IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 4, April 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104223 |

Efficient Detection of Congestive Heart Failure Detection through Machine Learning

Professor. A.N.Gedam[*1]Parth Adesara[*2], Anil Biradar[*3], Pratik Lokhande[*4], Abhishek Galve[*5]

Professor, Dept. of Computer Engineering, AISSMS polytechnic Pune, Maharashtra, India[*1]

UG Student, Dept. of Computer Engineering AISSMS polytechnic Pune, Maharashtra, India[*2*3*4*]

ABSTRACT: This paper presents a majority voting ensemble method that is able to predict the possible presence of heart disease in humans. The prediction is based on simple affordable medical tests conducted in any local clinic. Moreover, the aim of this project is to provide more confidence and accuracy to the Doctor's diagnosis since the model is trained using real-life data of healthy and ill patients. The model classifies the patient based on the majority vote of several machine learning models in order to provide more accurate solutions than having only one model. Finally, this approach produced an accuracy of 90% based on the hard voting ensemble model. . We also proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyperparameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers.

KEYWORDS: Machine learning; Majority Voting ensemble method; heart disease; UCI dataset; classification.

I. INTRODUCTION

In the present era, heart disease rates have dramatically increased to become the leading cause of death in the United States upon adults due to the widespread of unhealthy habits. These include a declination in physical activity since the technology trend is moving towards replacing human physical activity and unhealthy eating habits which are directly linked to increasing the risk of having heart diseases. Starting off with the definition of a Heart Disease, according to the National Heart, Lung, and Blood Institute states that heart disease is a disruption to the heart's normal electrical system and pumping functions. Where the disease makes it harder for the heart muscle to pump blood efficiently. Furthermore, according to the World Health Organization (WHO), 17.9 million people die each year from cardiovascular diseases which correspond to 31% of all deaths around the world [3]. This incurs the need of having an affordable system that is able to give a preliminary assessment of a patient based on relatively simple medical tests that are affordable to everyone. To conduct the training and testing of the machine learning model, the Cleveland dataset from the well-known UCI repository was used since it is an authenticated dataset that is widely used for training and testing in machine learning models . The dataset contains 303 instances and 14 attributes that are based on well-known factors that are thought to correlate with risks of heart diseases. The approach presented in this paper uses the hard voting ensemble method which is a technique where multiple machine learning models are combined and the prediction result is based on the majority vote from all models. This technique is used in order to improve the overall prediction J48 model that is 56.76% with a total build model time of 0.04 seconds. Finally, the work in deploys various machine learning models in order to investigate the highest performance metric (Accuracy, Sensitivity, Specificity, and Kappa). The machine learning algorithms involve Random Forest, Logistic Regression and Artificial Neural Network. Cross-Industry Standard Process for Data mining technique (CRISP-DM) is used to find insights and meaningful information from the data. The CRISP-DM involves six stages that were followed in this research. Moreover, the accuracies obtained from the models used were as follows; 80.9 % for the Random forest, 79.78% for the Artificial Neural Network, and 85.39% for the Logistic Regression.

II. RELATED WORK

In the field of heart disease detection, a variety of techniques regarding data preprocessing and model variation has been used. The work presented in [5] used the same dataset as this paper but different machine learning models were implemented. Three discrete classifier models were built which included Support Vector Machine (SVM) classifier, naïve Bayes algorithm, and

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

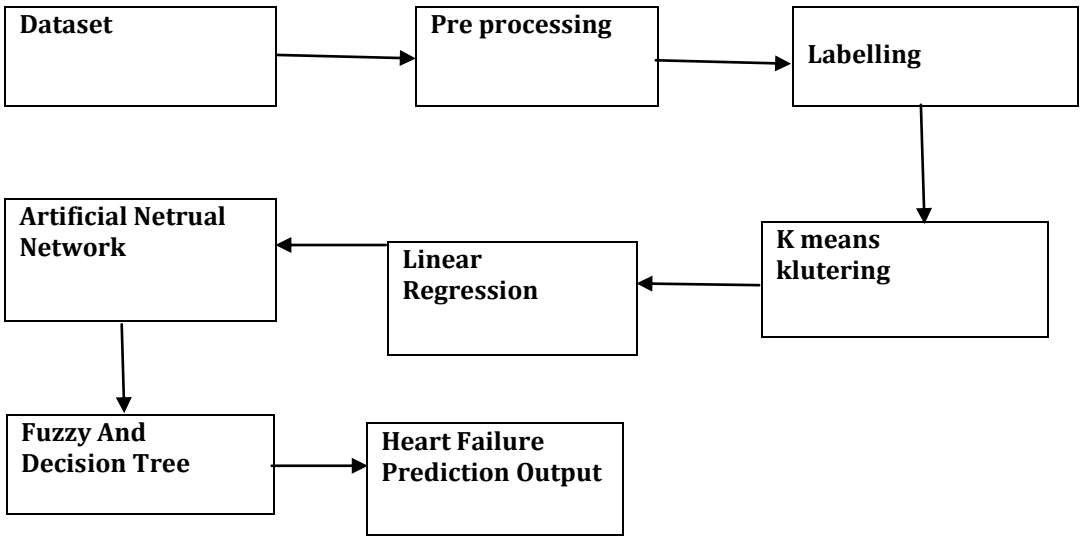
| DOI: 10.15680/IJIRCCE.2023.1104223 |

C4.5. The prediction of the heart disease was conducted based on each of these models discretely and produced a maximum accuracy of 84.12% in the SVM machine learning model.

III. METHODOLOGY

The system has been designed for the identification of heart disease. The performances of various machine learning classifiers for HD identification have been checked on selected features. The standard state of art algorithms of features selection includes Relief, MRMR, LASSO, and LLBFS are utilized for features selection. We also proposed FCMIM algorithm for features selection. The performance of the classifiers evaluated on selected features sets which are selected by the state of the art FS algorithms and proposed FCMIM algorithm. The LOSO technique of cross-validation also used for best model evaluation. The model's performance measuring metrics include accuracy, specificity, sensitivity, MCC and processing time is automatically calculated for classifiers evaluation. The proposed system methodology is organized.

SYSTEM ARCHITECTURE



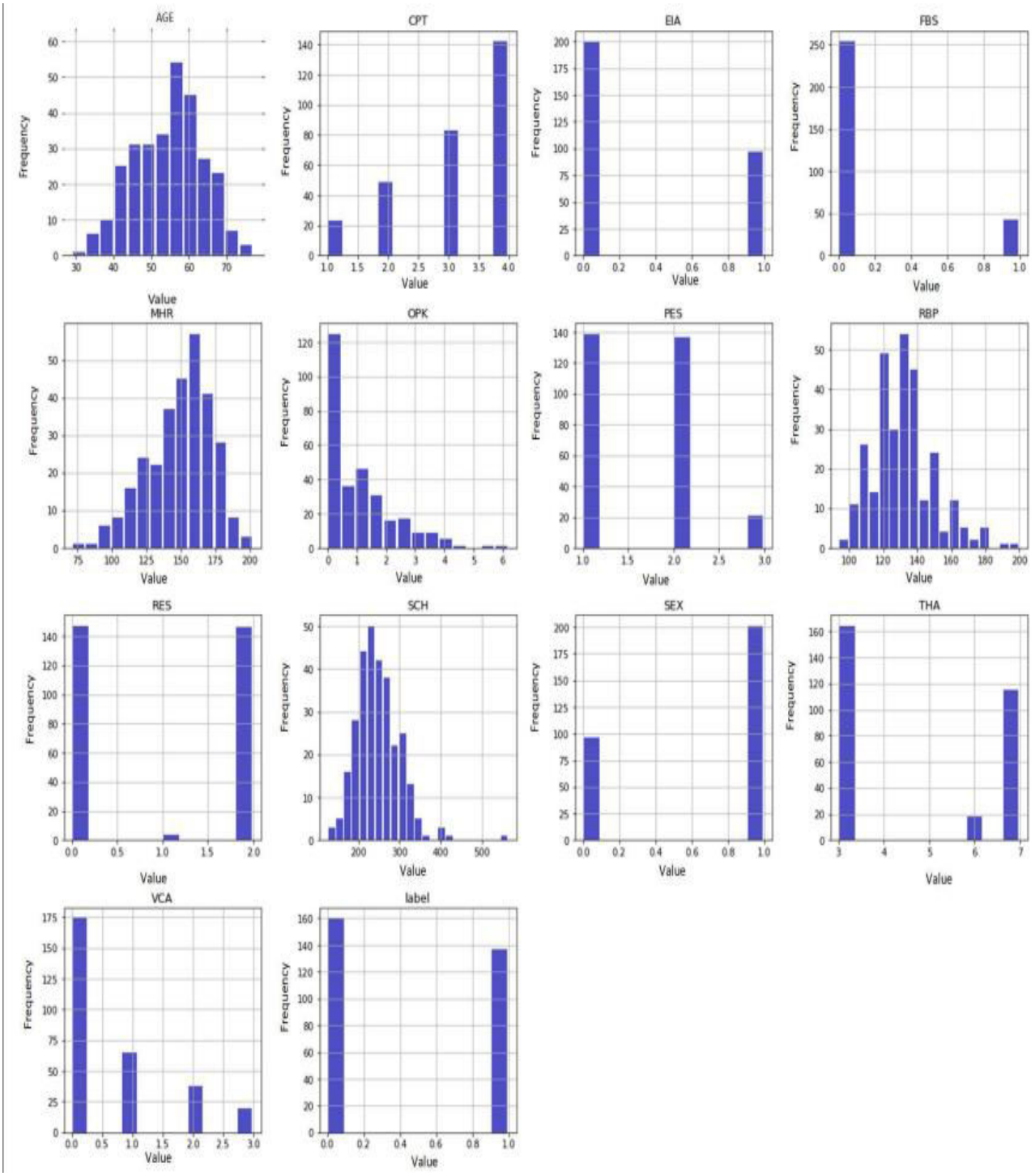
S.no	Feature code	Min-Max	Means, ± Standard division
1	AGE	29.000000-77.000000	54.542088, ± 9.049736
2	SEX	0.000000-1.000000	0.676768, ± 0.468500
3	CPT	1.000000-4.000000	3.158249, ± 0.964859
4	RBP	94.000000-200.000000	131.693603, ± 17.762806
5	SCH	126.000000-564.000000	247.350168, ± 51.997583
6	FBS	0.000000-1.000000	0.144781, ± 0.352474
7	RES	0.000000-2.000000	0.996633, ± 0.994914
8	MHR	71.000000-202.000000	149.599327, ± 22.941562
9	EIA	0.000000-1.000000	0.326599, ± 0.469761
10	OPK	0.000000-6.200000	1.055556, ± 1.166123
11	PES	1.000000-3.000000	1.602694, ± 0.618187
12	VCA	0.000000-3.000000	0.676768, ±0.938965
13	THA	3.000000-7.000000	4.730640, ± 1.938629
14	LB	Heart disease patient=1, Healthy=0	

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104223 |



IV. RESULTS AND DISCUSSION

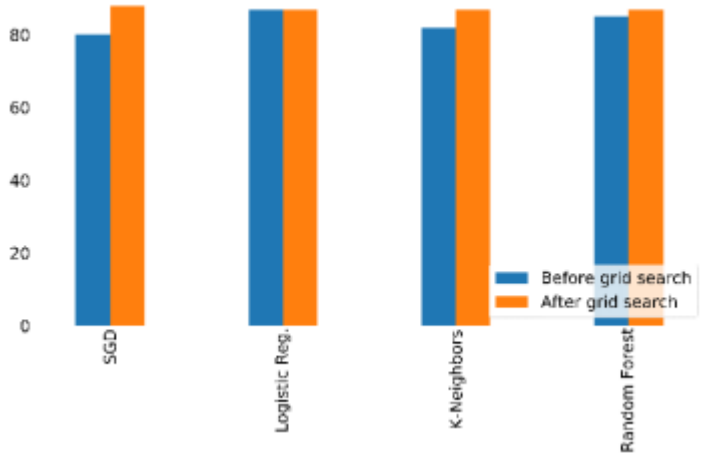
Starting off with the SGD classifier, the prediction was run on the test set which is considered unseen data that the model has never prevailed. The first test was run on the default parameters of the classifier and produced an accuracy of 80%. Then after running a GridsearchCV, the optimized parameters based on cross-validation were found and the accuracy increased to 88%. Figure 10 shows the confusion matrix obtained from this model. The model was built using

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104223 |

the default parameters and the classification occurred based on the unseen test set. The accuracy came out to be 87% and after conducting GridsearchCV the accuracy remained the same since the default parameters came out to be the same as the optimized parameters. Figure shows the confusion matrix of this model. shows how running the GridsearchCV which is based on the cross-validation technique improves the accuracy of every model. This shows the need to fine-tune the parameters of any machine learning algorithm.



To further investigate the models built, a receiver operating characteristic curve (ROC) was plotted as shown in Figure 16 for all of the models involved in this project.

V. CONCLUSION

In conclusion, this paper presented a machine learning ensemble technique that combined multiple machine learning techniques in order to provide a more accurate and robust model for predicting the possibility of having a heart disease. The Ensemble model achieved 90% accuracy, which exceeds the accuracy of each individual classifier. In this study, an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Machine learning classifiers include LR, K-NN, ANN, SVM, NB, and DT are used in the designing of the system. Four standard feature selection algorithms including Relief, MRMR, LASSO, LLBFS, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. LOSO cross-validation method is used in the system for the best hyperparameters selection. This project presented a new approach to heart disease classification, using the Random Forest machine learning algorithm and attributes based on clinical data and patient test results. It reached an overall accuracy of 84.448%. The highest accuracy was reached while using an additional 10 times cross-validation in the process and it outperforms other machine learning techniques using the same database [25]. Using the Random Forests algorithm without the cross validation secured an overall accuracy of 82.895%

REFERENCES

[1] D. Swain et al, "A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning", International Conference on Advanced Computation and Telecommunication (ICACAT), 2018.
[2] T. Kasbe and R. Pippal, "Design of Heart Disease Diagnosis System using Fuzzy Logic", International Conference on Energy, Communication, Data Analytics, and Soft Computing, ICECDS, 2017.
[3] S. Bhatia et al, "SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features", World Congress on Engineering and Computer Science, 2008.
[4] M. Raihan et al, "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", 19th International Conference on Computer and Information Technology, December 18-20, 2016.
[5] C. Malode et al, "Soft set and Fuzzy Rules enabled SVM Approach for Heart Attack Risk Classification among Adolescents", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 |

|| Volume 11, Issue 4, April 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1104223 |

[6] P. Sasidharan et al, "Wearable Cardiorespiratory Monitoring Device for Heart Attack Prediction", International Conference on Communication and Signal Processing, April 4-6, 2019.
[7] B. Bölümü et al, "Heart Attack Risk Detection Using Bagging Classifier", 24th Signal Processing and Communication Application Conference (SIU), 2016.
[8] M. Mustapha et al, "Mobile Heart Rate Monitor for Myocardial Infarction Patients", 6th ICT International Student Project Conference (ICT-ISPC), 2017.
[9] A. Hodge et al, "Wireless Heart Rate Monitoring and Vigilant System", 3rd International Conference for Convergence in Technology (I2CT), 2018.
[10] A. Narin et al, "Effect of Linear and Non-Linear Measurements of Heart Rate Variability in Prediction of PAF Attack", 25th Signal Processing and Communications Applications Conference (SIU), 2017.
[11] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, 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. 304–310, Aug. 1989.
[12] J. H. Gennari, P. Langley, and D. Fisher, ''Models of incremental concept formation,'' Artif. Intell., vol. 40, nos. 1–3, pp. 11–61, Sep. 1989.
[13] Y. Li, T. Li, and H. Liu, ''Recent advances in feature selection and its applications,'' Knowl. Inf. Syst., vol. 53, no. 3, pp. 551–577, Dec. 2017.
[14] J. Li and H. Liu, ''Challenges of feature selection for big data analytics,'' IEEE Intell. Syst., vol. 32, no. 2, pp. 9–15, Mar. 2017.
[15] 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. 3941–3954, Nov. 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

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