CatLOrf: A new Ensemble model using Optimized Classifiers for Cardiovascular Disease Prediction

Abhinav Dahiya[1], Kamaldeep Joshi[1], Rajkumar Yadav[1], Rainu Nandal[1], Komal Rani Tehlan[1] and Deepika Kumari[2]

[1]Department of Computer Science and Engineering, University Institute of Engineering & Technology, Maharshi Dayanand University, Rohtak – Haryana, 124001, INDIA.

[2]Department of Commerce, GCW, Lakhan Majra, Rohtak – Haryana, 124001, INDIA.

Email Id: abhinav.rs.uiet@mdurohtak.ac.in, kamalmintwal@gmail.com, rajyadav76@rediff.com, rainu_nandal@yahoo.com , tehlankomal@gmail.com, deepikakumari096@gmail.com

ABSTRACT - Heart disease is one of the leading causes of death in the globe at present. WHO reports that heart diseases are the leading cause of death worldwide. In clinical data analysis, heart disease prediction is a significant impediment. Machine learning (ML) is a helpful decision-making and forecasting tool for the healthcare industry's vast amounts of data. This study offers *CatLOrf*, a new approach based on ensemble learning that combines three distinct classification algorithms: *Cat*Boost, *L*ogistic regression, and an *O*ptimized *r*andom *f*orest. The test accuracy of CatLOrf is 93.55%, and the test recall is 95.45%. We utilized the kaggle "Heart Disease Analysis and Prediction" dataset, which includes 14 features and over 303 records. In addition, we conducted a thorough exploratory data analysis to determine the significance of each feature in triggering heart disease.

Index Terms - Cardiovascular disease, CatLOrf, Machine learning, Optimization, Ensemble.

I. INTRODUCTION

The phrase "cardiovascular disease" refers to conditions that can damage the heart or blood vessels, including coronary heart disease (clogged arteries) and other conditions that can lead to heart attack, stroke, non-inherited heart defects, and peripheral artery illness. According to WHO [1], in 2016, CVDs were responsible for an estimated 17.9 million fatalities, or 31 percent of all global deaths. Eighty-five percent of these deaths were attributable to heart attacks and strokes. In India in 2016, Non-Communicable Diseases were responsible for 63 percent of all deaths, with cardiovascular diseases accounting for 27 percent of those deaths. Cardiovascular diseases are responsible for 45 percent of deaths between the ages of 40 and 69. Because of the multiple risk factors that can contribute to heart diseases, such as diabetes, high blood pressure, excessive cholesterol, abnormal pulse rate, and many more [2], it can be difficult to diagnose heart disease in its early stages. The term "heart attack" refers to a complex clinical condition when the heart cannot pump enough blood to the rest of the body. A heart attack happens when the flow of blood rich in oxygen to a section of the heart is restricted, leading to a lack of oxygen in that portion of the heart. The plaque will burst and release cholesterol and other chemicals into circulation after a cardiac failure.

In recent years, the use of computer-assisted diagnosis approaches based on artificial intelligence-based solutions has increased [3]. This is because machine learning algorithms can quickly identify and categorize significant trends and patterns in the data [4]. Various methods, including Decision Trees (DT), K - Nearest Neighbor Algorithm (KNN), Random Forest(RF), and Naive Bayes (NB), are used to classify the severity of the condition. The complexity of cardiovascular disease necessitates a cautious approach to treatment. Data mining and classification are essential in heart disease prediction and analysis.

This research suggests CatLOrf, an ensemble solution based on machine learning [5], to forecast the possibility of cardiac failure. We started by implementing several Machine Learning algorithms, including Logistic Regression, KNN, Support Vector Machine, Random Forest, Decision Tree, Gaussian Naive Bayes, Adaptive Boosting, and CatBoost. After that, we constructed the CatLOrf model that incorporatesthree of these models, and it was this model produced the most accurate results:95.45% recall and 93.55% accuracy on test data. In addition, we conducted a comprehensive exploratory data analysis, performed suitable scaling, removed outliers, and gained insight into the association between various factorsand a heart attack.

The remaining sections are organized as follows: The second section includes information about the dataset, followed by section three, where comprehensive information about the employed dataset is done. The fourth section contains data pre-processing. Section five discusses the proposed model. The sixth section discusses the results obtained, followed by the last section, the seventh, which concludes the whole paper. Figure 1 shows the complete flow of action for the remaining paper.

II. DATASET

In this study, we use the "Heart Attack Analysis & Prediction " dataset version 2 provided by Kaggle [6]. To classify cardiac disease in this research, we utilized a Kaggle notebook. Table 1 displays detailed information on the dataset's features. The dataset has no missing values among the 303 (165 for healthy patients and 138 for heart disease) patient records. The process begins with comprehensive exploratory data analysis, followed by data pre-processing, feature selection, identifying and removing outliers

*e*ISSN 1303-5150

classification, performance evaluation, and optimization to improve classification models' accuracy.

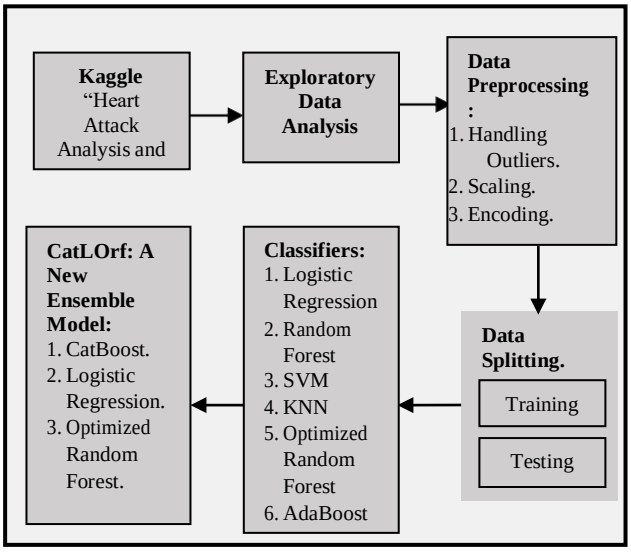

Fig. 1 Flow diagram of the course of actions.

TABLE I
Detailed Information of Features.

Features	Description	Numbers of Unique Values
Age	Age of subject in years. (Numerical)	41
Sex	Gender of the subject. (Categorical)	02
CP	Chest Pain, i.e., Asymptomatic, non-anginal pain, atypical angina, typical angina. (Categorical)	04
Trestbps	Blood pressure of the subject arriving at the hospital in mm of Hg. (Numerical)	49
Chol	The cholesterol of the subject is in mg/dl. (Numerical)	152
FBS	Fasting blood sugar of the subject. (Categorical)	02
RestECG	Resting electrocardiographic report of the subject. (Standard, ST-T wave abnormality, probable left ventricular hypertrophy). (Categorical)	03
Thalach	Maximum heart rate of the subject. (Numerical)	91
Exang	Angina induced by exercise. (Categorical)	02
Oldpeak	ST depression induced by exercise relative to rest. (Numerical)	40
Slope	The slope of the peak exercise ST segment. (Categorical)	03
CA	The number of significant vessels colored by Fluoroscopy.(Categorical)	05
Thal	Thalium stress result of the subject. (Categorical)	04
Target	Diagnosis of heart disease. (Categorical)	02

III. EXPLORATORY DATA ANALYSIS

We carried out a thorough exploratory data analysis on the characteristics outlined above. As a result, we obtained significant insights into these factors' impact on cardiovascular disease. Figure 2 shows a heatmap for the correlation between features.

A. AGE

There exist some intriguing analyses. Generally, it is believed that the incidence of heart attacks increases with age. But, according to the datasets, as the age increases, the number of individuals at risk for heart attack reduces. The correlation between age and target attribute is negative, which is -0.230.

B. TRTBPS

Predicting whether a patient will have a heart attack based on resting blood pressure data is difficult. The correlation between TRTBPS and the target attribute is -0.14, which is negative and very little. Hence this attribute is dropped.

C. CHOL

According to analysis, a cholesterol level between 200 and 250 is dangerous for individuals. After the value of 250, it becomes impossible to identify the risk of heart attack among patients. The correlation between the variables is -0.085239, which is extremely low. Hence it is dropped.

D. THALACH

The patient's likelihood of suffering a heart attack is proportional to the patient's maximum heart rate. There is a positive and moderate correlation of 0.4217 between the variables.

E. OLDPEAK

If the value of this variable is between 0 and 1.5, the likelihood of suffering a heart attack increases significantly. The correlation coefficient between the two variables is -0.430696, negative and moderate.

F. SEX

Patients with a high risk of heart attack among females constitute nearly more than half of those with low risk. In conclusion, the risk of heart attack is higher for female patients. The correlation between the two variables is -0.280937.

G. CP

Patients with asymptomatic pain are less likely to have a heart attack. For the other three categories of pain, the chance of having a heart attack is about three times that of not having one. The correlation between the two variables is 0.433798, which is moderately strong.

H. FBS

Those with a blood glucose observation value of 0 had a marginally increased risk of a heart attack. Patients with a fasting glucose level of less than 120 mg/dL are at risk. There is a -0.028046 correlation between the two variables. Hence it is dropped.

I. REST_ECG

Patients with a "REST_ECG" value of 1 should be closely monitored because the danger of having a heart attack is about twice as great as the risk of avoiding one. There is a correlation of 0.137230 between the two variables.

J. EXANG

If the patient's "exang" variable is 1, indicating that he has exercise-related angina, the likelihood that he will not have a heart attack is more significant. Pain caused by exercise has nothing to do with a heart attack. There is an inverse correlation of -0.436757 between the two variables.

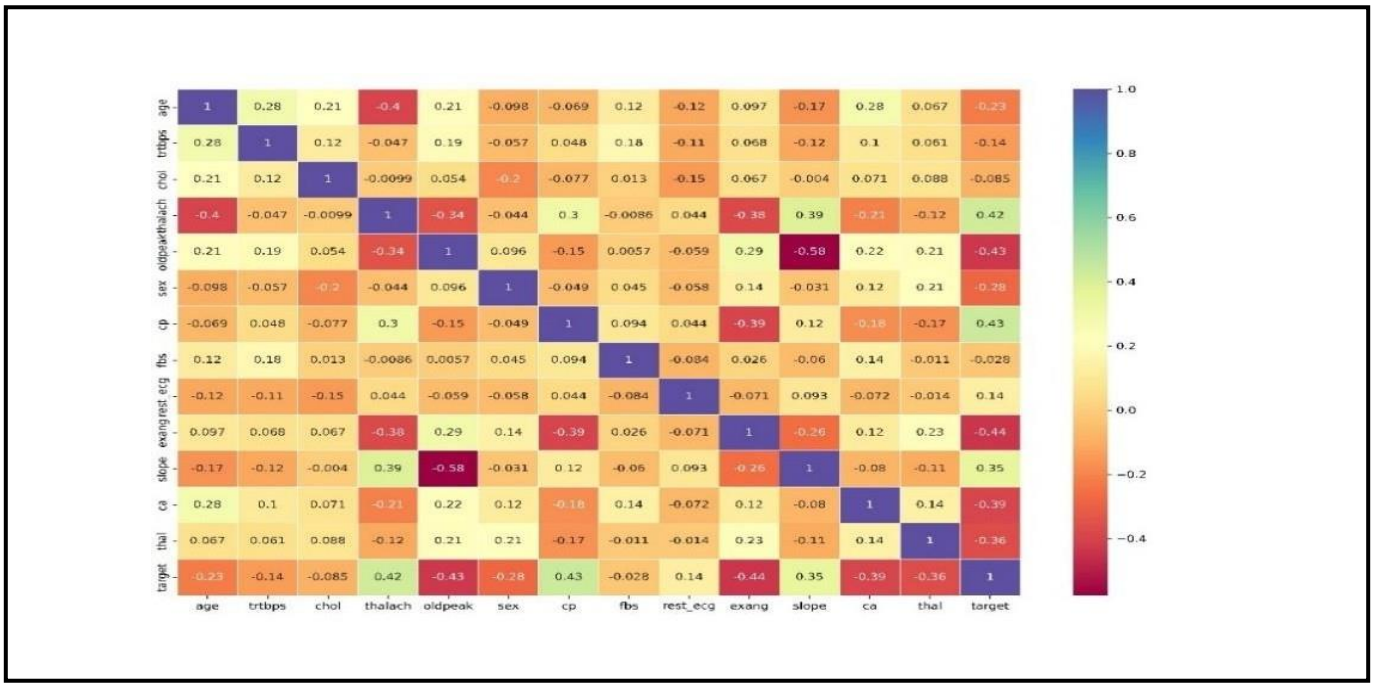

Fig.2 Heatmap for feature's correlation.

K. SLOPE

The patients with a "slope" variable of 2 require special consideration because patients with an observation value of 2 are three times more likely to suffer a heart attack than those without such a value. The correlation between the two variables is 0.345873.

L. CA

Patients with an observation value of 0 are nearly three times more likely to suffer a heart attack than those with a value of 1. Patients with "ca" observation values of 1, 2, and 3 are twice as likely to suffer a heart attack. The correlation between the two variables is -0.363322.

M. THAL

Patients with an observation value of 2 are three times more likely to suffer a heart attack than those without an observation value of 2. The correlation between the two variables is -0.363322.

IV. DATA PREPROCESSING

After comprehensive exploratory data analysis, the pre-processing of the data starts. Firstly, outliers will be handled, followed by scaling and encoding.

A. OUTLIERS

Figure 3 depicts a boxplot for numerical variables. After analyzing these boxplots, the following conclusions could be made regarding outliers [7]. The boxplot for the "age" variable is positioned in the middle of the whiskers. Hence "age" variable has no outliers. The "trtbps" variable contains more outliers than the others. For the "thalach" variable, very few outliers exist. There is one outlier in the mustache region. Within the "oldpeak" variable, there is no such thing as a lower whisker. The bottom of the box is at the lower mustache level. There are some outliers at the top of the mustache level. Outliers for "trtbps" and "oldpeak" variables are corrected using winsorization. Winsorization [8] is the practice of substituting the extreme values of statistical data to reduce the impact of outliers on calculations or outcomes derived from the data. There was only one outlier for the "thalach" variable, which dropped from the dataset.

Figure 4 and Figure 5 show the boxplot for "oldpeak" and "trtbps" variables after winsorization having 0 outliers.

B. SCALING AND ENCODING

After removing the outliers, One Hot encoding [9] is applied to categorical features, and the RobustScalar method is used to scale all numerical features. Robust Scalar [10] scales features using robust statistics against outliers. This method does away with the median and instead scales the data within the range of the first to third quartiles or the 25th to 75th quantiles, whichever is lower.

C. TRAINING AND TESTING

The whole dataset is divided into training and test subset in the ratio of 9:1. To determine the best effective algorithm, we trained many machine learning models on the dataset and then applied hyperparameter optimization to the Random Forest method.

The following metrics were used to assess the effectiveness of these various classification approaches:

1) *Accuracy [11]:* Accuracy is the proportion of correct predictions made by the model or the number of accurate forecasts out of total predictions.
2) *Recall [11]*: Recall refers to the percentage of true positives accurately identified across all predictions.

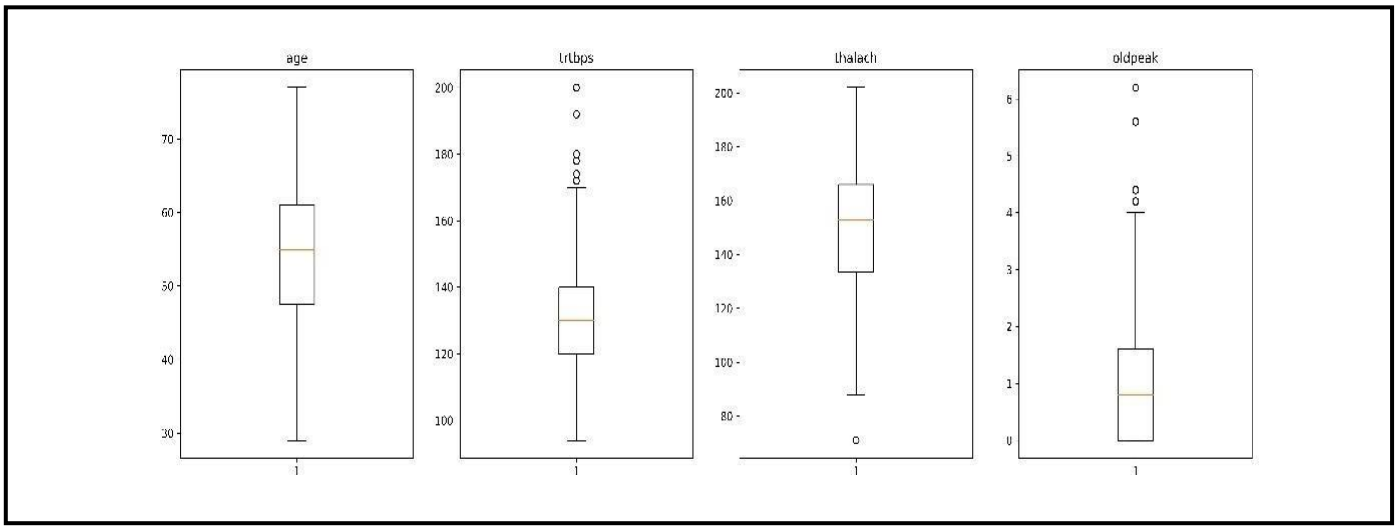

Fig. 3 Boxplot of numerical features.

3) *Precision [11]:* It represents the fraction of actual positives relative to all anticipated positives.
4) *F1 score [11]:* The F1 score is the harmonic mean of precision and recall.

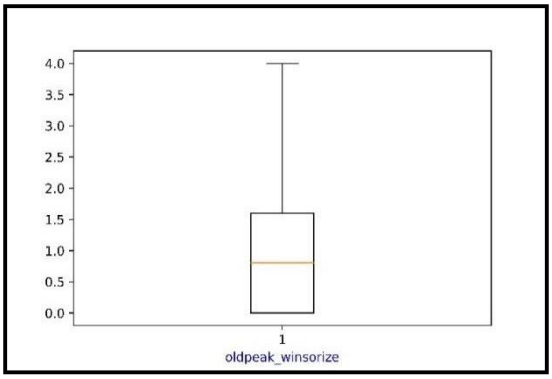

Fig. 4 Boxplot for "oldpeak" feature.

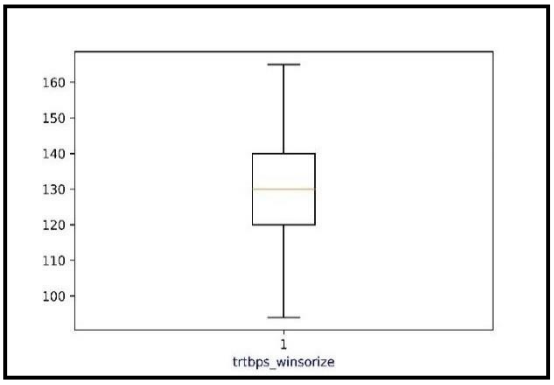

Fig. 5 Boxplot for "trtbps" feature.

V. Proposed Model

We investigated various methods, ranging from straightforward classifiers like Logistic Regression to more complex boosting and bagging algorithms like AdaBoost and CatBoost [12]. The classifiers we used are outlined in Table 2, along with their accuracy, recall, precision levels, and F1 scores on the test data. Figure 6 shows the ROC curve graph for all applied classifiers and their AUC values. The optimized random forest classifier and CatBoost classifier outperform other classifiers.

TABLE 2
Performance metrics for various applied machine learning algorithms.

Classifier	Accuracy	Precision	Recall	F1 Score	AUC score
Logistic Regression	87.10 %	87.50 %	95.45 %	91.30 %	0.88
Decision Tree	83.87 %	94.74 %	81.82 %	87.80 %	0.85
SVM	87.10 %	90.91 %	90.91 %	90.91 %	0.89
Gaussian Naïve Bayes	83.87 %	86.96 %	90.91 %	88.89 %	0.82
KNN	90.32 %	91.30 %	95.45 %	93.33 %	0.89
Random Forest	83.87 %	90.48 %	86.36 %	88.37 %	0.92
AdaBoost	90.32 %	91.30 %	95.45 %	93.33 %	0.87
CatBoost	90.32 %	95.24 %	90.91 %	93.02 %	0.92
Optimized Random Forest Model	90.32 %	95.24 %	90.91 %	93.02 %	0.93

The Random Forest model gives an accuracy of 83.87 % and an AUC score of 0.92. After applying 10-fold cross-validation [13], the random forest classifier was found to

underfit. Hence, hyperparameter optimization is applied. In hyper-parameter optimization [14], optimum values of parameters are found for given classifiers. The SKlearn's GridSearchCV module is used to calculate the optimum values of the parameters. Table 3 shows the optimized values for Sklearn's RandomForestClassifier. Using optimized values for RandomForestClassifier, a Hybrid Optimized Random Forest classifier is achieved with enhanced accuracy, precision, recall, F1 score, and AUC score.

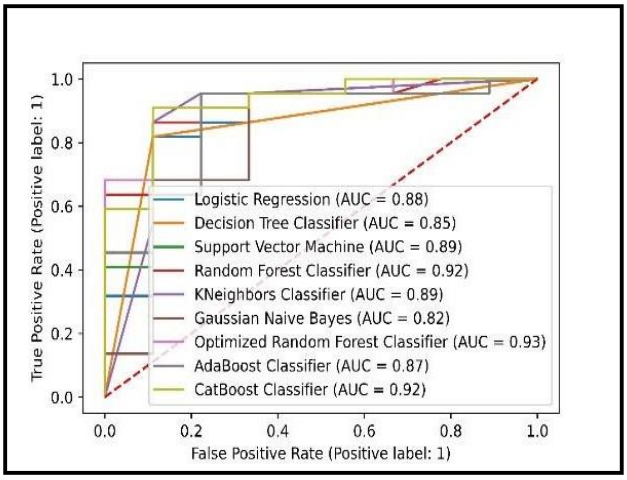

Fig. 6 ROC curve graph

The results of multiple machine learning models can be combined into a single prediction using an ensemble model. It can help resolve bias and variance, showing difficult-to-converge scattered findings and bias representing miscalibration or inefficiency in targeting necessary results. Both of these issues can be difficult to fix. An ensemble can make more accurate forecasts and perform better than any individual model [15]. An ensemble model CatLOrf is created using the CatBoost model, the Logistic regression model, and the Optimized random forest model. The Proposed CatLOrf is a hard voting-based ensemble that takes the result from the majority of classifiers.

TABLE 3
Optimized values for Random Forest Classifier.

Parameter	Value
bootstrap	True
criterion	entropy
max_features	auto
n_estimators	200
random_state	5

Table 4 illustrates our ensemble model's overall effectiveness based on the tests' results.

TABLE 4
Performance metrics for the proposed ensemble model.

Parameters	CatLOrf model
Accuracy	93.55%
Precision	95.45%
Recall	95.45%
F1 score	95.45%

VI. RESULTS

This article stands out from the crowd because it presents CatLOrf, a new ensemble-based pipeline for diagnosing heart attacks, and demonstrates the transformative potential of ensemble learning. It also suggests that ensemble models can be used even when the data is exceptionally imbalanced because they predict based on the properties of numerous models combined. The confusion matrix is depicted in Figure 7, showing the number of accurate and erroneous test data predictions. The ROC curve for classifiers used in CatLOrf is shown in Figure 8 with their respective AUC score.

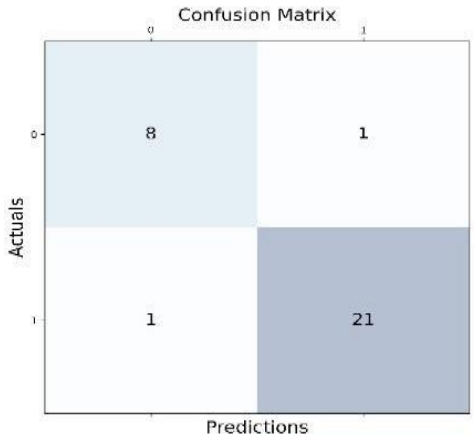

Fig. 7 Confusion matrix of proposed ensemble model.

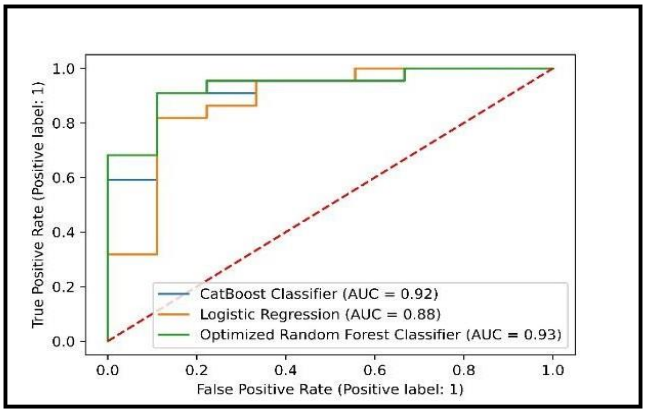

Fig. 8 ROC curve for classifiers used in CatLOrf.

eISSN 1303-5150

VII. Conclusion

This study offers CatLOrf, a novel ensemble approach that uses multiple machine learning classifiers to predict the risk of a heart attack. We used Kaggle based heart attack dataset. A comprehensive data analysis is performed on features. We investigated the connection between several variables and the role those variables play in the likelihood of a heart attack. We applied One Hot encoding on categorical features and robust scaling on numerical data. We identify features having outliers and successfully remove them. We trained several different ML algorithms on these data to determine the most effective and efficient model. We concluded that the most effective model is one in which three algorithms are combined to form an ensemble model. This CatLOrf model was accurate 93.55 percent of the time and had a recall rate of 95.45 percent. This technique can be used to anticipate a patient's risk of heart failure well in advance to prevent later complications.

References

[1] World Health Organization. (2019). World health statistics overview 2019: Monitoring health for the SDGs, sustainable development goals (No. WHO/DAD/2019.1). World Health Organization
[2] MayoClinic, "heart attack," June162020.Accessed on: October 01, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-attack/symptoms-causes/syc-20373106.
[3] Romiti, S., Vinciguerra, M., Saade, W., Anso Cortajarena, I., & Greco, E. (2020). Artificial intelligence (AI) and cardiovascular diseases: an unexpected alliance. Cardiology Research and Practice, 2020.
[4] Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153.
[5] Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning* (pp. 1-34). Springer, Boston, MA.
[6] Rashik Rehman. (2017). Heart Attack Analysis & Predictiondataset, version 2. Retrieved October 02, 2022, from https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.
[7] Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S., &Kavsek, B. (2000, August). Informal identification of outliers in medical data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology* (Vol. 1, pp. 20-24).
[8] Ghosh, D., & Vogt, A. (2012, July). Outliers: An evaluation of methodologies. In Joint statistical meetings (Vol. 2012).
[9] Rodríguez, P., Bautista, M. A., Gonzalez, J., & Escalera, S. (2018). Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, *75*, 21-31.
[10] Cao, X. H., Stojkovic, I., & Obradovic, Z. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. BMC Bioinformatics, 17(1), 1-10.
[11] Yacouby, R., & Axman, D. (2020, November). Probabilistic extension of precision, recall, and F1 score for a more thorough evaluation of classification models. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 79-91).
[12] Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *41*(3), 552-568.
[13] Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137-146.
[14] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295-316.
[15] Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. Journal of Climate, 9(7), 1518-1530.