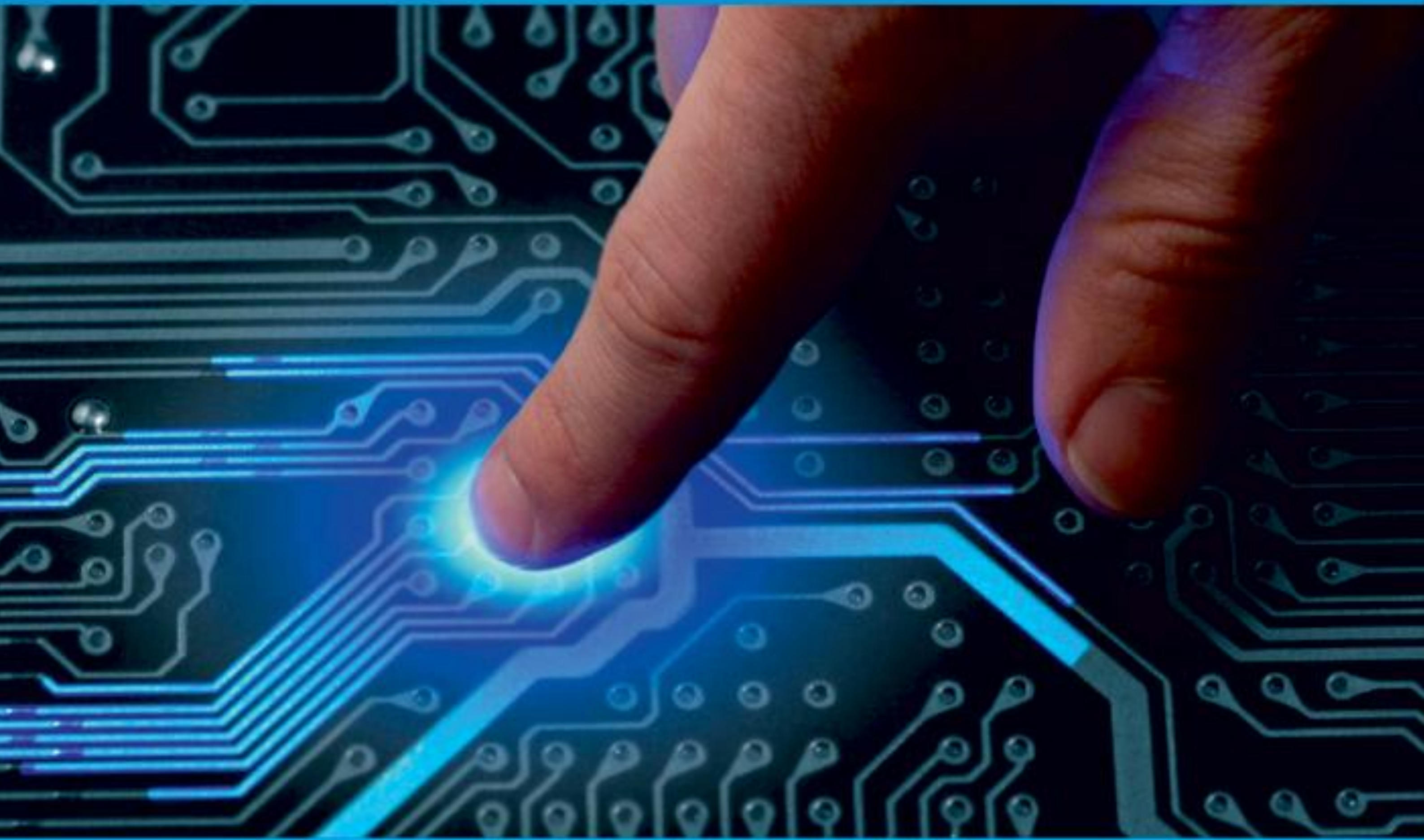
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

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

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005232 |

Heart Disease Prediction Using Machine Learning Techniques DBSCAN, SMOTE-ENN and XGBoost

Sharon Suprabha, Tabassum Ara, Manju Sri, Poorna Chandra, Maddala Jeevan Babu

UG Students, Dept. of C.S.E, Vasireddy Venkatadri Institute Of Technology, Affiliated to Jawaharlal Nehru Technological University, Kakinada, A.P., India

Assistant Professor, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh, India

ABSTRACT: Heart disease is the most unpredictable causes of death in today's world because of its uncertainty in occurrence. Prediction of heart disease is still a big question and difficult challenge for people in the area of clinical data analysis as it has to be predicted early for a human to survive. The main idea of this project is to propose an effective heart disease prediction model (HDPM) which consists of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to detect and eliminate the outliers, a hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN) to balance the training data distribution and XGBoost to predict heart disease. Two publicly available datasets (Statlog and Cleveland) were used to build the model and compare the results with those of other traditional Machine Learning Models(Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT),Kth Nearest Neighbors (KNN) and Random Forest (RF)). The results revealed that the proposed model outperformed other models significantly and can be used effectively against the prediction of heart disease.

KEYWORDS: Logistic Regression(LR), Decision Tree(DT), Kth Nearest Neighbors(kNN), Support Vector Machine(SVM), Random Forest(RF), XGBoost, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE-ENN).

I. INTRODUCTION

Heart Disease is a Cardio Vascular Disease (CVD) that remains the number one cause of death globally and contributes to approximately 30% of all global deaths .The American Heart Association reported that nearly half of American adults are affected by CVDs, equating to nearly 121.5 million adults . Heart disease is a condition when plaque on arterial walls can block the flow of blood and cause a heart attack or stroke. Several risk factors that can lead to heart disease include unhealthy diet, physical inactivity, and excessive use of tobacco and alcohol. The early heart disease identification of high-risk individuals and the improved diagnosis using a prediction model have generally been recommended to reduce the fatality rate and improve the decision-making for further prevention and treatment .

Machine learning-based clinical decision making have recently been applied in healthcare area. Previous studies have shown that machine learning algorithms (MLAs) such as fire flyalgorithm , backpropagation neural network (BPNN) , multilayer perceptron (MLP) , logistic regression (LR) , support vector machine (SVM) , and
random forest (RF) have been successfully used to help as decision making tools for heart disease prediction based onindividual data.

In this project, Heart Disease Prediction is done through hybrid model. As previous studies stated that hybrid isgoing to produce maximum results than linear individual models like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost and Kth Nearest Neighbor. For hybrid model of Heart Disease Prediction three Machine Learning Algorithms (MLA) are going to be used to produce most accurate results. The Clustering Algorithms of Unsupervised Machine Learning are DBSCAN –Density Based Spatial Clustering of Application with Noise and SMOTE – ENN- Synthetic Minority Oversampling Technique- Edited Nearest Neighbor, then a Gradient Boosting Algorithm is used which is XGBoost.

II. RELATED WORK

Long et al. (2015): [1] Several studies have reported the development of heart dis- ease diagnosis based on machine learning models with the aim of providing an HDPM with enhanced performance. Two publicly available heart disease datasets, namely Stat- log and Cleveland, have been widely used to compare the performance of prediction models among researchers. For Statlog dataset, a heart disease clinical decision support system based on chaos firefly algorithm and rough sets-based attribute reduction was developed. The rough sets were used to reduce the number of attributes while the chaos firefly algorithm was used to classify the disease. The results revealed that the proposed model achieved the highest performance among all the models with accuracy, sensitivity, and specificity of 88.3%, 84.9%, and 93.3%, respectively.

Nahato et al. (2015): [2] The combination of rough sets-based attributes selection and BPNN (RS-BPNN) was proposed with the selected attributes, the pro- posedRS-BPNNachievedaccuracyofupto90.4%.

Verma et al. (2016): [3] Developed a hybrid prediction model based on correlation feature subset (CFS), particle swam optimization (PSO), K- means clustering and MLP. The results showed that the proposed hybrid model achieved accuracy of up to 90.28%.

Dwivedi et al. (2018): [4] Compared six machine learning models (ANN, SVM, LR, k-nearest neighbour (kNN), classification tree and NB) with various performance metrics.

Haq et al. (2018): [5] Performed a comparative study on a hybrid model based on various feature selection techniques (relief, minimal- redundancy- maximal-relevance (mRMR), least absolute shrinkage and selection operator (LASSO)) and machine learning models (LR, kNN, ANN, SVM, DT, NB and RF).

Their study revealed that the features reduction affects the performance of the models. The study concluded that a combination of Relief- based feature selection and LR-based machine learning algorithm (MLA) provides higher accuracy (up to 89%) as compared with other combinations used in the study.

Gupta et al. (2020): [10] Developed a machine intelligence framework consisting of factor analysis of mixed data (FAMD) and RF-based MLA. The FAMD was used to find the relevant features and the RF to predict the disease. The experimental results showed that the proposed method outperformed other models and previous study results by

achieving the accuracy, sensitivity, and specificity of up to 93.44%, 89.28%, and 96.96%, respectively. None of the fore mentioned previous studies have applied outlier detection and data balancing method to improve the accuracy of classification model, especially for the case of heart disease datasets. Thus, in this study we used outlier detection and data balancing methods to improve the model performance.

III. PROPOSED SYSTEM

A. *Design Considerations:*

There are thirteen features and one target as below:

age: The person's age in years

sex: The person's sex (1 = male, 0 = female)

cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)

trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)

chol: The person's cholesterol measurement in mg/dl

fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)

restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

thalach: The person's maximum heart rate achieved

exang: Exercise induced angina (1 = yes; 0 = no)

oldpeak: ST depression induced by exercise relative to rest

slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)

ca: The number of major vessels (0-3)

thal: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

target: Heart disease (0 = no, 1 = yes)

B. *Description of the Proposed Algorithm:*

Aim of the proposed algorithm is to get maximum accuracy using DBSCAN, SMOTE-ENN and XGBoost Algorithms.

Step 1: DBSCAN Model

DBSCAN is a clustering algorithm that is used to remove outliers in the data. So, in this proposed model this DBSCAN is used as preprocessing technique to incorporated dataset.

DBSCAN basically depends on two parameters. The two parameters are:

1. Min points (minpts)
2. Epsilon (eps)

1. Min points (minpts): These are the points that are present in single cluster. In other words min points represent size of the cluster.

2. Epsilon (eps): Epsilon is the radius that is required to collect all the points that belong to single cluster. Epsilon is set in increasing fashion. Firstly few points are considered in one cluster with one epsilon later epsilon value is increased and then it is fixed.

Here is the working example of DBSCAN Algorithm:

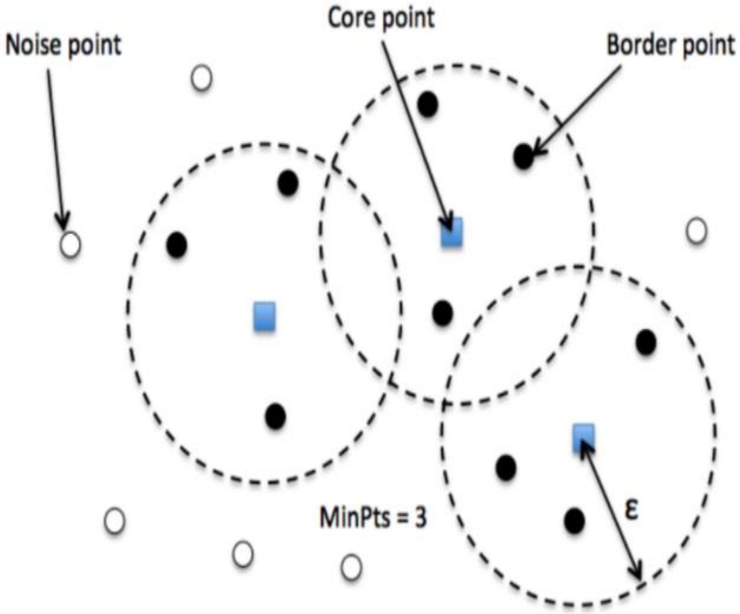


Figure 1: DBSCAN working example

Step 2: SMOTE – ENN Model

This model is developed by Batista et al (2004) , this method combines the SMOTE ability to give synthetic examples for minority class and ENN ability to delete few observations from both classes that are identified as different class from observation's class and its K-nearest neighbour majority class.

The steps for SMOTE-ENN Techniques:

1. (**Start of SMOTE**) Choose random data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Repeat step number 2–3 until the desired proportion of minority class is met. (**End of SMOTE**)

5. (**Start of ENN**) Determine K, as the number of nearest neighbors. If not determined, then K=3.

6. Find the K-nearest neighbor of the observation among the other observations in the dataset, then return the majority class from the K-nearest neighbor.

7. If the class of the observation and the majority class from the observation's K-nearest neighboris different, then the observation and its K-nearest neighbor are deleted from the dataset.

8. Repeat step 2 and 3 until the desired proportion of each class is fulfilled. (**End of ENN**)

Step – 3: XGBoost Model

XGBoost is a decision tree based ensemble machine learning algorithm. It uses Gradient Boosting Framework.XGBoost algorithm was developed as a research project at the University of Washington. Tianqi Chen and Carlos Guestrin presented their paper at SIGKDD Conference in 2016 and caught the Machine Learning world by fire. Since its introduction, this algorithm has not only been credited with winning numerous Kaggle competitions but also for being the driving force under the hood for several cutting-edge industry applications. As a result, there is a strong community of data scientists contributing to the XGBoost open source projects with ~350 contributors and ~3,600 commits on GitHub. The algorithm differentiates itself in the following ways:

1. A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.

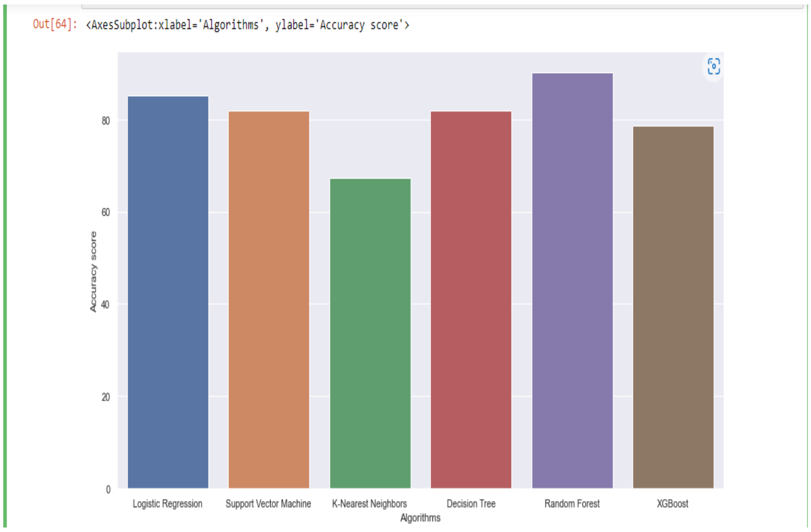
2. Portability: Runs smoothly on Windows, Linux, and OS X.

3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.

4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems

IV. RESULTS

If popular Machine Learning Algorithms are used then the accuracy results are as shown below.The below figure is a Comparison graph for Machine Learning Algorithms like Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, Kth- Nearest Neighbor and XGBoost.



```
Out[64]: <AxesSubplot:xlabel='Algorithms', ylabel='Accuracy score'>
```

Figure 2: Comparison Graph of Machine Learning Algorithms

As observed above figure Off all the Algorithms Random Forest gained highest accuracy in prediction of Heart Disease.

But proposed model produces a high accuracy of 97% in prediction of disease. In this model preprocessing of data Plays a vital role with removal of outliers and balancing the minority classes in the dataset. Boosting the Accuracy by using Gradient Boosting Algorithm XGBoost has gained a high accuracy.

As Heart Disease Prediction should be more accurate than normal predictions, this model achieved almost of the maximum accuracy.

V. Conclusion and Future Work

We proposed an effective heart disease prediction model (HDPM) for heart disease diagnosis by integrating DBSCAN, SMOTE-ENN, and XGBoost- based MLA to improve prediction accuracy. The DBSCAN was applied to detect and remove the outlier data, SMOTE-ENN was used to balance the unbalanced training dataset and XGBoost

MLA was adopted to learn and generate the prediction model. Two publicly available datasets of heart disease were utilized by produce the generalized prediction model. We performed evaluation analysis of our proposed model with other classification models and the results from previous studies. In addition, we presented the statistical evaluation to confirm the significant of our model as compared to other models. The experimental results confirmed that the proposed model achieved better performance than that of state-of-the-art models and previous study results, by achieving an accuracy up to 95.90% and 98.40% for datasets I and II, respectively. In addition, the statistical-based analysis result also showed the significant improvement for the proposed model as compared with the other models. Furthermore, we also proposed HDPM to diagnose the subjects'/patients' heart disease status effectively and efficiently. Thus, the developed HDPM is expected to help clinicians to diagnose patients and improving heart disease clinical decision making effectively and efficiently We have also ensured and considered the comparison of other data sampling with the model hyper-parameters and broader medical datasets. In addition, a comparison and analysis study with different outlier detection methods is

investigated. Furthermore, with the increasing concerns about privacy, security and time-sensitive applications, edge computing and edge device concepts could be further studied with the goal of improving the medical clinical decision support system.

References

[1] N. C. Long, P. Meesad, and H. Unger, ''A highly accurate firefly basedalgorithm for heart disease prediction,'' Expert Syst. Appl., vol. 42, no. 21, pp.8221–8231,Nov.2015,doi:10.1016/j.eswa.2015.06.024.

[2] K. B. Nahato, K. N. Harichandran, and K. Arputharaj, ''Knowledge miningfrom clinical datasets using rough sets and backpropagation neural network,''Comput.Math.MethodsMed.,vol.2015,pp.1–13,Mar.2015,doi: 10.1155/2015/460189.

[3] L. Verma, S. Srivastava, and P. C. Negi, ''A hybrid data mining model topredict coronary artery disease cases using non-invasive clinical data,'' J. Med.Syst.,vol.40,no.7,p.178,Jul.2016,doi:10.1007/s10916-016-0536-z.

[4] A.K.Dwivedi,''Performanceevaluation of differentmachinelearningtechniques for prediction of heart disease,'' Neural Comput. Appl., vol. 29, no.10,pp.685–693,May2018,doi:10.1007/s00521-016-2604-1.

[5] A. U. Haq, J. P. Li, M. H. Memon,S. Nazir, and R. Sun, ''A hybridintelligent system framework for the prediction of heart disease using machinelearning algorithms,'' Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018, doi:10.1155/2018/3860146.

[6] C. B. C. Latha and S. C. Jeeva, ''Improving the accuracy of prediction ofheart disease risk based on ensemble classification techniques,'' Inform. Med.Unlocked,vol.16,Jan.2019,Art.no.100203,doi: 10.1016/j.imu.2019.100203.

[7]L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R.Nour, and S. A. C. Bukhari, ''An optimized stacked support vector machinesbased expert system for the effective prediction of heart failure,'' IEEE Access,vol.7,pp.54007–54014,2019,doi:10.1109/ ACCESS.2019.2909969.

[8]WorldHealthOrganization.(2017).CardiovascularDiseases(CVDs).[Online].Available:https://www.who.int/health-topics/cardiovasculardiseases/.

[9] E. J. Benjamin et al., ''Heart disease and stroke statistics—2019 update: Areport from the American heart association,'' Circulation, vol. 139, no. 10, pp.e56–e528,Mar.2019,doi:10.1161/CIR.0000000000000659.

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