INTERNATIONAL JOURNAL · ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

IJIEMR Transactions, online available on 28th May 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue= Spl Issue 04

DOI: 10.48047/IJIEMR/V11/SPL ISSUE 04/16

Title **An Improved Machine Learning Framework for Cardiovascular Disease Prediction**

Volume 11, SPL ISSUE 04, Pages: 141-150

Paper Authors

Sarath Chandra Banala, V. Sowjanya

An Improved Machine Learning Framework for Cardiovascular Disease Prediction

[1]Sarath Chandra Banala, [2]V. Sowjanya
[1,2]Department of Computer Science and Engineering,
PSCMR College of Engineering & Technology, Vijayawada, Andhra Pradesh, India
Email: sarathit862@gmail.com, sowjanyav@pscmr.ac.in

Abstract:
Cardiovascular illnesses are the most life-threatening syndromes in the world, with the greatest fatality rate. They have become quite common throughout time, and they are now overstretching the healthcare systems of countries. High blood pressure, family history, stress, age, gender, cholesterol, BMI, and an unhealthy lifestyle are all key risk factors for cardiovascular disease. Researchers have proposed numerous ways for early diagnosis based on these criteria. However, due to the inherent criticality and life-threatening dangers of cardiovascular disorders, the accuracy of offered procedures and approaches requires some refinement. A Machine Learning based Cardiovascular Disease Diagnosis framework is proposed in this paper for the accurate prediction of cardiovascular disorders. The approach mean replacement technique and Synthetic Minority Over-sampling Technique (SMOTE). Following that, the Feature Importance approach is used to choose features. Finally, for improved prediction accuracy, an ensemble of Logistic Regression and K-Nearest Neighbour (KNN) classifiers is presented. The benchmark datasets are used to validate the framework for achieving better accuracy. Finally, a comparison of framework to existing state-of-the-art approaches with more accurate than existing state-of-the-art approaches.

Keywords: Machine Learning; ensemble; SMOTE; cardiovascular prediction; feature importance.

I. INTRODUCTION

The current era's hectic pace leads to an unhealthy lifestyle that generates anxiety and despair. To cope with these conditions, people often turn to excessive smoking, drinking, and drug use. All of these factors play a role in the development of a variety of severe diseases, such as cardiovascular disease and cancer. Early detection of these disorders is critical so that preventative steps can be implemented before something terrible occurs. A disorder that damages the heart and the blood vessels is referred to as cardiovascular disease Coronary heart disease, stroke/transient ischemic attack (TIA/ Mini-stroke), peripheral arterial disease, and aortic disease are the four main kinds of CVDs [1].

CVDs are associated with hypertension, smoke, diabetic, body mass (Bms), fat, age, and family history. For several people, these factors are different. CVDs are brought on by a number of factors, like youth, genetics, strain, as well as an addictive personality. The main problem is to accurately forecast these diseases in a timely manner so that mortality can be decreased through appropriate medicine and other countermeasures.

Using various datasets and approaches, researchers have suggested multiple algorithms in order to anticipate cardiovascular diseases. Heart disease [2], Cleve- land [3], Framingham [4], and cardiovascular disease [5] are some of the most commonly used datasets for CVD prediction. These datasets are made up of several attributes that are used to predict CVDs. The factors that lead to cardiovascular disease include both changeable and non-changeable risks. The Framingham dataset [4], which is gathered against these parameters, is one of the most well-known datasets. This data collection has been utilised by several researchers to validate their prediction models. In the given study context, various ML and DL based algorithms for the diagnosis of cardiovascular disease were developed. Researchers, on the other hand, are the focus is on feature selection approaches and classification algorithms, with the issue of class imbalance being overlooked. The problem of class imbalance has a significant impact on the classification algorithm's accuracy. Furthermore, when data is not balanced, during prediction, a large number of characteristics are required. This considerably increases the computing complexity of the solution, rendering it unsuitable for use in a real-world setting. Furthermore, existing feature selection algorithms must be improved in order to reduce computing while keeping a reasonable level of accuracy. In the same way, existing classifier results need to be enhanced in order to produce reliable findings.

To summarise, a cardiovascular ML framework with unified machine learning illnesses is desperately needed, Data balancing, feature selection, and classification enhancement are all done in a systematic way. Fig1 shows a high-level representation of the framework. In contrast to previous research, which has mostly absorbed on feature selection and standard classification approaches, by addressing missing values and data, the framework tries to boost overall correctness that is imbalanced**.**

Missing values were handled by replacing the missing value with the average of all the values of a relevant characteristic. Framework proposes Synthetic Minority Over-sampling Technique (SMOTE) to deal with data imbalance.

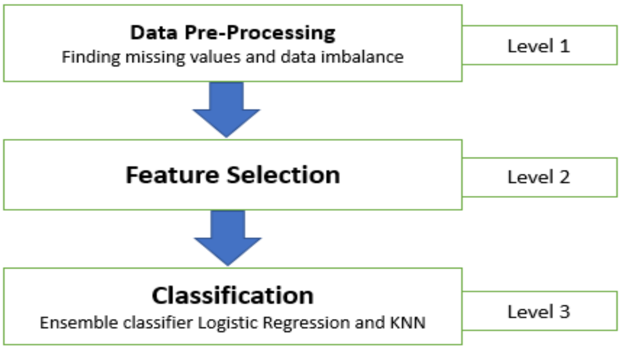
Data Pre-Processing
Finding missing values and data imbalance
Level 1
Feature Selection
Level 2
Classification
Ensemble classifier Logistic Regression and KNN
Level 3

Fig 1. Framework for Cardiovascular Diagnosis Based on ML.

When the data is balanced, To select the best set of features, the framework employs the feature importance technique. Finally, An ensemble of Logistic Regression and K-Nearest Neighbor (KNN) models is presented for enhanced prediction. Three benchmark dataset Framingham are used to validate the framework.

II. LITERATURE REVIEW

The heart and blood arteries are affected by cardiovascular diseases. Cardiovascular diseases come in a variety of forms and can have a wide range of effects on the human body. High blood pressure, smoking, diabetes, body mass index (BMI), cholesterol, age, family history, and other risk factors have been identified as potential causes of CVDs [10]. CVDs are caused by a variety of variables, including age, gender, stress, and an unhealthy lifestyle [9].

In a word, whatever the reason, the most essential thing is to detect cardiovascular diseases as soon as possible. Similarly, the researchers have employed a variety of datasets to validate their proposed methodologies. Cleveland [3], Framingham [4], and the heart disease dataset [2] have all received a lot of attention.

The properties of these datasets are mostly the same. The experimental setup, or how the dataset is acquired, is the major variation in these datasets. The research on CVD prediction that was done by deserving academics, as well as the necessary datasets, are presented in the following paragraphs.

The Cleveland dataset was utilised by Tanvi et al. [10] to predict heart disorders. The model was trained using 14 characteristics as part of the prediction process. From all these models, Decision Tree has had the best accuracy.

On the cardiac disease dataset [2] accessible in the UCI repository, Singh et al. [11] used various categorization techniques. The use of Logistic Regression, which was the most accurate of all the other models, yielded 87.1 percent accuracy. Amanda et al [12] used ten distinct features from a South African dataset on cardiac disease. The dataset is subjected to three distinct models: Decision Tree, Nave Bayes, and SVM, which are then evaluated using the Confusion Matrix. Naive Bayes generated the best outcomes of the three models.

The dataset utilised by Ketut et al. [13] was obtained from Harapan Kita Hospital. For the prediction of cardiac disorders, 18 factors were extracted from this dataset. The accuracy of KNN with and without parameter weighting was 75.11 percent and 74.0 percent, respectively, in this investigation.

For the forecast of disease, researchers have utilised a variety of approaches and procedures using the Framingham dataset. The Random Forest algorithm was proposed by Rubini et al. [6] for the prediction of cardiac disease.

Hoda et al. [7] proposed utilising KNN and Random Forest to predict CVDs. KNN and Random Forest were utilised to classify the data in the suggested method. The accuracy of KNN was found to be 66.7 percent, whereas the Random Forest method was found to be 63.4 percent accurate. We are employing the Framingham dataset in our study with necessary variables to aid in the prediction of these diseases.

There are 16 features in the dataset that can be used to make a prediction. The Framingham dataset contains information from three generations of people, including those who took part in the original study.

III. PROPOSED SOLUTION

For the prediction of cardiovascular disorders, a unique ML based Cardiovascular Disease Diagnosis Framework was presented. The purpose of this research is to develop a ML model that can correctly identify cardiovascular illnesses from patient clinical data. The steps in the proposed strategy are as follows: 1) Outlier elimination, missing value replacement, and data imbalance class handling are all part of data preparation Using SMOTE, 2) feature importance technique is used for feature selection, 3) combination of logistic regression and KNN is used for ensemble classification. The trained model is then used to make predictions. Our primary goal is to get good outcomes with limited features and less computing complexity. Fig 2 shows framework will be presented now.

Data preparation is done first in the proposed method. The data is largely searched for probable outliers during the data pre-processing step.

The majority of outliers are considered noise, which has no effect on the data's significance and hurts the model's performance. If the data contains missing values, the model cannot be trained on such data because to the restricted amount of training samples. This has an impact on the model's accuracy. As a result, in our proposed framework, missing values are replaced by the average of all values of the associated property.

This allows for the retention of training data without introducing additional data to the dataset, reducing the risk of overfitting. After resolving the problems of outliers and missing values using the mean replacement technique, the next phase of pre-processing in our proposed framework is to handle the problem of imbalanced class using SMOTE Technique and K- Means clustering algorithms.

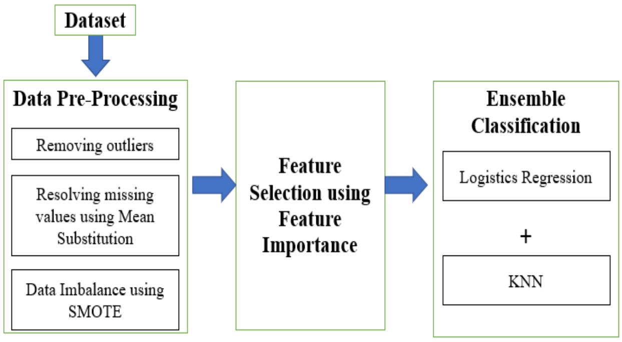
Dataset
Data Pre-Processing
Removing outliers
Resolving missing values using Mean Substitution
Data Imbalance using SMOTE
Feature Selection using Feature Importance
Ensemble Classification
Logistics Regression
+
KNN

Fig 2. Flowchart of the Cardiovascular Diagnosis Perdition Framework based on ML.

By locating the minority class's k nearest neighbour, SMOTE improves the sample size of the minority class. Then, at random, one of the k closest neighbours is chosen to boost the smaller class samples. This approach can generate as many smaller class samples as needed.

This method generates samples that are very identical to the original, enhancing reliability and lowering randomness. Finally, the framework recommends an ensemble for classification utilising the boosting technique.

The model learns from its prior experiences and can therefore deliver desirable results. Our proposed ensemble is based on two models: logistic regression and K-nearest neighbour. Using logistic regression, a set of independent factors is used to predict the categorical dependent variable. It's been used to solve a range of prediction problems in the medical industry. K-Nearest Neighbor (KNN) is a good choice for datasets with a large number of samples. It also works well with the attribute number. Finally, using k-fold cross-validation, our framework assesses the model's correctness. It's a method for determining the performance of a trained model when it's put to the test with real data. The 'k' option determines how many folds the data must be split into.

IV. EVALUATION OF PROPOSED SOLUTION

1) DATASET

We've used the Framingham [15] data set to demonstrate the framework's applicability. Data collection was done in three phases. This round, that took place in 1948, gathered information from 5209 participants aged 30 to 62. In 1971, 5124 people took place in the second stage, and they're all given the same examination. Those are the children of the first-round draftees. Finally, data from the first cohort's 3rd generation was obtained.

Table1 Attributes and its type of Dataset

Attribute	Type
Sex	Nominal
Age	Continuous
Education	Continuous
Current Smoker	Nominal
Cigarettes per day	Continuous
BP Meds	Nominal
Prevalent Stroke	Nominal
Prevalent Hyp	Nominal
Diabetes	Nominal
Total Cholo	Continuous
Sys BP	Continuous
Dia BP	Continuous
BMI	Continuous
Heart Rate	Continuous
Glusoce	Continuous
Ten-year CHD	Nominal

The Framingham dataset contains information from three generations of people. Participants with cardiovascular diseases are represented by Class 1 samples, whereas those without cardiovascular disorders are represented by Class 2 samples. Table1 shows 16 attributes in this data collection.

2) FINDING OUTLIERS, MISSING VALUES AND IMBALANCED DATA

Outliers are regarded noise in the data and have an impact on the model's accuracy. For the reduction of outliers, we used Boxplot. Boxplots depict the smallest value, first quadrant (Q1), second quadrant (Q2), third quadrant (Q3), and highest value. When these have points are plotted, they form a box-like graph, and any point that lies outside of this box is called an outlier. The boxplot of the "Total cholesterol" (totChol) characteristic is shown in Fig. 3.

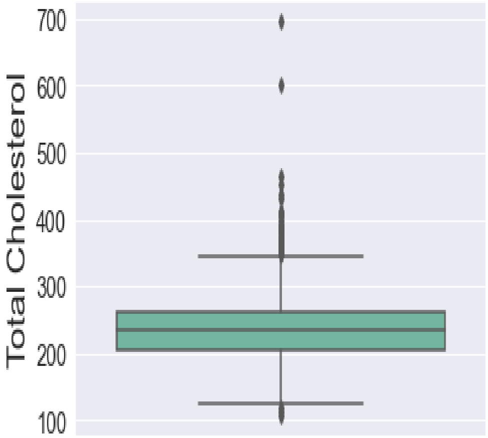

Fig3: Boxplot of total cholesterol column

Missing values in data can occur for a variety of causes, including measuring device malfunctions, human mistake, and incorrect measuring units, among others. Before training the model, any missing data can be addressed. since they impair the learning algorithm's accuracy. As a consequence, missing data should be handled in a really way that as little data as possible is lost. The missing values in the Framingham dataset are managed by our proposed system's pre-processing stage.

Table2 Missing Values

Attribute	Missing Value
Sex	0
Age	0
Education	105
Current Smoker	0
Cigarettes per day	29
BP Meds	53
Prevalent Stroke	0
Prevalent Hyp	0
Diabetes	0
Total Cholo	50
Sys BP	0
Dia BP	0
BMI	20
Heart Rate	1
Glusoce	367
Ten-year CHD	0

Table 2 lists the properties as well as the values that are lacking for each one. As mentioned in our technique, the mean of all associated attribute values is used to replace all attribute data of a single property.

3) Data Imbalance

Researchers have encountered class imbalance as a serious issue affecting accuracy in numerous machine learning problems. When the samples from different classes are not equal, a problem arises. The class imbalance problem also exists in the Framingham dataset (Table 2). In the Framingham dataset, there are 644 examples of class 1 and 3596 samples of class 2, indicating that the dataset is significantly imbalanced, as seen in Fig. 4.

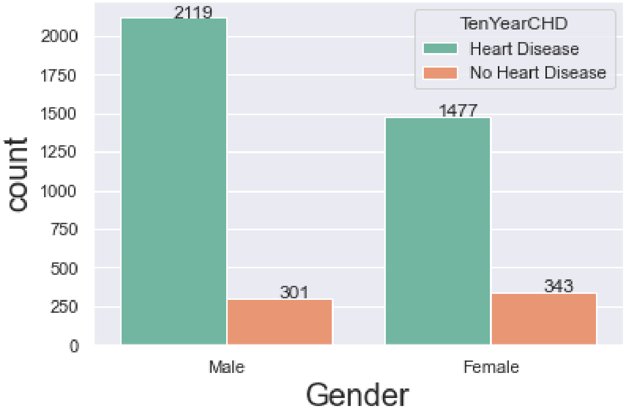

Fig4: Data Imbalance

The oversampling strategy entails randomly copying minority class samples to boost the amount of minority class samples, hence balancing the minority and majority class sample sizes.

The **synthetic minority oversampling technique** (SMOTE), which generates synthetic training instances for the minority class using linear interpolation, is one of the most often used oversampling strategies to handle imbalance problems.

For every instance in the minority class, several of the k-nearest neighbours is randomly selected to construct these synthetic training instances. Fig 5 shows the flowchart of SMOTE. The data is rebuilt just after oversampling process, and the modified data can be classified using a variety of methodologies. The following is the basic idea:

Step 1. Measuring the Euclidean distance among x and any minority classes A yields the k-nearest neighbours of x for each $x \in A$.

Step 2. The uneven proportion determines the sampling rate N. For each $x \in A$, N examples x_1 , x_2 ,x_3,….. x_N ($N \leq K$) are chosen at random from its k-nearest neighbours and used to create the class A_1.

Step 3. For each example $x \in A_1$ (k=1,2….N) , To create a new example, use Eq(1) the formula below:

$$x_{new} = x + rand(0,1) * |x-x_k| \qquad (1)$$

where rand (0,1) denotes a randomized values 0 or 1.

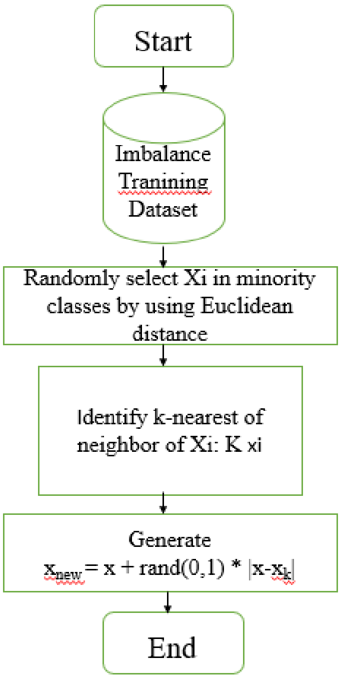
Start
Imbalance Tranining Dataset
Randomly select Xi in minority classes by using Euclidean distance
Identify k-nearest of neighbor of Xi: K xi
Generate
$x_{new} = x + rand(0,1) * |x-x_k|$
End

Fig5: Flowchart of the SMOTE

Our Framework works used SMOTE to resolve data imbalance while choosing the classes of dataset.

4) FEATURE IMPORTANCE

Feature importance assigns a score to each data feature; the greater the score, the more essential or relevant the feature is for forecasting. Feature significance is a built-in class in Tree-Based Classifiers; our suggested framework extracts the most important features using the SelectKBest class.

The weighted average of the node impurity multiplied by the chance of accessing that node determines the relevance of a characteristic.

Table 3. Specs Score

10	sysBP	727.935535
1	age	319.266019
9	totChol	235.502392
4	cigsPerDay	209.897040
11	diaBP	152.748563
7	prevalentHyp	92.048736
8	diabetes	39.144944
5	BPMeds	30.759595
0	male	18.899930
6	prevalentStroke	16.109887
12	BMI	15.227367
2	education	6.318253
13	heartRate	4.232372
3	currentSmoker	0.811334

The 'feature importance' technique has been applied in our proposed framework in this regard. Table 3 shows specs score of our dataset. Figure 6 depicts the findings obtained using this technique. By removing the duplicate feature from the data, feature importance lowers overfitting. Because the data chosen is not duplicated or misleading, it contributes to increased accuracy.

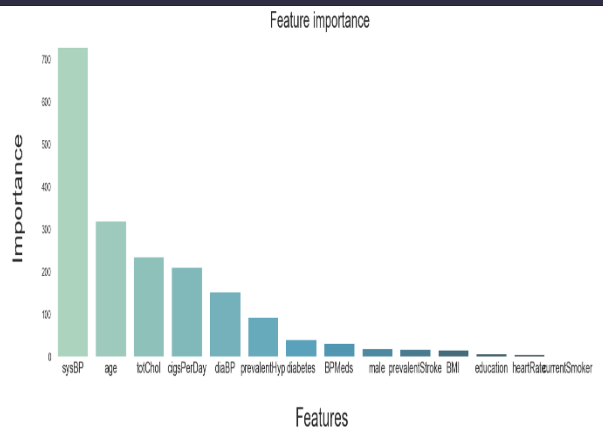

Fig6: Feature Importance

Features with high score are selected for find probability. In this, selected features for applying models to predict more accuracy are SysBP , Age , totChol , cigsperday ,diaBP.

1) Applying LR

When the variable is categorical, logistic regression is now the most appropriate regression method to use (binary). Logistic regression is just a predictive analysis, just like all other regression methods. Logistic regression is a statistical technique for describing and understanding relationships among one dependent categorical variable and nominal, ordinal, interval, or proportion independent variables.

Consider a categorical dependent variable Y and really want to model the probability p(Y = 1|X = x) as just a function of x; any latent variables in the function must be estimated. The outcome of linear regression is squeezed between 0 and 1 using logistic regression. Eq(2) shows a logistic functions to find probablility.

We know that the value of (P) will either exceed 1 or fall below 0, and that the limit of 0 or 1. To solve this problem, we use P's "odds." The proportion of the likelihood of success to the probability of failure is known as odds. By limiting the range, we are reducing the quantity of data points, and as a result, our correlation will fall.

In logistic regression, a logistic function of the odds which serves for predicting target variable:

Formulae for logistic function as follows

log(odds)=logit(P)=ln(P1−P)

consider dependent variable, CHR from our dataset and augment a regression equation for the independent variables such as SysBP, age, totChol, cigPerDay, and diaBP. Then a logistic regression:

logit(p)=a+b1x1+b2x2+b3x3+…

least-squares regression, the relationship between the logit(P) and X is assumed to be linear.

Calculation

$$P = \frac{\exp(\mathrm{a} + \mathrm{b1x1} + \mathrm{b2x2} + \mathrm{b3x3} + \cdots)}{1 + \exp(\mathrm{a} + \mathrm{b1x1} + \mathrm{b2x2} + \mathrm{b3x3} + \cdots)}$$

where :
P = the probability of predicting a value to target(CHR) vriable,

exp = the exponential function (approx. 2.72),

a = the constant (or intercept) of the equation and,

b = the coefficient (or slope) of the predictor variables.

~~$P = \frac{1}{1+e^{-\beta_0+\beta_1 x}}$ (2)~~

Here P is the probability of predicting target variable belongs to 0 or 1. β_0 and β_1 are regression coefficients.

We're using the P function since we're trying to predict probability rather than the log of odds. To do so, multiply both sides by the exponent and thereafter solve for P.

We should anticipate Y = 1 when p 0.5 as well as Y = 0 when p 0.5 to reduce the rate of misclassification.

This means that if $\beta0 + \beta1x$ is non-negative, you should assume 1 and 0 otherwise. As a result of logistic regression, we have a linear classifier. The resolution of $\beta0 + \beta1x = 0$, which would be a point that x is belongs to one dimensional if x is two dimensional, is the decision boundary between the two anticipated classes.

Logistic regression is a statistical model that separates every sample into two categories (Yes/No). It's a way to predict a categorical response variable from a set of independent variables. The LR Model achieved 94.3 percent accuracy on the selected features.

2) Applying KNN

One of the most fundamental adaptive algorithms being used supervised learning is the K-Nearest Neighbour(KNN) approach. In supervised learning, the training data is being labelled and found unknown sample, the model forecasts it using a trained model. KNN performs effectively on datasets with just a large number of samples. It works well with numeric properties as well.

The training dataset is directly used by KNN to create predictions. For each new instance (x), predictions are formed by scanning the complete training set for the K closest instances (neighbours) and summing the outcome variable for K instances.

The average output variable in regression, or the modal category value in classification, could be used. A distance metric is used to identify which of the K examples inside the training data are closest to the new input.

Euclidean distance is the most widely used distance measure for input variables with real values. Euclidean distance is measured with formulae of Eq(3) , for all input variables of k with square root of sum of squared differences between every new point(a) and existing point(b). Formulae for Euclidean distance (ED),

ED (a, b) = sqrt (sum((a-b) ^2))   (3)

The 'k' value is picked, and the distance between the k closest neighbours is determined using that value. Euclidean geometry is widely used. The distance between neighbours is measured. The number k is set to 5. This approach achieves an accuracy of 84.19 percent.

3) Applying DECISION TREE

The supervised learning category includes the decision tree method. They can be used to address problems involving regression and classification. The problem is solved using the tree representation, whereby each leaf node correlates to a class label and characteristics are expressed on the tree's interior node. The selection of attribute for root node within every level is the most difficult task in Decision Tree. Attribute selection is the term for this process. There are two popular methods for selecting attributes.

In machine learning, the Decision Tree algorithm works no attribute-based parameter technique. If there is a single attribute that really can simply segregate data and improve decision-making, it works well. range of the root node poses a hurdle in this approach. When the root node is chosen carefully, the algorithm's computational complexity is reduced, and it becomes extremely effective. This model has a 74.3 percent accuracy rating.

4) Applying ENSEMBLE CLASSIFIER

Ensemble is a strategy for improving the accuracy of outcomes by combining different ML algorithms. Ensemble has substantially contributed in the improvement of accuracy. In our proposed Framework classification, we use an ensemble of LR and KNN. Prediction accuracy of 99.16 percent was attained with this ensemble. We utilised a boosting strategy in the ensemble to produce improved forecasts in the imminent.

V. Result Analysis:

Our proposed machine learning based Cardiovascular Disease Diagnosis framework has analysed the accuracy of the ML models with-out SMOTE and with SMOTE technique. Accuracy of the models with SMOTE is good but the recall of the minority class is very less, i.e. the model is more depends on majority class.

In order to improve recall from both the minority and majority of classes, we need to do data imbalance problem of minority class using SMOTE. SMOTE generates the virtual training records by linear interpolation for the minority class. Above ML models gives accuracy after applying SMOTE technique. The detailed analysis is listed below Table4 with and with out SMOTE.

Table 4 Result Analysis

ML Model	With SMOTE		With-out SMOTE	
	Recall	Accuracy	Recall	Accuracy
LR	0.84	0.94	0.72	0.99
	0.88		0.97	
KNN	0.95	0.84	0.75	0.865
	0.94		0.96	
DT	0.72	0.74	0.72	0.84
	0.97		0.97	
ENSEMBLE	0.99	0.99	0.99	0.99
	0.71		0.32	

CONCLUSION AND FUTURE WORK

The framework is divided into four primary phases, the first of which deals with the mean replacement technique for addressing missing values. In the second phase, the Synthetic Minority Over-sampling Technique is used to correct the data imbalance problem. Feature selection is done in the third phase utilising the feature significance technique. Finally, for improved prediction by using an collaborative of Logistic Regression and KNN is presented. The framework is validated using three benchmark datasets, including Framingham, with accuracies of 99.16 percent. Outlier detection is used to reduce data noise, and then incorrect values are addressed., are all processes in our suggested framework that boost data reliability. The data is then balanced to prevent the model from overfitting or underfitting. The model's computational complexity is reduced by the feature selection stage. All of these processes work together to improve the algorithm's categorization accuracy. It combines unique pre-processing and feature selection stages on the one hand, and employs an innovative ensemble on the other.

REFERENCES

[1] 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.

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

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

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

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

[6] 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

[7] H. A. G. Elsayed and L. Syed, ``An automatic early risk classi_cation of hard coronary heart diseases using Framingham scoring model,'' in *Proc. 2nd Int. Conf. Internet Things,Data Cloud Comput.*, Mar. 2017, pp. 1_8.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.