e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

HEART DISEASE PREDICTION AND COMPARATIVE ANALYSIS USING SUPERVISED MACHINE LEARNING ALGORITHMS

Jomy Mary Mathew[*1]

[*1]Student, Department Of Information Technology, B.K Birla College Of Arts, Science And Commerce (Autonomous) Kalyan, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS30238

ABSTRACT

The heart is the most vital organ in the human body. Millions of people worldwide pass away due to diseases that affect the human heart. To prevent this, it is necessary to detect the disease beforehand. Prediction of such cardiovascular disease is a highly challenging task. Prediction of heart disease needs accuracy, preciseness, and correctness. Any fault or ignorance in this can lead to life-threatening situations. Machine Learning is the most nowadays used technology for such types of predictions. In this paper, we find the accuracy by using machine learning algorithms for predicting the possibility of heart disease and also by comparing the accuracy of the algorithms. Machine Learning Algorithms such as random forest, support vector machine (SVM), Decision tree, KNN, and logistic regression are proposed in this paper. All these algorithms are used to predict whether the patient has heart disease or not on the UCI dataset. For checking performance evaluation various methods like precision, recall, and f1 are used. The main objective is to find out a suitable ML technique that is accurate for the prediction of heart disease.

Keywords: Heart Disease, Machine Learning, Classification Algorithms, Random Forest, Logistic Regression, Accuracy.

I. INTRODUCTION

The heart is the most vital organ in a human being. It is the one that is responsible for the circulation of blood throughout the body. Any defect in the heart can lead to distress in the other parts of the body. It can also lead to life-threatening situations for the person. So, it becomes a major thing for predicting such types of diseases as soon as possible. Nowadays even when technology is on a rising scale, still many people are unknowingly dying of such diseases due to the late knowledge of the disease. According to World Health Organization, more than 10 million die due to heart diseases every single year around the world. A healthy lifestyle and the earliest detection are the only possible ways to prevent heart-related disease [5].

Thinking of prediction-based research Machine Learning (ML) is one of the most efficient technologies. ML is a branch of Artificial Intelligence (AI) which is one of the broad areas of learning where machines emulate human abilities, machine learning is a specific branch of AI [12]. Machine Learning can be categorized into supervised, unsupervised, and reinforcement. For this paper, we have focused on supervised machine learning algorithms. Algorithms such as Logistic regression, decision tree, random forest tree, support vector machine, and KNN are used on the UCI Cleveland dataset. This UCI Cleveland dataset is split into both a training set and a test set.

In this paper, the model calculates the accuracy of five different machine learning algorithms on the UCI dataset, and also to evaluate the performance of these algorithms, precision, recall, and F1 score are also analyzed. Machine Learning can be used to detect, diagnose, and predict many disorders in the medical industry. The primary aim of this study is to give the cardiologist a tool to detect cardiac problems at the initial stages [15].

II. LITERATURE REVIEW

There is ample work that has been carried out in this field. This paper has proposed a hybrid HRFLM approach that combines the characteristics of Random Forest and Linear Method. By using the ML algorithms like decision tree, and SVM and combining the random forest model with a linear model, almost 88.7% of accuracy has been predicted on the UCI dataset [2].

Sonam Nikhar and A.M. Karandikar have presented a paper regarding the heart disease prediction system with different classifier techniques for prediction of the heart disease. The paper shows a comparison between the naive bayes classifier and decision tree and came up with the result that the decision tree has better accuracy

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

compared to naive bayes. Also, for improving the naive bayes classifier they have used the trees constructed by C4.5, and this technique is called as Selective naive bayes classifier [4].

Jaymin Patel et al. have proposed a paper regarding heart disease prediction using ML and data mining techniques. The paper has compared the J48 tree technique to ML algorithms like LMT and Random Forest and came up with an outcome as J48 had the most accuracy i.e., around 56.76% compared to others [8].

Theresa Princy. R, et al, have proposed a survey including different classification algorithms used for prediction of the heart disease. A prediction method using the KNN algorithm and ID3 algorithm has been proposed in the paper. Using both the algorithms the risk of heart disease was detected and the accuracy level for different attributes.

Nagaraj M. Lutimath, et al, have performed a heart disease prediction using classification algorithms like SVM and naive bayes. SVM algorithm with radial kernel has better accuracy than the naive bayes algorithm. For performance analyses of the dataset Mean Absolute Error (MAE), Sum of Squared Error (SSE), and Root Mean Squared Error (RMSE) are performed.

All the above research papers lead to finding a better ML algorithm that has higher accuracy than compared to others, for helping in predicting heart disease using attributes in a dataset.

III. METHODOLOGY

The proposed model predicts the possibility of heart disease using supervised machine learning algorithms on the UCI Cleveland dataset. The paper also compares the accuracy of these algorithms and validates using methods like Precision, Recall, and F1 Score. The model can be used to check which algorithm is the most efficient one among others. The data from the dataset is fed into the algorithm using a python programming language which later on calculates the accuracy which can be understood using data visualization technique, i.e., confusion matrix.

1. Data Collection

The dataset used in the model is of UCI Cleveland from the Kaggle website. The dataset consists of 304 records and 14 attributes where all attributes are numeric-valued. This dataset is split into a training set and a test set in which 80% is taken as training and 20% as test data. As the dataset might contain missing data, the model uses preprocessed data that is already clean and contains no missing data.

2. Data Preprocessing

Data Preprocessing refers to the steps applied to the dataset before implementing the machine learning algorithms to the dataset [9]. A raw dataset can be unclean and have many errors, missing data, noise, etc. which cannot be suited for the process. So, the data must be preprocessed before using these algorithms. Various data preprocessing methods include Data cleaning, data integration, data normalization, feature selection, etc. For this analysis, a preprocessed dataset has been used. Table 2 shows all the attributes involved in the dataset and each attribute are in the form of numerical values. Figure 1 shows the proposed model used in the paper. Initially after the data preprocessing the data divided into both test and train set. Classification machine learning algorithms are used on train set for training the model. After that the dataset is tested using the test set and checks whether the model is correctly predicting the outcomes, then checks the accuracy of the outcome.

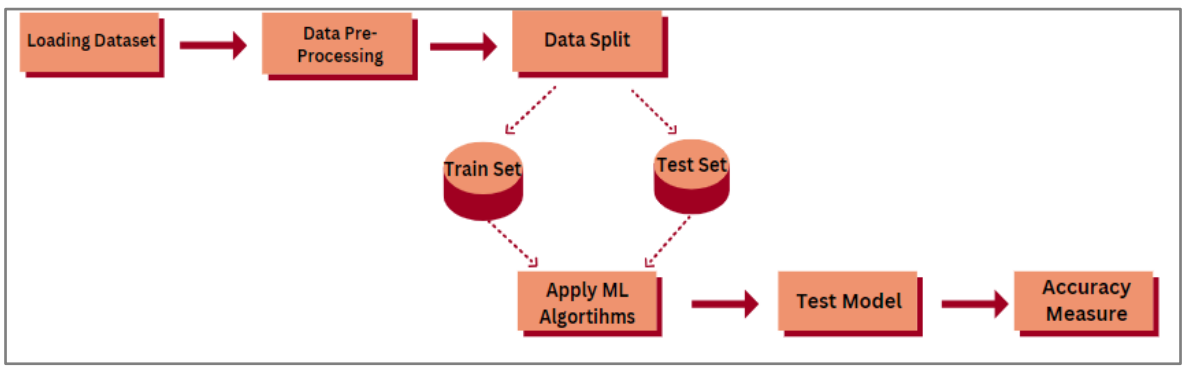

Figure 1: Machine Learning Process

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Table 1: UCI Dataset Attributes

SI. No.	Attribute Description	Distinct Values of Attributes
1.	Age – represents the age of a person	Different Values between 29 & 71
2.	Sex – represents the gender of the person (0-Female, 1-Male)	0 and 1
3.	CP – Chest pain type Type 1: typical angina Type 2: atypical angina Type 3: non–anginal pain Type 4: asymptomatic	0,1,2,3
4.	Trestbps – Resting blood pressure (in mm Hg on admission to hospital)	Different values between 94 and 200
5.	Chol – serum cholesterol in mg/dl	Values between 126 and 564
6.	FBS – Fasting blood sugar	0,1
7.	Resting – Resting electrocardiographic result	0,1
8.	Thalach – Maximum Heart rate achieved	Values between 71 & 202
9.	Exang – Exercise included angina 1-yes, 0-no	0,1
10.	Oldpeak – ST depression introduced by exercise relative to rest	0 to 6.2
11.	Slope – The slope of the peak exercise ST segment	0,1,2
12.	Ca – Number of major vessels	0,1,2,3
13.	Thal – test is required for the patient suffering from pain in the chest or difficulty in breathing. There are 4 kinds of values that represent the Thallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. This dataset has binary classification i.e., two classes (0,1). In class "0" represent there is less possibility of heart disease whereas "1" represents high chances of heart disease. The value "0" Or "1" depends on the other 13 attributes.	0,1

IV. MACHINE LEARNING

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that allows applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Figure 1 shows the steps and the flow involved in the machine learning process.

Machine Learning can be categorized into 3 types:

a. Supervised Machine Learning

This type of machine learning is used when the model is provided with the labeled data and on basis of that data, machines predict the output. The training data provided to the machine acts as a teacher which teaches the machine to predict the output. The data is divided into training and test data, the training data is used to teach the machine, and using test data the machine is tested if it is giving the right predicted output. This supervised learning can be further classified into regression and classification methods.

Algorithms that come under supervised learning are:

- Logistic Regression
- Support Vector Machine (SVM)
- Naïve Bayes
- Decision Tree, Random Forest

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

b. Unsupervised Machine Learning

Unsupervised machine learning is a type of ML where the machine is provided with the set of unlabeled data and the machine is expected to provide the output i.e., predicted data. The model is not supervised using training data, instead model itself finds the patterns from the given data. The algorithm will predict the output using grouping the images or data according to the similarities between the data. This can be further classified into two more types, clustering and dimensionality reduction.

Algorithms that come under unsupervised ML are:

- Hierarchical
- Mean Shift
- Density-based

c. Reinforcement Learning

This technique is in the middle of supervised and unsupervised learning, where the model improves its performance as it interacts with the environment. Hence, learn how to correct its mistakes. It ought to get the correct result through examination and trying out different possible outcomes [3]. Reinforced learning is the agent's ability to interact with the environment and find out the outcome. It is based on the "hit and trial" concept [12]

For this paper, we have used classification-based supervised learning algorithms:

A. Logistic Regression
B. Decision Tree
C. Random Forest
D. Support Vector Machine
E. KNN

Classification ML Algorithms:

Table 01 shows all the attributes that are fed into the classification algorithms like logistic regression, decision tree, random forest, support vector machine, and KNN. The data set is split into both train and test data, where the train set is taken as 80% and the test set as 20%.

A. Logistic Regression

Logistic Regression is a type of classification supervised machine learning algorithm mostly used for binary classification. Logistic regression has mainly two parts dependent and independent variables.

The outcome of this can be either Yes or No, 0 or 1, etc. In logistic regression instead of fitting a straight line or hyperplane, the logistic regression algorithm uses the logistic function to predict the output of a linear equation between 0 and 1 [5]. Based on the independent variables (X) the dependent variable (Y) is predicted. It can be represented as, P(Y) = X.

B. Decision Tree

A decision tree classifier is a type of supervised machine learning algorithm which is both classification and regression-based. It is a tree-structured based classifier where internal nodes are the features of a particular dataset and each leaf node represents the outcome. The main thing that comes in a decision tree is to select the best attributes for the root node and sub-nodes. For selecting the attribute two techniques can be used i.e., Information gain and Gini index.

Information Gain is the measure of changes in entropy after segmentation of a dataset based on the attributes.

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)

Entropy is the measure of impurity in a given attribute. It can be calculated as,

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no) Where,

S= Total number of samples, P(yes)= probability of yes, P(no)= probability of no

C. Random Forest

Random Forest is based on the concept of ensemble learning which is a process of combining multiple classifiers to improve the performance of a particular model. Random forest is a supervised ML algorithm that

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

contains a combination of decision trees of a dataset and takes the average to improve the accuracy of the dataset.

D. K-Nearest Neighbor

K-Nearest Neighbor or KNN algorithm is based on the method where it classifies the similar case or data in a dataset that lies in the same categories. KNN algorithm stores all the available data in separate categories and when new data comes it is classified into a similar category.

E. Support Vector Machine

Support Vector Machine or SVM is one of the supervised ML algorithms used for both classification and regression problems. The main aim of SVM is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category. The best line that is fitted is called the hyperplane. The SVM chooses the data points that are closer to the hyperplane which is called a support vector.

V. RESULTS AND ANALYSIS

By applying the above algorithms in the UCI Heart Disease Dataset, we have the following results and analysis. For evaluating the performance of the model, we have used performance metrics which are accuracy, precision, recall, F1 score, and for model visualization confusion matrix.

A. Precision, Recall, and F1 Score

For evaluating the performance of the model, we have certain parameters like precision, recall, and F1 score which are based mainly on these four values:

1. True Positive (TP): The predicted outcome is true, and it is true in reality, as well. (Total No. of people with heart disease)
2. True Negative (TN): The predicted outcome is false, and it is false in reality. (Total No. of people with no heart disease)
3. False Positive (FP): The prediction outcomes are true, but they are false in actuality. (People with No Heart disease but predicted falsely as they have the disease)
4. False Negative (FN): The predictions are false, and they are true in actuality. (People with heart disease but predicted as they don't have the heart disease)

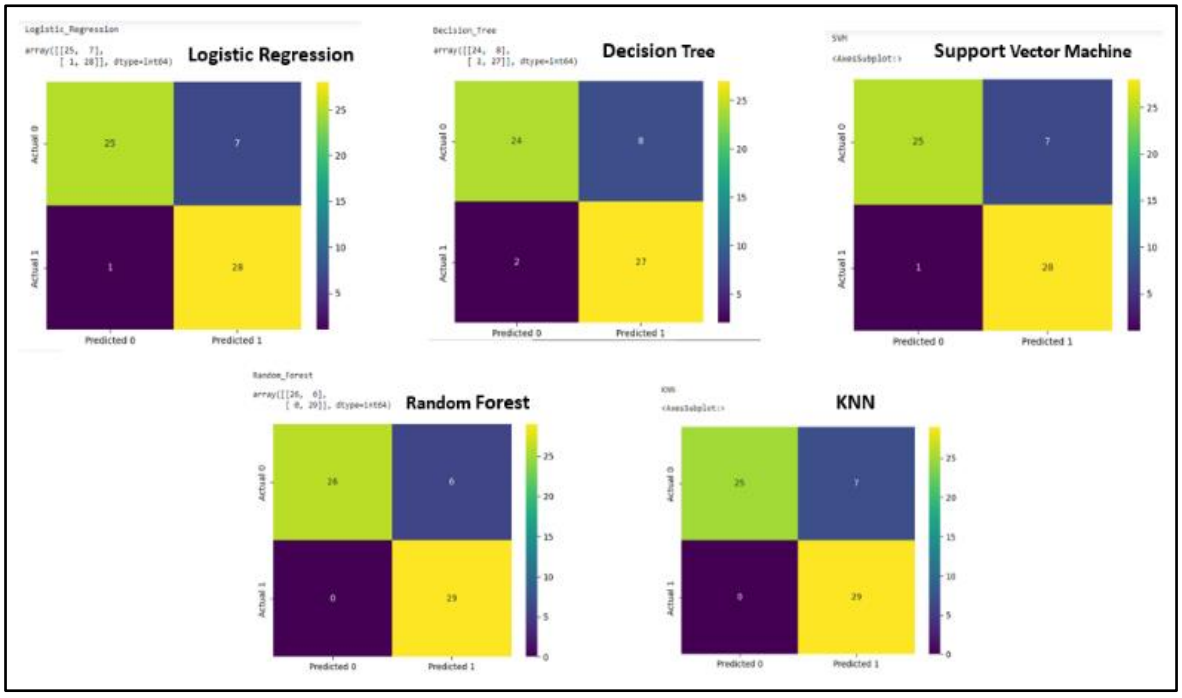

Figure 2: Confusion Matrix

B. Accuracy

By applying the above algorithms to the dataset, we get the accuracy as follows:

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

We can see that Random Forest has the highest accuracy among all the other algorithms used. Random Forest gives about 90% of accurate results as compared to the other algorithms.

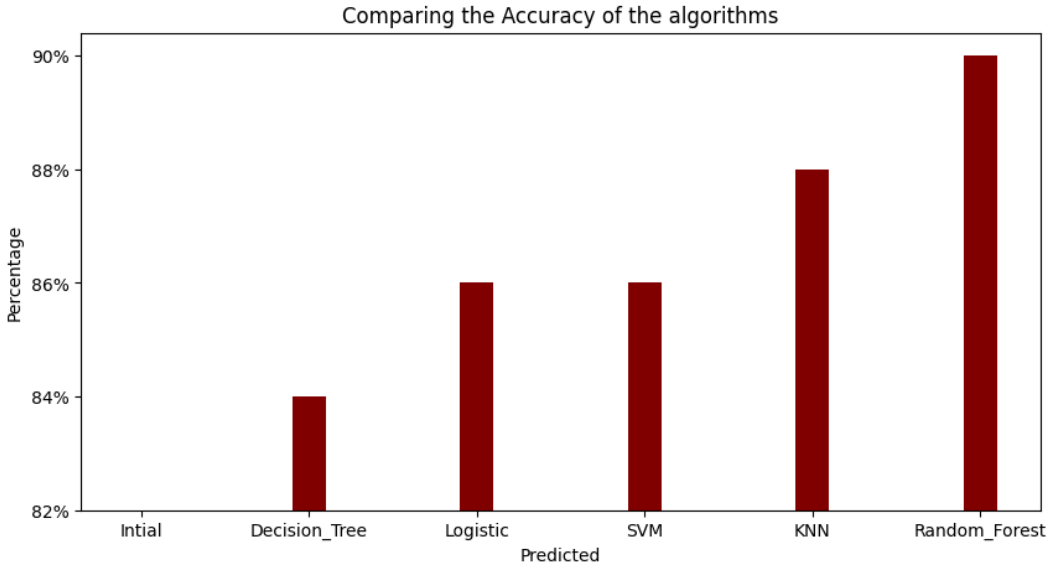

Figure 3: Accuracy Measure

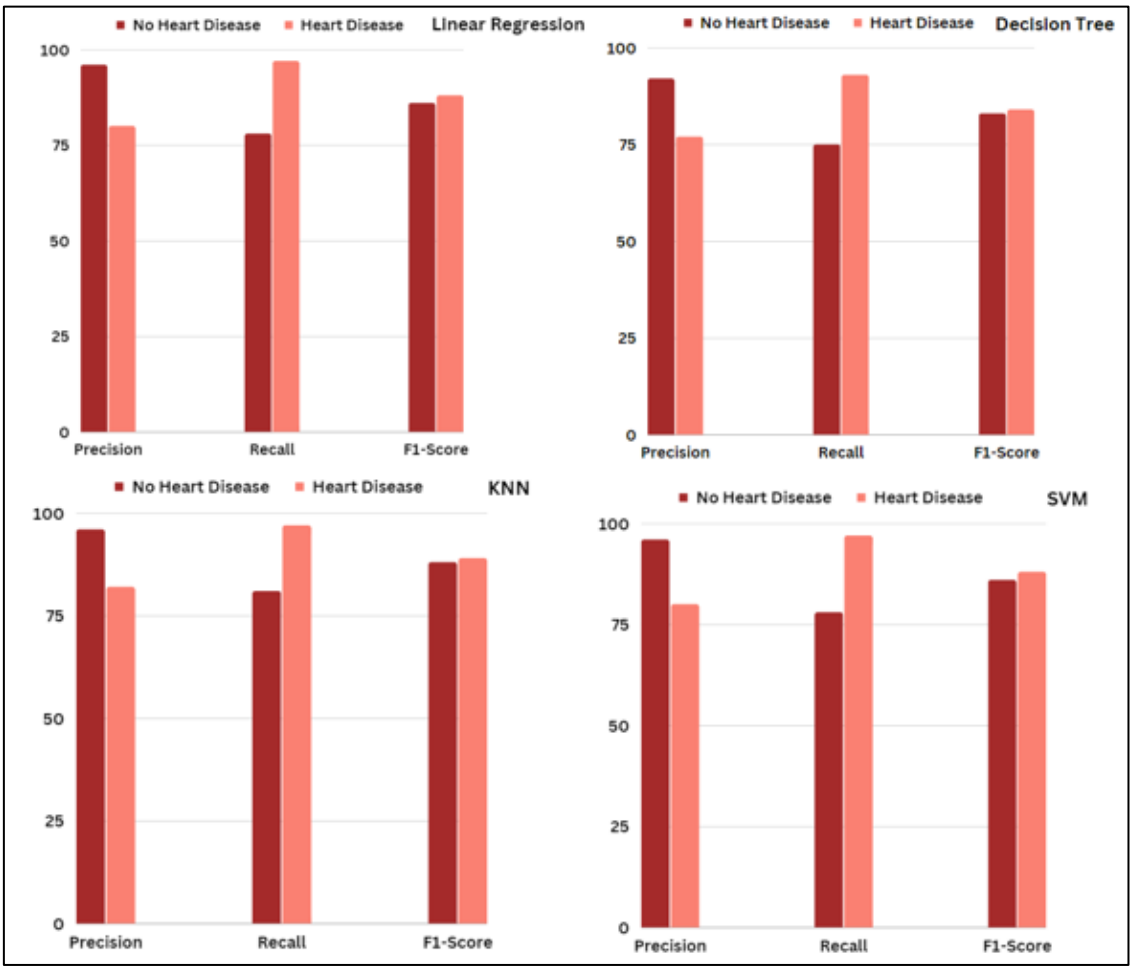

Figure 4: Precision, Recall and F1-Score comparison

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Table 2. Precision, Recall, F1-Score

Algorithm	Precision	Recall	F1-Score	Accuracy
Linear Regression	0.80	0.97	0.88	86%
Decision Tree	0.77	0.93	0.84	83%
Random Forest	0.85	0.84	0.92	91%
KNN	0.82	0.97	0.89	88%
SVM	0.80	0.97	0.88	86%

VI. CONCLUSION

As we have seen Random Forest is the most accurate over the particular dataset as compared to others. The accuracy and the performance evaluation techniques like precision, recall, and F1 score depends and changes their value according to the dataset passed to it. The paper finds the accuracy among the ML Algorithms and compares the best among them. Heart Disease when gone unchecked or the lack of knowledge of the disease can be life-threatening over time. This paper shows that using ML Algorithms for the prediction of such disease can be a helpful way to detect the disease as soon as possible. In the future, more algorithms or the accuracy of these algorithms can be increased based on the input or using a larger dataset.

ACKNOWLEDGEMENTS

A special thanks and gratitude to B.K Birla College (Autonomous), IT Department, and Prof. Swapna Augustine Nikale, B.K Birla College (Autonomous), Kalyan for providing this opportunity and mainly for the continued guidance and support for this paper.

VII. REFERENCES

[1] Uddin, S., Khan, A., Hossain, M. and Moni, M., 2019. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making, 19(1).

[2] Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE Access, 7, pp.81542-81554.

[3] Jindal, H., Agrawal, S., Khera, R., Jain, R. and Nagrath, P., 2021. Heart disease prediction using machine learning algorithms. IOP Conference Series: Materials Science and Engineering, 1022(1), p.012072.

[4] Nikhar, Sonam, and A. M. Karandikar. "Prediction of heart disease using machine learning algorithms." International Journal of Advanced Engineering, Management and Science 2.6 (2016): 239484.

[5] Apurb Rajdhan, Avi Agarwal, Milan Sai and Dundigalla Ravi, Dr. Poonam Ghuli, 2020. Heart Disease Prediction using Machine Learning. International Journal of Engineering Research and Technology, V9(04).

[6] Rindhe, B. U., Ahire, N., Patil, R., Gagare, S., & Darade, M. (2021). Heart Disease Prediction Using Machine Learning. Heart Disease, 5(1).

[7] Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

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

[9] Aljanabi, M., Qutqut, M. H., & Hijjawi, M. (2018). Machine learning classification techniques for heart

disease prediction: a review. International Journal of Engineering & Technology, 7(4), 5373-5379.

[10] Lutimath, N. M., Chethan, C., & Pol, B. S. (2019). Prediction of heart disease using machine learning. International Journal of Recent Technology and Engineering, 8(2), 474-477.

[11] Raju, K., Dara, S., Vidyarthi, A., Gupta, V. and Khan, B., 2022. Smart Heart Disease Prediction System with IoT and Fog Computing Sectors Enabled by Cascaded Deep Learning Model. Computational Intelligence and Neuroscience, 2022, pp.1-22.

[12] Singh, A. and Kumar, R., 2020. Heart Disease Prediction Using Machine Learning Algorithms. 2020 International Conference on Electrical and Electronics Engineering (ICE3),.

[13] Raja, M., Anurag, M., Reddy, C. and Sirisala, N., 2021. Machine Learning Based Heart Disease Prediction System. 2021 International Conference on Computer Communication and Informatics (ICCCI).

[14] Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. International Journal of Recent Technology and Engineering, 8(1), 944-950.

[15] Thomas, J., & Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In 2016 international conference on circuit, power and computing technologies (ICCPCT) (pp. 1-5). IEEE.