iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** X **Month of publication:** October 2020

DOI: https://doi.org/10.22214/ijraset.2020.31917

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

Heart Disease Prediction using Machine Learning Techniques: A Survey

M. Merla Agnes Mary[1], Dr. T. Lucia Agnes Beena[2]
[1]*Research Scholar,* [2]*Assistant Professor, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India*

Abstract: ***Medical diligences generate a huge amount of data. Cardiovascular Disease is becoming a common disease leads to death now a days. Hence it is important to predict as early as possible. Heart disease data are stored in a database in large amount. Different Machine Learning techniques (SVM, Logistic Regression, Neural Network, KNN, RF, Naïve Bayes, DT, and GDBT- Bagging Tree) can be used to classify the data in the database. This paper concentrates on various ML techniques that are used to predict the heart disease by using dataset. The accuracy, sensitivity, specificity and Area under Curve (AUC) are calculated for various techniques. The result shows that Artificial Neural Network (ANN) algorithm with SAE produce the highest accuracy of 90% respectively. Ensemble method of techniques is also used to improve an accuracy of heart disease.***
Keywords: Heart disease, Support Vector Machine, Logistic Regression, Neural Network, K Nearest Neighbor, Random Forest, Naïve Bayes, Decision Tree, GDBT- Bagging Tree

I. INTRODUCTION

Day by day amount of data is increasing in medical industry and storing data become difficult. In health industry heart disease is one of the main disease that cause many deaths with sequence of diseases involving the circulatory system, including angina pectoris, myocardial infarction, coronary heart disease, heart failure, arrhythmia and related to atherosclerosis etc.

The World Health Organization (WHO) has estimated that 17.7 million deaths have occurred worldwide due to cardiovascular diseases (WHO, 2017). CVDs are the number one cause of death globally. Further people die annually from CVDs than from any other causes [1].

A result of heart disease and blood vessels, and It is estimated that by 2030 that number will increase to 23 million. For decades, cardiovascular diseases have been a significant part of the leading causes of disease in the population [2]. In healthcare field, due the huge amount of data (big data) generated from multiple areas, multiple sources such as streaming machines, advanced healthcare systems, high throughput instruments, sensor networks, internet of things, mobile application, data collection and processing is becoming very common these days [4].

In conformity with global trends, the upsurge in the number of deaths from cardiovascular disease in India is influenced by population growth and aging without the decline in age specific death rates deserved in several other countries [6]. Heart disease is also one of the top ten leading causes of death globally and it is also the second cause of deaths in Japan. Advances in computer technology and an increasing interest in artificial intelligence have raised the research and development [7].

Most heart diseases cannot be detected by primitive ECG process so for that many preventive sensors or devices such as PCG, EGM, are invented [14].

In order to resolve these complexities in invasive-based diagnosing of heart disease, a noninvasive medical decision support system based on machine learning predictive models are developed by various researchers. Due to these machine-learning-based expert medical decision system, the ratio of heart disease death is observed to be decreased [8]. The heart disease databases are available at the University of California Irvine (UCI) repository.

This data has been available since 1988 and used by many researchers in heart disease prediction research Cleveland dataset has been used in machine learning. It consists of 303 instances of the record with 14 attributes, 13 being the independent variable [16].

In this paper, the use of different machine learning algorithms in the prediction of heart disease such as RF, SVM, Neural Network, KNN, DT, Naïve Bayes, BT and Logistic Regression are discussed.

Table 1: The main terminologies mentioned in this paper (include abbreviations)

List of terminologies (method and indices)			
ECG	Electrocardiography	CART	Classification And Regression Tree
PCG	Phonocardiogram	BN	Bayes Net
EGM	Electromyogram	PART	Projective Adaptive Resonance Theory
KNN	K-Nearest Neighbour	RBF	Radial Basis Function Kernel
SVM	Support Vector Machine	MCC	Matthews Correlation Coefficient
RF	Random Forest	AUC	Area Under the Curve
DT	Decision Tree	LASSO	Least Absolute Shrinkage and Selection Operator
BT	Bagging Tree	mRMR	Minimal-Redundancy-Maximal-Relevance
GBDT	Gradient Boosting Decision Tree	PSO	Particle Swarm Optimization
SAE	Sparse autoencoder	PCA	Principal Component Analysis
ANN	Artificial Neural Network	MLPNN	Multilayer Perceptron Neural Network

Rest of the paper is organized as follows: Section II describes about the heart disease. In Section III state an art of heart disease along with the dataset. Different algorithms of machine learning are elaborated in Section IV. The Section V analyzes the comparison of the machine learning algorithms and the paper ends with the conclusion in Section VI.

II. HEART DISEASE

The heart is important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient work of the heart. The term Heart disease refers to disease of heart & blood vessel system within it. Some types of heart disease are coronary heart disease, angina pectoris, congestive heart failure, cardiomyopathy, congenital heart disease, arrhythmias, myocarditis, heart attack and heart cancer. Among these diseases cardiovascular disease or coronary heart disease is very dangers disease. Some reasons of heart disease are

A. Age
B. Smoking
C. Sugar
D. Obesity
E. Depression
F. Hyper tension
G. High blood
H. Cholesterol
I. Poor diet
J. Family history
K. Physical inactivity

There are many types of heart disease that affect different parts of the organ and occur in different ways.

1) Congenital Heart Disease
2) Arrhythmia
3) Coronary Artery Disease
4) Dilated Cardiomyopathy
5) Myocardial Infarction
6) Heart Failure
7) Hypertrophic Cardiomyopathy
8) Mitral Regurgitation
9) Mitral valve Prolapse
10) Pulmonary stenosis

The Symptoms of Heart disease vary from an individual to individual. The common symptoms are chest pain (angina pectoris), Indigestion, Heart burn, Stomach ache, sweating and nausea.

III. LITERATURE SURVEY

Shafenoor Aminet.al [1] proposed the comparison of two ensemble algorithm (Naïve Bayes and Logistic Regression) with vote method and achieved an accuracy of 87.41% by using nine attributes for the implementation. Among the nine attributes chest pain (cp) is used by most of the researchers. The prediction system is developed as intelligent heart disease prediction system.

Damir Imamovic, Elmir Babovic [2] implement three types of data mining algorithms like Decision Tree, Logistic Regression and Neural Networks and get F1 measure as 76.62%, 80.17% and 83.12% respectively. They conclude that the Neural Network is best among the three. The accuracy is recorded as 65%, 69% and 74.34% , precision is described as prediction of (alive and dead) as (81.31% of alive,26.67% of dead), (83.04 of alive,32.50% of dead), (86.49% of alive,41.46% of dead) and recall is described as True (alive and dead) with (72.50% of alive,37.50% of dead), (77.50% of alive,40.62% of dead), (80.00% of alive ,53.12% of dead) respectively for the three algorithms.

Ibomoiye et.al. [3] have compare more machine learning algorithms and proposed a new method by combine SAE+ANN and achieved the accuracy of 90%, precision of 89%, recall with 91%and F1 measure as 90%. SAE network used (100, 75, 50, and 25) layers in encoder and decoder.

Khalil Maalmi et.al [4] proposed the better framework for early detection of heart disease by comparing Apache Spark and other traditional frameworks. First Spark MLib is used with Spark Streaming by the classifier random forest for predicting the heart disease. Second Spark Cassandra is used to store large volume of data. Random forest classifier is used along with the Apache Spark to evaluate the execution time of the records. First Apache Spark and other frameworks is compared to estimate the time taken to the build the random forest, Second both are compared to estimate the time taken to test the random forest classifier. Execution, time taken by the Apache Spark is faster than the other traditional frameworks. The execution time taken by Apache Spark is 0.7s and traditional framework took 1.2s for test a 20, 00,000 records.

Latha, Jeeva [5] employed ensemble methods (Boosting, Bagging, Stacking and Voting) along with classification techniques (NB, PART, MLP, BN, C4.5) The feature set is selected as FS1, FS2, FS3, FS4, FS5 and FS6 with 13 attributes. Therefore n value is 13. The combination of empty set, is represented as 2^n-1. The classification methods are combined with the voting mehtod, and gets majority of vote with the ensembling of accuracy 85.45% by algorithms (NB+PART+MLP+BN+C4.5.

Imran et.al. [6] proposed four data mining techniques for heart prediction. The proposed work used Linear SVM, RBF SVM, KNN and Naïve Bayes algorithm. Among the four the highest accuracy (87.114%) is produced by RBF SVM.

Alberto Palacios Pawlovsky [7] used KNN algorithm with three and five distances (Euclid, Manhattan, Chebyshev, Sorensen, Canberra and Mahalanobis). The each distance gives the average accuracy when the KNN used. Among the distance Mahalanobis give better performance with 84.5% of accuracy. In ensemble methods the weights are used in 3-distance and 5-distance. The best accuracy of raw data is found by fixing classification set size is 90% with k value =7 in raw data gives 72.1% and the best accuracy of normalized data is increased by fixed classification data 90% with k value=52 and accuracy is 83.4% at last the highest accuracy of 84.8% is achieved ty the standardized data at 90% of classification set with vw method and k value is 50.

Amin Ul Haq, Jian Ping Li [8] used seven popular machine learning algorithms, three feature selection (FS) algorithms (Relief, mRMR, and LASSO with k-fold cross-validation) and measured the seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, MCC, and execution time. Proposed SVM RBF and Logistic regression are observed as best ML algorithms and accuracy is recorded for c=10 and g=0.0001. They observed that (88% of accuracy for SVM RBF and 87% of accuracy for Logistic regression. For SVM RBF 96% of specificity, 75% of sensitivity, 88% of MCC and 89% AUC are observed respectively.

Tejali Mhatre et al. [10] proposed the evolutionary NN by genetic algorithm act as a weight optimization engine for backpropagation network. Genetic algorithm uses a direct metaphor of natural behavior, work with a population of individual strings. The fitness function also use for the best fit and worst fit individuals to selected and then duplicate the best fit with the worst fit. The genetic based neural network achieved 75% of training accuracy and 78.7% of testing classification accuracy.

Uma [11] run through linear classifier as a Naive Bayes+PSO an efficient evolutionary computation technique and Feature Selection. The predictive model with Naïve Bayes 79.12% of accuracy is recorded, by proposed model Naive Bayes+PSO 87.91% of accuracy is recorded by 100 iterations. Proposed approach (NB+PSO) is compared with NB+GA. Using GA, accuracy is recorded as 86.29%.

Li Yang, Jing Yan [12] Several methods were used to build prediction model including multivariate regression model, CART, Naïve Bayes, BT, Ada Boost and RF. The multivariate regression model used as a benchmark for performance and documented AUC = 0.7143. The results showed that the Naïve Bayes=0.7074, BT=0.7448, Ada Boost=0.7662, RF=0.7872 and CART=0.7025. From other methods RF was superior to other methods with an AUC of 0.787.

K. Mathan et al. [13] proposed a calculation for decision trees classification and applied with neural networks and Gini index prediction and recorded 87.89% of accuracy. For Gain ratio and Gini index an equal width, entropy and frequency are calculated by decision tree without voting and with voting. The Gain ratio and Neural Network used with the decision tree and recorded accuracy (85%, 87.89%), Specificity (89, 87%.65%) and Sensitivity (75%, 85.6%).

Yumna Farooq et.al. [14] propose a novel method that comprises machine learning algorithms for the early prediction of heart disease. Stratified K-fold cross validation is used with random forest and accuracy achieved of 86.94% which outperforms compare with Hoeffding tree method reported accuracy of 85.43%.

Table 2: Data Description (Feature information of the Cleveland dataset)

S.No	Attribute Name	Description	Range of Values
1	Age	Age of the person in years	29 to 79
2	Sex	Gender of the person [1: Male, 0: Female]	0, 1
3	cp	3 Cp Chest pain type [1-Typical Type 1 Angina 2- Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)	1, 2, 3, 4
4	Trestbps	Resting Blood Pressure in mm Hg	94 to 200
5	Chol	Serum cholesterol in mg/dl	5 126 to 564
6	Fbs	Fasting Blood Sugar in mg/dl	0, 1
7	Restecg	Resting Electrocardiographic Results	0, 1, 2
8	Thalach	Maximum Heart Rate Achieved	71 to 202
9	Exang	Exercise Induced Angina	0, 1
10	OldPeak	ST depression induced by exercise relative to rest	1 to 3
11	Slope	Slope of the Peak Exercise ST segment	1, 2, 3
12	Ca	Number of major vessels colored by fluoroscopy	0 to 3
13	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3, 6, 7
14	Num	Class Attribute	0 or 1

Xiaoming et.al. [15] combined fuzzy logic and Bootstrap Aggregating (Bagging) algorithm based on GBDT algorithm and proposed Fuzzy-Bagging-GBDT. DT algorithm is also used to compare for accuracy by including 14 parameters. The confusion matrix is calculated. The documented accuracy is 87%, sensitivity is 92%, and specificity is 86% and AUC 87%. The ROC curves are calculated for each method and fuzzy-bagging-GBDT record the highest accuracy.

Akansh et.al. [16] implement Naïve Bayes algorithm combine with other classification algorithm for highest accuracy and NB produced 84.43% of accuracy before optimization and 88.16% after optimization

A. Data Set

Most of the researchers used UCI machine learning repository to collect heart disease data. There are three databases. They are Cleveland, Hungary, and Switzerland. The most used database among the three is Cleveland database it contains 303 records and 76 attributes. The researchers mostly used 14 important parameters. Few important parameters are age, heart rate, cholesterol and blood pressure. The data set description is shown in Table 2.

IV. PERFORMANCE EVALUATION METRICS

The confusion matrix helps practitioners to form a clear idea of whether the results have a high performance. In order to check the performance of the classifiers, various performance evaluation metrics were used in this research. We used confusion matrix, every observation in the testing set is predicted in exactly one box. It is 2 × 2 matrix because there are 2 inactivity classes as Fig 3. The different performance metrics were calculated using a confusion matrix.

		Actual Outcome		
		Positive	Negative	
Prediction Model Outcome	Positive	True positive (TP)	False Positive (FP)	Positive Predictive Value (Precision)=TP/(TP+FP)
	Negative	False Negative (FN)	True Negative (TN)	Negative Prediction Value TN/(TN+FN)
		Sensitivity(Recall)= TP/(TP+FN)	Specificity= TN/(FP+FN)	Accuracy (TP+TN)/(TP+FP+TN+FN)

Figure 1. Confusion Matrix

1) *Classification Accuracy*: Accuracy shows the overall performance of the classification system and it is calculated using equation 1.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

2) *Classification Error*: it is the overall incorrect classification of the classification model which is calculated using equation 2

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \quad (2)$$

3) *Sensitivity*: it is the ratio of the recently classified heart patients to the total number of heart patients. Sensitivity of the classifier for detecting positive instances is known as "true positive rate." It is calculated using the equation 3

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

4) *Specificity*: is a diagnostic test of negative and the person healthy is mathematically calculated using the equation 4

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

5) *Precision*: the equation 5 is used to calculate the precision value for person health. $Pecision = \frac{TP}{TP+FP}$ (5)

6) *F1 Score*: The F1 score considered a harmonic average between precision in recall is defined by equation 6

$$F-measure = \frac{2\times Precision\times Recall}{Recall\times Precision} \quad (6)$$

7) *MCC:* Matthews's correlation coefficient (MCC) was introduced by Brian W. Matthews to predict the performance of protein secondary structure. The results of MCC are in percentage. Therefore, MCC becomes a widely used performance metric in medical research for imbalanced data expressed by equation 7 $MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ (7)

8) *Kappa:* Kappa measures the percentage of agreement between two raters. The formula to calculate Kappa is represented by K where *p*0 is the percent of agreement among raters, as in Eq. 8 and *pc* is the chance agreement.

$$K = p\frac{p_0-p_c}{1-p_c} \quad (8)$$

9) *ROC (Receiver Operator Characteristic):* It is a probability curve indicating the capability of a model between classes. The ROC curve shows trade-off between True Positive Rate (TPR) and the False Positive Rate (FPR). AUC (Area under the Curve) closer to 1 would be able to perfectly differentiate the two classes in the case of binary classification. Therefore, AUC closer to 1 is better predictive measure.

$$AU-ROC = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \quad (9)$$

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue X Oct 2020- Available at www.ijraset.com

V. RESULTS AND DISCUSSION

In this section the comparison results of various Machine Learning Algorithm for each techniques and ensemble methods of algorithms are furnished. Figure 2 provides the comparison result of various ML algorithm. Figure 3 represents the best accuracy of ensemble method by using the 14 attributes from UCI repository heart disease dataset of heart disease data given by various researcher is displayed.

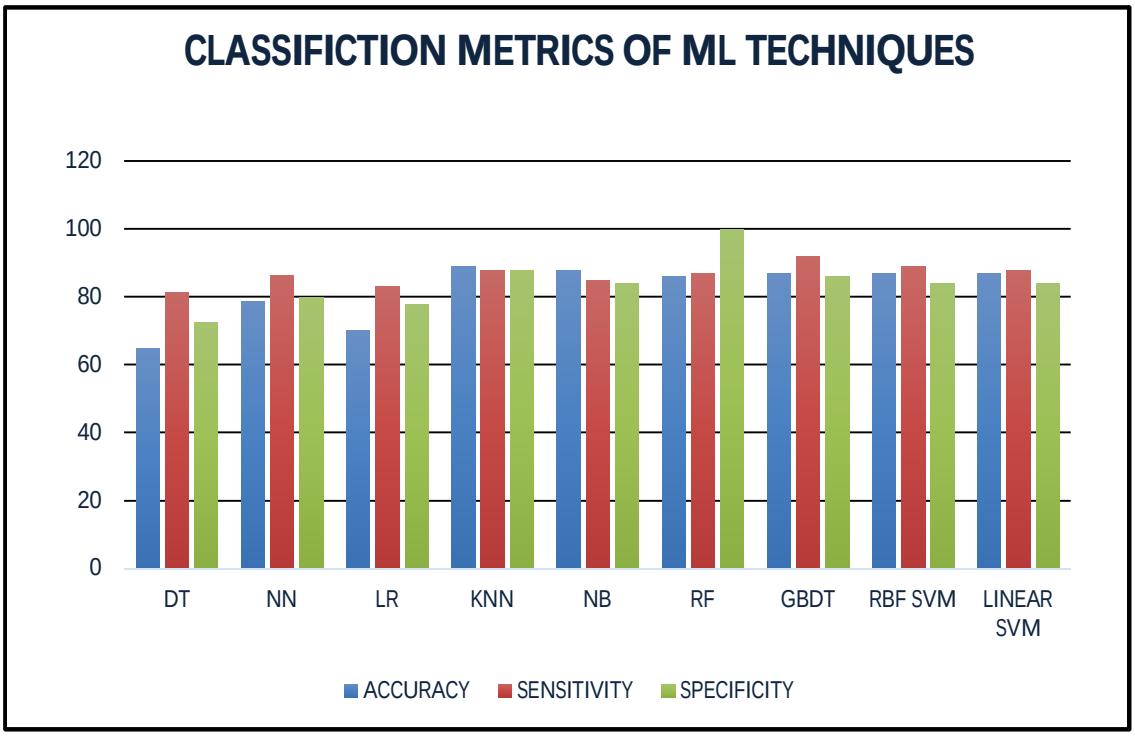


Figure 2 Classification metrics of ML algorithms

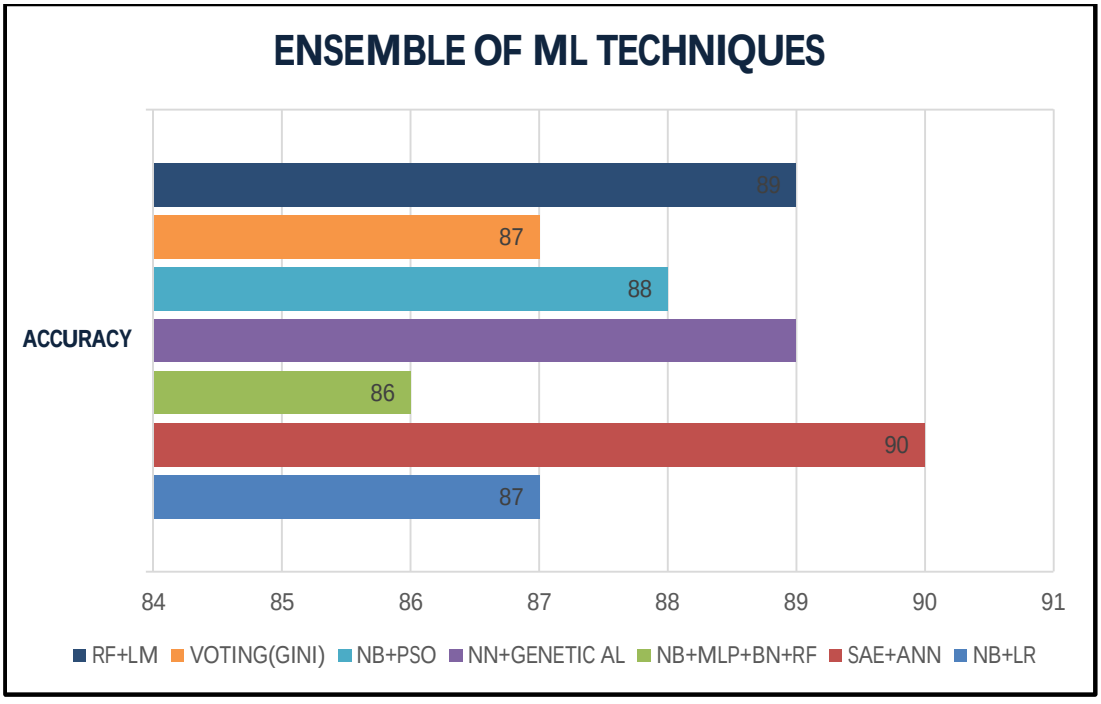


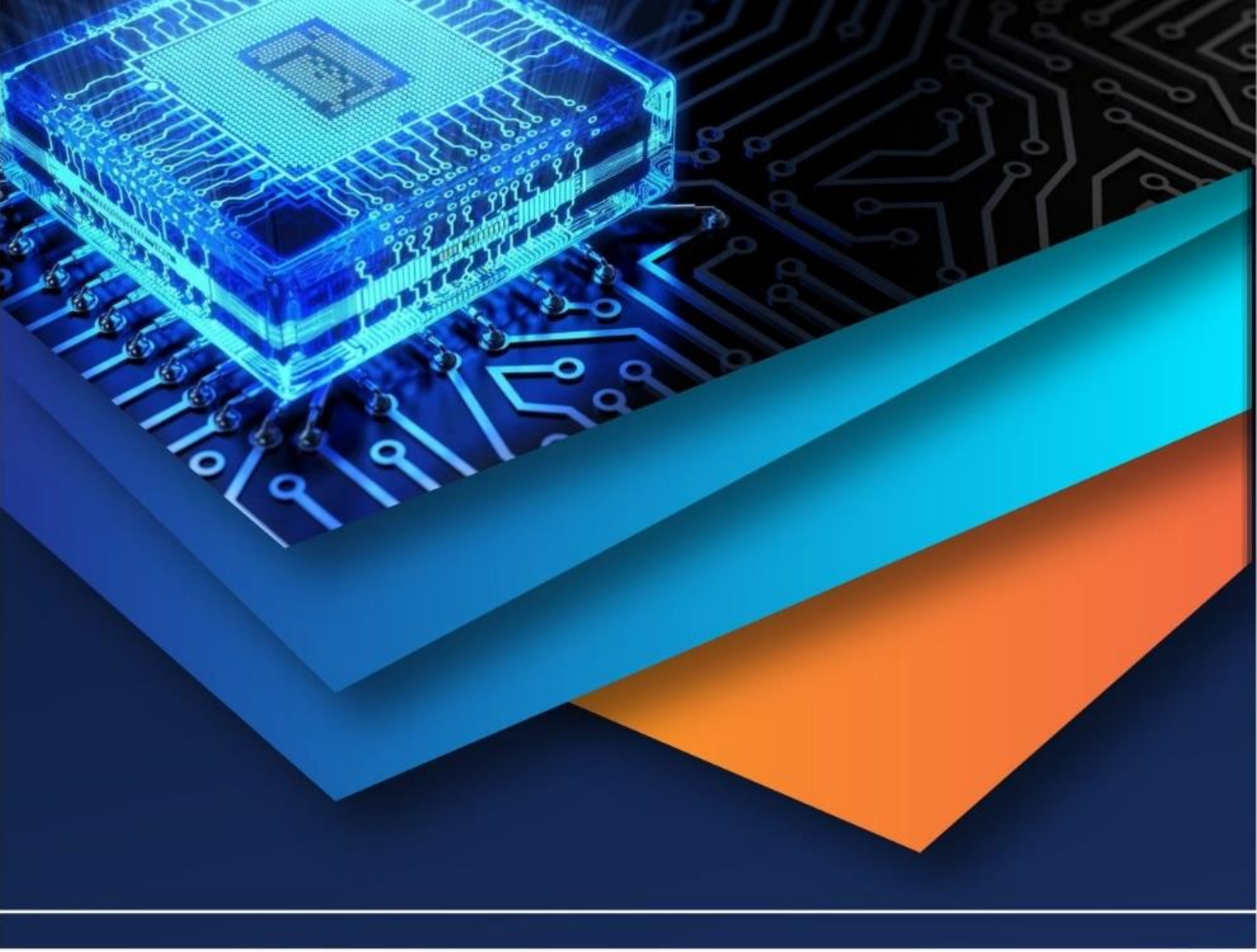
Figure 3. Accuracy of ensemble method of ML algorithms

VI. CONCLUSION

The medical industry is huge area that produce large amount of data every second and it enforce the use of big data. The inquiries made in the stored data can save many life when the prediction is done earlier so the researcher used various techniques to predict the data saved in healthcare database. Machine learning technique is one of the techniques used for prediction. This paper discussed different algorithms and the accuracy recorded by them. And many researchers used combination of algorithm and formulate ensemble method of algorithm and the best accuracy recorded is 90% by ANN+SAE and NN+ Genetic Algorithm and Random Forest + Linear Method produce 89% of accuracy set respectively.

REFERENCES

[1] Shafenoor Amin, M., Kia Chiam, Y., Dewi Varathan, Kasturi Dewi Varathan., "Identification of significant features and data mining techniques in predicting heart disease", S0736-5853(18)30887-6-Telematics and Informatics, November 2018.
[2] Damir Imamovic, Elmir Babovic, Nina Bijedic, "Prediction of mortality in patients with cardiovascular disease using data mining methods", 19th International Symposium INFOTEH-JAHORINA (IEEE), 20 March 2020.
[3] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang,Improved sparse autoencoder based artificial neural network approach for prediction of heart disease", Informatics in Medicine Unlocked 18-100307(ELSEVIER) ,2020
[4] Khalil Maalmi , Abderrahmane Ed-daoudy, " Real-time machine learning for early detection of heart disease using big data approach", 978-1-5386-7850-3/19/$31.00 (IEEE), 2019.
[5] C. Beulah Christalin Latha, S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked 18 - 100203(ELSEVIER), 2019
[6] Imran Mirza, Arnav Mahapatra, Daryl Rego, Kenneth Mascarenhas, "Human Heart Disease Prediction Using Data Mining Techniques", Department of Computer Engineering, Don Bosco Institute of Technology, Mumbai-70, India, 2018.
[7] Alberto Palacios Pawlovsky , "An Ensemble Based on Distances for a kNN Method for Heart Disease Diagnosis", Google scholar, 2018.
[8] Amin Ul Haq, Jian Ping Li Muhammad Hammad Memon, Shah Nazir, Ruinan Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms", Mobile Information Systems, Volume 2018, Article ID 3860146, (Hindawi),2018.
[9] Hamidreza Ashrafi Esfahani, Morteza Ghazanfari," Cardiovascular disease detection using a new ensemble classifier", IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2017.
[10] ejali Mhatre , Satishkumar Varma, " Heart Disease Prediction using Evolutionary based Artificial Neural Network", International Journal of Engineering Research & Technology (IJERT) - ISSN: 2278-0181, 2019. .
[11] Uma N Dulhare, Uma, "Prediction system for heart disease using Naive Bayes and particle swarm optimization", **DOI:** 10.4066/biomedicalresearch.29-18-620,Biomedical Research, May 21, 2018.
[12] Li Yang, Haibin Wu, Xiaoqing Jin, Pinpin Zheng, Shiyun Hu, Xiaoling Xu, Wei Yu & Jing Yan, " Study of cardiovascular disease prediction model based on random forest in eastern China", Scientific Reports 10:5245, 2020.
[13] K. Mathan, Priyan Malarvizhi Kumar, Parthasarathy Panchatcharam, Gunasekaran Manogaran, R.Varadharajan, "A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease", Springer Science + Business Media, LLC, part of Springer Nature, 2 April 2018.
[14] Yumna Farooq, Muhammad Affan Alim, Shamsheela, Habib, Abdul Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model", 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) 978-1-7281-4970-7/20 (IEEE), 2020.
[15] Xiaoming Yuan1, Xue Wang, Jianchao Han, Jiemin Liu, Haiyan Chen, Kuan Zhang, and Qiang Ye , "A High Accuracy Integrated Bagging-Fuzzy-GBDT Prediction Algorithm for Heart Disease Diagnosis", IEEE/CIC International Conference on Communications in China (ICCC),2019.
[16] Akansh Gupta, Lokesh Kumar, Rachna Jain and Preeti Nagrath," Heart Disease Prediction Using Classification (Naive Bayes)", Proceedings of First International Conference on Computing, Communications, and Cyber-Security, Springer Nature Singapore Pte Ltd, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY