SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V12I3P111

Volume 12 Issue 3, 100-113, March 2025

Original Article

Improving Coronary Heart Disease Prediction Using Random Forest with a Modified Minority Synthetic Oversampling Technique on an Imbalanced Dataset

M. Janaki Ramudu[1], K. Narasimha Raju[2], A. Krishna Mohan[3]

[1] *Department of Computer Science & Engineering, JNTUK, Kakinada, Andhra Pradesh, India.*
[2]*Department of CSE, GVP College of Engineering, Visakhapatnam, Andhra Pradesh, India.*
[3] *Department of CSE & Director SMS, JNTUK, Kakinada, Andhra Pradesh, India.*

[1]*Corresponding Author : mjanakiramster@gmail.com*

Received: 18 December 2024 Revised: 15 March 2025 Accepted: 20 March 2025 Published: 31 March 2025

Abstract *- Coronary Heart Diseases (CHDs) are the leading cause of death, with a fatal rate increasing every year. Around 80 million females and 110 million males are afflicted by this illness across the globe. Early detection and accurate risk assessment of this disease remain crucial in medical research. Many researchers are working on this issue, but it remains challenging. The proposed technique predicts CHD by applying the Modified Minority Synthetic Over-Sampling Technique (MMSOT) to balance the data and classify the data using the Random Forest (RF) and grid search techniques to fine-tune the hyperparameters. The proposed technique achieved decent performance on the Comprehensive Heart Disease Dataset, with an accuracy of 94.84%, ROC-AUC of 98.15%, Sensitivity of 95.00%, Specificity of 94.70%, F1-Score of 94.61%, Precision (PPV) of 94.21%, and NPV of 95.42%, outperforming baseline models.*

Keywords *- Coronary Heart Disease, Grid Search, Machine Learning Techniques, MMSOT, SMOTE.*

1. Introduction

Coronary Heart Disease (CHD) is a heart disease that damages the heart and eventually results in death [1]. It occurs due to the buildup of fatty substances in the arteries when the coronary arteries struggle to supply oxygen-rich blood to the heart [2]. Even in developed countries like the United States, approximately 7 lakh people died due to this disease in 2022, which is almost 1 in 5 deaths [3, 4].

This disease affects over 110 million men and 80 million women globally [5]. Early detection and accurate risk assessment remain crucial in lessening the devasting effects of CHD despite advancements in medical research and therapeutic approaches [6]. Traditional healthcare systems have struggled to satisfy patient needs, resulting in unreliable outcomes. Modern medical equipment and technologies include internal applications for gathering and storing precise patient data, which serves as a valuable resource for Machine Learning (ML) predictions [7]. ML algorithms have the adaptability to extract knowledge from data for CHD risk assessment. These algorithms can unveil intricate relationships that underpin CHD risk by ingesting and processing a myriad of patient attributes, including demographics, medical history, laboratory results, and imaging data.

1.1. Background on SMOTE

In the existing approach, SMOTE initially chooses a sample of the dataset's minority classes at random. Next, it looks through the other minority class samples to identify the minority sample's k-nearest neighbors (Using Euclidean distance). One neighbor is chosen at random from these k-nearest neighbors. Lastly, SMOTE interpolates between the randomly selected minority sample and its neighbor to create a synthetic sample. Usually, a point is randomly selected along the line connecting the two samples to interpolate. The interpolation formula is:

$$X_{new} = X_{oldMin} + r * (X_n - X_{oldMin}) \quad (1)$$

X_{new} is the new synthetic sample, X_{oldMin} is the existing minority sample, r is any random factor, and X_n is the randomly selected neighbor. The main problem in SMOTE is that if the initially chosen minority sample is located in a region heavily dominated by majority-class samples, there's a high chance that the synthetic sample generated from this point could still lie close to or within the majority-class's decision boundary. As a result, during the testing process, the model might incorrectly predict this synthetic sample as part of the majority class, leading to misclassification.

1.2. Research Gap

Many researchers have worked on detecting the risk of CHD by applying ML algorithms, but they still need methods to improve accuracy by utilizing balanced datasets.

1.3. Problem Statement

CHD is a global health issue affecting millions of people. Accurate CHD risk assessment is critical to finding high-risk patients and establishing preventative measures. Developing techniques to balance the data for applying ML algorithms in assessing CHD risk is a major problem.

1.4. Objectives

- To develop a classification model to predict CHD.
- A new data balancing technique will be applied to make the model efficient.
- To improve the accuracy of the existing models.

1.5. Novel Contributions of the Work

- Integrate the Modified Minority Synthetic Oversampling (MMSOT) technique to balance the data predicting CHD.
- Applied Random Forest model along with Grid Search for hyperparameter tuning.
- Achieved high accuracy in CHD risk assessment.

The entire study is described in 5 segments. Segment 2 provides a detailed literature review, and Segment 3 describes the methodology; Segment 4 discusses data and result analysis. Finally, Segment 5 concludes by reviewing the findings made during the investigation and providing suggestions for future studies.

2. Literature Review

This section provides an overview of various approaches that use different models to predict cardiac disease. ML [8] has transformed healthcare by allowing data-driven decision-making, boosting diagnosis accuracy, and optimizing treatment strategies. In cardiovascular disorders, notably Coronary Heart Disease (CHD), ML models may evaluate massive volumes of patient data, including medical history, lifestyle factors, and clinical test results, to uncover patterns that traditional methods may miss. ML improves patient outcomes by facilitating early diagnosis, risk assessment, and individualized treatment planning through the use of predictive analytics. RF, LR, and SVM algorithms were employed to classify CHD. The dataset underwent balancing using the SMOTE method, and hyperparameter optimization was carried out through 10-fold cross-validation. The accuracy achieved by the RF model was reported as 0.929 [9]. Accuracy is still a problem. Four classification algorithms in ML, namely DT, RF, SVM, and Neural Networks (NN) were employed for CHD prediction. SVM achieved an AUC of 0.75 [10]. A grid search was proposed to fine-tune the hyperparameters in combination with statistical methods [11] for risk assessment of CHD.

The SMOTE technique is used to balance the data, followed by the feature selection technique, and finally, the application of ML techniques. LR achieved better accuracy among multiple classification models. However, the technique did not handle missing values and outliers. Therefore, preprocessing is necessary to improve the performance of any classifier [12].

NB, SVM, and DT were applied for the analysis of CHD with 10-fold cross-validation. The empirical study utilized the South African Heart Disease dataset, containing 462 instances, although small. Among NB, SVM, and DT, Naïve Bayes exhibited better performance in detecting CHD [13]. A comparative study on predicting CHD was conducted using various classification techniques. The recursive feature elimination method and the Boruta method were employed for feature selection, while ROS and SMOTE techniques were utilized to balance the data. Various classification techniques were applied, with the RF Classifier notably achieving an accuracy of 88% [14]. Resampling techniques, including ROS and SMOTE, were applied in conjunction with various classification methods. Notably, the findings revealed that using SMOTE with the Naïve Bayes classifier yielded a higher accuracy of 81.73%, surpassing the 81.12% accuracy achieved with ROS and Naïve Bayes [15].

KNN, SVM, DT, LR, and RF techniques were applied to predict CHD. Results reveal the RF algorithm's superiority, with an 85.05% accuracy, highlighting its potential to enhance cardiovascular risk assessment [16]. The SMOTE resampling technique was introduced to improve classifier performance and sensitivity in detecting minority classes. Using this strategy, the minority class is over-sampled while the majority class is under-sampled [17]. MLP with grid search for hyperparameter tuning achieved the highest accuracy (87.28%) for the prediction of CHD [18].

A comparative study for predicting CHD was conducted, employing various ML classification techniques, including KNN, NB, DT, and RF [19]. This study utilized the "Cleveland dataset", comprising 303 samples and 76 features, with only 14 features considered for evaluation. KNN demonstrated superior performance among all classification techniques, achieving an accuracy of 90.79%. While this performance is notable for a balanced dataset, the model's effectiveness on an imbalanced dataset remains uncertain.

A Hard Voting (HV) classifier comprising LR, RF, MLP, and GNB classifiers was introduced. Before this, the dataset was balanced using Random Under Sampling (RUS), and their proposed technique later achieved improved accuracy at 88.42% [20]. A comparative study was proposed to predict coronary Artery Disease (CAD) using the SVM and ANN models. In their investigation, SVM demonstrated strong performance in predicting CAD [21]. A Logistic Regression model for cardiac disease classification was proposed. The

UCI dataset was used for their study. The model's performance was enhanced through meticulous pre-processing, including data cleaning, handling missing values, and selecting features based on positive correlations. Various training-testing ratios were explored, with the 90:10 split achieving 87.10% accuracy [22]. RF and Extra Trees Classifier (ETC) are the "Tree-based models" that have shown superior predictive accuracy, particularly when combined with feature ranking methods. The Synthetic Minority Oversampling Technique (SMOTE) effectively addresses class imbalance, enhancing model performance. Significant predictors identified include age, creatinine levels, and ejection fraction, which improve classification outcomes. This study builds on these findings by evaluating multiple classifiers, including Logistic Regression and Adaptive Boosting, with results demonstrating that ETC achieves the highest accuracy of 0.9262 [23]. Some of the challenges encountered in these works include class imbalance problems, difficulty in achieving satisfactory accuracy levels, and a notable absence of representation for key performance measures. The summary of the literature is presented in Table 1.

Table 1. Overview of literature

Citation	Year	Technique	Performance Measures (in %)	Dataset	Remarks
[18]	2023	RF, DT, MLP, XGB with cross-validation	Out of all, MLP performed well Accuracy = 87.28	"Kaggle cardiovascular disease dataset"	Accuracy is the problem
[9]	2022	RF, LR, and SVM with 3-repeats, 10-fold repeated cross-validation + SMOTE resampling	Out of all, RF performed well Accuracy = 92.9	"Heart Disease Dataset" from the IEEE-Data Port database	Accuracy is the problem
[22]	2022	Feature Selection + LR	Accuracy = 87.1 (when splitting ratio is 90:10)	"UCI dataset"	Need to improve accuracy at 80:20 splitting ratio.
[11]	2022	RF with a grid search model	Recall = 90.2 F1-score = 82.1 Accuracy = 86	"Framingham Heart Study dataset" from Kaggle	Dataset is unbalanced
[16]	2021	KNN, SVM, DT, LR, and RF with 10-fold cross-validation	Out of all, RF performed well Accuracy = 85.05	"Framingham Heart Study dataset" from the Kaggle	Dataset is unbalanced, but no resampling technique is used
[23]	2021	ETC with SMOTE	Accuracy=0.9262 Precision=0.93 Recall=0.93 F1-Score=0.93	"Heart-failure-clinical-records-dataset" from the UCI ML repository	Small dataset is used for experiments
[19]	2020	KNN, NB, DT, and RF	Out of all, KNN performed well Accuracy = 90.79	"Cleveland dataset" from UCI machine learning repository	Accuracy and data balancing are the problems.
[10]	2019	DT, RF, SVM, and NN	Out of all, SVM performed well Accuracy = 75	"Framingham Heart Study dataset" from the Kaggle repository	Accuracy is the problem.
[21]	2019	SVM and ANN model	SVM performed well PPV = 87.1 Sensitivity=92.32 Specificity=74.42	The research sample was collected from AJA University of Medical Sciences affiliated colleges.	Accuracy is the problem.

3. Methodology

CHD is a prevalent and life-threatening cardiovascular condition that demands effective risk prediction and early intervention. Traditional risk assessment methods based on clinical and demographic data may have limitations regarding accuracy and predictive power. This work uses ML and data balance techniques to create a reliable and accurate model for CHD risk prediction. This model aims to optimize the model's hyperparameters and evaluate its performance using cross-validation techniques. Data collection, data preprocessing, model design and evaluation are the crucial steps.

3.1. Data Collection and Cleaning

The suggested model uses 3 datasets, including Framingham, Z-Alizadeh Sani, and Comprehensive Heart Disease Dataset (Statlog + Cleveland + Hungary dataset).

These datasets were obtained from Kaggle, and feature engineering was performed to choose the best features.

3.1.1. Framingham Dataset

The "Framingham dataset" [24] comprises 4,240 records and 16 features. Among the 4,240 records, 644 are labeled as "yes", while the remaining 3,596 are labeled as "no." For simplicity, the dataset's features are categorized as "Nominal", abbreviated as "Nom", and "Continuous" abbreviated as "Contn" in Table 2.

The Framingham dataset has been extensively used in heart disease research and provides well-established features, making it a suitable choice for detailed analysis in this study. Hence, its attributes are explicitly designated, while the other datasets are not elaborated. However, this study conducted research on all three datasets.

Table 2. Framingham data set attributes description

	Variable Name	Description	Type of Attribute
Demographic	"MALE"	Male or Female (0 or 1)	Nom
	"AGE"	Patient's age (32 to 70 yrs)	Contn
Behavioral	"EDUCATION"	Education Levels: 1 to 4	Contn
	"CURRENTSMOKER"	Whether the patient smokes now or not (0 or 1)	Nom
	"CIGSPERDAY"	Daily intake of cigarettes (0 to 70 per day)	Contn
Medical History	"BPMEDS"	Whether the patient took blood pressure medicine or not (0 or 1)	Nom
	"PREVALENTSTROKE"	1, if the patient had a stroke in the past, otherwise it is 0	Nom
	"PREVALENTHYP"	High blood pressure 1, otherwise 0	Nom
	"DIABETES"	Diabetic patient 1, otherwise 0	Nom
	"TOTCHOL"	Total cholesterol measurement in mg/dL	Contn
	"SYSBP"	Systolic blood pressure, mmHg	Contn
	"DIABP"	Diastolic blood pressure, mmHg	Contn
	"BMI"	Body Mass Index, weight(kg)/height(m^2)	Contn
	"HEARTRATE"	Heart rate measured in beats/minute	Contn
	"GLUCOSE"	Glucose level, mg/dL	Contn
Target Variable	"TENYEARCHD"	Is there a 10-year risk of CHD for the patient? (1 represent yes, 0 means no)	Nom

3.1.2. Z-Alizadeh Sani Dataset

The "Z-Alizadeh Sani dataset" [25] contains 303 samples and 55 features. The class label "Cath" has two possible values: "Cad" and "Normal".

3.1.3. Comprehensive Heart Disease Dataset

The "Comprehensive Heart Disease dataset" [26] comprises 1,190 records and 12 features.

The class label "target" is binary, with values "1" indicating CHD (629 entries) and "0" indicating non-CHD (561 entries).

3.2. Data Preprocessing

After loading the data, preprocessing begins by identifying missing values in the dataset, as shown in Figure 1.

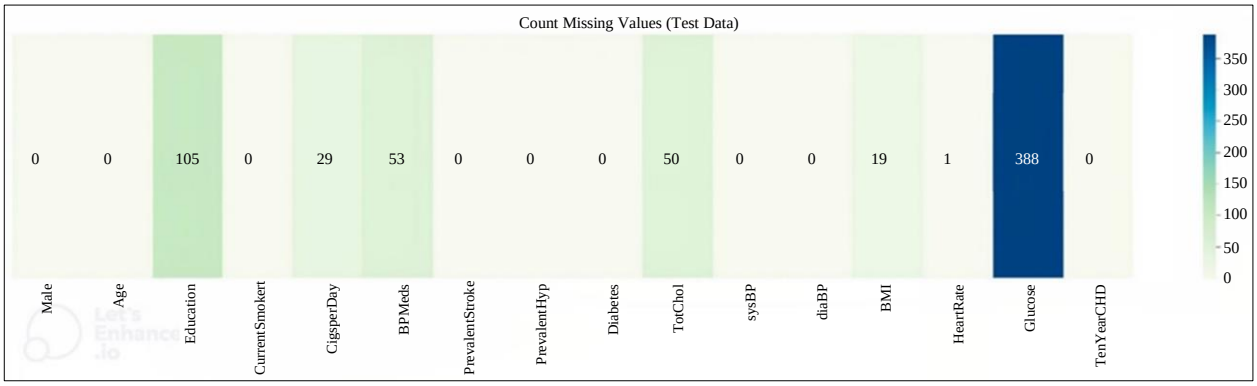

Fig 1. Missing values in framingham dataset

Here, three different approaches are considered, are

- Mean/Median Imputation: Use the mean or median of the relevant feature to replace missing numerical values. Despite being a straightforward imputation technique, it could not be appropriate if the missing values are not MCAR.
- Mode Imputation: Replace missing categorical data with the mode.
- Remove Rows: Consider eliminating certain rows if the dataset's total number of missing values is relatively low. However, use this approach cautiously, which may result in data loss.

Based on the above analysis, only glucose has the highest number of missing values, almost ~10%, so mean imputation is used to fill in missing values. The remaining records have a very low percentage of values, so the remaining records are removed. Table 3 describes the statistical properties of the attributes in the dataset. In the following table, the attributes "male, age, education, current smoker, cigsPerDay, BPMeds, prevalentStroke, prevalentHyp, diabetes, totChol, sysBP, diaBP, BMI, heartRate, glucose, TenYearCHD" are renamed as "Gn, Age, Edu, CS, CPD, BPM, PStr, PHyp, Diab, TCh, sysBP, DBP, BMI, HR, Gl, and TYCHD" respectively.

Table 3. Statistical properties of the data

Index	Gn	Age	Edu	CS	CPD	BPM	PStr	PHyp	Diab	TCh	sysBP	DBP	BMI	HR	Gl	TYCHD
Count	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989	3989
Mean	0.43	49.47	1.98	0.49	9.02	0.03	0.01	0.31	0.03	236.6	132.23	82.87	25.77	75.87	81.86	0.15
Std	0.5	8.53	1.02	0.5	11.92	0.17	0.07	0.46	0.16	44.02	21.94	11.88	4.08	12.09	22.89	0.36
Min	0	32	1	0	0	0	0	0	0	113	83.5	48	15.54	44	40	0
25%	0	42	1	0	0	0	0	0	0	206	117	75	23.06	68	72	0
50%	0	49	2	0	0	0	0	0	0	234	128	82	25.38	75	79	0
75%	1	56	3	1	20	0	0	1	0	263	143.5	89.5	27.99	83	85	0
Max	1	70	4	1	70	1	1	1	1	600	295	142.5	56.8	143	394	1

All these datasets are split into two sections: a test portion and a training component. While the training phase creates a model that predicts heart disease, the test dataset portion is used to evaluate classifiers, as shown in Figure 2. Data cleansing involves addressing missing values in the dataset by filling them with the mean value for numerical features. There are no missing values in categorical features. Subsequently, Exploratory Data Analysis (EDA) is applied for data visualization to observe the correlation among all features. The feature matrix undergoes standardization using a standard scalar; the resulting mean and standard deviation are approximately 0 and 1, respectively.

3.3. Proposed Model

The overall architecture consists of MMSOT, Grid Search, and Random classifier, as shown in Figure 3.

The MMSOT technique is employed to balance the data described in Algorithm 1. The MMSOT technique is used to generate synthetic samples from existing minority samples.

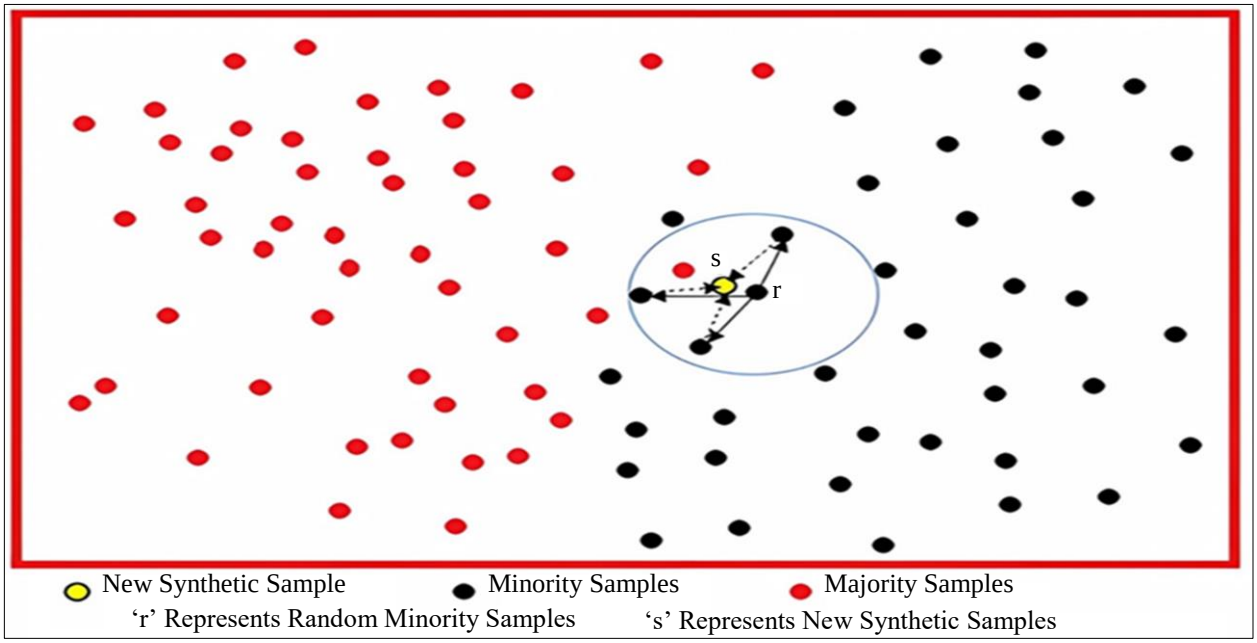

Fig. 2 Modified minority synthetic oversampling technique (Let K=3)

The value of k is determined based on 1% to 10% of the minority sample count, depending on the percentage of minority samples in the dataset. In the context of the MMSOT algorithm, the k value was selected using grid search to optimize model performance. This approach allowed for the systematic evaluation of various k values, ensuring that the chosen parameter effectively contributed to improving classification outcomes. The algorithm finds the k nearest neighbors for a randomly selected minority sample and generates a synthetic sample by taking the average of these k neighbors. This technique generates synthetic samples by averaging the k nearest neighbors. Figure 3 describes how this technique is applied to the given dataset.

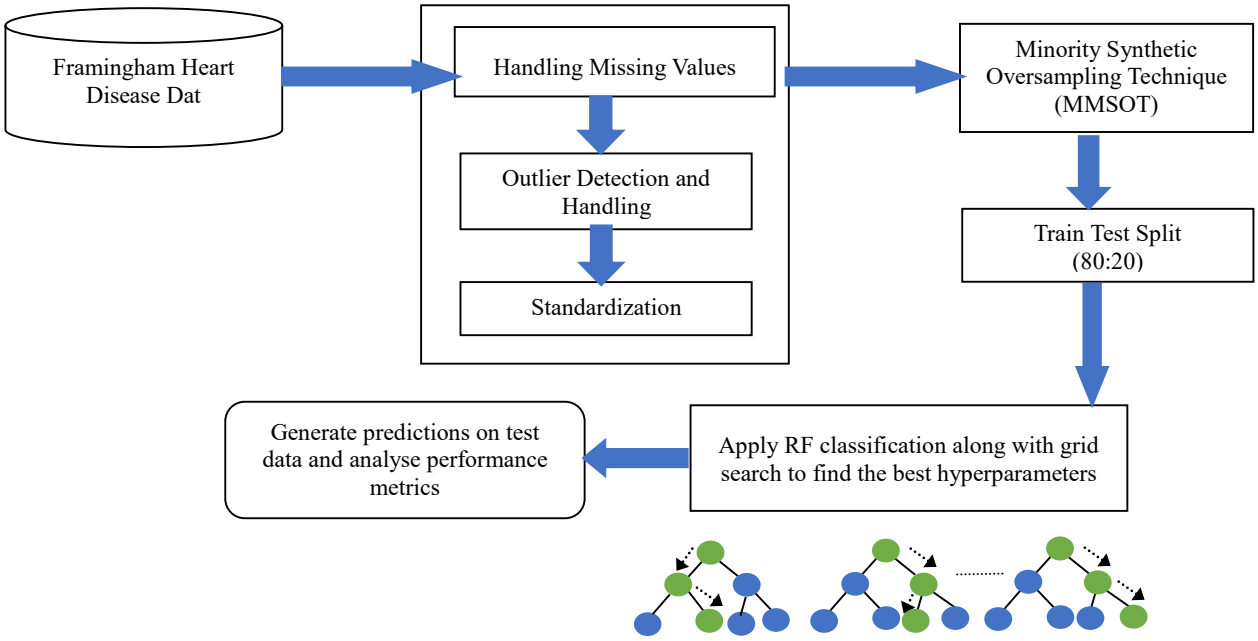

Fig. 3 The architecture of the proposed model (RF + MMSOT)

After that, Grid Search is employed to find the best hyper-parameters, and then a Random Forest classifier is applied. The pseudo-code/algorithm 2 for the proposed model is described below. The following algorithm describes how MMSOT is applied to balance the unbalanced data.

Algorithm 1: MMSOT

1. Input:
   a. X_minority: Array of minority class samples, where each sample is represented as a vector x_i.
   b. k_neighbors: Number of nearest neighbors to consider.
2. Initialization:
   a. Create a NearestNeighbors model with k_neighbors + 1 neighbors.
3. Fit the Model:
   a. Fit the $NearestNeighbors$ model on $X\_minority$.
4. Randomly Select Sample:
   a. Randomly select an index i from the range $[0, \text{len}(\text{X\_minority}) - 1]$.
   b. Retrieve the selected sample X_i from X_minority.
5. Find Nearest Neighbors:
   a. Use the NearestNeighbors model to find the k_neighbors + 1 nearest neighbors of x_i.
   b. Extract the indices of these nearest neighbors, excluding i. Let these indices be$\{j_1, j_2, \ldots, j_{k_neighbors}\}$.
6. Generate Synthetic Sample by Averaging the Nearest Neighbors:
   a. Retrieve the nearest neighbors $\{x_{j_1}, x_{j_2}, \ldots, x_{j_{k\_neighbors}}\}$ from X_minority.
   b. Compute the average vector of these nearest neighbors:

$$x_{new} = \frac{1}{k\_neighbors} \sum_{i=1}^{k\_neighbors} x_{j_i}$$

7. Round the values of x_{new}Convert it to integer type if needed, and then return the new synthetic sample.

The following algorithm outlines the application of RF with the grid search algorithm.

Pseudo Code/Algorithm 2: Proposed Model (RF + MMSOT)

1. Load the dataset
2. Use standard scalar to standardize features
3. Apply MMSOT to balance the class distribution
4. Split the data in an 80:20 ratio between training and test sets.
5. best_n_estimators, best_max_depth = Grid_search (estimator, param_grid, scoring, cv)
6. Create an RF classifier with obtained hyperparameters in step 5

var=1
while var<= best_n_estimators do

- At each node, randomly pick Z features from the total number of D features. Usually, the Z value $\sqrt{D}$ is taken.

$$D = |\{x_1, x_2, x_3, \ldots., x_n\}|$$

Where $x_1, x_2, x_3, \ldots. x_n$ are features in the dataset. D is the cardinality of the set containing the individual features in the dataset.

depth=1
while depth != max_depth do

- Compute the Gini index for each potential split using Z selected subset of features based on the following Equation

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2$$

- Choose the split with the lowest Gini Index
- Increase the depth, $depth = depth + 1$

end while

$$var = var + 1$$

end while

7. Repeat the above process to create a forest of k-trees
8. For each tree in the forest of k-trees, find the output and use the majority voting method to predict the final result using the following formula

$$H(x) = \arg\max_{y} \sum_{i=1}^{k} I(\, h_i(x) = Y)$$

Where H(x) is the final predicted class, k is the number of trees in a random forest, and $h_i(x)$ is the predicted class by the i[th] tree. $I(h_i(x) = Y)$ is an indicator function that equals 1 if $h_i(x) = Y$ and 0 otherwise.

$$I(h_i(x) = Y) = \begin{cases} 1\ if\ h_i(x) = Y \\ 0 \quad Otherwise \end{cases}$$

9. Make predictions on test data.
10. Calculate various performance metrics.

This model uses the Gini index as an impurity measure to construct a tree. Mathematically, it could be written as shown in Equation (2).

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2 \quad (2)$$

If the predicted class is binary, with values such as YES (Y) or NO (N), the Gini Index is rewritten, as shown in Equation (3).

$$Gini\ Index = 1 - [(P_Y)^2 + (P_N)^2] \quad (3)$$

Where P_Y and P_N are probabilities of YES and NO classes, respectively.

4. Experimental Setup and Result Analysis

4.1. Experimental Setup

Google Colab was used to conduct the analysis. For data manipulation, analysis, and visualization, the work made substantial use of Python tools such as scikit-learn (sklearn), Matplotlib, Pandas, and NumPy. Sklearn library is used for data preprocessing, model training, and evaluation; the Matplotlib library is used for creating visualizations such as plots and charts; Panda's library is used for data manipulation and analysis; and NumPy library is used for numerical operations.

The investigation was carried out on a machine running Windows 11 with an Intel (R) CoreTM i3-8130U CPU running at 2.21 GHz and 8GB of RAM. All 3 datasets are imbalanced; the study addresses this issue by applying the MMSOT technique, oversampling the minority class to achieve a balanced distribution. After preprocessing and addressing the imbalance, the dataset is partitioned into two segments, allocating 80% for training and 20% for testing purposes. The next step involves applying different classifiers, coupled with grid search, to perform hyperparameter tuning and achieve optimal classification.

4.1.1. Correlation Matrix

A correlation among attributes would be described with the help of the heatmap shown in Figure 4. Utilizing a heatmap facilitates the visualization of the influence of independent features on dependent variables. Furthermore, it aids in identifying the features most strongly associated with the dependent variable. Later, analyzed the Framingham dataset concerning numerical attributes and observed the strength of the relationship between these features and the class label based on bar plot and KDE (Kernel Density Estimation) plot statistical visualizations as shown in Figure 5.

Next, analyzed the Framingham dataset with categorical features and observed the relationship between categorical variables and the binary variable TenYearCHD in the dataset. This statistical analysis shows two plots for each categorical variable; the first plot shows a Pie chart, calculates the counts and proportion of unique values, and visualizes the distribution of unique values within the categorical variable.

The other plot represents a Stacked bar chart for TenYearCHD, which is a count plot that stacks bars to show the distribution of TenYearCHD within each categorical variable such as education, male, currentSmoker, BPMeds, prevalentStroke, prevalentHyp and diabetes are shown in the following Figure 6 to Figure 12.

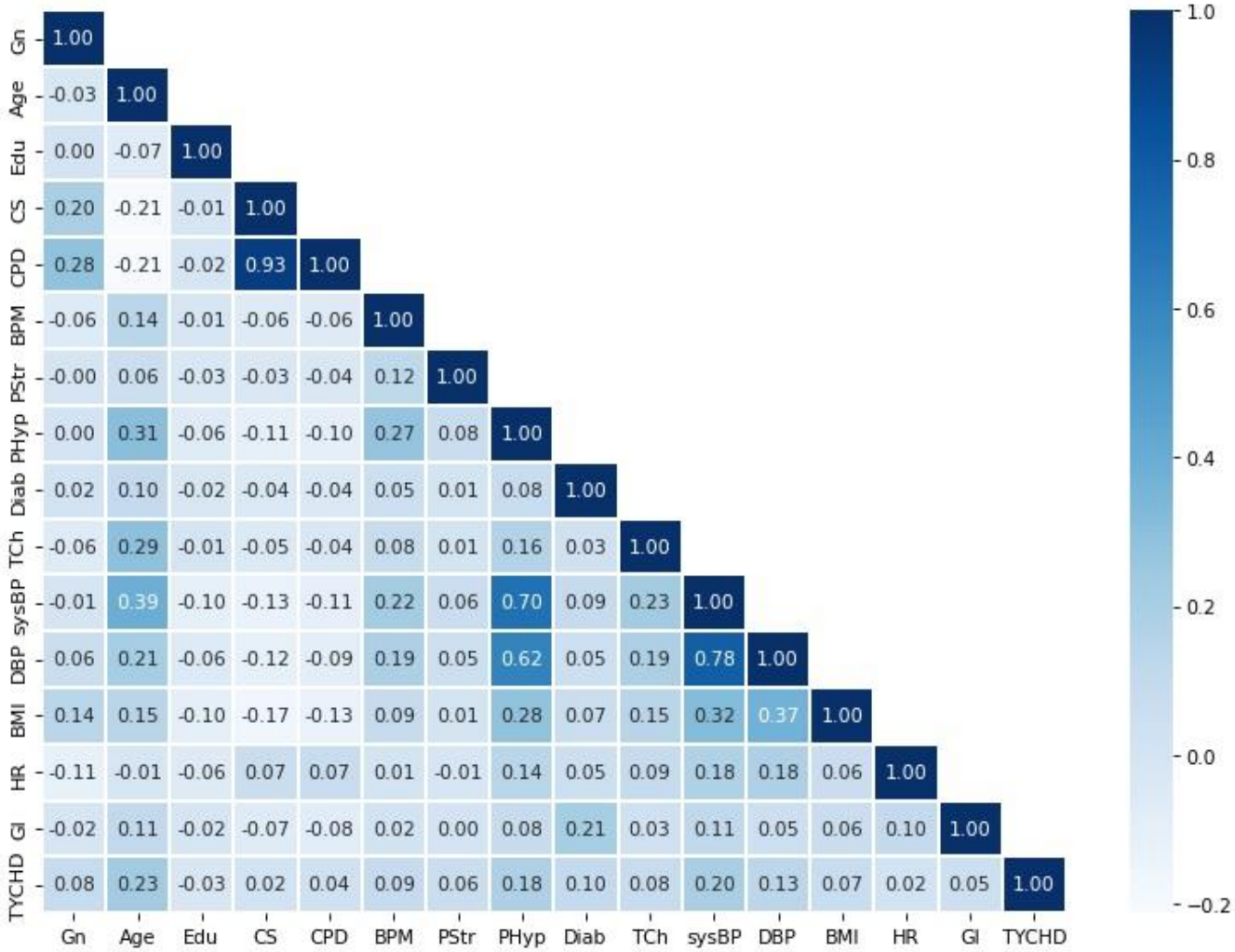

Fig. 4 A correlation matrix with Heatmap

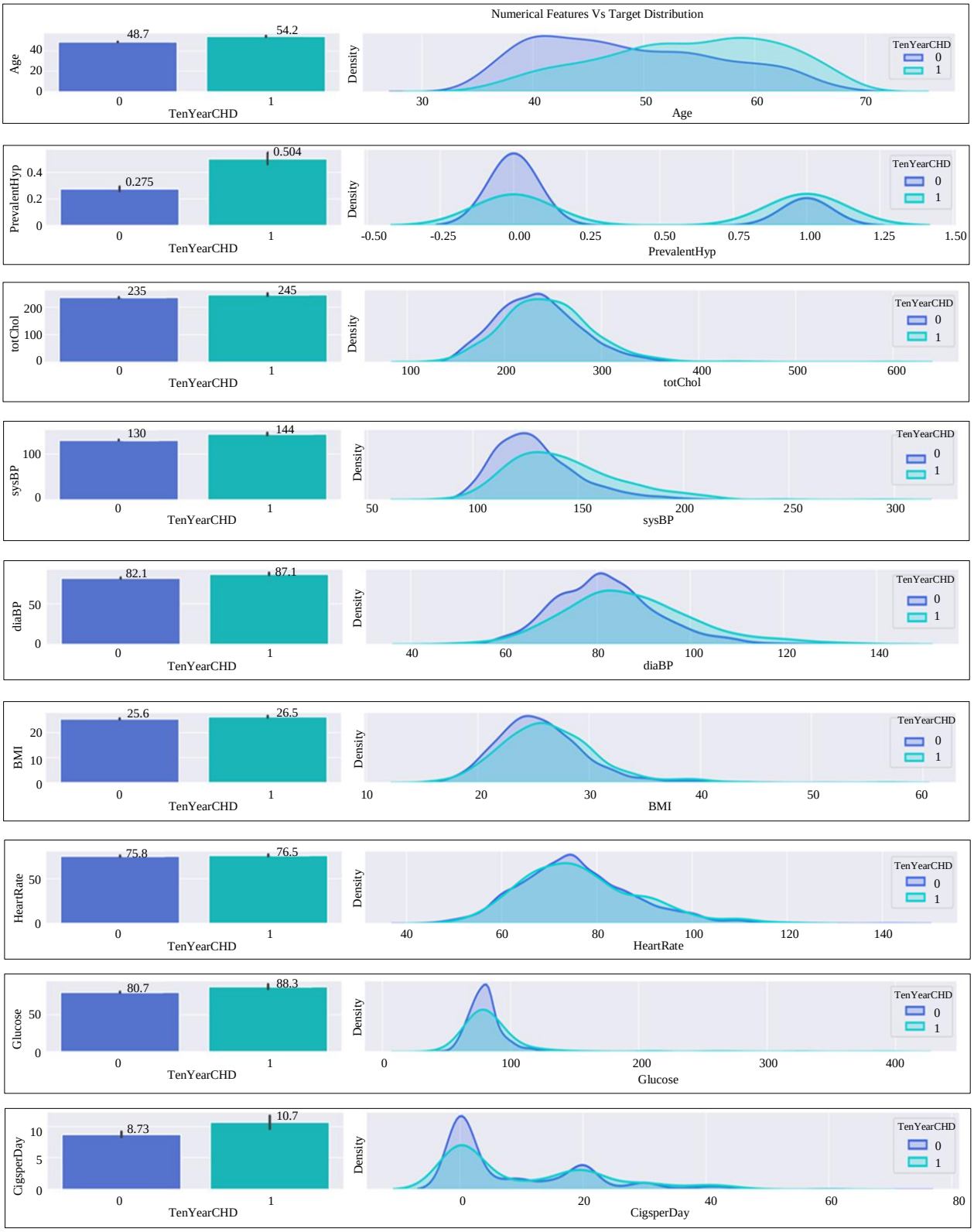

Fig. 5 Numerical features vs Target distribution (TenYearCHD)

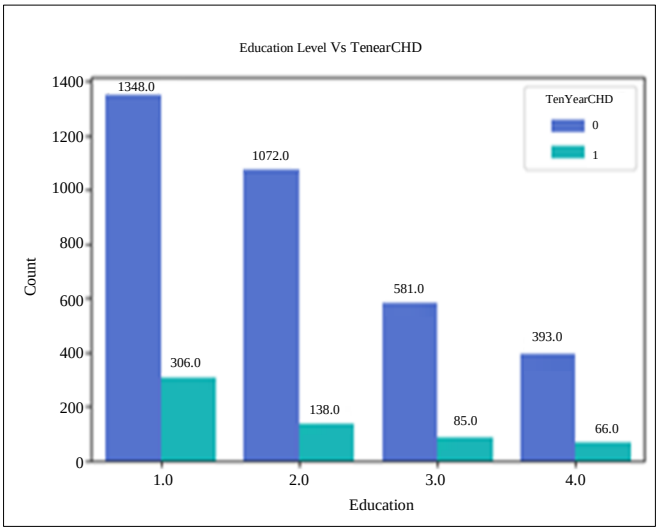

Fig. 6 Analysis of education vs TenYearCHD

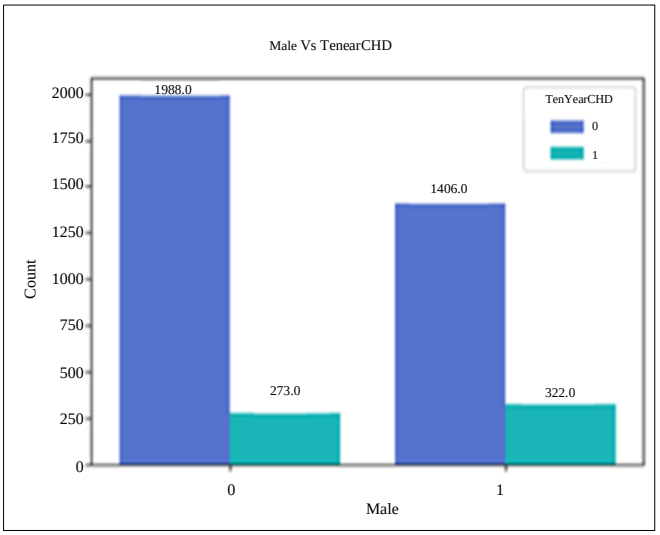

Fig. 7 Analysis of gender vs TenYearCHD

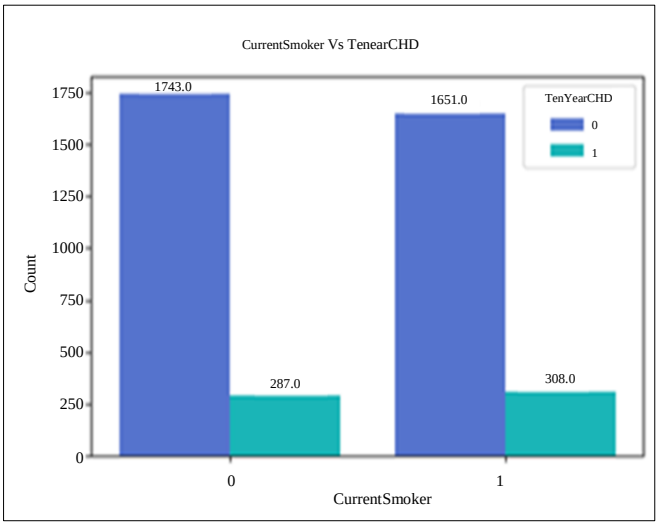

Fig. 8 Analysis of smoking vs TenYearCHD

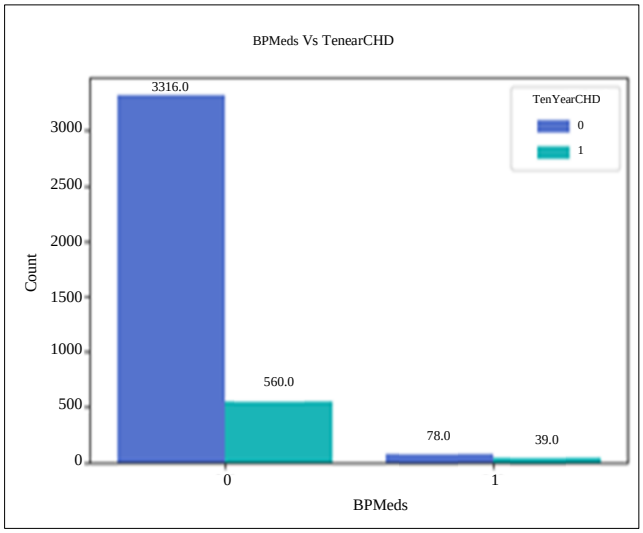

Fig. 9 Analysis of BPMeds vs TenYearCHD

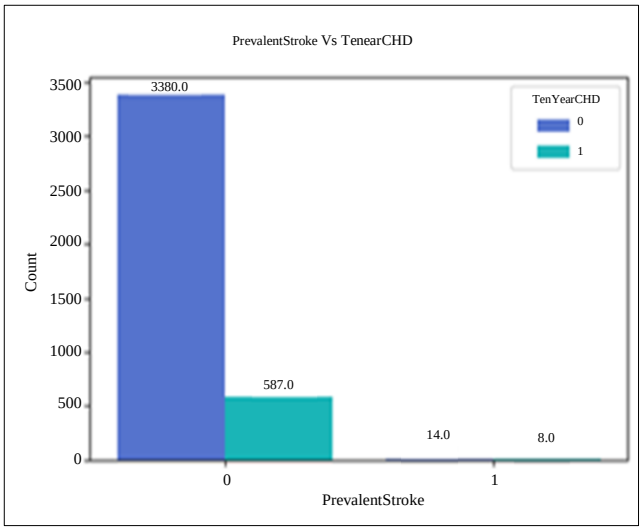

Fig. 10 Analysis of PrevalentStroke vs TenYearCHD

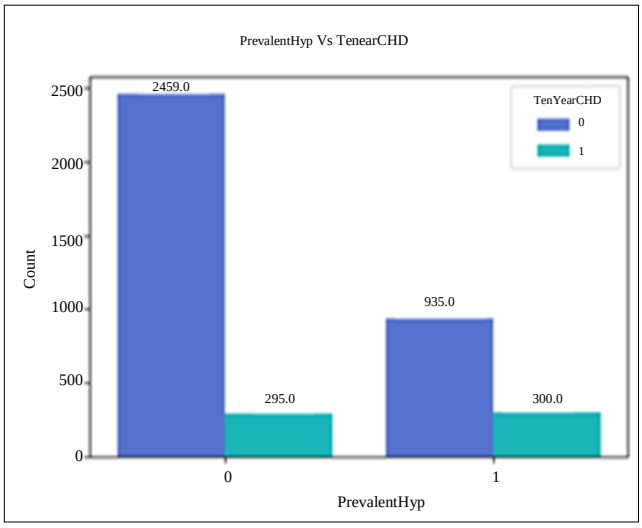

Fig. 11 Analysis of PrevalentHyp vs TenYearCHD

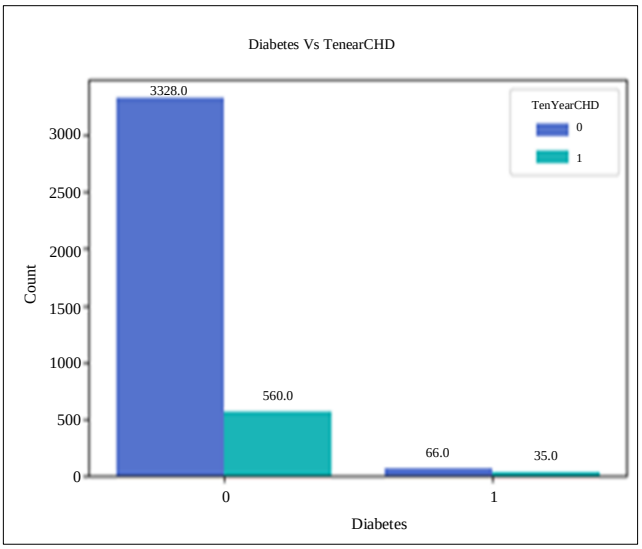

Fig. 12 Analysis of diabetes vs TenYearCHD

4.2. Performance Measures

This section evaluates the efficacy of ML classifiers through a range of assessment metrics, such as accuracy, ROC AUC, recall, specificity, PPV, NPV, and F1-score. Evaluate the performance of various classifiers with the help of various performance measures.

Accuracy: It is the percentage of accurately predicted cases-both TP and TN-out of all the instances in the dataset, which is known as accuracy. Accuracy in mathematical terms is represented as shown in Equation (4):

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions (TP+TN)}}{\text{Total Number of Predictions (TP+TN+FP+FN)}} \quad (4)$$

- Where TP means the total count of instances where the classifier correctly identifies something as positive when it is positive.
- TN means the total count of instances where the classifier correctly identifies something as negative when it is negative.
- FP means the total count of instances where the classifier correctly identifies something as positive when it is negative.
- FN means the total count of instances where the classifier correctly identifies something as negative when it is positive.

Sensitivity: It evaluates how well a classification model can identify positive occurrences among all the dataset's actual positive cases. Mathematically, it could be represented as Equation (5).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (5)$$

Specificity: It evaluates how well a classification model can identify negative occurrences among all the dataset's actual negative cases. It could be represented in Equation (6).

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (6)$$

PPV (Precision): This metric assesses how well the classification model predicts positive outcomes. It could be represented in Equation (7).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

NPV: It evaluates a classification model's ability to make accurate negative predictions. It could be represented in Equation (8).

$$\text{NPV} = \frac{\text{TN}}{\text{TN+FN}} \quad (8)$$

F1-Score: It is a harmonic mean of obtained precision and recall. It could be represented in Equation (9).

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision*Recall}}{\text{Precision+Recall}} \quad (9)$$

4.3. Result Analysis with Dynamic k Value

Based on the results obtained from different ML classifiers, the proposed model demonstrated superior performance, yielding higher accuracy, ROC AUC, sensitivity, specificity, F1-score, PPV, and NPV compared to all other classifiers. The following Table 4 shows the performance of the proposed model.

For different values of K in MMSOT, the following metrics were given: When K = 1%, it means the k value is 1% of total minority samples; at this stage, it gave better performance. Where k is essentially a dynamically adjusted parameter that changes based on the size of the minority class. In general, the default value for K in SMOTE resampling is 5.

Table 4. Analysis of MMSOT at various K values applied to the framingham dataset

K value	Test Accu	ROC AUC	Sen	Spe	PPV	NPV	F1-score
K=1%	0.934	0.976	0.907	0.961	0.960	0.910	0.933
K=2%	0.912	0.971	0.865	0.961	0.958	0.874	0.909
K=3%	0.914	0.968	0.860	0.969	0.966	0.871	0.909
K=4%	0.905	0.965	0.853	0.958	0.954	0.864	0.901

K=5%	0.912	0.967	0.866	0.960	0.957	0.875	0.909
K=6%	0.914	0.963	0.865	0.966	0.963	0.874	0.911
K=7%	0.913	0.964	0.857	0.970	0.967	0.869	0.909
K=8%	0.911	0.961	0.851	0.972	0.969	0.865	0.906
K=9%	0.918	0.967	0.866	0.970	0.967	0.876	0.914
K=10%	0.910	0.960	0.852	0.970	0.967	0.865	0.906

4.4. Comparison of Existing Approaches with Proposed Work

Table 5. Comparison of existing approaches with proposed work on framingham dataset

Technique/ Model	Performance Metrics						
	Test Accuracy	ROC AUC	Sensitivity/ Recall	Specificity	F1-Score	PPV / Precision	NPV
LR [22]	86.72	73,15	86.87	66.67	92.85	99.71	3.7
MLP [18]	85.83	64.95	86.47	14.29	92.37	99.13	0.93
KNN [19]	83.33	50.92	95.36	6.48	30.21	17.95	86.69
SVM [10]	86.46	47.64	-	100	-	-	86.47
ETC + SMOTE [23]	73.64	82.71	72.20	75.11	73.48	74.81	72.52
RF + SMOTE + Grid Search [9]	89.10	95.97	91.99	86.14	89.52	87.17	91.31
Proposed Model (RF + MMSOT + Grid Search)	93.37	97.58	90.68	96.13	93.26	95.99	90.97

Table 6. Comparison of existing approaches with proposed work on Comprehensive Heart Disease dataset (Statlog + Cleveland + Hungary dataset)

Technique/ Model	Performance Metrics						
	Test Accuracy	ROC AUC	Sensitivity/ Recall	Specificity	F1-Score	PPV / Precision	NPV
LR [22]	79.83	89.16	81.40	77.98	81.40	81.40	77.98
MLP [18]	87.82	94.96	89.15	86.23	88.80	88.46	87.03
KNN [19]	86.55	92.56	89.15	83.49	87.79	86.47	86.67
SVM [10]	80.25	88.20	82.17	77.98	81.85	81.54	78.70
ETC + SMOTE [23]	87.39	95.18	89.15	85.32	88.46	87.79	86.92
RF + SMOTE + Grid Search [9]	93.65	97.59	94.17	93.18	93.39	92.62	94.62
Proposed Model (RF + MMSOT + Grid Search)	94.84	98.15	95.00	94.70	94.61	94.21	95.42

Table 7. Comparison of existing approaches with proposed work on the Z-Alizadeh Sani dataset

Technique/ Model	Performance Metrics						
	Test Accuracy	ROC AUC	Sensitivity/ Recall	Specificity	F1-Score	PPV / Precision	NPV
LR [22]	85.25	92.65	93.18	64.71	90.11	87.23	78.57
MLP [18]	86.89	90.78	93.18	70.59	91.11	89.13	80.00
KNN [19]	81.97	86.30	88.64	64.71	87.64	86.67	68.75
SVM [10]	85.25	91.84	90.90	70.59	89.89	88.89	75.00
ETC + SMOTE [23]	89.66	97.30	88.37	90.90	89.41	90.48	88.89
RF + SMOTE + Grid Search [9]	93.10	98.26	90.70	95.45	95.12	91.30	92.86
Proposed Model (RF + MMSOT + Grid Search)	94.84	98.15	95.00	94.70	94.21	95.42	94.61

Figure 13 to Figure 15 show the ROC AUC comparison of baseline models with a proposed model on 3 different datasets.

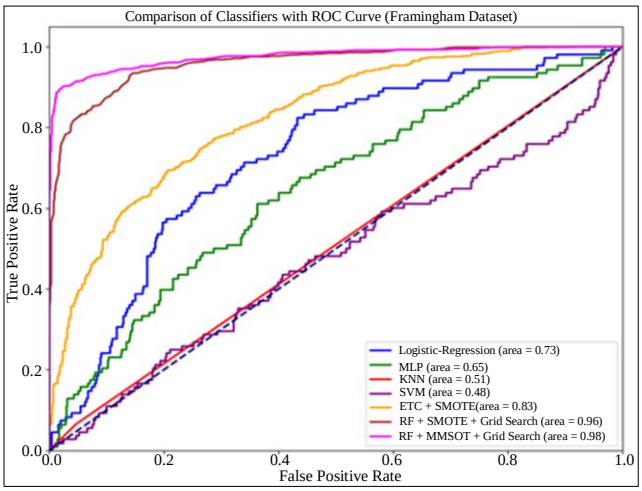

Fig. 13 ROC AUC comparison of baseline models with the proposed model on framingham dataset

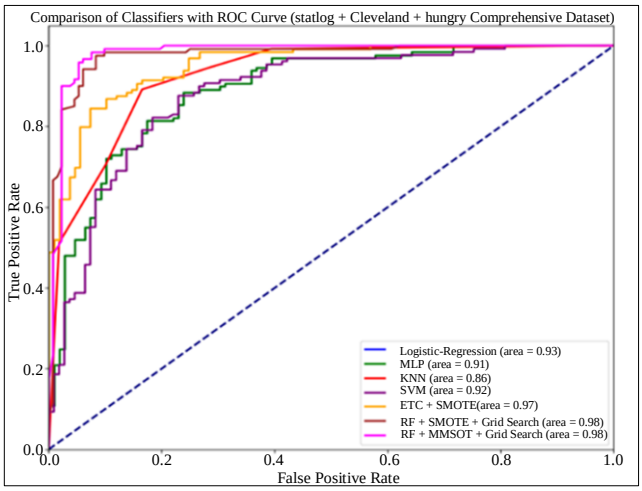

Fig. 14 ROC AUC comparison of baseline models with the proposed model on statlog + Cleveland + Hungary comprehensive dataset

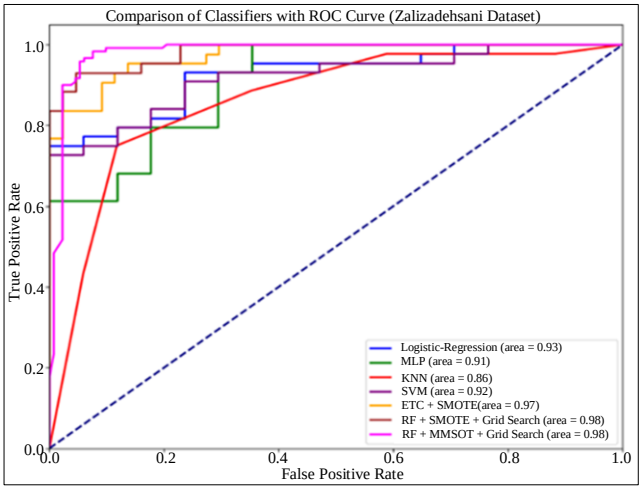

Fig. 15 ROC AUC comparison of baseline models with the proposed model on the Z-Alizadeh Sani dataset

5. Conclusion and Future Scope

CHD is one of the serious health problems and leading cause of most deaths. Early prediction of this disease is crucial to avail better treatment. The proposed model integrates MMSOT, RF and Grid Search techniques. The MMSOT technique provides better class balancing for the gathered data. The performance measures such as ROC AUC, accuracy, and other important metrics consistently showed superior results with the proposed model over the state-of-the-art. This result implies that MMSOT is one of the best competitive models among others, which improves the model's ability to generalize the difficulties related to class imbalance. The proposed model can be deployed in the healthcare sector to provide early prediction of CHD. The present work is limited to 3 datasets and ML algorithms. In future, advanced deep-learning techniques will be explored to enhance heart disease prediction further. IoT is a new emerging technology that controls and monitors data from remote areas. In the future, it can be integrated into IoT devices.

References

[1] Roth, G. A. et al. "Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update from the GBD 2019 Study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982-3021, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[2] Cardiovascular Diseases, World Health Organization, 2019. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[3] National Center for Health Statistics, Mortality Data on CDC WONDER, CDC WONDER Database, 2018. [Online]. Available: https://wonder.cdc.gov/mcd.html

[4] Seth S. Martin et al., "2024 Heart Disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association," *Circulation*, vol. 149, no. 8, pp. e347-e913, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Adel Bashatah, Wajid Syed, and Mohmood Basil A. Al-Rawi, "Knowledge of Cardiovascular Disease Risk Factors and Its Primary Prevention Practices Among the Saudi Public - A Questionnaire-Based Cross-Sectional Study," *International Journal of General Medicine*, vol. 16, pp. 4745-4756, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Maedeh Amini, Farid Zayeri, and Masoud Salehi, "Trend Analysis of Cardiovascular Disease Mortality, Incidence, and Mortality-To-Incidence Ratio: Results from Global Burden of Disease Study 2017," *BMC Public Health*, vol. 21, no. 1, pp. 2-12, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mithun Sarker, "Revolutionizing Healthcare: The Role of Machine Learning in the Health Sector," *Journal of Artificial Intelligence General Science*, vol. 2, no. 1, pp. 36-61, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 2017. [Publisher Link]
[9] Rüstem Yılmaz, and Fatma Hilal Yagin, "Early Detection of Coronary Heart Disease Based on Machine Learning Methods," *Medical Records*, vol. 4, no. 1, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[10] Juan-Jose Beunza et al., "Comparison of Machine Learning Algorithms for Clinical Event Prediction (Risk of Coronary Heart Disease)," *Journal of Biomedical Informatics*, vol. 97, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[11] S. Prabu et al., "Grid Search for Predicting Coronary Heart Disease by Tuning Hyper-Parameters," *Computer Systems Science and Engineering*, vol. 43, no. 2, pp. 737-749, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[12] Kelvin Kwakye, and Emmanuel Dadzie, "Machine Learning-Based Classification Algorithms for the prediction of CHD," *arXiv Preprint*, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[13] Amanda H. Gonsalves et al., "Prediction of Coronary Heart Disease Using Machine Learning: An Experimental Analysis," *Proceedings of the 3rd International Conference on Deep Learning Technologies*, Xiamen, China, pp. 51-56, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[14] K. Nirmala Devi, S. Suruthi, and S. Shanthi, "Coronary Artery Disease Prediction using Machine Learning Techniques," *8th International Conference on Advanced Computing and Communication Systems*, Coimbatore, India, pp. 1029-1034, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[15] Nur Silviyah Rahmi et al., "SMOTE Classification and Random Oversampling Naive Bayes in Imbalanced Data: (Case Study of Early Detection of Cervical Cancer in Indonesia)," *IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, Yogyakarta, Indonesia, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[16] Walaa Adel Mahmoud, Mohamed Aborizka, and Fathy Amer, "Heart Disease Prediction Using Machine Learning and Data Mining Techniques: Application of Framingham Dataset," *IDOSR Journal of Computer and Applied Sciences*, vol. 6, no. 1, pp. 66-73, 2021. [Google Scholar] [Publisher Link]
[17] Nitesh V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. [CrossRef] [Google Scholar] [Publisher Link]
[18] Chintan M. Bhatt et al., "Effective Heart Disease Prediction Using Machine Learning Techniques," *Algorithms*, vol. 16, no. 2, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]
[19] Devansh Shah, Samir Patel, and Santosh Kumar Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science*, vol. 1, no. 6, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[20] Al-Zadid Sultan Bin Habib, and Tanpia Tasnim, "An Ensemble Hard Voting Model for Cardiovascular Disease Prediction," *2nd International Conference on Sustainable Technologies for Industry 4.0*, Dhaka, Bangladesh, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[21] Haleh Ayatollahi, Leila Gholamhosseini, and Masoud Salehi, "Predicting Coronary Artery Disease: A Comparison between Two Data Mining Algorithms," *BMC Public Health*, vol. 19, no. 1, pp. 1-9, 2019. [CrossRef] [Google Scholar] [Publisher Link]
[22] G. Ambrish, "Logistic Regression Technique for Prediction of Cardiovascular Disease," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 127-130, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[23] Abid Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol. 9, pp. 39707-39716, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[24] Ashish Bhardwaj, Framingham Heart Study Dataset, Kaggle, 2022. [Online]. Available: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset/
[25] Manu Siddhartha, Heart Disease Dataset (Comprehensive), Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final
[26] Yu Lin Hsu, Z-Alizadeh Sani Dataset (2).Csv, Kaggle, 2018. [Online]. Available:https://www.kaggle.com/datasets/tanyachi99/zalizadeh-sani-dataset-2csv