Development of Supervised Machine Learning Models in Healthcare to Predict Heart Attack

Anshika Aggarwal[1], Bhavya[2], Aryan[3], Pankaj Kansal[4]
Department of CSE (AI&ML)[1,2,3,4]
Panipat Institute of Engineering and Technology, Panipat, India[1,2,3,4]

Abstract: *Cardiac arrest and heart-attack cases are rising sharply year after year, and World Health Organization (WHO) figures place the resulting global death toll above 18 million annually. This study aims to build machine learning models capable of predicting heart disease with improved accuracy. Five supervised algorithms were implemented for this purpose: Random Forest (RF), Support Vector Machine (SVM), Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN). Random Forest delivered the strongest result at 90.16% accuracy, Logistic Regression followed at 85.25%, and KNN trailed the group at 67.21%. Precision and recall were also derived from the confusion matrix for each model. Since near-perfect accuracy can signal overfitting rather than genuine predictive strength, models exceeding 90% accuracy were treated here as the reliability benchmark.*

Keywords:- Random Forest, Decision Tree, SVM, Machine Learning, Heart Disease, KNN, Logistic Regression.

Introduction

Cardiovascular disease is one of the most significant problems in today's health care, and accounts for a significant proportion of deaths annually due to heart attack and cardiac arrest. There are several reasons for this, but the most prominent include poor eating habits and the stress that comes with modern life (including work stress, monetary difficulties, etc.). Delayed diagnosis also worsens the situation and consequently, the annual mortality rate due to heart attacks continues to rise at a rapid pace [1] [2].

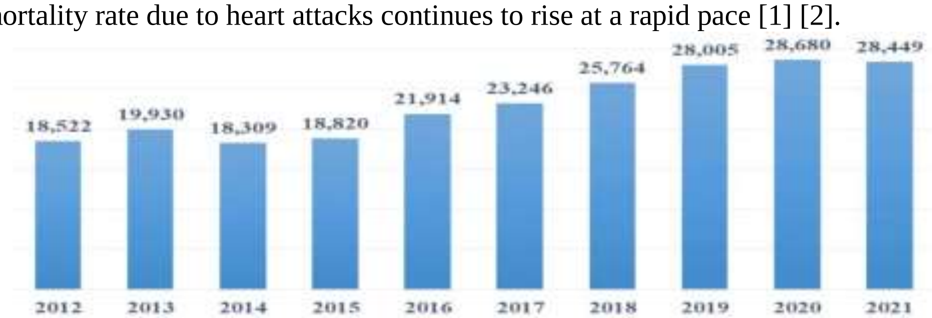

Figure 1: Deaths due to heart attack in India [in 2022, 70 per cent of heart attack deaths occurred in the 30–60 age group].

Several different classification methods (Naive Bayes + Laplace smoothing) have been used to predict heart diseases to date and machine learning algorithms in general have yielded real-time observations and results to support this task [3]. There is also a wide body of related work that investigates machine learning and data-mining methods to predict heart and cardiovascular disease [4], [7]–[11], [13]–[25]. Clinicians will usually use a number of diagnostic tests – summarized in a variety of statistical measures – to help reduce the chances that someone has heart disease. The term 'heart' comes from the Latin word for the organ that is clinically called the cardiac muscle. Fig. 2 is the anatomical structure of the heart. A hollow organ made of muscle tissue, with a partition between four chambers, two on top and two below (atria and ventricles). Blood entering the heart goes through the atria and is pumped through the ventricles to circulate throughout the body of the body [5]. Oxygenated blood is blood which carries oxygen, which helps the body to carry out its functions.

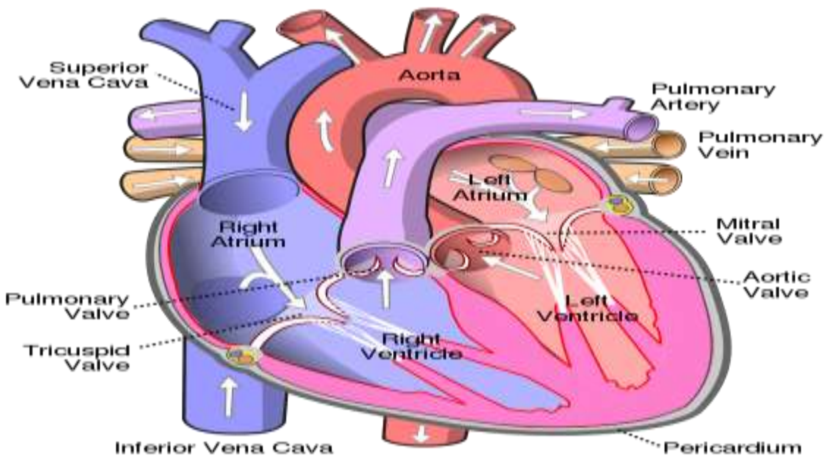

Figure 2: Anatomic structure of the heart.

Heart disease is generally categorized according to abnormalities within the blood vessels themselves — irregular white blood cell activity, altered platelet counts, and fatty deposits that obstruct normal blood flow to the heart are among the most common underlying causes [6].

Methodology

Use parameters from Kaggle to create this study, as this is the data that has parameters for patients. All data used is pre-processed to ensure that there are no inconsistencies that may impact the performance of the algorithms. The data set was then split into 70% training data set and 30% test data set, the model was then fitted on the training data set. The latter was then used to develop a successful prediction technique for heart-disease using five supervised learning algorithms (Logistic Regression, KNN, SVM, Decision Tree and Random Forest).

A. Logistic Regression

Logistic regression is a statistical technique used to analyze data with a binary dependent variable, with its parameters estimated directly from the underlying logistic model. Technically, it expresses the probability of an event as a linear combination of the independent variables. It isn't a classifier in the strict sense — rather, it produces an output probability from given inputs — but it is commonly turned into one by applying a cut-off threshold: values falling below the threshold are assigned to one class and those above it to the other. Two main variants exist: Multinomial Logistic Regression (MLR), which works with absolute values and sorts outputs into more than two categories, and Ordinal Logistic Regression (OLR), which arranges those multiple outputs in a meaningful order. In both cases, the underlying likelihood of an event remains a linear combination of the independent variables.

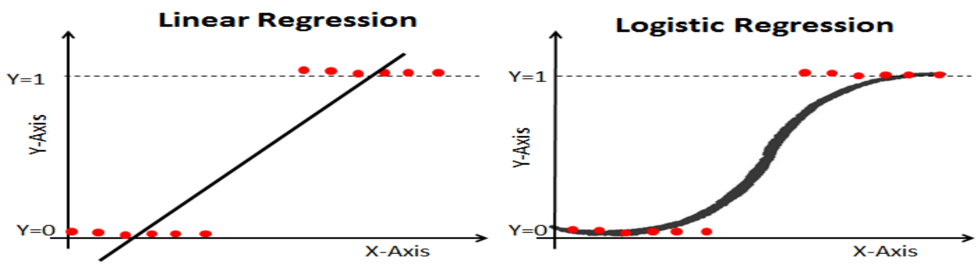

Figure 3: Linear regression versus logistic regression.

The sigmoid function is used as the activation function for logistic regression and is defined as:

$f(x) = 1 / (1 + e^{-x})$ (1)

Logistic regression is expressed as:

$y = e^{(b0 + b1x)} / (1 + e^{(b0 + b1x)})$ (2)

where x is the input value, y is the predicted output, b0 is the bias/intercept term, and b1 is the coefficient for input x.

B. Decision Tree

Decision trees are widely applied in healthcare settings and are built around core concepts such as the root node, internal branches, and leaf nodes. Each tree arrives at its own class prediction, with different trees formed by varying the parameters considered. A voting classifier then combines these individual trees into a single, more accurate final model — this ensemble strategy works for both classification and regression tasks. Multiple trees are grown during training, and their individual regression-based outputs are aggregated to produce forecasts. The main benefit of this approach is lower variance, along with a strong ability to combine multiple input features when making predictions.

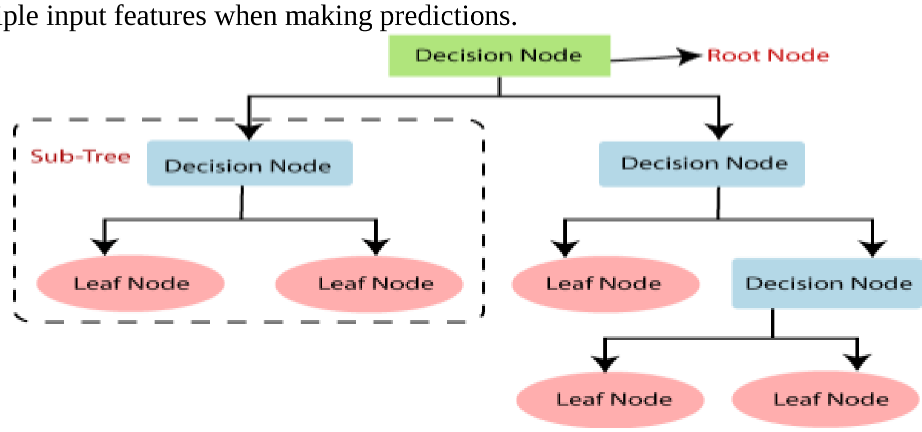

Figure 4: Decision tree algorithm flowchart.

C. K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is one of the simplest supervised learning algorithms available. It works by comparing a new data point against existing data and placing it into whichever class it most closely resembles. Because it keeps the entire training dataset in memory and classifies new points using a similarity measure, KNN can technically handle both classification and regression, though classification is by far its more common use. Its main drawback is that it tends to be slow and does not actually learn patterns from the training data in the way other algorithms do.

Euclidean = $\sqrt{(\Sigma(x_i - y_i)^2)}$, i = 1…k (3)

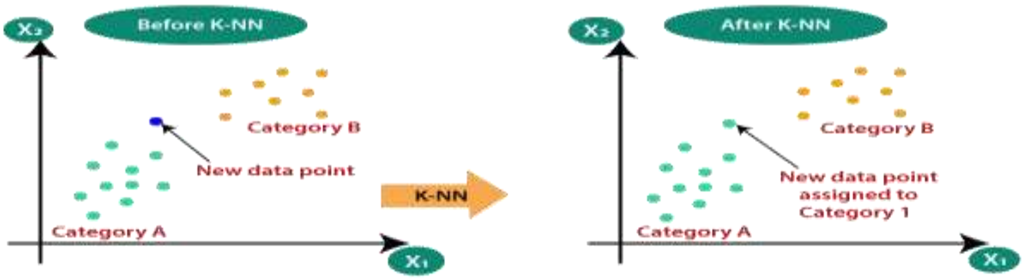

Figure 5: KNN machine learning algorithm.

D. Random Forest

Random Forest builds on the decision tree concept by combining a large collection of individual trees into a single ensemble. Each tree makes its own class prediction based on a different subset of parameters, and a voting mechanism merges these predictions into one optimized final model — a strategy that applies to both classification and regression problems. Because many trees are trained in parallel and their regression-based outputs pooled together, the resulting model tends to have lower variance and can capture relationships across many input features effectively. The main trade-off is interpretability: understanding exactly why a random forest reached a particular classification is considerably harder than doing so for a single tree.

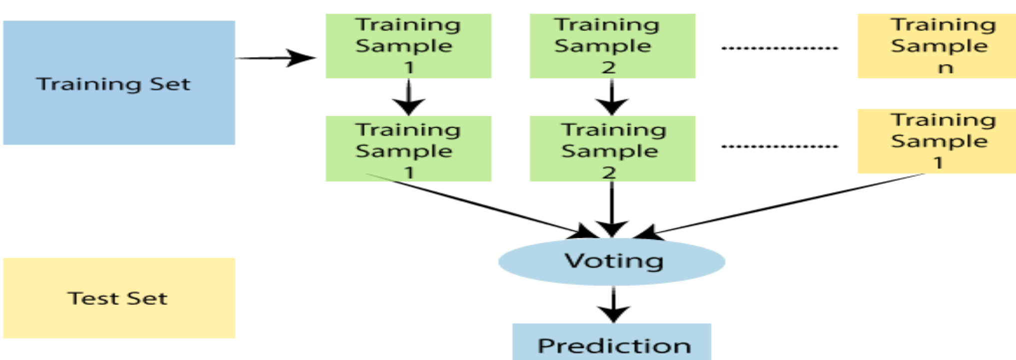

Figure 6: Plot of a simplified random forest.

Entropy quantifies the likelihood of a particular outcome and is used to decide how each node should branch while constructing a decision tree:

Entropy = $\Sigma -p_i \cdot \log_2(p_i)$, i = 1…c (4)

E. Support Vector Machine (SVM)

Non-linear SVM is among the most widely used algorithms for handling unlabelled data and finds application across numerous industries. Given a labelled training set, it identifies an optimal hyperplane that can then be used to classify new incoming instances. In a two-dimensional space, this hyperplane is simply a dividing line, with each class occupying one side of it [12]. The analysis carried out in this work relies on Support Vector Machines (SVMs), where the algorithm identifies relevant features starting from a collection of sample feature sets drawn from the dataset.

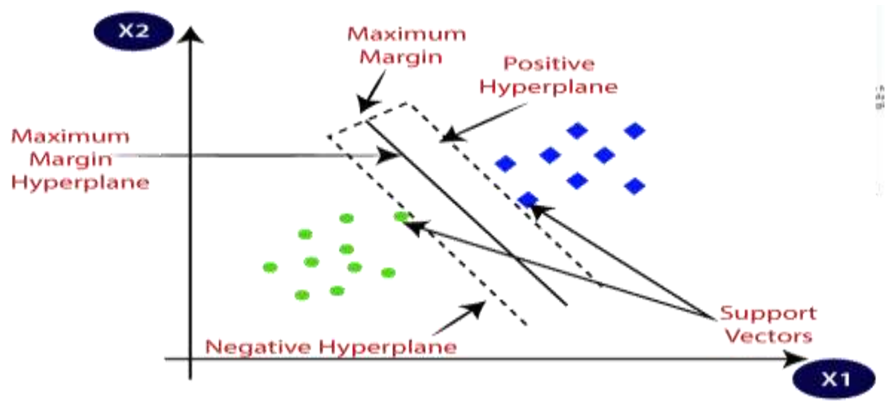

Figure 7: Support vector machine (SVM) algorithm.

The two-dimensional linearly separable data can be separated by a line whose function is y = ax + b. Renaming x as x1 and y as x2 gives:

$$ax1 - x2 + b = 0 \quad (5)$$

Defining x = (x1, x2) and w = (a, −1) gives:

$$w \cdot x + b = 0 \quad (6)$$

This relation, derived here for two dimensions, generalizes to any number of dimensions and defines the equation of the hyperplane.

Results and Discussion

The dataset used in this study is obtained from the standard Kaggle dataset, which has a variety of human health parameters. The data was pre-processed to correct for any potential inconsistencies that would negatively impact the accuracy of the algorithms, then split the data into a training set (70%) and a test set (30%) – the ratio may vary, but the idea is to produce the most accurate model possible. Finally, five supervised ML algorithms (LR, KNN, SVM, DT, and RF) were used to create an optimal heart-disease prediction approach. All experiments were carried out in Google Colab using Python. The respective outcomes of each of the implemented algorithms are given below:

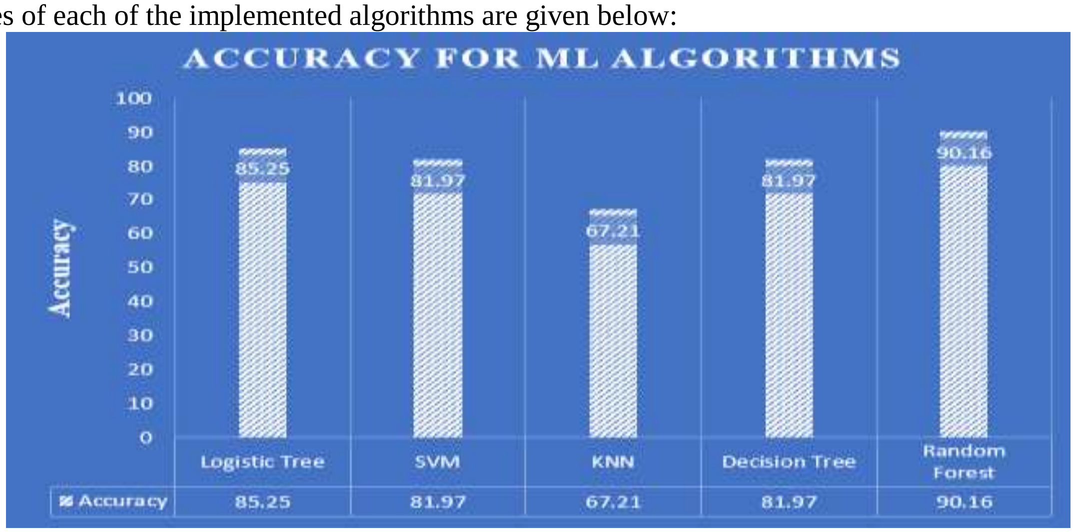

Figure 8: Comparative analysis of accuracy among implemented algorithms.

The best results were obtained by the Random Forest model (90.16%) and the weakest by the KNN model (67.21%). The results of Decision Tree and SVM were 81.97%. It is important to note that an exceptionally high accuracy figure might also indicate that the algorithm is simply memorizing the patterns of the specific dataset, and not actually predicting the data well; the threshold for good and excellent accuracy is as follows: over 80% accuracy is a red flag for overfitting the data; above 90% accuracy is considered a good-to excellent-accuracy score.

Figure 9: Precision and recall values for the ML algorithms.

Precision and recall were also computed; Random Forest achieved a precision of 91.67 and a recall of 92.5.

Conclusion

We first gave an overview of the range of input devices and methods for alphanumeric data. We then had a closer look at touch-typing as input method and highlighted its benefits. We found that the trend goes towards retaining the original keyboard metaphor as closely as possible. We have designed a new Gadget i.e., Sun-Key. In this device we have removed some of the drawbacks of previously explained virtual keyboards.

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 Int. Conf. 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 Int. Conf. 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 Int. Conf. (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, 2019. 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 Proc. 2018 IEEE Int. Conf. on Current Trends toward Converging Technologies, Coimbatore, India, 2018.

[7] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," in Proc. 2nd Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA 2018), IEEE, 2018, ISBN: 978-1-5386-0965-1.
[8] S. Zhang and Y.-L. S. A., "Deep learning-based recommender system: a survey and new perspectives," 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," 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," Int. J. Innovative Research in Science, Engineering and Technology, vol. 3, no. 2, pp. 277–281, 2016.
[11] S. Aydin, "Comparison and evaluation of data mining techniques in the diagnosis of heart disease," Indian J. 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 Trans. 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," J. 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," J. 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," J. Fuzzy Systems, vol. 2, no. 6, pp. 1377–1382, 2016.
[16] S. Prabhavathi, "Analysis and prediction of various heart diseases using DNFS techniques," Int. J. 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," Int. J. 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," Int. J. 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 J. 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 Naive Bayes," Int. J. Innovative Research in Computer and Communication Engineering, vol. 3, 2015.
[22] M. Vafaie and M. Ataei, "Heart disease prediction based on ECG signals classification using a genetic-fuzzy system," 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," 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," Scientific World Journal, vol. 9, no. 8, pp. 1897–1904, 2014.
[25] D. Chandna, "Diagnosis of heart disease using data mining algorithm," Int. J. Computer Science and Information Technologies, vol. 5, no. 2, pp. 1678–1680, 2014..