IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 10, 2021, accepted June 25, 2021, date of publication July 20, 2021, date of current version August 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3098688

An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases

AQSA RAHIM[1], YAWAR RASHEED[1], FAROOQUE AZAM[1], MUHAMMAD WASEEM ANWAR[1], MUHAMMAD ABDUL RAHIM[1], AND ABDUL WAHAB MUZAFFAR[2]

[1]Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan

[2]College of Computing and Informatics, Saudi Electronic University, Riyadh 11673, Saudi Arabia

Corresponding author: Muhammad Waseem Anwar (waseemanwar@ceme.nust.edu.pk)

ABSTRACT Cardiovascular diseases are considered as the most life-threatening syndromes with the highest mortality rate globally. Over a period of time, they have become very common and are now overstretching the healthcare systems of countries. The major factors of cardiovascular diseases are high blood pressure, family history, stress, age, gender, cholesterol, Body Mass Index (BMI), and unhealthy lifestyle. Based on these factors, researchers have proposed various approaches for early diagnosis. However, the accuracy of proposed techniques and approaches needs certain improvements due to the inherent criticality and life threatening risks of cardiovascular diseases. In this article, a MaLCaDD (**Ma**chine **L**earning based **Ca**rdiovascular **D**isease **D**iagnosis) framework is proposed for the effective prediction of cardiovascular diseases with high precision. Particularly, the framework first deals with the missing values (via mean replacement technique) and data imbalance (via Synthetic Minority Over-sampling Technique - SMOTE). Subsequently, Feature Importance technique is utilized for feature selection. Finally, an ensemble of Logistic Regression and K-Nearest Neighbor (KNN) classifiers is proposed for prediction with higher accuracy. The validation of framework is performed through three benchmark datasets (i.e. Framingham, Heart Disease and Cleveland) and the accuracies of 99.1%, 98.0% and 95.5 % are achieved respectively. Finally, the comparative analysis proves that MaLCaDD predictions are more accurate (with reduced set of features) as compared to the existing state-of-the-art approaches. Therefore, MaLCaDD is highly reliable and can be applied in real environment for the early diagnosis of cardiovascular diseases.

INDEX TERMS Cardiovascular diseases, machine learning, cardiovascular prediction, ensemble, SMOTE, feature importance, MaLCaDD framework.

I. INTRODUCTION

The busy schedule of the modern era leads to an unhealthy life style which causes anxiety and depression. In order to overcome these conditions, there is a tendency to resort to excessive smoking, drinking and taking drugs. All these things are the root cause of many dangerous diseases including cardiovascular diseases, cancer etc. [1]. According to the World Health Organization (WHO), cardiovascular diseases (CVDs) have the highest number of death rates, globally. Almost 31% of the world's deaths are because of the CVDs [2]. The early prediction of these kinds of diseases is very important so that precautionary measures could be taken before something serious happens.

The associate editor coordinating the review of this manuscript and approving it for publication was Wentao Fan.

Cardiovascular Diseases (CVDs) is a term which is used to describe a condition that affects the heart or blood vessels. Four main types of CVDs include Coronary Heart Disease, Stroke/ Transient ischemic attack (known as TIA/ Mini-Stroke), Peripheral arterial disease and Aortic disease [3]. The exact cause of CVDs is unknown still; some risk factors are responsible for these diseases including high blood pressure, smoking, diabetes, body mass index (BMI), cholesterol, age, family history, etc. These factors are different for different people. Age, gender, Stress, and unhealthy lifestyle are also some of the major factors which are responsible for the CVDs [4]. The major challenge is to timely predict these diseases with high accuracy so that mortality rate may be reduced through effective medication and other counter measures.

Over the period of time, researchers proposed several algorithms for the prediction of CVDs. In Spain, SCORE

risk chart and REGICOR risk score [5] are being used for cardiovascular risk prevention on the recommendation of the Spanish Society of Cardiology. Similarly, In United States researchers have assessed five methods for predicting mortality with congestive heart failure (CHF) [6]. A prospective cardiovascular münster (PROCAM) study was also developed for the prediction of CVDs [7]. One of the common methods for the prediction of cardiovascular disease was proposed by Jain et al [8] using ECG signals. Image-based diagnosis can also be made using cardiac imaging for the prediction of CVDs [9]. In the same way, some researchers have used other features for the prediction of cardiovascular diseases.

During past decade, researchers have proposed many algorithms for the prediction of CVDs by using different datasets and techniques. The common datasets which are used for the prediction of CVDs include: heart disease [13], Cleveland [14], Framingham [15] and Cardiovascular Disease [16]. These datasets consist of different attributes that are used for the prediction of the CVDs. The factors which are involved in cardiovascular disease include modifiable and non-modifiable risk factors. Non-modifiable are the one that cannot be changed such as age ethnicity, and family history. Whereas, the modifiable risk factors such as Smoking, unhealthy lifestyle, blood pressure, and cholesterol can be changed and controlled by taking certain precautions and medication. Many datasets have been shaped by taking into account these attributes and a lot of effort has been done by the researchers on these datasets. One of the renowned datasets is the Framingham dataset [15] which is collected against these attributes. Many researchers have used this data set in order to validate their prediction frameworks. Nitten *et al.* [17] presented a prediction algorithm using 22 features. Afterward, machine learning algorithms were applied and based on their computation time and accuracies, proposed an ensemble method for the prediction of cardiovascular disease. Rubini *et al.* [18] proposed a solution for the prediction of heart disease using the Random Forest algorithm. The proposed algorithm was compared with different classifiers including Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) and it was demonstrated that the proposed algorithm i.e. Random Forest achieved an accuracy of 84.81%. Hoda *et al.* [19] has used the Framingham scoring model for the validation of their framework. The algorithms which were included in the experimentation are KNN and random forest and it was observed that accuracy given by the KNN (66.7%) was relatively higher than that of the Random Forest (63.49%). So KNN was considered as the proposed algorithm.

In the given research context, different machine learning [10] and deep learning [11] based techniques were developed for the prediction of CVDs. However, the focus of researchers is on feature selection techniques and classification algorithms while ignoring the issue of class imbalance. The problem of class imbalance highly affects the accuracy of the classification algorithm. Furthermore, the large number of features are required for prediction when data is not balanced. This significantly increases the computational complexity and make the solution impractical for real environment. In addition to this, the improvement of existing feature selection techniques are required to reduce the computational complexity while achieving acceptable accuracy. Similarly, the accuracies of the existing classifiers need to be improved to achieve reliable results. To summarize, there is a strong need of an integrated machine learning framework for cardiovascular diseases where data balancing, optimum feature selection and improved classification should be achieved in a systematic way. This not only leads to improve the predictions for cardiovascular diseases but also reduces the computational complexities. To the best of our knowledge, such integrated framework for cardiovascular diseases is hard to find in literature and industrial projects.

To achieve aforementioned objectives, this article presents a MaLCaDD (**M**achine **L**earning based **C**ardiovascular **D**isease **D**iagnosis) framework. It is fully functional on major cardiovascular diseases factors like high blood pressure, family history, stress, age, gender, cholesterol, Body mass index (BMI), and unhealthy lifestyle. The overview of MaLCaDD is shown in Fig. 1. Particularly, in contrast to the existing studies where researchers have mainly focused on various feature selection and traditional classification methods, MaLCaDD intends to improve the overall accuracy via handling of missing values and imbalanced data. The missing values have been handled by replacing the missing value with the mean of all the values of a corresponding feature. In order to deal with data imbalance, MaLCaDD proposes Synthetic Minority Over-sampling Technique (SMOTE). Once data become balanced, MaLCaDD employs feature importance technique for the selection of optimum set of features. Finally, ensemble of Logistic Regression and K-Nearest Neighbor (KNN) is proposed for improved prediction as shown in Fig. 1. The contributions of this article are summarized as:

1) An integrated machine learning framework is proposed where data balancing, feature selection and classification are targeted altogether for the improved and early prediction of cardiovascular diseases. Particularly, missing values and data balancing are managed through mean and SMOTE respectively. Moreover, optimum feature sets are derived using feature importance technique. Finally, improved prediction is achieved through the ensemble of logistic regression and KNN classifiers.
2) The implementation of framework is carried out in PYTHON using different libraries. The framework is publically available [56] at GitHub repository.
3) The validation of framework is performed through three benchmark datasets (i.e., Framingham, Heart Disease and Cleveland) and the accuracies of 99.1%, 98.0% and 95.5 % are achieved respectively. This ensures the broader application of proposed framework on multiple cardiovascular datasets.
4) Finally, proposed framework is comprehensively compared with state-of-the-art studies. The results prove

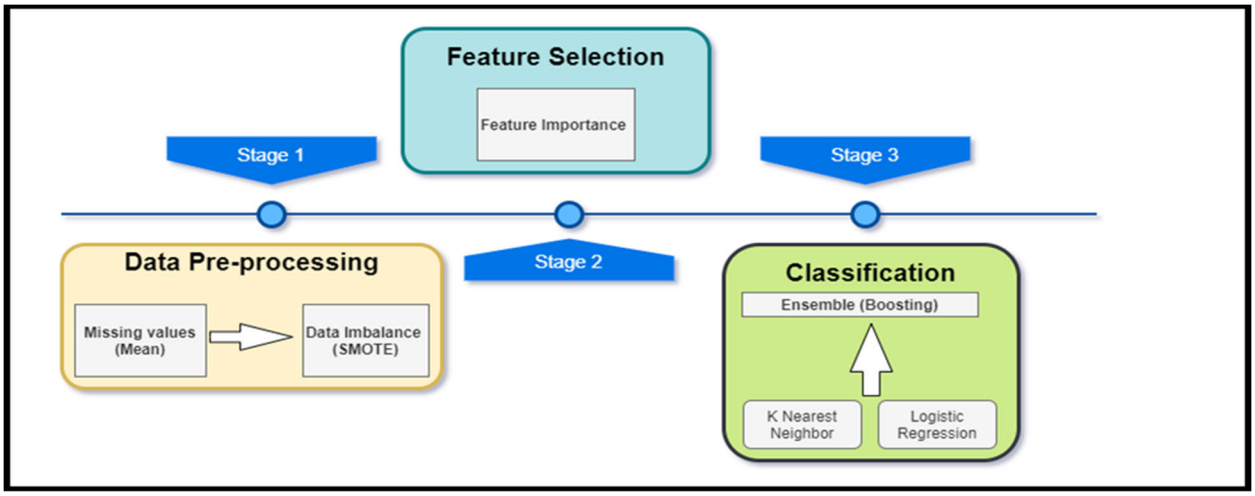

FIGURE 1. Machine Learning based Cardiovascular Disease Diagnosis (MaLCaDD) Framework.

that proposed framework outperforms the state-of-the-art studies by achieving the improved accuracy with reduced set of features.

Rest of the paper is arranged as follows: Section II presents Literature Review and Research Gap. The framework is proposed in Section III and experimental evaluation is performed in Section IV. The comparative analysis of proposed framework with state-of-the-art is discussed in Section V. Finally, Section VI presents the conclusion and future work.

II. PRELIMINARIES

In this section, we discuss the existing work relevant to the domain under discussion. We also explore the datasets which have been used by the researchers for making predictions. Particularly, Section A highlights the literature review and Section B reports the research gap.

A. LITERATURE REVIEW

Cardiovascular Diseases (CVDs) affect the heart and blood vessels. There are numerous types of CVDs that may have a variety of effect over a human body. Certain risk factors have been identified which may cause CVDs which include high blood pressure, smoking, diabetes, body mass index (BMI), cholesterol, age, family history, etc. [20]. Age, gender, Stress, and unhealthy lifestyle are also some of the major factors which are responsible for the CVDs [21]. In a nutshell, whatever is the cause, the most important thing is the timely / early identification CVDs.

Researchers have proposed various approaches for the prediction of CVDs on the basis of above-mentioned risk factors. Similarly, variety of datasets have been used by the researchers for the validation of their proposed approaches. Mostly, Cleveland [14], Framingham [15] and Heart Disease dataset [13] have been in lime light. Mostly, the attributes of these datasets are similar. However, the major difference in these datasets is of the experimental setup i.e. how the dataset is collected. In the succeeding paras, we present the research that has been undertaken by worthy researchers in connection with prediction of the CVDs along with the corresponding data sets. Tanvi *et al.* [22] have used the Cleveland dataset for the prediction of heart diseases. As part of prediction, 14 features were used for the training of the model. Moreover, classifiers of Decision Tree, MARS, Random Forest and TMGA were used for making predictions. Out of all these models, Decision Tree has rendered the highest accuracy of 93.24% and the time taken was 112.36 sec. Singh *et al.* [23] have applied different classification algorithms on the heart disease dataset [13] which is available in the University of California Irvine (UCI) repository. In the proposed solution, authors have used a backward selection method for feature selection. By using this method, 11 significant features were selected which were then used to train algorithms including Logistic Regression, SVM and pruning decision tree. 87.1% accuracy was achieved by applying Logistic Regression which was highest among all of the other models. Amanda et al [24] have used 10 different features from the South African heart disease dataset. Three different models i.e., Decision Tree, Naïve Bayes and SVM are applied on the dataset and then these models are evaluated using Confusion Matrix. Out of these three models, Naive Bayes produced the best results. Ketut *et al.* [25] have used a dataset that was collected from Harapan Kita Hospital. From this dataset, 18 parameters were used for the prediction of heart diseases. In this study, KNN was applied with and without parameter weighting and achieved an accuracy of 75.11% and 74.0% respectively. In addition to that, Naïve Bayes and SVM were also applied to the same dataset but the results achieved from those models were not very significant. Randa *et al.* [26] have used the Collective Heart Disease dataset (CAD) for the prediction of heart valve diseases. 13 features were used for the training of the model. It was observed that the best results were achieved with the accuracy of 92.0% when the dataset was trained with Naive Bayes classification algorithm.

TABLE 1. Comparison of machine learning algorithms with different datasets.

Author	Year	Dataset	Features	Classifier	Accuracy
Randa [26]	2016	CAD	13	Naïve Bayes	92.0 %
Tanvi [22]	2017	Cleveland	14	Decision tree	93.2 %
Rubini [18]	2019	Framingham	10	Random forest	84.8 %
Divya [28]	2019	Framingham	13	Random forest	96.8 %
Nitten [17]	2018	Framingham	22	SVM	90.2 %
Hoda[19]	2017	Framingham	9	KNN	66.7 %
Shen [27]	2014	Framingham	10	CART	-
Ketut [25]	2018	Harapan Kita Hospital (HKH)	14	KNN	75.1 %
Singh [23]	2018	Heart Disease	11	Logistic regression	87.1 %
Amanda [24]	2019	South African Heart Disease	10	Naïve Bayes	82 .0%

Researches have used different methods and algorithms on the Framingham dataset as well for the prediction of the cardiovascular disease. Rubini *et al.* [18] proposed a solution for the prediction of heart disease using the Random Forest algorithm. The proposed algorithm was compared with different classifiers including Logistic Regression, Naïve Bayes and SVM. In the proposed paper, Random Forest achieved an accuracy of 84.81%. Prediction of CVDs using KNN and Random Forest was propounded by Hoda *et al.* [19]. In the proposed approach, KNN and Random Forest were used for classification. It was observed that KNN has achieved an accuracy of 66.7% whereas; Random Forest algorithm achieved 63.4% accuracy. Shen *et al.* [27] proposed a Clustered CART Framework to manage the large number of discovered rules. For training and testing, 10 features were used. These features were then preprocessed i.e. continuous features were discretized and the missing values were replaced with global means. Afterward, an existing ARM tool was applied for the generation of CARS for the prediction of chronic disease. Divya *et al.* [28] propounded a solution for the automatic prediction of heart diseases. Firstly, machine learning algorithms were used, thereafter; the ensemble method was created for the final cross-validation results using machine learning and ensemble method. Nitten *et al.* [17] presented a prediction algorithm using 22 features. Afterward, machine learning algorithms were applied and based on their computation time and accuracies, an ensemble method was proposed for the prediction of cardiovascular diseases. Table 1 shows the Comparison of Machine Learning Algorithm with different datasets.

The dataset which we are using in our research is the Framingham dataset as it is one of the frequently used datasets and it contains all the necessary attributes which can help us in the prediction of these diseases. Framingham dataset [15] was collected in three different phases. The first phase was conducted back in 1948 in which the data was collected from 5209 men and women with ages ranging from 30 to 62. The second phase was conducted in 1971, in which 5124 people participated and were asked to go through the same examination process. These were the second generation of the people who participated in the first phase. Finally, in April 2002 the final phase was undertaken in which the data was collected from the third generation of the original cohort. The dataset consists of 16 features that may be used for the prediction. The Framingham dataset has data from the people of three different generations including participants who originally participated, their children, and grandchildren [29]. This makes the dataset very reliable and it is expected that highest accuracy may be achieved by using this dataset for prediction.

B. RESEARCH GAP

Most of the researchers have focused on improving the accuracy of prediction via various feature selection and traditional classification methods. Whereas, missing values in the data that highly affects the accuracy of the model (as missing values in the data reduces the samples of the data which results in an ineffective model) have little been catered for. In addition to that, the problem of class imbalance (in which samples of one class is relatively less than the samples of the other class / classes) has also not been amply focused in previous researches for improving accuracy. Consequently, there is a dire need that problems of missing values and class imbalance must be catered for before suggesting any classification mechanism.

In a similar manner, features of the dataset highly affect the accuracy and computational complexity of machine learning process. Therefore, the prime thing in any machine learning process is to select the right subset of features while performing feature extraction. Although, researchers have already proposed various feature selection techniques and classification algorithms for the given datasets, the feature selection process needs a definite improvement so that right subset of the features are selected that contribute in reliable prediction with improved accuracy. In order to reduce the computational complexity of proposed classifiers, there is a dire need that the feature selection technique should bring out minimum number of essential features that could help in making reliable predictions. Moreover, for improved accuracy the classification algorithm also need to be efficient.

Hence, such a versatile framework that is applicable on wide variety of datasets, takes into account the problems of missing values/ imbalanced class and performs reliable predictions (with minimal features and reduced computational complexity) is hard to find in literature.

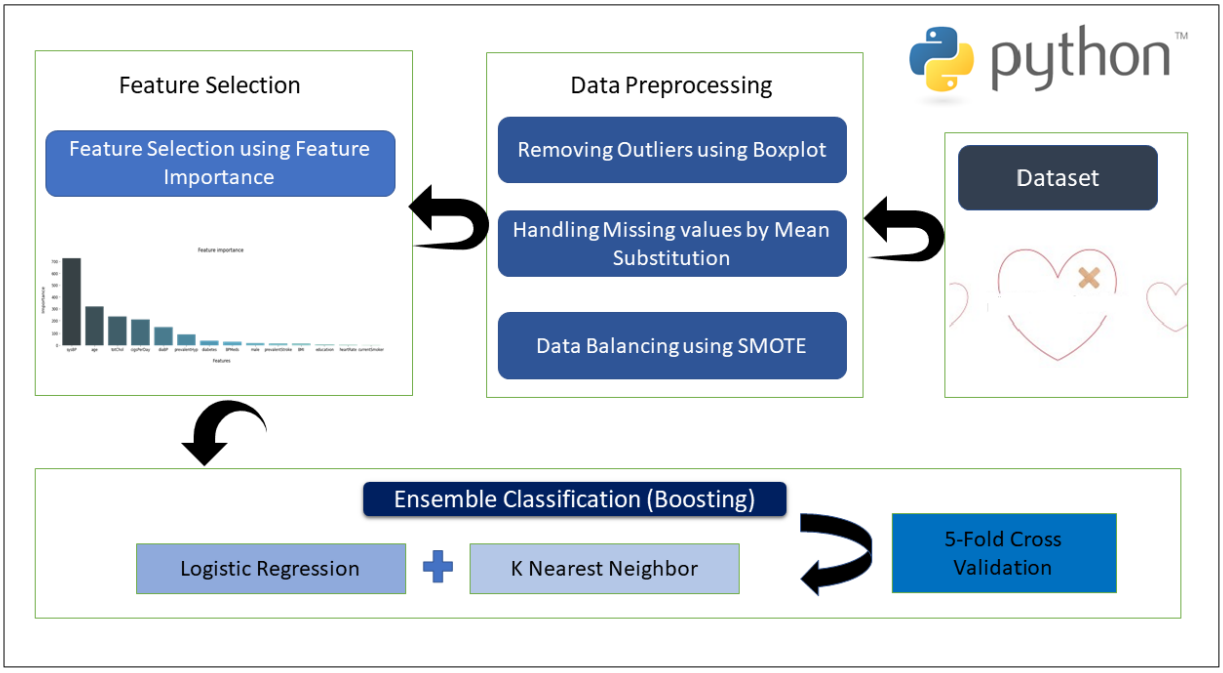

FIGURE 2. Flowchart of (MaLCaDD) Framework.

III. PROPOSED SOLUTION

In this paper, we have proposed a novel Machine Learning based Cardiovascular Disease Diagnosis (MaLCaDD) Framework for the prediction of cardiovascular diseases. The objective of this paper is to present an accurate machine learning model which can identify cardiovascular diseases reliably, based on the patients' clinical parameters. The proposed approach is based on steps which are: 1) Data preprocessing that includes outlier removal, replacing missing values and handling imbalance class. 2) Feature selection using feature importance technique, and 3) Classification using an Ensemble of Logistic Regression and KNN as shown in Fig. 2. Thereafter, model training is performed and the trained model will then be used to make predictions. Our main aim is to get the improved results with reduced number of features / computational complexity. Now we will present our framework in detail. In the proposed MaLCaDD Framework, initially data preprocessing is undertaken. In the data preprocessing step, primarily the data is explored for possible outliers. Outlier is that sample in the dataset which deviates from the normal behavior of the dataset [30].

The majority of the outliers in the data are considered as noise which adds no value to the significance of data and negatively affects the performance of the model. Raveendrababu *et al.* [31] have demonstrated that removing the outliers from the data helps in achieving the improved results and proposed various methodologies for outlier removal. Our proposed framework uses Boxplot for the removal of outliers. Boxplots are used to give five numbers summary i.e. minimum, first quartile (Q1), median (Q2), third quartile (Q3) and maximum. When these five points are plotted it forms a box like structure and any point which is lying outside this box is considered as an outlier. After removing the outliers, our proposed framework checks for the missing values in the data. Missing values in the data are very common which can exist because of the faulty instrument or human error [32]. If there are missing value in the data then model cannot be efficiently trained on that data due to the reduced number of training samples. This affects the accuracy of the model. That is why our proposed framework replaces the missing values by the mean of all the values of the respective attribute. Mean substitution is one of the most common methods used by various researchers (e.g., Dodeen et al [33]) for replacing the missing values. This helps to retain training data without adding further information to the dataset which reduces the chances of overfitting. After resolving the issues of outliers (via boxplot technique) and missing values (via mean replacement technique), the final step of preprocessing in our proposed framework is to resolve the problem of imbalanced class. This problem is handled by using the Synthetic Minority Over-sampling Technique (SMOTE), which is based on the Canopy and K-means clustering [34]. SMOTE increases the samples of the minority class by finding its k nearest neighbor. Then one of the k nearest neighbors is selected at random to increase the minority class samples [35]. This procedure can be used for creating as many minority class samples as needed. The samples created using this technique are very much close to the original sample which increases the reliability and reduces randomness. SMOTE has efficiently been used by researchers like D. Yue et al [36].

After performing the data preprocessing, MaLCaDD framework suggests the process of feature selection, as the right subset of features can increase the accuracy of the model

and reduces the computation time and complexity. For feature selection we are using 'Feature importance' technique. Our proposed approach selects the most relevant features based on the p-score of the attribute. Feature importance technique suggests calculation of a score for each attribute of the data; the higher the score the more important or relevant the feature becomes towards the final prediction. Feature importance reduces overfitting by removing the redundant features from the data. Since the final features which are selected are not redundant and misleading so an improved accuracy is achieved. Feature importance has widely been used for selecting the relevant features e.g., Debadri *et al.* [37] selected the top features in his research using this technique and got promising results.

Finally, MaLCaDD proposes an ensemble for classification with boosting technique. The model learns from its previous experience which helps in achieving better results in the future. Our proposed ensemble [56] is based on two models i.e. Logistic Regression and K-Nearest Neighbor. Logistic Regression is used for predicting the categorical dependent variable using a given set of independent variables. It has been used widely in different prediction problems in the field of health care. K-Nearest Neighbor (KNN) performs well on datasets with a greater number of samples [39]. It also gives good results for the numeric attribute. Finally, our framework analyzes the accuracy of the model using k-fold cross-validation. It is technique which gives the performance of the trained model when it is tested on test data. The parameter 'k' decides that in how many folds we need to split the given data [40].

IV. EXPERIMENTAL EVALUATION OF PROPOSED SOLUTION

In this section, we have presented the application of our proposed framework on the selected data set and reported the achieved results. Section A presents the experimental setup. Section B discusses the techniques which are used for handling the missing values and imbalanced data. Section C explains the feature selection and Section D discusses classification of dataset using our proposed ensemble. Finally, Section E presents the results of the different classifiers.

A. EXPERIMENTAL SETUP

This section is further divided into two sub sections. First sub section provides an insight about the dataset which we have used and second one elaborates the tools and techniques which have been used for this research.

1) DATASET

The primary dataset which we have used to demonstrate the applicability of our framework is Framingham [15] data set. Dataset was collected in three different rounds. The first round was conducted in 1948 when the data was collected from 5209 men and women with ages ranging from 30 to 62. The second phase was conducted in 1971 in which 5124 people participated and were asked to go through the same examination process. These were the second generation of the people who patriated in the first round. Finally, in April 2002 the final round was conducted in which the data was collected from the third generation of the original cohort. The Framingham dataset has data from the people of three different generations including participants who originally participated, their children and grandchildren. This makes the dataset very reliable and gives good results when used for prediction. It consists of 4240 samples which contains 644 values from class 1 (Yes) and 3596 from class 2 (No). Here the samples of Class 1 represent the instances of those participants who are suffering from cardiovascular diseases and the instances of Class 2 are those who are not suffering from cardiovascular diseases. This data set has 15 attributes. The attributes in the given dataset provides variety of information including demographic/ behavioral as well as medical history and current medical condition. Our proposed framework uses these attributes for the prediction of CVDs. Table 2 reports the features of the dataset.

TABLE 2. Attributes of Framingham dataset.

Attribute	Values
Sex (Nominal)	Male=1or female=0
Age (continuous)	Age of patient in the whole number
Education (continuous)	Values=1-4 , Some High School=1 , High School or GED=2 , Some College or Vocational School = 3 , college=4
Current Smoker (Nominal)	Yes=1 or No=0
Cigarettes per day (Continuous)	Number of cigarettes smoked per day
BP Meds (Nominal)	Yes=1 or No=0 was BP patient or not
Prevalent Stroke (Nominal)	Yes=1 or No=0 was Stroke patient or not
Prevalent Hyp (Nominal)	Yes=1 or No=0, whether the patient was hypertensive
Diabetes (Nominal)	Yes=1 or No=0 was he a diabetes patient
Tot Chol (Continuous)	Total cholesterol level
Sys BP (Continuous)	Systolic Blood pressure
Dia BP (Continuous)	Diastolic Blood Pressure
BMI (Continuous)	Body Mass Index
Heart Rate (Continuous)	Heart rate or pulse rate
Glucose (Continuous)	Glucose Level
Ten-year CHD (Nominal) (Target Attribute)	Yes=1 No=2, the 10-year risk of coronary Heart Disease (CHD)

2) TOOLS AND TECHNOLOGIES

Exploratory analysis and data processing are performed using Python [41]. Table 3 shows the libraries which are used in this research.

'Pandas' library is used to analyze the data in connection with missing values as well as outliers. 'Seaborn' and 'matplotlib' are used for data visualization and plotting which helps in data preprocessing. For balancing the dataset, we have used the library of 'imblearn'. The selection of

TABLE 3. Libraries used for implementation.

Library	Purpose
Pandas	Data manipulation and analysis [42]
Numpy	High-level mathematical functions
Seaborn	Data visualization and graphs [43]
matplotlib	Plotting
Sklearn	Classification Algorithms
imblearn	SMOTE

relevant features and use of classification algorithms has been made possible using 'sklearn' library. 'Numpy' helps to use high-level mathematical functions. In our research, we have calculated accuracies of the model using this library.

B. PREPROCESSING–EXPERIMENTATION FOR OUTLIERS, MISSING VALUES AND IMBALANCED DATA

This section is further divided into two sub sections. The first sub section reports the removal of outliers and handling of missing values in the dataset whereas, the second one discusses the handling of imbalance class.

1) OUTLIERS AND MISSING VALUES

Outlier is considered as noise in the data and affects the accuracy of the model. We have used Boxplot for the removal of outliers. Boxplot represents five numbers summary i.e., minimum value, first quartile (Q1), median (Q2), third quartile (Q3) and maximum value. When these five points are plotted, a box like graph is plotted and the point that is lying outside this box is considered as an outlier. Fig. 3 shows the boxplot of ''Total cholesterol'' (totChol) attribute.

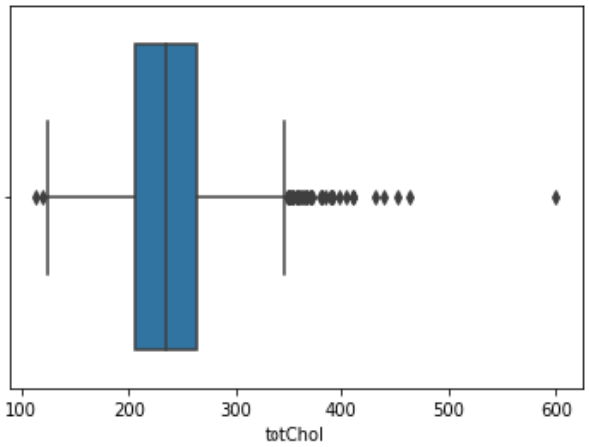

FIGURE 3. Boxplot of "Total Cholesterol" Attribute.

Missing values in the data can exist because of various reasons including the faults of the measuring instrument, human error or inconsistent measuring unit etc. Before training the model, missing values should be handled as it affects the accuracy of the learning algorithm [44]. Therefore, missing values should be handled in such a way that the minimum amount of data should be lost. Framingham dataset also contains missing values which are handled in the preprocessing stage of our proposed framework. Table 4 shows the attributes along with the missing values against each attribute.

TABLE 4. Count of missing values.

Attributes	Count of Missing Values
Male	0
Age	0
Education	105
CurrentSmoker	0
CigsPerDay	29
BPMeds	53
PrevalentStroke	0
PrevalentHyp	0
Diabetes	0
totChol	50
SysBP	0
diaBP	0
BMI	19
Heartrate	1
Glucose	388
TenYearCHD	0

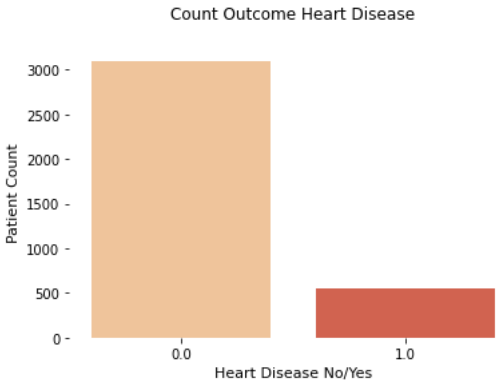

FIGURE 4. Data Imbalance.

As proposed in our framework, all the missing values of specific attribute (in the dataset) are replaced by the mean of all the values of corresponding attribute. For example: the mean of the attribute 'glucose' comes out to be 81.9. It means that all the 388 missing values of the glucose attribute will be replaced by 81.9.

The reason for using the mean substitution is that it increases the samples in our data without adding further information. In this manner, it contributes towards making more informed prediction / decision.

2) DATA IMBALANCE

Class imbalance is a major issue affecting accuracy, which researchers have faced in the various machine learning problems [45]. The problem occurs when the samples of various classes are not equal to one another. Framingham dataset also has the class imbalance problem (Table 4). The number of samples of class 1 in Framingham dataset is 644 while the number of samples of class 2 is 3596, which shows that the dataset is highly imbalanced as shown in Fig. 4.

Our proposed framework deals with this specific issue using 'Synthetic Minority Over-sampling Technique' (SMOTE). Using SMOTE, our proposed framework makes the number of samples of both the classes of Framingham dataset equal.

C. EXPERIMENTATION FOR FEATURE SELECTION

Feature selection has become an important part of machine learning [47] especially for the datasets where there are a large number of features and a greater number of samples. In feature selection, relevant features are selected so that the efficiency of the algorithm increases with the decrease in computational time and complexity. Hence, Feature selection is a very critical step in machine learning and the accuracy of the algorithm is highly dependent on it.

Since, the Framingham dataset contains 15 attributes out of which our proposed framework suggests selection of most relevant features. In our proposed framework, 'feature importance' technique has been used in this regard. The results achieved by using this technique are shown in Fig. 5. Feature importance reduces overfitting by removing the redundant feature from the data. As the data which is selected is not redundant and misleading, so it contributes towards better accuracy. Also, less data is used for training which makes the proposed scheme computationally inexpensive.

Feature importance provides a score for each feature of the data; the higher the score, more important or relevant is the feature towards prediction. Feature importance is an inbuilt class that comes with Tree-Based Classifiers; our proposed framework uses SelectKBest class to extract the most relevant features. The importance of feature is determined with the increase in the node impurity weighted with the probability of reaching that node. The probability of a node can be calculated using (1).

$$Node\ Probability = \frac{No\ of\ samples\ reaching\ that\ node}{Total\ number\ of\ samples} \quad (1)$$

Importance of the feature depends on the value of node probability. Higher the value, more important the feature will be. Features with a score greater than 100 are selected for further computation. The important features selected from the dataset are given in Table 5.

TABLE 5. Selected features.

Attributes	Score
SysBP	727.935535
Age	319.266019
totChol	235.502392
CigsPerDay	209.897040
diaBP	152.748563
SysBP	727.935535

The attribute 'Education' is manually removed from the dataset as it will not contribute much in the prediction of CVDs. After the feature selection, our framework suggests classification via ensemble. The selected features are given as input to the proposed ensemble as discussed in the proceeding sub section.

D. EXPERIMENTATION–CLASSIFICATION VIA ENSEMBLE

Ensemble is the process of combining various machine learning algorithms in order to achieve the results with improved accuracy. Ensemble has been used widely in the field of medical sciences and it has helped a lot in getting better accuracies [48]. Our proposed MaLCaDD Framework classifies using an ensemble of 'Logistic Regression' and 'KNN (K Nearest Neighbor)'. An accuracy of 99.1% has been achieved in prediction using this ensemble. In the ensemble, we have used boosting technique which means that the model learns from its previous errors to make better predictions in the future.

E. EXPERIMENTATION WITH VARIOUS CLASSIFIERS

In order to demonstrate the improved accuracy of our proposed ensemble, we have also experimented with various other classifiers. It is important to highlight that the selected features and all the experimental setup remained the same. Fig. 6 shows a comparison of the results obtained via proposed ensemble and those obtained after applying various other machine learning classifiers on the same selected features of Framingham dataset.

1) EXPERIMENTATION USING LOGISTIC REGRESSION

Logistic Regression is a statistical model that is used as a binary classifier [49], which classifies each sample into two classes (Yes/No). It is used for predicting the categorical dependent variable using a given set of independent variables. An accuracy of 94.2 % has been achieved using Logistic Regression Model on the selected features.

2) EXPERIMENTATION USING KNN

K-Nearest Neighbor (KNN) is one of the simplest parametric algorithms that is included in the supervised learning. In supervised learning, the training data is labeled. When an unseen sample is encountered, the model predicts that sample with the help of trained model. KNN performs well on datasets with a greater number of samples [51]. It also gives good results for the numeric attributes. Value of 'k' is decided and based on that value, distance of k nearest neighbors is taken into consideration. In common practice Euclidean Distance is taken between neighbors. However, Manhattan and Minkowski Distance may also be measured for neighbors. Framingham dataset has 4240 instances which include numeric attributes as well. The value of k is set as 5. The accuracy achieved using this technique is 83.4%.

3) EXPERIMENTATION USING DECISION TREE

Decision Tree is a non–parametric algorithm and is considered a classical machine learning algorithm. It performs well in situations where there is a single attribute that can easily split the data and helps in decision making [50].

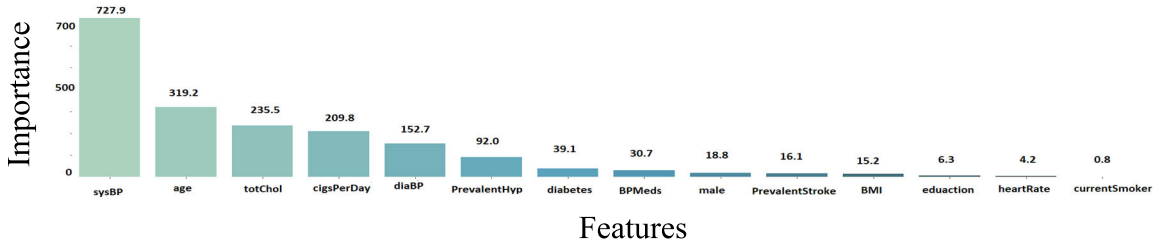

FIGURE 5. Feature Importance.

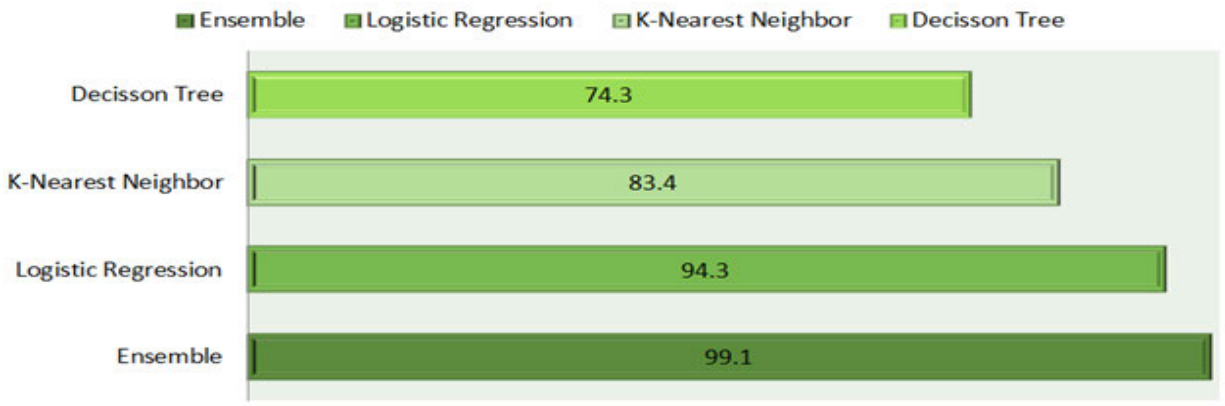

FIGURE 6. Comparison - Results of Various Classifiers.

The challenge in this algorithm is to select the root node. If the root node is selected wisely it reduces the computational complexity of the algorithm and makes it very efficient. The decision tree is easy to understand and the reader can interpret some information from the tree. The accuracy achieved using this model is 74.3%.

V. COMPARATIVE ANALYSIS AND DISCUSSION

In this paper, we have proposed a Machine Learning based Cardiovascular Disease Diagnosis framework i.e. MaLCaDD Framework. This framework can be reliably used for the early diagnosis and prediction of cardiovascular disease due to the reasons of improved accuracy and less computational complexity as well as reduced number of features required for making predictions. Primarily, much improved accuracy is achieved by handling missing values and imbalanced data. Missing values have been replaced with the mean of the all the values of respective attribute whereas the problem of data imbalance is resolved using 'SMOTE'. For efficient features selection, our proposed framework incorporates 'Feature Importance' technique due to which the number of features required for making prediction have been reduced to greater extent. This reduces the computational complexity of the solution. Finally, for increasing the classification accuracy an ensemble of Logistic Regression and KNN with 5-fold cross validation has been suggested. Our proposed framework has the distinction to achieve an accuracy of 99.1 % while classifying Framingham dataset.

In order to demonstrate the wide applicability, MaLCaDD Framework has also been applied on other datasets as well including Heart Disease Dataset [13] and Cleveland Dataset [14]. Heart Disease Dataset [13] is taken from the University of California Irvine (UCI) machine learning repository. This dataset is collected at 4 different locations and it contains 76 attributes. The target attribute shows the presence of disease in case of value 1 and absence of disease in case of value 0. Table 6 shows the attributes of the Heart Disease Dataset.

Many researchers have applied their proposed solutions on this dataset for the prediction of heart diseases and related problems e.g., Jan et. al [52] proposed an approach based on ensemble. In their approach, authors normalized the attributes of the dataset under discussion and removed the missing values. For the purpose of feature reduction, authors used attribute selection or feature sub-setting and selected 13 features. Their proposed ensemble includes five different classifiers i.e., Naïve Bayesian, neural network, SVM, decision tree-based RF algorithm and regression analysis. The overall accuracy which was achieved was 93%. Similarly, Khateeb *et al.* [53] presented a framework in which SMOTE technique is used for handling the imbalance class problem and then feature reduction is performed. For classification, Naïve Bayes is used which has given an accuracy of 79.2%. It is pertinent to mention that our proposed framework i.e. MaLCaDD Framework has also been applied on the same dataset (i.e. Heart Disease Dataset) in which we have used 14 out of 76 attributes (as 14 attributes are related to heart diseases) and the achieved accuracy is 98.0%.

In the similar manner, another dataset which has been classified using our proposed framework is 'Cleveland

TABLE 6. Attributes of heart disease dataset.

Attribute	Values
Sex (Discrete)	Male=1or female=0
Age (continuous)	Age of patient in years
Trestbps (continuous)	Resting blood pressure
Cp (Discrete)	Chest pain type 1= typical angina 2=atypical angina, 3=non-anginal pain 4=asymptomatic
Chol (Continuous)	Serum Cholesterol in mg/dl
Fbs (Discrete)	Fasting blood sugar> 120 mg/dl 1=true, 0=false
Restecg (Discrete)	Resting Electrocardiography results 0= Normal 1= ST-T wave abnormality 2= showing probable or define left ventricular hypertrophy by Estes' criteria
Thalach (Continuous)	Maximum heart rate achieved
Exang (Discrete)	Exercise induced angina 1=yes 0=no
Old peak ST (Continuous)	Depression induced by exercise relative to rest
Slope (Discrete)	The slope of the peak exercise segment 1=up sloping 2= flat 3= down sloping
Ca (Discrete)	Number of major vessels colored by fluoroscopy that ranges between 0 and 3
Thal (Discrete)	3=Normal 6=fixed defect 7=reversible defect
Target Attribute (Discrete)	Yes=1 or No=0

TABLE 7. Attributes of heart cleveland dataset.

Attribute	Values
Id (Discrete)	Unique identifier
Age (Discrete)	Age of patient in days
Gender (Discrete)	Women=1 , men=2
Height (Discrete)	In cm
Weight (Continuous)	In kg
Ap_hi (Discrete)	Systolic blood pressure
Ap_low (Discrete)	Diastolic blood pressure
Cholesterol (Discrete)	1=normal, 2=above normal, 3=well above normal
Gluc (Discrete)	1=normal, 2=above normal, 3=well above normal
smoke (binary)	whether patient smokes or not (yes=1, No=0)
Alcohol (binary)	whether patient drinks or not (yes=1, No=0)
Active (binary)	Physical activity (yes=1, No=0)
Cardio (binary)	Presence or absence of cardiovascular disease (Yes=1 or No=0)

Dataset' [14] taken from the UCI repository. This dataset contains the records of 7000 patients with no missing or null values (both discrete and continuous values are present). The dataset contains total of 13 attributes out of which 12 attributes are independent, and 1 is a dependent attribute. Table 7 shows the chosen attributes from the Cleveland dataset.

In addition to Framingham Dataset, application of our proposed 'MaLCADD' Framework on 'Heart Disease' / 'Clevelands' datasets and the achievement of promising results amply demonstrate that our proposed framework can be reliably used for the classification of cardiovascular diseases.

Now, we will present a comparative analysis of the accuracy achieved via our proposed MaLCaDD Framework over Framingham, Heart Disease and Cleveland datasets and the accuracies reported to have been achieved by various authors over same datasets. Table 8 reports this comparison. In the proposed framework, 5-folds are used for the validation. However, we have computed the results for 3-folds and 10-folds as well. When using 3-folds we achieved the accuracy of 99.0%,98.0% and 95.5% for Framingahm, Heart Disease and Cleveland dataset respectively. Similarly, for 10-folds we achieved the accuracy of 99.1%, 98.0% and 99.5% % for Framingahm, Heart Disease and Cleveland dataset respectively.

It may be noticed from Table 8 that the accuracy achieved by Divya *et al.* [28] is 96.8% for Framingham dataset whereas, the accuracy achieved by our proposed MaLCaDD framework is 99.1 %. Moreover, much reduced number of features (i.e. 5 Features) are used for making such an accurate prediction. With such reduced number of features, certainly the computational complexity has also reduced. To summarize, the proposed framework predictions are more accurate (with reduced set of features) as compared to existing state-of-the-art approaches.

A. DISCUSSION AND THEORETICAL ANALYSIS

So far, we have done a comparative analysis of the techniques which we have used for handling missing values, data imbalance as well as feature selection with the various techniques that are currently in vogue for the specific purposes. This justifies the use of these techniques as part of our framework. For example, our proposed framework handles the missing values by replacing them with the mean of all the values of the corresponding attribute. However, various other techniques as reported by Kang *et al.* [55] also exist for handling missing values. To counter verify our results we have applied those methods and reported the results / accuracy in Table 9. As may be noticed that the accuracy achieved using mean substitution is the highest.

Similarly, the problem of data imbalance is resolved using the SMOTE technique as discussed in section IV (B). However, various other methods also exist to deal this problem [34]. To counter verify our result we have applied those techniques and analyzed the results. It can be solved by increasing the samples of the minority class also known as oversampling. The samples are added randomly which makes the data unreliable and redundant. Similarly, by reducing the samples of the majority class, the classes can be balanced. This is known as under-sampling in which some of the data might be lost. The comparison is reported in Table 10. It can be seen that the accuracy of 99.1% is achieved when SMOTE is used for balancing the class which makes it a reliable solution for the given problem. Similarly, the important features

TABLE 8. Comparison of MaLCaDD framework on different datasets.

Author	Year	Data set	Data Imbalance	Feature selection	Features	Classifier	Validation type	Accuracy Achieved
Framingham Dataset [15]								
Rubini [18]	2019	Framingham	Under-sampling	Automatic Feature Selection	10	Random Forest	3-fold	84.8%
Divya [28]	2019	Framingham	Over sampling	Automatic Feature Selection	13	Random Forest	3-fold	96.8%
Nitten [17]	2018	Framingham	No	All features were used	22	SVM	-	90.2%
Hoda [19]	2017	Framingham	No	Feature importance	9	KNN	-	66.7%
Shen [27]	2014	Framingham	-	-	10	Clustered CART Framework	-	-
MaLCaDD Framework (Proposed)	**2021**	**Framingham**	**SMOTE**	**Feature importance**	**5**	**Ensemble**	**3-fold**	**99.0%**
							5-fold	**99.1%**
							10-fold	**99.1 %**
Heart Disease Dataset [13]								
Jan et. al [52]	2018	Heart Disease	-	Attribute Selection	13	Ensemble	10-fold	95.5%
Khateeb et al. [53]	2017	Heart Disease	SMOTE	Feature reduction	14	Naïve Bayes	-	79.2%
MaLCaDD Framework (Proposed)	**2021**	**Heart Disease**	**SMOTE**	**Feature importance**	**5**	**Ensemble**	**3-fold**	**98.0%**
							5-fold	**98.0%**
							10-fold	**98.0%**
Cleveland Dataset [14]								
Tanvi et. al [22]	2017	Cleveland	-	-	All features	Decision tree	-	93.2%
Tama et al. [54]	2020	Cleveland	-	Correlation based	10	Ensemble	10-fold	88.0%
MaLCaDD Framework (Proposed)	**2021**	**Cleveland**	**SMOTE**	**Feature importance**	**5**	**Ensemble**	**3-fold**	**95.5%**
							5-fold	**95.5%**
							10-fold	**95.5%**

are selected from the dataset using the feature importance technique as discussed in section IV (C).

There are some other methods as well which are used for selecting the features of interest. To counter verify our results achieved using feature importance technique, we have applied those techniques as well. One of them is the correlation-based feature selection. In correlation-based feature selection, every attribute is ranked according to the Heuristic Evaluation Function based on correlation.

After these features are selected, they are given as an input to different classification models (Random Forest, SVM, Logistic Regression, Decision Tree and KNN). Different parameters are used in the training of the models. In case of SVM, two parameters are used: kernel and c. The kernel parameter is used to transform the data into required form by using mathematical functions. Similarly, c is the ''penalty parameter''. In our case, it is not required to transform the data into some other domain, therefore, we have used ''Linear'' kernel with the value of c=10000. Secondly, in case of Random forest, it includes parameters: N_estimators = 100 (which tells the total number of trees in the forest), M_features = 5 (which tells the maximum number of features that is required for splitting a node) and Bootstrap was carried out using sampling with replacement. Table 11 reports a comparison of the achieved results. The best result are achieved using ''Ensemble''.

TABLE 9. Handling missing values.

Missing Values	Classifier	Accuracy
Remove	Logistic Regression	90.0%
Median	Logistic Regression	92.1%
Mean	Logistic Regression	95.0%
Remove	KNN	83.4%
Median	KNN	85.0%
Mean	KNN	86.6%
Remove	Ensemble	95.5%
Median	Ensemble	93.2%
Mean	**Ensemble**	**99.1%**

The datasets in this study give better results using a non-parametric model. For example, K-NN is a memory-based non-parametric approach. It immediately adapts on new training data. It allows to respond quickly to real-time changes in the input. On the other hand, K-NN has few disadvantages as well, for example, the efficiency or speed of algorithm declines as dataset grows. However, in this study, we have limited datasets. Similarly, K-NN algorithm struggles to

TABLE 10. Handling imbalance data.

Data Imbalance	Feature Selection	Accuracy
Oversampling	All Features	90.5%
Oversampling	Correlation-based	89.5%
Oversampling	Feature Importance	91.0%
Under-sampling	All Features	90.4%
Under-sampling	Feature Importance	88.0%
Under-sampling	Feature Importance	89.0%
SMOTE	All Features	91.5%
SMOTE	Correlation-based	94.6%
SMOTE	**Feature Importance**	**99.1%**

TABLE 11. Comparison of different models.

Model	Accuracy
Random Forest	82.0%
SVM	79.0%
Logistic Regression	94.2%
Decision Tree	74.3%
KNN	83.4%
Ensemble	**99.1%**

predict the output of new data point as number of variables grows, however, we have small number of input variables in this study. Therefore, the selection of K-NN algorithm is appropriate in the proposed framework. In contrast to K-NN, PAM (Partition Around Medoids) is not sensitive to outliers. Moreover, it is more complex than K-NN. Furthermore, a good preprocessed data is available in this study. Therefore, KNN algorithm is more suitable in the proposed framework as compared to PAM.

The non-parametric statistics are important to analyze data distribution characteristics for the efficacy of results. For example, on comparing two independent samples, when the outcome is not normally distributed and the samples are small, a nonparametric test is appropriate. We have non-normal distribution of the important features which are selected for the training of model in the proposed framework. For example, the feature like ''number of cigarettes per day'' is not normally distributed as most people are not smokers. Similar is the case with ''glucose'' feature. In this research, the samples are small as well as the outcome is an ordinal variable or a rank. Therefore, the effectiveness of proposed framework is further confirmed by performing Mann Whitney U test.

The computational complexity is highly important especially in case of large datasets. The proposed framework consists of three stages including i) Data Pre-Processing ii) Feature Selection and iii) Classification. In Data Preprocessing, outliers in the data are detected using box plot which is one of the most common and accurate way of detecting the outliers in case of symmetric data. The Framingham dataset was symmetric that's why we have used this method for the detection of outliers. The outliers are removed from the data and data becomes noise free. If a point is outside the certain threshold which is defined by Inter Quartile Range (IQR) than it is considered as an outlier. This process is similar to the linear search. So the Big O notation for this step is O (n) as it includes searching for the point/element under a certain threshold value. In the next step of data preprocessing, missing values are handled using mean substitution. This step involves the average/mean function so it have linear complexity and Big O notation is O (n). In the last step of data Pre-Processing, data is balanced using Synthetic Minority Oversampling Technique (SMOTE) to make samples of both classes equal. In this KNN is implemented, so the Big O notation for this step is O (np). In Feature Selection, important and relevant features are selected using Feature Importance which is an inbuilt class that comes with Tree-Based Classifiers; and Big O notation for this step is O (p). In the classification phase, we have proposed an ensemble, in which we have used Logistic Regression and K-Nearest Neighbor. The Big O notation for Logistic Regression is O (p) and for K-Nearest Neighbor is O (np) where n=number of training samples and p= number of features. To conclude, the complexity of proposed framework is acceptable in all three major phases i.e. Data Pre-Processing, Feature Selection and Classification.

VI. CONCLUSION AND FUTURE WORK

This article presents MaLCaDD (**Ma**chine **L**earning based **Ca**rdiovascular **D**isease **D**iagnosis) framework for the early prediction and diagnosis of cardiovascular diseases. The framework is based on four major phases where first phase deals with the handling of missing values via mean replacement technique. In second phase, data imbalance issue is resolved via Synthetic Minority Over-sampling Technique (SMOTE). In third phase, feature selection is performed using feature importance technique. Finally, ensemble of Logistic Regression (LR) and K-Nearest Neighbor (KNN) is proposed for improved prediction. The implementation of MaLCaDD is carried out in Python and it is publically available at GitHub repository. The validation of framework is performed through three benchmark datasets (i.e. Framingham, Heart Disease and Cleveland) and the accuracies of 99.1%, 98.0% and 95.5 % are achieved respectively. The comparative analysis proves that MaLCaDD outperforms the state-of-the-art studies by achieving the improved accuracy with reduced set of features. Therefore, MaLCaDD is highly reliable and can be applied in real environment for the early diagnosis of cardiovascular diseases effectively.

The preprocessing steps in our proposed framework increases the reliability of data which includes outlier detection to remove the noise in the data followed by handling the missing values in the data. After that, data is balanced to avoid overfitting or underfitting of the model. The feature selection step reduces the computational complexity of the model. All these steps together improves the classification accuracy of the algorithm. The validity of proposed framework on three benchmark datasets demonstrates that our framework is an innovative and reliable framework. On one hand it combines the innovative pre-processing and feature selection

steps and on the other hand it applies an innovative ensemble. MaLCaDD Framework not only achieves higher accuracy but can be also be reliably applied on wide variety of datasets for prediction of cardiovascular diseases.

Currently, the evaluation of MaLCaDD is performed through state-of-the-art datasets. We are now working with different hospitals to assess the applicability of MaLCaDD in real environment. In this regard, we intend to share the real time evaluation results of MaLCaDD on different patients in future.

REFERENCES

[1] S. Ambekar and R. Phalnikar, "Disease risk prediction by using convolutional neural network," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–5.

[2] *Health Stats 2017 by World Health Organization (WHO)*. Accessed: Mar. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(CVDs)

[3] D. Mousa, N. Zayed, and I. A. Yassine, "Automatic cardiac MRI localization method," in *Proc. Cairo Int. Biomed. Eng. Conf. (CIBEC)*, Giza, Egypt, Dec. 2014, pp. 153–157.

[4] L.-N. Pu, Z. Zhao, and Y.-T. Zhang, "Investigation on cardiovascular risk prediction using genetic information," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 5, pp. 795–808, Sep. 2012.

[5] V. Gil-Guillen, D. Orozco-Beltran, A. Maiques-Galan, J. Aznar-Vicente, J. Navarro, L. Cea-Calvo, F. Quirze-Andrés, J. Redon, and J. Merino-Sanchez, "Agreement between REGICOR and SCORE scales in identifying high cardiovascular risk in the Spanish population," *Revista Espanola de Cardiol.*, vol. 60, no. 10, p. 1042, 2007.

[6] A. Khandoker, Y. Al Zaabi, and H. Jelinek, "What can tone and entropy tell us about risk of cardiovascular diseases?" in *Proc. Comput. Cardiology Conf. (CinC)*, Dec. 2019, pp. 1–4.

[7] G. Assmann, P. Cullen, and H. Schulte, "Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study," *Circulation*, vol. 105, no. 3, pp. 310–315, 2002.

[8] S. K. Jain and B. Bhaumik, "An ultra low power ECG signal processor design for cardiovascular disease detection," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 857–860.

[9] C. Martin-Isla, V. M. Campello, C. Izquierdo, Z. Raisi-Estabragh, B. Baeßler, S. E. Petersen, and K. Lekadir, "Image-based cardiac diagnosis with machine learning: A review," *Frontiers Cardiovascular Med.*, vol. 7, p. 1, Jan. 2020.

[10] R. K. Sevakula, W. T. M. Au-Yeung, J. P. Singh, E. K. Heist, E. M. Isselbacher, and A. A. Armoundas, "State-of-the-Art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system," *J. Amer. Heart Assoc.*, vol. 9, no. 4, 2020, Art. no. e013924.

[11] R. Shadmi, V. Mazo, O. Bregman-Amitai, and E. Elnekave, "Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 24–28.

[12] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," in *Proc. 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. (ICEEICT)*, Sep. 2016, pp. 1–5.

[13] *Heart Disease Dataset by UCI*. Accessed: Oct. 25, 2020. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[14] *Cleveland Dataset by KEEL*. Accessed: Nov. 15, 2020. [Online]. Available: https://sci2s.ugr.es/keel/dataset.php?cod=57

[15] *Framingham Dataset by Kaggle*. Accessed: Nov. 20, 2020. [Online]. Available: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

[16] *Cardiovascular Disease by Kaggle*. Accessed: Oct. 15, 2020. [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[17] N. S. Rajliwall, R. Davey, and G. Chetty, "Machine learning based models for cardiovascular risk prediction," in *Proc. Int. Conf. Mach. Learn. Data Eng. (iCMLDE)*, Dec. 2018, pp. 142–148.

[18] P. E. Rubini, C. A. Subasini, A. V. Katharine, V. Kumaresan, S. G. Kumar, and T. M. Nithya, "A cardiovascular disease prediction using machine learning algorithms," *Ann. Romanian Soc. Cell Biol.*, vol. 25, no. 2, pp. 904–912, 2021. [Online]. Available: https://www.annalsofrscb.ro/index.php/journal/article/view/1040

[19] H. A. G. Elsayed and L. Syed, "An automatic early risk classification of hard coronary heart diseases using Framingham scoring model," in *Proc. 2nd Int. Conf. Internet Things, Data Cloud Comput.*, Mar. 2017, pp. 1–8.

[20] E. D. Frohlich and P. J. Quinlan, "Coronary heart disease risk factors: Public impact of initial and later-announced risks," *Ochsner J.*, vol. 14, no. 4, p. 532, 2014.

[21] R. Hajar, "Risk factors for coronary artery disease: Historical perspectives," *Heart Views*, vol. 18, no. 3, p. 109, 2017.

[22] T. Sharma, S. Verma, and Kavita, "Prediction of heart disease using cleveland dataset: A machine learning approach," *Int. J. Recent Res. Aspects*, vol. 4, no. 3, pp. 17–21, 2017.

[23] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *Proc. 4th Int. Conf. Comput. Commun. Autom. (ICCCA)*, Dec. 2018, pp. 1–4.

[24] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning: An experimental analysis," in *Proc. 3rd Int. Conf. Deep Learn. Technol.*, 2019, pp. 51–56.

[25] I. K. A. Enriko, M. Suryanegara, and D. Gunawan, "Heart disease diagnosis system with K-nearest neighbors method using real clinical medical records," in *Proc. 4th Int. Conf. Frontiers Educ. Technol.*, 2018, pp. 127–131.

[26] E. B. Randa, "An ensemble model for Heart disease data sets: A generalized model," in *Proc. 10th Int. Conf. Inform. Syst.*, 2016, pp. 191–196.

[27] S. Song, J. Warren, and P. Riddle, "Developing high risk clusters for chronic disease events with classification association rule mining," in *Proc. 7th Australas. Workshop Health Inform. Knowl. Manage.*, vol. 153, 2014, pp. 69–78.

[28] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," in *Proc. IEEE Region 10 Conf. (TENCON)*, Oct. 2019, pp. 367–372.

[29] *Framingham Heart Study by National Heart, Lung, and Blood Institute (NIH)*. Accessed: Nov. 2020. [Online]. Available: https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs

[30] M. D. A. Pranatha, N. Pramaita, M. Sudarma, and I. M. O. Widyantara, "Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting," in *Proc. 4th Int. Conf. Wireless Telematics (ICWT)*, Jul. 2018, pp. 1–4.

[31] L. S. Reddy. D, D. B. Raveendrababu, and D. A. Govardhan, "A model for improving classifier accuracy using outlier analysis," *Int. J. Comput. Technol.*, vol. 7, no. 1, pp. 500–509, May 2013.

[32] C. R. Padgett, C. E. Skilbeck, and M. J. Summers, "Missing data: The importance and impact of missing data from clinical research," *Brain Impairment*, vol. 15, no. 1, pp. 1–9, May 2014.

[33] H. M. Dodeen, "Effectiveness of valid mean substitution in treating missing data in attitude assessment," *Assessment Eval. Higher Educ.*, vol. 28, no. 5, pp. 505–513, Oct. 2003.

[34] H. Lee, S. Jung, M. Kim, and S. Kim, "Synthetic minority over-sampling technique based on fuzzy c-means clustering for imbalanced data," in *Proc. Int. Conf. Fuzzy Theory Appl. (iFUZZY)*, Nov. 2017, pp. 1–6.

[35] C. Guo, Y. Ma, Z. Xu, M. Cao, and Q. Yao, "An improved oversampling method for imbalanced data–SMOTE based on Canopy and K-means," in *Proc. Chin. Automat. Congr. (CAC)*, Nov. 2019, pp. 1467–1469.

[36] Y. Ge, D. Yue, and L. Chen, "Prediction of wind turbine blades icing based on MBK-SMOTE and random forest in imbalanced data set," in *Proc. IEEE Conf. Energy Internet Energy Syst. Integr. (EI2)*, Nov. 2017, pp. 1–6.

[37] D. Dutta, D. Paul, and P. Ghosh, "Analysing feature importances for diabetes prediction using machine learning," in *Proc. IEEE 9th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Vancouver, BC, Canada, Nov. 2018, pp. 924–928.

[38] J. A. Quesada, A. Lopez-Pineda, V. F. Gil-Guillén, R. Durazo-Arvizu, D. Orozco-Beltrán, A. López-Domenech, and C. Carratalá-Munuera, "Machine learning to predict cardiovascular risk," *Int. J. Clin. Pract.*, vol. 73, no. 10, Oct. 2019, Art. no. e13389.

[39] J. Huang, Y. Wei, J. Yi, and M. Liu, "An improved kNN based on class contribution and feature weighting," in *Proc. 10th Int. Conf. Measuring Technol. Mechatronics Autom. (ICMTMA)*, Changsha, China, Feb. 2018, pp. 313–316.

[40] F. Y. Ahmed, Y. H. Ali, and S. M. Shamsuddin, "Using K-fold cross validation proposed models for spikeprop learning enhancements," *Int. J. Eng. Technol.*, vol. 7, nos. 4–11, pp. 145–151, 2018.

[41] *Python*. Accessed: Feb. 2021. [Online]. Available: https://colab.research.google.com/notebooks/intro.ipynb

[42] N. Pandey, P. K. Patnaik, and S. Gupta, "Data pre-processing for machine learning models using Python libraries," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 1995–1999, 2020.
[43] A. Oberoi and R. Chauhan, "Visualizing data using Matplotlib and Seaborn libraries in Python for data science," *Int. J. Sci. Res. Publication*, vol. 9, no. 3, p. 8733, 2019.
[44] A. R. Kapil, "Methods of missing value treatment and their effect on the accuracy of classification models," Tech. Rep., 2018, doi: 10.13140/RG.2.2.31137.86881.
[45] K. R. Weiss and T. M. Khoshgoftaar, "Comparing transfer learning and traditional learning under domain class imbalance," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Cancun, Mexico, Dec. 2017, pp. 337–343.
[46] X.-Y. Liu, S.-T. Wang, and M.-L. Zhang, "Transfer synthetic over-sampling for class-imbalance learning with limited minority class data," *Frontiers Comput. Sci.*, vol. 13, no. 5, pp. 996–1009, Oct. 2019.
[47] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, J. E. Gubernatis, and T. Lookman, "Importance of feature selection in machine learning and adaptive design for materials," in *Materials Discovery and Design*. Cham, Switzerland: Springer, 2018, pp. 59–79.
[48] X. Y. Gao, A. Amin Ali, H. H. Shaban Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, Feb. 2021, Art. no. 6663455.
[49] J. Klimaszewski, M. Sklyar, and M. Korzeń, "Learning ℓ^1-penalized logistic regressions with smooth approximation," in *Proc. IEEE Int. Conf. Innov. Intell. Syst. Appl. (INISTA)*, Gdynia, Poland, Jul. 2017, pp. 126–130.
[50] P. Subarkah, A. N. Ikhsan, and A. Setyanto, "The effect of the number of attributes on the selection of study program using classification and regression trees algorithms," in *Proc. 3rd Int. Conf. Inf. Technol., Inf. Syst. Electr. Eng. (ICITISEE)*, Yogyakarta, Indonesia, Nov. 2018, pp. 1–5.
[51] C. Isaksson and M. H. Dunham, "A comparative study of outlier detection algorithms," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit.* Berlin, Germany: Springer, 2009, pp. 440–453.
[52] M. Jan, A. A. Awan, M. S. Khalid, and S. Nisar, "Ensemble approach for developing a smart heart disease prediction system using classification algorithms," *Res. Rep. Clin. Cardiol.*, vol. 9, pp. 33–45, Dec. 2018.
[53] N. Khateeb and M. Usman, "Efficient heart disease prediction system using K-nearest neighbor classification technique," in *Proc. Int. Conf. Big Data Internet Thing (BDIOT)*, 2017, pp. 21–26.
[54] B. A. Tama, S. Im, and S. Lee, "Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble," *BioMed Res. Int.*, vol. 2020, Apr. 2020, Art. no. 9816142.
[55] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, p. 402, 2013.
[56] *MalCaDD Framework's Code and User Manual*. Accessed: Jan. 2, 2021. [Online]. Available: https://github.com/20-8/Machine-Learning

AQSA RAHIM received the bachelor's degree in computer engineering from the Department of Computer and Software Engineering, National University of Sciences and Technology, Islamabad, Pakistan, in 2019, where she is currently pursuing the master's degree in software engineering. Her research interests include data engineering, signal processing, machine learning, and deep learning.

YAWAR RASHEED received the bachelor's degree in computer engineering from the Department of Computer Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2007, and the master's degree in computer software engineering from the Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, NUST, in 2020. His major research interests include model driven software engineering, model driven architecture, and software requirements engineering.

FAROOQUE AZAM is currently a Key Faculty Member with the Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering (CEME), NUST, Pakistan. He has been involved in post graduate teaching and research, since 2007. His research interests include model driven software engineering, model driven testing, model driven embedded applications, model driven web engineering, and software design and architectures.

MUHAMMAD WASEEM ANWAR received the Ph.D. degree in software engineering from the Department of Computer and Software Engineering, CEME, National University of Sciences and Technology, Pakistan. He is a Senior Researcher and an Industry Practitioner in the field of model based system engineering (MBSE) for embedded and control systems. He is associated with Model Driven Software Engineering (MDSE) and Artificial Intelligence (AI) research groups, CEME, NUST. His major research interest includes model based system engineering (MBSE) for complex and large systems.

MUHAMMAD ABDUL RAHIM received the bachelor's degree from the University of Engineering and Technology, Peshawar, Pakistan, in 1991, and the master's degree in electrical engineering from the Department of Electrical Engineering, National University of Sciences and Technology, Islamabad, Pakistan, in 2010. He is associated with Artificial Intelligence (AI) research group, CEME, NUST. His research interests include control system engineering and electronics communication.

ABDUL WAHAB MUZAFFAR received the Ph.D. degree in software engineering from the National University of Sciences and Technology (NUST) Islamabad, Pakistan, in 2017. He is currently an Assistant Professor with Saudi Electronic University, Saudi Arabia. His research interests include model-driven software engineering, data and text mining, bio informatics, and machine learning.