Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

# HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

**Dr. Aniket Siddhaling Kothawale,** Assistant Professor, Electronics Department, Tuljaram Chaturchand College Baramati 413102. kothawaleaniket71@gmail.com
**Dr. Ashok Eknath Kalange,** HOD & Professor, Physics Department, Tuljaram Chaturchand College Baramati 413102.
**Mr. Chandrashekhar Swami,** Assistant Professor, Statistics Department, Tuljaram Chaturchand College Baramati 413102.
**Mr. Sandip Kakade,** Assistant Professor, Physics Department, Tuljaram Chaturchand College Baramati 413102.

**Abstract:** The world is facing one of its most difficult challenges: heart disease. Predicting cardiovascular disease is difficult in medical data analysis. Machine learning (ML) can make predictions from large amounts of healthcare and hospital data. This research work predicts coronary heart disease and provides risk information to the patient. The prediction model uses many options and categorization methods. This is done by comparing the accuracy of various algorithms to the outputs of the hybrid system version separately and selecting the most accurate method for prediction. In this paper various machine learning techniques such as Logistic Regression, K Nearest Neighbour, Random Forest Classifier, Decision Tree, XG Boost Classifier, and Support Vector Machine have been utilized to predict heart disease based on their accuracy, recall, precision, and F1-Score. The simulation results have been obtained using Python.

**Key words**: Heart Disease, Machine Learning, Accuracy.

## INTRODUCTION:

There are several conditions that contribute to the uncertainty that is associated with the description of cardiac disease [1]. These conditions include high blood pressure, excessive cholesterol, an irregular pulse rate, and many others. A variety of data mining and neural network techniques have been utilized to investigate the severity of human cardiac disease. To classify the severity of the illness, several different approaches are utilized, such as the K-Nearest Neighbour Algorithm, Decision Trees, Logistic Regression, and the Support Vector Machine. It is necessary to exercise caution when managing heart disease because it is a complicated illness [2]. If this is not done, serious cardiovascular issues or even death may result. The perspectives of medical research and data mining need to be combined to discover the many different types of metabolic diseases [3]. The analysis of large amounts of data and the prediction of cardiovascular illness are two areas in which sequence-based data mining is particularly useful. In many parts of the world, the concept of machine learning is frequently put into practice. Because it will make it easier for medical professionals to diagnose patients in a shorter amount of time, it has the potential to improve the healthcare industry. The objective of this text is to apply the heart disease dataset to evaluate and anticipate whether a patient has coronary heart disease. In other words, the purpose of this text is to apply system learning to anticipate coronary heart disease. It is possible that the healthcare industry will be able to develop more rapidly and successfully because of this prediction, which will save time.

A dangerous diet, a lack of physical exercise, the use of cigarettes, and the consumption of alcohol in excessive amounts are the most primary behavioral risk factors for coronary heart disease and stroke. Behavioral risk factors can also cause individuals to experience additional signs and symptoms, such as high blood pressure, high blood glucose, high blood lipids, and obesity or weight problems. It is also possible for individuals to have these symptoms. These "intermediate threat elements" can be identified in primary care settings and indicate an increased risk of cardiovascular events such as heart attacks, strokes, heart failure, and other adverse outcomes.

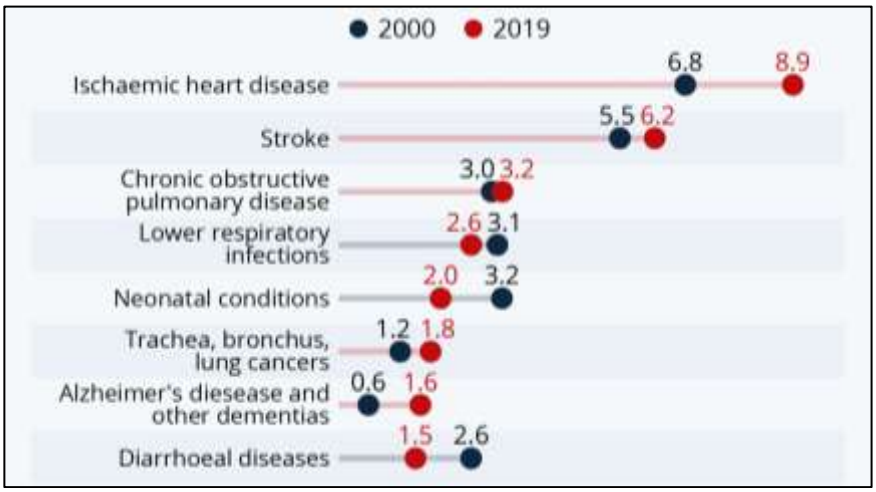



Figure 1 Causes of death that are the most prevalent around the world (total number of deaths in millions)

Risk factors that can be modified or eliminated by adopting specific activities are referred to as modifiable risk factors or controlled risk factors [4]. Variable risk factors are also known as controlled risk factors. The World Health Organization provided definitions of risk factors. Figure 2 illustrates a few different risk factors that can be controlled for cardiovascular disease:

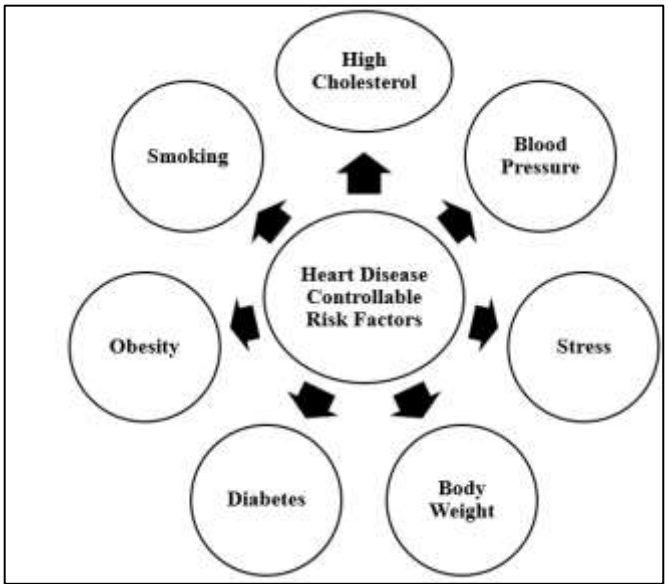



Figure 2 Heart Disease Risk Factors

**1.1 Classification of Machine Learning :** Machine learning is mainly classified in two categories: Supervised and Unsupervised Technique.

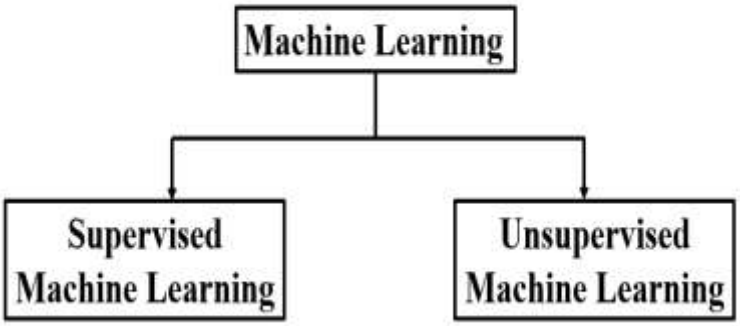



Figure 3 Machine Learning Techniques

**(A) Supervised Machine Learning Techniques:** For generating the training model, these methods make use of evidence that is largely known and verified. Figure 4 illustrates the classification of supervised learning in its various forms [5]. In the process of supervised

learning, the predictive models are constructed through the application of classification and regression techniques.

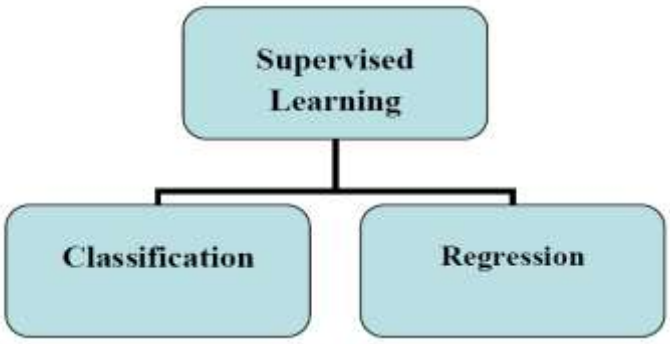



Figure 4 Supervised Machine Learning Techniques

**(B) Unsupervised Machine Learning Techniques:** These methods do not involve the use of training datasets to supervise the implementation of models [6]. The process of learning some novel concepts in the human mind is analogous to the techniques that are being discussed here.

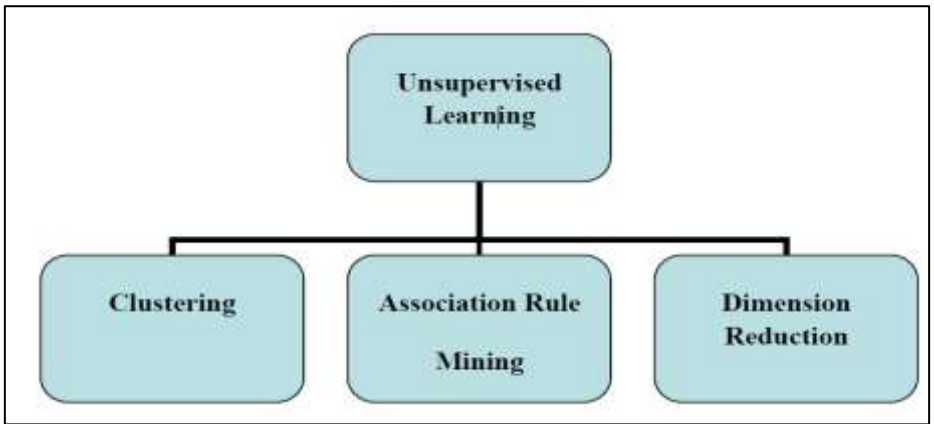



Figure 5 Unsupervised Machine Learning Techniques

**2. Related Work:** SVM and Naive Bayes were used to analyse heart disease prediction systems in [7]. This system classifies medical data as none, low, normal, high, and unusually high. To improve categorization (training) and prediction (testing), the system uses the database update method. In [8], the authors examined cardiac disease prediction methods. Fuzzy systems and support vector machine algorithms can quickly diagnose cardiac illness, aiding endurance therapy. Performance indicators included faster categorization and sensitivity. The authors developed a cardiac classification and prediction method to evaluate coronary artery disease for medical decision-making in [9]. The Cleveland Hart dataset archive was used. Accuracy tests showed significant improvement. The heart is an essential organ contained within the human body. Through the blood vessels that make up the circulatory system, this organ is responsible for pumping blood [10]. It is essential for physical mobility that oxygen be transported throughout the body, and the blood plays a role in this process. The coronary heart beats one hundred thousand times every single day. Frequently, conditions that affect the heart are referred to as cardiovascular diseases (CVDs) [11]. One of the most common causes of death around the world is coronary heart disease. According to the World Health Organisation (WHO), heart disease affects both males and females in the same proportion. In 2016, heart disease was responsible for the deaths of 17.9 million people around the world, which is equivalent to 31% of all deaths that occurred worldwide. Stroke and coronary heart attacks were responsible for 85 percent of those deaths in 2016, according to the World Health Organisation. In the article [12], the authors propose that multiple linear regression is an effective method for predicting the risk of coronary heart disease. This prediction is in line with the Heart Disease Prediction Model, which makes use of different algorithms. The research utilised a raw statistics series that was repeated one thousand times and contained ten functions that were pre-specified as

being exclusive. Since it is far from the consequences that the regression algorithm's end is the maximum in comparison to other algorithms, the statistics are separated into stages of statistical division. Seventy percent of the statistics are used to teach the system, and thirty percent of the statistics are used for making experiments. The authors of [13] proposed targeting methods for detecting persistent disorder with the assistance of extracting statistics from beyond fitness facts through the utilisation of Nave Bayes, Decision Trees, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). These methods were proposed regarding the detection of persistent disorders. The SVM precision rate that was obtained from this test was of the highest possible quality, and the Nave Bayes method also provided the most accurate results. Regarding the prediction of coronary heart disease, non-linear class strategies, which agree with [14], are most likely utilised. Within the scope of this study, the application of various information mining algorithms for the purpose of identifying cardiac illness was also thoroughly examined. This includes the storage of enormous amounts of data across many nodes through the utilisation of HDFS, as well as the appearance of the prediction set of rules through the utilisation of SVM across many nodes at the same time. It is utilised in a manner that is comparable, with processing times that are significantly faster than the norm. The use of the UCI machine learning dataset for the purpose of predicting cardiac disease has garnered a lot of the attention that it deserves. A discussion of some of the data mining techniques that have been utilised to achieve varying degrees of precision is presented in the following example. There are a few machine learning techniques that have been discussed in [15] for their potential utility in the classification of cardiac diseases. Extensive research was conducted to investigate the effectiveness of the Decision Tree, KNN, and K-Means algorithms for classification using the data collected. Based on the findings of this study, it was determined that the decision tree produced the most accurate results. Furthermore, it was concluded that the tool could be improved by employing a variety of different approaches.

**3. Proposed Method:** The proposed model for the heart disease prediction is shown in Fig. 6. There are six distinct classification models that have been compiled into the hybrid model that has been proposed. Random Forest, Decision Tree, Logistic Regression, Support Vector Machine, KNN, and XGBoost are some of the models that are featured in this collection. Following the incorporation of six distinct categorization models and the comparison of those models, the performance can be taken into consideration.

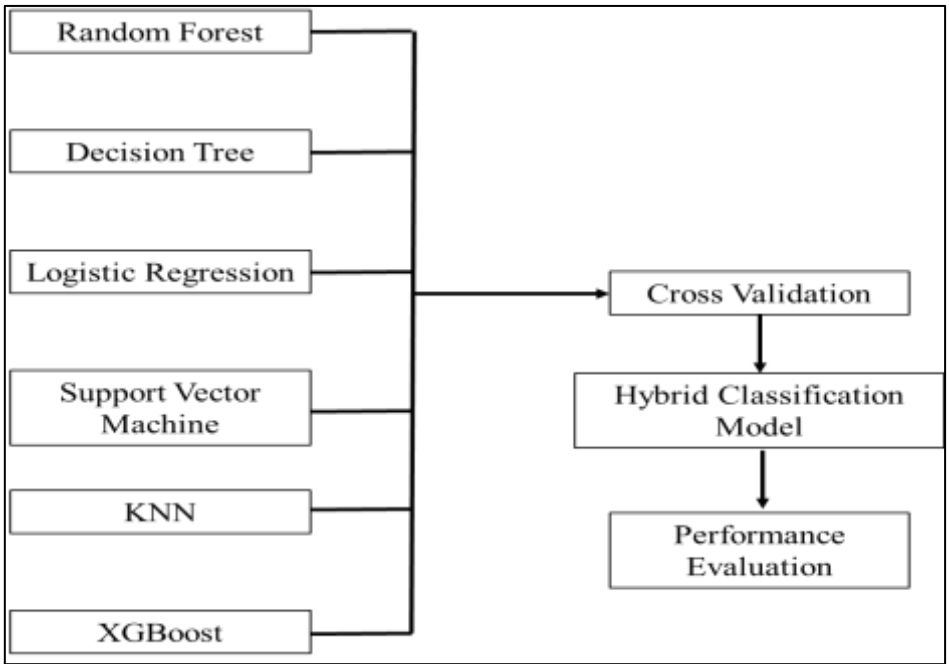



Fig. 6 Proposed hybrid classification model

**4. Result and Discussion:** A confusion matrix is utilized for the purpose of performance measurement in the implementation of machine learning classification. There is a type of table that is used to assist in determining how effective the classification model is when applied to a set of test

data for which the actual values are already known. It does this by contrasting the actual classes with the ones that were predicted to provide a visual representation of how accurate a classifier is. The binary confusion matrix is composed of the following squares, which are named as follows:

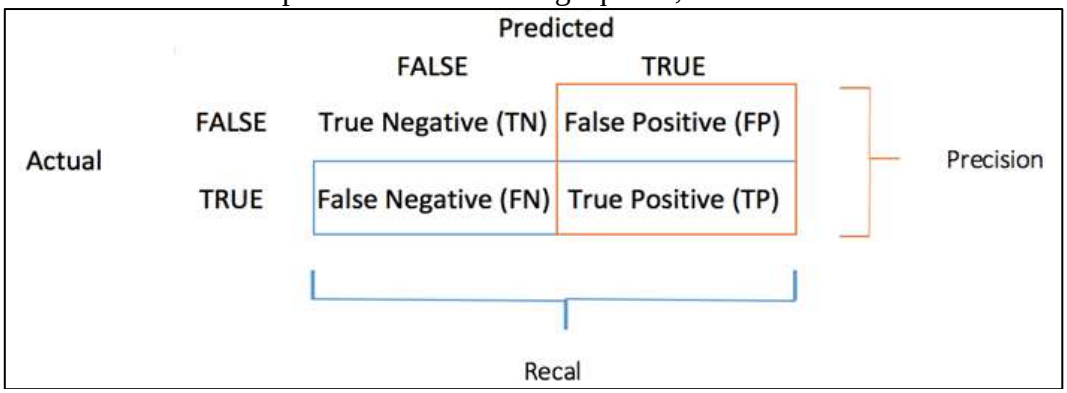


Figure 7 Confusion Matrix

TP: True Positive: Predicted values matched actual positives
FP: A positive was predicted incorrectly. i.e., negative values are predicted positively.
FN: Not true. Negative: Positive is assumed to be negative.
TN: true negative: correctly predicted negative values.

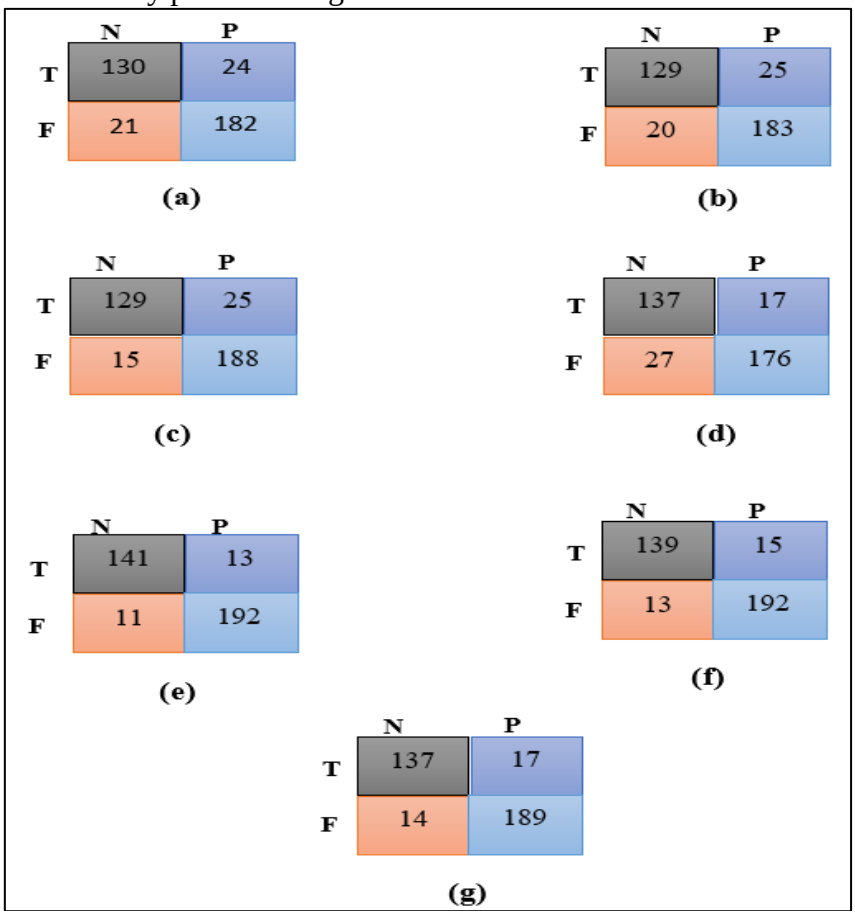
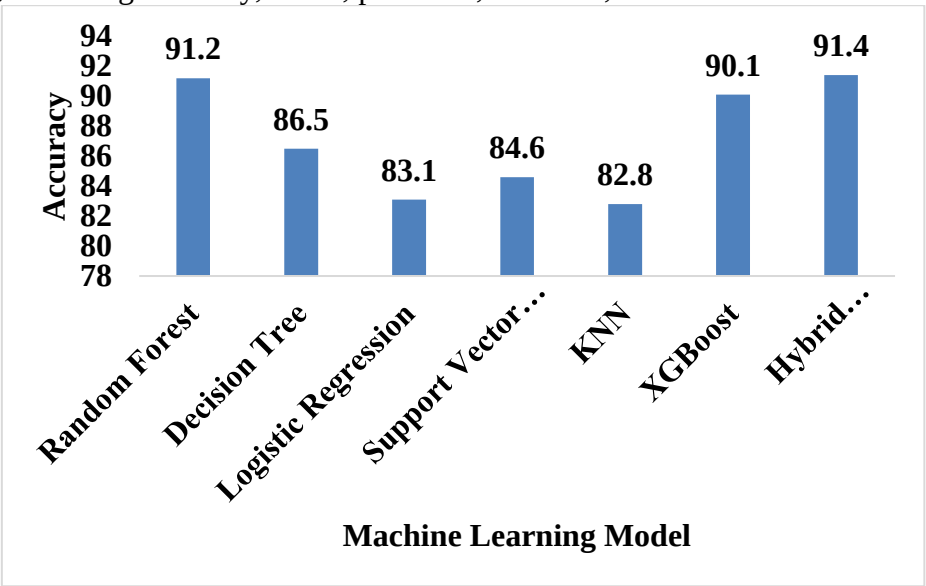


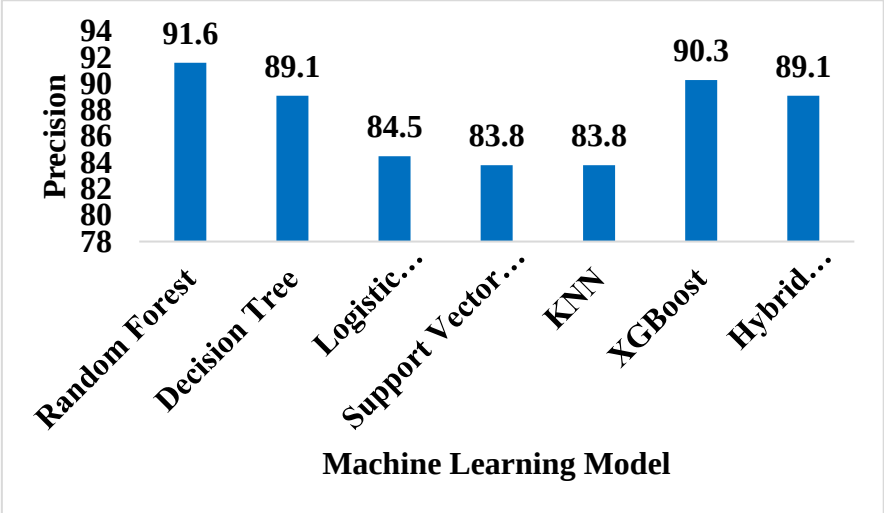
Figure 8 Confusion Matrix of (a) Logistic Regression (b) K-Nearest Neighbors (c) Support Vector Machine (d) Decision Tree Classifiers (e) Random Forest (f) XGBoost (g) Hybrid Classifier

A hybrid classification model that has been suggested is applied to the test data to provide a prediction regarding the class label. This is accomplished using a machine learning classification model. An evaluation of the classifier's performance is carried out by comparing it to the performance of other traditional methods, such as decision trees, support vector machines, and other

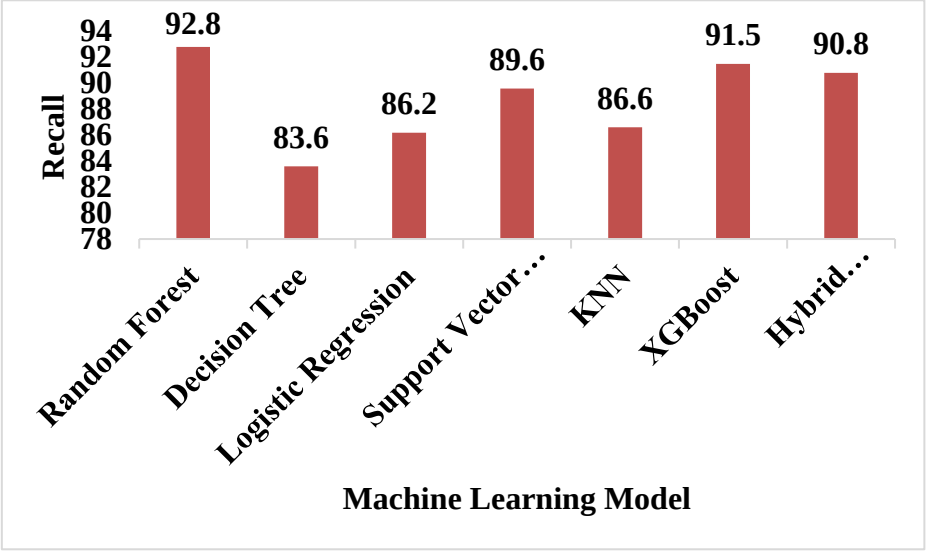
techniques. The comparison is carried out with the assistance of a wide variety of parameter measurements, including accuracy, recall, precision, F1-score, and confusion matrix.



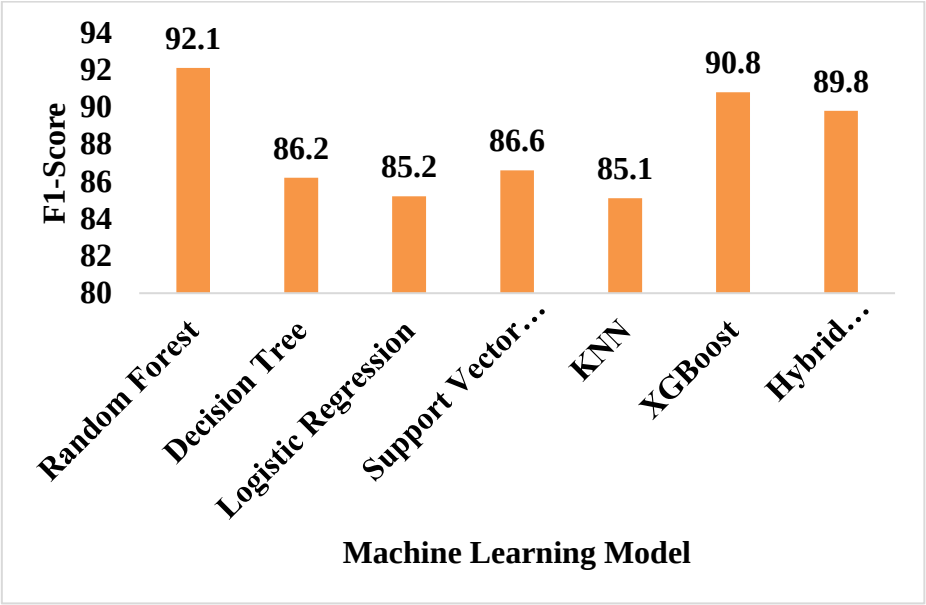

(a)




(b)




(c)




(d)

Figure 9 Comparison of different Machine Learning Models for (a) Accuracy, (b) Precision, (c) Recall (d) F1-score

## 5. Conclusion

The heart disease predictor has been designed in this paper to predict early detection of heart disease, so that a patient can be treated on time. In this paper, the proposed classification model has been compared to existing methods based on F1 score, precision, recall, and accuracy. The F1 score measures how well the proposed classification recalls information. Performance analysis compares proposed classifier results to established classifier results. The UCI repository heart disease dataset was used to test the suggested classifiers' performance metrics. The hybrid classification approach proposed in this thesis outperformed current algorithms.

## References

[1] Z. Gao et al., “Developing and Validating an Emergency Triage Model Using Machine Learning Algorithms with Medical Big Data,” Risk Manag. Healthc. Policy, vol. 15, pp. 1545–1551, 2022, doi: 10.2147/RMHP.S355176.

[2] A. Kumar, R. Gupta, and R. Bhandari, “WoS Bibliometric-based Review on Serverless Computing model,” PDGC 2022 - 2022 7th Int. Conf. Parallel, Distrib. Grid Comput., pp. 600–605, 2022, doi: 10.1109/PDGC56933.2022.10053142.

[3] A. Pyrros et al., “Predicting Prolonged Hospitalization and Supplemental Oxygenation in Patients with COVID-19 Infection from Ambulatory Chest Radiographs using Deep Learning,” Acad. Radiol., vol. 28, no. 8, pp. 1151–1158, Aug. 2021, doi: 10.1016/j.acra.2021.05.002.

[4] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, “Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes,” IEEE ACCESS, vol. 8, pp. 120537–120547, 2020, doi: 10.1109/ACCESS.2020.3005540.

[5] V. Sharma, S. Goel, A. K. Jain, A. Vajpayee, R. Bhandari, and R. G. Tiwari, “Machine Learning based Classifier Models for Detection of Celestial Objects,” 2023 3rd Int. Conf. Intell. Technol. CONIT 2023, pp. 1–7, 2023, doi: 10.1109/CONIT59222.2023.10205666.

[6] D. Kumari and R. Bhandari, “Machine Learning Classification Techniques to investigate Parkinson’s disease,” 7th Int. Conf. Trends Electron. Informatics, ICOEI 2023 - Proc., no. Icoei, pp. 1162–1168, 2023, doi: 10.1109/ICOEI56765.2023.10125933.

[7] S. Sakr et al., “Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project,” BMC Med. Inform. Decis. Mak., vol. 17, Dec. 2017, doi: 10.1186/s12911-017-0566-6.

[8] K. Bahani, M. Moujabbir, and M. Ramdani, “An accurate fuzzy rule-based classification systems for heart disease diagnosis,” Sci. AFRICAN, vol. 14, Nov. 2021, doi: 10.1016/j.sciaf.2021.e01019.

[9] S. Weichwald et al., “Improving 1-year mortality prediction in ACS patients using machine learning,” Eur. Hear. JOURNAL-ACUTE Cardiovasc. CARE, vol. 10, no. 8, pp. 855–865, Oct. 2021, doi: 10.1093/ehjacc/zuab030.
[10] T. Takura, K. H. Goto, and A. Honda, “Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases,” BMC Med., vol. 19, no. 1, Jan. 2021, doi: 10.1186/s12916-020-01874-6.
[11] G. Konstantonis et al., “Cardiovascular disease detection using machine learning and carotid/femoral arterial imaging frameworks in rheumatoid arthritis patients,” Rheumatol. Int., vol. 42, no. 2, pp. 215–239, Feb. 2022, doi: 10.1007/s00296-021-05062-4.
[12] J. R. A. Solares et al., “Long-Term Exposure to Elevated Systolic Blood Pressure in Predicting Incident Cardiovascular Disease: Evidence From Large-Scale Routine Electronic Health Records,” J. Am. Heart Assoc., vol. 8, no. 12, Jun. 2019, doi: 10.1161/JAHA.119.012129.
[13] C. Sowmiya and P. Sumitra, “Morality prediction model in cardiovascular disease with significant feature selection and hybrid KNN classification technique,” Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 5497–5502, 2019, doi: 10.35940/ijitee.K2344.1081219.
[14] F. Drenos, E. Grossi, M. Buscema, and S. E. Humphries, “Networks in Coronary Heart Disease Genetics As a Step towards Systems Epidemiology,” PLoS One, vol. 10, no. 5, May 2015, doi: 10.1371/journal.pone.0125876.
[15] R. Jothiramalingam, A. Jude, and D. J. Hemanth, “Review of Computational Techniques for the Analysis of Abnormal Patterns of ECG Signal Provoked by Cardiac Disease,” C. Model. Eng. \& Sci., vol. 128, no. 3, pp. 875–906, 2021, doi: 10.32604/cmes.2021.016485.