2021 6th International Conference on Inventive Computation Technologies (ICICT) | 978-1-7281-8501-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICT50816.2021.9358733

Heart Failure Prediction by Feature Ranking Analysis in Machine Learning

Pushpavathi T P
Dept of CSE,MS Ramaiah
University of Applied Sciences,
Bengaluru, Karnataka, India.
acepushpa@gmail.com

Santhosh Kumari
Dept of CSE,MS Ramaiah
University of Applied
Sciences,Bengaluru, Karnataka,
India.
santhoshi.cs.et@msruas.ac.in

Kubra N K
M.Tech in Computer Science
and Networking.Dept. of CSE,
MS Ramaiah University of
Applied Sciences,Bengaluru,
Karnataka, India.
kubrank27@gmail.com

Abstract— **Heart disease is one of the major cause of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the field of clinical data analysis. With the advanced development in machine learning (ML), artificial intelligence (AI) and data science has been shown to be effective in assisting in decision making and predictions from the large quantity of data produced by the healthcare industry.ML approaches has brought lot of improvements and broadens the study in medical field which recognizes patterns in the human body by using various algorithms and correlation techniques. One such reality is coronary heart disease, various studies gives impression into predicting heart disease with ML techniques. Initially ML was used to find degree of heart failure, but also used to identify significant features that affects the heart disease by using correlation techniques. There are many features/factors that lead to heart disease like age, blood pressure, sodium creatinine, ejection fraction etc. In this paper we propose a method to finding important features by applying machine learning techniques. The work is to design and develop prediction of heart disease by feature ranking machine learning. Hence ML has huge impact in saving lives and helping the doctors, widening the scope of research in actionable insights, drive complex decisions and to create innovative products for businesses to achieve key goals.**

Keywords—Heart disease prediction; Feature selection;Machine Learning; Feature ranking; prediction model, classification algorithms, cardiovascular disease (CVD).

1. INTRODUCTION

Heart is one of the vital organ of the human body which is prone to Cardiovascular Diseases (CVD). The illness in the heart and the blood is because of coronary heart diseases which is generally addressed as heart attacks, cerebral vascular diseases called as strokes, heart failures and few other pathological disorders come under Cardiovascular Diseases. Approximately millions of deaths due to CVDs. United Kingdom having the highest rise in the mortality rate initially from 50 years [1]. When the heart is not able to pump sufficient blood few other disorder are noticed in the human body such as diabetes, high blood pressure, etc.

Patients paper chart is determined by the device that is a digital version called as an electronic health record (EHR). This information obtained is secured, real time and patient centered record which makes it easy for the medical staff. This helps to recognize hidden data and brings a correlation between patient data which can be used for clinical and research practices. This process helps in eliminating the traditions.

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Four out of 5CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

Individuals at risk of CVD may demonstrate raised blood pressure, glucose, and lipids as well as overweight and obesity. These can all be easily measured in primary care facilities. Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. Access to essential noncommunicable disease medicines and basic health technologies in all primary health care facilities is essential to ensure that those in need receive treatment and counselling.

Millions of people worldwide struggle to control the risk factors that lead to cardiovascular disease, many others remain unaware that they are at high risk. A large number of heart attacks and strokes can be prevented by controlling major risk factors through lifestyle interventions and drug treatment where necessary. The risk factors for CVD include behavioral factors, such as tobacco use, an unhealthy diet, harmful use of alcohol and inadequate physical activity, and physiological factors, including high blood pressure (hypertension), high blood cholesterol and high blood sugar or glucose which are linked to underlying social determinants and drivers, such as ageing, income and urbanization[14].

Currently various machine learning algorithms are used to predict the severity of heart diseases. These algorithms are Decision Tree, Naïve Bayes etc., these algorithms makes use of limited data like age, sex etc and segregate into different class depending on which the models are fabricated. The major contribution of these process is the prediction of heart failure but not to determine the root causes. There is no such techniques to predict the precise matching between cause and symptoms. Few researches have been done on earnestness and need of bug report. Currently structure the point to point envision of requirements using deep CNN and best fitting model are used. There are number of correlation techniques used to determine the particular reason behind the myocardial infarction (MI). To identify the reason using the many machine learning approaches are used like Convolution Neural Network ,Naïve Bayes, Random forest and K-Nearest Neighbor (KNN) etc. For obtaining the exact outcome with accuracy, the review matrices investigation and execution analysis are performed.

Organization of the paper is as follows, section 2 reviews on related work, mathematical preliminaries are reviewed and proposed schemes in section 3. Machine learning models are discussed in section 4. Section 5 explains design and implementation of proposed method. Finally Results and conclusions are described in section 6.

2. LITERATURE SURVEY

Around 26 million people have hospitalized due to acute heart failure. Then there are many investigations and analysis were performed on therapeutic and equivocal results. Unfortunately the solution obtained are very less for providing a good therapy for heart failure but the care remains the same. A new innovation and evolution of a new approach has taken place for clinical trials and therapeutic decision making of heart failure. Researchers have made developments on the application of novel therapies such as determination of clinical profile for every patient. But these profile help to determine the laboratory variables which cannot be used for building a personalized treatment that improve the patient's health. Therefore the clinical profiles created should be based on the aetiologu and the responses of the patient [2].

Sentilkumar et.al form a feature in with the use of machine learning approaches to increase the accuracy in prediction the CVDs. This approach is called Hybrid-Random-Forest-Linear-Model (HRFLM). The classification of errors, precision, F measure, sensitivity and specificity with high accuracy are determined using the above mentioned approach. Finally author conveyed that HRFLM approach gives better accuracy in predicting heart disorders. The major disadvantage is only theoretical reasoning and data's are available but real world data sets are absent [3].

Authors in [4] specified about finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The methodologies used are hybrid forest with a linear model (HRFLM). Their findings in research are classification error, precision, F-measure, sensitivity and specificity, the highest accuracy is achieved by HRFLM classification method in comparison with existing methods. Based on the findings the conclusion made was that HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Limitations of this paper are that there are no investigations to real-world data sets instead of just theoretical approaches and simulations.

Authors in [2] have specified cardiac patient monitoring system using the concept of Internet of Things (IoT) with different physiological signal sensors and Arduino microcontroller. The methodologies used are SVM, Random Forest, Simple Logistic models platform (WEKA) and Arduino based microcontroller system. Their findings in research are Support Vector Machine had highest accuracy compared with the other algorithms system and continuous sensor monitoring. Based on the findings the conclusions are that the design validates the possibility of an integrated cardiac patient monitoring system which can be used by the patient at home environment and enables the patient with an online centralized monitoring system. The limitations found are Photo plethysmography (PPG) based blood pressure sensor module or electronic sphygmomanometer is absent hence not effective in transmitting real-time data to the server.

Authors in [5] made comparison of different machine learning algorithms for the prediction of coronary artery disease.The methodologies used are Smote algorithm, Support Vector Machine, Neural Network Algorithms,logistic regression.Their findings in research are that the Support Vector Machine performed better than the other two Algorithms as it had the highest accuracy. Drastic improvements in results were seen after applying SMOTE. Based on the findings the conclusion is that the performance of Support Vector Machine and Artificial Neural Network significantly improved when trained on the balanced data set however, the overall accuracy of Logistic Regression stayed the same on both sets of data the limitations found are working on a larger data set with more features could be a better extension to this research.

Authors in [4] built a predictive model that provides doctors and health care providers with personalized information .The methodologies used are Logistic Regression, Classification Tree with Bagging, Random Forest, Support Vector Machine and K-Nearest Neighbors. Their findings in research is that the RF model has highest sensitivity 0.9623, KNN model lowest sensitivity 0.9245.SVM model accuracy 0.8947, sensitivity 0.9434, specificity 0.7826, AUC 0.8868.Based on the findings the conclusion is that SVM model is able to predict the presence of CAD more effectively and accurately than other models with high accuracy,sensitivity,specificity and AUC.the limitations found are that further research might be necessary to improve in the performance of the machine learning algorithm before this method is translated into clinical solution.

Authors in [6] worked on effective classification for prediction of heart disease combining KNN with genetic algorithm.The methodologies used are K nearest neighbor(KNN),Genetic algorithm.Their findings in research is that the accuracy of heart disease data is decreased by 32% using cross validation,Accuracy of the heart disease is increased by 5% using and GA using full training data set and 15%improvement in accuracy for cross validation against KNN without GA.Based on the findings the conclusion is that integrating GA with KNN out performs the other methods with greater accuracy. the limitations found are that integrating GA along with SVM or random forest would give higher results.

Authors in[6][7] worked on prediction systems for Heart disease using more number of input attributes.The methodologies used are Neural Networks, Decision Trees, Naive Bayes.Their findings in research is that the Naive Bayes accuracy for 13 and 15 attributes: 94.44 and 90.74,Decision Tree accuracy for 13 and 15 attributes:96.66 and 99.62 neural Networks accuracy for 13&15 at:99.25 100.Based on the findings the conclusion is that Neural Networks provides accurate results as compare to Decision trees &Naive Bayes.the limitations found are that the system does not find the attribute which highly affects the heart failure from given attributes. Authors in [8] Emphasis on discovering patterns that explains the data to be interpreted by humans .Predictions in data mining in medical field to help patients.The methodologies used are simple cart,rep tree,naive baye's, bayesnet,j48 algorithms.Their findings in research are J48, SIMPLE CART, REPTREE ALGORITHM classification techniques,Bayes Net algorithm out-performed Naive Bayes algorithm. J48, SIMPLE CART and REPTREE provide more predictive accuracy than other algorithms.Based on the findings the conclusion is that predictive accuracy determined by REPTREE, J48 and Bayes Net algorithms propose parameters used are consistent indicator to predict the heart diseases. the limitations found are that more parameters must be considered for better prediction.

Authors in[9] worked on a heart disease prediction model based on the machine learning approach which enables predicting heart disease with 95% accuracy. The methodologies used are decision tree,svm,naive bayes,random forest, logistic regression,QDA.Their findings in research SVM with a peak accuracy 95%,LR and QDA (94%) achieved second-highest classification accuracy. Based on the findings the conclusion is that SVM outperformed all the other methods with 95% accuracy.the limitations found are that this system does not include critical factors for women such as menopause etc which can be the major factors affecting heart disease.

Authors in[10] devised an NN-based prediction of CHD risk using feature correlation analysis.The methodologies used are statistical analysis for KNHANES-VI dataset, FCA.Their findings in research is feature correlation analysis accuracy (82.51%) in a CHD prediction, useful than the FRS applied in the past. Based on the findings the conclusion is that the proposed model will improve the CHD risk and decision support for suitable treatment. Sex, hemoglobin, thyroid disease, H_B,H_C, and cirrhosis were not associated, whereas triglyceride and CRF were closely related to CHD.

the limitations found are that The data is based on Korean people, hence this accuracy would apply only to koreans and not to the world.

Authors in[11] used ImageNet Classification is used with Deep Convolutional method.The methodologies used are dropout technique,CNN,Neural Networks.Their findings in research is Setting output of each hidden neuron to 0 with 0.5 probability and those neurons are dropped out. Based on the findings the conclusion is that the performance reduces even if 1 layer of CNN is reduced.The limitations found are that results are improvised when faster GPU's and bigger data sets are available.

Authors in [12] discussed deep learning system to screen diseases. The methodologies used are RT-PCR detection, CT scan using CNN. Their findings in research is 3D CNN model, Image classification model, Noisy- Bayesian Function performed well to screen diseases. Based on the findings the conclusion is that the combination of Resnet, location and attributes leads to more accuracy. The limitations found are that improvised Image processing can be used for better analysis. Literature survey inspire us to use the deep learning methodologies which helps in tracing the bugs in prior stages itself. This helps to increase software qualities in market. The model uses high number of portions on the classification to upgrade the generality.

The deep learning models have proved to be a promising procedure for programmed highlight extraction which includes different layers in the design. It shows very efficiently classify various unpredictable information by using the middle level numerous hidden layers. Random forest known for its capacity of extracting and learning, is used in taking semantic highlights from the symbolic vectors extracted from the myocardial infarction also known as heart failure prediction.

3. MATHEMATICAL PRELIMINARIES

Correlation Analysis is evaluating the relationship between the two features of the dataset of the bug reports and achieve the correlation among the feature of the bugs. To the extent the nature of relationship, the estimation of the correlation coefficient fluctuates among +1 and - 1. An estimation of ±1 shows a perfect degree of connection between the two elements. As the correlation coefficient regard goes towards 0, the association between the two components of bug will be increasingly delicate. The degree of the correlation relationship if exposed as negative and positive coefficient with aid of its operator sign. In real time there are frequently used correlation techniques are Spearman, Pearson and the Kendall correlation. The item underneath licenses for the comfortable relationship measure.

Pearson's correlation: This correlation used as highly recommended techniques to estimate the relationship among the various feature of the data elements and bring high degree with respect to correlation. For example, in the money related trade, to measure how two stocks are related to each other, Pearson r association is used to evaluate the degree of association between the two. The point-biserial association is driven with the Pearson relationship condition beside that one of the components is dichotomous. The going with formula is used to figure the Pearson r correlation.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

Spearman's rank correlation: Spearman rank correlation is a non-parametric test that is used to check the degree of connection between two elements. The Spearman rank correlation test doesn't pass on any speculations about the course of the data and is the best possible correlation assessment when the components are evaluated on a scale that is in any occasion ordinal. The going with formula is used to figure the Spearman rank correlation:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ = Spearman rank correlation
d_i = the difference between the ranks of corresponding variables
n = number of observations

Kendall's rank correlation: This correlation evaluates the dependency among data features and its non-parametric in nature. If it consider two models, a and b, where every model size is n, we understand that the full-scale number of pairings with a b is n(n-

1)/2. The going with condition is used to find out the estimation of Kendall rank correlation:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad \text{------------} \quad (3)$$

n_c = number of concordant
n_d = number of discordant

Considering the positive component assurance is the most fundamental workmanship in AI which makes the enormous multifaceted nature between normal and good model. Feature getting sorted out is the course toward changing unsavory data into features subject to best correlation score that better area the secured issue to the sharp models, recognizing improved model precision on covered data.Next, training data set: data which is utilized for the training phase and test data set: data which is utilized for the testing phase guarantees that test set meets the two conditions: whether adequately gigantic to retrieve the enormous amount of results. Whether agent of the enlightening assortment all in all In a manner of speaking, not to pick a test set with startling characteristics in contrast with the planning set. Expecting that test set meets the previous two conditions, it will probably make a model that summarizes well to new data. Test set fills in as a mediator for new data.

4. Design and Implemetation

The data flow is relatively designed for the proper analysis of data. The data is loaded for correlation analysis, after finding relation between each feature correlation score is found out, based on the correlation score the features are ranked. Machine learning models are applied on the features selected. System design is an important development for any proposed system to be productive. Noteworthy level arrangement gives the system's raised level of perspective and functionality. Figure1shows the proposed framework and figure 2 shows the experiment workflow with dataset.

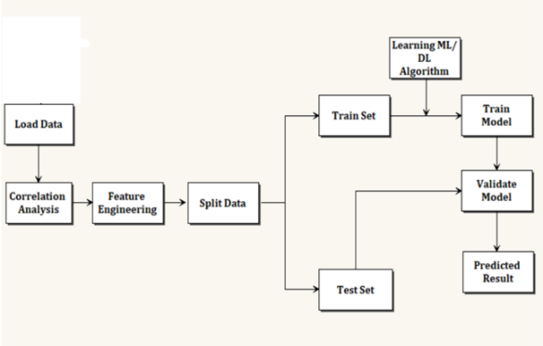


Fig 1: Framework of the proposed model

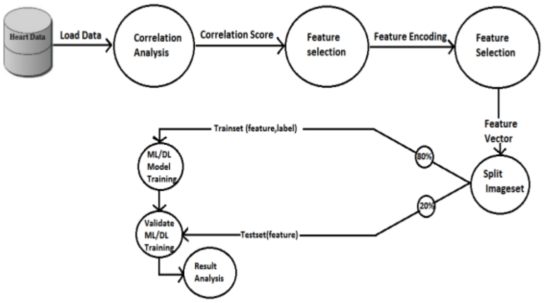
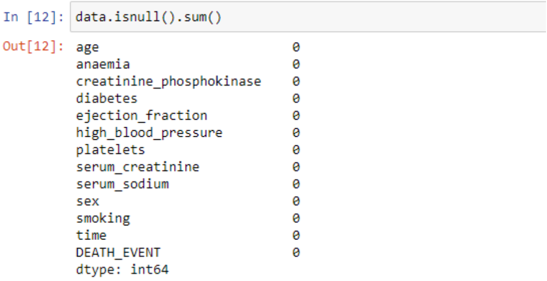


Fig 2: Experiment workflow with dataset

4.1 Data Preprocessing

Data Set: The proposed system consists of open source data sets namely extracted from kaggle about bug repository. The data set is accessed from https://www.kaggle.com/zeeshanmulla/heart-disease-dataset. These data sets were downloaded as comma-separated values (csv) files. The dataset used in this experiment is above mentioned link. Heart disease data is pre-processed after collection of various records. The dataset contains a total of 303 patient records, where 6 records are with some missing values. Those 6 records have been removed from the dataset and the remaining 297 patient records are used in pre-processing. The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to 1, else the value is set to 0 indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical records into diagnosis values. The results of data pre-processing for 297 patient records indicate that 137 records show the value of 1 establishing the presence of heart disease while the remaining 160 rejected the value of 0 indicating the absence of heart

disease.Pre-processing such as null value analysis over the data set before proceeding with correlation among the features. Null value analysis is to remove the jargons or any error in the data.
This data set was formed after biomedical analysis tests done for various patients taking into consideration the main features leading to heart attack. The medical tests were done by medical professionals based on features such as age, anaemia, creatinine phosphokinase as shown below.



```
In [12]:  data.isnull().sum()

Out[12]:  age                          0
          anaemia                      0
          creatinine_phosphokinase     0
          diabetes                     0
          ejection_fraction            0
          high_blood_pressure          0
          platelets                    0
          serum_creatinine             0
          serum_sodium                 0
          sex                          0
          smoking                      0
          time                         0
          DEATH_EVENT                  0
          dtype: int64
```

Fig 4:Null Analysis

4.2 Correlation Analysis-Pearson

Correlation techniques are done to compare the top features and then to relate how one feature is related to another feature. In this project we have related various featured with each other so that there is precise feature ranking. The first technique we used is Pearson. This technique measures the statistical relationship of the features based on the values obtained.

Followed by null analysis before proceeding the modeling correlation techniques are applied among all the feature of bug analysis using Pearson correlation figure 5.

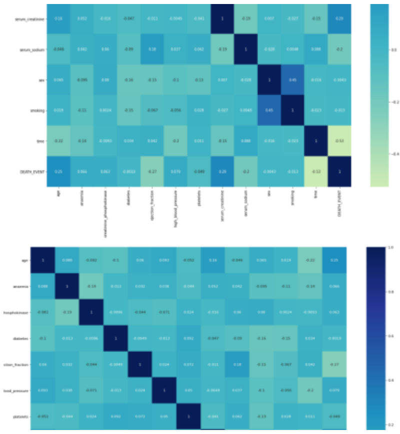
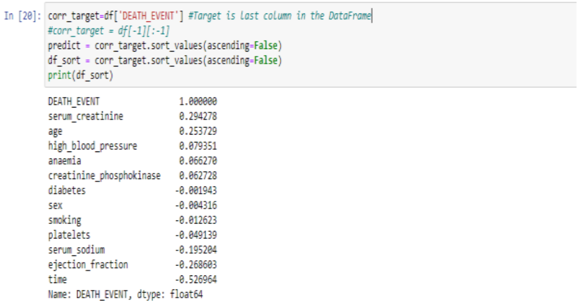


Figure 5. Pearson Correlation with all features

We analyze dataset to understand and summarize main characteristics of the data. Here the scale of measurement is taken in a ratio of 1:1, when we correlate, r=1, it's a positive correlation hence ranks the features accordingly. It helps in seeing what the data infers even before applying any modelling or hypothesis testing tasks. Hereby we implement the correlation of the target feature priority Fig 6 with all other feature of the bug so that respective features are selected for the model building.

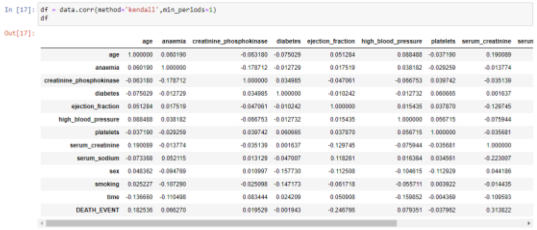


```
In [20]:  corr_target=df['DEATH_EVENT'] #Target is last column in the DataFrame
          #corr_target = df[-1][:-1]
          predict = corr_target.sort_values(ascending=False)
          df_sort = corr_target.sort_values(ascending=False)
          print(df_sort)

          DEATH_EVENT                 1.000000
          serum_creatinine            0.294278
          age                         0.253729
          high_blood_pressure         0.079351
          anaemia                     0.066270
          creatinine_phosphokinase    0.062728
          diabetes                   -0.001943
          sex                        -0.004316
          smoking                    -0.012623
          platelets                  -0.049139
          serum_sodium               -0.195204
          ejection_fraction          -0.268603
          time                       -0.526964
          Name: DEATH_EVENT, dtype: float64
```

Fig 6 Pearson correlation for priority Target

4.3 Correlation Analysis-kendall

In Kendall rank correlation we use a coefficient to measure the ordinal association between two measured features, here the rank comes high when features have similar rank or features. Followed by null analysis before proceeding the modeling correlation techniques are applied among all the feature of bug analysis using kendall correlation fig 7

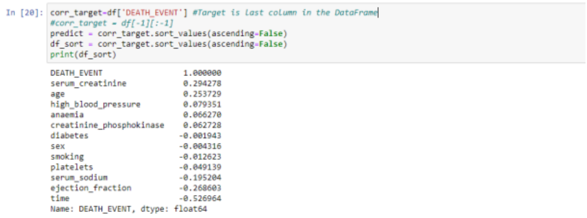


```
In [17]:  df = data.corr(method='kendall',min_periods=1)
          df

Out[17]:
```

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serun
age	1.000000	0.060190	-0.063180	-0.075029	0.051284	0.088488	-0.037190	0.190089	
anaemia	0.060190	1.000000	-0.178712	-0.012729	0.017519	0.038182	-0.029259	-0.013774	
creatinine_phosphokinase	-0.063180	-0.178712	1.000000	0.034985	-0.047061	-0.066753	0.039742	-0.035139	
diabetes	-0.075029	-0.012729	0.034985	1.000000	-0.010242	-0.012732	0.060665	0.001637	
ejection_fraction	0.051284	0.017519	-0.047061	-0.010242	1.000000	0.015435	0.037870	-0.129745	
high_blood_pressure	0.088488	0.038182	-0.066753	-0.012732	0.015435	1.000000	0.056715	-0.075944	
platelets	-0.037190	-0.029259	0.039742	0.060665	0.037870	0.056715	1.000000	-0.035681	
serum_creatinine	0.190089	-0.013774	-0.035139	0.001637	-0.129745	-0.075944	-0.035681	1.000000	
serum_sodium	-0.073368	0.052115	0.013129	-0.047007	0.118261	0.016364	0.034561	-0.223007	
sex	0.048362	-0.094769	0.010997	-0.157730	-0.112508	-0.104615	-0.112929	0.044186	
smoking	0.025227	-0.107290	-0.025098	-0.147173	-0.061718	-0.055711	0.003922	-0.014435	
time	-0.136660	-0.110498	0.083444	0.024209	0.050908	-0.159852	-0.004369	-0.109593	
DEATH_EVENT	0.182536	0.066270	0.019529	-0.001943	-0.246766	0.079351	-0.037962	0.313822	

Fig 7.Kendall Correlation with all features

Here every feature is correlated with each other and as show in the fig 7, there is a rank given on how one feature may be responsible for presence of another. Hereby we implement the correlation of the target feature priority Fig 8 with all other feature of the bug so that respective feature can be selected for the model building

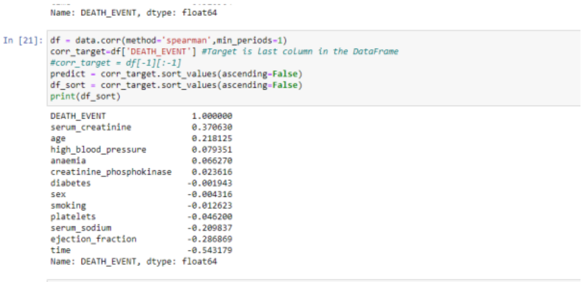


```
In [20]: corr_target=df['DEATH_EVENT'] #Target is last column in the DataFrame
         #corr_target = df[-1][:-1]
         predict = corr_target.sort_values(ascending=False)
         df_sort = corr_target.sort_values(ascending=False)
         print(df_sort)

         DEATH_EVENT                 1.000000
         serum_creatinine            0.294278
         age                         0.253729
         high_blood_pressure         0.079351
         anaemia                     0.066270
         creatinine_phosphokinase    0.062728
         diabetes                   -0.001943
         sex                        -0.004316
         smoking                    -0.012623
         platelets                  -0.049139
         serum_sodium               -0.195204
         ejection_fraction          -0.268603
         time                       -0.526964
         Name: DEATH_EVENT, dtype: float64
```

Fig 8. Kendall correlation for priority Target

4.4 Correlation Analysis-Spearman

Correlation in spearman is done using monotonic function. we correlate the features monotonically if the relationship between the feature are not linear especially when the data is roughly and elliptically distributed. We find the spearman coefficient r for both continuous and ordinal variables. Hereby we implement the Spearman correlation of the target feature priority Fig 9 with all other feature of the bug so that respective feature can be selected for the model building



```
         Name: DEATH_EVENT, dtype: float64

In [21]: df = data.corr(method='spearman',min_periods=1)
         corr_target=df['DEATH_EVENT'] #Target is last column in the DataFrame
         #corr_target = df[-1][:-1]
         predict = corr_target.sort_values(ascending=False)
         df_sort = corr_target.sort_values(ascending=False)
         print(df_sort)

         DEATH_EVENT                 1.000000
         serum_creatinine            0.370630
         age                         0.218125
         high_blood_pressure         0.079351
         anaemia                     0.066270
         creatinine_phosphokinase    0.023616
         diabetes                   -0.001943
         sex                        -0.004316
         smoking                    -0.012623
         platelets                  -0.046200
         serum_sodium               -0.209837
         ejection_fraction          -0.286869
         time                       -0.543179
         Name: DEATH_EVENT, dtype: float64
```

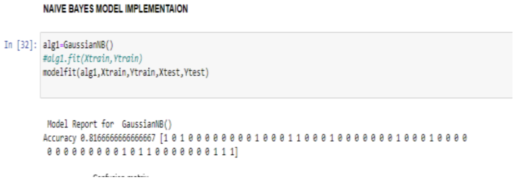
Fig 10 Spearman correlation for priority Target

5. RESULTS AND DISCUSSIONS

After 3 correlations we find that the features such as serum creatinine, age and high blood pressure are the features which highly affect heart failures and features such as ejection fraction and serum sodium have least impact on the heart failure. Base on this the doctors can work on the features having high affect and save the patient.

After all the feature engineering we implement various machine learning algorithm such as Naive Bayes,KNN using sklearn package.All the models are trained using Training data set(X_Train,Y_Train) using fit() function and test the model with X_Test data using predict() function .we developed a modelfit() which incorporate model building and accuracy calculation with confusion matrix as part of result analysis in the figure 11.

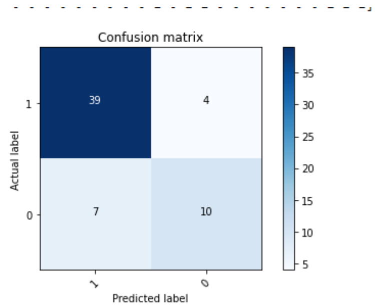
Naive Bayes Accuracy:



NAIVE BAYES MODEL IMPLEMENTAION

```
In [32]: alg1=GaussianNB()
         #alg1.fit(Xtrain,Ytrain)
         modelfit(alg1,Xtrain,Ytrain,Xtest,Ytest)

          Model Report for  GaussianNB()
         Accuracy 0.8166666666666667 [1 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
          0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 1 1]
```

Naive Bayes Confusion Matrix:

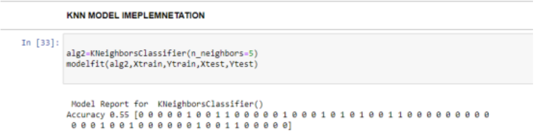


```
confusion Matrix [[39  4]
 [ 7 10]]
```

Fig 11 Machine Learning model

In tis confusion matrix 39 features match with training set and 10 do not match at all, 4 and 7 alternatively matches the features.

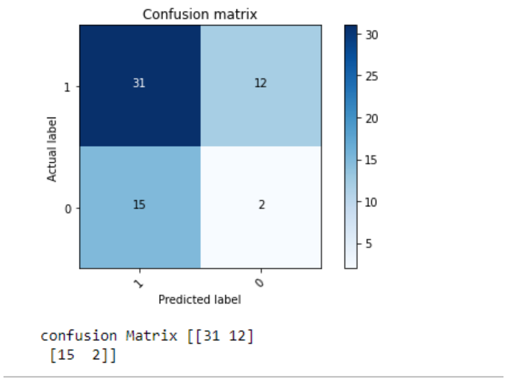
K nearest neighbor accuracy:



KNN MODEL IMEPLEMNETATION

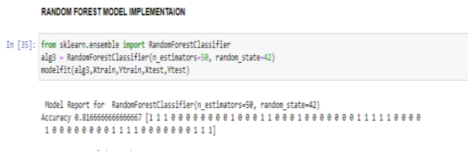
```
In [33]:
         alg2=KNeighborsClassifier(n_neighbors=5)
         modelfit(alg2,Xtrain,Ytrain,Xtest,Ytest)

          Model Report for  KNeighborsClassifier()
         Accuracy 0.55 [0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0
          0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0]
```

KNN confusion matrix:



```
confusion Matrix [[31 12]
 [15  2]]
```

Random forest accuracy:

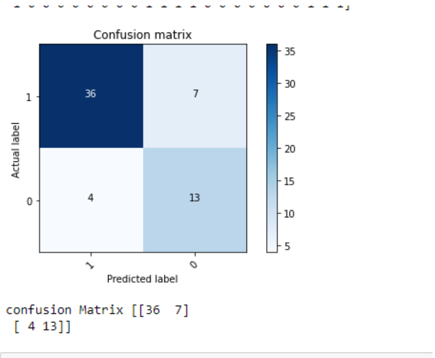


RANDOM FOREST MODEL IMPLEMENTAION

```
In [35]: from sklearn.ensemble import RandomForestClassifier
         alg3 = RandomForestClassifier(n_estimators=50, random_state=42)
         modelfit(alg3,Xtrain,Ytrain,Xtest,Ytest)
```

```
 Model Report for  RandomForestClassifier(n_estimators=50, random_state=42)
Accuracy 0.8166666666666667 [1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0
 1 0 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1]
```

Random forest confusion matrix:



```
confusion Matrix [[36  7]
 [ 4 13]]
```

Accuracy can be defined as performance measure. It is a measure of correctly predicted observations to the sum of the observations. If we have high accuracy in our model, then the model is performing better. Accuracy is calculated with the following equation 1.

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} \quad \text{Eq (1)}$$

Table 2: Priority accuracy

Classifier	Accuracy-Priority
	Target
KNN	55%
NaiveBAYES	81%
Random Forest	81.6%
CNN	99.6%

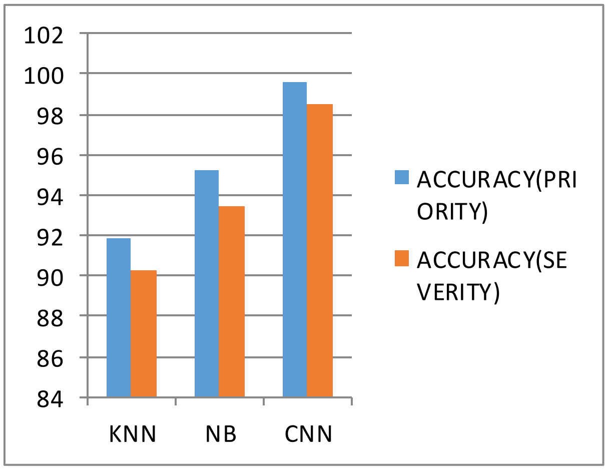


Fig 19 Comparison of ML and DL –Severity and Prirorty

From the above figure 19 out of the 3 algorithms CNN shows highest accuracy with 99.6% mentioning that it has predicted accuracy of those selected features more accurately compared to KNN, Naïve Bayes and random forest. CNN being the dep learning algorithm has various neural convolutions hence resulting in more efficiency. In machine learning techniques Random forest prediction was about 81.6% being the most accurate compared to the other ml techniques.

6. CONCLUSION

Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. The convolution neural system is successful in classification, another profound learning model has been proposed, called bug severity and priority classification by means of convolution neural network. We implemented multi model system whereas as we evaluate both machine learning algorithms like KNN Naive Bayes with deep learning algorithm. However, in future we improve single model approach for both priority and severity prediction of the bug reports. Finally, performance evaluation is carried out in terms of accuracy, precision and recall metrics and their performance is comparable. CNN proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques. Furthermore, new feature selection methods can be developed to get a broader perception of the features to increase the performance of heart disease prediction.

REFERENCE

1. M. S. Amin, Y. K. Chiam, K. D. Varathan, ''Identification of significant features and data mining techniques in predicting heart disease,'' Telematics Inform., vol. 36, pp. 82–93, Mar. 2019. [Online]. Available: https:// linkinghub.elsevier.com/retrieve/pii/S0736585318308876.
2. The Guardian. UK heart disease fatalities on the rise for first time in 50 years. https://www.theguardian.com/society/2019/may/13/heart_x0002_circulatory-disease-fatalities-on-rise-in-uk. Accessed 25 Oct 2019.
3. National Heart Lung and Blood Institute (NHLBI). Heart failure. https://www.nhlbi.nih.gov/health-topics/heart-failure. Accessed 20 June 2019.
4. S. Mohan et al.: Effective Heart Disease Prediction Using Hybrid ML Techniques, VOLUME 7, 2019,DOI 10.1109/ACCESS.2019.2923707
5. Dipto, I. , Islam,T. , Rahman, H. and Rahman, M. (2020) Comparison of Different Machine Learning Algorithms for the Prediction of Coronary Artery Disease. Journal of Data Analysis and Information Processing, 8, 41-68. doi: 10.4236/jdaip.2020.82003.
6. M. AkhiljabbaraB.L ,Deekshatulub, PritiChandrac, Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm, Elsevier,Procedia Technology,Volume 10, 2013, Pages 85-94, https://doi.org/10.1016/j.protcy.2013.12.340
7. Chaitrali S. Dangare ,Sulabha S. Apte 2012 A Data Mining Approach for Prediction of Heart Disease Using Neural Networks, International Journal of Computer Engineering and Technology (IJCET), Volume 3, Issue 3, October-December 2012.
8. M.C.S.Geetha, Dr.I.Elizabeth Shanthi, Ms.N. Sanfia Sehnaz Heart Disease Prediction using Machine Learning, international journal of engineering research & technology(IJERT), volume 09, issue 08 (august 2020).
9. Maruf Ahmed Tamal et,al 2019, Heart Disease Prediction based on External Factors: A Machine Learning Approach, International Journal of Advanced Computer Science and Applications(IJACSA), Volume 10 Issue 12, 2019.
10. Jae Kwon Kim and Sanggil Kang (2017), Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis, Hindawi Journal of Healthcare Engineering Volume 2017, Article ID 2780501.
11. Alex Krizhevsky,Ilya Sutskever,Neural Networks 2012. ImageNet classification with deep convolutional neural networks, NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 December 2012 Pages 1097–1105, https://dl.acm.org/doi/10.5555/2999134.2999257
12. Charmaine Butt, Jagpal Gill, David Chun, and Benson A. Babu, Deep learning system to screen coronavirus disease 2019 pneumonia, Appl Intell. 2020 Apr 22 : 1–7. doi: 10.1007/s10489-020-01714-3
13. https://towardsdatascience.com/heart-disease-prediction-73468d630cfc
14. https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1