BioGecko

Vol 12 Issue 02 2023
ISSN NO: 2230-5807

DETECTION OF CARDIOVASCULAR DISEASE USING MACHINE LEARNING

1Dr.Ch.Kavitha,2 Ch.Yasaswini,3 K.Gnana Deepika,4 G.Harika Siva Satyavathi .
1Professor and HOD, Department of IT,SRGEC,Gudlavalleru.
2Undergraduate Student,Department of IT,SRGEC,Gudlavalleru.
3Undergraduate Student,Department of IT,SRGEC,Gudlavalleru.
4Undergraduate Student,Department of IT,SRGEC,Gudlavalleru

ABSTRACT_ The most common cause of death in the world is cardiovascular disease (CVD). Based on clinical data, a machine learning (ML) system can predict CVD in the early stages to reduce mortality rates. Numerous studies have recently used various machine learning techniques to detect CVD or determine the severity of the patient's condition. Despite the positive outcomes of these studies, none of them concentrated on using optimisation techniques to enhance the ML model's performance for CVD detection and severity level classification. The Synthetic Minority Oversampling Technique (SMOTE), six different ML classifiers to determine the patient's status, and Hyperparameter Optimization (HPO) to determine the ideal hyperparameter for ML classifier in conjunction with SMOTE are all used in this study to provide an efficient method for handling the imbalance distribution issue.The model with all features was constructed and tested on two open datasets. The findings demonstrate that SMOTE and Extra Trees (ET) optimised using hyperband outperformed state-of-the-art works by achieving 99.2% and 98.52% in CVD detection, respectively, outperforming other models. Additionally, using the Cleveland dataset, the developed model converged to a severity classification accuracy of 95.73%. Doctors can determine a patient's current heart disease status using the suggested model. As a result, by starting early therapy, heart disease-related mortality can be avoided.

Keywords Cardiovascular disease, Machine Learning , Classification algorithm

1.INTRODUCTION

The leading cause of death worldwide is heart disease, also known as cardiovascular disease (CVD). The World Heart Federation found in a recent study that cardiovascular disease accounts for one in three deaths [1]. As indicated by World Wellbeing Association (WHO) insights, by 2030, more than 23.6 million individuals might pass on from CVD, primarily from strokes and cardiovascular breakdown [2]. Stress, alcohol, smoking, an unhealthy diet, an inactive lifestyle, and other related health issues like high blood pressure or diabetes are all potential causes of cardiovascular disease (CVD). However, the majority of diseases related to cardiovascular disease (CVD) are completely curable once detected early [3]. Heart failure diagnosis and prediction must be given a higher priority in this instance. New data analysis techniques may make it possible to diagnose cardiovascular disease (CVD) earlier [2].

In general, the lowest-risk level, highest-risk level, presence of primary lesions, and defective diagnosis can all be used to assess a person's health status. Due to natural factors, social and medical status, genetics, and individual lifestyle, the detection process may take longer than anticipated. However, better health outcomes can be achieved by anticipating and evaluating risk factors to prevent illness before it progresses to a more severe level [4]. As a result, a high-accuracy system for detecting heart failure levels using clinical data from heart disease (HD) has been developed. Clinical datasets were used in a number of studies to forecast cardiovascular disease (CVD). In any case, clinical datasets present significant troubles attributable to class unevenness and their high dimensionality. Consequently, employing machine learning without addressing these issues lowers the methods' efficiency and accuracy. Several ML systems and a focus on feature selection (FS) were used by previous researchers to predict CVD.

Long et al. [5] used a type-2 fuzzy logic system for HD detection and a heart disease decision system based on Rough Sets (RS) and the Choas firefly algorithm (CFARS-AR) to select the best features. The developed model was accurate at 88.3 percent. In [6], the authors use a backpropagation neural network (BPNN) and the RS to predict CVD. Moreover, Dwivedi played out a similar report on HD expectation utilizing different ML models, like counterfeit brain organization (ANN), Calculated Relapse (LR), grouping trees, and Innocent Bayes (NB). For CVD detection, the author concludes that LR performed better than the other models [7].

2.LITERATURE SURVEY

Haq et al. also conducted a comparison of various ML models, such as ANN, RF, and LR, with various FS methods, such as relief. The models' performance is impacted by removing features, according to the authors. The study came to the conclusion that LR with relief FS outperformed other models used in the same study by 89% [8]. Amin et al. used seven ML models, including ANN, LR, Decision Trees (DT), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), NB, and a hybrid model (vote with LR & NB) [9], in a comparative analysis that focused on the most important characteristics. The hybrid model had the highest accuracy (87.41 percent), according to the findings. To increase the accuracy of CVD prediction, a different study developed a hybrid model (HRFLM) that combines RF and a linear model. The led research was applied to various blends of highlights; [10] The proposed method was accurate up to 88.4%. A while later, Vijayashree et al. introduced a novel function that uses population diversity and an optimization technique to find the best weights. They also used SVM to improve the PSO fitness function and cut down on the number of attributes. Their method was therefore accurate up to 84.36% [11].

Ali and co. introduced two stacked SVMs, which improved the HD diagnosis procedure. The non-significant attributes were removed by the first SVM, and the presence and absence were predicted by the second SVM. Consequently, the developed model by Stacked had an accuracy of 92.22 percent [12]. In addition, Gupta et al. based a HD diagnosis machine intelligence framework on FAMD and RF. FAMD chose the critical highlights, and RF anticipated the nonappearance and presence of HD. The results demonstrated that the proposed model was accurate at 93.44 percent [13]. The authors of [14] conducted research using comparative analysis. The authors used multiple classifiers on various datasets. The other classifiers were outperformed by the conditional inference tree forest (cforest). Tama and others 15] introduced a heart disease prediction model with a two-tier ensemble and PSO-based feature selection. The created model arrived at exactness up to 93.55%. However, the above-mentioned studies have some limitations when it comes to detecting HD using the suggested methods due to the imbalanced clinical datasets.

Fitriyani and co developed a HD prediction method that used density-based spatial clustering of applications with noise (DBSCAN) and a hybrid synthetic minority over-sampling technique-edited nearest neighbor (SMOTE-ENN) to find and get rid of outliers and bring the distribution of the data back into balance. The XGBoost classifier also predicts the patient's condition. [16] The developed method was accurate to 95.9%. Waqar and co. proposed deep learning based on SMOTE to predict heart attacks. Without feature selection, the author balanced the dataset using the SMOTE technique. A deep neural network was used to train and test the balanced dataset to predict the absence and presence of a heart attack, achieving an accuracy of 96% [17]. In the past, Ishaq et al. utilized SMOTE to balance the distribution of the data and extremely randomized trees (ET) on selected parameters in order to predict patient survival using RF importance ranking [18].

Salari and co. introduced a hybrid approach that used a modified kNN, a backpropagation neural network, and a genetic algorithm (GA) for feature selection and severity classification. The outcomes uncovered that the technique acquired 62.1% precision [19]. In a similar setting, Wiharto et al. proposed a hybrid strategy for predicting HD's sickness level that uses SVM and the binary tree (BT). The study achieved an accuracy of 61.86% [20]. Khateeb et al. don't think so. The data were subjected to a re-sampling filter, and then kNN (IBK) was used to predict the severity level. Up to 79.20% accuracy was achieved by the model [21]. Magesh and Swarnalatha fostered an ideal FS strategy in view of bunch based DT learning (CDTL) to track down the huge highlights and afterward utilized the RF to foresee the seriousness level. For severity level predictions, the developed model was accurate at 89.3 percent [22]. A decision system for HD severity level prediction based on an ML-

based fusion approach was recently developed by the authors in [23]. The findings demonstrated that the developed model was able to predict severity levels with an accuracy of 75%.

3.PROPOSED WORK

This paper's primary contribution can be summed up as follows:

• This study suggested integrating SMOTE, ML classifiers, and HB method into an efficient decision support system. It is believed that combining these methods will increase the effectiveness of current approaches for predicting CVD using clinical datasets.

• Using an optimised classifier and a suitable number of synthesised samples, the HB method determines the best SMOTE hyperparameters to produce the highest prediction.

• The effects of tree-based, statistical-based, and regression-based models with various optimisation algorithms, such as HB, improved Particle Swarm Optimization (PSO), and are discussed in relation to predicting CVD. The Cleveland and Stat log datasets were used to assess the proposed model. To demonstrate the efficacy of the suggested model, numerous comparisons with earlier studies are made.

3.1 IMPLEMENTATION

CVD detection

Based on clinical data, a machine learning (ML) system can predict CVD in the early stages to reduce mortality rates. Numerous studies have recently used various machine learning techniques to detect CVD or determine the severity of the patient's condition. Despite the positive outcomes of these studies, none of them concentrated on using optimisation techniques to enhance the ML model's performance for CVD detection and severity level classification.

The RS and a back propagation neural network (BPNN) are combined by the authors in [6] to predict CVD. Additionally, Dwivedi conducted a comparison study on HD prediction using various ML models, including classification trees, nave bayes, logistic regression, and artificial neural networks (ANN). (NB). The author draws the conclusion that LR outperformed the other models for the detection of CVD.

Severity classification

The model with all features was constructed and tested on two open datasets. The findings demonstrate that SMOTE and Extra Trees (ET) optimised using hyper band outperformed state-of-the-art works by achieving 99.2% and 98.52% in CVD detection, respectively, and outperformed other models. Additionally, using the Cleveland dataset, the developed model converged to a severity classification accuracy of 95.73%. Doctors can determine a patient's current heart disease status using the suggested model. As a result, by starting early therapy, heart disease-related mortality can be avoided.

Hyper parameter optimization

Many decision-tree-based ensemble techniques have been created to improve model performance by combining different decision trees, including ET, RF, and XGBoost models. The bagging method incorporates numerous RF decision trees. With regard to building DTs from all samples and selecting feature sets at random, ET is a different tree-based ensemble learning method that is similar to RF. Additionally, while ET generates splits at random, RF optimises DT splits..

For binary data, LR is frequently used to determine the result of one or more features. The cost function of LR may vary depending on the regularisation method used. The three main types of regularisation are elastic-net, L1-norm, and L2-norm. The regularisation coefficient (C), which describes the regularisation strength, and solver, which stands for the optimisation algorithm, are the other two key components that define LR.

Extra trees

The Synthetic Minority Oversampling Technique (SMOTE), six different ML classifiers to determine the patient's status, and Hyper Parameter Optimization (HPO) to determine the best hyper parameter for ML classifier in conjunction with SMOTE are all used in this study to provide an efficient method for handling the imbalance distribution issue.

kNN, Extra Trees (ET), XGBoost, Gradient Descent (SGD), and LR are used to predict the severity of heart failure. Given that there are numerous ML hyper parameter optimum values for various datasets, the chosen classifiers were SMOTE and HB optimised to find the best performing hyper parameters.

3.2 ALGORITHM:

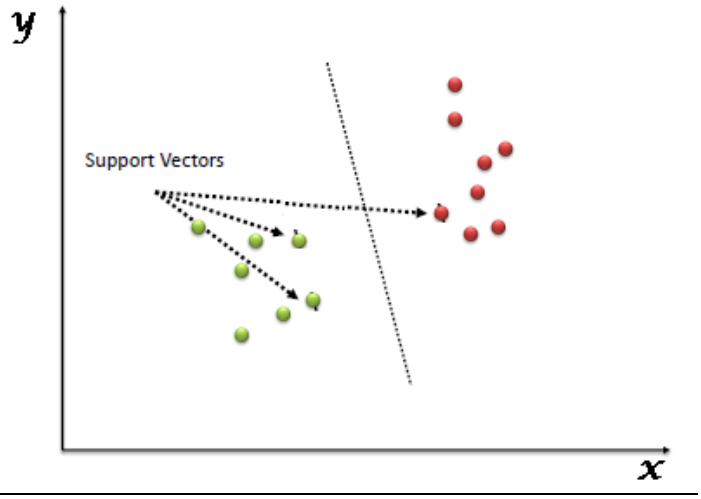
Machine Learning (ML) algorithms were used in numerous studies to predict CVD using clinical datasets. Clinical datasets still pose significant challenges because of class imbalance and high dimensionality. As a result, using machine learning without addressing these issues decreases the methods' efficiency and, consequently, their accuracy.

In addition, some researchers determined the severity of the patients using a hybrid approach. A hybrid approach with a genetic algorithm (GA) for feature selection, a modified kNN, and a back propagation neural network for severity classification was introduced by Salari et al.

Predicting CVD is discussed in relation to the effects of tree-based, statistical-based, and regression-based models with various optimisation algorithms, such as HB, improved Particle Swarm Optimization (PSO) [26], [27], GA, and RS [28].

3.2.1 Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).



Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

The followings are important concepts in SVM −

- **Support Vectors** − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.
- **Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
- **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

3.2.2 KNN

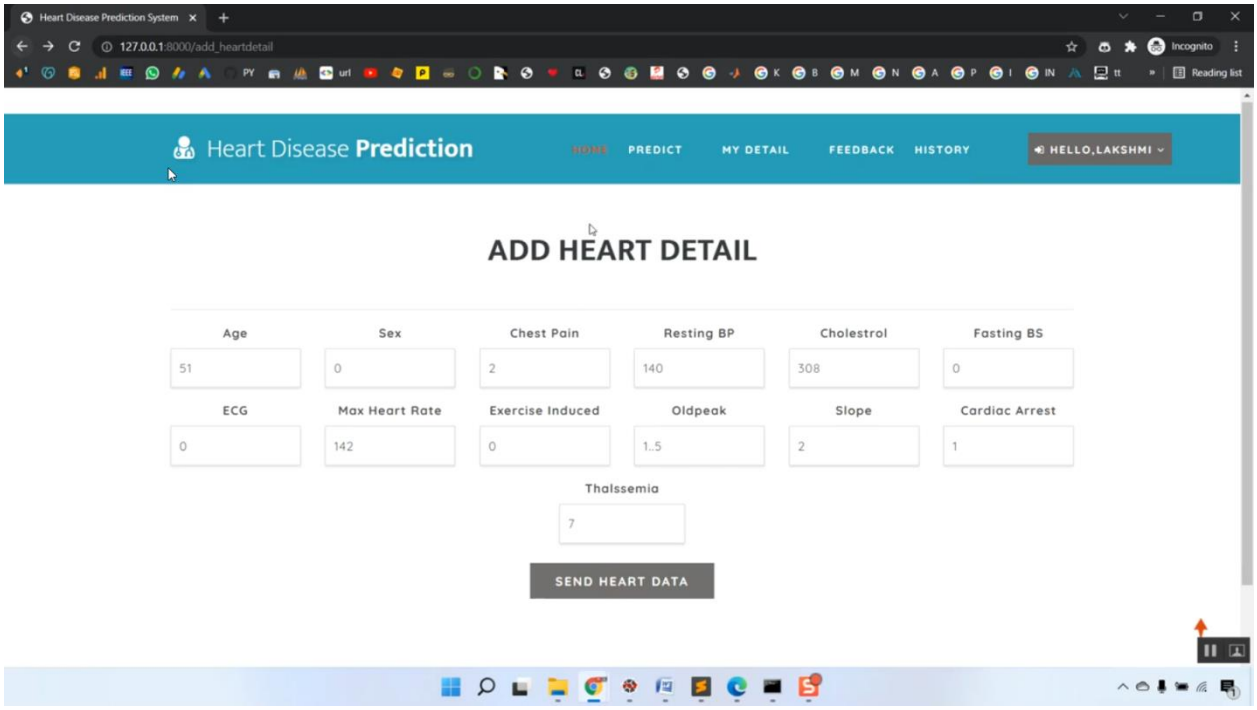
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

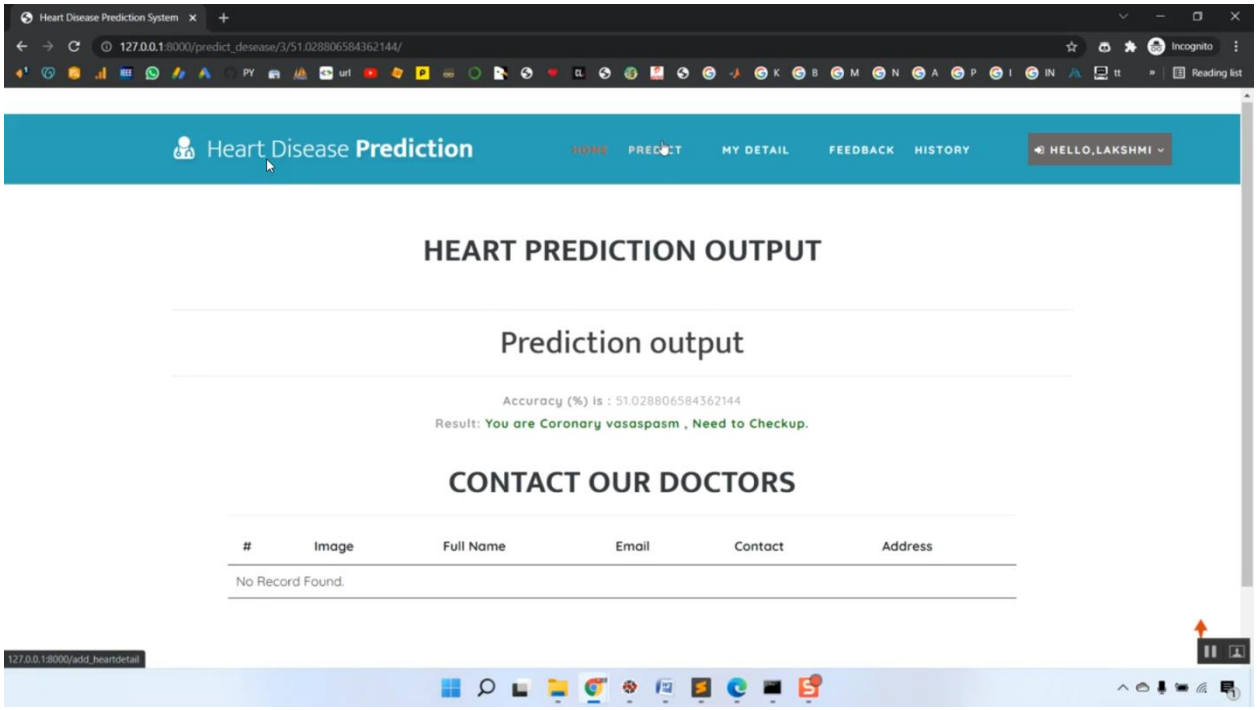
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

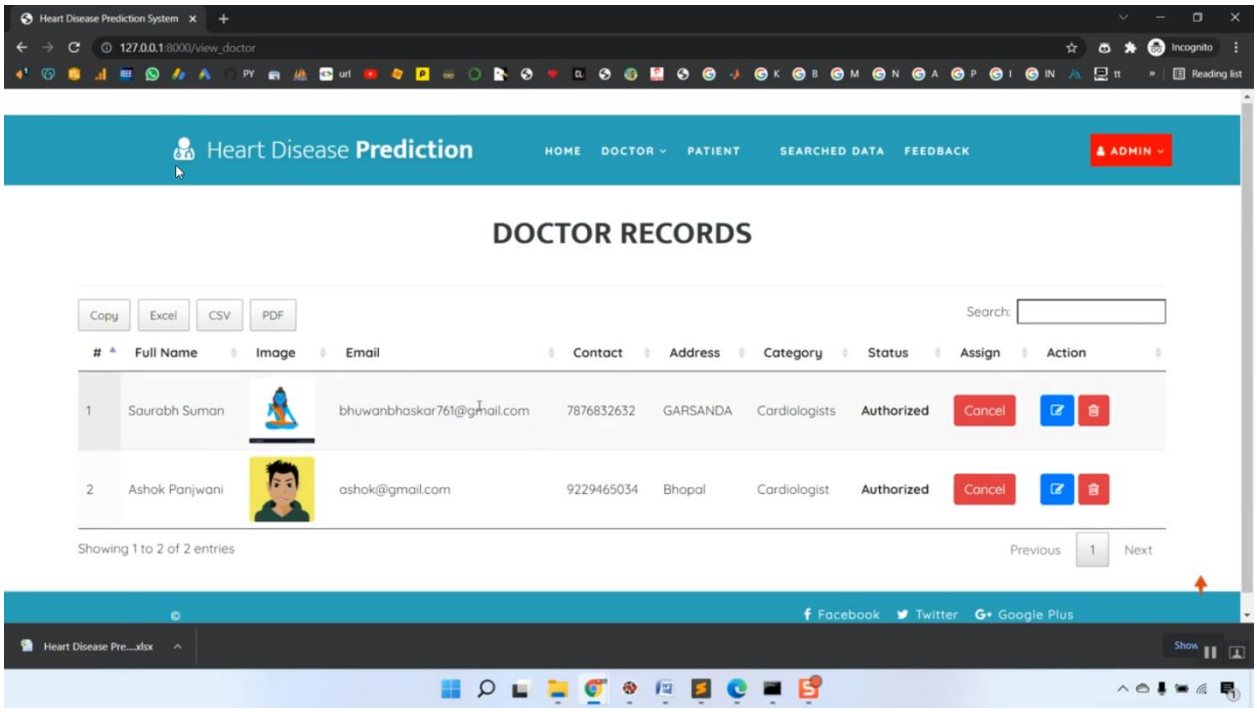
3.2.3 Decision Tree Classification Algorithm

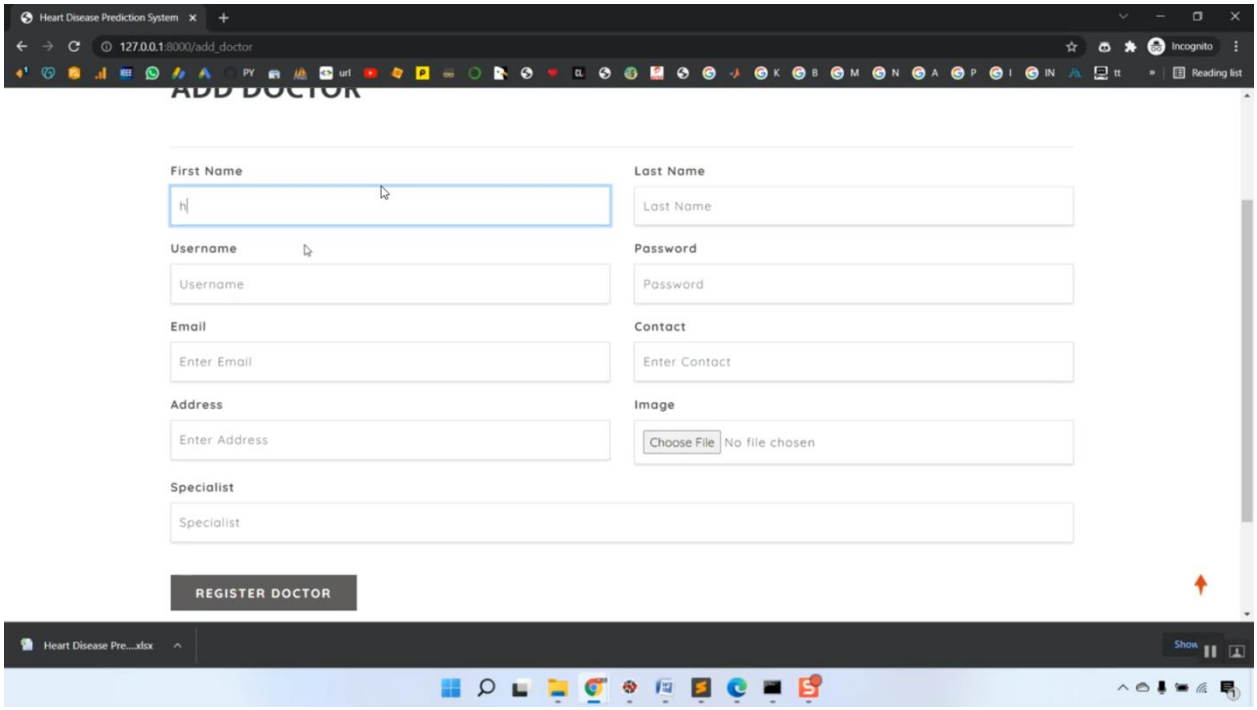
- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
- Below diagram explains the general structure of a decision tree:

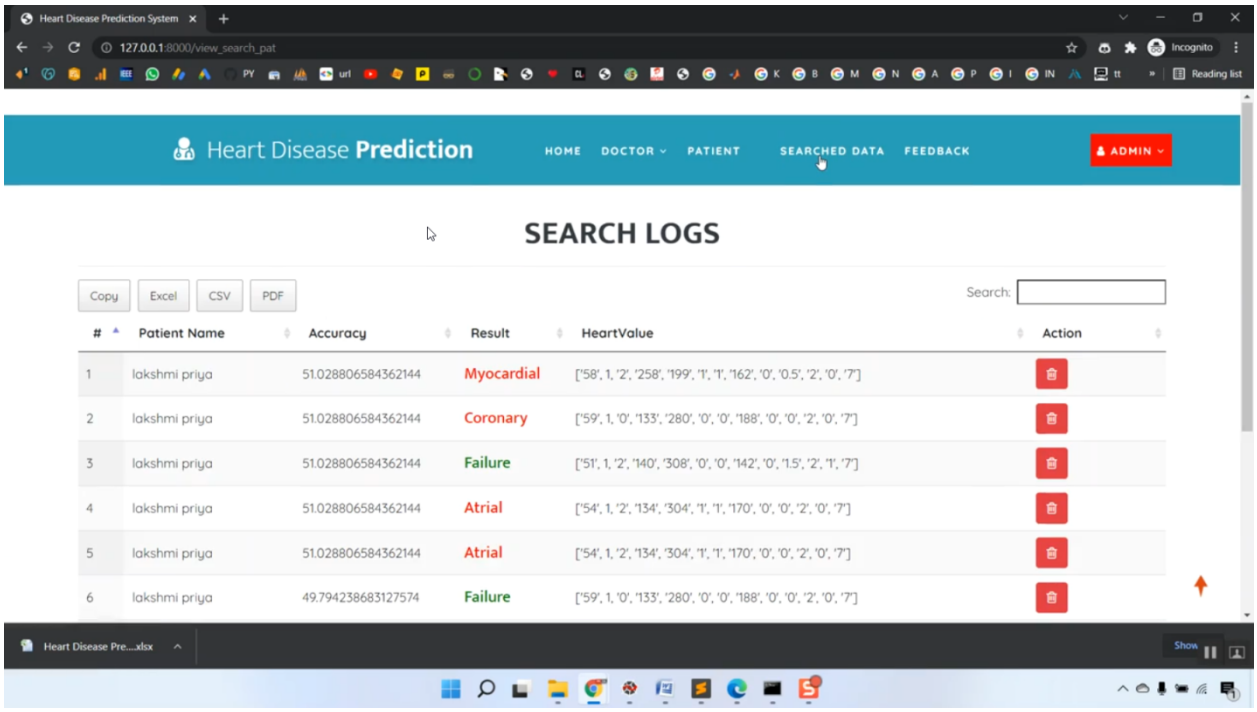
4.RESULTS AND DISCUSSION




Heart Disease Prediction System
127.0.0.1:8000/predict_desease/3/51.028806584362144/
Heart Disease Prediction
HOME
PREDICT
MY DETAIL
FEEDBACK
HISTORY
HELLO,LAKSHMI
HEART PREDICTION OUTPUT
Prediction output
Accuracy (%) is : 51.028806584362144
Result: You are Coronary vasaspasm , Need to Checkup.
CONTACT OUR DOCTORS
Image Full Name Email Contact Address
No Record Found.


Heart Disease Prediction System
127.0.0.1:8000/view_doctor
Heart Disease Prediction
HOME
DOCTOR
PATIENT
SEARCHED DATA
FEEDBACK
ADMIN
DOCTOR RECORDS
Copy
Excel
CSV
PDF
Search:
Full Name Image Email Contact Address Category Status Assign Action
1 Saurabh Suman bhuwanbhaskar761@gmail.com 7876832632 GARSANDA Cardiologists Authorized Cancel
2 Ashok Panjwani ashok@gmail.com 9229465034 Bhopal Cardiologist Authorized Cancel
Showing 1 to 2 of 2 entries
Previous 1 Next
Facebook
Twitter
Google Plus





5.CONCLUSION

In order to predict the presence or absence of CVD and categorise the severity level of the condition, this paper offers an accurate and effective decision support system based on ML modelling. The suggested model combines HB, ET, and SMOTE. Data balancing, classification, and hyperparameter optimisation all use those three techniques. The suggested model is assessed and compared to earlier research. We have provided a statistical assessment of the performance of our model using six performance metrics. The experimental findings demonstrate that HB optimisation has a greater

impact on increasing model accuracy than do tree-based models in producing results of higher quality. As a result, on both datasets, SMOTE and ET optimised by HB had the highest accuracy for binary and multiclass problems.The experimental results showed that, for the Cleveland dataset, our model outperforms state-of-the-art models in terms of accuracy, recall, f1-score, and MCC by up to 99.2%, 99.33%, 99%, and 0.983, respectively, and for the Statlog dataset, 98.52%, 98.08%, 98.08, and 0.969. For multiclass (severity level) prediction, our model achieved 95.73%, 96.35%, 95.73%, 95.78%, and 0.939 in terms of accuracy, precision, recall, f1-score, and MCC, respectively.

Future Work:

This research may be used by medical professionals to forecast heart failure and improve patient care. In order to improve the detection and severity level classification of various diseases using real-time clinical data for our future work, we aim to develop a general framework based on machine learning, including outlier detection and removal, feature selection, and feature selection.

REFERENCES:

[1] R. T. Selvi and I. Muthulakshmi, ‘‘An optimal artificial neural network based big data application for heart disease diagnosis and classification model,’’ J. Ambient Intell. Humanized Comput., vol. 12, no. 6, pp. 6129–6139, 2021.

[2] G. Bazoukis, S. Stavrakis, J. Zhou, S. C. Bollepalli, G. Tse, Q. Zhang, J. P. Singh, and A. A. Armoundas, ‘‘Machine learning versus conventional clinical methods in guiding management of heart failure patients—A systematic review,’’ Heart Failure Rev., vol. 26, no. 1, pp. 23–34, Jan. 2021.

[3] A. Makhlouf, I. Boudouane, N. Saadia, and A. R. Cherif, ‘‘Ambient assistance service for fall and heart problem detection,’’ J. Ambient Intell. Humanized Comput., vol. 10, no. 4, pp. 1527–1546, 2018.

[4] M. Chen, S. Gonzalez, V. Leung, Q. Zhang, and M. Li, ‘‘A 2G-RFIDbased e-healthcare system,’’ IEEE Wireless Commun. Mag., vol. 17, no. 1, pp. 37–43, Feb. 2010.

[5] N. C. Long, P. Meesad, and H. Unger, ‘‘A highly accurate firefly based algorithm for heart disease prediction,’’ Expert Syst. Appl., vol. 42, no. 21, pp. 8221–8231, 2015.

[6] K. B. Nahato, K. N. Harichandran, and K. Arputharaj, ‘‘Knowledge mining from clinical datasets using rough sets and backpropagation neural network,’’ Comput. Math. Methods Med., vol. 2015, pp. 1–13, Mar. 2015.

[7] A. K. Dwivedi, ‘‘Performance evaluation of different machine learning techniques for prediction of heart disease,’’ Neural Comput. Appl., vol. 29, no. 10, pp. 685–693, 2018.

[8] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, ‘‘A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms,’’ Mobile Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018.

[9] M. S. Amin, Y. K. Chiam, and K. D. Varathan, ‘‘Identification of significant features and data mining techniques in predicting heart disease,’’ Telematics Inform., vol. 36, pp. 82–93, Mar. 2019.

[10] S. Mohan, C. Thirumalai, and G. Srivastava, ‘‘Effective heart disease prediction using hybrid machine learning techniques,’’ IEEE Access, vol. 7, pp. 81542–81554, 2019.