IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# Evaluating Machine Learning Models For Accurate Cardiovascular Prediction

**[1]Pydi Lahari, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions**

**[2]Dr. B. Sreenivasa Rao, Professor at Miracle Educational Society Group of Institutions**

**[3]Kamath G B S Ramya, Assistant Professor at Miracle Educational Society Group of Institutions**

## ABSTRACT

Diagnosing heart disease is a medical problem that is both critical and has a time factor which has to be taken into consideration in order to be treated. This project looks into the performance analysis of four machine learning models which are SVM, KNN, Logistic Regression and XGBoost, with and without parameter tuning via GridSearchCV. The analysis set up is based on Hungarians Cleveland data set which has attributes for predicting whether the patient is likely to have heart problems. XGBoost computes higher accuracy among the algorithms but had prolonged computation time. The study therefore extends with Random Forest in that regard which parallels the accuracy of XGboost but decreased computation time. This project underlines the importance of parameter tuning in the improvement of model performance and identifies Random Forest as a low cost method which will result in faster and more accurate predictions for heart diseases.

**Keywords:** XGBoost, KNN, SVM

## INTRODUCTION:

The phenomenon of medical data availability brought into discussion the means and methods for improving the performance of life-saving health care on a larger scale. Cardiovascular disease (CVD) has been and still is one of the leading causes of global mortality, where myocardial infarction and strokes are common contributors. The WHO has reported that CVD caused 31% of all global deaths in the year of 2016, that figure is predicted to increase to 23.3 million deaths by 2030. Prevention is essential given the rise in high blood pressure, cholesterol and stress, and associated lifestyle habits. With growing data an enriched view of specific outcomes can be made. By helping find patterns in data, such models can facilitate the diagnostic process and be more proactive.

Heart disease has been detected using both supervised and unsupervised machine learning algorithms– which is a remarkable advantage for healthcare solutions focusing on reducing the mortality rate and economic efficiency of healthcare operations.

## GAP IDENTIFIED BASED ON LITERATURE SURVEY:

Given the discrepancy of patient data, and modeling algorithms, an accurate and a quick diagnosis of a heart ailment is complex. While there are logistic regression, KNN, SVM, and XGBoost as machine learning algorithms for possible classification tasks, there appear to be a handful of possible limitations.

Parameter tuning is a challenge many models faced, it is evident in previous research. GridSearchCV has become a way to improve algorithms based on its hyperparameters. Yet, XGBoost models have been shown to be accurate, they are also very slow, rendering them less useful for real-world situations.

Another gap is also related to the narrow search of different models which would be both accurate and fast enough. High accuracy comes at a cost, most of the research mentioned above does not take into account the time and resources needed.

This project addresses these gaps by systematically comparing algorithms with and without tuning using the datasets from Hungary Cleveland database. It also proposes Random Forest as a suitable substitute for XGBoost which yields the same prediction accuracy but is less costly in terms of computations. By providing improvements that are both accurate and time efficient, this study helps improve the heart disease diagnosis process.

## PROBLEM STATEMENT:

For heart disease prediction, timeliness and accuracy of the ML models used are crucial, but many current ML models have issues with performance and time efficiency.

### Key Challenges:

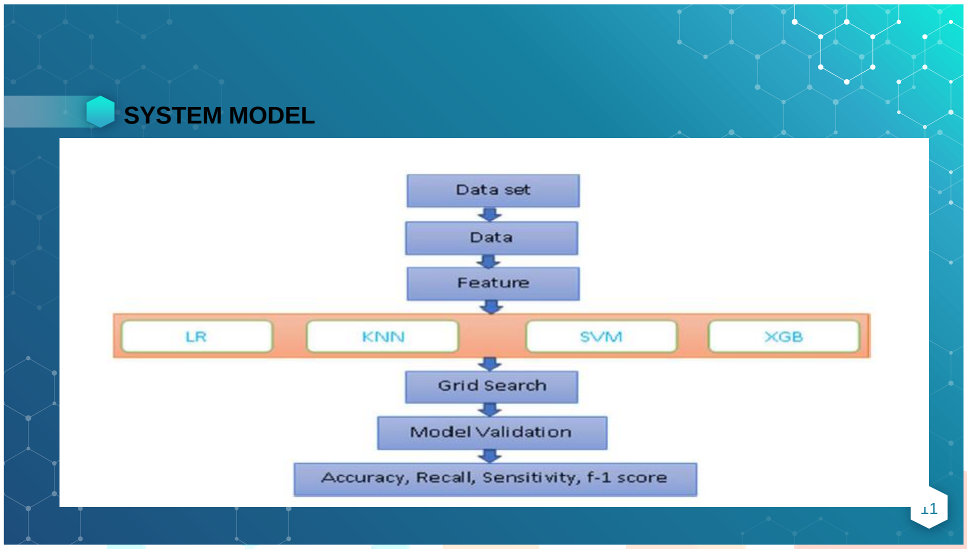
1. High accuracy of heart disease prediction on different patient databases.
2. Parameter tuning approaches in the form of GridSearchCV for machine learning optimizations.
3. The time cost of running such models such as XGBoost.
4. Finding good enough and fast enough solutions which can be used in practical situations.

## PROPOSED METHOD:

Logistic Regression, KNN, SVM, and XGBoost are implemented on the Hungary Cleveland heart disease dataset in this phase of the project. Such algorithms are tuned for optimal classification performance both with and without GridSearchCV parameter tuning. However, its time taken for computation stays unreasonable despite XGBoost attaining a high level accuracy of 100% as mentioned.

On the other hand, Random Forest is said to provide the same accuracy but takes a longer time. This problem is addressed by Random Forest. The stages include dataset cleaning , feature selection, model building , testing and evaluation. Measures like accuracy , precision, recall and computation time could be considered. This way there is always a trade-off between accuracy and speed, which provides a good alternative for diagnosing heart disease.

## ARCHITECTURE:



SYSTEM MODEL

Data set

Data

Feature

LR

KNN

SVM

XGB

Grid Search

Model Validation

Accuracy, Recall, Sensitivity, f-1 score

## DATASET:

Heart disease prediction is possible using the Hungary Cleveland dataset which was obtained from the UCI repository. It includes features such as age, sex, cholesterol levels, blood pressure, and ECG results. It involves normalizing values and missing data and also transforming the categorical variables to numerical variables.

This classification encases healthy patients and sick patients suffering with heart disease. Due to the detailed nature of the attributes, the dataset is ideal to conduct tests with the machine learning algorithms, allowing for greater assessment of model performance when parameters are set and altered. This guarantees solid observations on how effective algorithms are in predicting one's cardio health.

## METHODOLOGY:

### Data Preprocessing:

Load and preprocess the Hungary Cleveland dataset.

Rescale numerical variables and encode the categorical features.

Impute all necessary variables to complete the dataset in the best manner.

### Baseline Model Training:

Apply the data to the Logistic Regression, KNN, SVM and XGBoost, and make predictions.

Make evaluation of performance and accuracy parameters without tuning the parameters.

**Parameter Tuning with GridSearchCV:**

Fine-tune every model using combinations of parameters using GridSearchCV including for instance max_iter, distances, and the type of kernel.

Each model's performance and the computation time are each recorded.

**Performance Evaluation:**

Investigate the computations that were performed aside from tuning that include time, accuracy, precision, and recall.

Argue how parameter tuning affected the performance metrics of the model.

**Introduction of Random Forest:**

A farmer can use Random Forest instead of XGBoost when there is computational load.

Fine tune Random Farm and fit it along with oters and without it.

**Comparison and Visualization:**

Provide the performance measurements of all algorithms, accuracy and computation time of the fastest one are the primary goals.

Include graphical information depicting bar plots and confusion matrices for understanding.

**Validation on Test Data:**

Utilize the trained Random Forest model on the able data and check its validity on the predicted answers.

Lastly, prediction results made earlier should be put against reality in order to check validity.

## EVALUATION:

**Precision:**

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

**Recall (Sensitivity):**

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

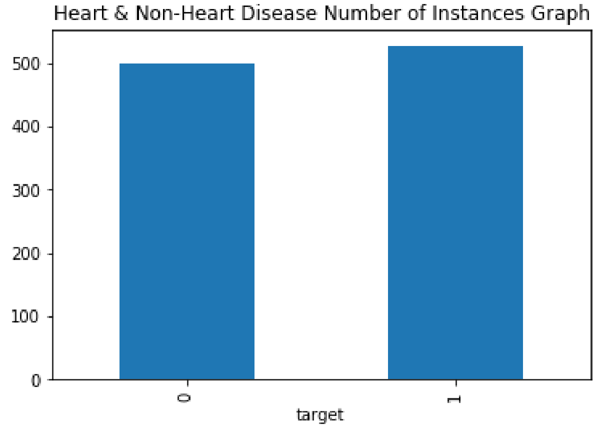
**F1 Score:**

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

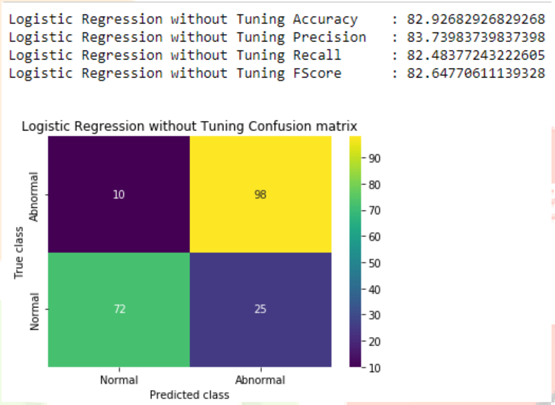
### Accuracy:

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

### RESULTS:

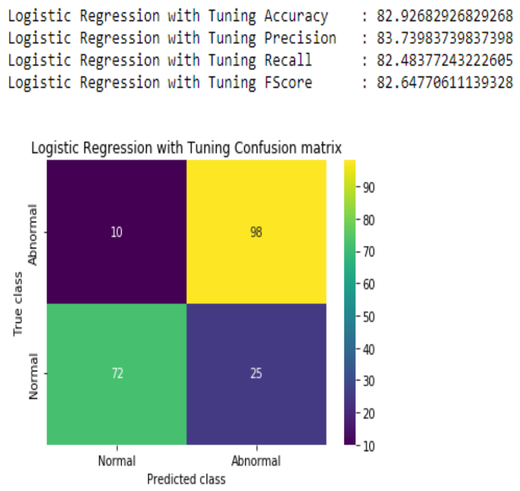


In above graph x-axis represents 0 as Healthy and 1 as Non healthy and y-axis represents count



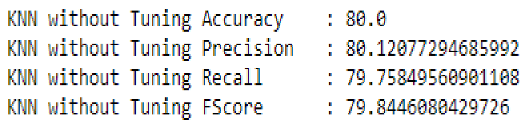
```
Logistic Regression without Tuning Accuracy    : 82.92682926829268
Logistic Regression without Tuning Precision   : 83.73983739837398
Logistic Regression without Tuning Recall      : 82.48377243222605
Logistic Regression without Tuning FScore      : 82.64770611139328
```

In above screen we are training logistic regression without tuning and we got its accuracy as 82%

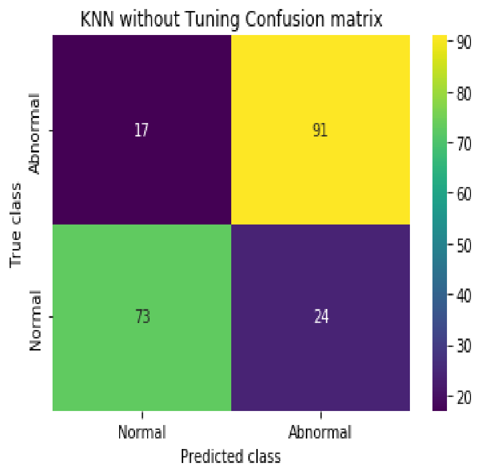


```
Logistic Regression with Tuning Accuracy    : 82.92682926829268
Logistic Regression with Tuning Precision   : 83.73983739837398
Logistic Regression with Tuning Recall      : 82.48377243222605
Logistic Regression with Tuning FScore      : 82.64770611139328
```

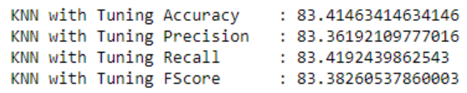
Logistic regression with Tuning parameters but we got same 82% accuracy



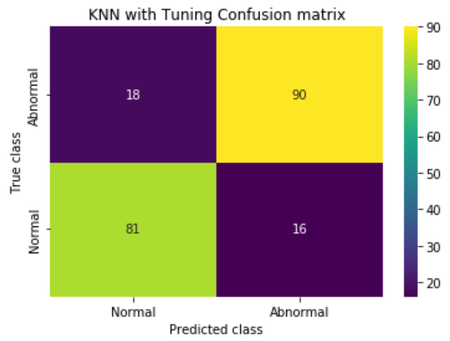
```
KNN without Tuning Accuracy   : 80.0
KNN without Tuning Precision  : 80.12077294685992
KNN without Tuning Recall     : 79.75849560901108
KNN without Tuning FScore     : 79.8446080429726
```



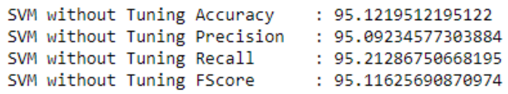
KNN without tuning and we got accuracy as 80%



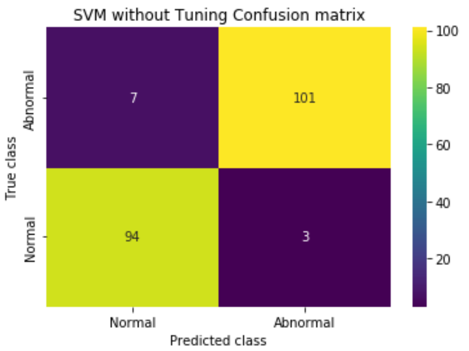
```
KNN with Tuning Accuracy    : 83.41463414634146
KNN with Tuning Precision   : 83.36192109777016
KNN with Tuning Recall      : 83.4192439862543
KNN with Tuning FScore      : 83.38260537860003
```



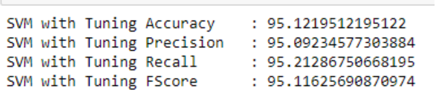
KNN with tuning parameters and we got increased accuracy as 83%



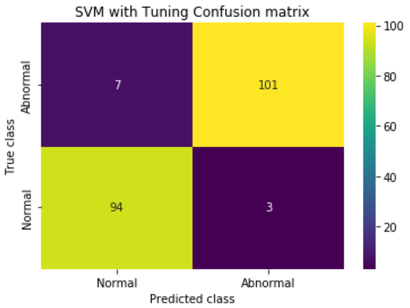
```
SVM without Tuning Accuracy    : 95.1219512195122
SVM without Tuning Precision   : 95.09234577303884
SVM without Tuning Recall      : 95.21286750668195
SVM without Tuning FScore      : 95.11625690870974
```



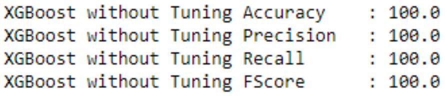
SVM without tuning and we got accuracy as 95%



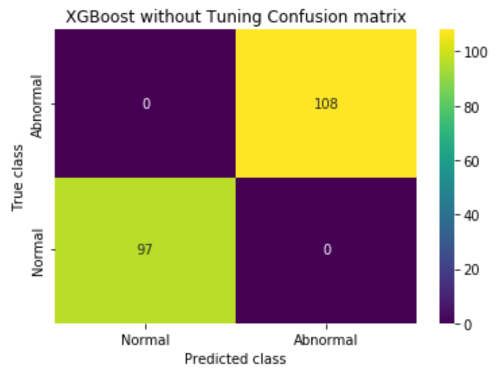
```
SVM with Tuning Accuracy    : 95.1219512195122
SVM with Tuning Precision   : 95.09234577303884
SVM with Tuning Recall      : 95.21286750668195
SVM with Tuning FScore      : 95.11625690870974
```



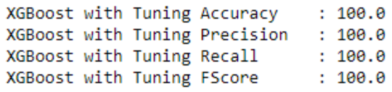
SVM with tuning and we got same accuracy as 95%



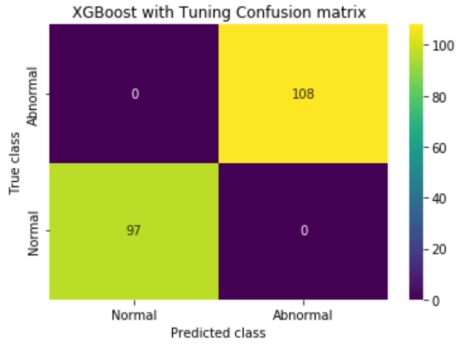
```
XGBoost without Tuning Accuracy    : 100.0
XGBoost without Tuning Precision   : 100.0
XGBoost without Tuning Recall      : 100.0
XGBoost without Tuning FScore      : 100.0
```

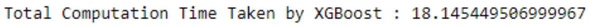


XGBOOST without tuning got 100% accuracy



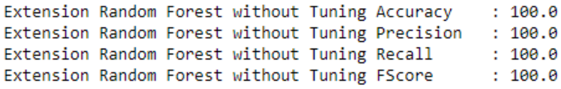
```
XGBoost with Tuning Accuracy    : 100.0
XGBoost with Tuning Precision   : 100.0
XGBoost with Tuning Recall      : 100.0
XGBoost with Tuning FScore      : 100.0
```



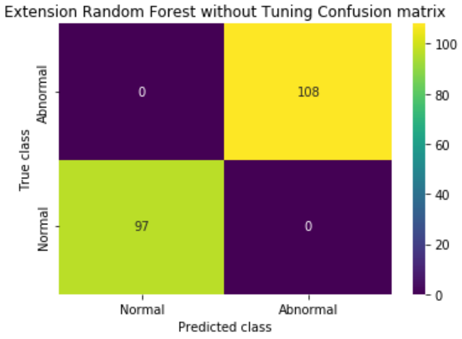


```
Total Computation Time Taken by XGBoost : 18.145449506999967
```

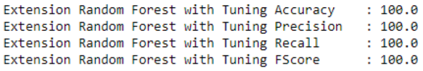
In above screen XGBOOST with tuning also got 100% accuracy computation time is 18 seconds



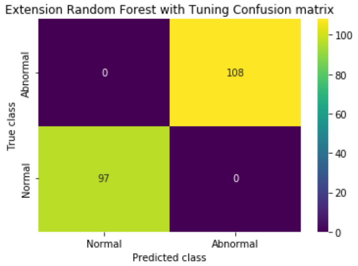
```
Extension Random Forest without Tuning Accuracy    : 100.0
Extension Random Forest without Tuning Precision   : 100.0
Extension Random Forest without Tuning Recall      : 100.0
Extension Random Forest without Tuning FScore      : 100.0
```

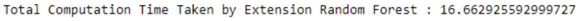


Random Forest without tuning and got accuracy as 100%



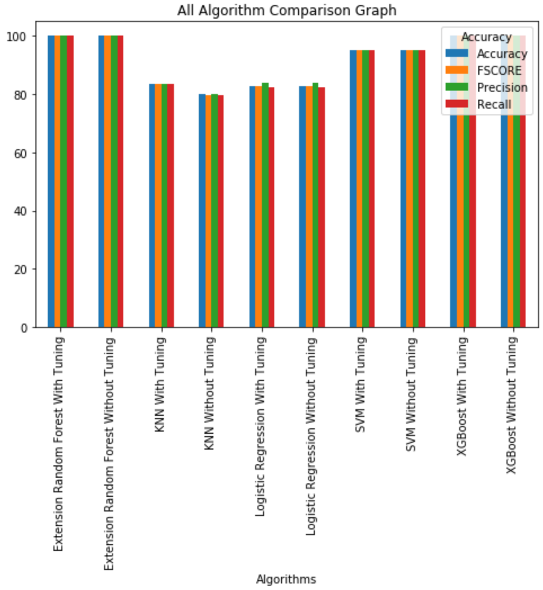
```
Extension Random Forest with Tuning Accuracy    : 100.0
Extension Random Forest with Tuning Precision   : 100.0
Extension Random Forest with Tuning Recall      : 100.0
Extension Random Forest with Tuning FScore      : 100.0
```





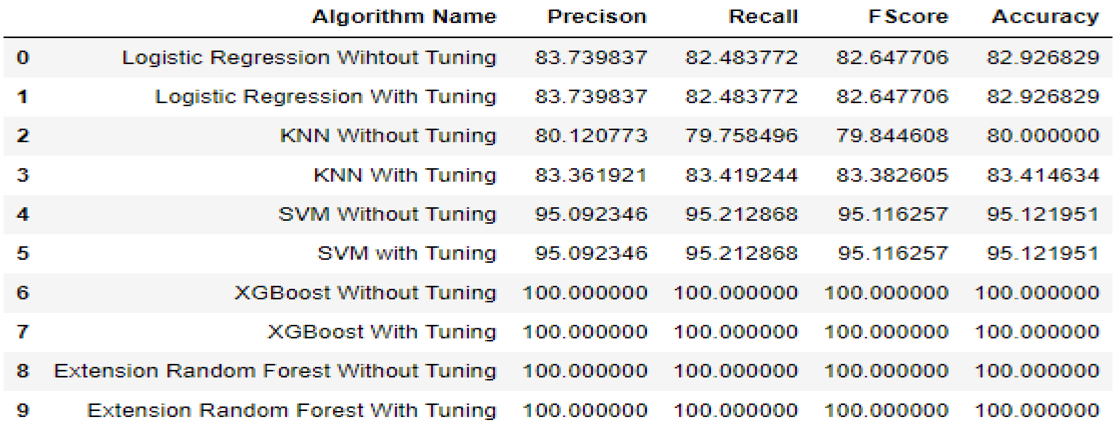
```
Total Computation Time Taken by Extension Random Forest : 16.662925592999727
```

Random Forest computation time is 16 seconds and XGBOOST took 18 seconds



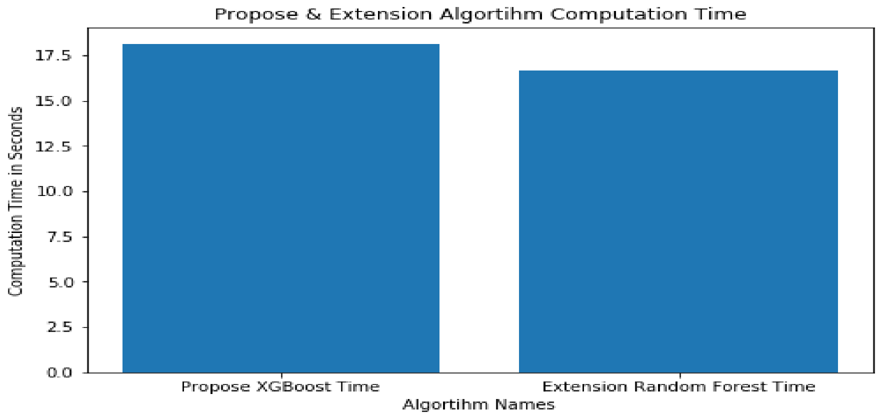
We can see XGBOOST and extension Random Forest got high accuracy

:



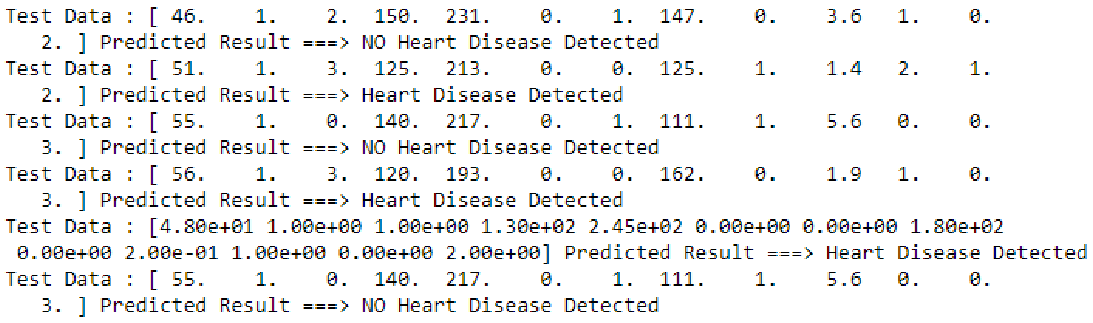
| | Algorithm Name | Precison | Recall | FScore | Accuracy |
|---|---|---|---|---|---|
| 0 | Logistic Regression Wihtout Tuning | 83.739837 | 82.483772 | 82.647706 | 82.926829 |
| 1 | Logistic Regression With Tuning | 83.739837 | 82.483772 | 82.647706 | 82.926829 |
| 2 | KNN Without Tuning | 80.120773 | 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



Computation time comparison between XGBOOST and Random Forest

**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
```

Predicted output as ‘Heart Disease or No Heart Disease’

## CONCLUSION

This research shows that the XGBoost classifier, despite attaining very high responsiveness in predicting heart disease, can be very expensive to use. The Random Forest does not require much time and is able to give similar accuracy. GridSearchCV improves the performance of the model significantly, but also increases the time of execution. The results highlight the tradeoffs between accuracy and efficiency in the machine learning application for medical problems. Through more efficient algorithms such as Random Forest, this research enhances the efficiency of heart disease diagnosis enabling better healthcare delivery and real-time diagnosis systems.

**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