IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Exploring Machine Learning For Efficient Human Heart Disease Detection

S.Saleem[1] Dr.P.Nageswara Rao[2]

[1]PG student, Vemu Institute of Technology, P. Kothakota.

[2]Professor,Vemu Institute of Technology, P.kothakota.

ABSTRACT

This project pioneers a novel approach to predicting cardiac disease using Machine Learning algorithms like LR, KNN, SVM, GBC, and the powerful Extreme Gradient Boosting Classifier (XGBoost) with GridSearchCV. Utilizing 5-fold cross-validation, it assesses performance across diverse datasets. The XGBoost Classifier with GridSearchCV achieves outstanding accuracy, hitting 100% in testing and 99.03% in training across multiple datasets, outperforming other algorithms and previous studies. Notably, the XGBoost Classifier without GridSearchCV also demonstrates strong accuracy. Furthermore, an extension employing Random Forest shows comparable accuracy with reduced computation time. This research underscores the efficacy of the proposed technique in advancing cardiac disease prediction.

Keywords: ML, cardiac disease

INTRODUCTION

Growing availability of medical data poses a challenge and opportunity for deriving innovative insights and solutions in healthcare. Diagnosing diseases, particularly cardiovascular disease (CVD), remains a significant challenge despite extensive research efforts. With CVD accounting for a substantial portion of global mortality, early diagnosis is crucial for effective treatment and reducing mortality rates. Machine learning offers a promising approach by analyzing vast amounts of data to identify patterns and develop predictive models. This project aims to leverage machine learning algorithms to predict cardiac disease, focusing on risk factors such as age, blood pressure, cholesterol level, and lifestyle choices. By optimizing and refining existing datasets, the study seeks to approach zero prediction error rates, advancing early detection and improving patient outcomes in cardiovascular health.

LITERATURE SURVEY

K. S. Reddy*et al*

Here the author emphasizes the increasing global burden of cardiovascular diseases (CVDs), particularly in developing countries undergoing rapid health transitions. Factors contributing to this rise include demographic changes, urbanization, globalization, and potential gene-environment interactions. Critical lifestyle factors like altered diets, reduced physical activity, and tobacco use further accelerate CVD epidemics. The author proposes a comprehensive public health response that integrates policies targeting multiple determinants of CVDs. This includes primordial, primary, and secondary prevention strategies to protect populations and individuals at risk. The focus is on nutrition-based preventive initiatives and collaborative efforts among communities, policymakers, and health professionals to ensure balanced development and promote cardiovascular health.

A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir*et al*

Since the author highlights the challenge of detecting cardiovascular diseases, particularly silent heart attacks, due to limited specialist doctors and frequent misdiagnoses. Despite efforts using medical data mining and machine learning, current prediction accuracies remain unsatisfactory. The paper proposes a heart attack prediction system leveraging Deep Learning techniques, specifically the Recurrent Neural Network (RNN). RNN, a potent classification algorithm within Deep Learning, aims to enhance prediction accuracy. The proposed model integrates deep learning and data mining to deliver more accurate results with minimal errors. This research serves as a foundation for developing an advanced heart attack prediction platform, advancing early detection and treatment.

C. D. Mathers and D. Loncar*et al*

As the author acknowledges the outdated global health projections from Murray and Lopez's 1996 study, which underestimated the impact of HIV/AIDS. Recognizing the need for current and accurate projections to inform international health policy and priority setting, the author proposes new mortality and burden of disease projections up to 2030. These projections are based on 2002 World Health Organization estimates and aim to provide a comprehensive understanding of future global health trends. The paper outlines the methodologies, assumptions, input data, and results, offering updated insights to guide informed decision-making and policy development in global health.

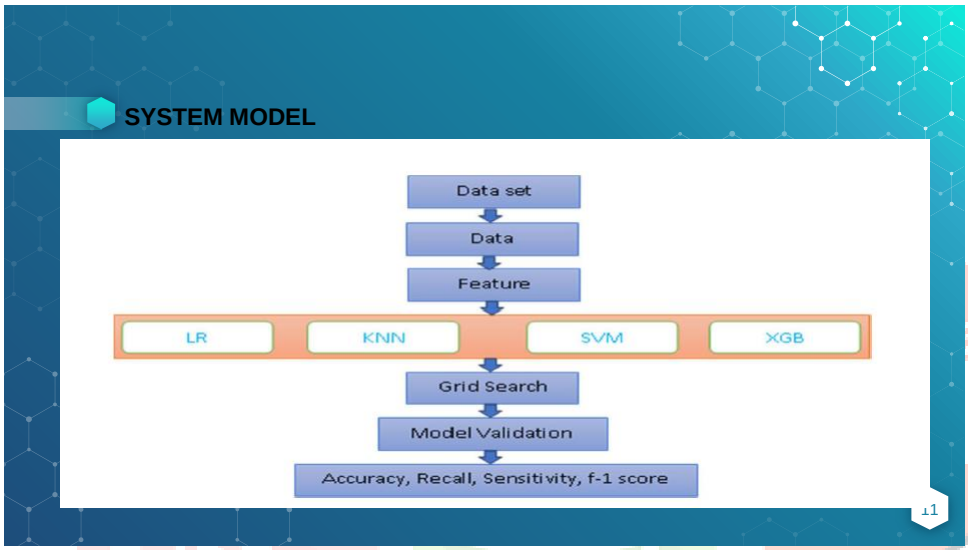
PROBLEM STATEMENT:

The challenge of diagnosing cardiac diseases amidst vast medical data prompts the quest for efficient analytical methods. Cardiovascular diseases (CVDs), including myocardial infarction, pose a significant global health threat, contributing to 31% of worldwide mortality, with forecasts predicting a substantial rise. Silent cardiac assaults, often undetectable by doctors, emphasize the critical need for early prediction. This scenario underscores the urgency to develop accurate and proactive diagnostic tools to control the rising impact of CVDs on global health.

PROPOSED METHOD:

Heart disease is one of the complicated and challenging disease which require lots of time and human experience for efficient diagnosis and timely and accurate detection of this disease can help in saving patient life. In propose paper author is evaluating performance of 4 machine learning algorithms with and without Tuning (Grid Search Parameters). Evaluated algorithms are SVM, KNN, Logistic Regression and XGBOOST. In all algorithms XGBOOST is giving 100% accuracy on both with and without tuning. This will create a unique model-creation technique for solving real-world problems.

ARCHITECTURE



HUNGARY CLEVELAND HEART DATASET DATASET

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

Defining class to normalize dataset and then we are reading and displaying dataset values

METHODOLOGY:

Data Preprocessing

Data Import First, we import necessary packages and libraries, including scikit-learn, XGBoost, and pandas. Then, we load the dataset containing information about patients, such as age, sex, cholesterol level, etc.

Data Cleaning and Normalization We handle missing values by replacing them with zero. Next, we normalize the dataset using StandardScaler to ensure all features are on a standard scale.

Exploratory Data Analysis (EDA)

Visualizing Data Distribution We visualize the distribution of heart disease cases versus non-cases in the dataset using a bar graph.

Model Training and Evaluation

Train-Test Split The dataset is split into training and testing sets with an 80-20 ratio.

Model Training and Evaluation We proceed to train and evaluate various machine learning models.

Logistic Regression Logistic Regression models are trained with and without hyperparameter tuning using GridSearchCV. Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Confusion matrices are plotted for visual assessment of classification results.

K-Nearest Neighbors (KNN) KNN models are trained with and without hyperparameter tuning using GridSearchCV. Model performance is evaluated and visualized.

Support Vector Machines (SVM) SVM models are trained with and without hyperparameter tuning using GridSearchCV. Model performance is evaluated and visualized.

XGBoost XGBoost models are trained with and without hyperparameter tuning using GridSearchCV. Model performance is evaluated and visualized.

Random Forest Random Forest models are trained with and without hyperparameter tuning using GridSearchCV. Model performance is evaluated and visualized.

Model Comparison

Comparison We compare the performance of all models based on accuracy, precision, recall, and F1-score. A bar graph is generated to visually compare the performance of different algorithms.

Computation Time Analysis

Computation Time The computation time for XGBoost and Random Forest models with and without tuning is measured. The computation time for each algorithm is visualized.

Prediction on Test Data

Prediction Finally, we load test data, preprocess and normalize it, and predict the presence of heart disease using the trained Random Forest model. The predicted results for each test sample are printed.

EVOLUTION:

Precision:

$$\text{Formula: Precision} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$$

Recall (Sensitivity):

$$\text{Formula: Recall} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$$

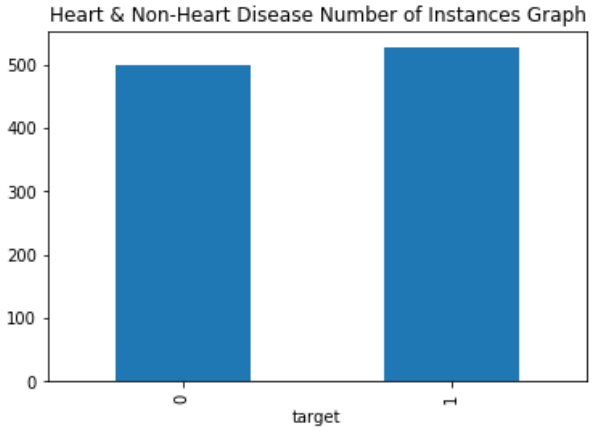
F1 Score:

$$\text{Formula: } F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

Accuracy:

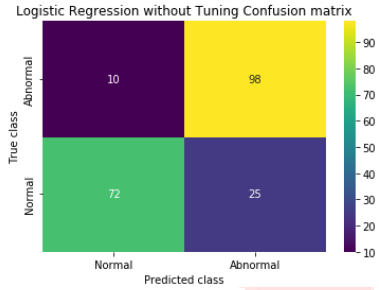
$$\text{Formula: Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

RESULTS:



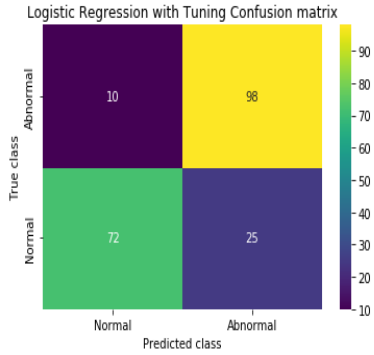
In above screen we are finding number of health and non-healthy patient' records found in dataset. In above graph x-axis represents 0 as Healthy and 1 as Non healthy and y-axis represents count

```
Logistic Regression without Tuning Accuracy    : 82.92682926829268
Logistic Regression without Tuning Precision   : 83.73983739837398
Logistic Regression without Tuning Recall      : 82.48377243222605
Logistic Regression without Tuning FScore      : 82.64770611139328
```



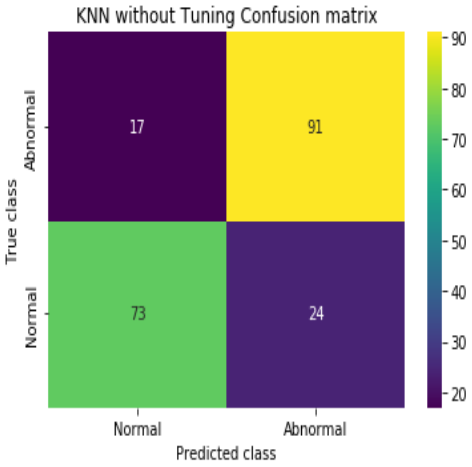
In above screen we are training logistic regression without tuning and we got its accuracy as 82% and in confusion matrix graph x-axis represents Predicted values and y-axis represents TRUE values where blue colour boxes contains incorrect prediction count and green and yellow represents correct prediction count

```
Logistic Regression with Tuning Accuracy    : 82.92682926829268
Logistic Regression with Tuning Precision   : 83.73983739837398
Logistic Regression with Tuning Recall      : 82.48377243222605
Logistic Regression with Tuning FScore      : 82.64770611139328
```



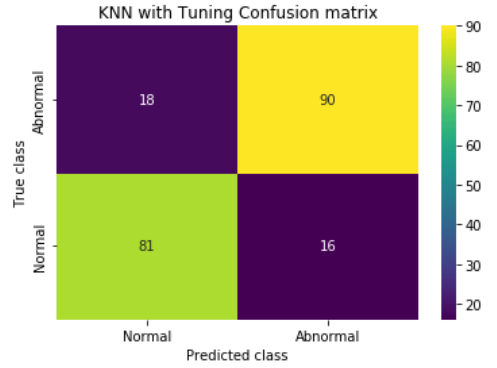
In above screen we are training Logistic regression with Tuning parameters but we got same 82% accuracy

```
KNN without Tuning Accuracy    : 80.0
KNN without Tuning Precision   : 80.12077294685992
KNN without Tuning Recall      : 79.75849560901108
KNN without Tuning FScore      : 79.8446080429726
```



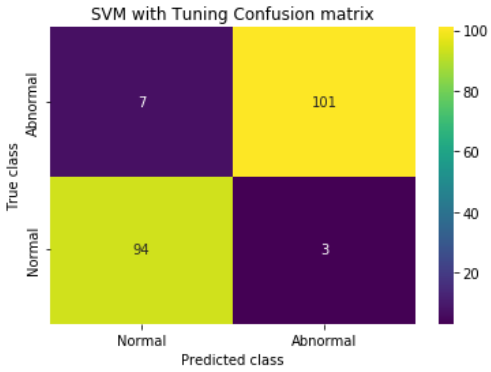
In above screen we are training KNN without tuning and we got accuracy as 80%

```
KNN with Tuning Accuracy    : 83.41463414634146
KNN with Tuning Precision   : 83.36192109777016
KNN with Tuning Recall      : 83.4192439862543
KNN with Tuning FScore      : 83.38260537860003
```



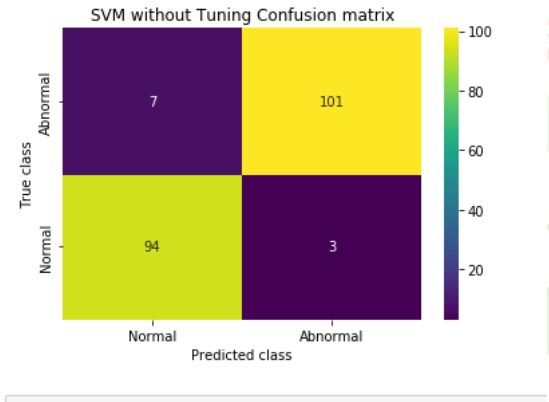
In above screen we are training KNN with tuning parameters and we got increased accuracy as 83%

```
SVM with Tuning Accuracy    : 95.1219512195122
SVM with Tuning Precision   : 95.09234577303884
SVM with Tuning Recall      : 95.21286750668195
SVM with Tuning FScore      : 95.11625690870974
```



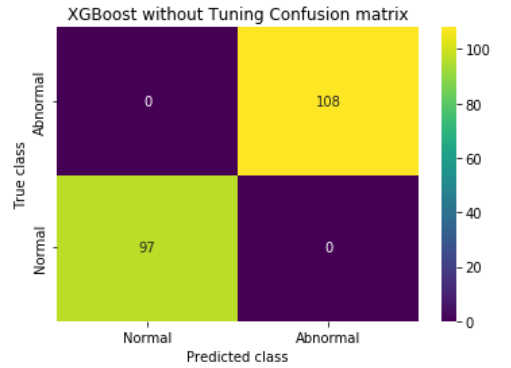
In above screen we are training SVM with tuning and we got same accuracy as 95%

```
SVM without Tuning Accuracy    : 95.1219512195122
SVM without Tuning Precision   : 95.09234577303884
SVM without Tuning Recall      : 95.21286750668195
SVM without Tuning FScore      : 95.11625690870974
```



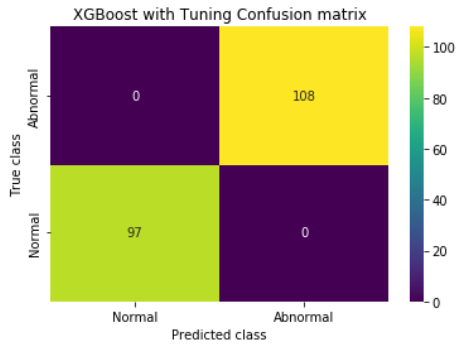
In above screen training SVM without tuning and we got accuracy as 95%

```
XGBoost without Tuning Accuracy    : 100.0
XGBoost without Tuning Precision   : 100.0
XGBoost without Tuning Recall      : 100.0
XGBoost without Tuning FScore      : 100.0
```



In above screen XGBOOST without tuning got 100% accuracy

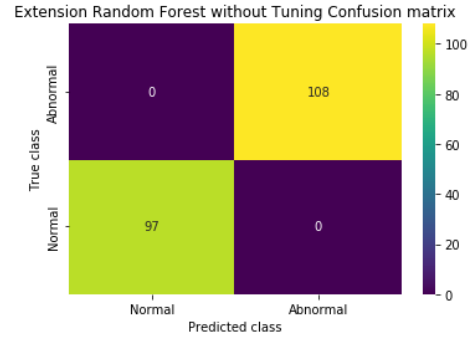
```
XGBoost with Tuning Accuracy    : 100.0
XGBoost with Tuning Precision   : 100.0
XGBoost with Tuning Recall      : 100.0
XGBoost with Tuning FScore      : 100.0
```




```
Total Computation Time Taken by XGBoost : 18.145449506999967
```

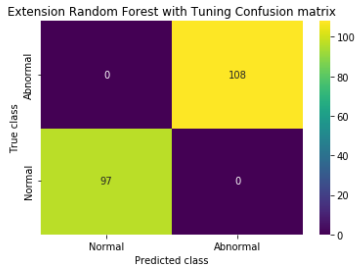
In above screen XGBOOST with tuning also got 100% accuracy and in blue colour text we can see its computation time is 18 seconds

```
Extension Random Forest without Tuning Accuracy    : 100.0
Extension Random Forest without Tuning Precision   : 100.0
Extension Random Forest without Tuning Recall      : 100.0
Extension Random Forest without Tuning FScore      : 100.0
```



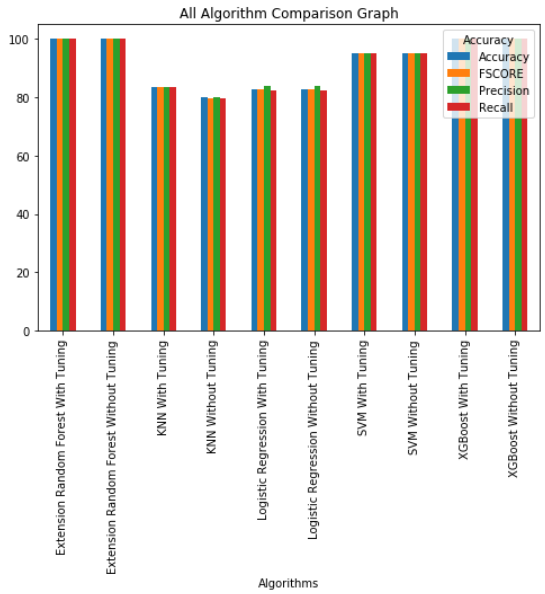
In above screen we are training extension Random Forest without tuning and got accuracy as 100%

```
Extension Random Forest with Tuning Accuracy    : 100.0
Extension Random Forest with Tuning Precision   : 100.0
Extension Random Forest with Tuning Recall      : 100.0
Extension Random Forest with Tuning FScore      : 100.0
```




```
Total Computation Time Taken by Extension Random Forest : 16.662925592999727
```

In above screen training Random Forest with tuning parameters and we got same accuracy as 100% similar to XGBOOST but Random Forest computation time is 16 seconds and XGBOOST took 18 seconds

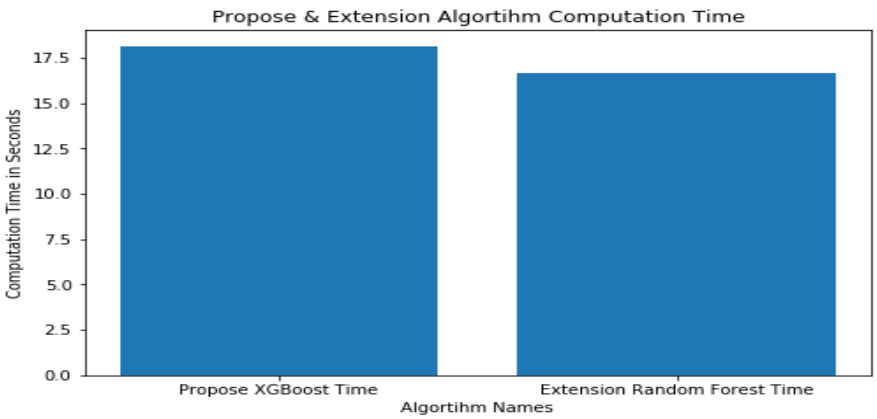


In above graph we are showing performance of all algorithms where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars. In above graph we can see XGBOOST and extension Random Forest got high accuracy

:

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Logistic Regression Wihtout Tuning	83.739837	82.483772	82.647706	82.926829
1	Logistic Regression With Tuning	83.739837	82.483772	82.647706	82.926829
2	KNN Without Tuning	80.120773	79.758496	79.844608	80.000000
3	KNN With Tuning	83.361921	83.419244	83.382605	83.414634
4	SVM Without Tuning	95.092346	95.212868	95.116257	95.121951
5	SVM with Tuning	95.092346	95.212868	95.116257	95.121951
6	XGBoost Without Tuning	100.000000	100.000000	100.000000	100.000000
7	XGBoost With Tuning	100.000000	100.000000	100.000000	100.000000
8	Extension Random Forest Without Tuning	100.000000	100.000000	100.000000	100.000000
9	Extension Random Forest With Tuning	100.000000	100.000000	100.000000	100.000000

Displaying all algorithms performance



In above graph we can see computation time comparison between XGBOOST and Random Forest and in above graph x-axis represents algorithm names and y-axis represents computation time in seconds

Prediction:

```
Test Data : [ 46.    1.    2.  150.  231.    0.    1.  147.    0.    3.6   1.    0.
   2. ] Predicted Result ===> NO Heart Disease Detected
Test Data : [ 51.    1.    3.  125.  213.    0.    0.  125.    1.    1.4   2.    1.
   2. ] Predicted Result ===> Heart Disease Detected
Test Data : [ 55.    1.    0.  140.  217.    0.    1.  111.    1.    5.6   0.    0.
   3. ] Predicted Result ===> NO Heart Disease Detected
Test Data : [ 56.    1.    3.  120.  193.    0.    0.  162.    0.    1.9   1.    0.
   3. ] Predicted Result ===> Heart Disease Detected
Test Data : [4.80e+01 1.00e+00 1.00e+00 1.30e+02 2.45e+02 0.00e+00 0.00e+00 1.80e+02
 0.00e+00 2.00e-01 1.00e+00 0.00e+00 2.00e+00] Predicted Result ===> Heart Disease Detected
Test Data : [ 55.    1.    0.  140.  217.    0.    1.  111.    1.    5.6   0.    0.
   3. ] Predicted Result ===> NO Heart Disease Detected
```

In above screen we are reading TEST data file and then random forest analysing that test data and giving prediction output. In above output in square bracket we can see test data values and after arrow =➔ symbol we can see predicted output as 'Heart Disease or No Heart Disease'

CONCLUSION

This study demonstrates the efficacy of machine learning algorithms in diagnosing heart diseases. XGBoost and Random Forest achieved 100% accuracy, with XGBoost requiring slightly more computation time. Through algorithm tuning, significant accuracy improvements were observed, particularly in KNN and SVM. These findings underscore the potential of machine learning in enhancing medical diagnosis, offering timely and accurate detection that can potentially save lives. However, further research is warranted to explore scalability and generalizability across diverse patient populations and healthcare settings. Overall, the study contributes to advancing medical diagnosis through innovative machine learning techniques.

REFERENCES:

[1] P. Drotár and Z. Smékal, ''Comparative study of machine learning techniques for supervised classification of biomedical data,'' ActaElectrotechnica Inf., vol. 14, no. 3, pp. 5–10, Sep. 2014, doi: 10.15546/aeei2014-0021.

[2] A. Levin, ''The clinical epidemiology of cardiovascular diseases in chronic kidney disease: Clinical epidemiology of cardiovascular disease in chronic kidney disease prior to dialysis,'' in Seminars in Dialysis, vol. 16, no. 2. Oxford, U.K.: Blackwell Science, Mar. 2003, pp. 101–105.

[3] K. S. Reddy, ''Cardiovascular diseases in the developing countries: Dimensions, determinants, dynamics and directions for public health action,'' Public Health Nutrition, vol. 5, no. 1, pp. 231–237, Feb. 2002.

[4] 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, p. 2395, 2018.

[5] C. D. Mathers and D. Loncar, ''Projections of global mortality and burden of disease from 2002 to 2030,'' PLoS Med., vol. 3, no. 11, p. e442, Nov. 2006.

[6] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, ''Heart disease prediction system using associative classification and genetic algorithm,'' in Proc. Int. Conf. Emerg. Trends Elect., Electron. Commun. Technol. (ICECIT), 2012, pp. 40–46.

[7] T. N. Sugathan, C. R. Soman, and K. Sankaranarayanan, ''Behavioural risk factors for non-communicable diseases among adults in Kerala, India,'' Indian J. Med. Res., vol. 127, no. 6, pp. 1–9, 2008.

[8] A. Ahmed and S. A. Hannan, ''Data mining techniques to find out heart diseases: An overview,'' Int. J. Innov. Technol. Exploring Eng., vol. 1, no. 4, pp. 18–23, 2012.

[9] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, ''MLaaS: Machine learning as a service,''

in Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2015, pp. 896–902.

[10] I. Castelli and E. Trentin, ‘‘Combination of supervised and unsupervised learning for training the activation functions of neural networks,’’ Pattern Recognit. Lett., vol. 37, pp. 178–191, Feb. 2014.

[11] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, ‘‘Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features,’’ Res. Cardiovascular Med., vol. 2, no. 3, p. 133, 2013.

[12] D. Tomar and S. Agarwal, ‘‘A survey on data mining approaches for healthcare,’’ Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241–266, 2013.

[13] Y. Er, ‘‘The classification of white wine and red wine according to their physicochemical qualities,’’ Int. J. Intell. Syst. Appl. Eng., vol. 4, no. 1, pp. 23–26, Dec. 2016.

[14] S. J. Pasha and E. S. Mohamed, ‘‘Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction,’’ IEEE Access, vol. 8, pp. 184087–184108, 2020.

[15] D. Swain, S. K. Pani, and D. Swain, ‘‘A metaphoric investigation on prediction of heart disease using machine learning,’’ in Proc. Int. Conf. Adv. Comput. Telecommun. (ICACAT), Bhopal, India, Dec. 2018, pp. 1–6.

[16] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, ‘‘Can machine-learning improve cardiovascular risk prediction using routine clinical data?’’ PLoS ONE, vol. 12, no. 4, Apr. 2017, Art. no. e0174944.

[17] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, ‘‘Machine learning techniques for heart disease datasets: A survey,’’ in Proc. 11th Int. Conf. Mach. Learn. Comput. (ICMLC), Zhuhai, China, 2019, pp. 27–35.

[18] S. Goel, A. Deep, S. Srivastava, and A. Tripathi, ‘‘Comparative analysis of various techniques for heart disease prediction,’’ in Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON), Mathura, India, Nov. 2019, pp. 88–94.

[19] V. Chaurasia and S. Pal, ‘‘Early prediction of heart diseases using data mining techniques,’’ Caribbean J. Sci. Technol., vol. 1, pp. 208–217, 2013.

[20] R. Alizadehsani, M. J. Hosseini, Z. A. Sani, A. Ghandeharioun, and R. Boghrati, ‘‘Diagnosis of coronary artery disease using cost-sensitive algorithms,’’ in Proc. IEEE 12th Int. Conf. Data Mining Workshops, Dec. 2012, pp. 9–16