An Analysis of SVM and KNN Based Machine Learning Approaches for Heart Attack Prediction

Chhaya, Chaman Kumar, Abhinav Dahiya, Kamaldeep Joshi, Rainu Nandal, Rajkumar Yadav
Department of Computer Science & Engineering
University Institute of Engineering & Technology,
Maharshi Dayanand University, Rohtak 124001, India
Email: chhayasharma.rs@gmail.com, chamans135@gmail.com, abhinav.rs.uiet@mdurohtak.ac.in, kamalmintwal@gmail.com, rainu.nandal@yahoo.com, rajyadav76@rediff.com

Abstract –
Heart attacks are one of the most severe health concerns in the modern world. Early detection of this disease is crucial because it has the potential to save more than 17.9 million lives annually. Heart attacks are a form of coronary artery disease which occurs when there is inadequate oxygenation in one or more heart regions. Several health factors and habits influence the risk of heart disorders. In recent years, this scientific field has received a lot of interest. While forecasting cardiac diseases has always been challenging and crucial for medical professionals, doing so at an early stage will benefit people everywhere so that appropriate preventative measures can be taken before the condition becomes severe. Over time, machine learning tools have been applied to the healthcare industry to forecast disease accurately. This paper studies K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) models to predict cardiac attacks accurately. The effectiveness of these models is assessed using K-Folds cross-validation and evaluation metrics like accuracy, precision, recall, F1 Score, and AUC. The KNN algorithm is proven more efficient and precise than SVM, with an accuracy score of 90.3% compared to SVM's only 87%.

***Index Terms* - SVM, KNN, Performance Metrics, Heart Attack Prediction**
DOI Number: 10.48047/NQ.2022.20.13.NQ88527 **Neuroquantology 2022; 20(13):4328-4336**

I.INTRODUCTION

Heart attacks are the leading cause of death globally. The World Health Organization (WHO) reported that heart attacks and strokes caused nearly 32% of all fatalities in 2019 and accounted for 85% of all fatalities globally. India accounts for one-fifth of the deaths caused by cardiovascular diseases worldwide. In 2016, 28.1% of all deaths were due to cardiovascular disease [1] [2].

A heart Attack, also known as myocardial infarction, is an acute event and usually occurs when one or more areas of the heart do not get enough oxygen. The most common reason for the blockage of oxygen supply to the heart is the building of plaque in the heart arteries. These plaques lead to blood clots that partially or totally block the blood flow in the heart's arteries and are the most common reason for a heart attack. Health factors such as lipid disorder, high blood pressure, metabolic syndrome, etc., and health behaviors such as smoking, alcohol consumption, dietary habits, physical inactivity, sedentary behaviors, etc. contribute to cardiovascular health.

Heart attack symptoms can start slowly, and they can be mild or severe and sudden. Symptoms of a heart attack include angina, discomfort or pain in arms, irregular heartbeats, excessive sweating, nausea, dizziness, weak pulse, shortness of breath, cold & clammy skin, etc. These symptoms can differ across individuals and between genders.

The fact that heart disorders can emerge silently and unexpectedly makes it challenging to diagnose them, along with the numerous risk factors that contribute to their development.

Artificial Intelligence (A.I.) has the potential to revolutionize healthcare practices. Machine Learning (ML) is a growing interdisciplinary field of A.I. Its techniques have shown outstanding results in healthcare domain. ML facilitates healthcare-centered intelligent solutions along with many benefits like enabling remote healthcare services to combat the medical person's shortage problem, minimizing the cost and duration of hospitalization, providing a better health infrastructure for rural & low-income areas, and so on. Because of their complex nature, cardiovascular disorders must be managed cautiously. Various machine learning algorithms have been employed to determine the causes and severity of heart attacks because of their potential to produce highly accurate predictive outcomes.

This paper explores K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) for the early prediction of heart attacks. A comparison is made between both methods based on evaluation metrics to determine the best model for medical use.

II.RELATED WORK

VrigeniyaS.C. and E. Ramaraj [3] have proposed an ML rule-based classification technique implemented using python. The data is gathered from Kaggle and tested with the proposed model before being used as input for the logistic regression algorithm (LRA). The model's strength is compared using evaluation metrics such as accuracy, F1 Score, precision, and recall. The results showed a higher accuracy of 86% compared to the existing LRA.

Yong Li et al. [4] have developed a simple architectural model of deep neural networks, CraftNet, to recognize handcraft features accurately. The proposed model combined multiple child classifiers with a customized structure to classify handicraft features by decision-directed acyclic graph and a mixed-loss function, P-S loss, for its optimization. CraftNet benefits from both deep learning methods and handcraft elements. The testing results on the MIT-BIH dataset demonstrated the sensitivity values of 88.16%, 85.37%, 94.53%, and 88.92% for four categories, as well as an increase in average sensitivity accuracy to 89.25% from 86.82%, demonstrating CraftNet's powerful recognition capability.

R. Sujatha et al. [5] have proposed a model to predict the survival chances of patients suffering from Heart failure using ML algorithms and scoring methods such as Information Gain, Gini Index, and Gain Ratio [5]. The classification accuracy obtained using S.M. is as follows: SVM (RBF with 83.6%, Sigmoid with 82.3%, Linear with 80.3%, Polynomial with 86.6%) & by using CFS-BFS technique: SVM (Sigmoid and RBF with 83.6%, Linear with 79.9%, Polynomial with 83.3%). The author has used "Multi-Kernel Support Vector Machine (MK-SVM) classifiers like radial base function (RBF), sigmoid, Linear, Polynomial."

Vinodhini V. et al. [6] have proposed an algorithmic model with ensemble techniques to improve conventional machine learning techniques. The EEG brainwave dataset from Kaggle is used for implementation. The suggested AWET algorithm's accuracy is 82%, which is 17% higher than the accuracy of SVM, Naïve Bayes, Random Forest, and Multilayer Perceptron classifiers.

Kumar Krishna et al. [7] have developed a mathematical model using a heart patient database. To model the patient's health and determine whether or not they have cardiac disease, curve fitting and artificial neural networks (ANN) are used. The proposed curve fitting model has been found to accurately recognize heart patients, with an R2-value of 0.6337, an MAE of 0.293, and an RMSE of 0.3688. In contrast, the ANN-based model accurately identifies patients with an R2-value of 08.491, an MAE of 0.20, and an RMSE of 0.267. These results demonstrated that for identifying cardiac patients, ANN's accuracy is found to be superior to curve fitting.

VermaLuxmi et al. [8] have suggested a novel hybrid method for CAD diagnosis, which assesses the presence of the illness and its severity using non-invasive clinical data. The correlation-based feature subset selection, K-means clustering techniques, and particle swarm optimization search method are used in the developed model to identify risk factors. The modeling of CAD instances employs multinomial logistic regression, multi-layer perceptron, C4.5, and fuzzy unordered rule induction method. The clinical data comprising 26 characteristics and 335 records gathered at the Department of Cardiology, IGMC, Shimla, India, and

Cleveland heart disease dataset are used to test this methodology. The MLR algorithm achieved the best prediction accuracy, which is 88.4%. The suggested model also raised the algorithm's accuracy for the Cleveland data from 8.3% to 11.4%.

Bodapati et al. [9] proposed a framework that leverages the advantage of the unsupervised information hidden in the data to improve the classifier accuracy. Using k-means & spectral clustering to avoid uncertainty, the augmented feature set is given as input to the classifier for disease prediction. Framingham and stat-log datasets are used to evaluate the effectiveness of the suggested algorithm & Logistic regression (L.R.), Naïve Bayes, and SVM for classification purposes. The experiment results showed that the proposed framework significantly improved the performance of the cardiac ailment prediction model by using the feature augmentation technique with cluster labels. The accuracy of the developed model on the Framingham dataset was as follows: 86.32% on L.R., 82.42% on Naïve Bayes, and 85.37% on SVM, while on the stat-log dataset, it was: 81.48% on L.R., 83.33% on Naïve Bayes and 85.18% on SVM.

K. Saikumar et al. [10] suggested a model that includes heart surgery procedures and quick diagnosis. The proposed approach makes use of C.T. angiography images. Technical image processing (TIP), Machine Learning (ML), and the Random Forest Optimization decision tree (RFO-DT) approach are used to identify problems in the region of the heart (ROH). This model has two stages: the first is preprocessing, and the second stage involves extracting D.T. and calculating probability values using the data set obtained from the IEEE data house website. The proposed model, RFO with A.I., provided the most accurate diagnosis findings by employing overlapping (O.V.) of 4.766%, coverage with the clinically significant piece of the vessel (O.T.) of 2.5%, traditional interior (A.I.) of 0.21%, and overlap until first plaque (OF) of 6.5%.

Shorewala V. [11] has evaluated prediction methods such as KNN, Binary Logistic Classification, and Naïve Bayes to forecast CHD using the risk factor strategy. The modeling methodologies were tested on the "Cardiovascular Disease Dataset," which included 70k patient data records of CAD patients. He has contrasted ensemble modeling methods with basic classifiers, including boosting, stacking, and bagging. The results were as followed: the bagged model had an average enhanced accuracy of 1.96% in contrast to their conventional counterparts, while the stacked models were found to be the most effective with a final accuracy of 75.1%. The highest AUC score was achieved by boosted models i.e., 0.73. Data-analytic techniques and K-Folds cross-validation were used to confirm the performance of the tested models.

III.EXPERIMENTAL SETUP & PARAMETERS FOR PERFORMANCE MEASURE

A. Dataset and Data Preprocessing

For this study, we used the data from the 'Heart Attack Analysis & Prediction' Dataset, Version 2, provided by Kaggle. Python was used to analyze, compute and visualize everything on Kaggle Notebook. 4330

The dataset contains 303 values with no missing values, as demonstrated in Fig. 1. Missing values are those that do not have any data stored in the variable for the observation.

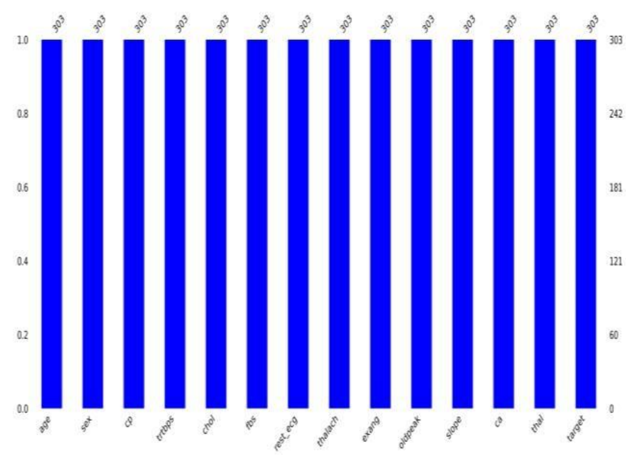


Fig. 1. Missing Values

Out of 303 records, 165 records were of healthy people, i.e., with no history of heart problems & 138 records were of patients with heart problems. The dataset consists of a set of 14 attributes. Table 1classifies the attributes, their description, and their types.

Table 1. Detailed Information of Dataset Attributes

Attribute	Description	Type
Age	Age of patient	Numeric
Ca	Fluoroscopy colored major vessels numbered 0-3	Numeric
Chol	Serum Cholesterol (mg/dl)	Numeric
Cp	Chest Pain	Nominal
Exang	Angina induced by exercise (0 denotes 'No' and 1 denotes 'Yes')	Nominal
FBS	Blood sugar levels during fasting > 120 mg/dl; 1 represents true, and 0 represents false.	Nominal
Oldpeak	ST depression induced by exercise as compared to the resting state	Numeric
Resting	ECG outcomes during resting	Nominal
Sex	Gender of the patient (F=0, M=1)	Nominal
Slope	S.T. segment calculation regarding slope while maximal exercise; States: 1. Unsloping, 2. Flat, 3. Downsloping	Nominal
Target	Diagnosis of heart disease demonstrated using 2 values: 0 and 1, where 0 = total absence & 1 = heart disorder	Nominal
Thal	Heart status demonstrated using three different numeral values: Normal = 3, fixed defect = 6, & reversible defect = 7.	Nominal
Thali	The maximum rate of heart accomplishment	Numeric
Trestbps	The blood pressure level of the patient at resting condition at the time of admission in the hospital (in mm/Hg)	Numeric

Out of 14 attributes, 3 attributes, namely Trestbps, Chol, and FBS, were removed because of very little correlation with the target feature, as shown in Fig. 2. The correlation matrix is converted into color labeling using a correlation heat map. The correlations heatmap illustrates the strength of correlations between numerical features. The remaining attributes are included in the heart attack prediction model, where one feature acts as a predictive attribute of the presence of cardiac disorder in the patient.

Fig. 2 Correlation diagram between attributes

The dataset is loaded, cleaned, and processed using preprocessing methods in the first stage. The subset of 14 attributes, namely Age, Ca, Sex, Trestbps, FBS, Cp, Chol, Thali, Thal, Resting, Oldpeak, Exang, target, and Slope, is chosen from the pre-processed dataset. Fig. 3 shows the workflow of the experiment.

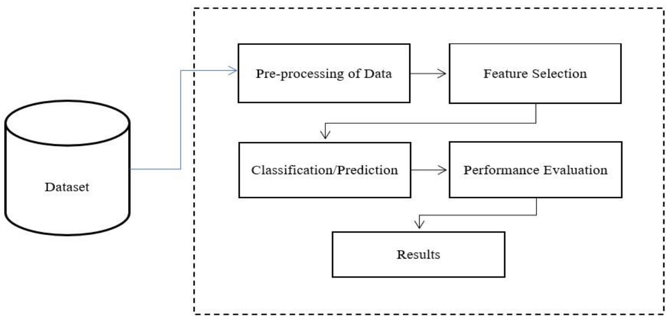


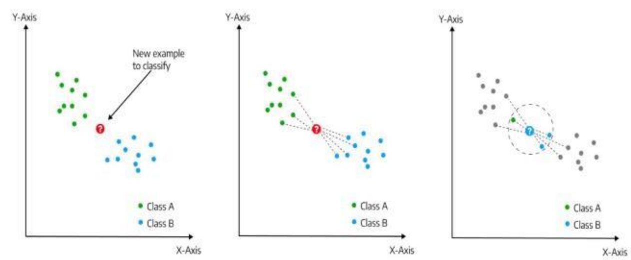
Fig. 3 Experiment workflow with the Kaggle dataset

To pre-process data, medical records are converted into diagnosis values. During the pre-processing stage, all 303 patient records are used. Multi-class variables are used to determine whether or not a heart disorder exists. The value is 1 if the patient has a heart condition; otherwise, it is 0, which indicates that the patient is healthy and has no cardiac problem.

Among 14 features of the employed dataset, two attributes, age & sex, are utilized to categorize the patient's personal data. The remaining 12 outputs are crucial because they include medical records necessary for diagnosis and understanding of the cardiac ailment severity. As indicated previously, KNN and SVM machine learning algorithms are employed for the experiment, and the evaluation is done using a confusion matrix. Many standard performance metrics such as accuracy, f-measure, recall, AUC, and precision are considered for computing the performance efficiency of both models. The recall, precision, F1 score, and accuracy calculations are based on the confusion matrix's four results (TP, TN, FP, and FN). Each algorithm's performance is assessed individually, and all outcomes are effectively documented for future searches.

Sklearn's Robust Scalar approach is utilized for numerical features while one-hot encoding is used for categorical information.

Fig. 4 KNN Model



B. Classification of Models 1)K-Nearest Neighbor (KNN) Algorithm

The KNN is a non-parametric supervised ML algorithm mainly used as a classification algorithm. However, it can be used in both regression and classification problems. It is known for its ease of interpretation and implementation. KNN predicts the classification of unlabeled data by considering the training data's labels and features. Two parameters are used in this algorithm –

(a) the distance between the query point and other points, &
(b) parameter k, which is used to decide the number of neighbors.

An appropriate value of k shall be chosen such that there is a balance between underfitting and overfitting. By default, Euclidean distance is used by the knn function, while alternative methods include Manhattan distance, Minkowski distance, and Hamming distance.

The KNN algorithm determines the value of new data sets using the feature similarity approach, which means that the value of the new data point is determined by how closely it resembles the elements in the training set. Fig. 4 demonstrates this.

2)Support Vector Machine (SVM) Algorithm SVMs are a set of supervised ML techniques used for classifying, regression, and recognizing outliers. It maximizes forecast accuracy while automatically avoiding over-fitting data utilizing ML theory. SVM is based on the structural risk minimization technique. Also, it can solve both linear and non-linear problems. The basic idea behind SVM is that it divides data into classes by constructing a line or hyperplane. The line which categorizes the two different datasets to their similar data types is called the "Maximum Margin Hyperplane." Fig. 5 shows an SVM model.

A support vector machine model consists of the following four concepts: (a) the maximum-margin hyperplane, (b) the separating hyperplane, (c) the kernel function, and (d) the soft margin.

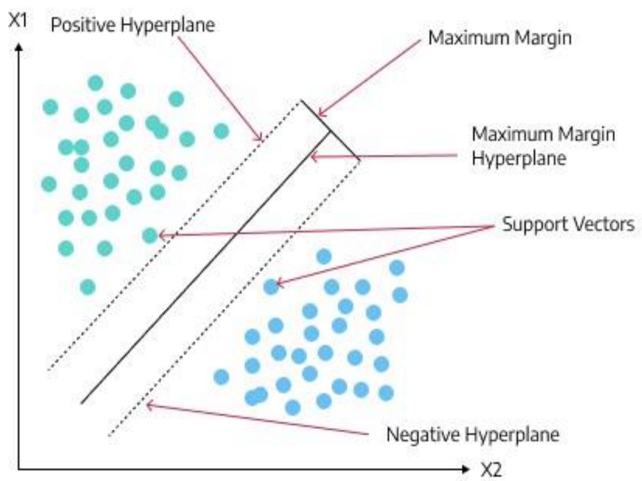


Fig. 5 SVM Model

C. Proposed Work

All numerical features are scaled with the Robust Scalar method, and qualitative features are classified with the One Hot encoding process. Rather than relying on the middle value, as is done with the median, the Robust Scaling technique scales the data between the lower quantiles (25th and 75th) or the first and third quartiles. In other way of putting it, the median is eliminated. Training and a test subset are extracted from the dataset using a test split of 0.1.

Sklearn'sKNeighbors Classifier is applied to the dataset. The number of neighbors' values is kept at 5. The proposed KNN model is compared with Sklearn's support vector machine model.

D. Classifier Validation Approach

Cross-validation is used for the evaluation of ML techniques to test their performances. In this, the dataset is divided into k number of subsets. The other set (k-1) is taken as the training set, and 1 subset out of k is employed as the test set each time. The model is tested on each test group, and this process goes on for all k groups. The best model is selected by choosing the highest out of all calculated scores. We have used k = 10 i.e., 10-fold cross-validation in this study.

E. Confusion Matrix

A confusion matrix is a method of describing the number of test cases that are classified correctly and incorrectly in matrix form to measure the classification model prediction. Here, P represents Positive prediction, N represents Negative prediction, and Truth values are represented by T for True and F for False.

True Positive (TP) - When both the expected and actual classes have positive values.

True Negative (TN) - When the projected and actual classes are both negative.

False Positive (FP) - When the projected class value is positive, but the actual class value is negative. False Negative (FN) - When the value of the projected class is no, the actual class value is yes.

Fig. 6 and 7 show the confusion metrics of the KNN and SVM models, respectively.

F. Evaluation Metrics

Evaluation metrics are those metrics that are used to measure the quality of ML or statistical models. In this paper, the following evaluation metrics are used:

Accuracy: Accuracy is calculated as the proportion of correctly identified test cases to all test case instances.

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

Precision: Precision is measured as the proportion of accurately predicted classifications to all positive classifications.

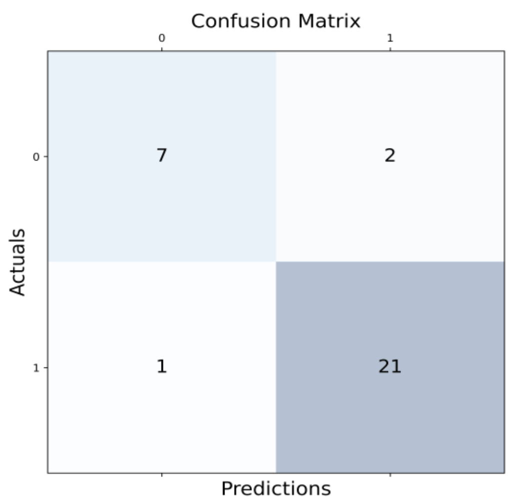


Fig. 6 KNN Confusion Matrix

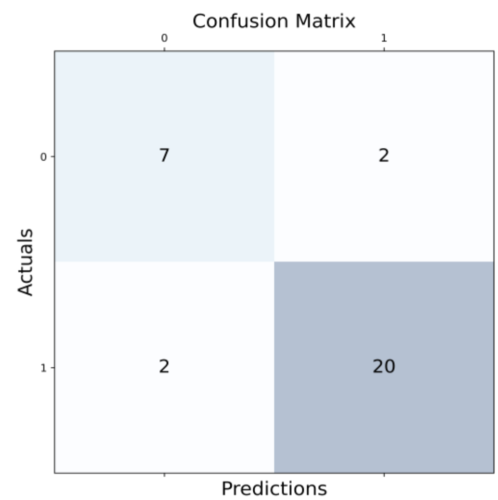


Fig. 7 SVM Confusion Matrix

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Recall: Recall measures the model's ability to predict positive outcomes correctly.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}$$

F1 Score: It is the harmonic mean of precision and recall, which supports a better understanding of the classifier's accuracy.

$$\text{F1 Score} = 2 * \left(\frac{\text{Precison} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

Area Under the ROC Curve (AUC-ROC):

The AUC-ROC curve is a performance metric that considers all potential classification criteria; the bigger the area, the better the model will perform.

Fig. 8 and 9 show the AUC of the KNN and SVM models, respectively.

IV. RESULT

The prediction model is created using 10 features. The scores for recall, F-measure, accuracy, precision, and AUC are compared inTable 2. The KNN algorithm's accuracy, at 90.32%, is higher than the SVM algorithm's accuracy, which is 87.10%.

Table 2. Comparison Table for SVM and KNN Model

Model	Accuracy	Precision	Fmeasure	Recall	AUC
KNN	90.32%	91.30%	93.33%	95.45 %	89 %
SVM	87.10%	90.91%	90.91%	90.91 %	89 %

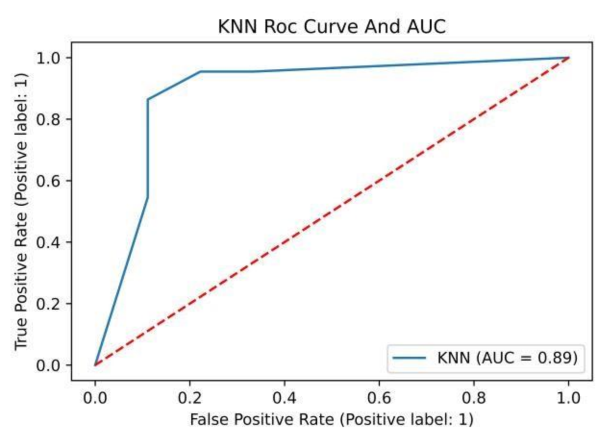


Fig. 8 AUC of K Nearest Neighbour (KNN)

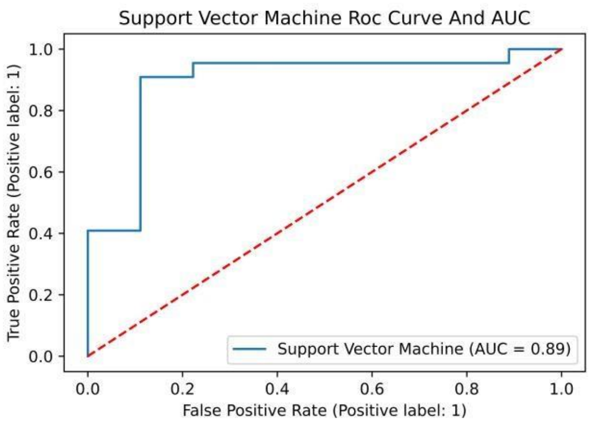


Fig. 9 AUC of Support Vector Machine (SVM)

V. CONCLUSION

Heart disease forecasting is very challenging because of the disease's complex nature and its various associated features. If the ailment is identified early and preventive steps are adopted and followed as soon as feasible, mortality rates can be dramatically reduced. Long-term life-saving and early diagnosis of irregularities in heart diseases can be made possible by identifying and processing the raw medical data of heart disorders. This paper analyzes the effectiveness of machine learning techniques SVM and KNN for the early forecast of cardiac attacks. We find that the KNN algorithm outperforms the SVM algorithm, with accuracy values of 90.32%, precision values of 91.30%, recall values of 95.45%, F-measure values of 93.33%, and AUC scores of 89% for the KNN algorithm, compared to accuracy values of 87.10%, precision values of 90.91%, recall values of 90.91%, and Fmeasure values of 90.91% for the SVM algorithm. Future studies in this field can be carried out using various feature selection approaches and ML technique combinations for better prediction results.

REFERENCES

[1] Jadhav, A. D., &Chobe, S. V. (2022). Risk Assessment of Cardiovascular Diseases Using kNN and Decision Tree Classifier. Revue D'IntelligenceArtificielle, 36(1), 155–161.

[2] Kumar, A. S., & Sinha, N. (2020). Cardiovascular disease in India: a 360-degree overview. Medical Journal, Armed Forces India, 76(1), 1-3.

[3] Virgeniya, S. C., &Ramaraj, E. (2019). Predictive analytics using rulebased classification and hybrid logistic regression (HLR) algorithm for decision making. International Journal of Scientific and Technology Research, 8(10), 1509-1513.

[4] Li, Y., He, Z., Wang, H., Li, B., Li, F., Gao, Y., & Ye, X. (2020). CraftNet: a deep learning ensemble to diagnose cardiovascular diseases. Biomedical Signal Processing and Control, 62, 102091,1-9.

[5] Sujatha, R., Chatterjee, J. M., Jhanjhi, N. Z., Tabbakh, T. A., &Almusaylim, Z. A. (2022). Heart Failure Patient Survival Analysis with Multi Kernel Support Vector Machine. Intelligent Automation and Soft Computing, 32(1), 115-129.

[6] Vinodhini, V., Vishalakshi, A., Chandrika, G. N., Sankar, S., &Ramasubbareddy, S. (2022). Predicting vasovagal syncope for paraplegia patients using average weighted ensemble technique. Journal of Mobile Multimedia, 135-162.

[7] Kumar, K., Kumar, N., Kumar, A., Mohammed, M. A., Al-Waisy, A. S., Jaber, M. M., ... & Al-Andoli, M. N. (2022). Identification of Cardiac Patients Based on the Medical Conditions Using Machine Learning Models. Computational Intelligence and Neuroscience, 2022, 1-15.

[8] Verma, L., Srivastava, S., &Negi, P. C. (2016). A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. Journal of medical systems, 40(7), 1-7.

[9] Bodapati, J. D., Sajja, V. R. K., Mundukur, N. B., &Veeranjaneyulu, N. (2019). Robust Cluster-then-label (RCTL) Approach for Heart Disease Prediction. *Ingénierie des Systèmes d Inf.*, *24*(3), 255-260.

[10] Saikumar, K., Rajesh, V., & SK, H. A. (2020). CAB for heart diagnosis with RFO artificial intelligence algorithm. International Journal of Research in Pharmaceutical Sciences, 11(1), 1199-1206.

[11] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 26(2021), 1-8.

[12] Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

[13] Manne, R., &Kantheti, S. C. (2021). Application of artificial intelligence in healthcare: chances and challenges. Current Journal of Applied Science and Technology, 40(6), 78-89.

[14] Chen, J. I. Z., &Hengjinda, P. (2021). Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. Journal of Artificial Intelligence, 3(01), 17-33.

[15] Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.

[16] Nikhar, S., &Karandikar, A. M. (2016). Prediction of heart disease using machine learning algorithms. International Journal of Advanced Engineering, Management and Science, 2(6), 239484, 617-621.

[17] Kumar, M. N., Koushik, K. V. S., & Deepak, K. (2018). Prediction of heart diseases using data mining and machine learning algorithms and tools. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 3(3), 887-898.

[18] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. Heart Disease, 7(1), 129-137.