2021 International Conference on Intelligent Technologies (CONIT)
Karnataka, India. June 25-27, 2021

Heart Failure Prediction using Machine Learning with Exploratory Data Analysis

Harsh Agrawal
Narsee Monjee Institute of Management Studies
Mumbai, India
harshsdw@gmail.com

Janki Chandiwala
Narsee Monjee Institute of Management Studies
Mumbai, India
janki.chandiwala@gmail.com

Sarvesh Agrawal
Narsee Monjee Institute of Management Studies
Mumbai, India
sarveshagrawal.profile@gmail.com

Yash Goyal
Narsee Monjee Institute of Management Studies
Mumbai, India
yashdineshgoyal@gmail.com

2021 International Conference on Intelligent Technologies (CONIT) | 978-1-7281-8583-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CONIT51480.2021.9498561

***Abstract*— According to WHO, cardiovascular diseases are the number 1 cause of death globally. It causes the death of more than 12 million people every year worldwide. The main issue that needs to be resolved is that one should be warned well before time to take precautionary measures. Thus, in this paper, we propose a radical solution based on ensemble learning combining 10 different classification algorithms namely AdaBoost, CatBoost, Decision Trees, KNN, Logistic regression, Light GBM, Gaussian Naïve Bayes, Random Forest, SVM and XGBoost. This ensemble model was able to achieve a test accuracy of 85.2% and test recall of 87.50%. We used the data collected from the Framingham Heart study which includes 15 attributes and 4200+ records. Moreover, we performed extensive Exploratory Data Analysis to understand the importance of each attribute in causing heart failure.**

Keywords*— *Classification, Data Visualization, Ensemble Modeling, Exploratory Data Analysis, Heart Attack, Machine Learning

I. INTRODUCTION

Cardiovascular disease is the term for all sorts of diseases that have an effect on the heart or blood vessels, as well as coronary heart condition (clogged arteries), which may cause heart attacks, stroke, noninheritable heart defects and peripheral artery unwellness. CVDs are the number 1 cause of death globally. An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke [2]. Estimates from the World Health Organization (WHO) show that by 2030, CVDs will be the main cause of death throughout India, accounting for more than 35% of all deaths [3]. Moreover, more than 800,000 people die of cardiovascular disease every year in the United States [1]. Heart Attack is a complex clinical syndrome that occurs due to the heart's inability to pump an adequate supply of blood to the body. All major body functions are disrupted without sufficient blood flow. One suffers a heart attack when the flow of oxygen rich blood is restricted to a certain section of heart which causes lack of enough oxygen. During a coronary failure, plaque will rupture and spill cholesterol and different substances into the blood. A grume forms at the positioning of the rupture. If the clot is giant, it will block blood flow through the arterial blood vessel, starving the heart of chemical element and nutrients (ischemia) that causes the heart muscles to die.

Three sorts of arteria malady will result in a heart failure. These are: ST section elevation infarct (STEMI), Non-ST section elevation infarct (NSTEMI) and arteria spasm. STEMI attack happens once the arteria is totally blocked, preventing blood from reaching an outsized space of the center. NSTEMI heart attack occurs when the coronary artery is partially blocked and blood flow is severely restricted. Coronary artery spasm happens once the arteries connected to the heart contract, limiting blood flow to the heart. These diseases can be caused depending on various factors such as age, gender, high blood pressure, high cholesterol, obesity, diabetes, sedentary lifestyle etc. [4].

In recent times, Computer-aided diagnosis using artificial intelligence-based solutions has become increasingly popular [5]. Machine learning-based algorithms have started to become the quality choice for medical data analysis and classifications because of the availability of a large amount of healthcare data. It is so because machine learning algorithms can easily identify meaningful trends and patterns in them and classify the data [6]. For example, in our case, we were able to draw inference from the results obtained by applying various ML algorithms that elderly people with high BP were most prone to heart failure. Similarly, it is being used to detect various diseases such as Kidney Failure [7], Ebola Outbreak [8], Covid-19 [9] etc.

This paper proposes a Machine Learning based Ensemble Solution to predict the probability of having heart failure in the next ten years. We began by implementing various ML algorithms: Logistic Regression, Gaussian Naïve Bayes, KNN, Support Vector Machine, Decision Tree, Random Forest, Adaptive Boosting, Light Gradient Boosting Machine, XGBoost and CatBoost and then we created an ensemble model combining all these models which gave the best results and we achieved 87.5% recall and 96.4% precision on test data. Adding to this, we also performed extensive Exploratory Data Analysis and gained insights on the correlation of various factors with a heart attack.

The rest of the paper is organized as follows: The first part provides detailed information about the dataset used, followed by extensive data analysis on this data. Then we describe the methodology used, followed by the result and the conclusion.

II. Dataset

In this paper, we used the dataset provide on Kaggle based on ongoing Framingham Heart Study (FHS) [17]. This study is being conducted on residents of Framingham, Massachusetts. The dataset provides the patients' information. It includes 4,238 records (3594 records for healthy class and 644 records for people at risk from CHD). We split the data in the train and test part by applying stratified split of 25% for test data. Hence our training set consists of 3178 records and the testing set consists of 1060 records. The purpose of the classification is to assess if the patient has a 10-year chance of potential coronary heart disease (CHD). The dataset consists of 15 attributes which include-

TABLE I. Attribute Description Table

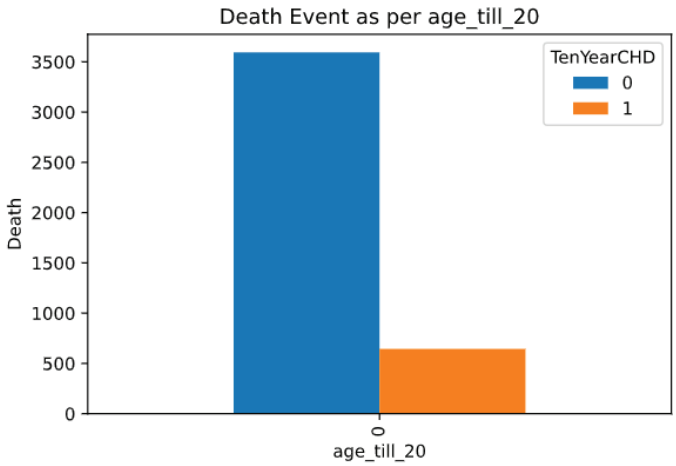
Attribute	Description
Age	Patient's Age (Continuous)
Blood Pressure Meds	If the patient was prescribed blood pressure medication (Nominal)
Body Mass Index	Patient's BMI (Continuous)
Cigarettes per day	Number of cigarettes consumed in a day (Continuous)
Current Smoker	If the patient smokes currently or not (Nominal)
Diabetes	If the patient suffers from diabetes (Nominal)
Diastolic Blood Pressure	Patient's Diastolic BP (Continuous)
Glucose	Patient's glucose level (Continuous)
Heart Rate	Patient's average heart rate (Continuous)
Prevalent Hypertension	If the patient is hypertensive or not (Nominal)
Prevalent Stroke	If the patient previously suffered from a stroke (Nominal)
Sex	Patient's gender (Nominal)
Systolic BP	Patient's Systolic BP (Continuous)
Total Cholesterol	Patient's average cholesterol level (Continuous)
Ten Year CHD	10-year risk of coronary heart disease (binary)

III. Exploratory Data Analysis

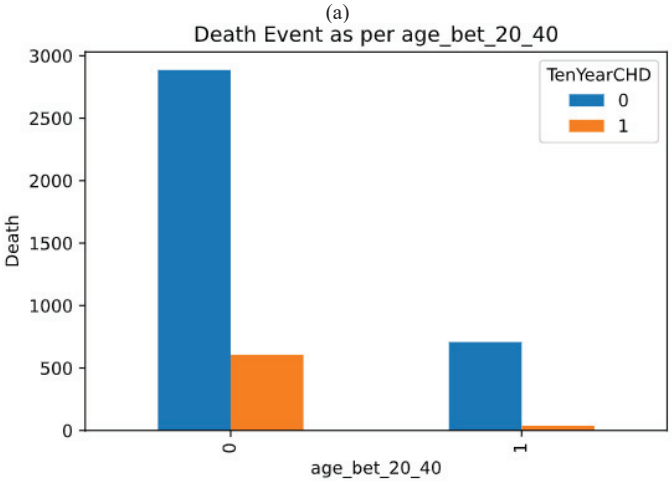
We performed extensive exploratory data analysis on the attributes mentioned above and obtained meaningful insights about the impact of these attributes on heart disease.

A. Age

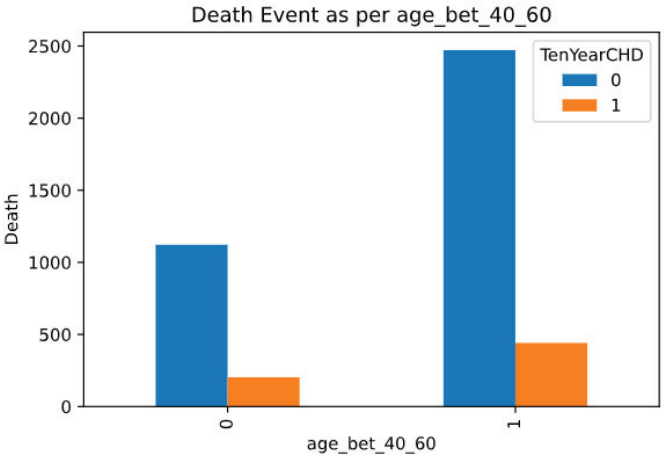
People within the given age range are depicted by 1 whereas 0 depicts the number of people not in that age range in Figure 1. The percentage of people who died and had an age between 20 to 40 is 5.09 whereas the percentage of people who died and had an age greater than 60 is 28.49. The mortality rate of people with age ranging from 40 to 60 is 15.13. This shows that younger people have a better immune system than elderly people.



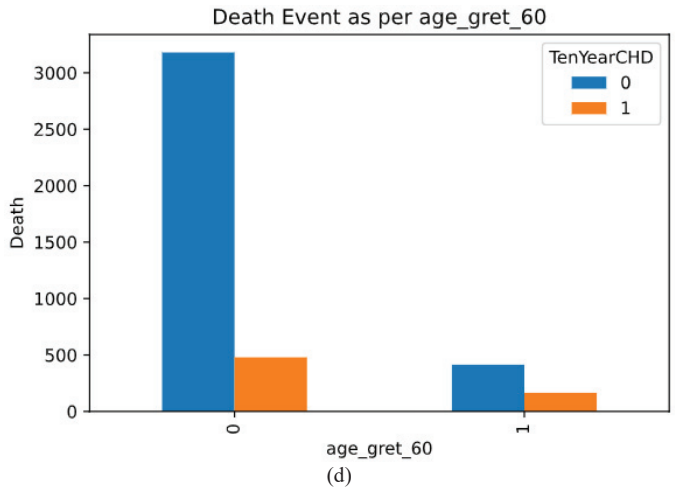
(a)



(b)



(c)



(d)

Fig. 1. Event of Death when age is – (a) less than 20, (b) between 20 to 40, (c) between 40 to 60, (d) greater than 60.

B. *Systolic and Diastolic Blood Pressure*

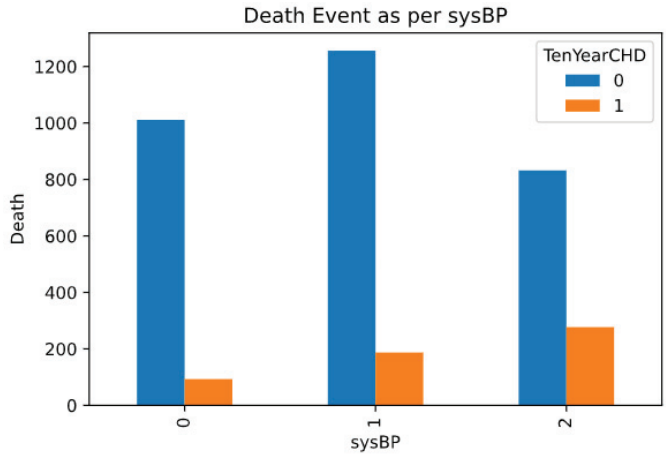


Fig. 2. Correlation of Event of Death with Systolic Blood Pressure

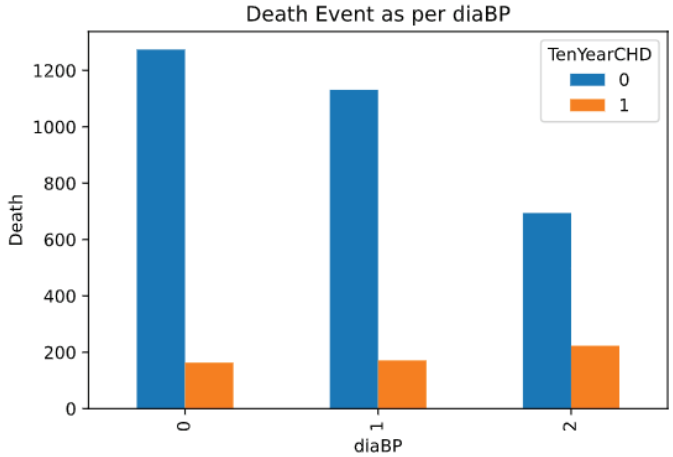


Fig. 3. Correlation of Event of Death with Diastolic Blood Pressure.

In Figure 2 and 3, low, normal and high values of systolic and diastolic blood pressure are represented by 0, 1 and 2 respectively. The normal range for systolic blood pressure is 120-140 mmHg while for diastolic blood pressure, it is 80-90 mmHg. The percentage of people with high systolic blood pressure who died is 24.97 while the percentage of people who died and had high diastolic blood pressure is 24.32. This means that people with high blood pressure are at a higher risk of suffering from a heart attack.

C. *Sex*

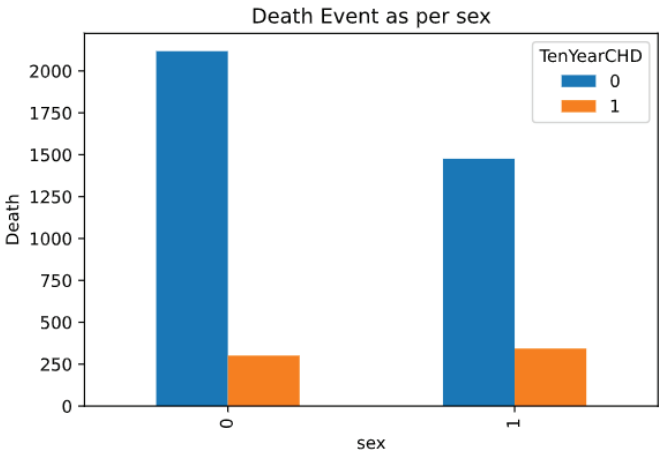


Fig. 4. Correlation of Event of Death with Sex.

In Figure 4, 0 represents the female and 1 represents the male. The percentage of people who died and are male is 18.86 while who died and are female is 12.44. This shows that males are at higher risk of suffering from a heart attack than females.

D. *Education*

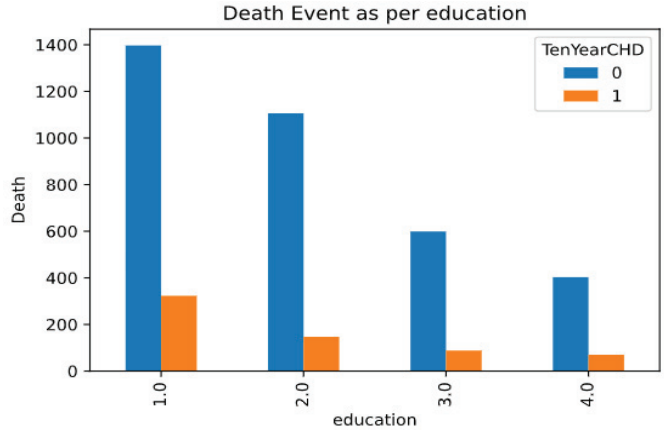


Fig. 5. Correlation of Event of Death with Education.

Education plays a very important role in maintaining a healthy lifestyle. Knowledge and awareness to make the right choices are thus key factors in predicting future possibility of getting a disease. This is also confirmed by the visualization in Figure 5. People with a very basic level of education have a more mortality rate than people with some level of education. The percentage of people who died and had some high school education is 18.78 while the percentage of graduates who died is 14.79.

E. Prevalent Hypertension

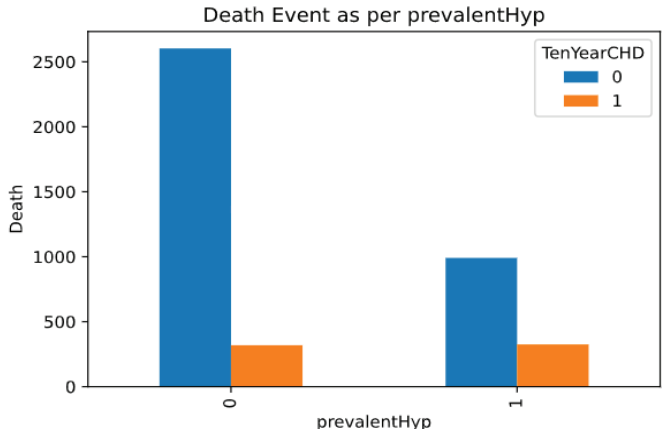


Fig. 6. Correlation of Event of Death with Prevalent Hypertension.

The percentage of people who died and didn't have the condition of prevalent Hypertension (represented by 0 in Figure 6) was 10.92 which is lesser than the percentage of people who died and had the condition of prevalent Hypertension (represented by 1 in Figure 6) which is 24.69. This shows that stress increases the chances of suffering from heart failure.

F. High Cholesterol and Gender

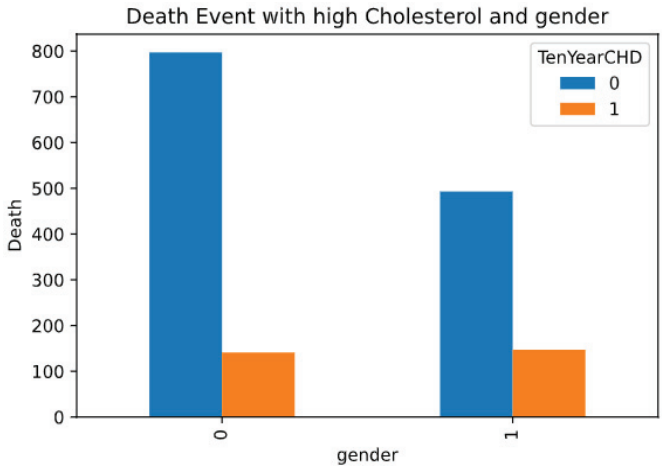


Fig. 7. Correlation of Event of Death with High Cholesterol and Gender.

In Figure 7, 0 represents the female and 1 represents the male. The percentage of people with high cholesterol who died and are male is 22.97% and females is 15.03%. Comparing Figure 4 and Figure 7, we observed that the mortality rate of males with high cholesterol increases by 4.11% and that of females increases by 2.59% with respect to those who don't have high cholesterol.

G. Systolic blood pressure and Cigarettes per day

In figure 2, the mortality rate of people dying due to high blood pressure is 24.97% which is 10.61% less than the mortality rate of people who smoke more than 20 cigarettes per day and had high Blood Pressure. This shows that smoking more than a packet per day on average can be harmful if one has high systolic blood pressure.

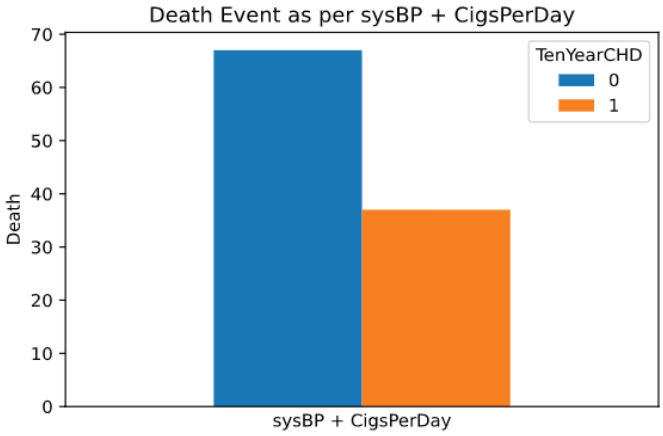


Fig. 8. Correlation of Event of Death with Systolic blood pressure and Cigarettes per day.

IV. Methodology

A. Data Preprocessing

The data has a lot of missing values which can reduce the statistical power of the algorithms used and can cause biases in the estimates, leading to invalid conclusions. We found 645 missing values in the dataset. So, to deal with this problem, we tried many strategies like dropping the Null values, imputing them with median, imputing them with median and imputing the with a frequently occurring constant. We finally used the mean of the column to replace the null values for each respective column because it gave us the best values for the chosen evaluation metrics.

We also scaled the data using sklearn's StandardScaler. Standard Scaling removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. We used this to normalize our data.

$$X_{new} = X_i - X_{mean} / \sigma \quad (1)$$

(1) represents the formula for standard scaling where

X_{mean} is the mean of training samples,

σ is standard deviation of the training samples,

X_i is the value that needs to be scaled,

X_{new} is the new value in place of Xi.

B. Training and Testing models

In order to find the most efficient algorithm, we trained several models on the dataset using existing machine learning

classification algorithms and then we devised an ensemble model using these algorithms.

These classification methods were evaluated using the following performance metrics-

1 Accuracy- It is the fraction of predictions our model got right i.e., the number of correct predictions out of all predictions.

2 Recall- It is the proportion of actual positives that were classified correctly from all predictions.

3 Precision- It is the proportion of actual positives out of all predicted positives.

In the medical domain, a False positive (people with low or no risk of heart failure) is not as harmful as a False Negative (people with a high risk of heart failure but remain undetected). Hence, recall is the most important metric among these for us.

We tested various algorithms ranging from basic classifiers like Logistic Regression to complex boosting and bagging algorithms like XGBoost. Table 2 lists all the classifiers we used with their accuracy and recall on test data.

An ensemble is a machine learning model that combines the predictions from two or more models. An ensemble can make better predictions and achieve better performance than any single contributing model. It can reduce the spread or dispersion of the predictions and model performance, especially when data is imbalanced which is our case as well. It can help in dealing with both bias and variance - variance representing scattered results that are difficult to converge, and bias representing miscalibration or error in targeting necessary results.

TABLE II. TEST ACCURACY AND RECALL VALUES FOR VARIOUS CLASSIFIERS

Sr. no.	Classifier	Accuracy (%)	Recall (%)
1	AdaBoost	85.80	86.62
2	CatBoost	85.80	86.41
3	Decision Trees	84.40	86.23
4	k-nearest neighbour (k=10)	85.80	86.22
5	Logistic regression	**86.40**	86.50
6	Light GBM	85.60	85.56
7	Gaussian Naïve Bayes	82.50	**87.06**
8	Random Forest	86.20	86.13
9	Support Vector Machine	85.90	86.23
10	XG Boost	86.10	86.53

So, we designed an ensemble model using all the classifiers mentioned in Table 2, taking Gaussian Naïve Bayes classifier as the meta classifier because the model devised resulted in the highest recall. A meta classifier is simply a classifier that makes the final prediction among all the predictions by using those predictions as features. So, it takes classes predicted by various classifiers and picks the final one as the result.

Table 3 shows the final performance of our ensemble model on the test data. It achieved a Recall score of 87.5% which is better than all the single classifiers.

TABLE III. VALUES OF PERFORMANCE METRICS FOR OUR DEVISED MODEL

Performance Metrics	Value (%)
Training accuracy	87.80
Testing accuracy	**85.20**
Recall	**87.50**
Precision	96.47
F1-score	91.76

V. RESULT

Regarding the related work, one very recent study for detecting the risk of chronic heart failure from heart sounds was presented by Martin Gjoreski et. al [10]. They considered data of 947 subjects available in six public datasets and one CHF dataset. The proposed method managed to achieve a score of 89.3%, 9.1% higher than the baseline method. Average accuracy of the method is 92.9% with an error of 7.1%. Baseline Random Forest achieved an accuracy of 74.2%, while baseline Logistic Regression achieved an accuracy of 64.1%. This method was able to achieve an overall accuracy of 92.9%.

Faruk Bulut et. al [11] proposed a supervised learning-based approach and achieved an accuracy of 81.7%, 75.8% and 79.7% on SVM, KNN (Neighbors = 1) and Decision Tree respectively. However, using the bagging method with a decision tree with 10, 20 and 30 datasets they achieved an accuracy of 87.5%, 94.5% and 95.5% respectively.

C. M. Chethan Malode et. al [12] focused towards adolescent's heart attack risk detection and classification. They proposed a Soft Set Fuzzy Enabled SVM Classifier which outperformed the conventional SVM Classifier by 5-7% in accuracy while it was outperformed by the latter in delay and efficiency.

Cameron R. Olsen et. al [13] used the EHR data from clinics with 15,492 heart failure patients. Their deep neural network algorithm (AUC 0.880) outperformed supervised learning methods such as linear and logistic regression, decision trees, SVMs, Random forest, Naive Bayes classifiers and KNN.

T. Obasi et. al [14] implemented Random Forest, Bayesian Classification and Logistic Regression on a dataset consisting of 18 features and 1990 observations. Extensive Data Visualization and Analytics were carried out on this dataset. Random Forest classifier achieved 92.44% accuracy while Naïve Bayes Classifier and Logistic Regression achieved 61.96%, and 59.7% respectively.

Renu Narain et. al [15] proposed a system based on Quantum Neural Network which achieved 98.57% accuracy in predicting Cardiovascular disease risk which outperformed the commonly used Framingham risk score.

Abhay Kishore [16] aimed to implement a method that determined the future possibilities of heart disease to a particular user using the past inputs of the disease outcomes. Following this, they proposed a Gated Recurrent Unit based Recurrent Neural Network which achieved an AUC score of 92%.

This paper stands out by providing a novel Ensemble-based pipeline to detect Heart Failure and shows the potential to which ensemble learning can transform results. It also shows that ensemble models can be used when the data is highly imbalanced as it considers features from various models to make predictions.

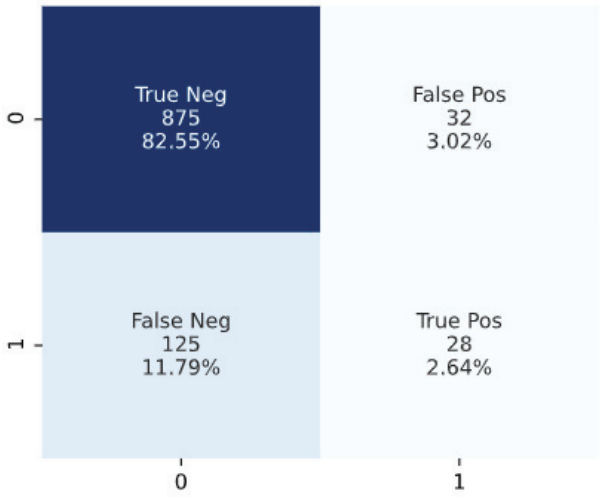


Fig. 9. Confusion Matrix on test data

Figure 9 shows the confusion matrix which displays the number of correct and incorrect predictions on the test data.

VI. CONCLUSION

This paper proposes a novel ensemble learning-based model that was created using various machine learning classifiers to predict the risk of heart failure in the next ten years. We used the data collected from the famous Framingham heart study and performed extensive exploratory data analysis in which we observed the correlation between various attributes and their contribution in causing a heart attack. We trained numerous Machine Learning Algorithms on this data to find the most efficient and effective model in terms of the highest recall. We concluded that the most efficient model is where we combine all the ten algorithms to develop an ensemble model. This ensemble model achieved an accuracy of 85.2% and a recall of 87.5%. This system can be used to predict the heart failure risk of a patient well in advance to avoid any serious consequences later. Our project can also be used by the fitness industry in many areas- be it the fast-growing trend of smart wearables like watches and bands or in gyms where the machines continuously monitor heart rate and takes into consideration various factors.

REFERENCES

[1] "NATIONAL HEART, LUNG, AND BLOOD INSTITUTE", Accessed on: January 15, 2021. [Online]. Available: https://www.nhlbi.nih.gov/

[2] World Health Organization, "Cardiovascular diseases (CVDs)", May 17 2017. Accessed on: Januray 15,2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[3] Coronary Artery Disease in Asian Indians, "Tsunami of Heart Disease ". Accessed on: January 16, 2021. [Online]. Available: https://cadiresearch.org/topic/asian-indian-heart-disease/cadi-india/tsunami

[4] Mayo Clinic, "Heart Attack", June 16 2020. Accessed on: January 16 ,2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.

[5] S. Romiti, M. Vinciguerra, W. Saade, I. Anso Cortajarena, and E. Greco, "Artificial Intelligence (AI) and Cardiovascular Diseases: An Unexpected Alliance," Cardiol. Res. Pract., vol. 2020, p. 4972346, 2020, doi: 10.1155/2020/4972346.

[6] J. Wiens and E. S. Shenoy, "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology," Clin. Infect. Dis., vol. 66, no. 1, pp. 149–153, 2018, doi: 10.1093/cid/cix731.

[7] Kilvia L. De Almeida, Lucília Lessa, Anny Peixoto, Rafael Gomes, Joaquim Celestino. Kidney Failure Detection Using Machine Learning Techniques. 8th International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE 2020), Candy E. Sansores, Universidad del Caribe, Mexico, Nazim Agoulmine, IBISC Lab, University of Evry - Paris-Saclay University, Jan 2020, Cancún, Mexico. pp.1–8. ffhal-02495264

[8] A. Colubri et al., "Machine-learning Prognostic Models from the 2014-16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications.," EClinicalMedicine, vol. 11, pp. 54–64, 2019, doi: 10.1016/j.eclinm.2019.06.003..

[9] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," Chaos, Solitons and Fractals, vol. 139, no. June, 2020, doi: 10.1016/j.chaos.2020.110059.

[10] M. Gjoreski, A. Gradišek, B. Budna, M. Gams and G. Poglajen, "Machine Learning and End-to-End Deep Learning for the Detection of Chronic Heart Failure From Heart Sounds," in IEEE Access, vol. 8, pp. 20313-20324, 2020, doi: 10.1109/ACCESS.2020.2968900.

[11] F. Bulut, "Heart attack risk detection using Bagging classifier," 2016 24th Signal Processing and Communication Application Conference (SIU), Zonguldak, Turkey, 2016, pp. 2013-2016, doi: 10.1109/SIU.2016.7496164.

[12] C. M. Chethan Malode, K. Bhargavi, B. G. Gunasheela, G. Kavana and R. Sushmitha, "Soft set and Fuzzy Rules Enabled SVM Approach for Heart Attack Risk Classification Among Adolescents," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697650..

[13] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure: Machine learning in heart failure," Am. Heart J., vol. 229, pp. 1–17, 2020, doi: 10.1016/j.ahj.2020.07.009.

[14] T. Obasi and M. Omair Shafiq, "Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 2393-2402, doi: 10.1109/BigData47090.2019.9005488.

[15] R. Narain, S. Saxena, and A. K. Goyal, "Cardiovascular risk prediction: A comparative study of framingham and quantum neural network based approach," Patient Prefer. Adherence, vol. 10, pp. 1259–1270, 2016, doi: 10.2147/PPA.S108203.

[16] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, "Heart Attack Prediction Using Deep Learning," Int. Res. J. Eng. Technol., vol. 5, no. 4, pp. 4420–4423, 2018

[17] Aman Ajmera. (2017). Framingham Heart study dataset, version 1. Retrieved December 15,2020 from

[18] https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset