iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** V **Month of publication:** May 2021

DOI: https://doi.org/10.22214/ijraset.2021.34735

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

E-Healthcare System for the Diagnosis of Heart Disease using MSSO-ANFIS

P. Pugalendhi[1], Mr. V. Sornagopal[2], Mrs. K. Periyarselvam[3], Dr. P. Sivakumar[4]
[1]*PG Student,* [2, 3, 4]*Associate Professor, Department of ECE, GRT Institute of Engineering and Technology, Tiruttani, India*

Abstract: ***Machine Learning is used in a variety of fields all over the world. There is no exception in the healthcare industry. Machine Learning can help forecast the existence or absence of heart problems. Heart disease is a complex condition that affects many people around the world in healthcare. In the field of cardiology, prompt and correct diagnosis of cardiac disease is critical. This system is used for diagnosing cardiac illness that is efficient, correct, and it is based on machine learning techniques. To solve the feature selection problem, the suggested system uses a novel Fast Conditional Mutual Information feature selection technique. The feature selection techniques are used to boost the classification accuracy and lower the classification system's execution time. The goal of this study article is to predict the likelihood of patients having heart disease. Gradient achieves the highest accuracy score, according to the results.***
Keywords: Machine Learning, Prediction, LT, Naïve bias, Decision Tree

I. INTRODUCTION

According to the World Health Organization, heart disease kills 17 million people worldwide each year. The global burden of cardiovascular disease has been rapidly increasing in recent years. Many types of studies have been carried out in an attempt to define the most important factors in heart disease and to precisely forecast the overall risk. Heart disease is also referred to as a "silent killer" because it causes death without causing noticeable symptoms. Early detection of cardiac disease is critical for implementing lifestyle modifications in high-risk people and, as a result, reducing consequences. This study tries to predict future heart illness by evaluating patient data and using machine-learning algorithms to classify whether they have heart disease or not.

Data mining is becoming increasingly popular in the healthcare sector, where there is a need for an effective analytical process for finding unknown and useful information.

Cardiovascular disease (sometimes known as heart disease) refers to a group of illnesses affecting the heart and blood vessels (arteries, capillaries, and veins). Even though cardiovascular disease is the leading cause of death worldwide, cardiovascular mortality rates have declined in several high-income countries since the 1970s. At the same time, cardiovascular mortality and disease in low and middle-income nations have risen dramatically.

Although cardiovascular disease primarily affects older persons, its antecedents, particularly atherosclerosis, begin in childhood, necessitating primary prevention efforts as early as childhood. As a result, there is a greater emphasis on lowering risk factors to prevent atherosclerosis. Age, gender, high blood pressure, high serum cholesterol levels, smoking, excessive alcohol consumption, sugar consumption, family history, obesity, lack of physical activity, psychosocial factors, diabetes mellitus, air pollution, and tobacco use are all risk factors for heart disease, according to evidence. According to the Globe Health Statistics 2012 report, one in every three persons in the world has high blood pressure, a condition that accounts for over half of all stroke and heart disease deaths. In many nations, including India, heart disease is the leading cause of death. In the United States, one person dies from heart disease every 34 seconds. Diagnosis is a difficult and vital activity that must be completed correctly and quickly.

The computing capability of both phones and access points is predicted to be sufficient enough to execute NOMA algorithms by 2020 when 5G networks are planned to be implemented. The modulation strategy is orthogonal frequency division multiplexing (OFDM), and the multiple access strategy is NOMA. Orthogonal frequency division multiple access (OFDMA) is utilized in standard 4G networks as a natural extension of OFDM, where information for each user is assigned to a subset of subcarriers. In NOMA, on the other hand, each user has access to all of the subcarriers. For two users, Figure 1 depicts spectrum sharing for OFDMA and NOMA. Both uplink and downlink transmissions are included under this idea. The major goal of this project is to perform user clustering and power distribution in such a way that the network's energy efficiency is maximized. To increase the spectral efficiency of a network in which two pairs of users share the same channels for sending data. To eliminate interference from other users whose signals are broadcast on the same spectrum. The sequential convex approximation is used to deal with non-convexity. To reduce network interference and improve network efficiency.

II. RELATED WORKS

The system was developed by authors A.U. Haq, J. Li, M.H. Memon, J. Khan, and S.M. Marium, who used a Sequential backward selection feature approach to choose an important number of features for the best classification accuracy. To train and test the classifier performance, the supervised learning classifier K-NN was utilized in the system for classification with a training and testing split technique. Preprocessing, feature selection, training/testing split, classifier, and classifier performance measurement metrics are the five steps in the suggested system technique.

Author U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun propose a hybrid ensemble model for heart disease detection and prediction that focuses on predicting labels of each SPECT picture based on feature vectors and labels assigned by base classifiers. The features of the proposed model's layout are described in this section to aid understanding of the proposed framework. The proposed hybrid ensemble model is depicted schematically. The partitioning module, inner classifiers module, and fuser module are the three modules that make up the system. The original dataset is sent to the partitioning module, which divides it into train and test subsets and prepares it for the following module. Different classification algorithms are applied to the train and test datasets in the inner classifiers module to produce input data for the fuser module, which considers the results of base classifiers alongside the initial feature vector of samples for building and adjusting components of the final classifier. The remainder of this section contains a summary of each component.

Author X. Liu, X. Wang, Q. SU, M. Zhang, Y. Zhu, Q. Wang, and Q. Wang tells the three advantages of combining Relief F and RS (RFRS) approach as an integrated feature selection system for heart disease diagnosis. The RFRS approach is more effective in removing extraneous and duplicate features. The Relief F algorithm can choose relevant features for disease diagnosis, but the selected relevant features may still contain duplicated characteristics. In such circumstances, the RS reduction method can be used to reduce any remaining redundant features, thus overcoming the Relief F algorithm's constraint.

III.PROPOSED METHODOLOGY

In this study, the proposed methodology is a machine learning-based diagnostics technique for detecting HD. For the detection of HD, machine learning predictive models such as LR, K-NN, SVM, DT, and NB are applied. For feature selection, a Fast Conditional Mutual Information (FCMIM) features selection technique was also developed. Aside from that, many performance assessment measures have been employed to evaluate the performance of classifiers. On the Cleveland HD dataset, the proposed approach was tested.

A. Pseudo-Code of Proposed Heart Disease Diagnosis System

1) Begin
2) Using preprocessing approaches, preparing a heart disease dataset
3) Selection of features utilizing current state-of-the-art and proposed FCMIM FS algorithms
4) Use the training dataset to train the classifiers
5) Validate using a test dataset
6) Calculates metrics for performance evaluation
7) End

Determine whether a patient needs to be diagnosed with heart disease. This is a one-of-a-kind situation. Positive (+) = 1 patient with a heart condition. Negative (-) Signifies a patient who has not been diagnosed with heart disease. Experiment with different Classification Models to determine which one gives you the best results. Examine our data for trends and relationships. Decide which characteristics are most essential to Positive/Negative Heart Disease diagnosis.

B. Features & Predictor

Our Predictor (Y, Positive or Negative Heart Disease Diagnosis) is based on 13 characteristics (X):

1) age (#)
2) sex: 1 denotes a male, 0 denotes a female (Binary)
3) (cp)type of chest pain (4 values -Ordinal): Value 1 indicates conventional angina, Value 2 indicates atypical angina, Value 3 indicates non-anginal discomfort, and Value 4 indicates silent angina.
4) resting blood pressure (#) (trestbps)
5) (chol) serum cholesterol in milligrams per deciliter (#)
6) (fbs)fasting blood sugar > 120 mg/dl (Binary) (1 = true; 0 = false

7) (restecg) resting electrocardiography results (values 0,1,2)
8) (thalach) maximum heart rate achieved (#)
9) (exang) exercise induced angina (binary) (1 = yes; 0 = no)
10) (oldpeak) = ST depression induced by exercise relative to rest (#)
11) (slope) of the peak exercise ST-segment (Ordinal) (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
12) (ca) number of major vessels (0–3, Ordinal) colored by fluoroscopy
13) (thal) maximum heart rate achieved — (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect
Note: Our data has 3 types of data:
Continuous (#): which is quantitative data that can be measured
Ordinal Data: Categorical data that has an order to it (0,1,2,3, etc)
Binary Data: data whose unit can take on only two possible states (0 &1)

C. Data Wrangling
It looks that the two binary outputs are in proper proportion.

D. Exploratory Data Analysis Correlations
The Correlation Matrix displays all the variables' correlations. You may determine whether something is positively or negatively connected with our prediction in a matter of seconds (target). Calculate correlation matrix. We can see there is a positive correlation between chest pain (CP) & target (our predictor). This makes sense because the more chest pain you have, the more likely you are to develop heart disease. CP (chest pain) is a four-valued ordinal feature: Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic. Furthermore, we observe a negative relationship between our predictor and exercise-induced angina (exang). This makes sense because your heart wants more blood while you exercise, but narrower arteries restrict blood flow. Pair plots are also a great way to immediately see the correlations between all variables. But you'll see that I'm only using continuous columns from our data because there are so many features that it's tough to see them all. As a result, I'm going to build a pair plot using solely our continuous features. When the ventricle is at rest and so repolarized, ST-segment depression occurs. This Heart Disease can be caused by a trace in the ST-segment that is excessively low below the baseline. This backs up the graph above because persons with low ST Depression have a higher risk of heart disease. While a high ST depression is considered normal & healthy. The "slope" color denotes the peak workout ST-segment, with values ranging from 0 to 2 (upsloping, flat, and downsloping). The three slope categories are distributed equally among both positive and negative heart disease patients.

E. Violin & Box Plots
The Box & Violin plots have the advantage of displaying the data's basic statistics as well as its distribution. These graphs are frequently used to compare the distribution of a variable across different categories.
It displays the median, interquartile range, and Turkey's fence. (the minimum, the first quartile (Q1), the median, the third quartile (Q3), and the maximum.)
As can be seen, the overall shape and distribution of negative and positive patients are dramatically different. Positive individuals have a lower median for ST depression and consequently a wide range of data between 0 and 2, whereas negative individuals have a range between 1 and 3. In addition, we don’t see many differences between male & female target outcomes.
Positive patients have a higher median for ST depression, whereas negative individuals have a lower median for ST depression. Furthermore, we don't notice many differences between male and female target outcomes, except males having slightly wider ST Depression ranges.

F. Filtering Data by Positive & Negative Heart Disease patient
When we compare the means of positive and negative patients, we can observe that there are significant disparities in many of our 13 Features. We can see those positive patients have a higher maximal heart rate attained (thalach) average when we look at the details. Furthermore, when compared to rest, positive patients experience around 1/3rd the degree of ST depression generated by activity (oldpeak).
Machine Learning + Predictive Analytics Prepare Data for Modelling
Simply remember ASN when preparing data for modeling (Assign, Split, Normalize). Assign the 13 features to X, and our classification prediction, y, to the last column.

G. Modeling /Training

Now we'll use the training set to train multiple Classification Models and evaluate which ones get the best results. We will compare the accuracy of Logistic Regression, K-NN (k-Nearest Neighbours), SVM (Support Vector Machine), Naive Bayes Classifier, Decision Trees, Random Forest, and XGBoost. Note: these are all supervised learning models.

H. Interpret Confusion Matrix

The number of True Positives in our data is 21, whereas the number of True Negatives is 28.
The number of errors is 9 and 3.
There are nine Type 1 errors (Untrue Positives), which occur when you forecast a positive outcome, but it turns out to be false.
There are three types of Type 2 errors (False Negatives): you predicted a negative, and it turned out to be false.
As a result, the accuracy is calculated as # Correct Predicted/ # Total.
In other words, the number of true positives, false negatives, false positives, and true negatives is represented as TP, FN, FP, and TN.
Accuracy = (TP + TN)/(TP + TN + FP + FN).
Accuracy = (21+28)/(21+28+9+3) = 0.80 = 80% accuracy

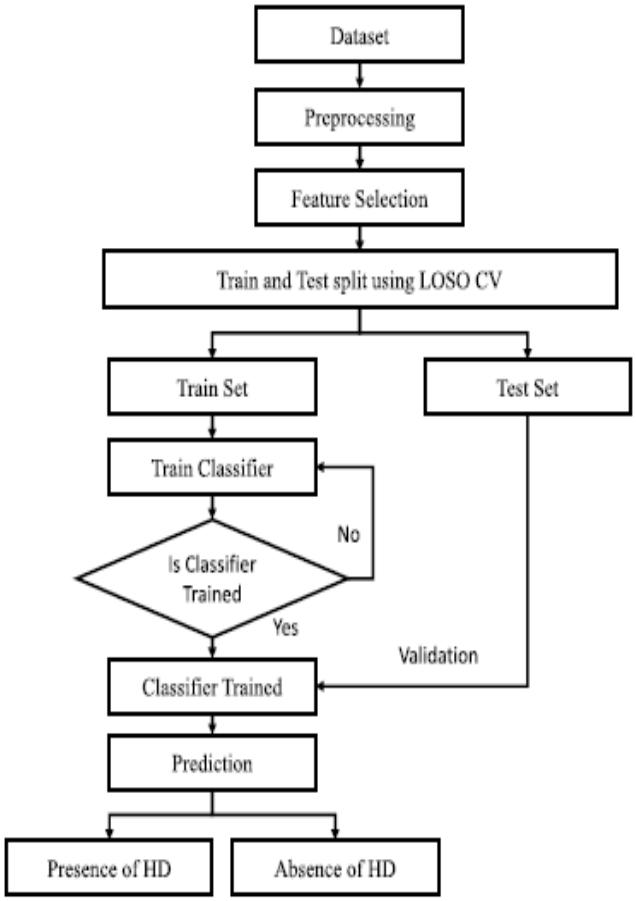


Figure 3.1 Proposed System Flow Diagram

IV. RESULTS AND DISCUSSIONS

On the Kaggle website, the dataset is open to the public. Age, sex, type of chest pain, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, exercise induced ST depression, the slope of the peak exercise, number of major vessels, and target ranging from 1 to 2, where 1 is the absence of angina and 2 is the presence of heart disease. The data collection is in CSV (Comma Separated Value) format, which is then converted to a data frame using the panda's Python tool. The education data is irrelevant to an individual's heart illness, so it is removed. Pre-processing and experiments are then carried out on this dataset.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue V May 2021- Available at www.ijraset.com

A. Data Preparation

In this work, the Cleveland Heart Disease dataset is used for testing purposes. There were 303 occurrences and 75 attributes when this data set was designed, but all published experiments only use a subset of 14 of these. We pre-processed the data set in this study, and six samples were excluded due to missing values. There are 297 samples left in the dataset with 13 characteristics and one output label. The absence of HD and the presence of HD are described by two classes in the output label. As a result, a 297*13 features matrix of extracted features is created.

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4235	0	48	2.0	1	20.0	NaN	0	0	0	248.0	131.0	72.0	22.00	84.0	86.0
4236	0	44	1.0	1	15.0	0.0	0	0	0	210.0	126.5	87.0	19.16	86.0	NaN
4237	0	52	2.0	0	0.0	0.0	0	0	0	269.0	133.5	83.0	21.47	80.0	107.0
4238	1	40	3.0	0	0.0	0.0	0	1	0	185.0	141.0	98.0	25.60	67.0	72.0
4239	0	39	3.0	1	30.0	0.0	0	0	0	196.0	133.0	86.0	20.91	85.0	80.0

4240 rows × 16 columns

Figure Dataset with Sample Attributes

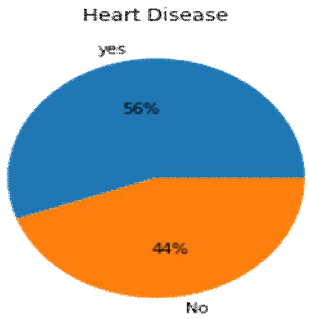


Figure. Heart Disease

B. Exploratory Analysis

Visualization of the Correlation Matrix before feature selection. It demonstrates that no single feature has a strong relationship with our desired value. In addition, some traits have a negative association with the goal value, while others have a positive correlation. Plots and bar graphs were also used to illustrate the data.

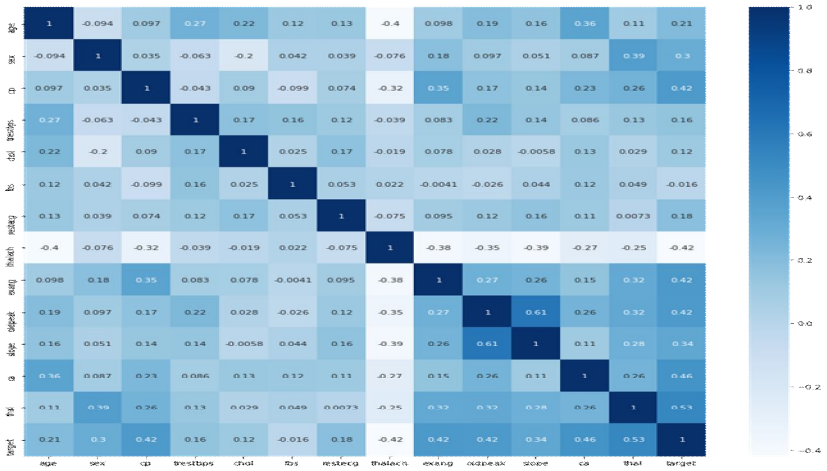
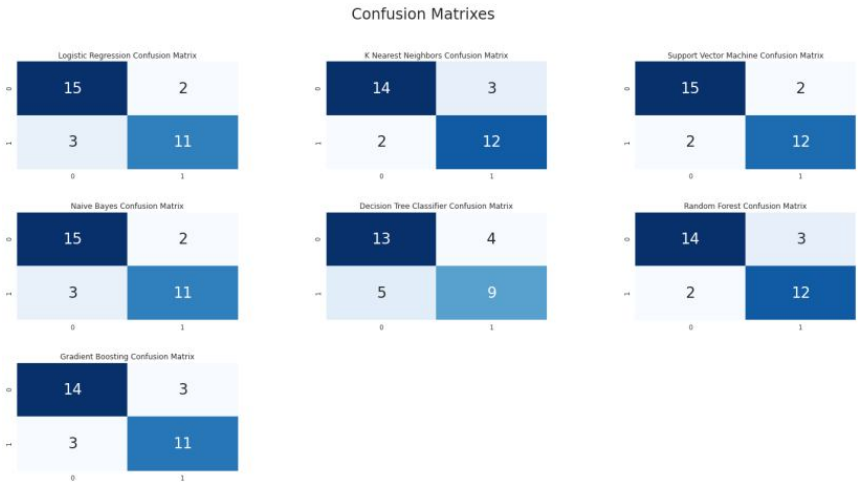


Figure Correlation Matrix Visualization



Confusion Matrix

Training MultinomialNB
Training MultinomialNB finished in 0.00 sec

Training DecisionTreeClassifier
Training DecisionTreeClassifier finished in 0.00 sec

Training LinearSVC
Training LinearSVC finished in 0.00 sec

Training AdaBoostClassifier
Training AdaBoostClassifier finished in 0.07 sec

Training RandomForestClassifier
Training RandomForestClassifier finished in 0.14 sec

Training BaggingClassifier
Training BaggingClassifier finished in 0.02 sec

Training LogisticRegression
Training LogisticRegression finished in 0.01 sec

Training SGDClassifier
Training SGDClassifier finished in 0.00 sec

SGDClassifier

Optimized Model

Best Parameters: {'alpha': 0.0003, 'max_iter': 3000}
Accuracy: 0.7209
F1-score: 0.7600
Precision: 0.7917
Recall: 0.7308
LogisticRegression

Optimized Model

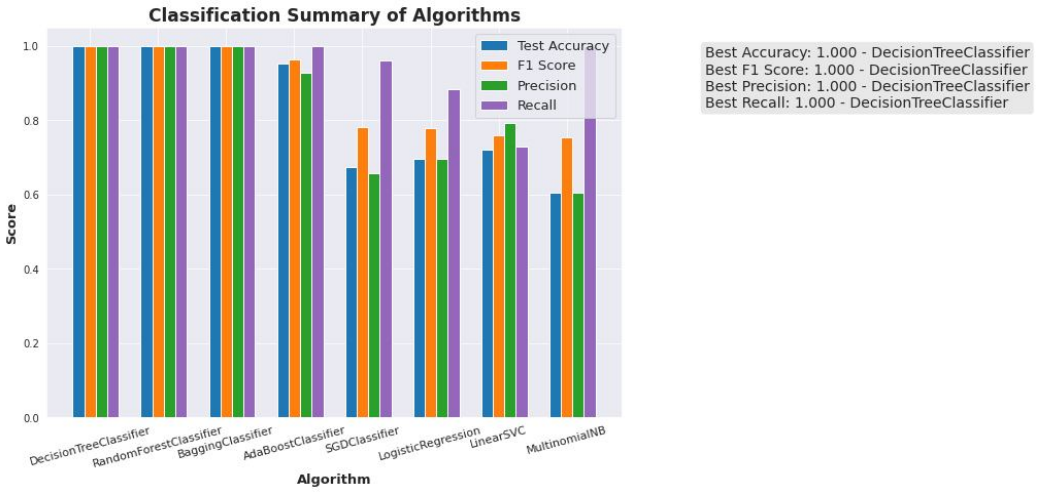
Best Parameters: {'C': 1.2}
Accuracy: 0.6744
F1-score: 0.7586
Precision: 0.6875
Recall: 0.8462
DecisionTreeClassifier

Optimized Model

Best Parameters: {'min_samples_leaf': 8, 'min_samples_split': 2}
Accuracy: 0.8605
F1-score: 0.8846
Precision: 0.8846
Recall: 0.8846
RandomForestClassifier

Optimized Model

Best Parameters: {'min_samples_leaf': 5, 'min_samples_split': 2, 'n_estimators': 50}
Accuracy: 0.9535
F1-score: 0.9630
Precision: 0.9286
Recall: 1.0000

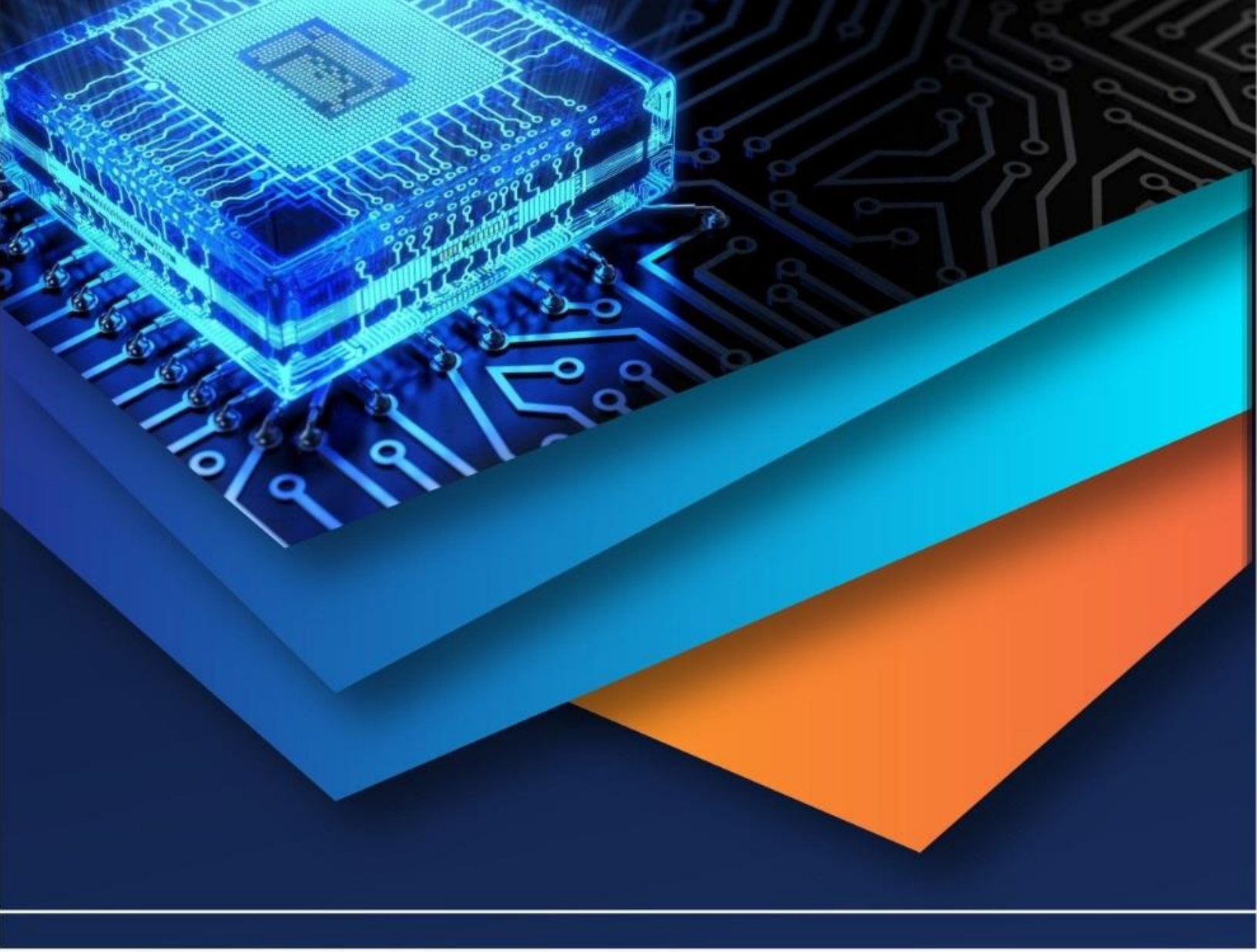


V. CONCLUSIONS

In this work, an effective machine learning-based diagnostics system for the diagnosis of heart disease was built. Chest pain type (CP), the highest heart rate obtained (thalach), number of major vessels (ca), and ST depression generated by exercise relative to rest were the top four important features that allowed us to identify between a positive and negative diagnosis out of the 13 features we looked at rest (oldpeak). Patients with Heart Disease can now be classified using our machine learning system. We can now accurately diagnose patients and provide them with the assistance to recover. By diagnosing detecting these features early, we may prevent worse symptoms from arising later. With an accuracy of 80%, our Random Forest algorithm is the most accurate. Any accuracy of more than 70% is considered good, but be cautious because excessively high accuracy may be too good to be true (an example of Overfitting). As a result, 80 percent accuracy is desirable. To improve it, we can train on models and anticipate the types of cardiovascular problems that users would face, as well as employ more advanced models.

REFERENCES

[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ``Epidemiology and risk profile of heart failure," Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.
[2] M. Durairaj and N. Ramasamy, ``A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate, " Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255_260, 2016.
[3] L. A. Allen, L.W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus, ``Decision making in advanceNUd heart failure: A scientific statement from the American heart association," Circulation, vol. 125, no. 15, pp. 1928_1952, 2012.
[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ``Innovative artificial neural networks- based decision support system for heart diseases diagnosis," J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.
[5] Q. K. Al-Shayea, ``Artificial neural networks in medical diagnosis," Int. J. Comput. Sci. Issues, vol. 8, no. 2, pp. 150_154, 2011.
[6] J. Lopez-Sendon, ``The heart failure epidemic," Medicographia, vol. 33, no. 4, pp. 363_369, 2011.
[7] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ``Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quanti_cation of average Parkinson's disease symptom severity," J. Roy. Soc. Interface, vol. 8, no. 59, pp. 842_855, 2011.
[8] S. I. Ansarullah and P. Kumar, ``A systematic literature review on cardiovascular disorder identification using knowledge mining and machine learning method," Int. J. Recent Technol. Eng., vol. 7, no. 6S, pp. 10091015, 2019.
[9] S. Nazir, S. Shahzad, S. Mahfooz, and M. Nazir, ``Fuzzy logic based decision support system for component security evaluation," Int. Arab J. Inf. Technol., vol. 15, no. 2, pp. 224231, 2018.
[10] R. Detrano, A. Janosi, W. Steinbrunn, M. Psterer, J.-J. Schmid, S. Sandhu, K. H. Guppy, S. Lee, and V. Froelicher, ``International application of a new probability algorithm for the diagnosis of coronary artery disease," Amer. J. Cardiol., vol. 64, no. 5, pp. 304310, Aug. 1989.
[11] J. H. Gennari, P. Langley, and D. Fisher, ``Models of incremental concept formation," Artif. Intell., vol. 40, nos. 13, pp. 1161, Sep. 1989.
[12] Y. Li, T. Li, and H. Liu, ``Recent advances in feature selection and its applications," Knowl. Inf. Syst., vol. 53, no. 3, pp. 551577, Dec. 2017.
[13] J. Li and H. Liu, ``Challenges of feature selection for big data analytics," IEEE Intell. Syst., vol. 32, no. 2, pp. 915, Mar. 2017.
[14] L. Zhu, J. Shen, L. Xie, and Z. Cheng, ``Unsupervised topic hypergraph hashing for efficient mobile image retrieval," IEEE Trans. Cybern., vol. 47, no. 11, pp. 39413954, Nov. 2017.
[15] S. Raschka, ``Model evaluation, model selection, and algorithm selection in machine learning," 2018, arXiv:1811.12808. [Online]. Available: http://arxiv.org/abs/1811.12808
[16] S. Palaniappan and R. Awang, ``Intelligent heart disease prediction system using data mining techniques," in Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl., Mar. 2008, pp. 108115.
[17] E. O. Olaniyi, O. K. Oyedotun, and K. Adnan, ``Heart diseases diagnosis using neural networks arbitration," Int. J. Intell. Syst. Appl., vol. 7, no. 12, p. 72, 2015.
[18] R. Das, I. Turkoglu, and A. Sengur, ``Effective diagnosis of heart disease through neural networks ensembles," Expert Syst. Appl., vol. 36, no. 4, pp. 76757680, May 2009.
[19] O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, ``An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," Expert Syst. Appl., vol. 68, pp. 163172, Feb. 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY