IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004006 |

Heart Disease Prediction Using Machine Learning Algorithm

Prof. V.B.Bhagat, Dr.V. M. Thakare, Sujata M. Borkar, Kajal C. Palkhade,

Samiksha P. Thakre, Swati M. Fokmare

Department of Computer Science and Engineering, P.R. Pote(Patil) College of Engineering & Management, Amravati, India

ABSTRACT: Heart disease prediction is one among the foremost complicated tasks in medical field. As heart condition prediction may be a complex task, there is a requirement to automate the prediction process to avoid risks related to it and alert the patient well beforehand. This paper makes use of heart condition dataset available in UCI machine learning repository. The proposed work predicts the probabilities of heart condition and classifies patient's risk level by implementing different data processing techniques like SVM, XGBOOST, Decision Tree, Logistic Regression and Random Forest. Thus, this project presents a comparative study by analysing the performance of various machine learning algorithms. The trial result verifies that XGBOOST algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented.

I. INTRODUCTION

Heart disease describes a range of conditions that affect your heart. Today, cardiovascular diseases are the leading cause of death worldwide with 17.9 million deaths annually, as per the World Health Organization reports . Various unhealthy activities are the reason for the increase in the risk of heart disease like high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. There are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2kg daily. Nowadays it is well known that machine learning and artificial intelligence are playing a huge role in the medical industry. We can use different machine learning and deep learning models to diagnose the disease and classify or predict the results. A complete genomic data analysis can easily be done using machine learning models. Models can be trained for knowledge pandemic predictions and also medical records can be transformed and analyzed more deeply for better prediction.

II. PROPOSED WORK

The heart disease prediction can be performed by following the procedure which is similar to Fig.1 which specifies the research methodology for building a classification model required for the prediction of the heart diseases in patients. The model forms a fundamental procedure for carrying out the heart disease prediction using any machine learning techniques. In order to make predictions, a classifier needs to be trained with the records and then produce a classification model which is fed with a new unknown record and the prediction is made.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 4, April 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1004006 |

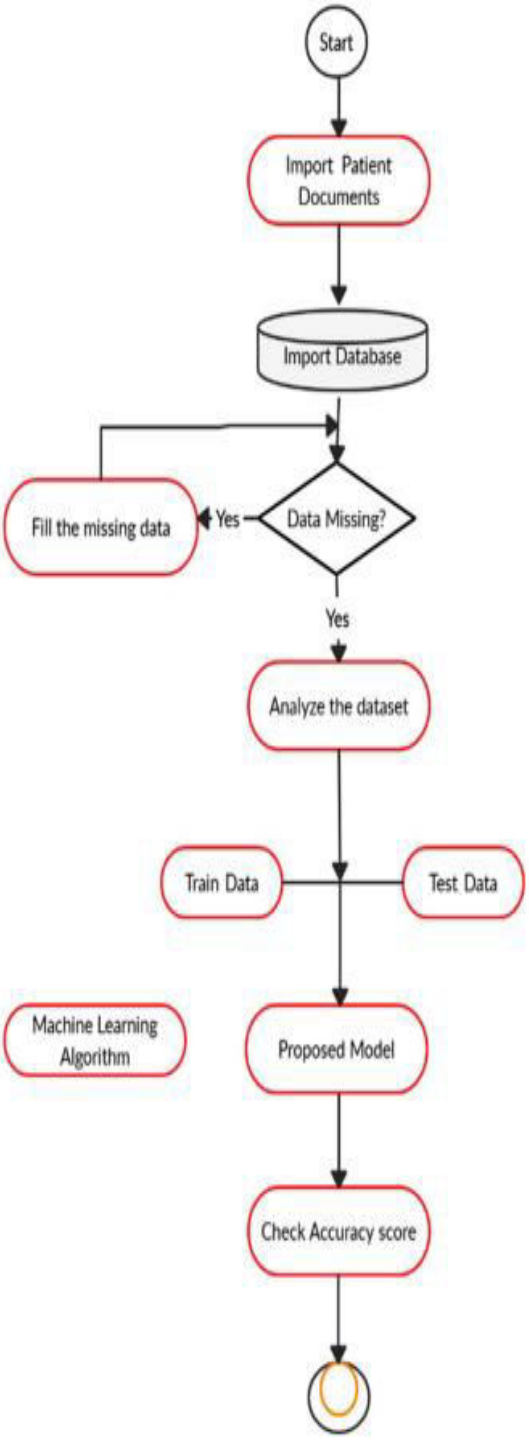


Fig.1. Proposed Methodology

III. LITERATURE REVIEW

The summary of the literature review can be seen in Table 1. Several approaches have been performed on this popular dataset, but the accuracy obtained by all the approaches is more with time computations.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com| |Impact Factor: 8.165

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003006 |

Author	Year	Findings
P. Melillo et. al	2013	Two public Holster databases were used for finding high-risk and low-risk patients. Cart algorithm is applied for the classification purpose.
M. M. A. Rahhal et. al	2016	ECG approach is used by consulting various domain experts and then the MIT-BIH arrhythmia database as well as two other databases called IN CART, and SVDB, respectively.
G. Guidi et. al	2014	Neural Networks, SVM, Fuzzy System approach, is used and Random Forest is used as a classifier, for the prediction of heart failure by using a clinical decision support system.
R. Zhang et. al	2017	A support vector machine is used for the classification purpose of the clinical data which is matched with the codes of the New York Heart Association, further findings are left for other researchers.
G. Parthiban et. al	2012	Diabetes is one of the main concerns for heart disease. The classifiers used are Naïve Bayes and SVM for extracting important features and classification purposes.
E. Keogh and A. Mueen et. al	2012	How to break the curse of dimensionality using PCA, SVM, and other classifiers and reduce features.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com| |Impact Factor: 8.165

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003006 |

IV. METHODOLOGY

Description of the Dataset

The dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them.

- **Age** – Age of patient in years, sex – (1 = male; 0 = female).
- **Cp** – chest pain type.
- **Trestbps** – resting blood pressure (in mm Hg on admission to the hospital), the normal range is 120/80, if you have a normal blood pressure reading, it is fine but it is a little higher than it should be, and you should try to lower it, make healthy changes to your lifestyle.
- **Chol** – serum cholesterol, shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
- **Fbs** – fasting blood sugar larger 120mg/dl (1 true), less than 100 mg/dL (5.6 mmol/L) is normal, 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
- **Restecg** – resting electrocardiographic results.
- **Thalach** – maximum heart rate achieved, the maximum heart rate is 220 minus your age.
- **Exang** – exercise-induced angina (1 yes), Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
- **Oldpeak** – ST depression induced by exercise relative to rest.
- **Slope** – the slope of the peak exercise ST segment.
- **Ca** – number of major vessels (0-3) colored by fluoroscopy.
- **Thal** – no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).

Target (T) – No Disease = 0 and Disease = 1, (angiographic disease status).

V. PRE-PROCESSING OF THE DATASET

The dataset does not have any null values. But, many outliers needed to be handled properly, and also the dataset is not properly distributed. Two approaches were used. One without outliers and Feature selection process and directly applying the Data to the Machine learning algorithms, and the results which were achieved were not promising. But after using the Normal distribution of Dataset for overcoming the Overfitting problem and then applying Isolation Forest for the outlier's detection, the results achieved are quite promising. Various plotting techniques were used for checking the skewness of the data, outlier detection, and the distribution of the data. All these pre-processing techniques play an important role when passing the data for classification or prediction purposes.

VI. FEATURE SELECTION

For selecting the features and only choosing the important feature, the Lasso algorithm is used which is a part of embedded methods while performing Feature selection. It shows better predictive accuracy than filter methods. It renders good feature subsets for the used algorithm. And then for selecting the selected features, select from the model which is a part of feature selection in the scikit-learn library.

VII.MACHINE LEARNING CLASSIFIERS PROPOSED

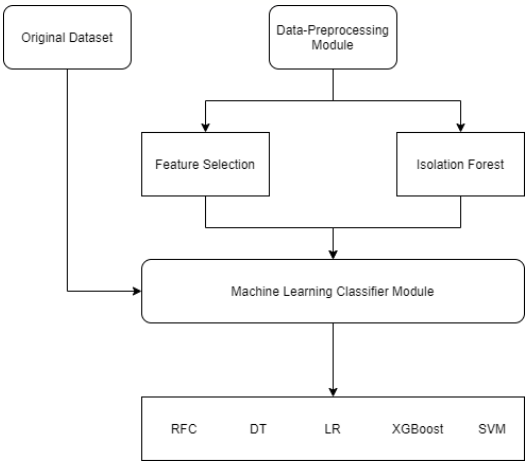
The proposed approach was applied to the dataset in which firstly the dataset was properly analysed and then different machine learning algorithms consisting of linear model selection in which Logistic Regression was used and then for also focusing on neighbour selection technique KNeighbors Classifier was used, then tree-based technique like Decision Tree Classifier was used, and then a very popular and most popular technique of Ensemble methods Random Forest Classifier was used, also for checking the high dimensionality of the data and handling it Support Vector Machine was used, another approach which also works on ensemble method and Decision tree method combination is XGBoost classifier.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com| |Impact Factor: 8.165

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003006 |

Figure 6: 1st Schematic diagram of the proposed model

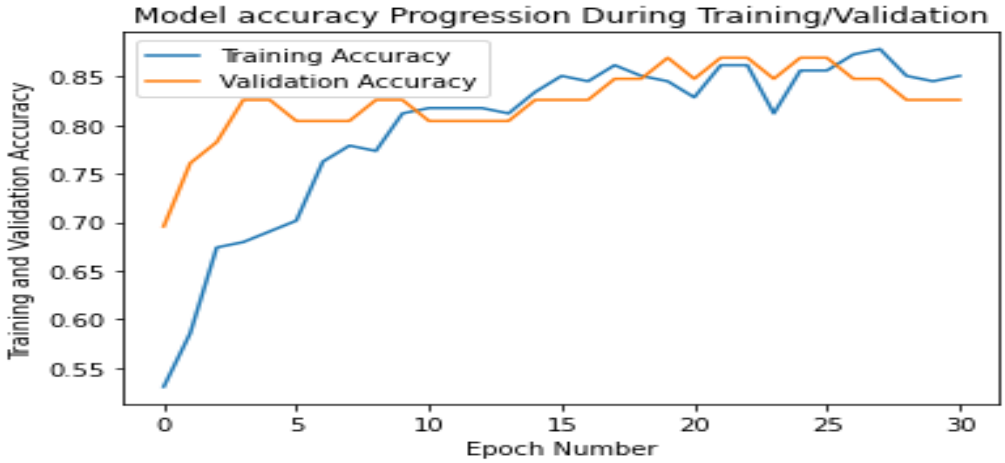


VIII. DISCUSSION AND CONCLUSION

In this paper, we proposed three methods in which comparative analysis was done and promising results were achieved. The conclusion which we found is that Machine learning algorithms performed better in this analysis.

Figure 1

Many researchers have previously also suggested that we should use ML where the dataset is not that large, which is proved in this paper. The methods which are used for comparison are confusion matrix, precision, specificity, sensitivity, and F1 score.



For the 13 features which were in the dataset, KNeighbors classifier performed better in the ML approach when data pre-processing is applied. The computational time also reduced which is helpful when deploying a model. It was also found out that the dataset should be Normalised, otherwise the training model gets over- fitted sometimes and the accuracy which is achieved is not sufficient when a model is evaluated for real-world data problems which can vary drastically to the dataset on which the model was trained.

IX. RESULTS

By applying different Machine learning algorithms and then using deep learning to see what difference comes when it is applied to the Data. Three approaches were used, in the first approach normal dataset which is acquired is directly

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com| |Impact Factor: 8.165

|| Volume 10, Issue 3, March 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1003006 |

used for classification, in the second approach, taking care of the data with Feature Selection and no outliers detection, the results which are achieved are quite promising and then in the third approach the dataset was Normalized with taking care of the outliers and Feature selection, the results achieved are much better than the previous techniques and when compared with other research accuracies, our results are quite promising.

REFERENCES

[1] "Cardiovascular diseases." [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1. [Accessed: 06-Oct-2020].

[2] "Classes of Heart Failure | American Heart Association." [Online]. Available: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure. [Accessed: 06-Oct-2020].

[3] "Heart Failure | American Heart Association." [Online]. Available: https://www.heart.org/en/health-topics/heart-failure. [Accessed: 06-Oct-2020].

[4] "Shalev-Shwartz: Understanding machine learning: From... - Google Scholar." [Online]. Available: https://scholar.google.com/scholar_lookup?title=Understanding machine learning%3A from theory to algorithms&author=S. Shalev-Shwartz&publication_year=2016. [Accessed: 06-Oct-2020].

[5] "Hastie: The elements of statistical learning: data... - Google Scholar." [Online]. Available: https://scholar.google.com/scholar_lookup?title=The elements of statistical learning%3A data mining%2C inference%2C and prediction&author=T. Hastie&publication_year=2017. [Accessed: 06-Oct-2020].

[6] "Marsland: Machine learning: an algorithmic perspective - Google Scholar." [Online]. Available: https://scholar.google.com/scholar_lookup?title=Machine Learning%3A an algorithmic perspective&author=S. Marsland&publication_year=2015. [Accessed: 06-Oct-2020].

[7] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," IEEE J. Biomed. Heal. Informatics, vol. 17, no. 3, pp. 727–733, 2013.

[8] M. M. A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," Inf. Sci. (Ny)., vol. 345, pp. 340–354, Jun. 2016.

[9] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE J. Biomed. Heal. Informatics, vol. 18, no. 6, pp. 1750–1756, Nov. 2014.

[10] R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York heart association classification from clinical notes," in Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, 2017, vol. 2017-January, pp. 1296–1299.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details