Comparative Analysis of Machine Learning Algorithms and Data Mining Techniques for Predicting the Existence of Heart Disease

Nourah Alotaibi[1], Mona Alzahrani[2]
Department of Information and Computer Science (ICS)
KFUPM
Dhahran, Saudi Arabia

***Abstract*—Heart diseases are considered one of the leading causes of death globally over the world. They are difficult to be predicted by a specialist physician as it is not an easy task which requires greater knowledge and expertise for prediction. With the variety of machine learning and deep learning algorithms, there exist many recent studies in the state of the art that have been done remarkable and practical works for predicting the presence of heart diseases. However, some of these works were affected by various drawbacks. Hence, this work aims to compare and analyze different classifiers, pre-processing, and dimensionality reduction techniques (feature selection and feature extraction) and study their effect on the prediction of heart diseases existence. Therefore, based on the resulting performance of several conducted experiments on the well-known Cleveland heart disease dataset, the findings of this study are: 1) the most significant subset of features to predict the existence of heart diseases are PES, EIA, CPT, MHR, THA, VCA, and OPK, 2) Naïve Bayes classifier gave the best performance prediction, and 3) Chi-squared feature selection was the data mining technique that reduced the number of features while maintained the same improved performance for predicting the presence of heart disease.**

Keywords*—Heart disease; feature selection; feature extraction; dimensionality reduction; Chi-squared; Naïve Bayes; Cleveland dataset

I. Introduction

Cardiovascular diseases (CVDs) [1] are when the heart and blood vessels affected by some diseases like coronary heart disease and heart failure disease. The statistics in Saudi Arabia that were collected over the past 40 years indicate that the deaths have increased from CVDs. Moreover, according to the World Health Organization (WHO)[1], 17.9 million deaths every year resulted of CVDs including different heart diseases such as cardiovascular disease, valvular heart disease, heart defects, heart infections or cardiomyopathy [2].

Correctly predicting a diagnosis, including predicting the Existence of Heart Disease (EHD), is essential to patient-centered care, equally in choosing healing plans and notifying patients as a basis for shared decision making [3]. In recent years, there have been plenty of studies on EHD prediction. However, EHD prediction research has passed through three different stages along with history. In 1979, two researchers [4] combined diverse results gained from examinations like stress electrocardiography and cardiotocography, and others into a diagnostic decision about the likelihood of getting a disease in a particular patient through Bayes' Theorem . While in 1998, the second stage started when Wilson et al. [5] established a new direction concerning heart diseases estimation by utilizing risk factor classes with the aid of logistic approaches and regression calculations. Nowadays, several researchers have developed various machine learning algorithms [3, 6–10] to predict the EHD on the publicly available datasets which are the focus of this work.

However, the machine learning-based studies were affected by various drawbacks. For example, using datasets without handling the imbalance classes [6, 10, 11], important features such as age were manually excluded from the experiments [6, 9, 10], no comparisons were made in terms of prediction methods [6] or the used dataset [6, 11, 12], various platforms were used for assessments [10], some important details were missing or not clear such as the number of selected features [7] or the number of samples per class [3, 8, 11]. Even more, surprisingly, no coloration exists between the selected features among these works [3, 6–10] even though some of them were using the same dataset. In addition, different machine learning algorithms were selected as the best classifiers in various works according to their experimented data mining approaches. To sum up, there is still a need to experimentally evaluate different classifiers with different data mining approaches to gain a final decision. Hence, inspired by the development of several machine learning-based models for improving the EHD prediction, this study contributes to the literature by providing a work that handles these drawbacks as follows:

- Study the effect of different balancing solutions such as oversampling and undersampling for EHD prediction to handle the imbalance classes issue that exist on Cleveland heart disease dataset.
- Explore the most significant subset of features for the EHD prediction by analyzing different data mining techniques.
- Investigate the best performed classifier for the EHD prediction by comparing different machine learning algorithms.
- Achieve the highest performance for EHD prediction compared to recent related works that experimented

[1]World Health Organization. Global Health Observatory: Cardiovascular Diseases-Country Statistics. Retrieved on March 11, 2022 from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1.

on the well-known Cleveland heart disease dataset by training the best performed classifier with the most significant subset of features.

Consequently, the following set of research questions were explored:

- RQ1: What are the best-performing feature sets for the EHD prediction?
- RQ2: How can the prediction performance be improved using data mining?
- RQ3: What is the most appropriate classifier using the selected features for EHD prediction?

The remaining of this paper is organized as the following: Section II presents the related works which provide several EHD prediction machine learning-based models. Section III explains the proposed methodology. Section IV demonstrates the achieved outcomes and tackles the discussion in light of the findings. Finally, Section V concludes this study and discuses future works.

II. Literature Review

Several heart diagnosis studies in the state of the art [3, 6–14] have been done extraordinary works that contributed by providing different prediction approaches. These studies could be categorized based on the targeted prediction such as Heart Failure (HF) prediction [7], mortality or hospitalization prediction of the HF patient [3, 6], and EHD prediction [8–14]. In addition, they also could be categorized in terms of the investigated learning technique, whether it is supervised learning [3, 8, 11, 12], ensemble learning [6, 12], deep learning [7, 9] or even hybrid learning [10]. Table I summaries the recent related works [3, 6, 7, 10–12] in terms of their experimented datasets, pre-processing techniques, learning methodologies, performance evaluation and drawbacks.

For HF prediction, Maragatham et al. [7] took advantage of the availability of the intensive substantial historical information in Electronic Health Record (EHR) and related time stamped data, in which, the authors inspected whether the usage of deep learning would improve the model performance for early HF diagnosis. The tested data consists of 365,446 patients, where 4289 of them had HF. The examination of time stamped EHRs aided in identifying the relations between numerous diagnosis events and predicting when a patient is being examined for a disease. Medical concept vectors and Long Short-Term Memory (LSTM) network were used to determine the diagnosis events and HF prediction. The proposed model was trained using one-hot vectors and grouped code vectors. As an activation function, SiLU and tanh were used in the hidden layers, while in the output layer, Softmax was used. For weight optimization through the network, Bridgeout, a regularization technique was used. K-nearest neighbour (KNN) and SVM were implemented using Python Scikit-Learn version 0.16.1 while Theano 0.7 was used to implement LSTM network, multilayer perceptron (MLP), and Logistic Regression (LR) models. They conducted two different experiments according to the length of the prediction and observation windows; and the performance was compared to well-known supervised approaches such as LR, MLP, SVM and KNN. Specifically, the first experiment, when using 12-months and 6-months as observation and prediction windows respectively, gave an AUC of 0.797, and when using 18-months and 0-months gave 0.894 AUC.

While for mortality prediction of the HF patient, Adler et al. [6] developed MARKER-HF, a tool that computes a risk score between -1 and +1 to predict mortality of hospitalized and ambulatory HF patients as high risk or low risk mortality using ensemble learning. MARKER-HF is based on a machine learning model that is trained using AdaBoost which is a boosted decision tree algorithm that is implemented in the TMVA toolkit. They used eight features to train their models with data of 5822 patients taken from the A systems BIOlogy Study to TAilored Treatment in Chronic Heart Failure (BIOSTAT-CHF) project, University of California San Diego (UCSD) and San Francisco (UCSF) Medical Centers. MARKER-HF results showed its ability to predict mortality consistently in three different datasets. For mortality prediction with a 95% confidence interval for the area under the ROC curve, they achieved AUC=0.88 using UCSD, AUC=0.84 using UCSF, and AUC=0.81 using BIOSTAT-CHF.

Moreover, Angraal et al. [3] compared five different supervised learning classifiers not only for mortality prediction but also for hospitalization of HF outpatients with preserved ejection fraction (HFpEF) through three years of follow-up. They trained the following five methods: two LR, one with a forward selection and one with a lasso regularization for feature selection, gradient descent boosting, Random Forest (RF) and Support Victor machine (SVM). They used a total of 86 candidate features to train their models, including demographic, clinical, laboratory and electrocardiography data, and KCCQ scores obtained from the patients. These patients' data are taken from the Treatment of Preserved Cardiac Function Heart Failure with an Aldosterone Antagonist (TOPCAT) trail. The authors used 5-fold cross-validation to divide the learning set into five subsets, where 80% of them were used for training and 20% for testing. They experimentally proved that RF is the best classifier with 95% confidence interval achieved AUC=0.72, and Brier score=0.17 for mortality prediction; and AUC=0.76 and Brier score=0.19 for HF hospitalization prediction. Even more, they found that the best features to predict mortality are the body mass index, BUN levels, and KCCQ scores; where BUN levels, hemoglobin level (H) and KCCQ scores are the best features to predict HF hospitalization.

On the other hand, most of the heart diagnosis studies were focused on predicting the EHD [8–14]. Ananey-Obiri and Sarku [11] investigated the traditional supervised learning algorithms and data mining techniques for EHD prediction. They investigated the following three supervised algorithms: decision tree (DT), LR, Gaussian Naïve Bayes (NB). They experimented the well-known Cleveland heart disease datasets with all of its 13 features shown in Table II which are age, sex, Resting Blood pressure (RBP), Chest Pain Type (CPT), Serum Cholesterol (SCH), Fasting Blood Sugar (FBS), Maximum Heart Rate achieved (MHR), Resting Electrocardiographic Results (RES), Exercise Induced Angina (EIA), Peak Exercise Slope (PES), Old Peak (OPK), Thallium Scan (THA) and number of major Vessels Colored by Fluoroscopy (VCA). The tested dataset contained 287 observations out of 303 after the duplicated, missing values and outliers were removed as a pre-

processing step. In addition, they used feature normalization as a feature scaling technique, and single value decomposition (SVD) as feature extraction to reduce the number of features from 13 to 4. The data was labeled as absent or present of heart diseases. The authors used 10-fold cross-validation to divide the learning set into ten subsets, where nine of them were used for training and one of them for testing. The reported results were 79.31% accuracy and 0.81 AUC for DT model, 76% accuracy and 0.87 AUC for Gaussian NB model, and 82.75% accuracy and 0.86 AUC for LR model. Moreover, in the work conducted by Reddy et al. [12], ten different supervised and ensemble learning techniques were tested for EHD prediction. These techniques include NB, LR, Sequential minimal optimization (SMO), bootstrap aggregation, AdaBoost, JRip, RF, and KNN. Cleveland dataset was tested with 303 samples and 13 pre-mentioned features. Three different feature selection methods were used to enhance the performance, which are chi-squared, BestFirst search method and ReliefF. 11 out of 13 features were selected with the best performance result. As an evaluation scheme, 10-fold cross-validation was applied. The best result was 85.15% accuracy which was obtained using the SMO classifier with Chi-Squared feature selection technique. Even more, a hybrid learning approach was adapted by Abdeldjouad et al. [10], in which a hybrid approach of various machine learning methods was proposed to predict EHD. These methods include AdaBoostM1, LR, Fuzzy Unordered Rule Induction (FURIA), Multi-Objective Evolutionary Fuzzy Classifier (MOEFC), Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML), and Genetic Fuzzy System-LogitBoost (GFS-LB). They also experimented Cleveland database for training and assessing their methods using 10-fold cross-validation. Two models were built in which the first model used AdaBoostM1, LR, and MOEFC with 14 features, and reduced to 12 by removing the personal information (e.g. age and sex) with the wrapper feature selection method. While the second model used FURIA, GFS-LB, and FH-GBML and reduced the features to 6 with Principal Component Analysis (PCA) as a dimensionality reduction technique. The first model was selected using the majority voting as the final best-performed model with 80.20% accuracy. For conducting this work, the Keel tool was used for feature selection, while the Weka tool was used for feature extraction.

However, the EHD prediction studies [8–14] have some drawbacks. For example, using datasets without handling the imbalance classes [6, 10, 11], important features such as age, were manually excluded [6, 9, 10], no comparisons were made in terms of prediction methods [6] or the used dataset [6, 11, 12], various platforms were used for assessments [10], some important details were missing or not clear such as the number of selected features [7] or the number of samples per class [3, 8, 11]. Even more, surprisingly, no coloration exists between the selected features among these works [3, 6–10] even though some of them were using the same dataset. In other words, the significant factors that cause variance in recent proposed works' performance are still not fully investigated. To sum up, there is still a need to experimentally evaluate different classifiers with different data mining approaches to gain a final decision. Hence, this study's primary goal is to compare and analyze different classifiers and data mining techniques (feature selection and feature extraction) and their effect on improving the EHD prediction.

III. Methodology

This work handled the previously mentioned drawbacks, in which it explored numerous solutions including the usage of different sampling techniques to overcome imbalanced datasets issue represented in [6, 10] with different dimensionality reduction techniques. Moreover, all the important features were taken into consideration which were excluded manually by [6, 9, 10]. Furthermore, some of the previous works [3, 6, 7] used different sets of performance metrics which make the comparison quite difficult, so the proposed approach was assessed by considering a full set of well-known performance metrics with clarifies details regarding the selected features and the number of samples per class to tackle what was missing in [3, 7, 8]. In addition, different machine learning algorithms were selected as the best classifiers in various works according to their experimented data mining approaches. To summarize, there is still a need to experimentally evaluate different classifiers with different data mining approaches to gain a final decision.

Consequently, the methodology shown in Fig. 1 which consists of seven main stages, was designed to conduct this study to ensure the proposed solutions. The first stage is *data collection* where the Cleveland heart disease dataset (Clev) [15] was selected in Section III-A to be investigated in this work. While in the second stage, a *data pre-processing* is performed in Section III-B using some data balancing techniques to generate two more balanced versions of the dataset. In Section III-C, *dimensionality reduction* was done as the third stage using some data mining techniques such as *feature selection*, and *feature extraction* to reduce the number of features, improve the performance and avoid overfitting. In Section III-D the fifth stage, which is *evaluation scheme preparing* was detailed. Then, seven well-known *classification algorithms* are selected in Section III-E for the purpose of comparison. Lastly, in the seventh stage, the conducted *comparative experiments* were designed in Section III-F.

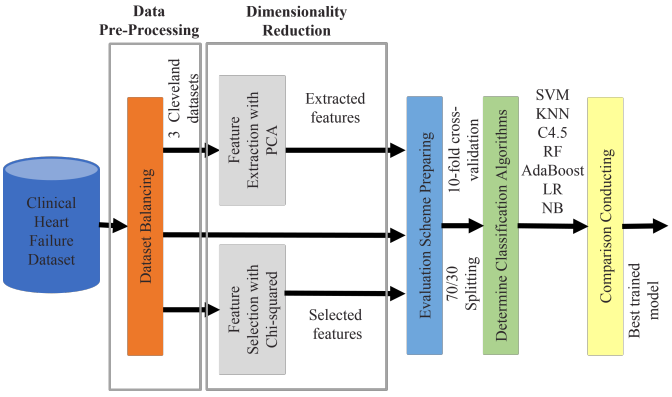

Fig. 1. The Proposed Methodology for EHD Prediction Comparison.

A. Dataset Collection

In this study, the well-known Clev heart disease dataset [15] was experimented since it is the most investigated dataset in this field by related works [8–11, 13, 14]. It is publicly available on an online machine learning and data mining repository of the University of California, Irvine (UCI). It contains

TABLE I. SUMMARY OF VARIOUS HF PREDICTION TECHNIQUES

Description	**Reference**	**Adler et al. [6]**	**Maragatham et al. [7]**	**Ananey-Obiri and Sarku [11]**	**Angraal et al. [3]**	**Reddy et al. [12]**	**Abdeldjouad et al. [10]**
	Year	2020	2019	2020	2020	2021	2020
	Main Goal	Mortality prediction of HF patients	HF prediction	EHD prediction	Mortality and hospitalization prediction of HF patients	EHD prediction	EHD prediction
Dataset	**Dataset Name**	UCSD	An arbitrary	Cleveland	TOPCAT	Cleveland	Cleveland
	# Samples	5822	4289	287	1,76	303	296
	Evaluation Scheme	50%|50% training|testing	6-fold cross-validation	10-fold cross-validation	5-fold cross-validation	10-fold cross-validation	10-fold cross-validation
Data Mining and Pre-Processing	**Technique**	- Exclusion of patients who had missing data, older than 80 years, with CIED device, died within 7 days of initial encounter, or had obvious medical record errors	Not clear	- Exclusion of patients who had missing or duplicated data - Feature scaling using normalization - Feature extraction using SVD	- Exclusion of features with 50% missing data - Feature selection using forward selection and a lasso regularization	- Feature selection using chi-squared, BestFirst search method and ReliefF	- Exclusion of patients who had missing data - Exclusion of personal features - Feature selection using a wrapper method - Feature extraction using PCA
	#Features	8	Not clear	4 out of 13	86	11 out of 13	13
	Features Names	Cr, RBP, H,BUN, platelets, WBC, RDW and albumin	Health info, tobacco usage, demographics and liquor consumption and lab test	Age, sex, RBP, CPT, SCH, FBS, MHR, RES, EIA, PES, OPK, THA and VCA	Demographic, clinical, laboratory and electrocardiography, and KCCQ scores features	Age, sex, RBP, CPT, SCH, FBS, MHR, RES, EIA, PES, OPK, THA and VCA	RBP, CPT, SCH, FBS, MHR, RES, EIA, PES, OPK, THA and VCA
Methodology	**Learning Category**	Ensemble learning	Deep learning	Supervised learning	Supervised and ensemble learning	Supervised and ensemble learning	Hybrid learning
	Prediction Methods	AdaBoost	LSTM	LR	LR with a forward selection, LR with a lasso regularization, RF, gradient descent boosting and SVM	SMO	A new hybrid approach of LR, AdaBoostM1, MOEFC, FURIA, GFS-LB and FH-GBML
Performance Evaluation	**Metrics**	ROC charts, and AUC	ROC charts, and AUC	CM, ACC, P, SE, F-score, ROC charts, and AUC	AUC, and Brier scores	ACC, MAE, SE, fallout, P, F-Score, SP, and ROC area	SE, SP, ACC, ER
	Results	AUC= 0.88	AUC= 0.894	ACC=82.75%, and AUC=0.86	Mortality prediction (AUC= 0.7, Brier score= 0.17) HF hospitalization prediction (AUC= 0.76, Brier score= 0.19)	ACC= 86.468%	ACC= 80.20%
Drawbacks	**In terms of** datasets, features, platforms, algorithms, and comparisons	- Imbalanced datasets - Exclusion of elderly patients above 80 years - No comparisons are made in terms of prediction methods	- Some important details are missing or not clear	- No. samples per class were not mentioned - Imbalanced datasets	- No. samples per class were not mentioned - No comparisons are made in terms of datasets - Dataset with missing data	- No comparisons are made in terms of the dataset	- Personal information is excluded manually (e.g age, sex) - Using various platforms - No comparisons are made in terms of the dataset - Imbalanced datasets

The evaluation metrics are P: Precision, CM: Confusion matrix, AUC: Area Under Curve, ROC: Receiver Operating Characteristic curve, ACC: Accuracy, SP: Specitivity, SE: Sensitivity, ER: Error Rate, MAE: Mean Absolute Error.
The features are RBP: Resting Blood pressure, CPT: Chest Pain Type, FBS: Fasting Blood Sugar, SCH: Serum Cholesterol, MHR: Maximum Heart Rate achieved, RES: Resting Electrocardiographic Results, EIA: Exercise Induced Angina, PES: Peak Exercise Slope, OPK: Old Peak, THA: Thallium Scan, VCA: Number of Major Vessels Colored by Fluoroscopy.

303 medical records of 165 patients with heart diseases and 138 are healthy. Moreover, it has 74 features, but 13 common features have been studied in the state of the art. Table II lists these features, their types and descriptions. The Clev originally contained five categorical classes (0, 1, 2, 3 and 4) where 0 refers to the absence of heart disease while the other four classes (1, 2, 3 and 4) refer to the presence of different heart diseases. However, most of the works [8–14] that used this dataset transferred these five categorical classes to a binary class for the purpose of distinguishing simplicity. The “class" field denotes the existence of heart disease in the patient. In which 0 means absence of heart disease (normal) and 1 means existence of heart disease.

B. Dataset Balancing

In the medical field, according to [10], the diagnosis of diseases is easier and quicker if data is balanced. The used Clev dataset contains 303 observations which present quite balanced positive and negative samples, with 165 and 138 observations respectively. But, machine learning techniques are very sensitive and the created prediction models are usually biased towards the larger class. However, the previous works [6, 10, 11] ignored this fact. Hence, to create an unbiased prediction model, reduce the gap between the two classes, and ensure equivalent balancing, dataset pre-processing is needed [9]. In this study, the class-imbalance problem in the original Cleveland dataset was solved by using oversampling and un-

TABLE II. DETAILED DESCRIPTION OF CLEVELAND DATASET'S FEATURES

#	Feature Name	Shortcut	Feature Type	Description
F1	Age	Age	Continuous	Age of the patient [years]
F2	Sex	Sex	Discrete	Sex of the patient [M: Male, F: Female]
F3	ChestPainType	CPT	Discrete	Chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
F4	RestingBP	RBP	Continuous	Blood pressure at rest [mm Hg]
F5	Cholesterol	SCH	Continuous	Serum cholesterol [mm/dl]
F6	FastingBS	FBS	Discrete	Fasting blood sugar [1: if FastingBS >120 mg/dl, 0: otherwise]
F7	RestingECG	RES	Discrete	Resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
F8	MaxHR	MHR	Continuous	Maximum heart rate achieved [Numeric value between 60 and 202]
F9	ExerciseAngina	EIA	Discrete	Exercise-induced angina [Y: Yes, N: No]
F10	Oldpeak	OPK	Continuous	Oldpeak = ST depression induced by exercise relative to rest [Numeric value measured in depression]
F11	ST_Slope	PES	Discrete	The slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
F12	MajorVessels	VCA	Continuous	The number of major vessels colored by flourosopy [0-3]
F13	ThalliumScan	THA	Discrete	Type of defect [3 = normal; 6 = fixed defect; 7 = reversible defect]

dersampling techniques to create two balanced versions of the dataset. The first version is (*Oversampled Cleveland*) which was generated using the Synthetic Minority Oversampling Technique (SMOTE) [16] which is popularly used in the medical field to deal with class imbalanced data. SMOTE adds more samples to the smaller class observations by generating random synthetic ones from its nearest neighbors using the Euclidean distance [17]. SMOTE increased the original Cleveland dataset from 303 to 330 observations (165 for each class). While the second version is (*Undersampled Cleveland*) which was simply generated by randomly reducing the number of larger class observations. Undersampling reduced the original Cleveland dataset from 303 to 276 observations (138 for each class).

C. Dimensionality Reduction

It decreases the number of input features (dimensions) of the original problem using specific techniques to improve the learning performance. These techniques are categorized as *feature selection* and *feature extraction*. The key difference between them is that the feature selection selects a subset of features from the original features, while the feature extraction uses the original features to create a new set of features [18]. In this study, one technique from each of these categories was experimented in the following illustration:

1) Feature Selection: it is the process of selecting a group of relevant features from the original features according to specific criteria [10]. Some of its main goals are: 1) reduce the algorithm's computational time, 2) identify the relevant features, 3) improve the prediction performance, and 4) avoid the overfitting by limiting the number of selected features; because the overfitting could affect the model to loss its robustness when the model is used to test new unseen data [6, 19]. In this study, the Chi-squared [20] was used as a feature selection technique to determine the most relevant features. Chi-squared was selected among other feature selection techniques since it improves most of the classifiers' performances and achieves remarkable results in this field [12, 14, 20]. This study starts with the 13 most common features and end up with only seven features after applying Chi-squared. The seven selected features are PES, EIA, CPT, MHR, THA, VCA, and OPK.

2) Feature Extraction: as one of the dimensionality reduction methods, it reduces the number of dataset's features in which the reduced features are represented by a set of new features [10]. The main goal of this technique is to use fewer features, which results in a simpler model that may have better performance with new unseen data. In this study, the principal component analysis (PCA) was used. The reason behind selecting PCA is because it is considered as one of the most famous dimensionality reduction and feature (components) extraction techniques for the medical applications [10]. PCA works by creating novel factors that have the best valuable information by capturing the highest variance of these features [21]. Using PCA, 2, 8 and 8 new sets of features were extracted for the original Clev, Oversampled Clev, and Undersampled Clev, respectively.

D. Evaluation Scheme Preparing

All the three Cleveland versions were evaluated by two schemes which are 10-fold cross-validation and 70%|30% for training|testing data splitting.

E. Classification Algorithms Selection

Machine learning and classification techniques are a group of computational models that can be used to solve many kinds of problems easily. Various applications of computational intelligence exist in the pathology and medicine field [22, 23]. This work, compared the following seven classifiers: SVM [24], KNN [25], C4.5 [26], RF [27], AdaBoost [28], NB [29] and LR [30]. SVM is a supervised learning technique that demonstrates superb performance in the medical field [31]. It depends on kernel functions that transfer all instances to a upper dimensional space intending to find a linear decision boundary for data partitioning [24]. KNN is a simple but effective method for classification [25]. C4.5 decision tree was selected due to its low complexity in implementation and excellent explanation [26]. In addition, a decision tree was investigated as the main classifier in several EHD prediction research. RF decision tree [27], is one of the popular techniques for pattern recognition which has been efficiently applied as a strong and widespread tool for predicting and classifying medical data. NB [29] is based on Bayes' Theorem with an assumption of independence among predictors.

In this study, SVM was implemented via the LibSVM library using nu-SVC as SVM type and linear kernel [24]. KNN was implemented using IBk library where the number of neighbors K to inspect equals 1. C4.5 decision tree algorithm was implemented using the J48 classifier (C4.5 release 8

implemented with Java) [26]. AdaBoost [28] is a short term of Adaptive Boosting and was implemented using AdaBoostM1.

F. Comparison Conducting and Performance Measurements

The comparison experiments were conducted based on the three versions of the Clev dataset (original, oversampled, and undersampled Clev), each with three copies: 1) without dimensionality reduction, 1) after applying Chi-squared, 3) after applying PCA; which makes them a total of nine datasets to be experimented. Each dataset was used to build training models using the seven classifiers which are SVM [24], KNN [25], C4.5 [26], RF [27], AdaBoost [28], NB [29] and LR [30]. So, a total of 63 models (9 datasets * 7 classifiers) were experimented. Each model was tested using two evaluation schemes; 10-fold cross-validation and 70%|30% splitting. Hence, the entire experiment ended up with a total of 126 trails (63 training models * 2 evaluation schemes).

The building training models were evaluated via six measures, which are Matthews correlation coefficient (MCC) [8], accuracy (ACC), recall/sensitivity (SE), precision, F-Score (FM) [11], specificity (SP), error rate (ER) [10], and area under the curve (AUC) [7]. MCC is a measure that is frequently utilized for assessing the quality of binary classification. It ranged from -1 to 1, in which 1 indicates an excellent prediction, 0 means the classification is no better than a random prediction, and -1 indicates full disagreement between prediction and observation. The Receiver Operating Characteristic (ROC) chart is used for additional investigation, where it consists of two rates, the true positive rate (TPR) versus the false positive rate (FPR) for various thresholds. The best ROC is the chart with more area under the curve (AUC). AUC equal to 1 presents the ideal ROC which means that the model can perform with 100% sensitivity and 100% specificity [8]. ACC refers to the fraction of accurately classified samples. SE is the fraction of correctly classified heart disease patients. It is also known as True Positive Rate (TPR) or recall, while SP is the correctly classified healthy subjects. These measures are formulated as the following[2]:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall/\ Sensitivity\ (SE) = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity\ (SP) = \frac{TN}{TN + FP} \tag{3}$$

$$F\text{-}Score = \frac{2TP}{2TP + FP + FN} \tag{4}$$

$$Accuracy\ (ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Error\ Rate\ (ER) = \frac{FP + FN}{P + N} \tag{6}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

[2]Confusion Matrix: https://en.wikipedia.org/wiki/Confusionmatrix

All the implemented experiments including oversampling, feature extraction, feature selection, and performance evaluation were done using various libraries in WEKA version 3.8.5 [32]. The oversampling was implemented using the *SMOTE* technique that was proposed by [16] under the supervised filters where the nearest neighbors parameter was set to 5. While feature extraction was implemented using *PrincipalComponents* as attribute evaluator and *Ranker* as the search method. On the other hand, feature selection was implemented using *ChiSquaredAttributeEval* as an attribute evaluator and *Ranker* as a search method as well.

IV. Results and Discussion

Tables III and IV summarize the splitting and cross-validation results respectively. These results only present the best performed classifiers of the nine datasets that have been obtained from 126 trials. While Fig. 2 and 3 present the splitting and cross-validation results in terms of accuracy.

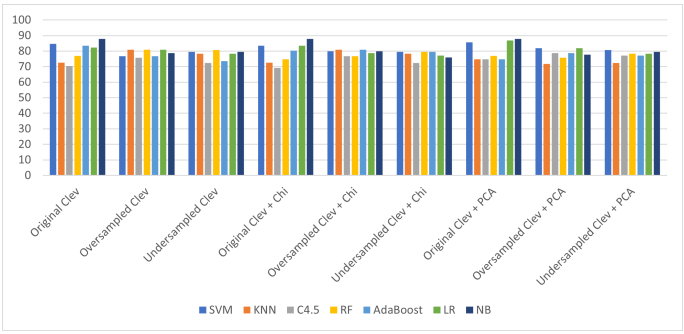

Fig. 2. The Accuracy of the Conducted Trails using 70%|30% Splitting.

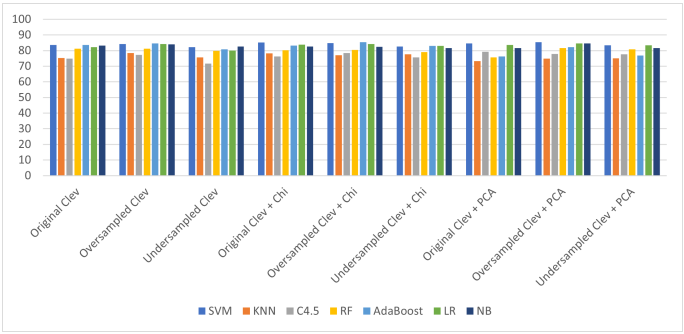

Fig. 3. The Accuracy of the Conducted Trails using 10-Fold Cross-Validation.

For *Original Clev dataset*, it can be noticed from the obtained accuracy in Fig. 2 of the splitting scheme that there are no significant differences between all the three versions of it (without processing, with Chi-squared, and with PCA) even though they contain different number of features. In fact, NB is the best performing classifier with ACC=87.91% and AUC=0.93, which is also summarized in Table III. C4.5 is the worst classifier, even when the number of features are changed. On the other hand, in terms of accuracy, when the cross-validation was used as shown in Fig. 3, SVM and AdaBoost were the best performing classifiers using all the 13 features with the same 83.5% accuracy. But SVM

TABLE III. THE SUMMARIZED RESULTS USING THE 70%|30% SPLITTING AS EVALUATION SCHEME

Exp #	Dataset	Data Mining Tech.	# of Features	Best Classifier	SP	SE/Recall	Precision	ACC	ER	FM	MCC	AUC
1	Original Clev	Without	13	NB	0.86	0.90	0.84	**87.91**	0.12	0.87	0.76	0.93
2		Chi-squared	7	NB	0.86	0.90	0.84	**87.91**	0.12	0.87	0.76	0.93
3		PCA	2	NB	0.86	0.90	0.84	**87.91**	0.12	0.87	0.76	0.93
4	Oversampled Clev	Without	13	LR RF	0.73	0.90	0.75	80.81	0.19	0.82	0.63	0.89
5		Chi-squared	7	AdaBoost	0.76	0.85	0.77	80.81	0.19	0.81	0.62	0.88
6		PCA	8	LR	0.73	0.92	0.76	**81.82**	0.18	0.83	0.65	0.88
7	Undersampled Clev	Without	13	RF	0.71	0.90	0.76	**80.72**	0.19	0.82	0.63	0.85
8		Chi-squared	7	RF	0.76	0.83	0.77	79.52	0.20	0.80	0.59	0.84
9		PCA	8	SVM	0.79	0.83	0.79	**80.72**	0.19	0.81	0.62	0.81

TABLE IV. THE SUMMARIZED RESULTS USING THE 10-FOLD CROSS-VALIDATION AS EVALUATION SCHEME

Exp. #	Dataset	Data Mining Tech.	# of Features	Best Classifier	SP	SE/Recall	Precision	ACC	ER	FM	MCC	AUC
1	Original Clev	Without	13	AdaBoost	0.78	0.88	0.83	83.50	0.17	0.85	0.67	0.88
2		Chi-squared	7	SVM	0.77	0.92	0.83	**85.15**	0.15	0.87	0.70	0.84
3		PCA	2	SVM	0.83	0.86	0.86	84.49	0.16	0.86	0.69	0.84
4	Oversampled Clev	Without	13	AdaBoost	0.85	0.84	0.85	84.55	0.15	0.84	0.69	0.85
5		Chi-squared	7	AdaBoost	0.85	0.86	0.85	**85.45**	0.15	0.86	0.71	0.92
6		PCA	8	SVM	0.85	0.85	0.85	**85.45**	0.15	0.85	0.71	0.85
7	Undersampled Clev	Without	13	NB	0.80	0.85	0.81	82.61	0.17	0.83	0.65	0.89
8		Chi-squared	7	AdaBoost	0.81	0.85	0.82	82.97	0.17	0.83	0.66	0.89
9		PCA	8	LR	0.82	0.85	0.82	**83.33**	0.17	0.84	0.67	0.89

TABLE V. THE MOST RECENT WORKS THAT STUDIED THE CLEVELAND DATASET

Ref.	Year	# of Features	Applied Technique	Classification Methods	Evaluation Scheme	ACC	AUC	MCC
[13]	2020	14	Data pre-processing	Artificial Neural Network (ANN)*, LR, SVM, KNN, NB and RF	10-fold cross-validation	85.86	-	-
[14]	2019	9	Feature selection	Majority voting of the weak classifiers	Splitting	85.48	-	-
[11]	2020	4	Data pre-processing, feature scaling, feature extraction, and outlier detection	Gaussian NB*, DT, and LR	10-fold cross-validation	82.75	0.87	-
[33]	2020	10	Data pre-processing, feature scaling, and data reduction	RF*, SVM, KNN, DT and LR	Splitting	85.71	0.87	-
Ours	2021	7	Data pre-processing, feature selection, feature scaling, and feature extraction	NB*, SVM, KNN, C4.5, RF, AdaBoost and LR	Splitting	**87.91**	**0.93**	**0.80**

* mean the classifier that gave the best performance in terms of accuracy.

outperformed the others when the features are reduced to 7 and 2 using Chi-squared (ACC=85.15% and AUC=0.84) and PCA (ACC=84.49% AUC=0.84).

For *Oversampled Clev dataset*, when the SMOTE technique is used to balance the data, the accuracy of the splitting scheme in Fig. 2 was significantly changed when the processing technique was changed (# of features) and with the evaluation scheme being changed. But the LR was the best classifier with ACC=81.82% when only the 8 features obtained from PCA were used. C4.5 and KNN gave the worst accuracy. However, in terms of AUCs, the RF and LR always outperformed other classifiers, even when the number of features being changed by 0.88 AUC for both classifiers. In addition, in terms of accuracy using cross-validation shown in Fig. 3, C4.5 and KNN again gave the worst accuracy. Where AdaBoost dominates the other classifiers by ACC=85.45%, when the original 8 or 7 features from Chi-squared and PCA were used, respectively. Moreover, in terms of AUCs, the best classifiers vary when the number of features being changed, but the worse classifiers were always C4.5 and KNN.

For *Undersampled Clev dataset*, Fig. 2 and 3, show that when the original Clev observations were reduced aiming for balancing, the obtained results were decreased compared with the above versions of the dataset. However, with the splitting scheme, it achieved 80.72% accuracy when using the 13 features with RF, and also when using the 8 features from PCA with SVM. Where with cross-validation scheme, it achieved 82.97% accuracy when using the 7 features from Chi-squared with AdaBoost. Furthermore, the LR and NB are the best classifiers in terms of AUCs, whatever the used features. Even with the splitting scheme, it achieved AUC=0.85 when using the 13 features with RF. Yet, with the cross-validation scheme, it achieved AUC=0.85 with whatever the used features.

According to the obtained results from the above experiments, the research questions were answered as follows: 1) seven features, which are PES, EIA, CPT, MHR, THA, VCA, and OPK, are the best performing features for predicting EHD, which were obtained from the Chi-squared feature selection technique; 2) it is noticeable that the best prediction performance ACC=87.91% and AUC=0.93 can be obtained whither the original features, the selected features, or the extracted features were used. Hence, the data mining techniques did not improve the prediction performance in terms of accuracy, however, they improved it in terms of reducing the number of features which lead to more computational efficiency; and 3) NB proves that it is the most appropriate classifier to be used with whatever feature sets to predict the EHD.

Furthermore, the availability of the Clev dataset allows many researchers to test their prediction models. For that reason, this work was compared to the recent related studies that used Clev dataset [11, 13, 14, 33]. Table V summarises their methods and obtained results along with this work. It is noticeable that these studies lack some important performance metrics such as MCC which play an important role in

reporting balance or imbalanced data. Moreover, the effect of using different applied techniques such as data pre-processing, feature selection, and feature extraction techniques were also compared. This comparison proved that this work outperforms those studies by achieving higher performance with a margin of 2.20%.

V. Conclusion and Future Work

EHD prediction is a field where researchers propose new techniques that hopefully can facilitate the diagnosis of the existence of heart diseases and enhance the decision-making operations of physicians. In this work, an experimental comparison was conducted between seven famous classifiers, which are SVM, KNN, C4.5, RF, AdaBoost, NB, and LR with different data mining techniques including feature extraction, and feature selection. This work utilized a famous heart disease dataset called Cleveland, aiming to undergo a deeper investigation of the effective techniques that could improve EHD prediction. A methodology of seven basic stages was proposed to conduct this study, including data collection, data pre-processing, and balancing techniques (oversampling and undersampling). Dimensionality reduction such as Chi-squared feature selection and PCA feature extraction techniques were also investigated. The main concern of this paper is not only enhancing the accuracy of weak classifiers but also investigating the famous Clev dataset closely and studying the influence of different pre-processing techniques, in addition, to determining the number of best features that work better with Clev. However, this work, like other Cleveland dataset-based works [11, 13, 14, 33], suffers from the limited number of observations that could be handled in the future works by merging it with another EHD prediction dataset. Moreover, this study could be extended by exploring more data pre-processing techniques such as outlier detection, applying deep learning techniques, and tuning the hyperparameters.

Acknowledgment

We gratefully thank Dr. Imane Boudellioua for her valuable suggestions and guidance, sharing her encouragement and useful feedback in completing this work.

References

[1] N. M. Aljefree, I. M. Shatwan, and N. M. Almoraie, "Association between nutrients intake and coronary heart disease among adults in saudi arabia: A case-control study," *PROGRESS IN NUTRITION*, vol. 23, no. 3, 2021.

[2] Heart disease. Accessed 19-October-2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[3] S. Angraal, B. J. Mortazavi, A. Gupta, R. Khera, T. Ahmad, N. R. Desai, D. L. Jacoby, F. A. Masoudi, J. A. Spertus, and H. M. Krumholz, "Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction," *JACC: Heart Failure*, vol. 8, no. 1, pp. 12–21, 2020.

[4] G. A. Diamond and J. S. Forrester, "Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease," *New England Journal of Medicine*, vol. 300, no. 24, pp. 1350–1358, 1979.

[5] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, 1998.

[6] E. D. Adler, A. A. Voors, L. Klein, F. Macheret, O. O. Braun, M. A. Urey, W. Zhu, I. Sama, M. Tadel, C. Campagnari *et al.*, "Improving risk prediction in heart failure using machine learning," *European journal of heart failure*, vol. 22, no. 1, pp. 139–147, 2020.

[7] G. Maragatham and S. Devi, "Lstm model for prediction of heart failure in big data," vol. 43, no. 5, 2019.

[8] L. Ali, A. Niamat, J. A. Khan, N. A. Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. C. Bukhari, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, pp. 54 007–54 014, 2019.

[9] L. Ali and S. Bukhari, "An approach based on mutually informed neural networks to optimize the generalization capabilities of decision support systems developed for heart failure prediction," *Irbm*, 2020.

[10] F. Z. Abdeldjouad, M. Brahami, and N. Matta, "A hybrid approach for heart disease diagnosis and prediction using machine learning techniques," in *International conference on smart homes and health telematics*. Springer, 2020, pp. 299–306.

[11] D. Ananey-Obiri and E. Sarku, "Predicting the presence of heart diseases using comparative data mining and machine learning algorithms," *International Journal of Computer Applications*, vol. 176, pp. 17–21, 2020.

[12] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "Heart disease risk prediction using machine learning classifiers with attribute evaluators," *Applied Sciences*, vol. 11, no. 18, p. 8352, 2021.

[13] I. Tougui, A. Jilbab, and J. El Mhamdi, "Heart disease classification using data mining tools and machine learning techniques," *Health and Technology*, vol. 10, pp. 1137–1144, 2020.

[14] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked*, vol. 16, p. 100203, 2019.

[15] M. Lichman. (2013) UCI machine learning repository. [Online]. Available: http://archive.ics.uci.edu/ml

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[17] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39 707–39 716, 2021.

[18] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in biology and medicine*, vol. 112, p. 103375, 2019.

[19] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, B. J. Chow, and G. Dwivedi, "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death," *PloS one*, vol. 14, no. 6, p. e0218760, 2019.

[20] L. Ali, A. Rahman, A. Khan, M. Zhou, A. Javeed, and J. A. Khan, "An automated diagnostic system for heart disease prediction based on x^2 statistical model and optimally configured deep neural network," *IEEE Access*, vol. 7, pp. 34 938–34 945, 2019.

[21] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and pca," *Informatics in Medicine Unlocked*, vol. 19, p. 100330, 2020.

[22] A. H. Shahid and M. Singh, "Computational intelligence techniques for medical diagnosis and prognosis: Problems and current developments," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 3, pp. 638–672, 2019.

[23] H. R. Tizhoosh and L. Pantanowitz, "Artificial intelligence and digital pathology: challenges and opportunities," *Journal of pathology informatics*, vol. 9, 2018.

[24] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[25] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37–66, 1991.

[26] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," 1994.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[29] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," *arXiv preprint arXiv:1302.4964*, 2013.

[30] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 41, no. 1, pp. 191–201, 1992.

[31] J. Zhi, J. Sun, Z. Wang, and W. Ding, "Support vector machine classifier for prediction of the metastasis of colorectal cancer," *International journal of molecular medicine*, vol. 41, no. 3, pp. 1419–1426, 2018.

[32] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[33] N. K. Kumar, G. S. Sindhu, D. K. Prashanthi, and A. S. Sulthana, "Analysis and prediction of cardio vascular disease using machine learning classifiers," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. IEEE, 2020, pp. 15–21.