ISSN No. 0976-5697

VOLUME 17 SPECIAL ISSUE 1, JULY 2026
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

Support Vector Machines and Ensemble Methods for Improved Heart Disease Classification

[1]Dr. Mukesh Choudhary
Professor,
Geetanjali Institute of Technical Studies,
Udaipur, Rajasthan, India.
Mukeshmewara84@gmail.com

[2]Dr. Sonam Mittal
Associate Professor,
Geetanjali Institute of Technical Studies,
Udaipur, Rajasthan, India.
sonammittalkota@gmail.com

[3]Dr. Vijendra Kumar Maurya
Associate Professor,
Geetanjali Institute of Technical Studies,
Udaipur, Rajasthan, India.
sonammittalkota@gmail.com

Abstract: Computing power and methodological advancements have brought to a significant deal of variety in the medical sciences, particularly in the area of human cardiac disease diagnosis. It has devastating consequences on human life and is now one of the most severe cardiac disorders in humans. Accurate and timely diagnosis of human cardiac sickness may significantly improve the patient's survival rate and prevent heart failure in its early stages. Manual techniques of diagnosing heart disease are prone to bias and examiner variability. When it comes to identifying and classifying individuals with heart disease from those without, machine learning algorithms are dependable and efficient resources. As per the suggested study, we used the heart disease dataset to analyze the performance of the machine learning algorithms with the help of different metrics, including sensitivity, specificity, F-measure, and classification accuracy. The algorithms were used to determine and forecast human cardiac disease. To do it, we first applied nine different machine learning classifiers to the final dataset (AB, LR, ET, MNB, CART, SVM, LDA, RF, and XGB) and compared their results before and after the change in the hyper parameter. We also do specific pre-processing, dataset standardization, and hyper parameter tweaking to ensure their correctness on the reference dataset for heart disease. We also used the industry-standard ML algorithms are trained and validated using the K-fold cross-validation approach. Finally, the results of the experiment demonstrated that by standardizing data and hyper-tuning the parameters of machine learning classifiers, the prediction classifiers' accuracy increased.

Keywords: Heart Disease Prediction, Machine Learning Classifiers, SVM, Hyper-parameter Tuning, Dataset Standardization, K-Fold Cross Validation, Classification Accuracy

I. INTRODUCTION

Data provided by the World Health Organization (WHO) indicates that in 2019 cardiovascular diseases (CVDs) resulted in 17.9 million deaths (including 32 percent of all deaths in the world) and a mortality rate of over 17.7 million [1] and [2].The Australian Institute of Health and Welfare (AIHW) reports that 42% of all deaths in Australia in 2018 were related to cardiovascular disease (CVD) [3]. Due to issues with accuracy and processing time, current approaches for diagnosing cardiac illnesses are not very good at early detection, thus researchers are trying to come up with a better approach [4]. Heart disease is notoriously difficult to diagnose and manage in areas without access to modern medical equipment and trained professionals [5]. An accurate diagnosis and prompt treatment may save many lives [6]. To diagnose cardiac conditions, a physician will take into account the patient medical history, the results of the physical examination, and any alarming symptoms. On the other hand, this diagnostic approach does not find nearly enough cases of cardiac disease to warrant further investigation. Investigating it is also costly and computationally challenging [7]. In order to allay these worries, we created a non-invasive prediction technique based on machine learning classifiers. Using artificial fuzzy logic and machine learning classifiers, an expert decision system efficiently diagnoses heart disorders. Consequently, there is a decrease in the death ratio [8, 9]. Other research made use of the Cleveland heart disease dataset. For machine learning prediction models to be trained and tested, the right data is required. Training and testing machine learning classifiers on a refined or standardized dataset improves their accuracy. Improved prediction model performance is also possible with the addition of pertinent and associated data elements. In order to ensure that machine learning classifiers are accurate, it is crucial to standardize data and pick features. In the literature, many researchers have employed various methods of prediction; nevertheless, these approaches fail to accurately anticipate the occurrence of cardiac illnesses. Improving the accuracy of machine learning classifiers requires standardizing the data. A set of standardized tests is required in order to predict when one will develop

heart disease. Finding the problem quickly is challenging. Other research made use of the Cleveland heart disease dataset. For machine learning prediction models to be trained and tested, the right data is required[10]. Computational classifiers, when trained with the right data, can be helpful in making accurate diagnoses [11]. When it comes to training and evaluating models, the majority of these approaches take use of publicly accessible datasets. The availability of these datasets has allowed researchers to explore new pathways for developing cutting-edge algorithms for predicting cardiovascular disease risk, and it has also improved the effectiveness of prediction models based on machine learning. These databases include details on potential dangers as well as the patient's health condition. Inconsistent and duplicate clinical datasets necessitate pre-processing when developing CVD predictive models [12]. Information about different risk factors (features) is provided when pre-processing is finished; features are selected based on parameters including high prevalence in the majority of populations, strong independent influence on heart disease, and controllability or treatment-ability to lower risks [13]. When trying to model CVD predictors, many research has used different risk variables or characteristics. The effectiveness of machine learning algorithms is maximized when they are trained on relevant datasets [12, 14]. The possible obstacles to precise prediction of cardiac diseases are the lack of medical data, predictive features, and implementations of the ML algorithm, as well as the in-depth analysis of these data. Our study tries to seal some of these knowledge gaps to enhance the model in predicting CVD. Also, the datasets applied in the past studies possess some limitations which should be overcome. Indicatively, techniques such as min-max scalar and standard scalar (SS) are applied in eliminating instances of missing values of features.

Our recommended machine learning heart disease prediction classifier makes use of the XG Boost (XGB), choice trees (CART), support vector machine (SVM), logistic regression (LR), linear discriminant analysis (LDA), Ada Boost (AB), and extra trees (ET) techniques. They include logistic regression (LR) and multinomial Naive Bayes (MNB) and random forest (RF). The reason behind the utilization of the Grid Search CV approach on standardization and hyperparameters is to find the best machine learning classifier value. Other than that, there are metrics like sensitivity, recall, accuracy, precision, and F-measures that are applied in order to measure the performance of the machine learning classifier. The Cleveland HD data was used to evaluate the proposed method. Moreover, the accuracy of the suggested machine learning classes has been equated with the state-of-the-art techniques that exist in the literature, which includes SVM, LR [19], and RF [20]. The key benefits of the proposed work are listed below:

(1) Prior to improving and standardizing the datasets, the authors attempt to address the dataset issue. +1 The datasets serve to train and test classifiers with a view of determining which classifiers give the best accuracy results.

(2) Secondly, the writers used the Grid Search CV technique to determine the optimal values for the hyperparameter.

(3) The third step in hyper parameter tuning for maximum accuracy is to use machine learning classifiers with optimal values.

(4) When it comes to accuracy, the suggested classifier (SVM) is top-notch.

II. Research Goals and Objectives

Creating a more accurate model for predicting cardiac problems is the primary motivation for this study. Timely patient identification, faster diagnosis, fewer heart attacks, and more lives saved are the particular goals.

III. Methodology

Here we lay out the phases of the proposed technique and explain how they define it, as seen in Figure 1.

(1) Picking a dataset from a repository of machine learning datasets is the first stage. There are many databases on the internet, and these are the Cleveland heart disease data, the Z-Alizadeh Sani data, Stat Log Heart, the Hungarian, the Long Beach VA data, and Kaggle Framingham data.

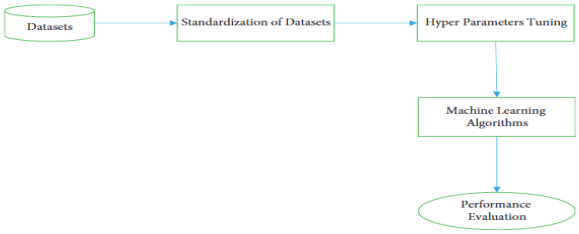
(2) The data sets acquired in the second stage were improved and standardized. The datasets in question had inaccurate values as they were not collected in a controlled setting. Data pre-processing is therefore a must-do for every data science or machine learning project. When a dataset's risk factors have different values, it's called data normalization. The temperature measurements in Celsius and Fahrenheit, for instance, are not interchangeable. Risk

variables are scaled and given values that indicate the variation between standard deviations from the mean when data is standardized. With a mean (μ) of 0 and a standard deviation (σ) of 1, it modifies the risk factor value to improve machine learning classifier performance. In mathematics, the standardization formula is (1).

$$Standardization\ of\ X = \frac{X - Mean\ of\ X}{Standard\ Deviation\ of\ X} \quad (1)$$

(3) The third step is hyperparameter tuning, which entails selecting the ideal hyperparameter value to get high accuracy. We used the Grid Search CV technique for this objective. If we want our machine learning classifiers to perform better, we tweak their hyperparameter values before we use them. It provides a unified setting for training and tuning hyperparameters of all machine learning algorithms. In order to get an accurate model, the full training dataset is used once the appropriate values for the hyper parameters have been determined. Training dataset is then used to find the most suitable values of the hyperparameters which can be optimized using a 10-fold CV. The CV process consists of setting hyperparameter values to find the highest classification accuracy. After that, machine algorithms Ada Boost, XG Boost, support vector machines, logistic regression, additional trees, multinomial Naive Bayes, classification and regression trees, random forests, and linear discriminant analysis are implemented on the data set in step 2. Lastly, a number of metrics, such as accuracy, precision, recall, and F-measure, are used to assess the prediction model's performance. The model that produces the best results is selected for each set of metrics, including prediction accuracy, precision, recall, and F-measures. The accuracy metric measures how well a classifier or machine learning model predicts the future. It may be expressed mathematically as (2)

$$Accuracy = \frac{true\ positive\ (TP) + true\ negative(TN)}{TP + TN + false\ negative\ (FN) + false\ positive(FP)} \quad (2)$$



The suggested method's block design is shown in Figure 1. Mobile Information Systems 5.

The fraction of true positives among all expected positives is known as accuracy. It may be expressed mathematically as (3).

$$precision = \frac{TP\ (true\ positives)}{TP\ (true\ positives) + FP(false\ positives)} \quad (3)$$

Recall measures how the overall amount of false-negative instances impacts the study of the total number of true-positive or real-positive examples. It may be expressed mathematically as (4).

$$Recall = \frac{TP(true\ positives)}{TP(true\ positives) + FN(false\ negatives)} \quad (4)$$

Preciseness and memory are harmonically averaged out in an F-Measure. It finds the sweet spot between recall and accuracy, which may be expressed mathematically as (5).

$$F\ -\ measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

IV. Data Collection

The most popular dataset in the study is the Cleveland heart disease dataset, which can be found on the UCI machine learning repository. Six of the 303 records in total include blanks. Only 13 of the 76 features in the original data set will be utilized in any published study; the remaining features indicate the condition's impact. The Z-Alizadeh Sani dataset is another often used dataset for forecasting by scholars. With 55 input parameters and a class label variable for every patient, it has data from 303 patients. Some of the datasets that the researchers

apply in their prediction method include StatLog Heart, Hungarian, Long Beach VA and Kaggle Framingham among others. The stat log entries (n=270) all have thirteen commonalities with Cleveland. Just as in the case of the Cleveland dataset, the Hungarian and Long Beach VA datasets comprising 274 entries with 14 characteristics will be provided by the UCI repository. The researchers employed a range of publicly available datasets for this investigation. Cleveland, Hungary, Switzerland, and other countries' databases were among them. The main source of data is the Cleveland Clinic Foundation [40], with additional data coming from +e StatLog datasets that may be viewed at [41]. The final data source is the Z-Alizadeh Sani dataset, which is available at [42]. It includes 303 data samples and 55 attributes. We used the Z-Alizadeh Sani dataset, a publically accessible resource, to acknowledge the assessment with literature. Table 1 lists the datasets and provides information on where they came from.

V. Experimental Results and Discussion

Here, we go over the outcomes of our experiments. After re-and standardizing the dataset, we pulled it from an open machine learning source. We used hyperparameter tweaking and machine learning classifiers after standardization. All of our classifiers are trained and tested using 10-fold cross-validation. Plus, uniform datasets allow for the analysis of classifier accuracy both before and after the fact. We depict the accuracy of the chosen classifiers for assessment purposes. Comparison of classifier accuracy using pre- and post-standardization data is shown in Figure 2. Figure 2 shows that MNB and SVM classifier accuracy dropped, although the accuracy of most machine learning algorithms (RF, CART, LDA, AB, LR, ET, and XGB) increased in the standardized data. In the standardized dataset, most of the classifiers including CART, ET and AB increased their accuracy significantly. Figure 1 shows that the ET and AB classifiers can best predict (90.16 percent).With an accuracy of just 59.01%, MNB demonstrates the worst overall performance. In addition, we evaluate the accuracy both before and after the dataset is standardised. Because of the dataset's normalization, ET and AB classifiers are able to get an accuracy of 90.16 percent. The experimental findings show that hyper parameter adjustment improved the classifier's accuracy. To get the most out of the chosen classifiers, we tweak their hyperparameter settings. Using 10-fold cross-validation, a set of accuracy is attained with different combinations of hyperparameters. Test split isn't a good choice since there aren't enough instances to train the model with our current set of training data. Consequently, 10-fold cross-validation is employed.

Table 1: Factors Associated with Heart Disease in Publicly Available Datasets

S. No.	Date Set	No. of sample	No. of Risk factors
1	Z-Alizadeh Sani	369	25
2	UCI	348	24
3	StatLog	398	40

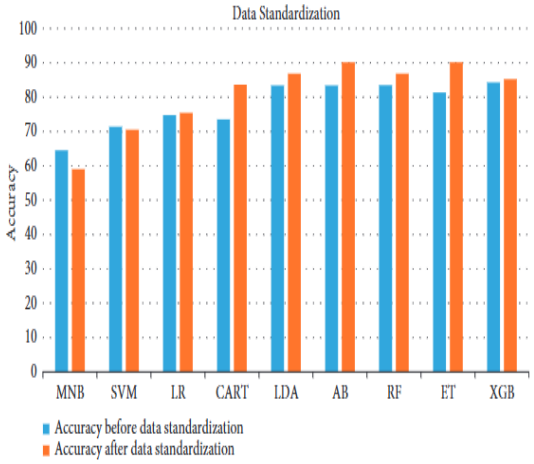


Figure 2: Pre- and post-data-standardization classifier accuracy.

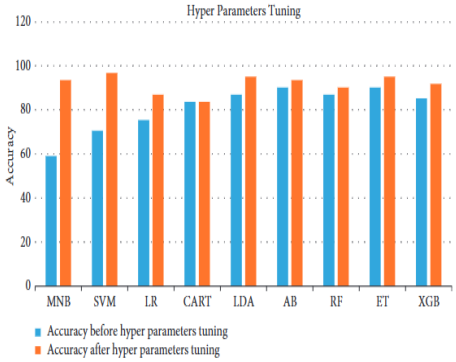


Fig. 3: Pre- and post-hyper parameter-tuning classifier accuracy in Mobile Information Systems

Figure 3 shows the classifiers' accuracy both before and after hyperparameter modification. Although hyper parameter adjustment enhanced the accuracy of most classifiers (MNB, RF, LR, LDA, AB, SVM, ET, and XGB), CART's accuracy remained

unchanged. The optimal hyperparameter combinations for several algorithms are shown in Table 2 for the purpose of improving their accuracy. Table 3 displays the classifiers' accuracy, recall, precision, and F-measure. Although MNB and XGB classifiers have a maximum recall of 100%, SVM has a maximum accuracy of 98% for positive classifications. In contrast, support vector machines (SVMs) often outperform the competition with recall rates of98%, precision rates of98%, F-measure rates of 98%, and accuracy rates of 96.76%. All classifiers have recalls over 90% and precisions above 80%. For the negative class, XGB and MNB reach a maximum accuracy of 100%. Cart has the lowest recall, F-measure, and accuracy of 61, 69 and 83.66 respectively and SVM and LDA has a maximum recall of 94.00. LR has a small precision of 78%. It is clear that the negative class was performing poorly based on its recall of 61% and F-measure of 69. SVM demonstrates relatively high results as far as negative classes are concerned since the recollection, precision, and F-measure are 94 percent and the accuracy is 96.72. The findings indicate that the SVM classifier was the best in terms of accuracy in the process of hyperparameter testing.

The optimal values for hyperparameter adjustment to enhance classifier accuracy are shown in Table 2.

Algorithm	Best Hyperparameter Values
SVM	Kernel: sigmoid, C:0.5
CART	Max features="auto," random state = 123, min samples split = 20, min samples leaf = 11
AB	N estimators = 50, learning rate = 0.05LDA
LDA	Shrinkage = "auto," solver = "lsqr"
GBM	estimators:250
RF	Criterion = "gini," n jobs= −1, min samples leaf = 2, min samples split= 5, n estimators 15, random state 123ET N jobs −1, min samples leaf =1, n estimators= 15, random state = 123,
ET	N jobs = −1, min samples leaf= 1, n estimators = 15, random state= 123, criterion= "gini," Min samples split= 6
XBoost	objective = "reg: linear," colsample bytree = 0.3, learning rate = 0.1, max depth = 15, alpha = 5, n estimators= 123

Table 3: Evaluation of classifiers based on accuracy, recall, F-measure, and precision

Results from pre- and post-standardization dataset comparisons show that most classifiers' accuracy improves as a result of dataset standardization; in fact, Some classifiers show a notable performance enhancement of up to 8.78% in accuracy. We found that most classifiers improved their accuracy when we compared their performance on normal and standardized datasets. hence, before implementing machine learning classifiers, standardizing the dataset is a helpful strategy for improving accuracy. Likewise, after hyperparameter adjusting the classifiers, we saw a considerable gain in accuracy. hence, algorithm tweaking is an additional helpful method for enhancing the algorithms' precision. We find that XGB and ET classifiers exhibit generally excellent accuracy when compared to other classifiers. But when it comes to changing the hyper parameters, SVM performs best, with a 96.72% success rate.

VI. Conclusions

The previous suggested systems' main flaw is that they become much less effective as the dataset size increases. While efficiently extracting dataset features helps improve machine learning, the fundamental issue is that data sets cannot be categorized well. One other thing is that up to a certain point, the accuracy of classifier predictions becomes better as the dataset gets bigger. However, beyond that point, it starts to hurt. The suggested approach states that cardiac disease prediction may be made more accurate and cost-effective by using machine learning methods. A variety of machine learning classifiers were used to predict cardiac disease; support vector machine (SVM) was able to reach an accuracy of 96.72%.

We plan to understand whether XG Boost can enhance the quality of pediatric heart disease prediction in the long term. Performance of classification of heart disease prediction will be great provided characteristics are managed appropriately. Results from future research using our suggested methodologies will be used as benchmarks for heart disease performance.

VII. REFERENCES

1. V. Manikantan and S. Latha, "Predicting the analysis of heart disease symptoms using medicinal data mining methods", International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.
2. Sellappan Palaniappan and Rafiah Awang, "Intelligent heart disease prediction system using data mining techniques", International Journal of Computer Science and Network Security, vol.8, no.8 pp. 343-350,2008.
3. K.Srinivas, Dr.G.Ragavendra and Dr. A. Govardhan," A Survey on prediction of

heart morbidity using data mining techniques",International Journal of Data Mining & Knowledge Management Process (IJDKP) vol.1, no.3, pp.14-34, May 2011.

4. G.Subbalakshmi, K.Ramesh and N.Chinna Rao," Decision support in heart disease prediction system using Naïve Bayes", ISSN: 0976-5166, vol. 2, no. 2.pp.170-176, 2011.
5. T.John Peter , K. Somasundaram, "An Empirical Study on Prediction of Heart Disease using classification data mining technique" IEEEInternational Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012.
6. Shamsher Bahadur Patel, Pramod Kumar Yadav and Dr. D. P.Shukla, "Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques",IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS),Volume 4, Issue 2 Jul. - Aug. 2013.
7. Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. I January - June 2007.
8. M.A.Nishara Banu, B.Gomathy, "Disease Forecasting System Using Data Mining Methods," International Conference on Intelligent Computing Applications,2014
9. Feixiang Huang, Shengyong Wang, and ChienChung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System", IEEE International Conference on Granular Computing,2012.
10. Chotirat "Ann" Ratanamahatana and Dimitrios Gunopulos,"Scaling up the Naive Bayesian Classifier:Using Decision Trees for Feature Selection", Computer Science Department University of California Riverside, CA 92521 1-909-787-5190
11. S. Goel, A. Deep, S. Srivastava, and A. Tripathi, ''Comparative anal- ysis of various techniques for heart disease prediction,'' in Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON), Mathura, India, Nov. 2019, pp. 88–94
12. A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, ''Machine learning techniques for heart disease prediction,'' Int. J. Sci. Technol. Res., vol. 8, no. 11, p. 97, Nov. 2019.
13. S. Mohan, C. Thirumalai, and G. Srivastava, ''Effective heart disease prediction using hybrid machine learning techniques,'' IEEE Access, vol. 7,pp. 81542–81554, 2019.
14. A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, ''Classification models for heart disease prediction using feature selection and PCA,'' Informat. Med. Unlocked, vol. 19, Jan. 2020, Art. no. 100330.
15. D. W. Hosmer, S. Lemeshow, and E. D. Cook, Applied Logistic Regression, 2nd ed. New York, NY, USA: Wiley, 2000.
16. E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi- Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, ''Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach,'' Pattern Recognit. Lett., vol. 133, pp. 33–40, May 2020
17. R. Atallah and A. Al-Mousa, ''Heart disease detection using machine learning majority voting ensemble method,'' in Proc. 2nd Int. Conf. new Trends Comput. Sci. (ICTCS), Oct. 2019, pp. 1–6.
18. A. Gupta, L. Kumar, R. Jain, and P. Nagrath, ''Heart disease pre-diction using classification (naive bayes),'' in Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (ICS). Singapore: Springer, 2020, pp. 561–573.
19. S. Mohan, C. thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning tech-niques," IEEE Access, vol. 7, pp. 81542–81554, 2019.
20. N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, andA. S. Sulthana, "Analysis and prediction of cardio vascular disease using machine learning classifiers," in Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 15–21,IEEE, Coimbatore, India, 2020