MACHINE LEARNING MODELS FOR PREDICTION OF CARDIOVASCULAR DISEASES

K.Sivaraman, V.Khanna

1Research Scholar, 2Professor, 1Department of Computer Science and Engineering, 2Department of Information Technology, Bharath Institute of Higher Education and Research Chennai-73, Tamil Nadu, India.

Corresponding Author Mail id: sivaraman2006@gmail.com

Abstract: Support Vector Machines (SVM) [9], Ada Boost (AB) [10], and Gradient Boosting (GB). Maximum Relevance, Minimum Redundancy (mRMR), Relief, and Least Absolute Shrinkage and Selection Operator (LASSO) are examples of fast correlation-based filters. The authors tested all attributes as well as the selected attributes generated by the above feature selection methods on the Cleveland heart disease dataset (CHDD) and the Hungarian heart disease dataset (HHDD). For their suggested framework, the authors were able to attain the greatest feasible model accuracy of 92.09 percent (all features) and 94.41 percent (selected features). The early detection of cardiac disease was demonstrated by several additional researchers.

Keywords: Machine Learning World Health Organization, cardiovascular disease, Support Vector Machines

1. Introduction

The term "cardiovascular disease" refers to a group of heart and blood vessel illnesses (CVD). Every year, more people die from CVDs than from any other cause in the world. According to the World Health Organization (WHO), 17.9 million people die each year from cardiovascular disease (CVD). It is estimated that it accounts for 31% of all deaths globally. In a CVD fact sheet, the WHO reported on another statistical analysis. By 2030, about 23.6 million people will die from CVDs, primarily heart disease and stroke, which will be the leading cause of mortality [1]. The greatest and simplest method to lower your risk of heart disease is to adjust your lifestyle behaviors, such as eating healthy foods and engaging in regular physical activity. Several tests are used by doctors to diagnose heart disease and stress tests during exercise [2]. These tests, on the other hand, are prone to side effects, take a long time to complete, and are costly. As a result, the requirement for early diagnosis of cardiac disease is unavoidable in order to protect patients at high risk from dying prematurely and to aid physicians in their prognosis.

Machine Learning (ML) is a rapidly expanding interdisciplinary topic that has recently attracted the attention of academics. The number of machine learning applications in the health-care industry is continually growing. In recent years, researchers and academicians have published various approaches for detecting cardiac problems. [3] proposed an intelligent framework for early detection of heart diseases, in which authors combined feature selection methods with Decision Tress (DT) [4], Random Forest (RF) [5, Logistic Regression (LR) [6, Nave Bayes (NB) [7, Artificial Neural Network (ANN) [8, Support Vector Machines (SVM)
[9], Ada Boost (AB) [10], and Gradient Boosting (GB) [11]. Maximum Relevance, Minimum Redundancy (mRMR) [13], Relief [14], and Least Absolute Shrinkage And Selection Operator (LASSO) [15] are examples of fast correlation-based filters. The authors tested all attributes as well as the selected attributes generated by the above feature selection methods on the Cleveland heart disease dataset (CHDD) and the Hungarian heart disease dataset (HHDD). For their suggested framework, the authors were able to attain the greatest feasible model accuracy of 92.09 percent (all features) and 94.41 percent (selected features). The early detection of cardiac disease was demonstrated by several additional researchers. Table 1 summarizes and reviews these state-of-the-art frameworks.

The state-of-the-art frameworks for cardiac disease detection are listed in Table 1. Table 1 shows the following observations.

- Several important datasets were used in the research, including CHDD, HHDD, Single Proton Emission Computed Tomography (SPECTF), and StatLog. Researchers like CHDD because it has more instances than any other publicly available database.

- In the identification of heart disease, feature engineering (FE) techniques take a back seat to data and machine learning (ML) models.

- The taxonomy of feature engineering (FE) in the listed state-of-the-art frameworks includes dealing with missing values, pervasive domain knowledge, introducing / removing dummy features, dealing with categorical features, creation of interactive and new features from existing raw features, and removal of unused / unwanted features.

- Several researchers experimented with different combinations of FE approaches to increase the performance of ML models.

- Current state-of-the-art results show that combining (assembling) ML models with appropriate FE approaches can outperform independent ML models.

- Various performance measures are used to evaluate the performance of the machine learning models and, as a result, the overall system.

- Accuracy, F1 score, recall, sensitivity, specificity, receiver operating characteristics curve (ROC) - area under the curve (AUC), true positive rate (TPR), and false positive rate (FPR) are the most commonly used metrics (FPR).

- In any of the aforementioned state-of-the-art frameworks, the Diagnostics Odds Ratio (DOR) and False Omission Rate (FOR) are not used to evaluate the performance of the models.

- Despite the fact that various frameworks for early diagnosis of cardiac disease have been developed, improvements in the model's robustness in terms of performance are still needed.
| Ref & Year | Dataset used for analysis | FE | ML Models used for comparative study | Performance metrics considered for analysis | Best Model (concluded) |
|------------|---------------------------|----|--------------------------------------|---------------------------------------------|------------------------|
| [19] 2015  | CHDD                      | Correlation, PCA | MLP, NB, J48, RF, SVM, AB, Boosted tree, and binary discriminant. | Classification accuracy, Specificity, Recall Precision, ROC, F-measure, MAE, RMSE | Ensemble-NB, AdaBoost, boosted tree |
| [21] 2016  | SPECT heart disease database | partitioning module, inner classifiers module and fuser module | AB, LR Boost, MLP, RF | sensitivity, specificity and classification accuracy | Hybrid ensemble model |
| [23] 2016  | CHDD | CFS, PSO | MLP, MLG, FURL, DT (C4.5) | Accuracy | Hybrid method (CFS + PSO + k means Clustering + MLP) |
| [16] 2016  | StatLog heart disease dataset | correlation | ANN SVM, kNN, CT, NB | Classification accuracy, sensitivity, specificity precision, negative predictive value (NPV), false positive rate (FPR), rate of misclassification (RMC), F1 measure | LR |
| [20] 2017  | Heart disease dataset, SPECTF dataset | BPSORS-AR | NB, SVM, ANN | Accuracy, Sensitivity and Specificity. | CFARS-AR |
| [19] 2018  | CHDD | Relief, mRMR, LASSO | LR, kNN, ANN, SVM (RBF kernel), SVM (linear kernel), NB, DT, RF | Specificity, Sensitivity, MCC, AUC | LR SVM (RBF) |
| [22] 2018  | CHDD | Relief, Correlation Filter, INFO gain Consistency, Chi-squared, Gain ratio, PSO | RF, NB, MLP, SVM | Accuracy | PSO with SVM |
| [18] 2018  | CHDD, HHDD, LBMC, and SUH | | | Accuracy, misclassification error, sensitivity (or recall), precision, specificity, F-score, ROC curve, AUC, and the K-S measure. | adaptive Boosting algorithm |
| [24] 2019  | Hospital in Beijing, China | XGBoost, one hot encoding | RF, GBDT and XGBoost | AUC, INE | XGBSVM hybrid model (XGBoost and SVM) |
| [25] 2019  | CHDD | L1 regularized SVM, L2 regularized SVM+ linear and RBF | AB, RF, Extra tree | Accuracy, sensitivity, specificity, MCC, ROC charts, AUC | HGSA- stacks two support vector machine (SVM) models |
| [26] 2019  | NSR-CAD, NSR-CHF and CAD-CHF | Nonlinear features and dimensional Reduction method as GDA, | GDA Method Having Kernel Function Like Gaussian, RBF and Polynomial + OS-ELM Having Sigmoid, Hardlim, RBF and Sine Activation | Mean ±SD, p Value | OS-ELM |
| [27] 2019  | CHDD | DT entropy | NB, GLM, LR, DL, DT, RF, GBT and SVM | accuracy, classification error, precision, F-measure, sensitivity and specificity | Hybrid HRFLM RF and Linear Method (LM) |
| [28] 2019  | CHDD, HHDD Switzerland SPECTF | MFSFSA, FFSA, RFSA and FSSA | A binary class RBF kernel-based SVM | accuracy, specificity and sensitivity | FSSA |
In this work, we propose a new framework for early detection of heart disease from CHDD.

### Understanding Dataset

Throughout our research, we used the Cleveland database (information about the creators can be found in Appendix-1). It can be found in the UCI machine learning repository. Cleveland database is a multivariate dataset containing 303 instances and 76 attributes. All of the published [Refs] experiments, on the other hand, use 14 of the 76 attributes as the subset of the Cleveland database. Table 2 contains descriptions of each attribute, the category of data contained in the attribute, the range of values, and the statistical summary. Table 2 shows that of the 13 CHDD attributes, A1, A4, A5, A8, and A10 are continuous data attributes, whereas the remainder are categorical data attributes (A2, A3, A6, A7, A9, A11, A12, and A13).
### Table 2: The synopsis of CHDD

| Attributes | Description | Attributes data category / Attribute’s values and its interpretation | Mean ± Std.dev |
|------------|-------------|---------------------------------------------------------------|----------------|
| sex (A2)   | Gender      | Categorical (Number of unique values = 2)                     |                |
|            |             | • 1 = male                                                   |                |
|            |             | • 0 = female                                                 |                |
| age (A1)   | Age in years | Continuous                                                   | 54.410596±9.040163 |
| trestbps (A4) | Blood pressure level (mm/Hg) (at admission time in the hospital) at rest | Continuous | 131.645695±17.612202 |
| cp (A3)    | Types of chest pain | Categorical (Number of unique values = 4)                    |                |
|            |             | • 1 = typical angina                                        |                |
|            |             | • 2 = atypical angina                                       |                |
|            |             | • 3 = non-anginal pain                                      |                |
|            |             | • 4 = asymptomatic                                          |                |
| fbs (A6)   | On fasting, the level of blood sugar level (mg/dl)           | Continuous     | 246.738411±51.858829 |
| chol (A5)  | serum cholesterols (mg/dl)                                  | Continuous     | 249.605960±22.912959 |
| thalach (A8) | Exercise Max Heart Rate Achieved | Continuous |                |
| restecg (A7) | Resting Electrocardiographic | Categorical |                |
|            |             | • 0 = normal                                                |                |
|            |             | • 1 = ST-T wave abnormal                                     |                |
|            |             | • 2 = left ventricular hypertrophy Estes)                    |                |
| oldpeak (A10) | ST depression induced by Exercise relative to Rest | Continuous | 1.035430±1.160723 |
| exang (A9) | Exercise Induced Angina                                   | Categorical (Number of unique values = 2)                    |                |
|            |             | • 1 = yes                                                   |                |
|            |             | • 0 = no                                                    |                |
| thal (A13) | Thalassemia                                            | Categorical (Number of unique values = 3)                    |                |
|            |             | • 3 = normal                                                |                |
|            |             | • 6 = fixed defect                                          |                |
|            |             | • 7 = reversible defect                                     |                |
| ca (A12)   | Major Vessels colored by Fluoroscopy                       | Categorical (Number of unique values = 4)                    |                |
|            |             | • 0 = fluoroscopy color 0                                   |                |
|            |             | • 1 = fluoroscopy color 1                                   |                |
|            |             | • 2 = fluoroscopy color 2                                   |                |
|            |             | • 3 = fluoroscopy color 3                                   |                |
| slope (A11) | Slope of Peak Exercise ST Segment | Categorical (Number of unique values = 3)                    |                |
|            |             | • 1 = upsloping                                             |                |
|            |             | • 2 = flat                                                  |                |
|            |             | • 3 = downsloping                                           |                |
| num (Outcome) | The diagnosis for heart disease | Categorical (Number of unique values = 2) |                |
|            |             | • 0 = absence of heart disease (class 0)                    |                |
|            |             | • x > 0 = presence of heart disease (class 1)               |                |

#### 3. Problem Formulation

The attribute “num” (outcome) denotes whether the patient has cardiac disease or not. It's an integer attribute with values ranging from 0 (no heart illness (class 0) through 1, 2, 3, and 4. (presence of heart disease in different states). The presence of heart disease in various states is addressed in this study as a single group dubbed "presence of heart disease" (class 1). As a result, the problem of heart disease is classified as a binary classification problem. The dataset under examination is fully dependent on early identification of cardiac problems
using machine learning techniques. The nature of the attributes in the dataset are not separable based on the classes, regardless of the datasets (used in the study of heart disease detection). Figure 1 depicts the population distribution of all CHDD traits in terms of classes. It is obvious from Figure 1 that all CHDD properties are not linearly separable. The difficulty in distinguishing between the presence of heart disease (class 0) and the absence of heart disease (class 1) is clearly demonstrated by the class wise data distribution of all CHDD features (figure 1). (class 0). Because of the nature of distribution data, detecting heart disease is a difficult task. We'll concentrate on supervised learning models as a data set with attributes and their corresponding classes in this research.

Figure 1: Class wise Data distribution of all features of CHDD

The CHDD of observations \( D = \{ (X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n) \} \) where \( X_n \) is an attribute vector and \( y_n \) is a class label (0 or 1), best subset (H) of D is arrived using different FE techniques and develop a binary classification model to learn from H to infer the value of yi given xi. FE includes dealing of outliers and missing values, data standardization, attributes selection and cross validation. LR, kNN, DT, NB, SVM, AB, XB and ensembling of different combination of the models are considered for performing evaluation of binary classification on the given dataset. In addition with sensitivity, specificity, AUC-ROC, F1 score and FOR, the independent prevalence metric DOR is also used to evaluate the performance of the models. The model providing more the number of best values of chosen metrics is considered as the best performing model.

4. Proposed Framework

The proposed framework is illustrated in Figure 2. Imputation of missing values and outlier removal processes are the integral part of the proposed framework. These two processes are inevitable steps to provide the quality data for the learning of ML model.
4.1. Dealing of missing values (IMV)

Missing values in a data set can be handled in one of two ways: they can be ignored or they can be accounted for. Dropping missing values results in a loss of information and precision for the data collected, as well as a reduction in the number of data available for training the model. Missing data is usually not fixed; rather, it is dealt with in order to deal with it. Impu does the missing data accounting. The task is to consider how to handle missing data using the imputation procedure that is provided for accounting missing data. When data is applied to the model, imputation by mean or median value progresses the model's performance metrics. The following is the formula for calculating the median for an odd number of observations of a feature in the data set (303 for CHDD).

\[
\text{Median} = \frac{n+1^{th}}{2} \text{ term}
\]

Where \( n \) is the number of observations. While analyzing the unique values of all the features in the CHDD under consideration, it was discovered that missing values ('?') exist in ca (A12) and thal (A12) (A13). Four values with '?' in A12, the feature that describes the Major Vessels colored by Fluoroscopy, are found in the index [165, 191, 286, 301]. Only one data has a result of '1' among the four values (heart disease). A12 is the categorical value (0 represents fluoroscopy color 0, 1 represents fluoroscopy color 1, 2 represents fluoroscopy color 2, and 3 represents fluoroscopy color 3). Two values with '?' in A13, the feature that describes Thalassemia, are discovered in the index [86, 265]. Only one data has a result of '1' among the two values (heart disease). The category value for A13 is 3 (normal), 6 (fixed defect), and 7 (reversible defect). Because both are categorical data, the missing values are imputed with median values.

4.2. Dealing of outliers

Outliers are measurements that are markedly different from the rest of the data. These outliers are highly sensitive to data distribution and have a direct impact on the learning process of machine learning classifiers. Figure 1 depicts the positive or negative skewness of the characteristics A1, A4, A5, A8, and A10 (asymmetry). These are continuous data qualities, which is intriguing. As a result, the continuous data variables in this study are subjected to outlier analysis. A thorough understanding of data is critical in health care domain data analysis. Data are rarely treated as outliers in health care measurement in this context, despite the fact that they are statistically defined as such. Outlier data, on the other hand, has a direct impact on the performance of machine learning classifiers. Blood pressure levels are classified as normal (120/80 mm/Hg), high (120-129 /80 mm/Hg), Stage 1 (130-139 / 80-89), Stage 2 (140/90 mm/Hg), and hypertensive crisis (> 180/>120 mm/Hg) by the World Health Organization (WHO). Human blood pressure should be tested if it is greater than 180 mm Hg, according to the standards (hypertensive). The attribute A4 in particular has a value of

![Figure 2: Proposed Framework](image-url)
200 mm/Hg at the 125th incidence of CHDD, which is statistically regarded an anomaly. Imputing this with any of the statistical parameters of the appropriate property will be meaningless in this case. Also, retaining this value in the dataset will affect the performance of the ML classifier. To overcome the aforementioned scenario, the outlier rejection (dropping) process is preferred in the proposed framework rather imputation of outliers with any of the statistical parameters. The process of dealing outliers used in the proposed framework algorithmically framed using Mahalanobis Distance (MD) metric [34] as given below.

**Algorithm: Dealing of Outliers**

Input: The data $[A_i]_{MxN}$, where $M$ is the number of rows containing outlier data and $N$ is the number of columns describing the multivariate features

Output: The data $[A_i]_{PxN}$ where $P$ is the number of rows of outlier data and $N$ is the number of columns describing the multivariate features

Load data $X$ and compute its mean $\bar{A_j} = \frac{1}{M} \sum_{i=1}^{M} A_i = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)$ and $j = 1$ to $N$ (1)

Compute $N \times N$ covariance matrix $C_{NxN} = \sum_{j=1}^{N} (A_i - \bar{A})(A_i - \bar{A})^T$ (2)

Compute inverse of covariance matrix $C^{-1} = \frac{1}{|C|} Adj(C)$ (3)

Compute Mahalanobis Distance (MD) for the input data $A_i$ using formula

$$D_i = \sqrt{(A_i - \bar{A})^T C^{-1} (A_i - \bar{A})}$$ (4)

Obtain threshold required for $D_i$ for statistical outlier detection using chi-square percent point function (ppf) as $D_i$ has a chi-square distribution with degrees of freedom ($p$).

$ppf$ (lower tail probability, $p$), where $p$ is degrees of freedom (number of attributes considered for distance measurement)

$[A_i]_{PxN}$ is stored with data containing $D_i > threshold$

The MD is a multivariate distance metric that measures the standard deviation of a data instance from the data distribution's mean. The attributes considered here are $A_1, A_4, A_5, A_8$, and $A_{10}$, and the actual correlation relationship between them is displayed in Figure 3. The covariance of $A_8$ (maximum heart rate) is high since it has a strong association with other qualities. MD adjusts the variance of the attributes to 1 and calculates the distance after transforming the attributes into uncorrelated variables. By dividing the large covariance ($C^{-1}$) as shown in equation 4, the distance is significantly minimized.

![High Positive](image1.png)

![No Correlation](image2.png)

![High Negative](image3.png)

**Figure 3** Correlation strength of the attributes with the target
The key to removing outliers is to fix the threshold. Because there are five qualities involved in the distance computation, the Chi-squared test statistics is used with $p=5$, where $5$ is the degree of freedom. The number of occurrences to be deleted was also determined for different values of the probability (lower tail of the distribution). Table 3 shows the number of instances discarded in the dataset as a function of the ppf value (threshold).

### Table 3: Lower Tail Probability and Ppf Value

| S.No | Lower Tail Probability | ppf value (threshold distance) | No. of instances dropped |
|------|------------------------|-------------------------------|--------------------------|
| 1    | 0.99                   | 15.086                        | 6                        |
| 2    | 0.97                   | 12.375                        | 13                       |
| 3    | 0.95                   | 11.07                         | 18                       |
| 4    | 0.93                   | 10.191                        | 21                       |
| 5    | 0.91                   | 09.521                        | 26                       |

**Figure 4.** MD plot with threshold line (outliers for the experiment values presented in Table 3 (row~1) is highlighted in orange)

The threshold distance lowers as the lower tail probability drops (see Table 3). As a result, additional data instances are removed. Using 15.086 as a cut-off value for identifying outliers, it is discovered that 6 data instances are identified as outliers (see Figure 4) and those instances are only eliminated to avoid removing more data instances.

The existence of outliers is indicated by heavy tails in the data distribution (Figure 5(a)) in the continuous data attributes. The remaining CHDD attributes are categorical data attributes that do not have any positive or negative skewness. The length of the lower tail is shortened and the skewness of the attribute's distribution is adjusted to zero mean when outliers are removed (see Figure 5(b)). Because only 6 data instances are removed, other statistically defined outliers are still found. In this literature, statistically identified outliers are dropped in favor of imputation of outliers when considering health care data.
4.3. Cross Validation and Grid Search Technique

The technique of cross validation (CV) is used to generalize a machine learning model for an independent dataset. The key to the optimum performance of the ML model with fresh data is its generalisation. K-fold CV is a popular CV strategy used by machine learning practitioners to manage the bias-variance trade-off of their models. Figure 6 depicts the fold CV, which is utilized in this literature. The CHDD is split in to K (K=5) folds and K-1 (4 folds) used in the training phase. The optimum hyperparameters of the ML model are found using a grid search strategy. The same percentage of samples for each class (presence of heart disease / non-presence of heart disease) was maintained using a Stratified K-fold CV, as shown in i. The model's performance is calculated in each fold, and the average of all five folds is used to calculate the final model's performance.

**Figure 5:** Data distribution of attributes with box plot (a) before outliers' removal (b) after outliers' removal
Figure 6: KCV and grid search of fine tuning of hyper parameters used for model’s performance evaluation

The optimum hyper parameters of the ML model are found using a grid search technique. It considers all of the model's hyper parameter combinations in detail. Table 4 lists the hyperparameters that were considered in the extensive experiments for several ML models.

| ML Model | Hyperparameters [values] considered |
|----------|-----------------------------------|
| LR       | penalty=['l1', 'l2', 'elasticnet'], C (inverse regularization)=[0.2,0.3, 0.6, 0.8, 1], solver=['newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga'] |
| DT       | criterion = ['gini','entropy'], splitter = ['best'], min. sample leaf = [1,2,3,4,5], min. samples split = [0,1,0,2,0,3,4,0,5,0,6,0,7,0,8,0,9,1,0] |
| RF       | criterion = ['gini','entropy'], number of estimators = [50,100,150,200,250] |
| NB       | variance smoothing = [1e-01,1e-02,1e-03,1e-04,1e-05,1e-06,1e-07,1e-08,1e-09,1e-10,1e-11,1e-12] |
| k-NN     | leaf size = [5,10,15,20,25,30,35,40,45,50], distance = [1,2], number of neighbors = [1,3,5,7,9,11,13,15,17,19,21,23,25,27,29,31,33,35,37,39,41,43,45,47,49] |
| SVM      | kernel=['linear','rbf','sigmoid'], C (regularization) = [0.3, 0.6, 0.8, 1], gamma = ['scale','auto'] |
| AB       | learning rate = [0.1,0.5,1.0], number of estimators = [10,50,100,200], algorithm = ['SAMME','SAMME.R'] |
| XB       | max. depth = [3, 4, 5], min. child weight = [1, 5, 10], gamma = [0.5, 1, 1.5, 2, 5], subsample=[0.5,1.0], coreset [0.6, 0.8, 1.0] |

5. Results and Discussions

5.1. Performance Evaluation – Environment and Metrics

The system utilized to develop the ML models and execute the lengthy model evaluation tests has an Intel (R) Core (TM) i7-4510U CPU @ 2.60 GHz, installed with RAM of 8.0 GB, and was running Windows 10 Operating System with Jupyter notebook and Python as the programming environment.

Experiments were carried out to assess the ML models. The ML models are assessed using a variety of measures, each of which provides a unique viewpoint on the study and evaluation of the models. Because the topic at hand is a binary classification problem with an unbalanced dataset, judging ML models just on accuracy (the proportion of accurate classifications) will lead to erroneous conclusions. The confusion matrix, which reports True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP), is used to calculate the majority of the metrics (FP). Other measures like Accuracy (Acc), F1 score (F1), Precision (Pre), Specificity (Spe), Sensitivity (Sen), Diagnostics Odds Ratio (DOR), and False Omission Rate
(F1OR) are calculated and reported using these four metrics. The parameters stated above are provided in their absolute values. In addition to these measures, the model's Area Under the Curve (AUC) with Receiver Operating Characteristics (ROC) is reported to evaluate the model's prediction ranking. These measures demonstrate the ML models' diagnostic capacity.

5.2. Performance Evaluation Results for ML model
In sections 4.1, 4.2, and 4.3, the outcomes of CHDD preprocessing are discussed. The results and robustness of the ML model are given in this section. The quantitative values of performance measures derived for several ML models are provided in Table 5, together with their standard deviations. The best performing classifier model is decided based on the performance metric in the last column of Table 5. The performance statistic with the highest value is highlighted (green colour). The two important metrics for the binary classification problem are accuracy and AUC, where accuracy is based on a threshold and AUC is a test of the quality of the internal value generated by the classifier. As CHDD is imbalance dataset, AUC is preferred over accuracy which helps in determine the robust classifier by providing insights over TPR and false FPR. Therefore, Table 6 is derived from Table 5, in a way to summarize AUC of the different models for different preprocessing. The best value of the performance metric is highlighted (green colour).

### Table 5: Performance metrics obtained for ML models under different pre-processing

| Preprocessing | Metric | LR | DT | RF | NB | kNN | SVM | AB | XB | Best Model |
|---------------|-------|----|----|----|----|-----|-----|----|----|------------|
| **IMV (Exp #1)** | AUC | 0.916±0.027 | 0.826±0.060 | 0.894±0.034 | 0.914±0.031 | 0.743±0.059 | 0.912±0.026 | 0.907±0.029 | 0.901±0.025 | LR |
| | Acc | 0.844±0.052 | 0.748±0.077 | 0.824±0.032 | 0.848±0.053 | 0.679±0.057 | 0.834±0.048 | 0.821±0.030 | 0.827±0.049 | NB |
| | Sen | 0.805±0.074 | 0.690±0.094 | 0.783±0.072 | 0.834±0.095 | 0.532±0.086 | 0.791±0.082 | 0.769±0.080 | 0.776±0.076 | NB |
| | Spe | 0.877±0.056 | 0.797±0.102 | 0.859±0.033 | 0.859±0.032 | 0.804±0.055 | 0.871±0.037 | 0.865±0.032 | 0.871±0.038 | LR |
| | Pre | 0.849±0.062 | 0.752±0.110 | 0.826±0.031 | 0.833±0.040 | 0.698±0.084 | 0.839±0.043 | 0.831±0.027 | 0.836±0.045 | LR |
| | DOR | 0.157±0.050 | 0.248±0.063 | 0.174±0.045 | 0.137±0.069 | 0.330±0.047 | 0.167±0.058 | 0.181±0.050 | 0.178±0.053 | NB |
| | FOR | 45.346±28.704 | 28.983±44.22 | 26.704±12.738 | 75.642±97.794 | 5.684±2.764 | 43.480±43.983 | 25.720±11.696 | 34.607±27.254 | NB |
| | LR | 0.825±0.062 | 0.716±0.088 | 0.803±0.043 | 0.832±0.065 | 0.602±0.079 | 0.813±0.058 | 0.796±0.043 | 0.804±0.059 | NB |

The investigation of Table 5 and Table 6 provides the following evidences. NB outperforms other models on 5 metrics (Acc, Sen, FOR, DOR, F1) and LR outperforms other models on 3

---

The investigation of Table 5 and Table 6 provides the following evidences. NB outperforms other models on 5 metrics (Acc, Sen, FOR, DOR, F1) and LR outperforms other models on 3
metrics (AUC, Spe, Pre) for IMV preprocessing (Table 5, Exp #1). NB outperforms other models on 4 metrics (Sen, FOR, DOR, F1), LR outperforms other models on 2 metrics (Spe, Pre) and LR & NB provides 2 metrics (AUC, Acc) with equal values for IMV+OR (Table 5, Exp #2). NB outperforms other models on 5 metrics (Acc, Sen, FOR, DOR, F1), LR outperforms other models on 2 metrics (AUC, Pre) and kNN outperforms other models on 1 metric (Spe) for IMV+OR+SS preprocessing (Table 5, Exp #3). LR and NB are outperforming all other models in respect of most of the performance metrics reported for different preprocessing methods applied on the dataset. Exceptionally kNN is outperforming the other models in respect of Spe for the preprocessing of IMV+OR+SS. LR, DT, NB, SVM and AB demonstrate the best AUC values for IMV preprocessing (Table 6). RF provides better AUC value for IMV+OR preprocessing (Table 6). kNN and XB exhibits the best AUC values for IMV+OR+SS preprocessing (Table 6). From the above discussion, it is evident that NB is performing better than all other models on CHDD. CHDD has more categorical attributes (see Table 2) which are uncorrelated and mutually exclusive. This is a key factor that driven NB to produce better performance.

Table 6: Summary of tuned best hyperparameters for the best preprocessing with best AUC values (highlighted)

| Preprocessing | IMV   | IMV+OR | IMV+OR +SS | Best preprocessing | Best hyperparameters                                      |
|---------------|-------|--------|------------|--------------------|-----------------------------------------------------------|
| Model         |       |        |            |                    |                                                           |
| LR            | 0.916±0.027 | 0.911±0.026 | 0.912±0.028 | IMV               | penalty=l2, C=1, solver=newton-cg                        |
| DT            | 0.826±0.060 | 0.822±0.041 | 0.819±0.043 | IMV               | criterion= entropy, min. samples leaf= 5, min. samples split= 0.1, splitter =best |
| RF            | 0.894±0.034 | 0.896±0.034 | 0.893±0.026 | IMV+OR            | criterion =entropy, number of estimators =150            |
| NB            | 0.914±0.031 | 0.911±0.032 | 0.910±0.033 | IMV               | Variance smoothing = 1e-06                              |
| kNN           | 0.743±0.050 | 0.731±0.058 | 0.909±0.033 | IMV+OR+SS         | leaf size = 10, p =1, No. of neighbors = 27             |
| SVM           | 0.743±0.050 | 0.909±0.029 | 0.910±0.029 | IMV               | kernel=linear, C = 0.6, gamma = scale                   |
| AB            | 0.907±0.029 | 0.871±0.093 | 0.871±0.093 | IMV               | learning rate =0.1, no. of estimators =200, algorithm =SAMME |
| XB            | 0.901±0.025 | 0.903±0.023 | 0.905±0.025 | IMV+OR+SS         | max. depth = 4, min. child weight = 5, gamma = 1.5, subsample= 1.0, colsample_bytree= 0.6 |

6. Conclusion
The tuned best hyperparameters (5-fold CV and grid search) are reported for the best possible AUC of all the models in Table 6. In OR process, data instances identified as outlier (statistically) are dropped. This majorly downgrade the performance of most of the models.
Subsequently, the addition of standardization of continuous attributes of CHDD is also not guaranteed the improvement of the models performance. It is observed from Table 4 and 5 that DT demonstrates underperformance when compared with the other models. Encoding categorical value (one-hot) is the major cause of DT’s underperformance as the encoding categorical attributes introducing sparsity in the CHDD which is not desirable for DT. Since DT is considered as the base learner for the boosting classifier (AB), the performance of AB is also not appreciable.

References

[1] World Health Organization (2017), Cardio Vascular Diseases fact sheet, Available Online: https://www.who.int/cardiovascular_diseases/about_cvd/en/

[2] Canters for disease control and prevention, Know the fact about Hear Disease, Available Online: https://www.cdc.gov/heartdisease/docs/consumered_heartdisease.pdf

[3] Muhammad, Y., Tahir, M., Hayat, M. et al. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Sci Rep 10, 19747 (2020). https://doi.org/10.1038/s41598-020-76635-9

[4] I. Jenhani, N. B. Amor, and Z. Elouedi, “Decision trees as possibilistic classifiers,” Int. J. Approx. Reasoning, vol. 48, no. 3, pp. 784-807, Aug. 2008. https://doi.org/10.1016/j.ijar.2007.12.002

[5] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001. https://doi.org/10.1023/A:1010933404324

[6] Irfana P. Bhatti, Heman D. Lohano, Zafar A. Pirzado and Imran A. Jafri, 2006. A Logistic Regression Analysis of the Ischemic Heart Disease Risk. Journal of Applied Sciences, 6: 785-788. https://dx.doi.org/10.3923/jas.2006.785.788

[7] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive bayes: Aggregating one-dependence estimators," Mach. Learn., vol. 58, no. 1, pp. 5-24, Jan. 2005. https://doi.org/10.1007/s10994-005-4258-6

[8] A. Reinhardt and T. Hubbard, "Using neural networks for prediction of the subcellular location of proteins," Nucleic Acids Res., vol. 26, no. 9, pp. 2230-2236, May 1998. https://doi.org/10.1093/nar/26.9.2230

[9] C. Cortes and V. Vapnik, “Support-vector networks," Mach. Learn., vol. 20, pp. 237-297, Sep. 1995.

[10] B. Kégl, "The return of AdaBoost.MH: Multi-class Hamming trees," 2013, Available Online: http://arxiv.org/abs/1312.6086

[11] Kartik Budholiya, Shailendra Kumar Shrivastava, Vivek Sharma, An optimized XGBoost based diagnostic system for effective prediction of heart disease, Journal of King Saud University - Computer and Information Sciences, 2020. https://doi.org/10.1016/j.jksuci.2020.10.013

[12] L. Yu and H. Liu. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In Proceedings of The Twentieth International Conference on Machine Learning (ICML-03), 856-863, Washington, D.C., August 21-24, (2003). https://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf

[13] Hanchuan Peng, Fuhui Long, and Chris Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp.1226-1238, (2005). https://doi.org/10.1109/TPAMI.2005.159

[14] KIRA, K. and RENDEL, L. The Feature Selection Problem: Traditional Methods and a new algorithm. Proc. Tenth National Conference on Artificial Intelligence, MIT Press, 129-134, (1992). https://www.aaai.org/Papers/AAAI/1992/AAAI92-020.pdf
[15] Tibshirani, Robert, "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1): 267–88, (1996). http://www.jstor.org/stable/2346178

[16] Nguyen Cong Long, Phayung Meesad, Herwig Unger, A highly accurate firefly based algorithm for heart disease prediction, Expert Systems with Applications, Volume 42, Issue 21, 30 November 2015, Pages 8221-8231
https://doi.org/10.1016/j.eswa.2015.06.024

[17] Verma, L., Srivastava, S. & Negi, P.C, A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data, J Med Syst., 40, 178 2016.
https://doi.org/10.1007/s10916-016-0536-z

[18] Kathleen H. Miao, Julia H. Miao, George J. Miao, Diagnosing Coronary Heart Disease Using Ensemble Machine Learning, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016, pp. 30–39
https://theias.org/Downloads/Volume7No10/Paper_4Diagnosing_Coronary_Heart_Disease_using_Ensemble_Machine_Learning.pdf

[19] Ashok Kumar Dwivedi, Performance evaluation of different machine learning techniques for prediction of heart disease, Neural Comput & Applic, Springer, 2016, https://doi.org/10.1007/s00521-016-2604-1

[20] Gupta N, Ahuja N, Malhotra S, Bala A, Kaur G. Intelligent heart disease prediction in cloud environment through ensembling. Expert Systems. 2017:e12207. https://doi.org/10.1111/exsy.12207

[21] Haq, A., Li, J., Memon, M., Nazir, S., & Sun, R, A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms. Mob. Inf. Syst., 2018, 3860146:1-3860146:21. https://doi.org/10.1155/2018%2F3860146

[22] J. Vijayashree H. Parveen Sultana, A Machine Learning Framework for Feature Selection in Heart Disease Classification Using Improved Particle Swarm Optimization with Support Vector Machine Classifier Programming and Computer Software, 2018, Vol. 44, No. 6, pp. 388–397
https://link.springer.com/article/10.1134/S0361768818060129

[23] Elham Nikookar, Ebrahim Naderi, Hybrid Ensemble Framework for Heart Disease Detection and Prediction, International Journal of Advanced Computer Science and Applications, Vol. 9, No. 5, 2018, pp. 243-248
https://theias.org/Downloads/Volume9No5/Paper_33-Hybrid_Ensemble_Framework_for_Heart_Disease.pdf

[24] Wenbing Chang , Yinglai Liu, Xueyi Wu, Yiying Xiao,Shenghan Zhou , Wen Cao, A New Hybrid XGBSVM Model: Application for Hypertensive Heart Disease, IEEE Access, Vol 7, 2019, pp- 175248- 175258.
https://doi.org/10.1109/ACCESS.2019.2957367

[25] Liaquat Ali , Awais Niamat, Javed Ali Khan, Noorbakhsh Amiri Golilzar, Xiong Xingzhong , Adeeb Noor, Redhwan Nour, Syed Ahmad Chan Bukhari, An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure, IEEE Access, Vol 7, 2019, pp- 54007 – 54014
https://doi.org/10.1109/ACCESS.2019.2909969

[26] Sulekha Saxena, Vijay Kumar Gupta, P. N. Hrisheekesh, Coronary Heart Disease Detection Using Nonlinear Features And Online Sequential Extreme Learning Machine, Biomedical Engineering: Applications, Basis and Communications, Vol. 31, No. 6 (2019) 1950046 (15 pages) https://doi.org/10.4015/S1016237219500467

[27] Senthilkumar Mohan, Chandrasagar Thirumalai, Gautam Srivastava, Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques,IEEE Access, Vol 7, 2019, pp- 81542- 81554
https://doi.org/10.1109/ACCESS.2019.2923707

[28] Syed Muhammad Saqlain, Muhammad Sher, Faiz Ali Shah, Imran Khan, Muhammad Usman Ashraf, Muhammad Awais,Anwar Ghani, Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines, Knowl Inf Syst (2019) 58:139–167
https://doi.org/10.1007/s10115-018-1185-y
[29] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour, An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection, IEEE Access, VOLUME 7, 2019, pp 180235 – 180243, https://doi.org/10.1109/ACCESS.2019.2952107
[30] Ankur Gupta, Rahul Kumar , Harkirat Singh Arora, Balasubramanian Raman, MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis, IEEE Access, Vol 8, 2020 pp 14659- 14674, https://doi.org/10.1109/ACCESS.2019.2962755
[31] Norma Latif Fitriyani, Muhammad Syafrudin,Ganjar Alfian, Jongtae Rhee, HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System, IEEE ACCESS, Vol 8, 2020 pp- 133034- 133050 https://doi.org/10.1109/ACCESS.2020.301051
[32] Jian Ping Li, Amin Ul Haq, Salah Ud Din, Jalaluddin Khan, Asif Khan, Abdus Saboor, Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare, IEEE ACCESS, Vol 8, 2020 pp-107562-107582 https://doi.org/10.1109/ACCESS.2020.3001149
[33] Dhyan Chandra Yadav, Veer Bahadur Singh, Saurabh Pal, Veer Bahadur Singh, Analysis of Heart Disease Using Parallel and Sequential Ensemble Methods With Feature Selection Techniques:Heart Disease Prediction,International Journal of Big Data and Analytics in Healthcare, Volume 6 • Issue 1 • January-June 2021. https://orcid.org/0000-0001-9545-7481.
[34] Mahalanobis, Prasanta Chandra "On the generalised distance in statistics" (PDF). Proceedings of the National Institute of Sciences of India.,1936, 2 (1): 49–55