Study on Heart Disease Prediction Using Machine Learning Algorithms

Sunaina[1], Deepak[2]
[1]MTech, CSE, Student, Sri Sai College of Engineering and Technology, Badani, Pathankot, Punjab
[2]Assistant Professor, CSE, Sri Sai College of Engineering and Technology, Badani, Pathankot, Punjab

Abstract - One of the most crucial components of human life is healthcare. One of the worst diseases and a major hindrance to the lives of many individuals worldwide is heart disease. The traditional methods have limitations in that they do not generalize well to new data that is not present in the training set. A significant discrepancy between training and test accuracy points to this. The code performs an exploratory data analysis (EDA), including visualizing the distribution of continuous and categorical features and their relationship with the target variable (heart disease presence). Data preprocessing techniques such as one-hot encoding and Box-Cox transformation are employed to prepare the data for modeling. Three machine learning models – Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) – are trained and evaluated using stratified k-fold cross-validation. Hyper parameter tuning is performed using Grid Search CV to optimize model performance. Model evaluation metrics include precision, recall, F1-score, and accuracy. Results indicate that the Random Forest model achieves the highest recall for the positive class (heart disease), making it the most suitable model for this application. Confusion matrices are visualized for each model to assess their predictive performance. This work demonstrates the effectiveness of machine learning models in heart disease prediction and provides insights into the selection of appropriate models based on desired performance metrics.

Keywords: Healthcare system, Heart disease, Machine learning, RF, SVM and KNN.

I. INTRODUCTION

In the past few years, especially in the wake of the Covid epidemic, there has been a significant amount of research in the field of healthcare. According to the World Health Organization [1], cardiac disorders are among the most lethal conditions that claim the lives of the most people worldwide. Additionally, it has been shown that various forms of heart disease account for more than 24% of fatalities in India [2]. Therefore, a mechanism for early diagnosis must be created in order to stop the deaths that are being caused by cardiac disorders. There are techniques for identifying heart disorders, such as angiography, but they are expensive and subject to adverse effects in the body of the patient. This stops these methods from being widely used in nations with sizable populations of the impoverished. Healthcare organizations are also looking for clinical testing that may be done inexpensively and without intrusion. Organizations can better serve the needs of millions of people worldwide by developing a computer-based decision support system for the diagnosis of various diseases. Research in many fields, especially medicine, has benefited from the quick development of machine learning algorithms. The accessibility of extensive data on medical diagnoses has aided in the training of these algorithms. These methods can be used to create the clinical support system, which aids in cutting costs and improving accuracy [3]. The organization of this paper is as follows. Section II describes the background of machine learning; Section III briefly discusses its applications and the SVM, RF and KNN architecture, section IV shows the literature survey in section V, we describe the proposed methodology. Experiments and results are discussed in Section VI. Finally, conclusions are presented in Section VII.

II MACHINE LEARNING

The practical application of artificial intelligence, in which the user can programme clever algorithms to produce results with more accuracy and generate predictions within a reasonable range. The programme can use the learned model on new datasets after training on the provided data. The idea is comparable to data mining and predictive modeling techniques, which explore deeply into databases to identify

patterns. On the other hand, unsupervised learning is employed for more complicated tasks and doesn't need to be trained using the necessary outcome data. The target of unsupervised learning is to group those datasets into sensible classes [4].

III. SUPPORT VECTOR MACHINES (SVM)

A Support Vector Machine (SVM) is a powerful machine learning algorithm widely used for both linear and nonlinear classification, as well as regression and outlier detection tasks. SVMs are highly adaptable, making them suitable for various applications such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are particularly effective because they focus on finding the maximum separating hyper plane between the different classes in the target feature, making them robust for both binary and multiclass classification.

Random Forest Algorithm

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce over fitting, and provide reliable forecasts in different environments.

K-Nearest Neighbor (KNN)

KNN is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection. The K-NN algorithm works by finding the K nearest neighbors to a given data point based on a distance metric, such as Euclidean distance. The class or value of the data point is then determined by the majority vote or average of the K neighbors. This approach allows the algorithm to adapt to different patterns and make predictions based on the local structure of the data.

IV.LITERATURE SURVEY

In 2018, Iancu [6] has developed a Mediative fuzzy logic technique for dealing with incompatible information that offered a solution during the existence of contradiction. The main intention of this work was to construct an expert system on the basis of fuzzy logic for diagnosing the probable heart disease for a patient. This implemented system was considered as an expansion of the benchmark “Mamdani fuzzy logic controller” and has comprised of “44 types of single input–single output rules”. The system has operated with input having 11variables and output having one variable.

In 2019, Amin [7] have intended to identify the considerable “features and data mining approaches” that enhanced the prediction accuracy of cardiovascular disease. Prediction methods were implemented based on diverse amalgamation of features, and “seven classification approaches NN, k-NN, Naive Bayes, Decision Tree, SVM, LR, and Vote (a hybrid approach with LR and Naïve Bayes)”. The investigational outcomes have explained that the heart disease prediction method that introduced by means of determined noteworthy “features and the data mining technique” (i.e. Vote) has attained a superior prediction accuracy of heart disease.

In 2019, Tao[8] have focused on introducing a quick and precise automatic IHD detection/localization approach. Initially, the features were extracted and were classified under three groups called: information theory features, “time-domain features”, and “frequency domain features”. Subsequent to this, diverse machine learning classifiers like DT, KNN, XGBoost, and SVM were compared. Three classifiers were selected in accordance with their best performance for determining the IHD case and were applied with ensemble method and finally analyzed with their results.

In 2020, Escamila[29] have utilized a machine learning (ML) as a solution for diminishing as well as comprehending the side effects identified with

coronary illness. The target behind this research was to dimensionality reduce the features and to explore the features associated with the coronary illness. Further, for performing this investigation, the authors have collected the data from the "UCI Machine Learning Repository called Heart Disease". The dataset contains "74 highlights and a name that were approved by six ML classifiers".

In 2020, Dutta [30] have proposed a proficient "neural network with convolutional layers" in order to categorize the class-imbalanced clinical data in a significant manner. The data was curate from the "National Health and Nutritional Examination Survey (NHANES) with the goal of predicting the occurrence of Coronary Heart Disease (CHD)". The authors have developed a two-step approach: at the initial stage, they have utilized "least absolute shrinkage and selection operator (LASSO) based feature weight assessment followed by majority-voting based identification of important features". Then, a fully connected layer was used to homogenize the important features before passing the outcomes to the progressive convolutional stages.

V. PROPOSED WORK

- Research Objectives

The major research objectives are as follows:

1. Study of existing models and approaches for prediction of heart disease.
2. To evaluate the impact of hyperparameter tuning on model performance, particularly focusing on maximizing recall for heart disease cases.
3. To compare the performance of the selected models based on relevant metrics (precision, recall, F1-score, accuracy) and identify the model that best balances overall performance with high recall for heart disease prediction.
4. To develop a robust and reliable heart disease prediction model with high recall, minimizing the risk of missing potential cases.

- Methodology

As the count of data is increasing, it is more complex to do processing and analysis and more particularly, maintaining the e-healthcare data. Under this circumstance, the prediction model seems to be more complicated when it comes on disease prediction. This proposal intends to propose a new heart disease prediction model that includes two phases: "Feature selection and Classification". In the first phase, we will implement feature selection algorithm. After this feature selection next process is classification.

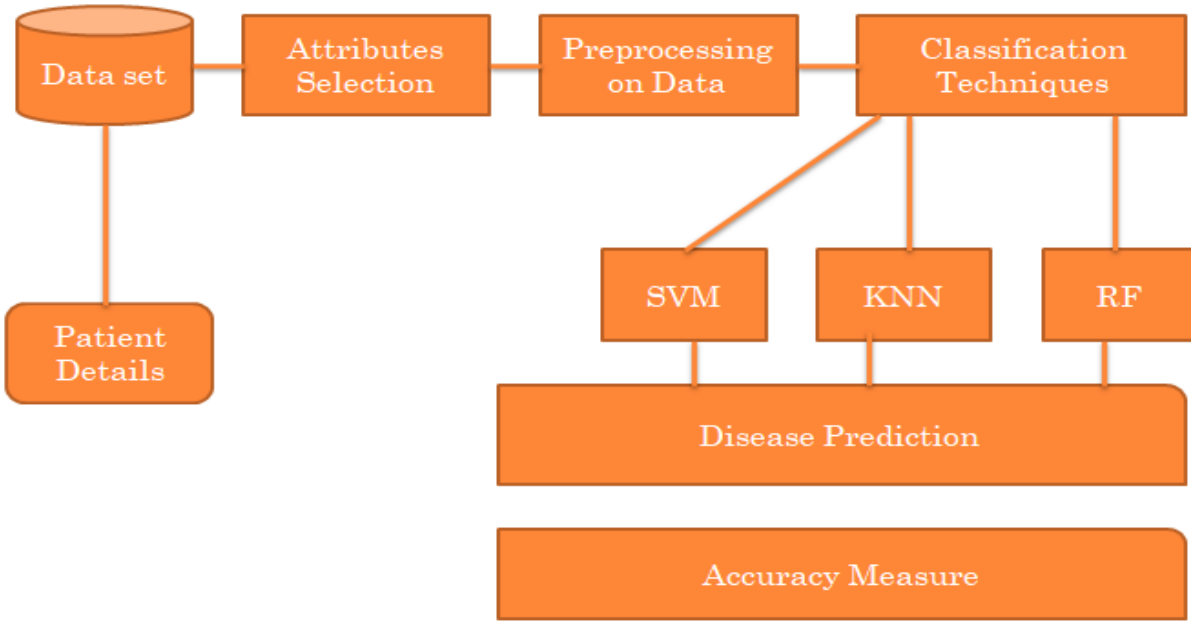

Figure 1: Proposed flowchart for the Methodology

The code follows a structured approach to predict heart disease using machine learning. The methodology can be divided into the following steps:

1. Data Loading and Exploration (EDA):

The code begins by loading the heart disease dataset from a CSV file using Pandas.EDA is performed to understand the data's characteristics. Histograms and summary statistics are used to examine the distribution of continuous features. Bar charts are used to visualize the distribution of categorical features. Relationships between features and the target variable (heart disease presence) are explored using bar plots and kernel density plots. The code checks for missing values in the dataset.

2. Data Preprocessing:

Categorical Feature Encoding: Categorical features are converted into numerical representations using one-hot encoding. This creates new binary columns for each category within a categorical feature.

Continuous Feature Transformation: The Box-Cox transformation is applied to continuous features to improve their normality and address potential skewness. This transformation helps optimize the performance of certain machine learning algorithms.

Data Splitting: The dataset is divided into training and testing sets using train_test_split. This ensures that the model is evaluated on unseen data. Stratification is used to maintain the proportion of heart disease cases in both sets.

3. Model Selection and Training:

Three machine learning models are chosen for prediction:

Random Forest: An ensemble method that combines multiple decision trees. Support Vector Machine (SVM): A powerful algorithm that finds an optimal hyperplane to separate data points into different classes. K-Nearest Neighbors (KNN): A simple algorithm that classifies data points based on the majority class among their nearest neighbors.

4. Hyper parameter Tuning:

Grid Search CV: This technique is used to systematically search for the best combination of hyper parameters for each model. It evaluates model performance across a range of hyper parameter values and selects the combination that yields the highest performance metric (in this case, recall for the positive class).

Stratified K Fold: This cross-validation method ensures that the class distribution is maintained across different folds during hyper parameter tuning.

5. Model Evaluation:

Classification Report: This report provides key metrics such as precision, recall, F1-score, and accuracy for each model.

Confusion Matrices: These matrices visualize the performance of each model by showing the number of true positives, true negatives, false positives, and false negatives.

Recall Comparison: The recall for the positive class (heart disease) is used as the primary metric to compare the performance of the three models. A bar chart is generated to visually represent the recall scores.

Overall, this methodology aims to develop a robust and accurate heart disease prediction model. Fig. 3 illustrates the proposed feature extraction & classification framework. Authors using the dataset as http://archive.ics.uci.edu/ml/datasets/statlog+(heart)"that is downloaded from Statlog (Heart) Data Set.

VI.RESULTS

After performing the machine learning approach for testing and training we find that accuracy of the SVM is much efficient as compare to other algorithms. Accuracy should be calculated with the support of confusion matrix of each algorithms as shown in Fig.6 and Fig.7 here number of count of TP, TN, FP, FN are given and using the equation(2) of accuracy, value has been calculated and it is conclude that SVM is best among them with 97% accuracy and the comparison is shown in TABLE.2.

Algorithm	Accuracy
Support Vector machine	97%
RF	88%
k-nearest neighbor	85%

Confusion matrix: An analysis of a machine learning model's performance on a set of test data is summarized by a confusion matrix. It is frequently used to assess how well categorization algorithms work. These models try to forecast a categorical label for each input event. The matrix shows how many true positives (TP), true negatives (TN), false positives

(FP), and false negatives (FN) the model generated using the test data. The relationship among false positive rate and the real positive rate is plotted on the ROC curve. The ROC curve for the suggested architecture is extremely close to the ideal curve, demonstrating the architecture's strong performance on the test set. Regarding each of these metrics for performance, the suggested architecture works admirably. Additionally, fresh data that is not part of the train or validation sets is used to validate the suggested design. Additionally, it performs well with fresh data. Additionally included is each feature's statistical significance for classification.

VII.CONCLUSION

Heart disease is a serious condition that can have severe complications including heart attacks. Data mining and machine learning techniques are crucial because they have the ability to accurately anticipate the occurrence of diseases. In this study, we present a method called Cardio Help that uses a machine learning algorithm called SVM (Support Vector Machine) to estimate the likelihood that a patient has cardiovascular disease. Heart is one of the essential and vital organs of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset whose attributes are shown in TABLE.1 and on the basis of confusion matrix, we find SVM is best one. For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

REFERENCE

[1] Santhana Krishnan J and Geetha S, “Prediction of Heart Disease using Machine Learning Algorithms” ICIICT, 2019.

[2] Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. KailashDevadkar, “Prediction of Heart Disease using Machine Learning”, Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2018.

[3] Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.

[4] Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using DataMining Techniques: A survey” International Journal of Advanced Research in Computer Science, IJARCS 2015-2019.

[5] Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication and Automation (ICCCA), 2018.

[6] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of HeartDisease Using K- Nearest Neighbor and Genetic Algorithm,” ProcediaTechnol., vol. 10, pp. 85–94, 2013.

[7] S. Kumra, R. Saxena, and S. Mehta, “An Extensive Review on Swarm Robotics,” pp. 140–145, 2009.

[8] Hazra, A., Mandal, S., Gupta, A. and Mukherjee, “A Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review” Advances in Computational Sciences and Technology, 2017.

[9] Rana Jagdev Singh “IoT Based Computer Generated Electromagnetic Radiation Detector Monitoring System” International Journal of Innovative Research in Technology” 212-214 Vol-11 Issue 6 November 2024

[10] Patel, J., Upadhyay, P. and Patel, “Heart Disease Prediction Using Machine learning and Data Mining Technique” Journals of ComputerScience & Electronics, 2016.

[11] https://archive.ics.uci.edu/ml/datasets/Heart+Disease