Exploring Machine Learning Techniques for Coronary Heart Disease Prediction

Hisham Khdair[1], Naga M Dasari[2]
International Institute of Business and Information Technology,
Federation University Associate
Adelaide, Australia

***Abstract*—Coronary Heart Disease (CHD) is one of the leading causes of death nowadays. Prediction of the disease at an early stage is crucial for many health care providers to protect their patients and save lives and costly hospitalization resources. The use of machine learning in the prediction of serious disease events using routine medical records has been successful in recent years. In this paper, a comparative analysis of different machine learning techniques that can accurately predict the occurrence of CHD events from clinical data was performed. Four machine learning classifiers, namely Logistic Regression, Support Vector Machine (SVM), K- Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP) Neural Networks were identified and applied to a dataset of 462 medical instances and 9 features as well as the class feature from the South African Heart Disease data retrieved from the KEEL repository. The dataset consists of 302 records of healthy patients and 160 records of patients who suffer from CHD. In order to handle the imbalanced classification problem, the K-means algorithm along with Synthetic Minority Oversampling TEchnique (SMOTE) was used in this study. The empirical results of applying the four machine learning classifiers on the oversampled dataset have been very promising. The results reported using different evaluation metrics showed that SVM has achieved the highest overall prediction performance.**

***Keywords*—Coronary heart disease; machine learning; prediction; classification**

I. Introduction

Heart disease refers to a wide range of conditions that affect the structure and function of the heart. CHD is one of the most common types of heart disease, and it is one of the leading causes of death around the world. CHD occurs when plaque builds up in the walls of the coronary arteries, it restricts blood flow to the heart muscle, and will eventually result in a heart attack. According to the Australian Institute of Health and Welfare (AIHW), CHD was Australia's leading cause of death in 2018, accounting for 17,500 deaths. This accounts for 11% of all deaths in Australia and 42% of all cardiovascular deaths [1].

Traditional risk factors for CHD are thought to be High-LDL cholesterol, low-HDL cholesterol, high blood pressure, diabetes mellitus, smoking, a family history of CHD, age, obesity, and an unhealthy lifestyle [2]. The estimated cost of CHD in 2015–16 in Australia was more than $2.2 billion. Private hospital services and public hospital admitted patient services accounted for a minimum cost of $813 million and $693 million, respectively. The burden on the Pharmaceutical Benefits Scheme (an Australian Government subsidy on medicine) for CHD was estimated to be around $218 million [1].

CHD can however be effectively managed with a change in lifestyle and adopting healthy habits, and hence save the high cost of medical treatment and hospitalization if early detected. With early detection of CHD, patients can have a range of treatments advised by doctors to reduce the risk of future heart problems and relieve or manage symptoms. In this context, electronic health records (EHRs, also called medical records (EMRs)) can be considered a useful resource of information to help medical practitioners in the detection or the prediction of CHD [3-6].

Advances in machine learning and artificial intelligence have motivated many scientists to use such technologies in the early detection of high-risk diseases such as heart diseases, diabetes, various types of cancer [7-9]. Machine learning applied to EHR can be a useful tool for predicting the CHD event with heart disease symptoms [10-12] as well as exploring the most significant clinical features and risk factors that may lead to heart attack and deaths. Clinicians and physicians can take advantage of machine learning for clinical feature ranking and unveil hidden and non-obvious correlations and relationships between patients' data. Several supervised machine learning classifiers were used for this purpose and have achieved success in this regard such as logistic regression, SVM, deep learning, KNN, decision tree [3, 13-15].

However, most of the machine learning models designed for the prediction of CHD have achieved modest accuracy [16], More recent models show some improvements but only in the prediction accuracy though [17, 18]. Moreover, the predicting variables of these models have limited interpretability [5, 12]. Even though scientists have identified a large number of predictors and indicators, there is still no consensus on such clinical features and their roles in affecting the occurrence of CHD [2, 19].

In this paper, we study a dataset of 462 medical records obtained from South African Heart Disease. The dataset is a quantitative sample of males in a heart-disease high-risk region of the Western Cape in South Africa-KEEL[20]. The objectives of this study are:

- To investigate machine learning techniques that achieve a high prediction performance in predicting CHD.
- To identify the most effective machine learning models that achieve the best prediction performance on the given dataset.
- And, to identify the best features that help in achieving

the best performance on the given dataset.

In this study, the machine learning classifiers that have been identified and utilized are Logistic Regression, SVM classifier, KNN, and MLP Neural Network. They were identified based on the literature as well as their performance on the given dataset and the suitability of the nature of the available data. The structure of the paper is as follows, we first discuss the related work in section 2, then we discuss the methodology in section 3, where we describe in detail the dataset, the exploratory data analysis, and the feature selection methods. The experimental framework is presented in section 4, results, discussion, and conclusion are discussed in sections 5 and 6 respectively.

II. Related Work

Several researchers have performed studies on routine clinical data (or EHR) obtained from primary health care centers or family practices to predict the occurrence of heart disease [3, 21-24, 46]. Electronic health records have been used or in combination with several machine learning algorithms to predict CHD [25, 26]. Machine learning algorithms have proved to be efficient techniques in predicting heart diseases [3, 18, 27, 47].

In a study performed on 378,256 instances of patient data obtained from UK family practices, the authors in [3] have used the machine learning algorithms logistic regression, random forest, gradient boosting machine, and neural networks. The authors have established that the algorithms have improved the prediction of heart disease, CHD. Improvements in accuracy according to AUC c-statistic are random forest +1.7%, logistic regression +3.2%, gradient boosting +3.3%, neural networks +3.6% when compared to baseline American Association of Cardiology (ACA) and American Heart Association method. In another study, the researchers in an experimental analysis [22] have applied several machine learning algorithms; Decision Tree, Naïve Bayes, K-nearest neighbors, SVM, Multi-Layer perceptron, radial basis function, and Single Conjunctive Rule Learner individually and in combination on the Cleveland dataset [28] which is available at University of California Irvine (UCI) machine learning repository. The authors have compared the algorithms using Precision, Recall, F-Measure, ROC, and accuracy. Support vector machine has provided the best results in the experiment with 84.15% accuracy and 0.897 F-measure. They have applied bagging, boosting and stacking methods to improve the results.

In recent work, the Cleveland and Statlog [29] datasets are further experimented with, by another set of researchers with impressive results [17]. Statlog dataset contains 270 samples with 150 absence of heart disease and 120 presence of it. Cleveland database contains 303 instances with 164 without heart disease and 139 positive cases. The algorithms used in this work were support vector machine, Logistic Regression, Naïve Bayes, deep neural networks, random forest, decision tree, and k-nearest neighbor. Their experiment results have shown that deep neural networks work better for Statlog database whereas SVM works better for the Cleveland database. However, the accuracy in both cases is very high, a fraction above 97%, which is signficantly high when compared to any other study. While the reported accuracy was very high at 97%, the other metrics such as precision, recall and specificity were not investigated which are important to measure the efficiency of a machine learning algorithm.

In experimenting with ensemble machine learning algorithms the authors in [27] have used 4 different datasets obtained from Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC), Long Beach Medical Center (LBMC), and Switzerland University Hospital (SUH) to predict CHD. The datasets contain 303, 294, 200, and 123 instances, respectively. All the patient instances were formatted uniformly with 76 attributes each out of which only 29 were used due to missing values. Adaptive boosting algorithm has been used for training and prediction. The experimental results produced accuracy and F-score for the different datasets in the order CCF – 80.14, 0.76; HIC- 89.12, 0.83; LBMC-77.78, 0.87; and SUH-96.72, 0.98.

In an experiment on deep learning, Baccouche et al. [30] have worked on heart disease data consisting of 900 samples with 149 attributes each, out of which 16% are related CHD instances. The data was obtained from Medica Norte Hospital, a Mexican hospital in Mexico. The authors have proposed an ensemble neural network framework with Bidirectional Long-Short Term Memory (BiLSTM) or Bidirectional Gated Recurrent Unit (BiGRU) with a CNN model with an accuracy rate of 91%.

Working on the dataset we are working on, Gonsalves et al. [16] performed experimental analysis using Decision Tree, Naïve Bayes, and support vector machine algorithms on WEKA tool. The accuracies obtained for all the three algorithms are above 70% with Naïve Bayes showing the highest with 71.5%. They have attributed the low accuracy to the small size of the dataset and the class imbalance problem in the dataset.

Based on the literature it is noticed that machine learning techniques such SVM, KNN, MLP Neural Networks, decision tree and boosting algorithms are widely used for predicting coronoary heart disease.

III. Methodology

We present the dataset we used for the experiment, exploratory data analysis, and feature selection methods used in this section.

A. Dataset

The dataset for this study has been retrieved from South African Heart Disease [20], which is a subset of a wider dataset. It has a total of 462 medical observations (instances) and 10 features, 9 as independent clinical features, and 1 is the target variable, a labeled binary class as 0 or 1, i.e., CHD event has been detected for the medical observations as positive or negative. The data is for a group of men from a high-risk area for heart disease in South Africa.

Each high-risk patient was monitored in the dataset and the features retrieved were as follows: systolic blood pressure (Sbp), cumulative tobacco in kg (Tobacco), bad cholesterol also known as low-density lipoprotein cholesterol (Ldl), adiposity, family history of heart disease (Famhist), type-A behavior (TypeA), Obesity, current alcohol consumption (Alcohol),

TABLE I. DESCRIPTION OF FEATURES IN THE DATASET

Feature	Explanation	Type and Range	Null Values
Systolic Blood Pressure	Blood pressure measure against the artery walls as the heart beats	Numerical [101, 218]	no
Tobacco	Accumulative tobacco in the body in (kg)	Numerical [0.0, 31.2]	no
LDL Cholesterol	low-density lipoprotein, also called bad cholesterol	Numerical [0.98, 15.33]	no
Adiposity	Adiposity is a measure of percentage of body fat	Numerical [6.74, 42.49]	no
Family History	Family history of heart disease	Binary [0, 1]	no
Type A Behavior	Type A behavior and personality	Numerical [13, 78]	no
Obesity	Weight-to-height ration measure (body mass index, bmi)	Numerical [0.0, 147.19]	no
Alcohol	Current alcohol consumption	Numerical [15, 64]	no
Age	Age of the patient	Numerical [15, 64]	no
CHD Event target	If Coronary heart disease was detected	Binary 0, 1	no

age at onset (Age), and coronary heart disease (Chd) (yes=1 or no=0).

B. Data pre-processing

The original dataset was in .dat format, we have converted it to .csv, and we edited the name of the columns to be more expressive. We have encoded the existing categorical text values in the original dataset into numerical values to be able to be fitted into machine learning models. The description of the features is shown in Table I.

C. Exploratory Data Analysis

The statistical quantitative characteristics of the dataset for numerical features are described in Table II. It can be noticed that the measurements for LDL Cholesterol, Obesity, and also Type A Behavior has slight differences in the mean value for patients with positive CHD event and negative event. The visualization of the counts of observations of the Family History binary class with respect to the negative CHD events and positive CHD events, as well as frequency of the target class CHD Event in the dataset, are shown in Fig. 1 and Fig.2 respectively. Out of 302 subjects without heart disease, 206 of them do not have CHD in the family history whereas 96 have the family history. For positive cases, 64 of them do not have CHD in the family history and 96 have CHD in the family history.

Fig. 3 presents the distribution of each feature's data based on CHD events with the minimum value, first quartile (Q1), median, third quartile (Q3), and the maximum value. The classes in many features seem overlapping, and several features record many outliers in the dataset. The distribution of the data is also skewed.

TABLE II. STATISTICAL CHARACTERISTICS OF THE DATASET

Feature	Full Sample		Full Sample		Full Sample	
	mean	std	mean	std	mean	std
Systolic Blood Pressure	138.33	20.5	143.74	23.68	135.46	17.98
Tobacco	3.64	4.59	5.52	5.57	2.63	3.61
LDL Cholesterol	4.74	2.07	5.49*	2.23	4.34*	1.87
Adiposity	25.41	9.82	54.49*	10.25	52.37*	9.52
Type A Behavior	53.1	9.82	54.49*	10.25	52.37*	9.52
Obesity	26.04	4.21	26.62*	4.39	25.74*	4.09
Alcohol	17.04	24.48	19.15	26.18	15.93	23.5
Age	42.82	14.61	50.29	10.65	38.85	14.88

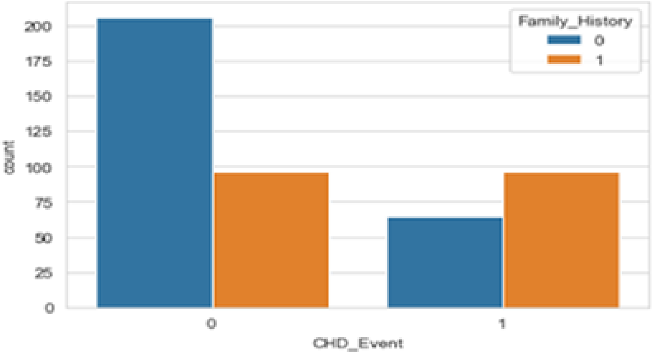

Fig. 1. Family History vs CHD_Event.

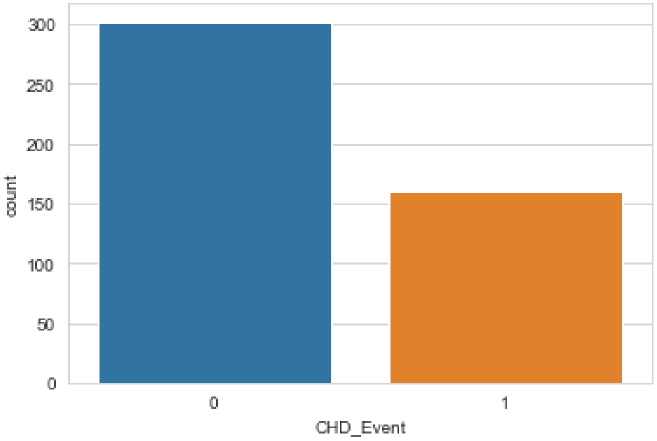

Fig. 2. Frequency of Positive and Negative CHD Events in the Dataset.

D. Feature Selection

The dataset includes many of the widely known risk factors or features that cause CHD, but we aim to rank which features are the most relevant to the target in predicting CHD and which features are the least relevant. This allows it to be further analyzed and interpreted by experts in the domain and could be used as the basis for gathering more or different data.

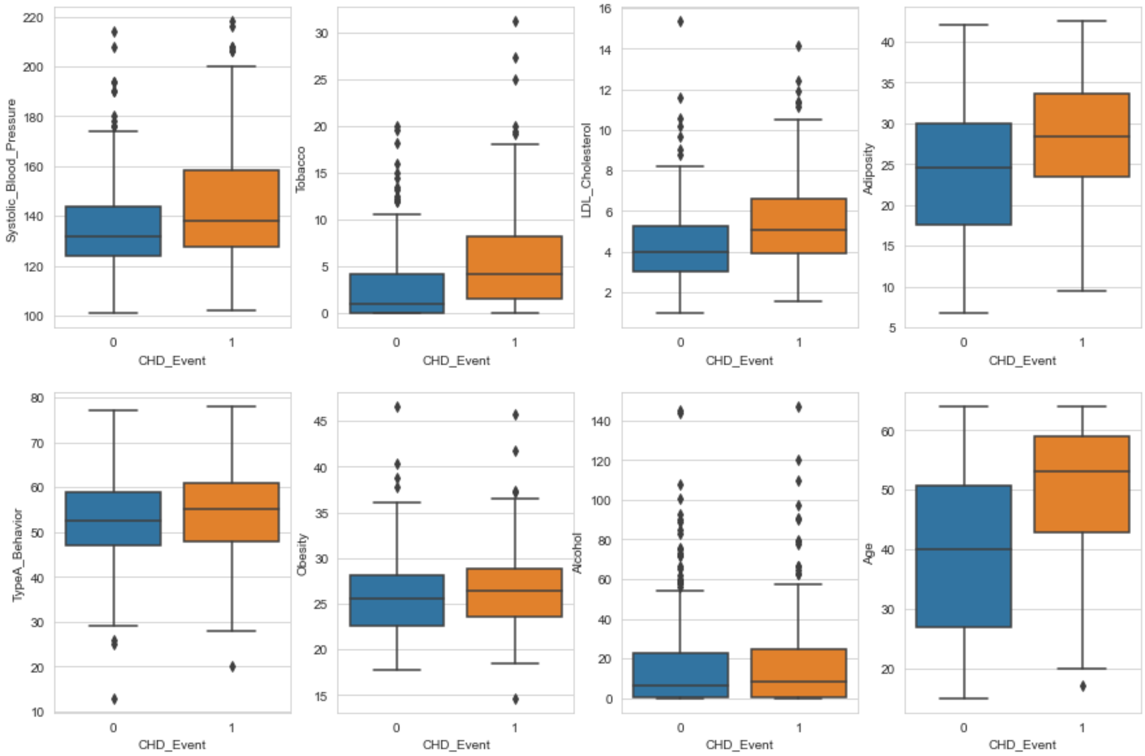

Fig. 3. Boxplot Representing the Distribution of Data Features with respect to CHD_Event.

1) ANOVA method: ANOVA stands for änalysis of variance änd it is a parametric statistical hypothesis test that determines whether the means of two or more samples of data (usually three or more) come from the same distribution or not [31]. An F-statistic, also known as an F-test, is a class of statistical tests that use a statistical test like ANOVA to measure the ratio between variance values. An ANOVA f-test is a type of F-statistic that uses the ANOVA method [32].

We used an implementation of the ANOVA f-test function from the scikit-learn machine learning library, which suits our classification problem task. Table III shows the scores of ANOVA f-test ranking of the features, i.e., the scores calculated for each input feature and the target variable (CHD Event) in descending order, the higher the score, the more important the feature.

2) Feature Importance: Statistical methods calculate the score of the feature ranking with relation to the features and the target variable, however, the importance and ranking of the features might be different when working together to predict the target variable. However, using machine learning methods for feature ranking provides insight into prediction models and which features are the most important and least important to the models when making a prediction.

Another method we used to compute a set of feature importance scores for our dataset is the permutation feature importance. The concept of Permutation Feature Importance was first introduced by Breiman [33] and applied to a random forest model. Permutation Feature Importance works by randomly changing the values of each feature column, one column at a time. It then evaluates the model.

TABLE III. ANOVA F-TEST RANKING

Feature	**Score**
Age	74.330
Tobacco	45.400
Family History	36.861
LDL Cholesteroal	34.197
Adiposity	31.756
Systolic Blood Pressure	17.674
Type A Behavior	4.948
Obesity	4.655
Alcohol	1.806

We have used the permutation_importance function from scikit-learn library with Random Forest as the fit model. We chose accuracy as the standard metric to measure performance in this context because this a classification problem. The ranking of the features is shown in Fig. 4.

The different methods of feature ranking showed that several features are most common as the most important features such as Age, Tobacco, LDL Cholesterol, Systolic Blood Pressure, Adiposity, and Family History. However, in our machine learning modeling experiments we used almost all the features and dropped Obesity as it has a high correlation with Adiposity, we used Pearson's correlation to calculate the

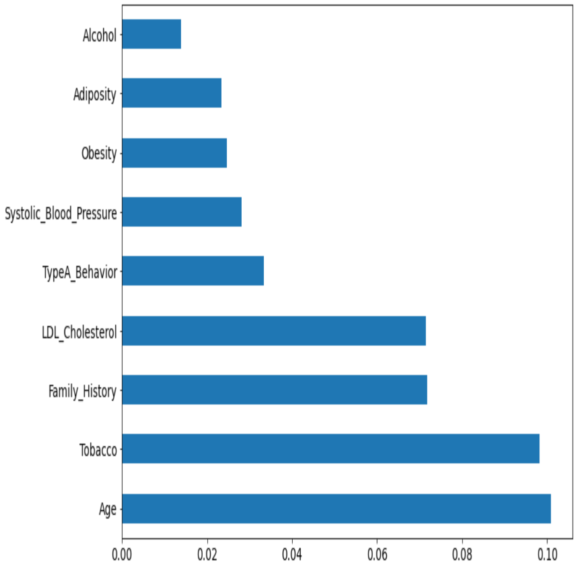

Fig. 4. Feature Importance using Random Forest.

score of the correlation between features and to drop the most correlated ones.

3) Pearson's Correlation Coefficient: Pearson's Correlation method is used for finding the feature correlation to remove the redundant features. As shown in Fig. 5, the correlation is represented as a number between -1 and 1, which indicates the extent to which two variables are related.

A correlation coefficient higher than 0.7 is considered strong and therefore one of the features can be dropped because this will affect the prediction accuracy. Given this, the obesity feature was dropped from the training dataset.

IV. Experimental Framework

In this study, we used scikit-learn Python library to conduct the experiments, the selected models, namely, Logistic Regression, SVM, KNN, and MLP Neural Network were applied on the dataset described in the last section, with 462 samples, 8 predictors (dropping Obesity feature) and 1 target variable.

Ten-fold stratified cross-validations were used for model training and testing. The stratified folds were used in these experiments because the dataset is imbalanced with evident imbalanced class distributions, as discussed earlier.

The machine learning techniques utilized for the prediction of CHD are set up as follows:

A. Logistic Regression

The logistic regression technique uses the logistic function [34] to model a binary dependent variable. The technique is capable of solving linear separable classes as well as complex problems. We have used the GridSearchCV function from scikit-learn library to find the optimal parameters, and the logistic regression was configured with 'lpfgs' solver, 'l2' penalty, and we set up 'C' to 0.25.

B. SVC

Based on the Support Vector Machine algorithm [35], this technique separates data points that belong to different classes with a decision boundary (hyperplane). The main parameter here is the kernel, it maps the observations into some feature space. With the help of GrisdSearchCV function, the kernel was set up to 'rbf', 'C' to 10 and we configured 'gamma' to auto.

C. KNN

KNN does not try to build an internal model, the computations are not done until the classification time. KNN stores instances of the training data in the features space and the class of an instance are determined based on the distance measure from its neighbors, Therefore, the most important parameter is the number of neighbors to be considered, here we set it up to 17. We used the elbow method [36] to calculate the optimal number of neighbors. And we set up 'minkowski' as the metric for the distance measure.

D. MLP Neural Networks

MLP is a neural network that consists of more than two layers with a number of neurons in each layer. We set up 3 layers with 50, 20, and 10 neurons consecutively. The activation function was set up to 'tanh' and the learning rate to '0.01'.

E. Classification of Evaluation Metrics

Accuracy, Precision (Positive predictive value), Recall (Sensitivity or True Positive rate), F1 score, in addition to Specificity (True Negative rate) were mainly used to evaluate the performance of the prediction models.

To calculate these, the confusion matrix is used to describe the performance of each predicted negative and positive class, as in Fig. 6.

Where:

- TN: is the total number of patients who correctly identified that they have no CHD.
- FN: is the total number of patients incorrectly identified that they have no CHD.
- TP: is the total number of patients correctly identified that they have CHD.
- FP: is the total number of patients incorrectly identified that they have CHD.

In imbalanced datasets precision, recall, and F1 score are often more important measures than accuracy. In this problem, even the accurate prediction of the CHD patients matters the most, i.e., high precision or high recall [37]. However, there is always a precision/recall trade-off, and in CHD prediction, high recall might be even preferred over high precision.

The aforementioned metrics can be calculated from the confusion matrix as follows:

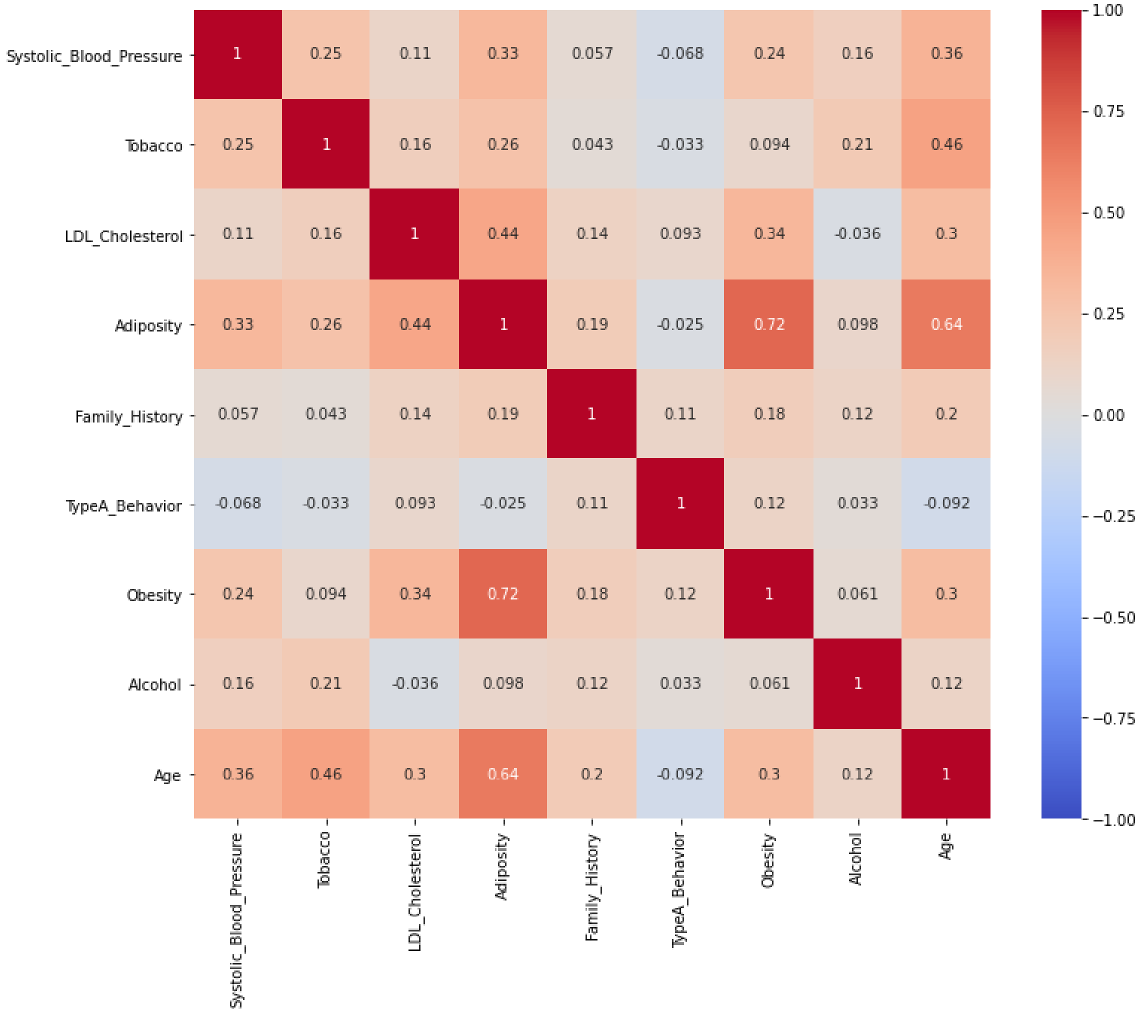

Fig. 5. Pearson's Correlation Matrix.

True			
0	TN	FP	
1	FN	TP	
	0	1	Predicted

Fig. 6. Confusion Matrix.

- Precision (Positive predictive value)

$$TP = \frac{TP}{TP + FP} \quad (1)$$

- Recall (Sensitivity or True positive rate)

$$recall = \frac{TP}{TP + FN} \quad (2)$$

- Specificity (True negative rate)

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

- F-score

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad (4)$$

- Accuracy

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

F. Data Oversampling

In general, many medical datasets show signs of imbalanced class distribution which greatly hampers the detection of rare events, as most classification methods implicitly assume an equal occurrence of classes [38, 39]. The dataset of this study is a very small size with a total of 462 instances, distributed into 302 negative CHD instances and 160 positive CHD instances.

In an imbalanced class distribution problem, the sample size is critical in evaluating the classification model, and the high error rate caused by the imbalanced class distribution decreases as the size of the training dataset increases [39]. Furthermore, in the dataset at hand, many variables are not linearly correlated, not linearly separable, and are complexly

TABLE IV. PREDICTION RESULTS - MEAN OF 10 FOLD CROSS-VALIDATION

Classifier	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**Specificity**
SVM	**0.738***	0.550	0.679	0.463	0.884
MLP Neural Network	0.734	0.553	0.661	0.475	0.871
KNN	0.732	0.504	**0.7***	0.394	**0.911***
Logistic Regression	0.727	**0.563***	0.633	**0.506***	0.844

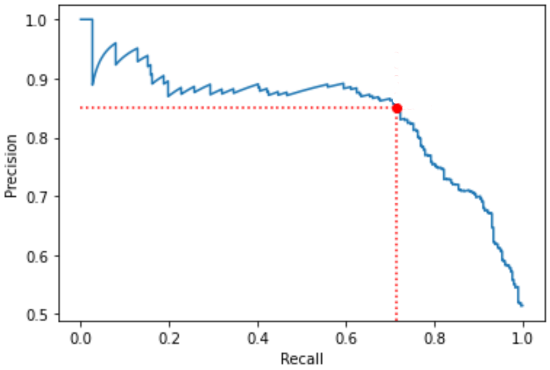

Fig. 7. Precision vs Recall in SVM.

overlapping, as discussed earlier. The authors in [40] have stated that not the imbalanced distribution of classes is the main problem in the classification with imbalanced data classification, but many characteristics, among them “the presence of small disjuncts, the lack of density in the training data, the overlapping between classes, the identification of noisy data”.

Several techniques have been introduced and used for handling the imbalanced datasets and improving the prediction [41-43]. In this study we have used Synthetic Minority Oversampling TEchnique (SMOTE) for short, this technique was first described by [41]. In particular, we have used the K-means SMOTE method [44, 45].

1) K-means SMOTE: SMOTE with the K-means method improves classification by producing minority class samples in safe areas of the input space. The method reduces noise while effectively addressing imbalances within and within samples. We used the KMeans SMOTE class from the imbalanced-learn Python library.

V. RESULTS

The mean accuracy results of applying the 10-fold stratified cross-validation on the dataset were obtained show that SVM slightly outperformed MLP neural network classifier, KNN, and Logistic Regression. The results were 73.8%, 73.4%, 73.2% and 72.7% respectively. However, as discussed before, accuracy alone is not the main concern here because this is an imbalanced dataset, the distribution of the labeled target class is unequal. The mean scores of applying 10-fold stratified cross-validation with F1 score, Precision, Recall, Specificity of the 4 classifiers are summarized in Table IV. The results show improvements in the accuracy compared to previous research results on the same dataset.

A. Results on the Oversampled Dataset

The dataset now has 604 samples with 302 instances with negative CHD events and 302 instances with positive CHD events. The mean scores of applying the classifiers 10-fold stratified cross-validation on the dataset after oversampling are summarized in Table 5. The classifiers are ordered based on the accuracy we also calculated the Matthews Correlation Coefficient (MCC) score, to highlight which classifier achieved good results in all 4 categories of the confusion matrix.

The results show major improvement on the recall score with an average of 32% for all classifiers. From 10 fold cross-validation, MLP neural network has recorded an average 80.3% of Recall score, while KNN has achieved an average 80% of Precision score and an average 85.8% of Specificity score. The overall accuracy for all classifiers has also been improved, SVM has achieved the highest overall accuracy and good results in all scores combined.

Fig. 7 shows the plot of the precision vs recall for SVM classifier, as it can be noticed the precision has dropped at around 78% of recall, so we can even create an SVM model with let’s say over 85% of Precision score with over 71% of recall by tunning the threshold of precision.

VI. DISCUSSION AND CONCLUSION

Our results show that, given sufficient data and proper selected clinical features, machine learning techniques are capable of predicting the occurrence of CHD events with high accuracy. The application of the four machine learning techniques SVM, KNN, MLP neural networks, and logistic regression using the South African Heart Disease dataset with the selected features reported roughly as high as 74% accuracy. While this shows a noticeable improvement in the prediction performance compared to previous researches on the same data, the main issue in this study was to resolve the imbalanced classification problem in the dataset and achieve even higher scores in Precision and Recall in particular, in addition to improving the overall prediction accuracy. Such a problem in the dataset was tackled by applying K-means SMOTE oversampling techniques, and as a result, the prediction performance of all prediction models has significantly enhanced, with an average improvement of 32% on the Recall score and an average improvement of 11% on the Precision score.

Among the four prediction techniques applied on the oversampled dataset in this study, SVM has obtained the best results in all the four confusion matrix categories, marginally followed by KNN, MLP neural network, and logistic regression respectively. However, from the usability standpoint, one might choose to use KNN as a prediction model for this problem, since KNN has obtained an 80% Precision score and around 86% Specificity score. Whereas MLP neural network has reported an 80% Recall score. Recent trends in prediction and classification are going toward using a combination of

TABLE V. PREDICTION RESULTS ON THE OVERSAMPLED DATASET- MEAN OF 10 FOLD CROSS VALIDATION

Classifier	**MCC**	**Accuracy**	**F1 Score**	**Precision**	**Recall**	**Specificity**
SVM	**+0.561**	**0.781***	0.780	0.784	0.776	0.785
MLP Neural Network	+0.549	0.774	0.781	0.760	**0.803***	0.745
KNN	+ 0.553	0.776	0.767	**0.8***	0.737	**0.858***
Logistic Regression	+0.538	0.769	0.775	0.755	0.795	0.742

prediction techniques for more accurate and more reliable outcomes. That is, it is a good idea in practice to use the SVM, KNN, MLP neural network classification models together for predicting the positive and negative CHD cases, as they strengthen and complement each other.

Our feature selection techniques have showed and confirmed that clinical features and risk factors such as, Tobacco, LDL Cholesterol, Systolic Blood Pressure, Adiposity, and Family History are among the most important features that help in the early detection and the prediction of the presence of CHD events from medical records. Medical practitioners can take advantage of the exploratory data analysis conducted on the dataset to show correlations and relationships between patients' data.

The success of machine learning relies heavily on the richness of the data representing the phenomenon under consideration. Even though the selected dataset has the most widely known features and risk factors for predicting CHD, with a rather rich set of features, more data and more variables can potentially help improve the prediction results. If additional external datasets with the same features from different regions had been available, we would have used it as a validation of our findings.

As future work, we are planning to apply our machine learning approach on other datasets of cardiovascular diseases, cancer, and infectious diseases. We are also preparing to deploy the models as a web service and integrate it in a web application to allow medical practitioners assess its usability in the real world.

REFERENCES

[1] AIHW. Coronary Heart Disease. 2020 [cited 2021 20 Feb]; Available from: https://www.aihw.gov.au/reports/australias-health/coronary-heart-disease

[2] Hajar, R., *Risk factors for coronary artery disease: historical perspectives.*, Heart views: the official journal of the Gulf Heart Association, 2017. 18(3): p. 109

[3] S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, *"Can machine-learning improve cardiovascular risk prediction using routine clinical data?"*, PloS one, 2017. 12(4): p. e0174944.

[4] P.K. Sahoo, S.K. Mohapatra, and S.L. Wu, *"Analyzing healthcare big data with prediction for future health condition"*, IEEE Access, 2016. 4: p. 9786-9799.

[5] D. Chicco, and G. Jurman, *"Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone"*, BMC medical informatics and decision making, 2020. 20(1): p. 16.

[6] S. Blecker, S.D. Katz, L.I. Howrwitz, and G. Kuperman, *"Comparison of approaches for heart failure case identification from electronic health record data"*, JAMA cardiology, 2016. 1(9): p. 1014-1020.

[7] M. Fatima, M. Pasha, *"Survey of machine learning algorithms for disease diagnostic"*, Journal of Intelligent Learning Systems and Applications, 2017. 9(01): p. 1.

[8] R. Sujatha, and A. Nithya, *"A Survey of Health Care Prediction Using Data Mining"*, International Journal of Innovative Research in Science, Engineering and Technology, 2016. 5(8): p. 14538.

[9] D. Kinge, and S.K. Gaikwad, *"Survey on data mining techniques for disease prediction"*, International Research Journal of Engineering and Technology (IRJET), 2018. 5(01): p. 630-636.

[10] K.G. Dinesh, K. Arumugaraj, K.D. Santosh, and V Mareeswari, *"Prediction of cardiovascular disease using machine learning algorithms"*, in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). 2018. IEEE.

[11] J.J. Beunza, E. Puertas, E. Garcia-Ovejero, G. Villalba, E. Condes, G. Koleva, C. Hurtado, and M.F. Landecho *"Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)"*, Journal of biomedical informatics, 2019. 97: p. 103257.

[12] Panicker, S.,*"Use of Machine Learning Techniques in Healthcare: A Brief Review of Cardiovascular Disease Classification"*. 2020.

[13] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, N.S. Naik, *"Prediction of coronary heart disease using supervised machine learning algorithms"*, TENCON 2019-2019 IEEE Region 10 Conference (TENCON). 2019. IEEE.

[14] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H. Rudd and M. van der Schaar, *"Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants"*, PloS one, 2019. 14(5): p. e0213653.

[15] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P.M. Kebria, F. Khozemeh, et al., *"Machine learning-based coronary artery disease diagnosis: A comprehensive review"*, Computers in biology and medicine, 2019. 111: p. 103346.

[16] A.H. Gonsalves, F. Thabtah, R.M.A Mohammad, G. Singh, *"Prediction of coronary heart disease using machine learning: an experimental analysis"*, in Proceedings of the 2019 3rd International Conference on Deep Learning Technologies. 2019.

[17] S.I. Ayon, M.M. Islam, and M.R. Hossain, *"Coronary artery heart disease prediction: a comparative study of computational intelligence techniques"*, IETE Journal of Research, 2020: p. 1-20.

[18] K.H. Miao, and J.H. Miao, *"Coronary heart disease diagnosis using deep neural networks"*, Int. J. Adv. Comput. Sci. Appl., 2018. 9(10): p. 1-8.

[19] G.D. Flora, and M.K. Nayak, *"A brief review of cardiovascular diseases, associated risk factors and current treatment regimes"*, Current pharmaceutical design, 2019. 25(38): p. 4063-4084.

[20] KEEL Data set. *"South African heart dataset"*, [cited 2021 18, February]; Available from: https://sci2s.ugr.es/keel/dataset.php?cod=184.

[21] R. Nakanishi, D. Dey, F. Commandeur, P. Slomka, J. Betancur, H. Gransar, C. Dailing, K. Osawa, D. Berman, and M. Budoff, *"Machine learning in predicting coronary heart disease and cardiovascular disease events: results from the multi-ethnic study of atherosclerosis (mesa)"*, Journal of the American College of Cardiology, 2018. 71(11S): p. A1483-A1483.

[22] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, *"A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease"*, 2017 IEEE symposium on computers and communications (ISCC). 2017. IEEE.

[23] S. Safdar, S. Zafar, N Zafar, N.F. Khan, *"Machine learning based decision support systems (DSS) for heart disease diagnosis: a review"*, Artificial Intelligence Review, 2018. 50(4): p. 597-623.

[24] K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T Dudley, and P.P. Sengupta, *"Machine learning in cardiovascular medicine: are we there yet?"*, Heart, 2018. 104(14): p. 1156-1164.

[25] C. Sowmiya, P. Sumitra. *"Analytical study of heart disease diagnosis using classification techniques"*, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). 2017. IEEE.

[26] E.B. Maini, B. Venkateswarlu, and A. Gupta. *"Applying machine learning algorithms to develop a universal cardiovascular disease prediction system"*, International Conference on Intelligent Data Communication Technologies and Internet of Things. 2018. Springer.

[27] K.H. Miao, J.H. Miao, and G.J. Miao, *"Diagnosing coronary heart disease using ensemble machine learning"*, Int J Adv Comput Sci Appl (IJACSA), 2016.

[28] *Heart Disease Data Set*, [cited 2021 20, March]; Available from: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[29] *Statlog (Heart) Data Set*, [cited 2021 20, March]; Available from: http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

[30] A. Baccouche, B. Garcia-Zapirain, C Castillo Olea, and A. Elmaghraby, *"Ensemble Deep Learning Models for Heart Disease Classification: A Case Study from Mexico"*. Information, 2020. 11(4): p. 207.

[31] B.G. Tabachnick, and L.S. Fidell, *"Experimental designs using ANOVA"*, 2007: Thomson/Brooks/Cole Belmont, CA.

[32] A. Bathke, *"The ANOVA F test can still be used in some balanced designs with unequal variances and nonnormal data"*, Journal of Statistical Planning and Inference, 2004. 126(2): p. 413-422.

[33] L. Breiman, *"Random forests"*, Machine learning, 2001. 45(1): p. 5-32.

[34] D.G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, *"Logistic regression"*, 2002: Springer.

[35] T. Joachims, *Svmlight: Support vector machine*, http://svmlight. joachims. org/, University of Dortmund, 1999. 19(4).

[36] P. Dangeti, *"Statistics for machine learning"*, 2017: Packt Publishing Ltd.

[37] A. Géron, *"Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems"*, 2019: O'Reilly Media.

[38] A. Ali, S.M. Shamsuddin, and A.L. Ralescu, *"Classification with class imbalance problem"*, Int. J. Advance Soft Compu. Appl, 2013. 5(3).

[39] M.A. Mazurowski, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, and G.D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Neural networks, 2008. 21(2-3): p. 427-436.

[40] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, *"An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics"*, Information sciences, 2013. 250: p. 113-141.

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

[42] W. Elazmeh, N. Japkowicz, and S. Matwin, *"Evaluating misclassifications in imbalanced data"*, in European Conference on Machine Learning. 2006. Springer.

[43] M.A. Maloof, *"Learning when data sets are imbalanced and when costs are unequal and unknown"*, ICML-2003 workshop on learning from imbalanced data sets II. 2003.

[44] G. Douzas, F. Bacao, and F. Last, *"Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE"*, Information Sciences, 2018. 465: p. 1-20.

[45] F. Last, G. Douzas, and F. Bacao, *"Oversampling for Imbalanced Learning Based on K-Means and SMOTE"*,.arXiv preprint arXiv:1711.00837, 2017.

[46] N. Kausar, A. Abdullah, B.B. Samir S., Palaniappan, B.S. AlGhamdi, and N. Dey, "Ensemble cluster algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease", Journal of Medical Imaging and Health Informatics, vol.6, number 1, February 2016, p.78-87(10).

[47] N. Kausar, S. Palaniappan, B.B. Samir, A. Abdullah, and N. Dey,*"Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients."*, Applications of intelligent optimization in biology and medicine, pp. 217-231. Springer, Cham, 2016.