw a line graph of x_1 vs x_2 to separate the classes when the dimensionlity D is 2 and a hyperplane on graphs of $x_1, x_2 \dots., x_D$ when $D > 2$. The hyperplane is described by w . x + b = 0 where w is normal to the hyperplane and $\frac{b}{||w||}$ is the perpendicular distance from the hyperplane to the origin (Tristan Fletcher)[8].

Performing kernel trick to increase dimensionality

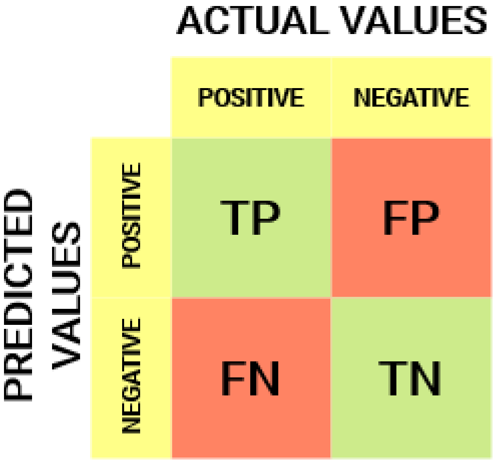
Given data $x_i \in R^D$ $where\ i \in (1,2,\dots.,n) and\ a\ kernel\ k: R^D \times R^D$ $a\ non\ linear$ transformation $\emptyset: R^D \rightarrow R^n$ $exists\ such\ that\ k(x_i, x_j) = \{\emptyset(x_i), \emptyset(x_j)\}$.(Jordan, Thibaux)[9]

RESULTS

In machine learning, model performance evaluation is key to assessing the effectiveness of a model. The evaluation of the three models employed in the classification of ECG signals is carried out by comparing the performance of each model using carefully selected performance metrics. The metrics to be considered are confusion matrix, precision, f1 score, accuracy score and reccall.

Confusion matrix:

A confusion matrix is a 2 dimensional representation of the classification accuracy of a model. It maps the number of true predictions to the total number of predictions for each class, highlighting where the true predictions and total predictions match. It displays the number of predictions true positives, true negatives, false positives and false negatives(Aniruddha Bhandari)[10].

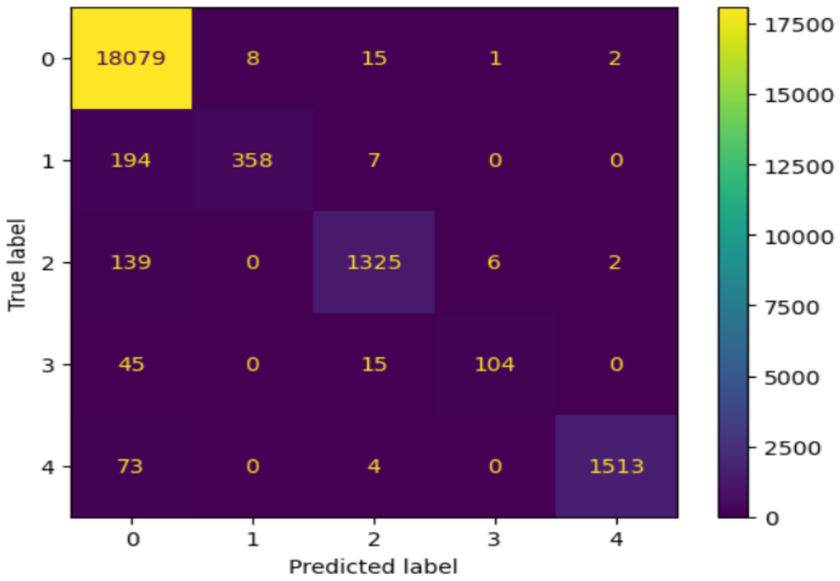


		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

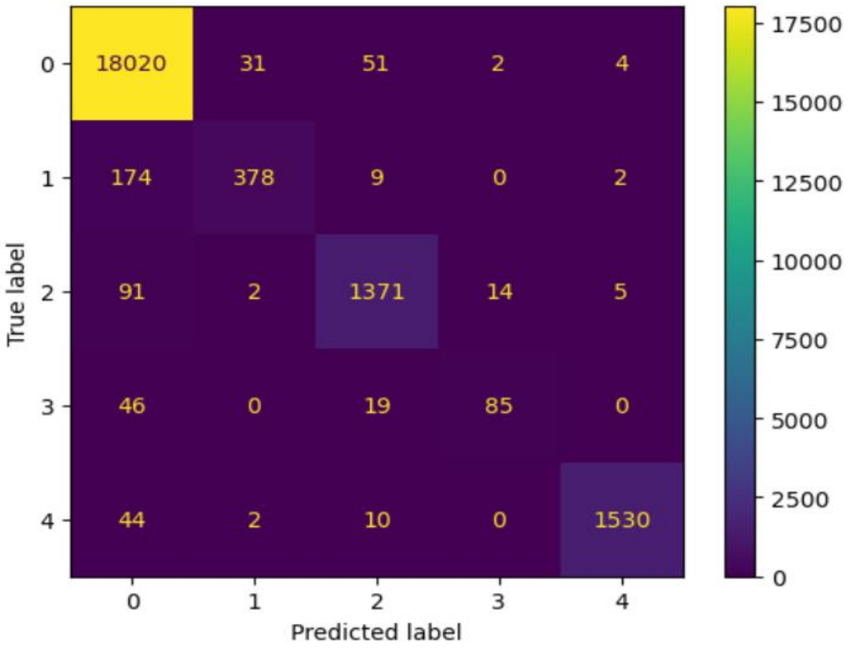
Figure 3 image of confusion matrix from analytics Vidhya [11]

The classes in the confusion matrices below conform to the ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4] classification of the ECG arrythmia data.

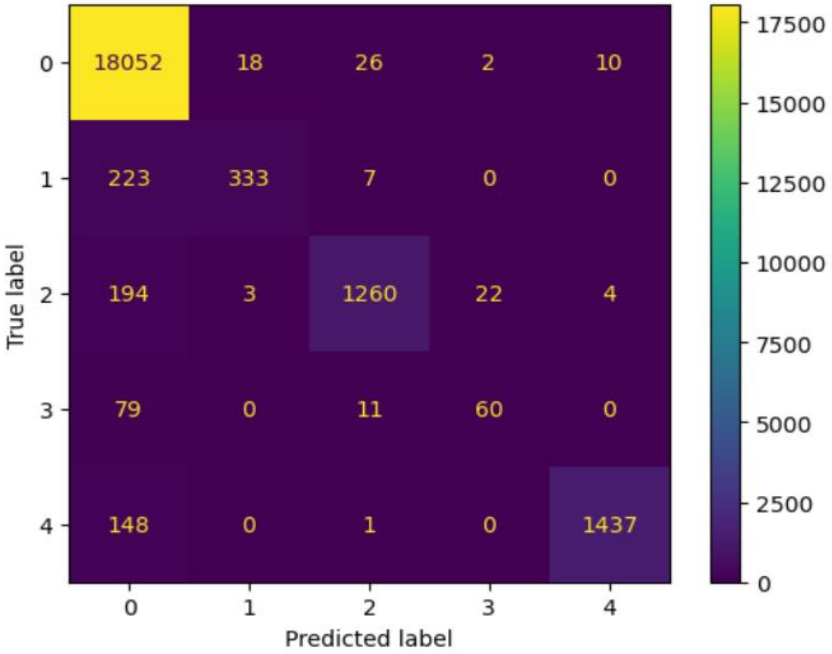
Random forest Confusion matrix:



Multi Layer Perceptron Confusion matrix:



SVM Confusion matrix:



The confusion matrices place the number of correct predictions diagonally and are all similar in their prediction accuracies with the most accurate being the random forest confusion matrix.

Accuracy

Accuracy is the total number of correct predictions out of all the preditions made.

It is the sum of the corrcet number of predictions correctly labelled as positive or correctlylabelled as negative divided by the total sum or predictions (Eugenio Zuccarelli)[12].

$$accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$$

Where TP is true positives, TN is true negatives, FP is false positives and FN is false negatives

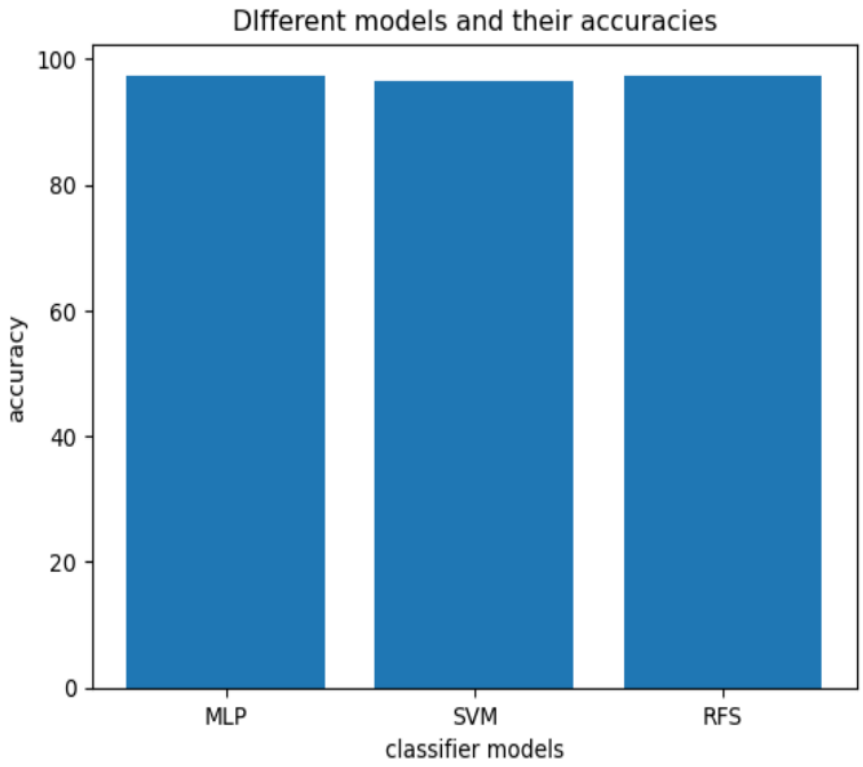


Figure 4 comparing all the model accuracies

The image above shows the comparison of accuracies between the three classifier models. The accuracy of all three classifier models are quite similar to each other.

Precision

Precision is usually known as a better prefomance metric for examining the classification ablilites of a classifier model on an imbalanced distribution of data like with the ECG dataset.

It is defined as the ratio of true positive predictions to the sum of all positive predictions(Eugenio Zuccarelli)[13].

$$precision = \frac{TP}{(TP\ +\ FP)}$$

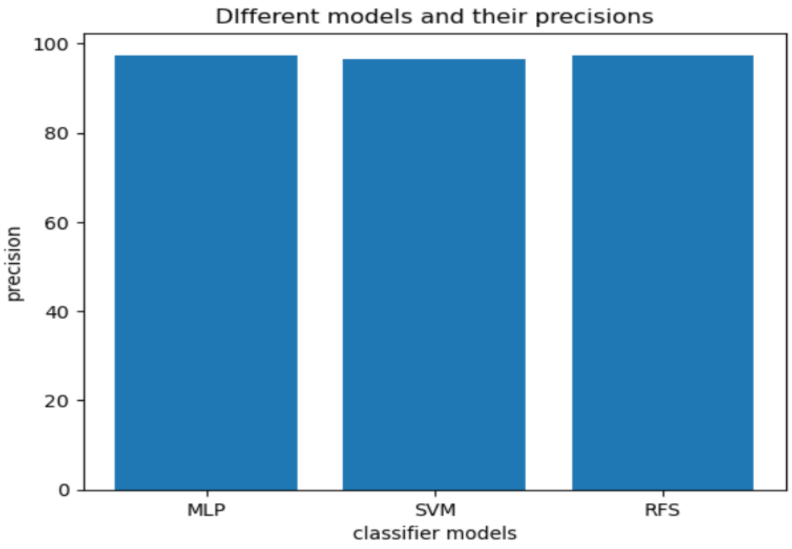


Figure 5 models and their precision scores

Recall

Recall is also a better performane metric for measuring thr performance of imblanced data than accuracy. It is the ratio of true positive predictions to the sum of true positives and false negatives i.e., wrong predictions (Eugenio Zuccarelli)[14].

$$recall = \frac{TP}{(TP\ +\ FN)}$$

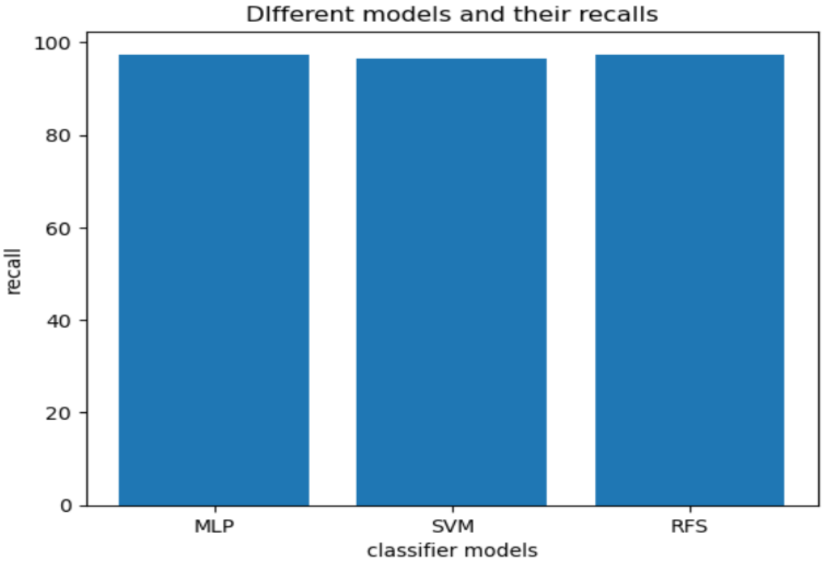


Figure 6 models and their recall scores

F1-score

The f1-score is also known as the hrmonic mean of the precision and recall metrics(Eugenio Zuccarelli)[15].

The formula for f1-score is given by

$$f1\ score = \frac{2 * precision * recall}{precision + recall}$$

The highest value of an f1-score is 1 and the lowest value is 0. All f1-scores have been multiplied by 100 to give their percentage value.

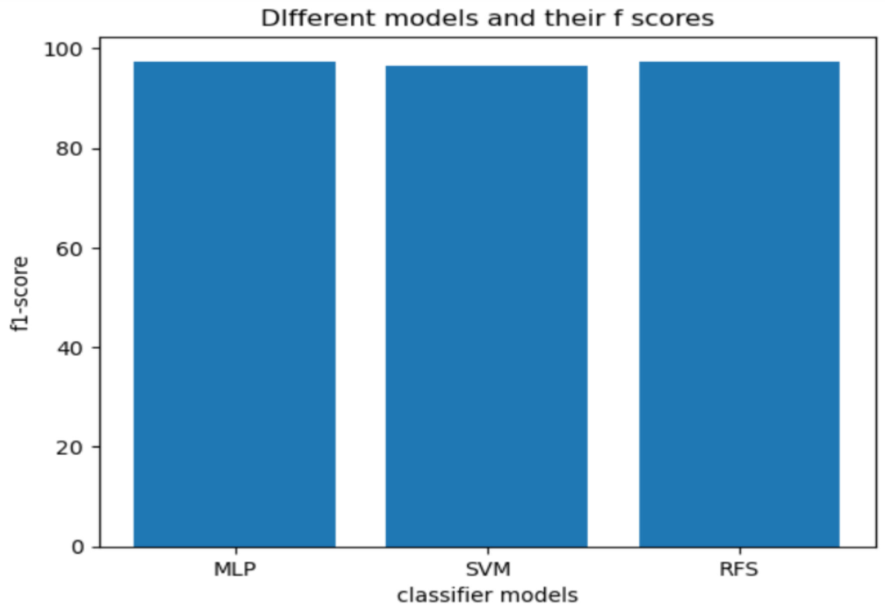
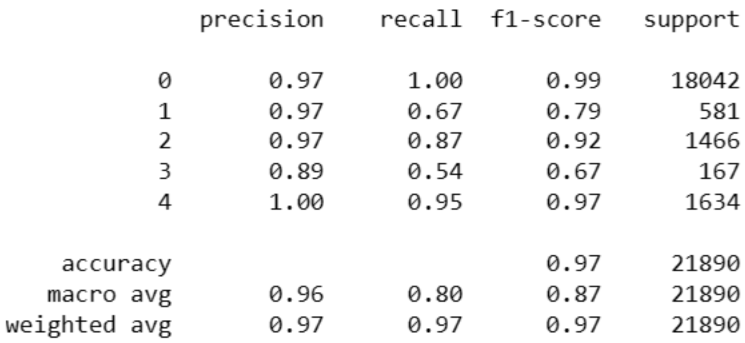


Figure 7 models and their f1 scores

Classification reports:

A classification report is a table that shows all the performance metrics for each model trained detailing how they performed in each class.

Below are the classification reports for each of the three classifier models.

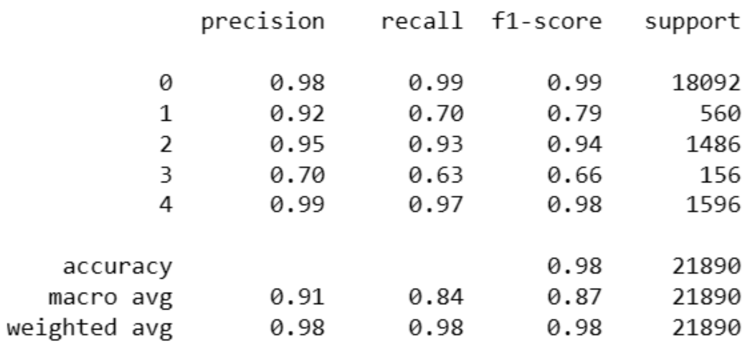


```
              precision    recall  f1-score   support

           0       0.97      1.00      0.99     18042
           1       0.97      0.67      0.79       581
           2       0.97      0.87      0.92      1466
           3       0.89      0.54      0.67       167
           4       1.00      0.95      0.97      1634

    accuracy                           0.97     21890
   macro avg       0.96      0.80      0.87     21890
weighted avg       0.97      0.97      0.97     21890
```

Figure 8 random forest classification report

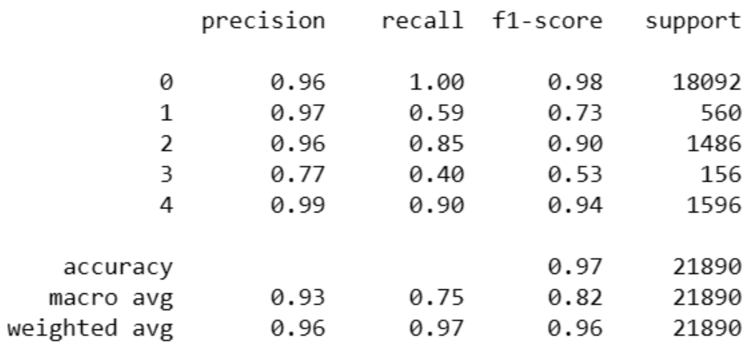


```
              precision    recall  f1-score   support

           0       0.98      0.99      0.99     18092
           1       0.92      0.70      0.79       560
           2       0.95      0.93      0.94      1486
           3       0.70      0.63      0.66       156
           4       0.99      0.97      0.98      1596

    accuracy                           0.98     21890
   macro avg       0.91      0.84      0.87     21890
weighted avg       0.98      0.98      0.98     21890
```

Figure 9 Multi Layer Perceptron Classification report



```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98     18092
           1       0.97      0.59      0.73       560
           2       0.96      0.85      0.90      1486
           3       0.77      0.40      0.53       156
           4       0.99      0.90      0.94      1596

    accuracy                           0.97     21890
   macro avg       0.93      0.75      0.82     21890
weighted avg       0.96      0.97      0.96     21890
```

Figure 10 Support Vector Machine Classification report

CONCLUSION

The data from the classification report shows that the highest accuracy recorded by a model is 0.98 or 98% which belongs to the MLP classifier. The other two classifiers have an accuracy score of 0.97 or 97% each. We can also observe that from the confusion matrices as well as the classfication reports, the performance metrics for class 3 which corresponds to fusion beats (F) are rather low compared to the confusion matrices and other performance metrics for the other classes. This is due to the fact that the dataset is highly unbalanced. A suggestion to improve this is to handle imbalanced data by using the upsampling technique where all the classes are made to have the same amount of data points as the largest class for better model perfromance on all classes. A second suggestion is to perfom label encoding where the numbers denoting the classes in the target variable y are replaced by the letters ['N': 0, 'S': 1, 'V': 2, 'F': 3, 'Q': 4] that adequately label them to help better visualise the data.

REFERENCES

1. SIGN 512. (2018). Cardiac arrhythmias in coronary heart disease: A national clinical guideline. Available at: https://www.sign.ac.uk/our-guidelines/cardiac-arrhythmias-in-coronary-heart-disease/ (Accessed: November 6, 2023).
2. World Health Organization (WHO). Available at: https://www.who.int/. (Accessed: November 6, 2023).
3. MIT-BIH Arrythmia Dataset. Available at: https://www.kaggle.com/datasets/shayanfazeli/heartbeat/data (Accessed: November 6, 2023).
4. MIT-BIH Arrhythmia Database v1.0.0 (physionet.org) (Accessed: November 10, 2023).
5. Sruthi E R. Analytics Vidhya, 2023. Available at https://www.analyticsvidhya.com/ (Accessed: December 5, 2023).
6. Adele Cutler. Random forests for classification and regression, 2010. (accessed: December5, 2023).
7. Prof. Alex Bronstein. Multi Layer Perceptron. Available at https://vistalab-technion.github.io/cs236605/lecture_notes/lecture_03/ (Accessed: December 7, 2023).
8. Tristan Fletcher. Support Vector Machines Explained (2008). (Accessed: December 9, 2023).
9. Micheal I. Jordan, Romain Thibaux. Advanced topics in learning and decision making. Available at lec3.pdf (berkeley.edu).(Accessed: December 10, 2023).
10. Aniruddha Bhandari. Analytics Vidhya, 2023. Available at https://www.analyticsvidhya.com/ (Accessed: December 10, 2023).
11. Analytics Vidhya, 2023. Available at https://www.analyticsvidhya.com/ (Accessed: December 10, 2023).
12. Eugenio Zuccarelli, Performance Metrics in Machine Learning Part1 (2020). Available at https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92. (Accessed: December 11, 2023)
13. Eugenio Zuccarelli, Performance Metrics in Machine Learning Part1 (2020). Available at https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92. (Accessed: December 11, 2023)
14. Eugenio Zuccarelli, Performance Metrics in Machine Learning Part1 (2020). Available at https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92. (Accessed: December 11, 2023)
15. Eugenio Zuccarelli, Performance Metrics in Machine Learning Part1 (2020). Available at https://towardsdatascience.com/performance-metrics-in-machine-learning-part-1-classification-6c6b8d8a8c92. (Accessed: December 11, 2023)
16.