Improving the Prediction Accuracy of Onset of Cardiovascular Diseases, using Ensemble Learning

Muralidharan Jayaraman[1], Dr. Shanmugavadivu Pichai[2]

{ jaymurali@gmail.com[1], psvadivu67@gmail.com[2]}

The Gandhigram Rural Institute – Deemeed to be University, Tamil Nadu, India[1,2]

Abstract. This work presents an approach for classifying cardiac and non-cardiac data extracted from a dataset comprising of 70,000 records. The methodology begins with preprocessing, eliminating noisy and inconsistent data points using the box plot-based outlier removal technique. Subsequently, training and testing sets are taken out of the cleansed dataset, for model evaluation using a variety of base classifiers, such as support vector machines, decision trees, and random forests, within the ensemble framework. The experimental result of proposed method reveals the accuracy of the ensemble classifier model in classifying cardiac and non-cardiac data with an accuracy of 88.39%, with a focus on minimizing both false positives and false negatives.

Keywords: Ensemble Learning, Cardiovascular Diseases Prediction.

1 Introduction

Substantial public health challenges are caused by Cardiovascular disease (CVD) throughout the world. Consequently, patients, their families, and the governments of affected nations have faced substantial socioeconomic burdens. Prediction models employing risk stratification can identify individuals at elevated CVD risk. This paper evaluates the existing algorithms, and a new methodology is deployed to improve the accuracy level of prediction.

2 Literature Review

Within the realm of cardiovascular research, machine learning (ML) algorithms excel in capturing the intricate interplays and nonlinear relationships among variables and outcomes, surpassing the capabilities of conventional statistical models (1). Numerous investigations (2–6) have concurred that Support Vector Machines (SVM), and Random Forests (RF) outperform traditional models within this domain. Chintan, et al introduced a machine learning based approach which uses k-modes clustering with Huang initialization to improve the accuracy of classification. ML models such as XG Boost, random forest, and multilayer perceptron were utilized and fine-tuned using GridSearchCV.

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343340

3 Proposed Methodology

The major contributions of the proposed work are: Proposal of IQR-based outlier removal, Selection and tuning of ML algorithms for better classification of cardiac and noncardiac classes, and Performance comparison of proposed classification model with recent models.

The process flow of the proposed methodology comprises of data collection, data pre-processing, data splitting, and classification as shown in Figure 1. The collected CHD dataset features were enhanced for better data quality, by removal of irrelevant information, and the introduction of an additional variable, namely BMI. Then the features are normalized, and the outliers are removed, thereby performance is improved for the proposed predictive model. The enhanced features are employed for training and testing using various ML models in order to establish the most effective predictive model.

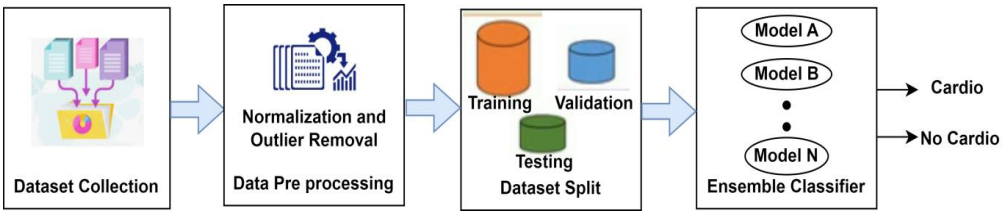

Fig. 1. Process Flow of Proposed Classification Technique

3.1 Dataset

The proposed work used the Cardiovascular Heart Disease (CHD) dataset which is available on the Kaggle website. The dataset has 70,000 patient records having 11 attributes, with cardio labels.

3.2 Data Pre-processing

Upon analysis of data, it is apparent that outliers exist within the dataset and it is analysed using the histogram information, which could be attributed to data entry errors. This technique facilitates the identification and elimination of data points that deviate significantly from the expected or normal range within a dataset. Box plots were generated for relevant variables, providing a visual representation of the data distribution.

The IQR is a robust measure of variability, less influenced by extreme values. Box plot is a valuable tool for understanding data distribution and variability, especially when dealing with data that may have outliers or a non-normal distribution. All instances of ap_hi, ap_lo, weight, and height that fell outside the range of 2.5% to 97.5% were manually removed. Consequently, after cleansing of data, the number of records decreased from 70,000 to 65,767.

3.3 Machine Learning Classifiers

The dataset was partitioned into two at 80:20 for training and testing, as per pareto principle. Respectively, the datasets are for training a predictive model, and evaluating the accuracy of the proposed model. ML classifiers were applied to the outlier removed dataset, for assessment of

the predictive capabilities. The accuracy of each classifier was rigorously evaluated based on a set of well-established metrics, including F-measure scores, accuracy, recall, and precision.

3.3.1 Ensemble Classifier (Adaboost M2)

This method leverages the collective intelligence of multiple base models. Adaboost M2 encompasses various strategies, such as Boosting, Bagging, and Stacking, each tailored to harness the individual model's strengths while easing out their inherent weaknesses.

Pseudo Code: Adaboost M2

Input:

- Training dataset D = $I\{(x_n, y_n)\}$, where x_n is a feature vector and y_n is the corresponding label (1(cardiac or 2(non cardiac)).

- The number of base classifiers t_n.

Initialization:

- Sample weights are initialized for each data point: $w_1 = 1/n$, $w_2 = 1/n$, wherein n refers to the number of data points.

- Initialize a list to store the base classifiers {H}

For t = 1 to t_n:

1. Train a weak classifier ht on the weighted training data D with weights $w_1,\ w_2,\ w_n$.
2. Calculate the weighted error rate εt of classifier ht on the training data:

$$J = \sum_{n=1}^{N} w_n\, I(y(x_n) \neq t_n)$$

3. Calculate the importance weight αt of classifier t_n.:

$$\alpha t_n = 0.5 * \ln(\frac{1-J}{J})$$

4. Update the weights for the training data points:

   For i = 1 to n:

$$w_n = w_n * \exp(\alpha t_n * y_n * x_n)$$

   Normalize the weights so that they sum up to 1: $w_1,\ w_2,\ w_n.$ = $\frac{w_1}{sum(\ w_1)}, \ldots\ldots\ldots \frac{w_n}{sum(\ w_n)}$

5. Add the weak classifier ht to the list of base classifiers: H.append

Output:

- The concluding ensemble classifier H(x) is a weighted combination of the weak classifiers:

$$H(x) = sign \sum_{t} (\alpha t_n * t_n)$$

Where:

- sign(z) returns -1 if z < 0, and 1 if z ≥ 0.

- t_n is the prediction of the nth weak classifier on input x.

- αt_n is the importance weight of the nth weak classifier. Top of Form

4 Results and Discussion

The model was implemented on a MATLAB, with the dataset comprised 70,000 feature values and 11 features, but after thorough cleansing and pre-processing, it became 68,329 rows with an incremental attribute of 13. Alongside these attributes, a single target output variable was included, where '1' and '0' indicate presence and absence respectively, of CVD (Cardiovascular Disease). The dataset exhibited a balance, with 35,000 patients diagnosed with CVD and 35,000 patients classified as normal. Given that all attributes were categorical in nature, an effort was made to enhance model efficiency by eliminating outliers. To facilitate model development and evaluation, the dataset was apportioned into training and testing sets having 80% and 20% of records respectively. The extracted features are classified using the Adaboost classifier.

As shown in Table 1, a variety of machine learning classifiers were applied to the normalized and raw datasets for detecting the disease following hyperparameter optimization and attained a better classification accuracy.

Table 1. Performance of the Proposed Model for Existing ML Classifier for Normalized and Raw Data

Existing ML Classifiers	Raw Data (Original)		Normalized Data	
	Training Accuracy (%)	Testing Accuracy (%)	Training Accuracy (%)	Testing Accuracy (%)
Naive Bayes	40	42.5	60	59.87
K-Nearest Neighbour Classifier	51.13	49.6	62.94	60.21
Decision Tree	49.57	47.9	64.87	63.54
Discriminate Analysis	52.68	50.27	64.6	65.98
SVM Classifier	72.14	72.18	75.14	79.15
XG Boost	59.15	52.87	73.25	75.12
LighGBM	60.12	61.54	74.23	72.64
Ensemble (Adaboost M2)	75.12	74.98	84.89	88.39

Table 2. Performance Measures of Adaboost M2 Classifier

Dataset Class	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
Cardio	82.6	78.3	80.39	84.89
Without Cardio	77.1	81.6	79.2	

The findings reveal that the ensemble Adaboost M2 classifier algorithm achieved the highest cross-validation accuracy at 84.89% as shown in Table 2, combined with other measures too being robust. The proposed approach is compared with other existing models as in Table 3.

Table 3. Performance Comparison of Proposed Approach with SOTA (CHD Dataset)

State-of-the-art Apporaches	Machine Learning Models	Accuracy (%)
Maiga, et al., 2019	-Random forest - Naive Bayes	73
	-Logistic regression	
	-KNN	
Waigi, at el., 2020	Decision Tree	72.77
Hana H. Alalawi, et al, 2021	Gradient Boosting	73
Shorewall, 2021	Stacking of KNN, Random Forest, and SVM outputs with Logistic Regression as the metaclassifier	75.1
Abdullah Alqahtani, et al, 2022	Machine Leaning [Stacked XGB, KNN, DT], Deep Learning DNN, KDNN, Majority Voting Ensemble algorithm	88.7
Bhatt, et al, 2023	MLP	87.23
Proposed	**Ensemble**	**88.39**

4.1 Research Outcome

• The research focuses on enabling advance detection of cardiovascular disease and the customization of treatment plans.

• Use of dataset with 70,000 patients and 11 variables is to avoid the risk of overfitting, and for robust model training and better generalization.

• Interquartile range-based outlier removal method was used, to clean the dataset and improve the accuracy of predictive models.

• Feature selection was done by assigning higher importance weights to informative features, improving accuracy and generalization through Adaboost M2 Classifier.

• Holistic view was taken to address challenges in disease prediction by combining data cleaning, feature selection, ensemble learning, and algorithm selection.

5 Conclusion

A robust methodology was applied to a dataset having 70,000 records to effectively utilize the importance of data pre-processing and ensemble classification. The pre-processing involved removal of noisy and inconsistent data points. Subsequent utilization of diverse base classifiers within an ensemble framework has showcased the experimental results, which underscore the vital role of data pre-processing and ensemble classification techniques in elevating the precision and reliability of cardiac anomaly detection. Finally, the impact of this work extends to better patient outcomes and improved healthcare quality, addressing the enduring challenge of reducing the morbidity and mortality associated with CVDs worldwide.

References

[1] Jiang Y, Zhang X, Ma R, Wang X, Liu J, Keerman M, Yan Y, Ma J, Song Y, Zhang J, He J, Guo S, Guo H. Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China. Clin Epidemiol. 2021 Jun 9;13:417-428. doi: 10.2147/CLEP.S313343. PMID: 34135637; PMCID: PMC8200454.

[2] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[3] https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[4] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, https://doi.org/10.48550/arXiv.1603.02754.
[5] Svetlana ulianova. Cardiovascular disease dataset. Retrieved from, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset; 2019, January 01.
[6] Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.