Index No: TL-26-36

Heart Disease Prediction Using Machine Learning Techniques: A Comparative Analysis

L. Gamage (1st Author)
Department of Software Engineering
University of Kelaniya
Kelaniya, Sri Lanka

Postgraduate Institute of Science
University of Peradeniya
Peradeniya, Sri Lanka
lgama191@kln.ac.lk

***Abstract* — In today's world, heart disease is one of the leading causes of death. In clinical data analysis, predicting heart disease is a difficult task. Machine Learning (ML) helps assist with the decision-making and prediction of large volumes of data generated by the healthcare industry. The main goal of this study is to find the best performance model and compare machine learning algorithms for predicting heart disease. This work applies supervised machine learning algorithms, namely Logistic Regression, Support Vector Machine, K-Nearest Neighbor, and Random Forest, to the Cleveland Heart Disease dataset to predict heart disease. Our experimental analysis using preprocessing steps and model hyperparameter tuning, Logistic Regression, Support Vector Machine, K- Nearest Neighbor and Random Forest achieved 90.16%, 86.89%,86.89%, and 85.25%, accuracies respectively. As a result, Logistic Regression classification outperforms other machine learning algorithms in predicting heart disease.**

Keywords* — *Machine Learning, Heart Disease Prediction, Logistic Regression, Support Vector Machine, K- Nearest Neighbor, Random Forest

I. INTRODUCTION

The heart is the most important organ in the human body. It is at the center of the human circulatory system, a network of blood vessels that delivers blood to every part of the human body. Blood carries oxygen and other essential nutrients that all body organs need to stay healthy and work properly. If it fails to function correctly, then the brain and various other organs will stop working, and within a few minutes, a person can die. In today’s world, people are getting very busy in their lives and work so much that they do not have time to take care of themselves. Due to this hectic lifestyle, they experience stress, anxiety, depression, and many more conditions. These factors contribute to people being ill and having severe diseases. There are many diseases such as cancer, diabetes, heart diseases, tuberculosis, Etc., which lead to the death of people each year. Due to a hectic lifestyle and not having good food habits, heart-related diseases increase among people more than any other disease. Also, some factors, diseases, and habits influence increasing the risk of having heart diseases. They are Age, Family History, Diabetes, High blood pressure, high cholesterol, being obese, depression, hypertension, smoking, or too much drinking and physical inactivity. So, taking care of the heart as well as other organs becomes a challenge in this era. Apart from the factors, there are different types of heart disease, such as Angina, Arrhythmia Fibrillation, Congenital heart disease, Coronary artery disease, Myocardial infarction, and Heart failure. To reduce the heart disease caused death, it is mandatory to early detection of heart disease and treatment on time. So, the Importance of Prediction of heart-related diseases can significantly impact the medical field and people’s lives. Now day’s heart disease is the main reason for deaths throughout the world. Many people die from heart disease than from any other disease. According to the World Health Organization, for the past 20 years, heart disease has been the leading cause of death at the global level[1]. The number of deaths from heart disease increased by more than 2 million since 2000, to nearly 9 million in 2019. Heart disease currently accounts for 16% of total deaths from all causes. Even though heart disease is the riskiest disease globally, people are unaware of the risks and symptoms of heart disease. Therefore, the prediction of heart disease is a significant concern of humankind. The foremost challenge facing healthcare organizations is the provision of quality services at a reasonable cost. Quality service indicates diagnosing patients correctly and administering effective treatments. Most of the clinical decisions are made based on the Doctor’s perception and practice rather than the hidden knowledge on the patient database. This practice may lead to mistakes and extreme medical expenses, which affect the quality of health care services. It is frequently difficult for medical practitioners to predict heart diseases as it requires experience and knowledge, which is very difficult to accomplish. In that case, machine learning techniques help the healthcare industry and the professionals in diagnosing heart diseases based on patients’ data.

In the medical domain, many medical datasets help make compelling predictions. These data can be exploited using machine learning techniques to extract hidden information from hidden patterns of datasets. Moreover, this extractive data will help to predict the medical diagnosis. The collected medical data are massive in size, and it can be noisy. This data which is too complicated for the human mind to understand can be easily explored using machine learning techniques. The future predictions will also help the doctors to diagnose the disease by discovering previous dataset patterns and taking the right step to treat the patients. It will save humankind and increase the quality of health care services. Several research studies have been conducted on predicting heart disease. They used several machine learning techniques to predict heart disease and achieved different results using different methods. [4]Their work on "Cardiovascular Disease Forecast using Machine Learning Paradigms" used machine learning algorithms and achieved 86.25% accuracy from the logistic regression model, giving the best accuracy among all four models. [9] Their work on" Heart Disease Prediction Using Machine Learning Algorithms" summarized the recent research with comparative results that have been done on heart disease prediction and also make analytical conclusions. Their experimental results show that the Decision Tree Classifier algorithm has the most precise and significant result compared to the others algorithm. Also, they mentioned that the predictions that are carried out with large datasets make better accuracy.

This study is for analyzing and finding an appropriate model for the heart disease data. Machine learning models are used to predict heart disease while conducting a comparative analysis on models. In this work, supervised machine learning algorithms, namely Logistic regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF), are used to predict heart disease and compare their performance.

II. OBJECTIVES

The goal of this study is to find out the best model that predicts whether the patient has heart disease. This model will provide important insights to doctors who can then adapt their diagnosis and treatment per-patient basis. This study is to be done on supervised machine learning classification.

The research objectives (ROs) of the research work are stated in the following.

RO1: To predict heart disease using selected four Supervised machine learning techniques and analyze the models; Logistic regression, Support Vector Machine, K-nearest neighbor, and Random Forest.

RO2: To find the best model that predicts heart disease more accurately among all the models.

RO3: To conduct a comparative analysis of machine learning models based on their performances.

III. METHODOLOGY

This section illustrates various resources and approaches that used in this study. Primarily, the description of the dataset is provided to understand how to work on it, followed by the preprocessing steps involved. Finally, the methodologies used for the experiment in this study are discussed.

In this section, the research materials and methodologies are presented and discussed in brief.

A. Dataset Description

In this work-study, the Cleveland Heart Disease dataset [4] has been collected from the UCI machine learning repository that has been used for both training and testing purposes. The dataset is a collection of medical analytical reports with values for 76 attributes and 303 rows, but this work considers a feature subset of 14 numerical valued attributes. The output level has two classes, where 0 represents not having heart disease, and one represents having heart disease. The information on the heart disease dataset is given

Table, where the attribute name and description are presented.

B. Data Preprocessing

In this step, data preprocessing is applied to identify the missing values, process the noisy, incomplete, irreverent, and inconsistent values, and remove some attributes' redundancy. Then separation, feature scaling, and normalization are performed to find the standard format of data. After data preparation, the dataset is divided into a training set (80% of data) and a test set (20% of data).

Table 1. Attributes of the dataset and their description

Attribute Name	Attribute Description
age	age in years
sex	1=Male,0=Female
cp	Chest pain type. 0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina
trestbos	Resting blood pressure (in mmHg)
chol	Serum Cholesterol (in mg/dl)
fbs	Fasting blood sugar> 120 mg/dl.1 = true; 0 = false
restecg	Resting Electrocardiography results. 0: showing probable or definite left ventricular hypertrophy by Estes' criteria, 1: normal, 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
thalach	Maximum heart rate achieved
exang	Exercise-induced angina ;(1 = yes; 0 = no)

Attribute Name	Attribute Description
oldpeak	Depression induced by exercise relative to rest
slope	The slope of the peak exercise segment. 0: down sloping. 1: flat. 2: upsloping
ca	Number of major vessels colour by fluoroscopy that ranges between 0 and 3
thal	Results of the blood flow were observed via the r adioactive dye. 0=Normal (Thallium test.) 1=fixed defect 2=reversible defect
target	0 = No heart disease, 1 = heart disease

C. Model Generation

In this stage, machine learning algorithms are applied to the training set to develop different classification models. Models are developed using four machine learning algorithms named LR, KNN, SVM and RF classifiers. Then using these four generated models, the test set is classified and evaluated the performance.

In machine learning, hyperparameter tuning is one of the most significant research issues. If the hyperparameters are tuned or optimized, then it is considered that the machine learning algorithms can give better performance. Models are developed with the help of grid search, and cross-validation approaches, hyperparameters are optimized and tuned. It considers cross-validation to guide the performance metrics. Grid search is an exhaustive search that can exercise to compute the optimal values of hyperparameters. It can build a model that generates every parameter combination and stores each combination of the model. The efforts and resources can be saved using this search. Then with the tuned parameters, the LR, KNN, SVM and RF classifier models are generated. After the generation of the classification model, the test set is applied to the proposed model with the tune hyperparameter and evaluated the test set's performance.

The flowchart of the model is shown in Fig. 1 below. It summarizes the steps for our proposed method.

D. Machine Learning Algorithms

This section explains the supervised algorithms of Machine learning that are used in this work. In the model generation process, Logistic regression (LR), K nearest neighbor (KNN), Support vector machine (SVM) and Random Forest (RF) classifiers are used as machine learning algorithms.

1) Logistics Regression

Logistic regression [5] is a supervised learning machine used for regression and classification problems. Logistic regression uses probability to predict the classification of categorical data. It is mainly used for a binary classification problem and a logistic function to predict a binary dependent variable [6].

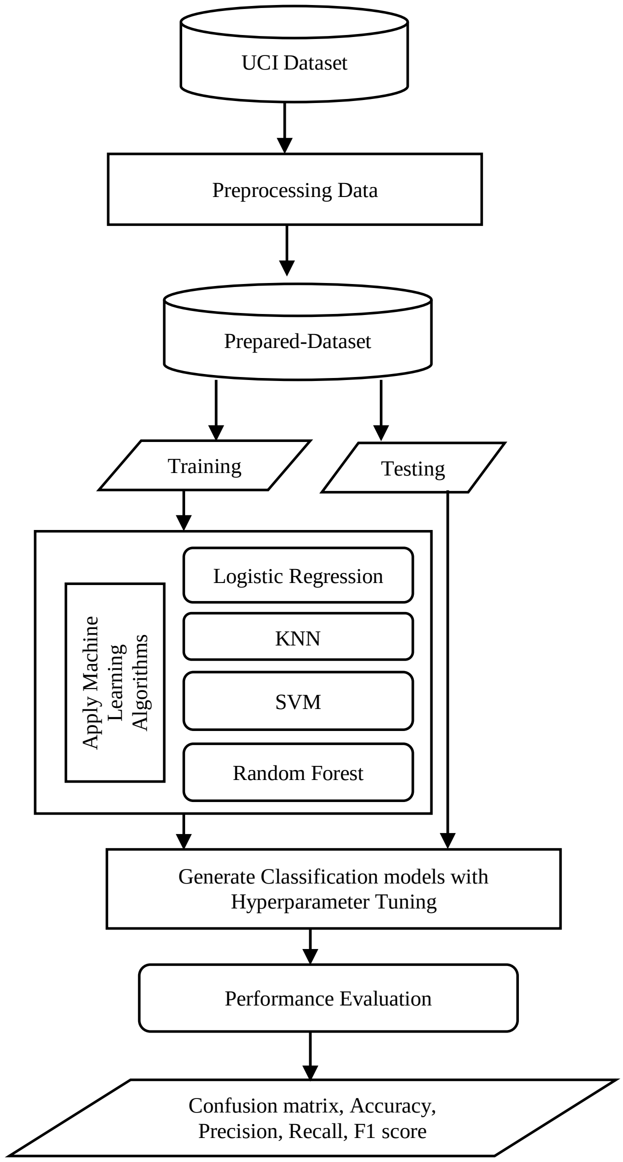
UCI Dataset
Preprocessing Data
Prepared-Dataset
Training
Testing
Apply Machine Learning Algorithms
Logistic Regression
KNN
SVM
Random Forest
Generate Classification models with Hyperparameter Tuning
Performance Evaluation
Confusion matrix, Accuracy, Precision, Recall, F1 score

Fig. 1. Flowchart of proposed work

2) KNN

K-Nearest Neighbor (KNN) [2] is also a supervised classification algorithm. It predicts the target class based on how similar that data is from other provided training data labels to the model. KNN is used widely in the machine learning classification problem. It is simple to understand and generates a non-parametric model that is applied to practical problems. It is a lazy learner or instance-based learner, which depend on the distance. It works well but does not learn any classification rule[5].

3) SVM

Support Vector Machine [6] is a prevalent supervised machine learning technique (having a pre-defined target variable) that can be used as a classifier and a predictor. For classification, it finds a hyper-plane in the feature space that differentiates between the classes. In SVM, a hyper-plane is created margin as wide as possible to separate different types of data or keep similar data of one type at one side and similar data of another type of data at another side of the margin [5].

4) Random Forest

Random forest (RF) [2]is a supervised machine learning algorithm. As the name predicts, it is a forest of randomly generated decision trees. It uses an approach bagging, where various learning models are combined to improve the overall results. To perform the bagging operation, it produces manifold decision trees and synthesizes them together to obtain a refined result. It is one of the finest machine learning algorithms. It uses a random subset of features by splitting a node to get the best feature that contributes the most to building the model.

E. Performance Evaluation

Performance evaluation of the proposed work is done based on the following measures.

1) Confusion Matrix

Confusion Matrix [2] is a matrix that is used to evaluate the performance of a model. The four terms associated with the confusion matrix which is used to determine the performance matrices are:

True Positive (TP): An outcome when the model correctly predicts the positive class

True Negative (TN): An outcome when the model correctly predicts the negative class

False Positive (FP): An outcome when the model incorrectly predicts the positive class

False Negative (FN): An outcome when the model incorrectly predicts the negative class

In this step, models evaluate the performance of the training set and test set and find the confusion matrix. Then the performance metrics of these two models have been calculated and assessed in terms of accuracy, precision, recall, and F1 Score with the help of the confusion matrix. The mathematical expressions of accuracy, precision, recall and F1-Score are shown in Equation (1), Equation (2), Equation (3) and Equation (4) respectively.

2) Accuracy

Accuracy is the ratio of the number of correct predictions given by the model to the total number of instances[2].

$$\text{Accuracy} = (TP+TN) / (TP+TN+FP+FN) \qquad (1)$$

3) Precision

Precision in this work measures the proportion of individuals predicted to be at risk of developing heart disease and had a risk of developing heart disease.

$$\text{Precision} = TP / (TP+FP) \qquad (2)$$

4) Recall

Recall, in this work, measures the proportion of individuals that were at risk of developing heart disease and were predicted by the algorithm to be at risk of developing heart disease.

$$\text{Recall} = TP / (TP+FN) \qquad (3)$$

5) F1-Score

F1 Score is the harmonic mean of precision and recall.

$$\text{F1 score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \qquad (4)$$

IV. RESULTS AND DISCUSSION

A. Results of Data Preprocessing

The heart disease dataset consists of 303 samples with 14 attributes, where 138 are healthy (0) instances and 165 instances (1) having heart disease. In the preprocessing step, the statistical operations have been performed to identify and remove the missing values and to find the maximum, minimum, mean, and standard deviation of each feature set.

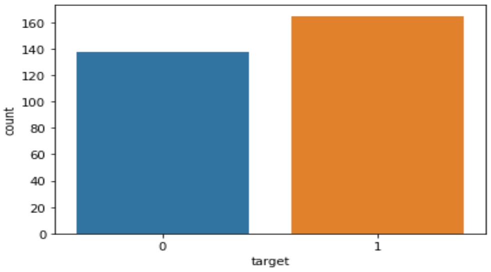

Fig.2.Count of target Variables

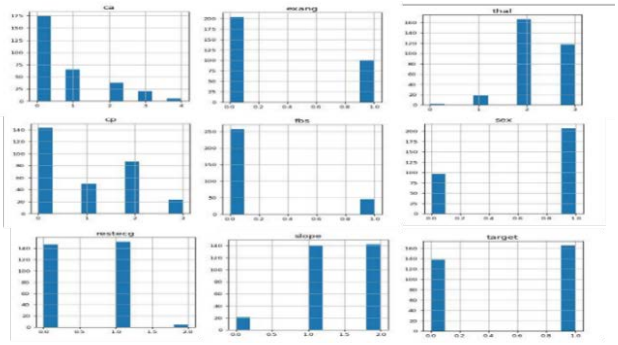

Fig.**3**. Histogram of categorical valued attributes

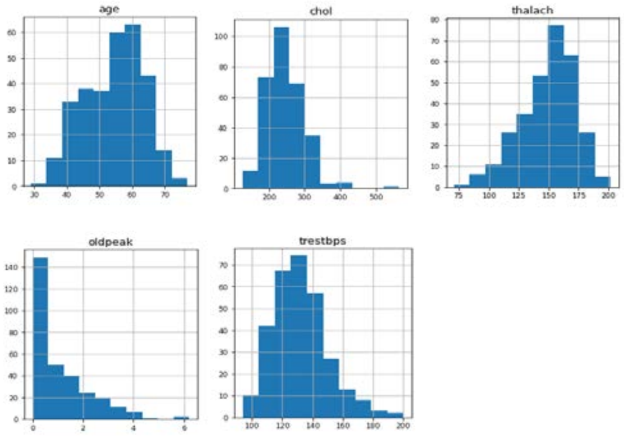

Fig. **4.** Histogram of continuous-valued attributes

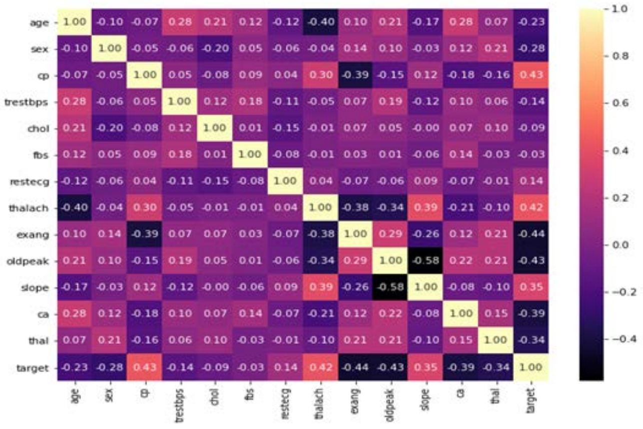

Fig. 5. The heat map for correlation features of the heart disease dataset

Consequently, several non-heart disease and heart disease patients are illustrated in **Fig. 2;** they are 138 and 165. This further confirms that the two classes are almost balanced, and it is good to proceed with data preprocessing. Then the histogram of categorical and continuous features has been plotted for easy and better understanding. The histogram plots are presented for the pattern and frequency distribution of categorical and continuous measurements of data. The distribution of each feature value is shown in Fig. and Fig. **4** as a histogram plot. It can help to identify the trend and patterns of data to understand the distribution of features.

The histogram plot of continuous variables shows how each feature and label is distributed along with different ranges, which further confirms the need for scaling data. The discrete plots in figure 5 show that each of these is categorical variables, and it confirms the need of converting them into numerical variables before training models.

Fig. 5 represents the heat map, which describes the co-relation among the features of the heart disease dataset. Here, different colors have been used to represent the values on the two-dimensional surface. It's clear that no one feature has a particularly strong relationship with the target variable, and categorical-valued attributes are more concentrated than continuous-valued attributes. The heat map of the heart disease dataset shows the hierarchical clustering and a general view of numeric data. After investigating the dataset, the categorical valued attributes have been converted into dummy attributes. Then, centering and scaling operations have been performed to standardize each feature by computing the relevant statistics on the dataset. The resultant dataset has been divided into a training set and a test set, with an 80% and 20% split.

B. Experiment Results of the Classifiers

1) Performance Evaluation of the Classifiers with Default Hyper-Parameters

In this experiment, the machine learning algorithms are applied with the default parameters. Table 2 shows the result of this system. In the training phase, Logistic Regression is fitted and executed the model with parameters of C=1 and solver= 'liblinear' and found 86.78%,86.33%,90.22% and 88.23% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this LR model and provides 90.16%, 93.30%,87.50% and 90.32% of accuracy, precision, recall, and F1 score, respectively. Again, in the training phase, KNN is fitted and executed the model with the parameters of no. of neighbor=7 and weights= 'uniform' and found 82.23%, 82.60%, 85.71% and 84.13% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this KNN model and provides 86.89%,90.00%,84.37%and 87.09%, of accuracy, precision, recall, and F1-score respectively

In another training model, SVM is fitted and executed with the parameters of C= 1.0, gamma= 0.01, and 'rbf' kernel, and found 82.23%,82.14%,86.46% and 84.25% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this SVM model and provides 83.61%, 89.28%, 78.13% and 83.33% of accuracy, precision, recall, and F1 score, respectively. In the last phase of training, Random Forest is fitted and executed with the parameter estimators = 100, min_samples_leaf=1, min_samples_split=2 and found 100%, 100%, 100%, and 100% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this Random Forest model and provides 81.97%,86.20%,78.12% and 81.97% of accuracy, precision, recall, and F1-score, respectively.

Table 2. Performance of evaluation and comparison of classification models on the training set and the test set

Algorithm	Training Dataset				Test Dataset			
	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)
LR	86.7	86.3	90.2	88.2	90.1	93.3	87.5	90.3
KNN	82.2	82.6	85.7	84.1	86.8	90.0	84.3	87.0
SVM	82.2	82.1	86.4	84.2	83.6	89.2	78.1	83.3
RF	100	100	100	100	81.9	86.2	78.1	81.9

2) Performance Evaluation of the Classifiers with Tuned Hyper-Parameters

In this experiment, the Grid search is used to find the optimal hyperparameters. After tuning the hyperparameters, the classification models are generated. Table 3 shows the result of the proposed system. In the training phase, Logistics Regression is fitted and executed with the tuned hyperparameters of C=1.08, penalty =' l2' and solver= 'liblinear' and found 86.78%, 86.33%, 90.22% and 88.23% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this LR model and provides 90.16%,93.33%,87.50% and 90.32% of accuracy, precision, recall, and F1 score, respectively. Again, in the training phase, KNN is fitted and executed with the tuned hyperparameters of no. of neighbor=5 and weights=' uniform', metric: 'manhattan' and found 88.43%, 87.76%, 91.72% and 89.70% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this KNN model and provides 86.89%,87.50%,87.50%, and 87.50% of accuracy, precision, recall, and F1 -core, respectively.

In another training model, SVM is fitted and executed with the tuned hyperparameters of C= 2.0, gamma= 0.1 and RBF kernel, and found 92.98%,92.03%,95.48% and 93.72% of accuracy, precision, recall, and F1 score respectively. The test set is predicted on this SVM model and provides 86.89%, 90.00%,84.37% and 84.37% accuracy, precision, recall, and F1 score, respectively. In the last phase of training, RF is fitted and executed with the tuned hyperparameters of 1000 no of estimators, two minimum samples of the leaf, and five minimum samples the split, maximum depth = 50, maximum features = square, and found 96.69%, 94.96%,99.24% and 97.05% of accuracy, precision, recall, and F1- score respectively. The test set is predicted on this RF model and provides 85.25%,89.65%,81.25% and 85.24% of accuracy, precision, recall, and F1-score, respectively.

Table 3. Performance of evaluation and comparison of classification models with a hyperparameter tuning approach on the training set and the test set

Algorithm	Training Dataset				Test Dataset			
	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1- score (%)
LR	86.7	86.3	90.2	88.2	90.1	93.3	87.5	90.3
KNN	88.4	87.7	91.7	89.7	86.8	87.5	87.5	87.5
SVM	92.9	92.0	95.5	93.7	86.8	90.0	84.3	84.3
RF	96.6	94.9	99.2	97.0	85.2	89.6	81.2	85.2

3) Performance Comparison of the Classifiers with default and tuned Hyper- Parameters

Table **4** represent the performance comparison between without and with the hyperparameters tuning approach of four machine learning algorithms.

Table 4. Accuracy comparison

Algorithm	Accuracy (%) of Training Dataset		Accuracy (%) of Test Dataset	
	Without parameter tuning	With Hyperparameter tuning	Without parameter tuning	With Hyperparameter tuning
LR	86.7	86.7	90.1	90.1
KNN	82.2	88.4	86.8	86.8
SVM	82.2	92.9	83.6	86.8
RF	100.0	96.6	81.9	85.2

These comparisons show that the models with tuned hyperparameters provide better accuracy results. It also noticed that the LR and KNN accuracies remain the same while SVM and Random Forest have increased their test accuracy.

V. CONCLUSION

Heart disease is one of the significant deaths anywhere in the world. Early detection of heart diseases will increase the survival rate; hence this work-study is intended to predict whether the patient has heart disease or not with the help of clinical data, which will assist the diagnosis process. However, machine learning techniques are helpful to predict the output from existing data. The results of this study confirm the application of machine learning algorithms in the prediction and early detection of heart disease. To our best understanding, the models built with tuned hyperparameters exhibit better accuracy than the models with default hyperparameters. The prediction accuracy of our proposed models reaches 90.16 % in heart disease detection using Logistic Regression, 86.89% in Support Vector Machine classifier,86.89% using KNN classifier and 85.25% using Random Forest Classifier. The experimental results show that the Logistic Regression algorithm predicts heart disease with the highest performance measures in terms of accuracy of 90.16% and other evaluating metrics.

The future aspect of this study will be to implement the model with the feature selection approach. Furthermore, new feature selection methods can be developed to get a broader perception of the significant features to increase the performance of heart disease prediction.

REFERENCES

[1] "WHO reveals leading causes of death and disability worldwide: 2000-2019." https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019 (accessed May 28, 2021).

[2] "UCI Machine Learning Repository: Heart Disease Data Set." https://archive.ics.uci.edu/ml/datasets/Heart+Disease (accessed May 31, 2021).

[3] R. Katarya and S. K. Meena, "Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[4] S. Islam, N. Jahan, and M. E. Khatun, "Cardiovascular Disease Forecast using Machine Learning Paradigms," in *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Mar. 2020, pp. 487–490. doi: 10.1109/ICCMC48092.2020.ICCMC-00091.

[5] E. Kabir Hashi and M. Shahid Uz Zaman, "Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction," *Journal of Applied Science & Process Engineering*, vol. 7, no. 2, 2020.

[6] V. v. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques: A survey," *International Journal of Engineering and Technology(UAE)*, vol. 7, no. 2.8 Special Issue 8, pp. 684–687, 2018, doi: 10.14419/ijet.v7i2.8.10557.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[8] Galgotias University. School of Computing Science and Engineering, Institute of Electrical and Electronics Engineers. Uttar Pradesh Section, and Institute of Electrical and Electronics Engineers, *2018 4th International Conference on Computing Communication and Automation (ICCCA).*

[9] V. Rachapudi, S. S. Vaddi, R. R. Karumuri, and S. Sripurapu, "Heart disease prediction using machine learning algorithms," *International Journal of Recent Technology and Engineering*, vol. 7, no. 5, pp. 805–809, Feb. 2019, doi: 10.32628/cseit206421.

[10] S. Anitha and N. Sridevi, "HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES," 2019. [Online]. Available: www.ijaconline.com,