Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Comparison of Machine Learning Models to Predict Heart Attack

Vivek Rathee[1], Renu Vadhera[2], Parveen Kumari[3]
M. Tech Scholar[1] – DPGITM Engineering College, Department of CSE & IT, Gurugram, Haryana, India
Assistant Professor[2,3]– DPGITM Engineering College, Department of CSE & IT, Gurugram, Haryana, India
vivekrathee@gmail.com[1], renu.csed@dpgitm.ac.in[2], parveen.csed@dpgitm.ac.in[3]

Article Info
Page Number: 1255-1264
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 February 2023
Revised: 24 March 2023
Accepted: 18 April 2023

Abstract
As According to a report from the World Health Organization (WHO), the number of cases of cardiac arrest and heart attacks is increasing exponentially each year. The mortality rate due to heart attacks exceeds 18 million people per year worldwide. The main goal of this work is to develop machine learning models that can predict heart disease with greater accuracy. In this work, there are total five different supervised machine learning algorithms implemented, namely Random Forest (RF), SVM, Logistic Regression, Decision Tree, and KNN, in order to achieve this objective. Among the algorithms we have implemented, Random Forest performed the best, achieving an accuracy of 90.16%, followed by Logistic Regression with an accuracy of 85.25%. The KNN algorithm yielded the lowest accuracy at 67.21%. In addition to accuracy, we also calculated two important parameters, precision and recall, from the confusion matrix. It is important to note that accuracy in machine learning models should not be too high, as this may indicate overfitting. Machine learning models with an accuracy of more than 90% are considered to be reliable.
Keywords: Random Forest, Decision Tree, SVM, Machine Learning, Heart Disease, KNN, Logistic Regression

1. Introduction

Heart diseases have become a major concern in modern medical circumstances, with a significant number of people losing their lives each year due to heart attack or cardiac arrest. There are multiple reasons behind heart attacks, but the unhealthy diets and stress associated with today's lifestyle, such as work pressure and business losses, are the primary contributors. The late detection of heart attacks is a major cause of mortality. The mortality rate due to heart attacks is increasing exponentially every year [1-2]. Several classification techniques, including the utilization of Naive Bayes and Laplace smoothing techniques, are available for heart disease prediction. Real-time implementation results and observations can be obtained using various machine learning algorithms [3]. Medical practitioners use diagnostic tests to reduce uncertainty about the presence of heart disease, and these tests are usually expressed using various statistical measures. This organ is composed of soft tissue muscle and is divided into four compartments separated by blood vessels and grouped into pairs of divisions called atrium and ventricle.

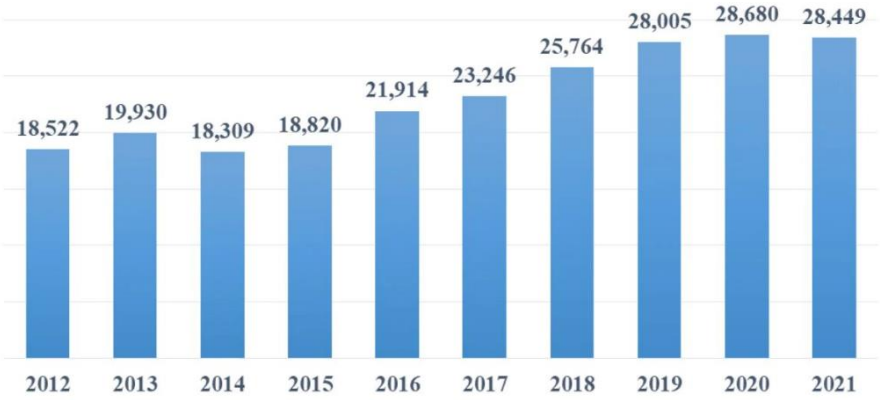

Fig.1. Death due to heart attack in India [in 2022, 70 per cent of heart attack deaths occurred in the 30-60 age group]

Atriums accumulate blood, and the ventricles play a crucial role in pumping the blood, which is then circulated throughout the body [5]. When the blood carries oxygen, it is known as oxygenated blood, which provides the body with the energy required for various functions. Figure 2 depicts the anatomic structure of the heart, which is clinically known as the cardiac organ, derived from the Latin word.

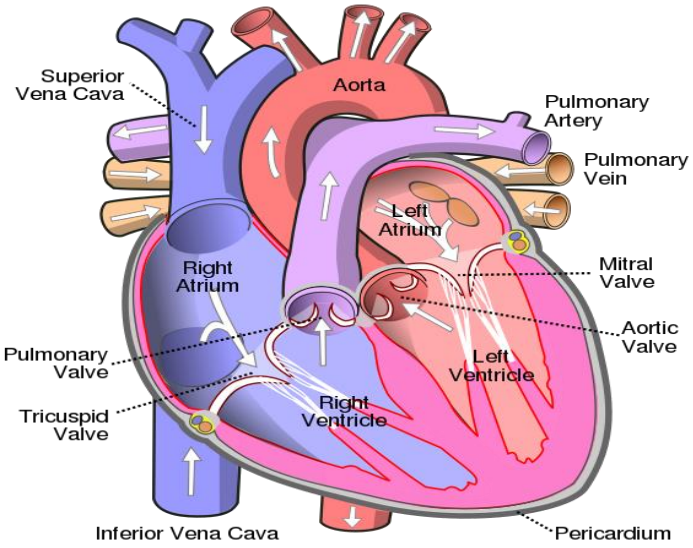

Fig.2. Anatomic Structure of Heart

The classification of heart diseases primarily focuses on different medical breakdowns of blood vessels, including pathologies such as white blood cell dysfunction, blood vessel platelet counts, and the presence of excessive fat substances in the blood vessels that disrupt the blood flow to the heart. These factors are prominent causes of heart disease [6].

2. Methodology

This research utilizes a standard dataset that was obtained from the Kaggle website. The dataset contains various parameters related to humans, or alternatively, a dataset may be used if

available. Prior to analysis, the dataset is pre-processed to resolve any ambiguities which may impact the accuracy of the algorithms. The pre-processed dataset is then divided into two parts: training (70%) and testing (30%). Machine learning models are then trained using the training dataset to predict the accuracy of heart disease. Subsequently, an optimal approach using supervised ML algorithms - Logistic Regression, KNN, SVM, Decision Tree, and Random Forest - is defined to predict heart diseases.

Logistic Regression: Logistic regression is a statistical method of data analysis applied to binary dependent variables. The logistic model parameter is estimated using logistic regression techniques. In technical terms, the probability of an event in a logistic model is a linear combination of independent variables. Although it is not generally a classification model, it models the output probability based on the given inputs. It is often used as a classifier by setting cut-off values. Values below the cut-off values belong to one class, while variables above the cut-off values belong to the other class. The two major types of logistic regression are Multinomial logistic regression (MLR) and Ordinal logistic regression (OLR). Multinomial logistic regression deals with absolute values, and it groups the output values into more than two categories. The process of ordering the multiple outputs produced by the multinomial regression model is called ordinal logistic regression. The likelihood of an event in a logistic model (LM) is a linear combination of independent variables using logistic regression techniques.

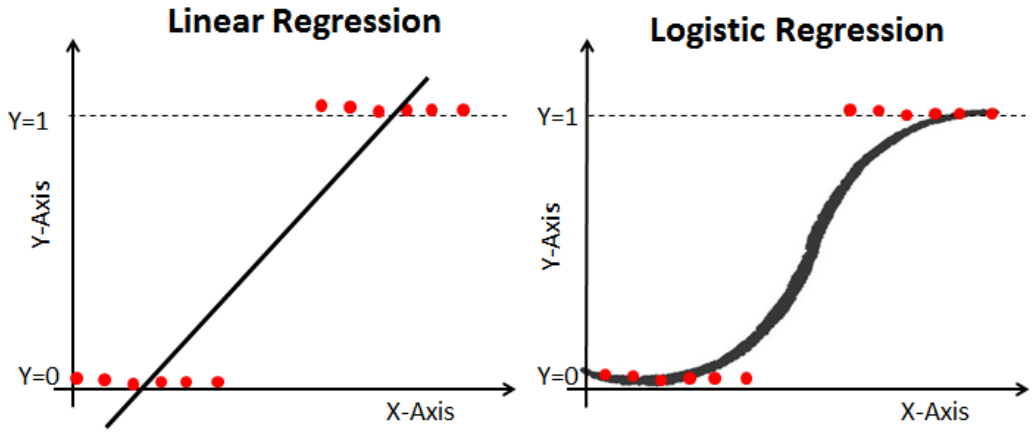

Fig.3. Linear Regression v/s Logistic Regression

The sigmoid function is referred to as an activation function for logistic regression and is defined as:

$f(x) = 1/1+e^{-x}$ …… (1)

The following equation represents logistic regression:

$y = e^{(b_0 + b_1 x)} / 1 + e^{(b_0 + b_1 x)}$ …… (2)

x = input value

y = predicted output

b0 = bias or intercept term

b1 = coefficient for input (x)

Decision Tree: The algorithm described here is typically implemented in the healthcare industry. The algorithm involves the use of keywords such as leaf node, root node, and branch. Each decision tree predicts a class, and different models are created by considering various parameters. In the end, a final model with optimal accuracy is designed using a voting classifier. The ensemble learning method is used for both regression and classification processes. During the training stage, multiple decision trees are constructed, and the outcomes of the individual trees are predicted using regression methods. This approach has the advantage of reduced variance and is capable of effectively correlating multiple features of the provided data for forecasting purposes.

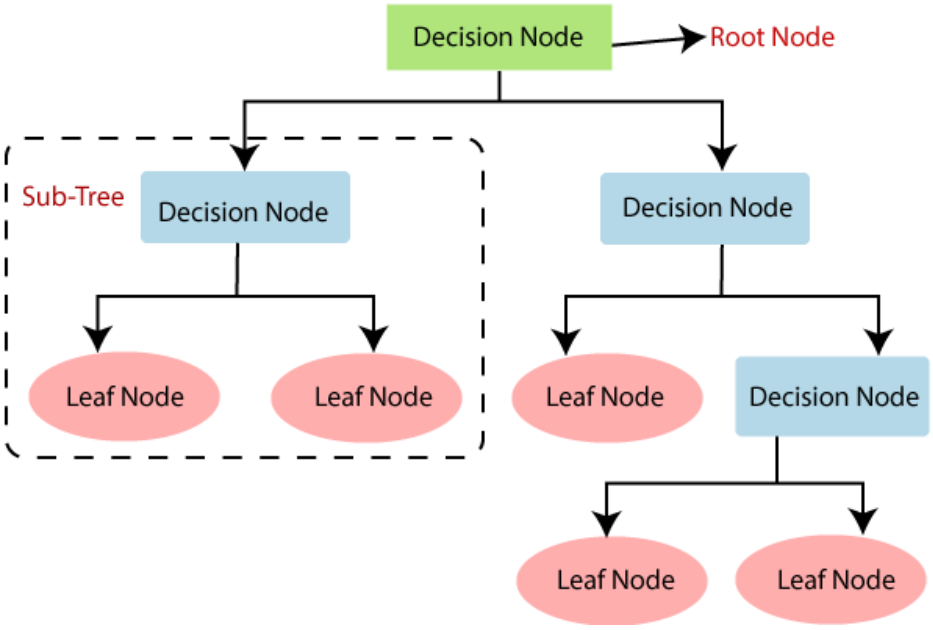

Fig.4. Decision Tree algorithm flowchart

KNN: One of the most basic supervised machine learning algorithms is the K-Nearest Neighbor (K-NN) algorithm. This algorithm relies on the similarity between novel data and existing data and assigns new data to the appropriate group based on its closest resemblance to existing classes. The K-NN algorithm stores all existing data and classifies a new data point based on a similarity index. This algorithm can be implemented for both classification and regression tasks, but in practice, it is more commonly used for classification. One of the disadvantages of the KNN algorithm is that it is slow and unable to learn from the provided dataset.

Distance Function:

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

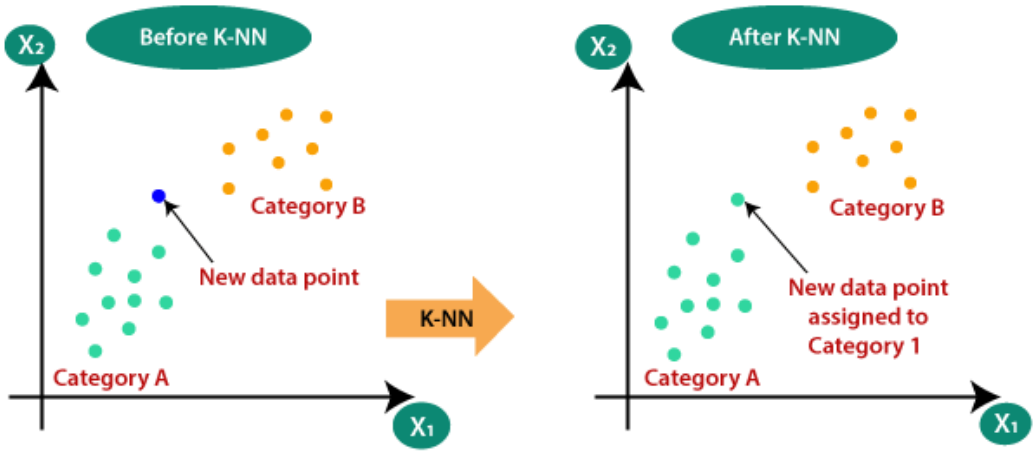

Fig. 5. KNN Machine Learning Algorithm

Random Forest: The random forest algorithm is composed of a large number of decision tree algorithms. Each decision tree predicts a class, and different models are created by considering various parameters. In the end, a final model with optimal accuracy is designed using a voting classifier. The ensemble learning method is used for both regression and classification processes. During the training stage, multiple decision trees are constructed, and the outcomes of the individual trees are predicted using regression methods. This approach has the advantage of reduced variance and is capable of effectively correlating multiple features of the data for forecasting purposes. However, one of the drawbacks of this algorithm is that it is often challenging to interpret the random forest classification algorithms.

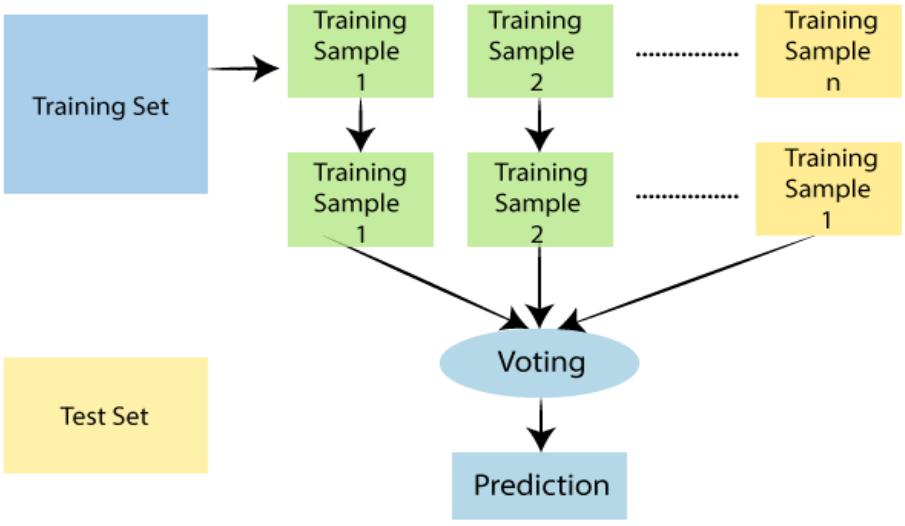

Fig. 6. Plot of Simplified Random Forest

Entropy uses the probability of a certain outcome in order to make a decision on how the node should branch and entropy to determine how nodes branch in a decision tree.

Entropy = $\sum_{i=1}^{c} -pi * log2\ (pi)$..... (4)

Support Vector Machine: 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 the new instances of the input data model. The hyperplane is a line that segregates the given hyperplane into two parts in a two-dimensional 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.

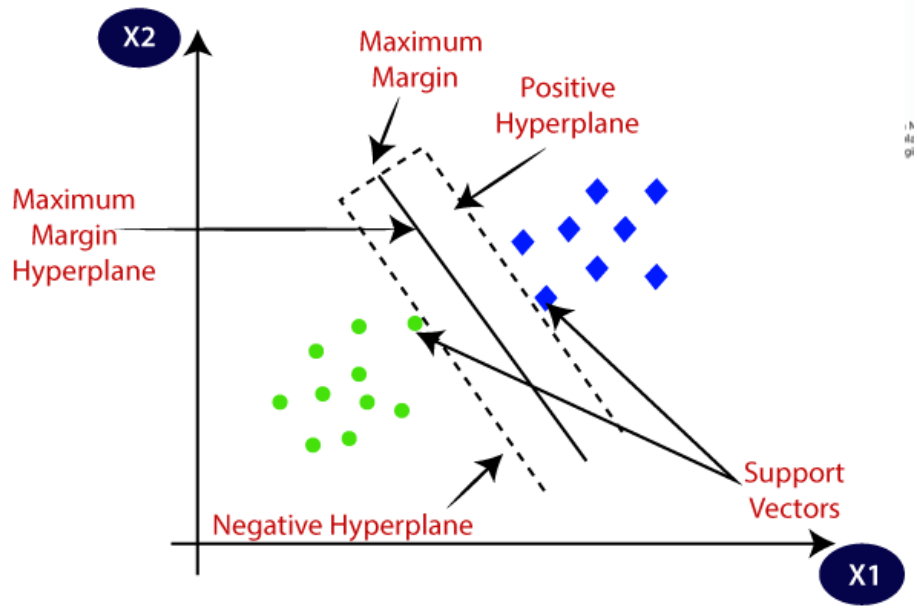

Fig.7. Support Vector Machine ML Algorithm

The two-dimensional linearly separable data can be separated by a line. The function of the line is y =a x + b. Rename x with x_1 and y with x_2 and we get:

$ax_1 - x_2 + b = 0$……. (5)

If we define **x** = (x1, x2) and **w** = (a, −1), we get:

w. x + b=0…… (6)

This equation is derived from two-dimensional vectors. But in fact, it also works for any number of dimensions. This is the equation of the hyperplane.

3. Result and Discussion

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 (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, KNN, SVM, Decision Tree and Random Forest. Research

work carried out using google colab using python language. In this section results of all implemented supervised machine algorithms are depicted:

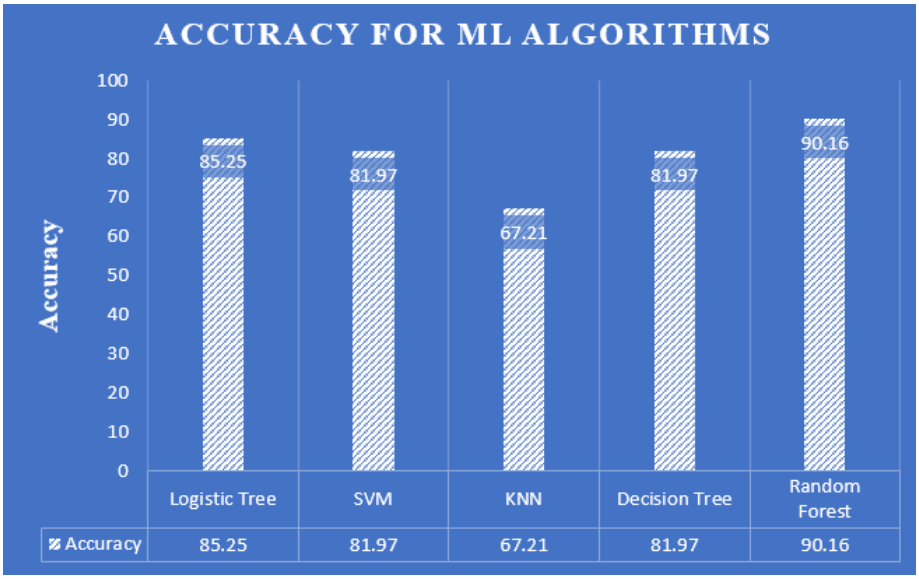

Fig. 8. Comparative analysis of accuracy among implemented algorithms

Among these algorithms, Random Forest performed better with an accuracy of 90.16%, while KNN had the lowest accuracy of 67.21%. Decision Tree and SVM both achieved an accuracy of 81.97%. It is important to note that accuracy should not be too high as it could mean overfitting the model to a specific dataset. An accuracy of more than 80% is considered decent, and anything above 90% is admirable.

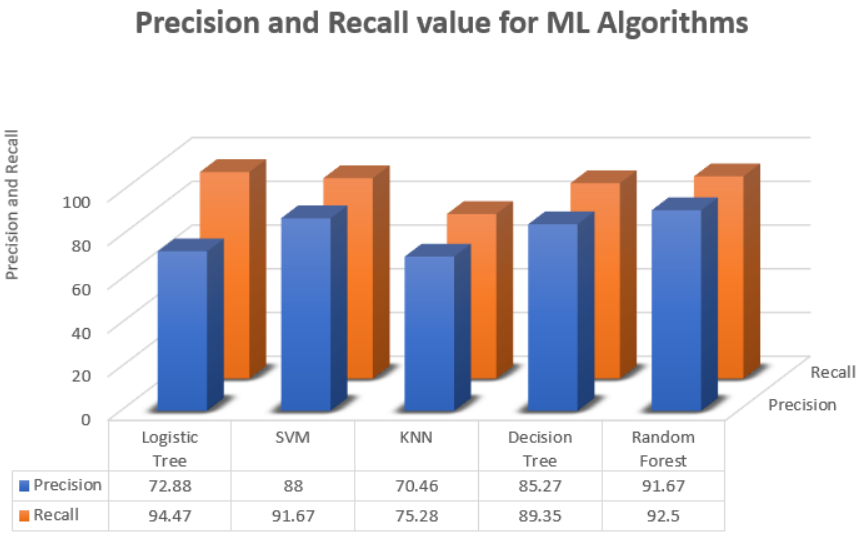

Fig.9. Precision and Recall value for ML Algorithms

Additionally, precision and recall were calculated and Random Forest had a precision value of 91.67 and recall value of 92.5.

4. Conclusion

Healthcare is a major concern in today's modern lifestyle and diagnosing diseases in their early stages can reduce fatality rates. Emerging fields like Soft Computing, Machine Learning (ML), and Deep Learning (DL) are playing a crucial role in healthcare by accurately predicting severe diseases. This research work used a standard dataset downloaded from Kaggle website containing various parameters related to humans. The data was pre-processed to resolve any ambiguities that could affect algorithm accuracy, and was divided into training (70%) and testing (30%) datasets. The study then implemented supervised ML algorithms, including Logistic Regression, KNN, SVM, Decision Tree, and Random Forest, to predict heart diseases. Among these algorithms, Random Forest performed better with an accuracy of 90.16%, while KNN had the lowest accuracy of 67.21%. Decision Tree and SVM both achieved an accuracy of 81.97%. It is important to note that accuracy should not be too high as it could mean overfitting the model to a specific dataset. An accuracy of more than 80% is considered decent, and anything above 90% is admirable. Additionally, precision and recall were calculated and Random Forest had a precision value of 91.67 and recall value of 92.5.

References

[1]. H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, “Heart disease prediction using machine learning algorithms,” in ICCRDA 2020, IOP Conf. Series: Materials Science and Engineering, vol. 1022, IOP Publishing, 2021, p. 012072. doi: 10.1088/1757-899X/1022/1/012072.

[2]. P. Motarwar, A. Duraphe, G. Suganya, and M. Premalatha, “Cognitive Approach for Heart Disease Prediction using Machine Learning,” in 2020 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, and M. Gupta, “Heart Disease Prediction using Machine Learning Techniques,” in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), IEEE, Dec. 18-19, 2020. doi: 10.1109/ICACCCN51052.2020.9362842.

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

[5]. S. Mohan, C. Thirumalai, and G. Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,” IEEE Access, vol. 7, Special Section on Smart Caching, Communications, Computing and Cybersecurity For Information-Centric Internet Of Things. doi: 10.1109/ACCESS.2019.2923707.

[6]. D. K. G, S. K. D, A. K, and V. Mareeswari, “Prediction of Cardiovascular Disease Using Machine Learning Algorithms,” in Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.

[7]. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of Heart Disease Using Machine Learning,” in Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), IEEE Conference, IEEE Xplore ISBN:978-1-5386-0965-1.

[8]. S. Zhang and Y.-L. S. A., “Deep learning-based recommender system: a survey and new perspectives,” Journal of ACM Computing Surveys, vol. 1, no. 1, pp. 1–35, 2017.

[9]. A. Khatami, A. Khosravi, and C. L., “Medical image analysis using wavelet transform and deep belief networks,” Journal of Expert Systems with Applications, vol. 3, no. 4, pp. 190–198, 2017.

[10]. A. Shetty and C. Naik, “Different data mining approaches for predicting heart disease,” International journal of innovative research in science, engineering and technology, vol. 3, no. 2, pp. 277–281, 2016.

[11] S. Aydin, "Comparison and evaluation data mining techniques in the diagnosis of heart disease," Indian Journal of Science and Technology, vol. 6, no. 1, pp. 420-423, 2016.

[12] N. Bayasi and Tekeste, "Low-power ECG-based processor for predicting ventricular arrhythmia," IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 24, no. 5, pp. 1962-1974, May 2016.

[13] B. Berikol and Yildiz, "Diagnosis of acute coronary syndrome with a support vector machine," Journal of Medical System, vol. 40, no. 4, pp. 11-18, 2016.

[14] Z. Wang, X. Liu, and J. G., "Identification of metabolic biomarkers in patients with type-2 diabetic coronary heart diseases based on metabolomic approach," Journal of Cardiovascular Diseases & Diagnosis, vol. 6, no. 30, pp. 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 Systems, vol. 2, no. 6, pp. 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, vol. 2, no. 7, pp. 678-684, 2016.

[17] R. Sali and 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, vol. 15, no. 1, pp. 312-327, 2016.

[18] P. K. Ghadge, "Intelligent heart attack prediction system using big data," International Journal of Recent Research in Mathematics, Computer Science and Information Technology, vol. 2, no. 2, pp. 73-77, 2016.

[19] G. Purusothaman and K. Krishnakumari, "A survey of data mining techniques on risk prediction: heart disease," Indian Journal of Science and Technology, vol. 8, no. 5, pp. 643-651, 2015.

[20] A. Richter, J. Listing, M. Schneider, T. Klopsch, A. Kapelle, J. Kaufmann, A. Zink, and A. Strangfeld, "Impact of treatment with biologic dmards on the risk of sepsis or mortality after serious infection in patients with rheumatoid arthritis," Annals of the Rheumatic Diseases, pp. 147-153, 2015.

[21] S. Sairabi and D. Mujawar, "Prediction of Heart Disease using Modified K-means and by using Naive Bayes," International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, 2015.

[22] M. Vafaie and M. Ataei, "Heart diseases prediction based on ECG signals classification using a genetic-fuzzy system," Journal of biomedical signal processing and control, vol. 14, no. 5, pp. 291–296, 2014.

[23] Z. Wang, "Study on qi-deficiency syndrome identification modes of Coronary heart disease based on metabolomic biomarkers," Journal of evidence-based complementary and alternative medicine, vol. 24, no. 16, pp. 192–198, 2014.

[24] X. Yang, M. Li, Y. Zhang, and J. Ning, "Cost-sensitive naive bayes classification of uncertain data," Journal of Scientific World, vol. 9, no. 8, pp. 1897–1904, 2014.

[25] D. Chandna, "Diagnosis of heart disease using data mining algorithm," International journal of computer science and information technologies, vol. 5, no. 2, pp. 1678–1680, 2014.