Innovative methods to apply on medical data for predicting Heart disease using Machine Learning approach

Dr Chandramouli VSA[1*], Dr P Devabalan[2], A D Devi[3], D Manoj[4],G R Kumar[5], B Sandeep[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering
[1,2,3,4,5,6]Bonam Venkata Chalamayya Engineering College (Autonomous), Odalarevu, A.P, India
[1*]chandramouli.ac@yahoo.com,
[2]devabalanme@gmail.com

Abstract: Accurately predicting the onset of cardiovascular disease is one of the most pressing problems in modern medicine. Medical experts devote a great deal of time trying to pin down what's causing this. The goal of using several algorithms, including LR, KNN, SVM, GBC, and the GridSearchCVs, is to forecast cardiac illness. The optimal strategy for hyperparameter testing is to use GridSearchCV in conjunction with the Extreme Gradient Boosting Classifier. We compare these results to those of previous studies that focused on cardiac prediction. For the purposes of feature selection and dimensionality reduction, principal component analysis (PCA) was employed. In order to obtain early prediction of heart disorders using data mining approaches, a number of machine learning classifiers, iterative dichotomization, decision tables, and classification and regression trees were utilised. Combination was performed on the log datasets originating from Long Beach, Stat, Switzerland, Virginia, and Cleveland. When the Relief and LASSO methods are applied, it becomes possible to choose features appropriately. The decision tree bagging technique (DTBM), the RFBM, the KNNBM, the ABCM, and the GBBM are all new hybrid classifiers that borrow training from traditional classifiers. With the help of several machine learning techniques, along with our model's accuracy, sensitivity, error rate, precision, and F1 score, we were able to determine the negative predictive value, false positive rate, and false negative rate. In order to facilitate comparisons, the findings are presented independently.

Keywords: Least Absolute Shrinkage and Selection Operator (LASSO) techniques. New hybrid classifiers likeDecision Tree Bagging Method (DTBM)

1. Introduction

According to the World Health Organisation (WHO), 31 percent of all deaths in 2016 were caused by cardiovascular disease (CVD), which impacts around 1.79 crore individuals. Of these fatalities, 85% are attributable to heart problems. Many lives could be saved if patients could get high-quality medical care and diagnoses quickly. Blood pressure is one of the most important factors [3].
There is no way for doctors to know when a patient may develop heart disease [4]. According to projections, 23.3 million deaths worldwide will be attributed to cardiovascular disease by the year 2030 [5]. The heart's veins transport oxygen-rich blood, and their constriction or blockage can cause heart disease and stroke [6]. Anxiety, stress, alcohol consumption, sedentary lifestyle, obesity, diabetes, high blood pressure, and high cholesterol can all harm the heart. The diagnosis of cardiovascular disease is influenced by these factors. High blood pressure causes the walls of the arteries to thicken, which can lead to blockage and, in extreme cases, an increase in mortality [7]. Using pattern extraction to get predictions from data is the primary objective of the algorithms proposed by the different academics. There is a correlation between the early detection and treatment of cardiac

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_135

sickness and improved survival or lower mortality rates for persons with the condition. When looking for abnormal narrowing of a heart artery, angiofigurey is a common tool to use. An evaluation of the symptoms, examination, and ECG characteristics was conducted using SMO, Naive Bayes, and Ensemble algorithms. The results showed an accuracy of 88.5% in predicting the existence of CAD [11].

2. Proposed System

We are utilising the Cleveland dataset from Hungary since it shares characteristics with the UCI dataset and the author of the proposed study used both datasets. Computer programme While training with all parameters will allow Grid Search to choose algorithm instances with greatest accuracy, tuning means continuously training the algorithm with all parameters (e.g., MAX ITERATIONS, distance functions, etc.).As an extension, we are utilising the Random Forest approach, which gives the same accuracy with less computation time, instead of XGBOOST, which gives 100% accuracy but has greater computing time in the proposed work.

3. Methodology

Gradient Boost Search Method:

Machine learning techniques like gradient boosting find use in many different areas, including classification and regression. It provides a model for making predictions by combining many weak prediction models, most often decision trees. The Extreme Gradient Boosting Classifier achieves the highest testing accuracy of 99.03% with GridSearchCV, outperforming the other three approaches. The Extreme Gradient Boosting Classifier achieved a 98.05% testing accuracy without GridSearchCV.

Grid search CV: Finding the best possible values for a collection of parameters in a grid is the job of Grid Search CV. It's just a method for cross-validation. It is necessary to input both the model and the parameters. Predictions are made after the optimal parameter values have been extracted.

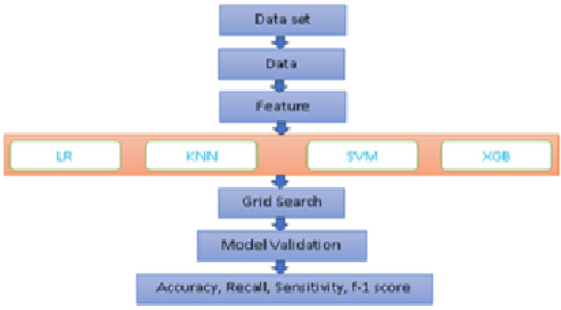

Fig.1. System architecture

Proposed methodology was implemented on publicly available datasets using python libraries. Proper preprocessing was done and results are narrated

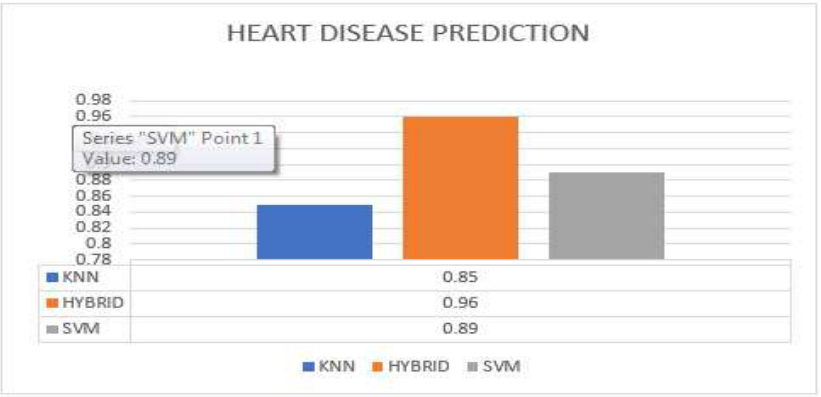

Fig.2 logistic regression training

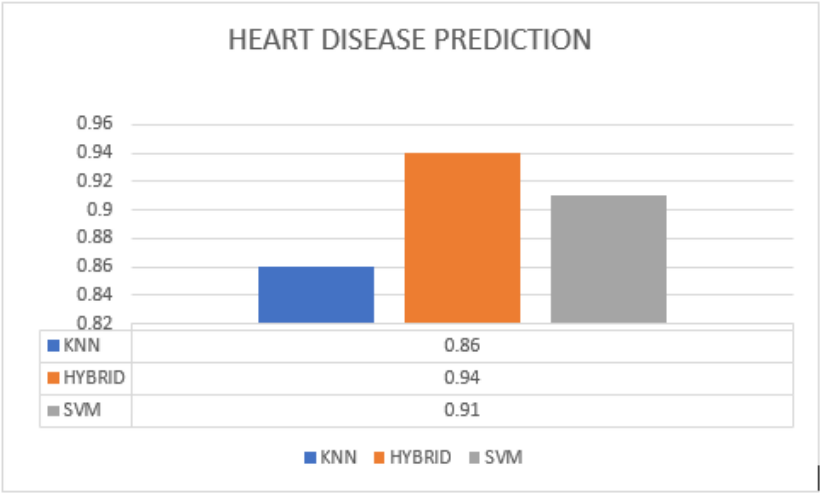

Fig.3.Training logistic regression with tweaking parameters

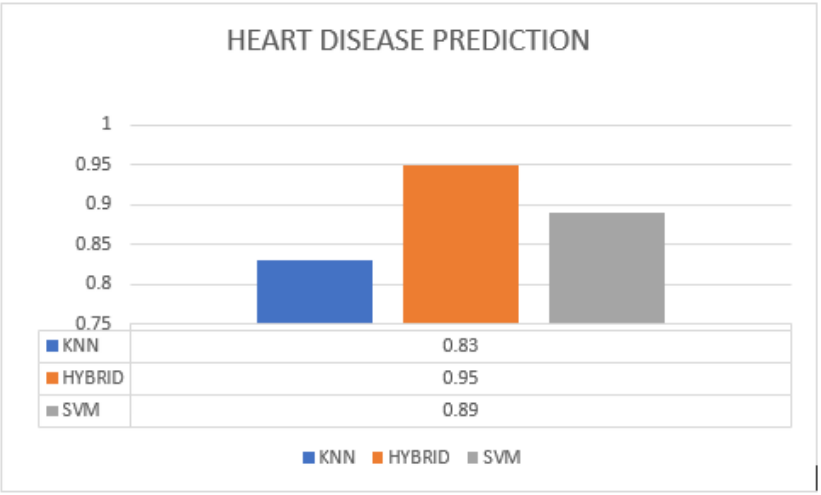

Fig.4. Training SVM with tuning

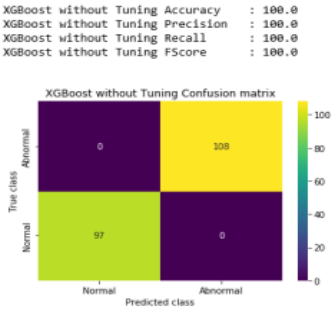

Fig.6. XGBOOST without tuning

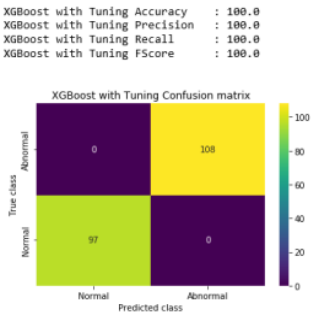

Fig. 7. XGBOOST With tuning

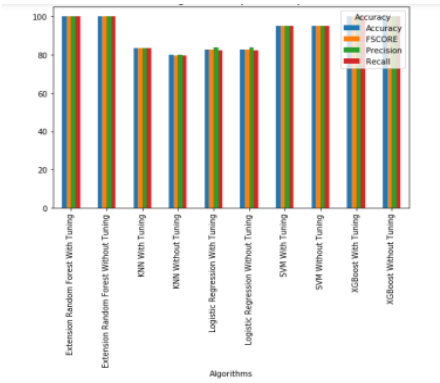

Fig. 8. Comparison of Performance Metrics

	Algorithm Name	Precison	Recall	FScore	Accuracy
0	Logistic Regression Wihtout Tuning	83.739837	82.483772	82.647706	82.926829
1	Logistic Regression With Tuning	83.739837	82.483772	82.647706	82.926829
2	KNN Without Tuning	80.120773	79.758496	79.844608	80.000000
3	KNN With Tuning	83.361921	83.419244	83.382605	83.414634
4	SVM Without Tuning	95.092346	95.212868	95.116257	95.121951
5	SVM with Tuning	95.092346	95.212868	95.116257	95.121951
6	XGBoost Without Tuning	100.000000	100.000000	100.000000	100.000000
7	XGBoost With Tuning	100.000000	100.000000	100.000000	100.000000
8	Extension Random Forest Without Tuning	100.000000	100.000000	100.000000	100.000000
9	Extension Random Forest With Tuning	100.000000	100.000000	100.000000	100.000000

Fig. 9. Algorithms performance in tabular format

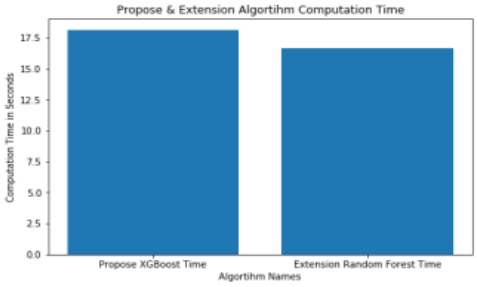

Fig. 10. Algorithms Computation Time

4. Conclusion

Accurately predicting the onset of cardiovascular disease is one of the most pressing problems in modern medicine. Medical experts devote a great deal of time trying to pin down what's causing this. In order to predict cardiac illnesses, this study used a wide variety of methods, including LR, KNN, SVM, GBC, and GridSearchCV. Our method of choice is a five-fold cross-validation strategy for validation. The comparative investigation makes use of these four methods. The models are evaluated using the following datasets: the Long

Beach V dataset, the UCI Heart Disease Dataset Kaggle, and datasets from Cleveland, Hungary, and Switzerland.

References

[1] P. Drotár and Z. Smékal, ‘‘Comparative study of machine learning techniques for supervised classification of biomedical data,’’ ActaElectrotechnica Inf., vol. 14, no. 3, pp. 5–10, Sep. 2014, doi: 10.15546/aeei2014-0021.

[2] A. Levin, ‘‘The clinical epidemiology of cardiovascular diseases in chronic kidney disease: Clinical epidemiology of cardiovascular disease in chronic kidney disease prior to dialysis,’’ in Seminars in Dialysis, vol. 16, no. 2. Oxford, U.K.: Blackwell Science, Mar. 2003, pp. 101–105.

[3] K. S. Reddy, ‘‘cardiovascular diseases in the developing countries: Dimensions, determinants, dynamics and directions for public health action,’’ Public Health Nutrition, vol. 5, no. 1, pp. 231–237, Feb. 2002.

[4] A. Kishore, A. Kumar, K. Singh, M. Punia, and Y. Hambir, ‘‘Heart attack prediction using deep learning,’’ Int. Res. J. Eng. Technol., vol. 5, no. 4, p. 2395, 2018.

[5] C. D. Mathers and D. Loncar, ‘‘Projections of global mortality and burden of disease from 2002 to 2030,’’ PLoS Med., vol. 3, no. 11, p. e442, Nov. 2006.

[6] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, ‘‘heart disease prediction system using associative classification and genetic algorithm,’’ in Proc. Int. Conf. Emerg. Trends Elect., Electron. Commun. Technol. (ICECIT), 2012, pp. 40–46.

[7] T. N. Sugathan, C. R. Soman, and K. Sankaranarayanan, ‘‘Behavioural risk factors for non-communicable diseases among adults in Kerala, India,’’ Indian J. Med. Res., vol. 127, no. 6, pp. 1–9, 2008.

[8] A. Ahmed and S. A. Hannan, ‘‘Data mining techniques to find out heart diseases: An overview,’’ Int. J. Innov. Technol. Exploring Eng., vol. 1, no. 4, pp. 18–23, 2012.

[9] M. Ribeiro, K. Grolinger, and M. A. M. Capretz, ‘‘MLaaS: Machine learning as a service,’’ in Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2015, pp. 896–902.

[10] I. Castelli and E. Trentin, ‘‘Combination of supervised and unsupervised learning for training the activation functions of neural networks,’’ Pattern Recognit. Lett., vol. 37, pp. 178–191, Feb. 2014.

[11] Z. Sani, R. Alizadehsani, J. Habibi, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, F. Khozeimeh, and F. Alizadeh-Sani, ‘‘Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiofigurey features,’’ Res. Cardiovascular Med., vol. 2, no. 3, p. 133, 2013.

[12] D. Tomar and S. Agarwal, ‘‘A survey on data mining approaches for healthcare,’’ Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241–266, 2013.

[13] Y. Er, ‘‘The classification of white wine and red wine according to their physicochemical qualities,’’ Int. J. Intell. Syst. Appl. Eng., vol. 4, no. 1, pp. 23–26, Dec. 2016.

[14] S. J. Pasha and E. S. Mohamed, ‘‘Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction,’’ IEEE Access, vol. 8, pp. 184087–184108, 2020.