Journal of Online Engineering Education ISSN: 2158-9658 Volume: 13 Issue: 2

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

Analysis of Machine Learning Models for Heart Disease Prediction using Different Algorithms

Neeraj[1], Jagtar Singh[2]

M. Tech Scholar[1], Assistant Professor[2] – NCCE, Israna, Department of ECE, Panipat, Haryana, India

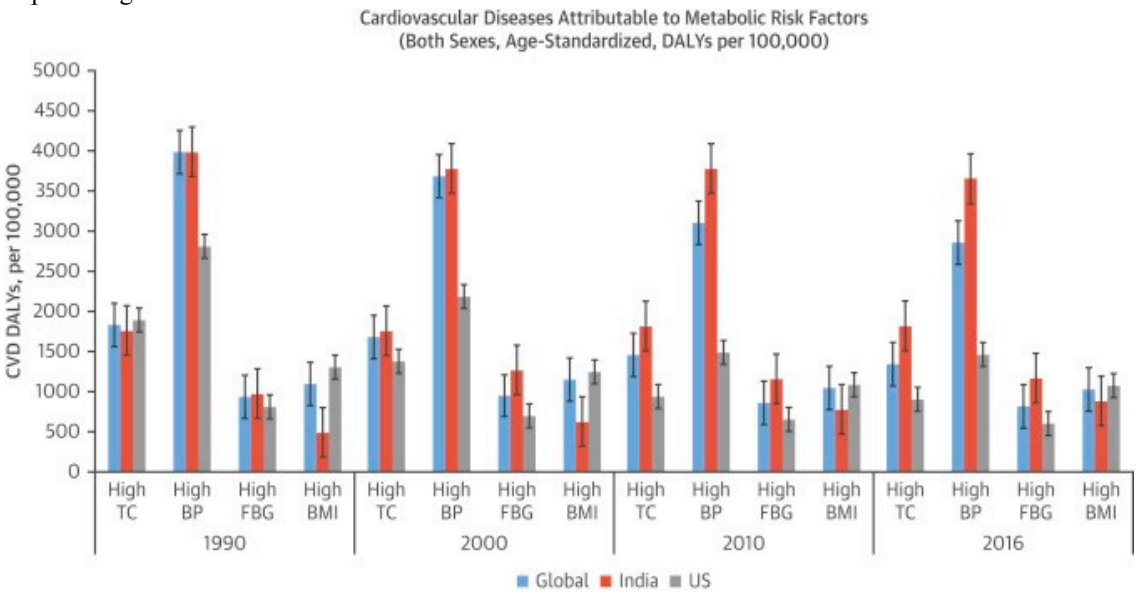
neeruahlawat4753@gmail.com[1], jagtar_nit@rediffmail.com[2]

Abstract: This research work seeks to explain the development of existing research on utilizing computational intelligence techniques in heart diseases diagnosis. This disorder extremely malignant ailment, over 1.7 billion demise all over the world. Diagnosis and treatment of heart diseases at early stage is the only solution otherwise it leads to fatality rate. With pace of time, new technologies emerging such as AI &ML, IoT and due to these advancement in science especially in healthcare, various types of severe disease can be diagnosed at early stage. The main objective of this work is to design machine learning models to predict heart disease with better accuracy. In our implemented work five different supervise ML (Machine Learning) algorithms are instigated which are Logistic Regression, KNN, SVM, Decision Tree and Bagging Classifier. Out of listed algorithm, SVM perform better and give the accuracy 93.40% and KNN gives the least accuracy 71.42%. Accuracy in machine learning models should not be so high otherwise it will be fall under over fitting. Machine learning models having accuracy more than 90 % is measured upright. One important thing is that accuracy should not be so high otherwise it may be possible that designed model is overfit for a specific dataset. Besides accuracy in this research article two parameters also calculated which are precision and recall from confusion matrix. Support vector machine algorithm gives precision value 88 and recall value 91.67.

Keywords: Bagging Classifier, Decision Tree, SVM, Machine Learning, Heart Disease, KNN, Logistic Regression

1. Introduction

In modern medical circumstances heart diseases is one of the prime concerns. Each year huge number of people lost their lives due to heart attack or cardiac arrest. There are various reasons of heart attack but today's life style which incorporate un-healthy diets and stress due to various reasons such as office work, business loss. Heart attack detection at last stage is one of the foremost causes of mortality. Mortality rate due to hear attack increasing exponentially each year. Mortality rate due to heart attack is more than 19 million per year around the globe. Developed countries somehow able to control in small context due to their good healthcare infrastructure. But developing countries like India with huge population unable to provide good treatment to its citizens.



CC BY NC SA

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

Fig.1. Cardiovascular disease in India compared with United States [2]

Real-time implementation results and observations can be seen using various ML algorithms. There is various classification technique with the utilization of naive bayes, Laplace smoothing techniques for heart disease prediction. Medical practitioners use diagnostic tests to reduce uncertainty about the presence of heart disease. These tests are usually expressed with various statistical measures.

Risk Factor of Heart Disease: There are numerous researchers who explain the factors involved in cardiac disease. Numerous considerations are intensifying the endanger of acquest cardiac disease. Figure 1.9 shows the controllable endanger components of cardiac discomfort. The components are recorded as cholesterol levels, gender, diabetes, smoking, drinking etc.

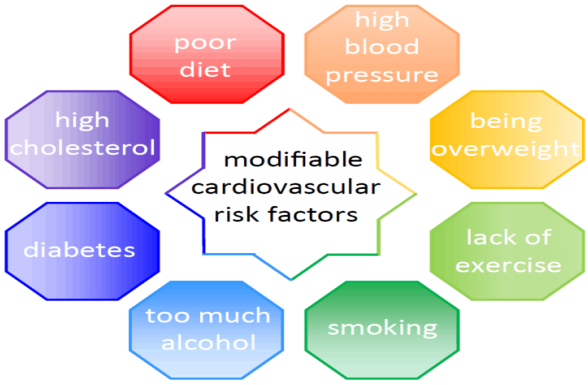


Fig.2. Risk Factors of Heart Disease [4]

The heart diseases classification mainly concentrates on various medical breakdowns of blood vessels, pathologies metrics such as white blood cell dysfunction, blood vessels platelets count and full of fat substance presents in the blood vessel which disturbs flow of blood to the heart are prominent cause of heart disease.

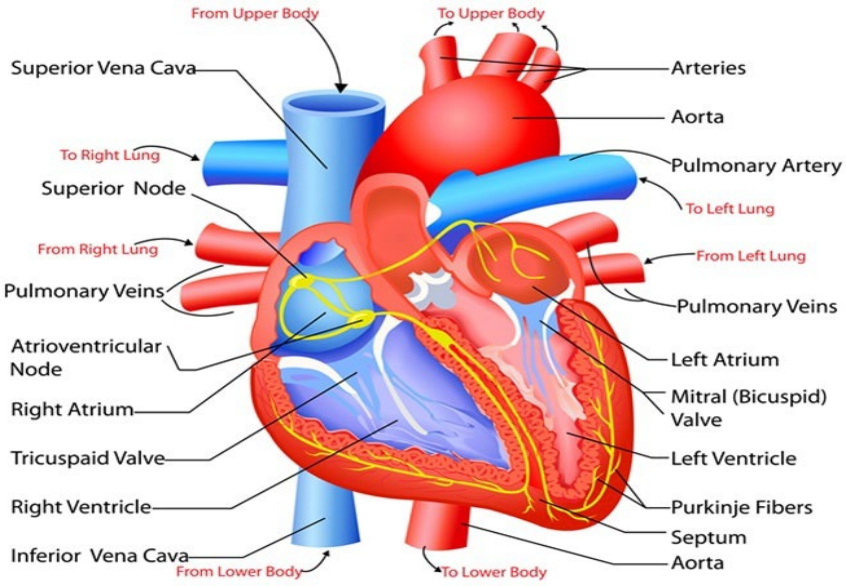


Fig.3. Anatomic Structure of Heart [5]

Figure 3 shows the anatomic structure of heart clinically known as cardiac which has derived from the Latin word. This organ made of soft tissue muscle contrived into 4 compartments unglued by blood vessels separated as pair of divisions. Respective divisions are called as artium and ventricle. Atriam accumulate blood flock together, after that ventricles pump play crucial role through which blood circulated in the body. When blood carrying oxygen then it is known as oxygenated blood which gives energy to the body for various function.

CC BY NC SA

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

2. Methodology

The health informatics combines mathematical models, algorithm and analysis to provide improved quality of healthcare services to the electronic health users. Across, the branch of health informatics machine learning plays a crucial role in health data analysis and management processes. This work mainly focuses on classification based on heart disease predictions. In this research work, standard dataset is used which is downloaded from Kaggle website having various parameters related to human or you can use your own as per availability. After that data is pre-processed so that if there is any ambiguity in data set can be resolved otherwise it will affect accuracy of algorithms. Data set is divided into training (70%) as well as testing (30%) data set. With this training data, machine learning models are prepared to predict the heart disease accuracy. Thereafter, defined an optimal approach to predict heart diseases using supervised ML algorithms: KNN, Logistic Regression, Decision Tree and SVM.

KNN: The simplest supervised machine learning algorithm is K-Nearest Neighbour. This algorithm based upon resemblance between novel data and existing data and put new data into appropriate group which is greatest like to existing classes. K-NN algorithm stores each and every existing data and categorizes a new data point based on similarity index. This algorithm can be implemented for both that is classification and regression. But in practice it is more implemented for sorting purpose. KNN algorithm is unable to generalize the training data in advance therefore for each prediction it scans the data then predict. This is the main reason that KNN is one of the slow algorithms in machine learning.

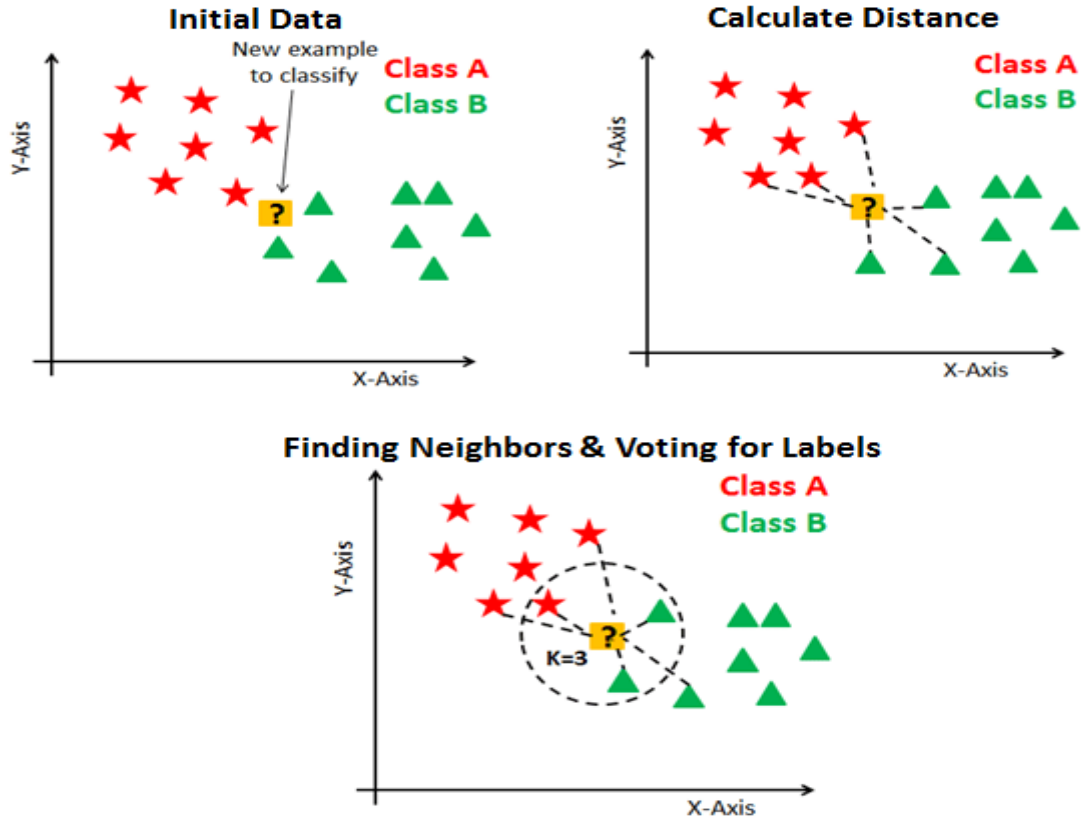


Fig.4. KNN Machine Learning Algorithm [7]

Logistic Regression: It is an arithmetical method of data examination applied across binary dependent variables. Logistic model parameter is estimated using the logistic regression techniques. In technical terms, in a logistic model likelihood of an incident is a direct amalgamation of independent variables. In wide-ranging, it is not a cataloguing system, but it just models the output probability through the given inputs. It is often used as a classifier by fixing the cut-off values. Thus, the values below threshold values fit to one class, and the variables above the cut of values belong to the other class. Multinomial logistic regression and ordinal (LR) logistic regression are the two major types of logistic regression. Multinomial logistic regression deals with absolute values, and it groups the output values in more than two categories. The process of ordering the multiple outputs produced by multinomial regression model is called the ordinal logistic regression

BY NC SA

Journal of Online Engineering Education ISSN: 2158-9658 Volume: 13 Issue: 2

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

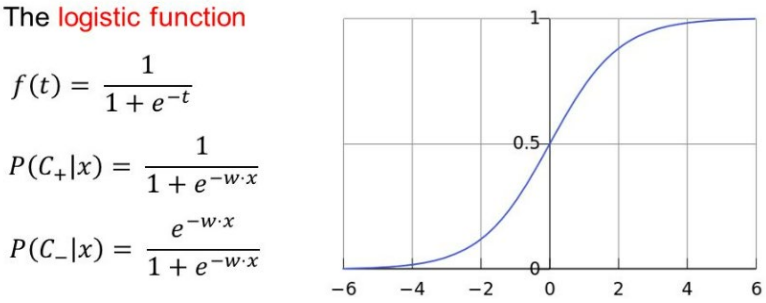


Fig.5. Logistic Regression Model

Support Vector Machine: There exist various machine learning algorithms to analyse the data for regression as well as classification purpose and SVM is one of them. Non-linear SVM approaches are the maximum widely implemented algorithm to deal with unlabelled data and used across several industrial applications. For any given set of data with labelled training samples, it outputs an optimal hyperplane. Which, further classifies novel illustrations of the input data model. The hyperplane is a line that segregates the given hyperplane into two parts in a 2D space. Each class resides at either side of the partitions [12]. Our examination work dependent on Support Vector Machines (SVMs). This calculation utilizes a SVM to perceive features. The calculation begins from an assortment of tests of features from data set. Support Vector Network is an alternate name of support vector machine.

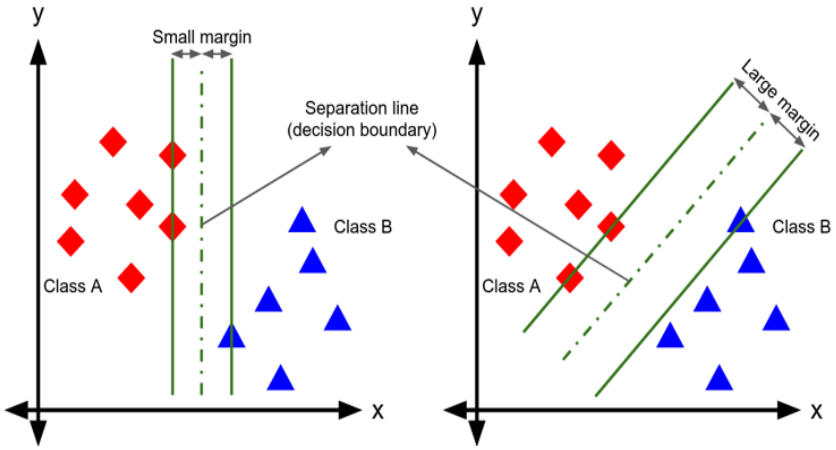


Fig.6. Functioning of SVM Algorithm

Decision Tree: Decision Tree is a supervised machine learning algorithm. This algorithm can be implemented for both types of problem classification as well as regression but in most of the cases best appropriate for classification problems. This algorithm generally implemented in healthcare domain. In this algorithm following are the key word like leaf node, root node and branch. Each decision tree predicts the class and by considering different parameters various models are created and in last using voting classifier a final model is designed which has optimum accuracy. As we know ensemble learning method performs regression and classification processes. In training stage, it constructs a multitude of the decision tree and predicts the outcomes of the individual trees using the regression methods. It has reduced variance and easily correlates the multiple features of provided data for forecast determinations. It is a supervised learning approach partitions the given dataset into training and the test data.

CC BY NC SA

Journal of Online Engineering Education ISSN: 2158-9658 Volume: 13 Issue: 2

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

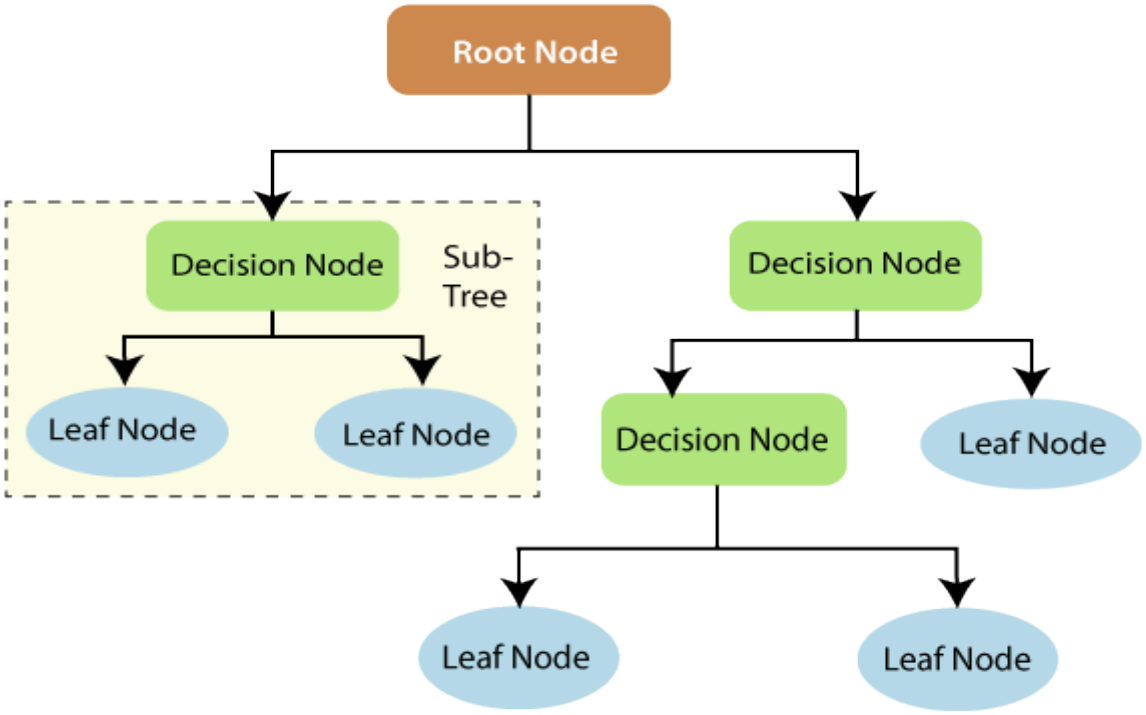


Fig.7. Decision Tree algorithm flowchart [10]

3. Result and Discussion

In the medical domain there exist various tedious tasks which are challenging like to estimate the severe diseases and proper control over it. In this implemented work, several supervised ML algorithms are implemented on standard dataset having 14 parameters related to human. After that data is pre-processed so that if there is any ambiguity in data set can be resolved otherwise it will affect accuracy of algorithms. Data set is divided into training (70%) as well as testing (30%) data set. Data set can be divided in any ratio but in the end, goal is to achieve optimized model where it can give maximum accuracy. Thereafter, defined an optimal approach to predict heart diseases using supervised ML algorithms:

- Logistic Regression
- Bagging Classifier
- KNN
- SVM
- Decision Tree

Implemented work executed using Jypyter Notebook/Google colab using python language. In this section results of all implemented supervised machine algorithms are depicted:

Table 1: Bagging Classifier Accuracy Analysis

	Sample Size	n_Estimator	Accuracy
Bagging Classifier Accuracy	**60**	**50**	**82.89**
	70	**100**	**85.52**
	80	**150**	**84.29**

CC BY NC SA

Journal of Online Engineering Education ISSN: 2158-9658 Volume: 13 Issue: 2

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

Fig.8. Accuracy of implemented algorithm for training data set

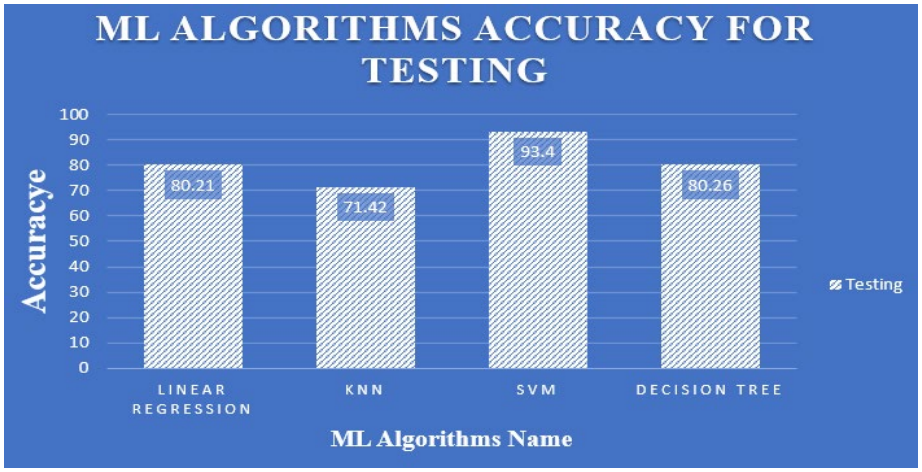


Fig.9. Accuracy of implemented algorithm for testing data set

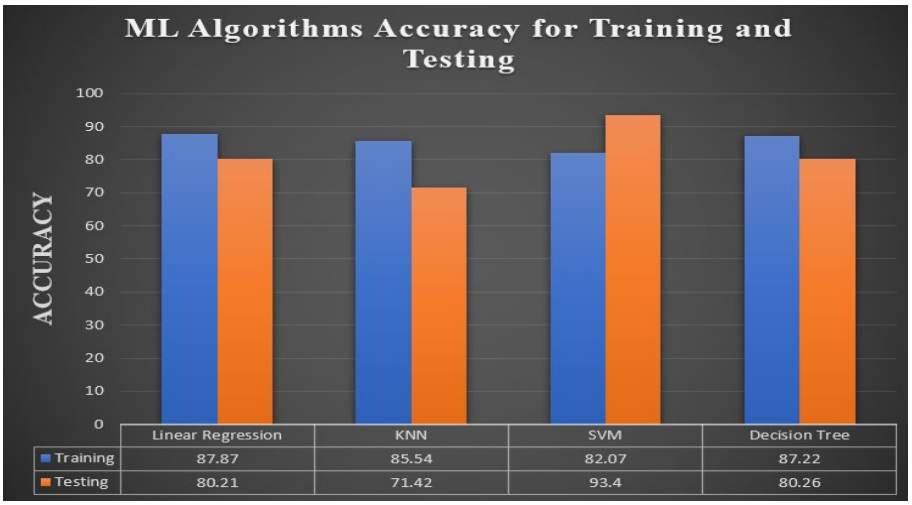


Fig.10. Accuracy result for implemented algorithm for training data set and testing data set

Table 2 Comparative analysis of different algorithm

Accuracy	Linear Regression	KNN	SVM	Decision Tree
Training	**87.87**	**85.54**	**82.07**	**87.22**
Testing	**80.21**	**71.42**	**93.40**	**80.26**

CC BY NC SA

Journal of Online Engineering Education ISSN: 2158-9658 Volume: 13 Issue: 2

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

Confusion Matrix

Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Precision:

$$Precision = \frac{TP}{TP + FP}$$

……Eq. No. (1)

Recall:

$$Recall = \frac{TP}{TP + FN}$$

……Eq. No. (2)

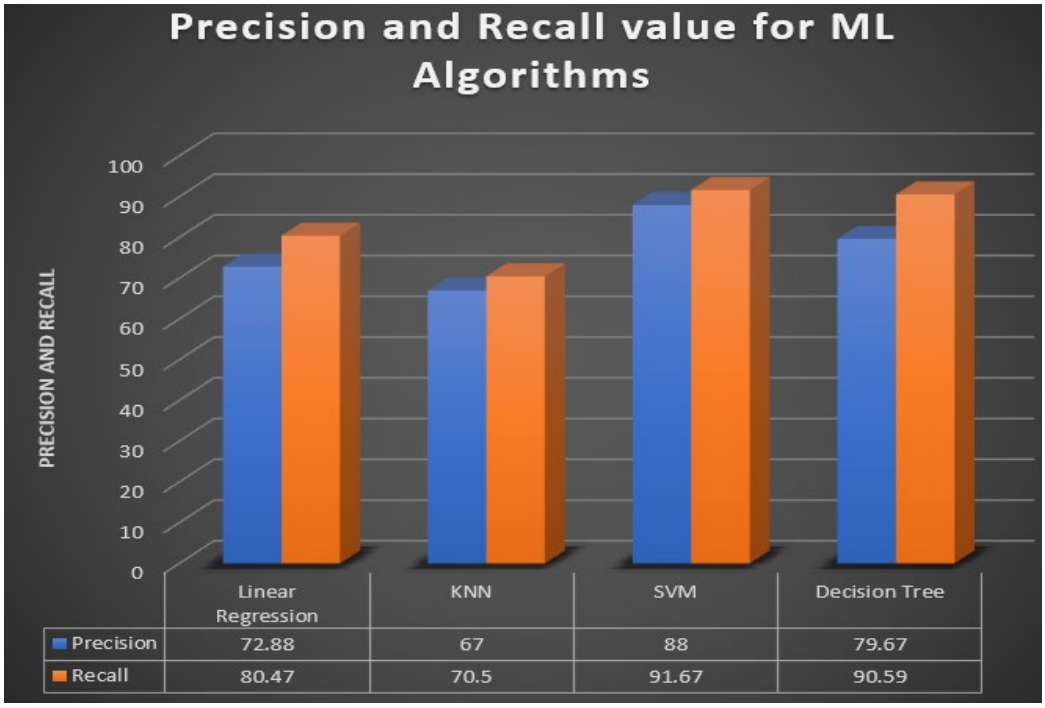


Fig.11. Precision and Recall value for ML Algorithms

Table 3: Comparative analysis of different algorithm for Precision and Recall

Parameters	Linear Regression	KNN	SVM	Decision Tree
Precision	**72.88**	**67**	**88**	**79.67**

CC BY NC SA

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

Recall	**80.47**	**70.5**	**91.67**	**90.59**

4. CONCLUSION

The recent advancement in medical technology, higher computational techniques, reduced cost of storage techniques and internet connectivity enables the digitalization of diagnostic systems in the present world. In these days machine learning playing a vital role in healthcare system to diagnose the various severe diseases precisely. In this research work, standard dataset is used which is downloaded from Kaggle website having various parameters related to human. After that data is pre-processed so that if there is any ambiguity in data set can be resolved otherwise it will affect accuracy of algorithms. Data set is divided into training as well as testing data set in desired ratio so that overall accuracy can be optimized. Thereafter, defined an optimal approach to predict heart diseases using supervised ML algorithms: Logistic Regression, KNN, SVM, Decision Tree and Bagging Classifier. Out of listed algorithm, SVM perform better and give the accuracy 93.40% and KNN gives the least accuracy 71.42%. One important thing is that accuracy should not be so high otherwise it may be possible that designed model is overfit for a specific dataset. Accuracy more than 80 % is measured good, and accuracy around 90% is admirable. Besides accuracy in this research article two parameters also calculated which are precision and recall. Support vector machine algorithm gives precision value 88 and recall value 91.67. Digitalization of medical data has generated a new era towards the diagnostic field. With the massive growth of digital information, these unprocessed patients' medical information is extremely essential to analyse, explore and utilize with various classification techniques.

REFERENCES

[1]. H. Jindal, S. Agrawal, R. Khera, R. Jain and P. Nagrath, "Heart disease prediction using machine learning algorithms", *ICCRDA 2020, IOP Conf. Series: Materials Science and Engineering*, 1022 (2021) 012072, DOI:10.1088/1757-899X/1022/1/012072.

[2]. P. Motarwar, A. Duraphe, G. Suganya, M. Premalatha, "Cognitive Approach for Heart Disease Prediction using Machine Learning", *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE),* IEEE, 2020, DOI: 10.1109/ic-ETITE47903.2020.242.

[3]. V. Sharma, S. Yadav, M. Gupta, "Heart Disease Prediction using Machine Learning Techniques", *2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, IEEE, 18-19 Dec. 2020, DOI: 10.1109/ICACCCN51052.2020.9362842.

[4]. A. Nikam, S. Bhandari, A. Mhaske, S. Mantri, "Cardiovascular Disease Prediction Using Machine Learning Models" *IEEE Pune Section International Conference (PuneCon)*, IEEE, 16-18 Dec. 2020, DOI: 10.1109/PuneCon50868.2020.9362367.

[5]. S. Mohan, C. Thirumalai, G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", *IEEE Access, Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things*, 2019, DOI: 10.1109/ACCESS.2019.2923707.

[6]. D. Kumar, S. Kumar, K. Arumugaraj, V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms", *IEEE International Conference on Current Trends toward Converging Technologies*, Coimbatore, 2018.

[7]. A. Gavhane, G. Kokkula, I. Pandya, K. Devadkar, "Prediction of Heart Disease Using Machine Learning", *Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology,* IEEE Conference, 2018.

[8]. Zhang, Shuai, Y.-L. S. A., "Deep learning-based recommender system: a survey and new perspectives", *Journal of ACM Computing Surveys* 1(1), 1–35, 2017.

[9]. A. Khatami, A. Khosravi, "Medical image analysis using wavelet transform and deep belief networks", *Journal of Expert Systems with Applications,* 3(4), 190–198, 2017.

[10]. A. Shetty, C. Naik, "Different data mining approaches for predicting heart disease", *International journal of innovative research in science, engineering and technology* 3(2), 277–281, 2016.

[11]. Aydin, S, "Comparison and evaluation data mining techniques in the diagnosis of heart disease", *Indian journal of science and technology*, 6(1), 420–423, 2016.

[12]. N. Bayasi, T. Tekeste, "Low-power ECG-based processor for predicting ventricular arrhythmia", *Journal of IEEE transactions on very large-scale integration systems,* 24(5), 1962–1974, 2016.

[13]. GB. Berikol, O. Yildiz, "Diagnosis of acute coronary syndrome with a support vector machine", *Journal of Medical System,* 40(4), 11–18, 2016.

CC BY NC SA

Article Received: 18 October 2022 **Revised**: 14 November 2022 **Accepted**: 24 November 2022

[14]. Z. Wang, X. Liu, "Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach", 6(30), 435–439, 2016.

[15]. M. Singh and Martins, "Building a cardiovascular disease predictive model using structural equation model and fuzzy cognitive map", *Journal of fuzzy system,* 02(6), 1377–1382, 2016.

[16]. S.Prabhavathi, "Analysis and prediction of various heart diseases using DNFS techniques", *International journal of innovations in scientific and engineering research,* 2(7), 678–684, 2016.

[17]. R. Sali, M. Shavandi, "A clinical decision support system based on support vector machine and binary particle swarm optimisation for Cardiovascular disease diagnosis", *International Journal of Data mining and Bio-informatics,* 15(1), 312–327, 2016.

[18]. P. Ghadge, K. Prajakta, "Intelligent heart attack prediction system using big data", *International journal of recent research in mathematics computer science and information technology,* 2(2), 73–77, 2016.

[19]. G. Purusothaman, K. Krishnakumari, "A survey of data mining techniques on risk prediction: heart disease", *Indian journal of science and technology,* 8(5), 643–651, 2015.

[20]. A. Richter, J. Listing, M. Schneider, T. Klopsch, A. Kapelle, J. Kaufmann, A. Zink, A. Strangfeld, "Impact of treatment with biologic dmards on the risk of sepsis or mortality after serious infection in patients with rheumatoid arthritis", *National Center for Biotechnology Information,* pp. 147–153, 2015.

[21]. S.H. Mujawar, P.R. Devale, "Prediction of Heart Disease using Modified K-means and by using Naive Bayes", *International Journal of Innovative Research in Computer and Communication Engineering,* 3(10), Oct 2015.

[22]. M. H. Vafaie, M. Ataei, "Heart diseases prediction based on ECG signals classification using a genetic-fuzzy system", *Journal of biomedical signal processing and control,* 14(5), 291–296, 2014.

[23]. J. Wang, W. Wang, "Study on qi-deficiency syndrome identification modes of Coronary heart disease based on metabolomic biomarkers", *Journal of evidence-based complementary and alternative medicine,* 24(16), 192–198, 2014.

[24]. X. Yang, M. Li, Y. Zhang, J. Ning, "Cost-sensitive naive bayes classification of uncertain data", *Journal of Scientific World,* 9(8), 1897–1904, 2014.

[25]. D. Chandna, "Diagnosis of heart disease using data mining algorithm", *International journal of computer science and information technologies,* 5(2), 1678–1680, 2014.

CC BY NC SA