A COMPARATIVE STUDY IN PREDICTING THE CARDIOVASCULAR DISEASE USING MACHINE LEARNING ALGORITHMS THROUGH SMOTE ANALYSIS TECHNIQUE

Dr Raghavendran C R,
Assistant Professor, Department of EEE,
Easwari Engineering College, Ramapuram, Chennai – 89.
raghavendran.cr@eec.srmrmp.edu.in
Dr E Kaliappan,
Professor & Head, Department of EEE,
Easwari Engineering College, Ramapuram, Chennai – 89.
hod.eee@eec.srmrmp.edu.in
Ms Harshini Manoharan,
Research Scholar, Department of Computer Science,
SRM Institute of Science and Technology, Ramapuram, Chennai – 89.
harshimanohar@gmail.com
Mr G Vignesh,
Assistant Professor, Department of EEE,
Easwari Engineering College, Ramapuram, Chennai – 89.
vignesh.g@eec.srmrmp.edu.in

ABSTRACT

In human life, healthcare plays a vital role and draws much of its attention in today's world with the advancement of technology and their role in aiding the healthcare sector. Irrespective of the race, gender and, demographic location humans are affected with innumerable number of diseases and still the causes of many remains unknown. It is high time for the medical practitioners in correctly diagnosing the diseases and providing them with proper medical assistance in this challenging environment. With the advent of machine learning techniques, medical practitioners are greatly benefitted through the assistance provided by the algorithm in accurately detecting the disease along with timely diagnosis. This work aims at predicting the cardiovascular diseases which are a group of diseases that affects heart and blood vessels by choosing the best model capable of accurate classification. The earlier methods of estimating the uncertainty levels of cardiovascular diseases helped in taking decisions to reduce the risk in high-risk patients. The prediction model is projected with mixtures of various options and a number of other classification techniques by comparing the accuracies of different algorithms. Our goal is to enhance the performance of the model by removing unnecessary and

insignificant attributes from the dataset and only collecting those that are most informative and useful for the classification task.

Keywords – Cardiovascular disease, K-Nearest Neighbor (KNN), Random Forest, Classification and Regression Tree (CART), Logistic Regression, Synthetic Minority Oversampling Techniques (SMOTE).

DOI Number: 10.48047/nq.2022.20.19.NQ99210 NeuroQuantology2022;20(19): 2497-2505

1. INTRODUCTION

Today's lifestyle has made humans to stick towards fast foods and unhealthy eating habits leading to various unknown diseases and complications. The challenging part of healthcare is that, still the root cause and treatment of many disease remains to be unknown. Globally, cardiovascular disease has been the leading cause of death with an estimate of 17.9 million people in 2019 [1,3]. Majority of the cardiovascular disease can be kept from its progression by inculcating physical activity, healthy eating habits, avoidance of alcohols and tobacco. The effects of these risk factors cause an increase in the blood pressure, blood glucose level, blood lipids and obesity. Hence it is a very critical task to treat cardiovascular beneficiaries as it requires timely medication and attention. Early detection of the disease helps the medical practitioners in provide timely diagnosis and recovery.

Heart disease has become a prevalent and deadly condition over the years due to fat suppression [2]. The overpressure in the human body causes this disease to develop. In the past, doctors might have misdiagnosed a patient's condition, but today's machine learning techniques are very effective at making predictions. Using a number of characteristics from the dataset, we can forecast cardiac illness [4,6]. The generated findings are contrasted with those of other models that have been used in the same domain, and it is discovered that they are better. Random Forest and Decision Tree are utilized to find patterns in the data of heart disease patients collected from the UCI laboratory [7,9].

Clinical data analysis faces a significant problem when predicting cardiovascular disease. With the use of machine learning (ML), it has been demonstrated that it is possible to make predictions and judgments from the vast amount of data generated by the healthcare sector. Additionally, we have observed the employment of ML approaches in recent advancements across several IoT domains (IoT). Only a few researches have looked into using ML to predict cardiac disease. By identifying the key features of the disease using machine learning technique we propose an unique approach to improve the precision of cardiovascular disease using SMOTE.

Machine learning algorithms have shown its importance greatly in the health care sector by providing early detection of the disease [10]. Through proper training of the model, desired and accurate output is obtained. Increasing the accuracy rate of the model decreases the error rate of the model. Dataset required in training the model is taken from Kaggle and the traditional machine learning algorithms like KNN, Random Forest, CART and Logistic Regression are applied to find out the best model [5]. Through this paper, we can access the uncertainty levels of the disease based on the attributes of the dataset.

2. PROPOSED SYSTEM

Several works have been proposed in predicting the cardiovascular disease through significant features using various machine learning algorithms [8]. This works scales the performance of the traditional machine learning algorithm and finds out the best one for prediction. The proposed methodology begins with the data preprocessing method by eliminating the incorrect values and tuning out the missing values through mean and median values. After obtained the processed data, the model is trained using the training dataset and tested for its ability in accurate prediction using the testing dataset. Finally a comparative analysis with the application of SMOTE and without the application of SMOTE technique is performed in determining the algorithm which outperforms the other algorithm.

2.1 Data Pre-Processing

The dataset required in carrying out the research was taken from Kaggle.com comprising of 1025 records with 13 attributes – age, sex, blood pressure, type of chest pain, blood glucose level, exercise induced angina, maximum heart rate, results of electrocardiography, cholesterol level, thal, slope of peak exercise ST segment, old peak and the number of blood vessels. The collected dataset is subjected for preprocessing that includes clearing out the missing values and ignore the incorrect ones. The preprocessed dataset is subjected to Synthetic Minority Oversampling Technique (SMOTE) analysis in reducing the biasness of the dataset. The model is trained with and without the application of SMOTE testing for its accuracy levels.

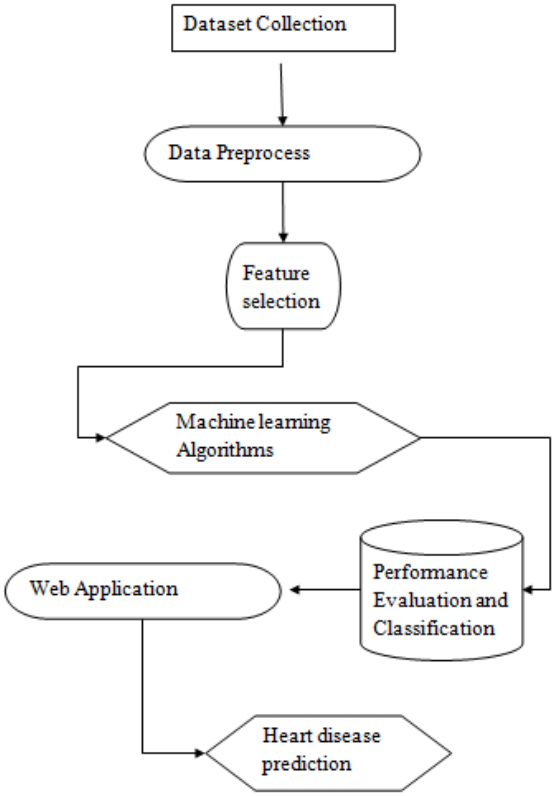

Fig 1. Architecture diagram of the proposed methodology

2.2 IMPLEMENTATION OF THE ALGORITHM

2.2.1 K-Nearest Neighbor (KNN)

K-Nearest Neighbor, a distance based classification algorithm is intended to classify the data points based on their proximity neighbors classes solving both classification and regression problems. The model is evaluated based on the following three aspects,

1. Ease to interpret output
2. Calculation time
3. Predictive Power

Table 1. Comparison of algorithm

Criteria	LR	CART	RF	KNN
Ease to interpret output	2	3	1	3
Calculation time	3	2	1	3
Predictive power	2	2	3	2

From table 1, it is clearly seen that the KNN algorithm fairs across all parameters of considerations. The results of the KNN algorithm are depicted in Fig 2. For the training dataset the value of K=1 since the distance between the closest point to the data point is one and the same yielding accurate results. This value of K, over fits the boundaries causing the error rate to decrease and reach a minimal value. After hitting the minimal value the value starts increasing. Hence to obtain the accurate K value, the training and the test dataset are segregated from the initial dataset. The error curve is plotted as shown in fig 3, this K value can be used for making the predictions.

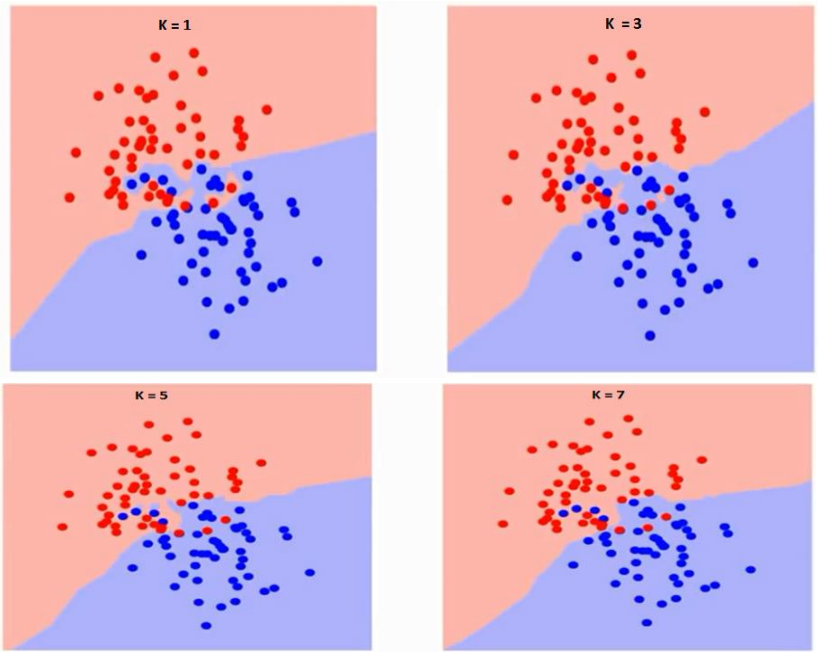

Fig 2. Results of KNN Algorithm

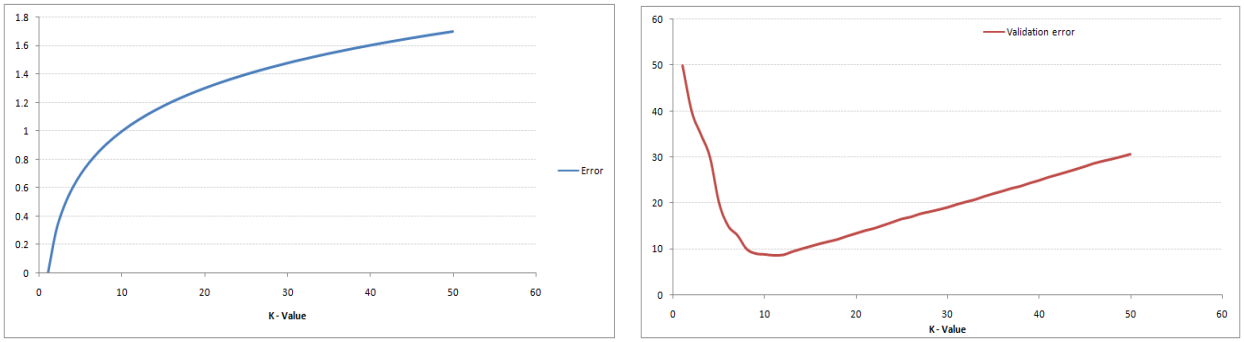

Fig 3. Error curve of KNN Algorithm

2.2.2 Random Forest

Random forest, an ensemble based technique comprises of several individual decision trees capable of precisely classifying the data points into its respective classes proving it to be a powerful technique in prediction. There are three key hyper parameters for random forest algorithms that must be set prior to training. Node size, tree count, and sampled feature count are a few of them. Uncorrelated models have the ability to generate ensemble forecasts that are more precise than any single prediction. Since the overall variance and prediction error are reduced by averaging uncorrelated trees, the classifier won't overfit the model.

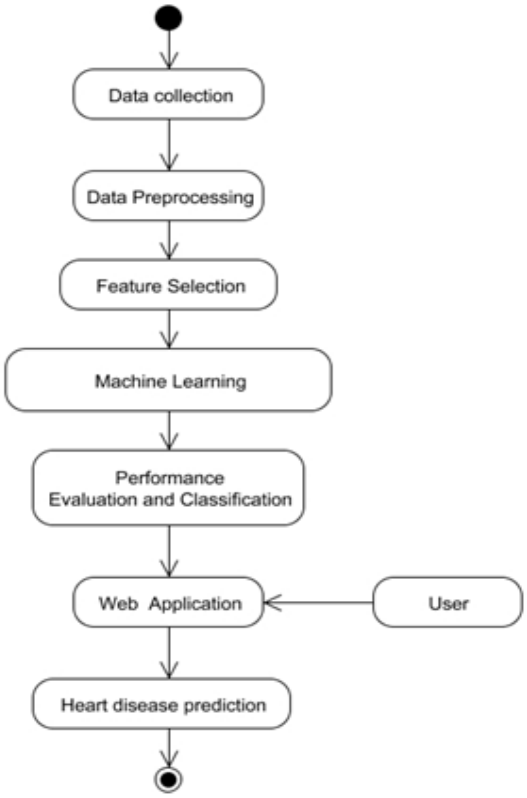

Fig 4. Data flow diagram of the proposed methodology

2.2.3 CART

When the dataset needs to be divided into classes that are a part of the response variable, classification trees are typically used. The dataset is divided using a classification tree to support the homogeneity of knowledge. Any classification or regression tree analysis has as its goal

creating a set of if-then scenarios that leave the precise prediction or categorization of a case. Regression and classification trees strive to provide precise forecasts or anticipated classifications, supported by an array of if-else scenarios. Compared to traditional decision trees, they often have a number of advantages. Many times, if-then statements are used to succinctly summaries the findings from classification and regression trees. This disproves the validity of the ensuing implicit presumptions.

2.2.4 Logistic Regression

Logistic Regression, a probability based algorithm is highly capable of predicting the relationship between the dependent and the independent variable bounded within the range of 0 to 1. Logit transformation is performed on the dataset where the probability of accurate prediction is divided by the probability of inaccurate predictions. The outcome of the LR is a binary value composed of 0 or 1. The logistic function is represented as follows,

$$Logit\ (pi) = \frac{1}{1+\exp(-pi)} \quad (1)$$

Within a logistic regression data analysis, the major drawback is the computation of log odds value. Exponentiating the beta estimates to create an odds ratio (OR) makes it easier to evaluate the data as a consequence.

3. RESULTS OF CLASSIFICATION ANALYSIS

The performance of the model was evaluated based on the computational metrics such as accuracy; precision and error. Model is trained using the pre-processed data given as the input that comprises of 70% of the data and the reaming 30% of the data goes out for testing the model for its performance.

Table 2. Performance of the classifier without SMOTE

ALGORITHMS	ACCURACY	SENSITIVITY	SPECIFICITY
KNN	89.2	78.3	79.3
RF	72.4	69.7	77.9
CART	74.5	65.4	78.4
LR	64.9	60.9	72.3

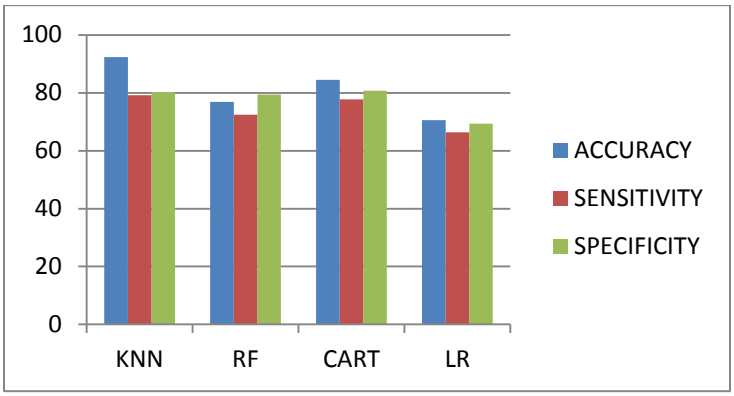

Fig 5. Graphical representation of classifiers performance without SMOTE

Table 3. Performance of the classifier with SMOTE

ALGORITHMS	ACCURACY	SENSITIVITY	SPECIFICITY
KNN	92.3	79.2	80.3
RF	76.9	72.4	79.4
CART	84.5	77.7	80.7
LR	70.6	66.4	69.3

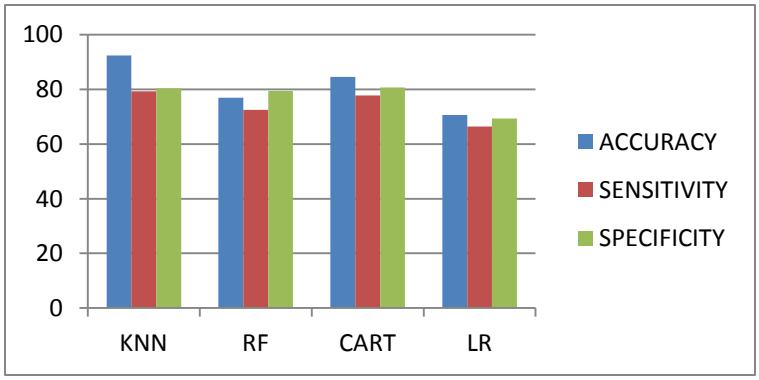

Fig 6. Graphical representation of classifiers performance with SMOTE

4. CONCLUSION

Based on the results obtained from the comparative study as shown in Fig 4,5 and table 2,3, it is evident that the KNN outperforms the other algorithms yielding an accuracy rate of 89.2 % in classifying the disease. The accurate of the model is further enhanced by applying SMOTE technique yielding the accuracy rate of 92.3 %. From this it is evident that the accuracy

of the model without SMOTE greatly differs from the model training with SMOTE technique, since SMOTE plays a vital role in reducing the biasness of the dataset.

References

[1] Kaan Uyar and Ahmet İlhan, "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks" in B.V ICTASC, Elsevier, 2017.

[2] Ashish Chhabbi, Lakhan Ahuja, Sahil Ahir and Y. K. Sharma, "Heart Disease Prediction Using Data Mining Techniques", *© IJRAT Special Issue National Conference "NCPC-2016"*, pp. 104-106, 19 March 2016.

[3] Berry JD, Lloyd-Jones DM, Garside DB, et al. Framingham risk score and prediction of coronary heart disease death in young men. Am Heart J. 2007;154(1):80–6.

[4] Theresa Princy and R, J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques", *© IEEE ICCPCT*, 2016.

[5] Kaur h Beant and Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", *©IJRITCC*, vol. 2, no. 10, pp. 3003-08, 2014.

[6]Kirmani, M.M., Ansarullah, S.I.: Prediction of heart disease using decision tree a data mining technique. IJCSN Int. J. Comput. Sci. Netw. **5**(6), 885–892 (2016)

[7] Salam Ismaeel, Ali Miri et al., "Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis", *IEEE Canada International Humanitarian Technology*

[8] Tahira Mahboob, Rida Irfan and Bazelah Ghaffar."Evaluating ensemble prediction of coronary heart disease using receiver operating characteristics" ©2017 IEEE

[9] Ammar Asjad Raja, Irfan-ul-Haq , Madiha Guftar Tamim Ahmed Khan "Intelligence syncope Disease Prediction Framework using DM-techniques" FTC 2016 –Future Technologies Conference 2016

[10] M.A. Jabbar, B.L.Deekshatulu, and Priti Chandra, "Intelligent heart disease prediction system using random forest and evolutionary approach", Journal of Network and Innovative Computing, Vol. 4, pp.174-184, 2016.