INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

HEART-PREP: Holistic Ensemble-based Algorithm for Robust Transformation and Preprocessing of Heart Disease Datasets

S Merlin Sofia[1]*, D Ravindran[2], G Arockia Sahaya Sheela[3]

1 Research Scholar, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620002, Tamil Nadu, India
2 Associate Professor, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620002, Tamil Nadu, India
3 Assistant Professor, Department of Computer Science, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli-620002, Tamil Nadu, India

OPEN ACCESS

Received: 15/08/2025

Accepted: 16/09/2025

Published: 08/10/2025

Citation: Merlin Sofia S, Ravindran D, Arockia Sahaya Sheela G (2025) HEART-PREP: Holistic Ensemble-based Algorithm for Robust Transformation and Preprocessing of Heart Disease Datasets. Indian Journal of Science and Technology 18(37): 2993-3004. https://doi.org/10.17485/IJST/v18i37.1435

* **Corresponding author**.

merlinsofia_ss1@mail.sjctni.edu

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Heart disease is one of the leading causes of morbidity and mortality worldwide, thus early detection and precise prediction are critical for efficient treatment and prevention. Attaining excellent predictive performance is strongly dependent on data quality, which is frequently hampered by missing values, outliers, and irrelevant features. While existing preprocessing approaches attempt to tackle these difficulties, they frequently have shortcomings like as bias, slowness, or less generalizability across datasets. **Objectives:** This study seeks to (1) create an effective preprocessing framework that incorporates dynamic feature selection, imputation, outlier removal, and scaling methods to enhance dataset quality; (2) improve the performance of machine learning classifiers such as SVM, KNN, DT, and RF in predicting heart disease; and (3) compare the suggested method to existing methods to show its dependability, scalability, and contribution to early heart disease detection. **Methods**: This study presents HEART-PREP, an ensemble-based preprocessing framework for heart disease prediction, which has been tested on five datasets (Cleveland, Hungary, Long Beach, Switzerland, and Statlog). Clinical criteria include age, gender, chest discomfort, blood pressure, cholesterol, ECG findings, heart rate, and thalassemia were evaluated. The framework includes RACE for outlier elimination, EMERALD for missing value imputation, Min-Max scaling for normalisation, and DYNAMIC for feature selection. Classification was done with SVM, KNN, DT, and RF, and the results were compared to existing techniques in terms of accuracy, precision, recall, and F1-score. **Finding**: HEART-PREP outperformed SVM (92.26%), KNN (91.07%), DT (92.86%), and RF (95.24%) in accuracy, recall, and F1-score. **Novelty**: HEART-PREP is unique in that it combines dynamic feature selection, powerful preprocessing, and scaling strategies to consistently exceed prior approaches with up to 95.24% accuracy in heart disease prediction.

Keywords: Preprocessing; Imputation; Outlier removal; Feature selection; Ensemble-based algorithms

1 Introduction

Heart disease is the primary cause of death worldwide, necessitating accurate and fast prediction approaches to aid in early detection and clinical decision-making[1]. Machine learning advancements have enabled the creation of predictive models that help healthcare workers analyze clinical datasets[2]. However, the performance of these models is strongly reliant on preprocessing, which serves as the foundation for accurate categorization. Preprocessing steps include data cleaning, normalization, handling missing values, finding outliers, and selecting informative characteristics[3].

Despite their relevance, existing preprocessing approaches have severe drawbacks. For example, standard imputation techniques like mean or median replacement can skew data distributions and create bias[4]. Similarly, traditional outlier detection strategies such as the Interquartile Range (IQR) might fail in noisy datasets or with irregular data patterns[5]. Feature selection strategies based only on statistical heuristics may ignore complicated feature relationships, resulting in unsatisfactory classification findings[6]. These deficiencies suggest that present preprocessing procedures are insufficient to maximize the performance of prediction models in heart disease categorization.

To fill these deficiencies, this paper introduces HEART-PREP (Holistic Ensemble-based Algorithm for Robust Transformation and Preprocessing of Heart Disease Datasets), an ensemble-driven preprocessing framework designed to improve the integrity, completeness, and quality of heart disease datasets. Unlike conventional approaches, HEART-PREP incorporates advanced ensemble procedures for missing value imputation, anomaly detection, and multi-criteria feature selection. HEART-PREP improves the accuracy, precision, recall, and F1-scores of popular classifiers like SVM, KNN, DT, and RF by systematically refining the dataset prior to classification. To confirm its efficacy, HEART-PREP is compared to standard preprocessing techniques and existing improved approaches, which show considerable improvements in predicting performance.

The HEART-PREP algorithm integrates numerous innovative methods:

- **EMERALD Algorithm:** Ensemble Technique for Regression-based Assessment of Lost Data utilizing Weighted Prediction, for robust missing data imputation.
- **RACE Algorithm:** Robust Statistics and Clustering Ensemble, for precise outlier detection and elimination.
- **DYNAMIC Algorithm:** Diverse Integrated Multi-criteria Ensemble Feature Selection, for optimal feature subset selection.

In this paper, we propose HEART-PREP, a comprehensive preprocessing framework, specifically customized for heart disease datasets, which contributes to the field. It intends to demonstrate superior performance than the methods available in literature, providing strong support for the identification of improvements in the discrimination and calibration of predictive models for clinical decision support systems and epidemiological studies focused on prognosis for treatment and management of heart diseases.

The present work is justified by the shortcomings of existing heart disease prediction systems, which frequently rely on isolated preprocessing procedures and lack consistency between classifiers. Previous research obtained great accuracy but were limited by dataset noise, redundant features, and imbalanced scaling, resulting in performance discrepancies. HEART-PREP solves these issues by proposing a complete preprocessing system that uses dynamic feature selection, Min-Max scaling, EMERALD, and RACE to improve data quality and classifier performance. This work establishes a dependable and generalizable approach by continuously enhancing accuracy, precision, recall, and F1-score across SVM, KNN, DT, and RF, adding both methodological innovation and practical benefit to medical data analytics.

Preprocessing of heart disease datasets has been the subject of a great deal of interest in recent years, due to the complexity and difficulty of Electronic Health Records. Sorkhabi et al.[7] state that EHRs represent an important form of Medical Big Data (MBD) and offer great potential for precision medicine advancement. But because they tend to be quite ”dirty”, these records often are not ideal candidates for direct mining. To address this, the authors present the systematic pre-processing technique PEPMED, which greatly improves data quality and usefulness by hybrid methods employing specific approaches to challenges presented by EHRs. To remedy the missingness (a prevalent problem in heart disease datasets), Zemlianyi and Angelpreethi[8] compare several imputation algorithms specifically designed for coronary heart disease data. They claim that the typical behavior of removing samples with missing values (for example, NumPy or Pandas) is an overkill and can discard important information.

Their focus is not to propose a new imputation theory, as a variety of imputations are already offered, such as mean/median imputation, k-nearest neighbors (KNN), multiple imputation by chained equations (MICE), and deep learning-based imputations. Their study shows that methods such as fillna_2steps_rg bring time advantages and higher classification accuracies, meaning that the use of purposeful imputation strategies could improve the data analysis outcomes by a large margin. Similarly, S. Parimala et al[9] presented a novel framework addressing missing data management in EHR using the GPLVM and SVM. Their method is not only accurate in estimating the missing values, but it also couples uncertainty in the predictions, making it

a robust solution for prediction-based predictions of heart failure hospital readmission prediction. Their results indicate higher accuracy of imputation and superior predictive power as opposed to traditional methods.

In a comparative study, Berkelmans and Angelpreethi et al.[10] Examine the effectiveness of various imputation techniques in clinical practice. Their findings reveal that simpler methods like median imputation perform comparably to more complex approaches, provided that critical predictor variables are not missing. This suggests that, in practical settings, straightforward imputation methods can be as effective as complex algorithms while being easier to implement and interpret.

Liu et al.[11] investigates the impact of feature selection on missing value imputation in medical datasets. They demonstrate that combining feature selection with imputation improves pattern recognition and predictive accuracy. Feature selection is crucial—genetic algorithms and information gain are appropriate for low-dimensional data, whereas decision trees are more effective for high-dimensional datasets.

In the realm of anomaly detection, Nanehkaran et al.[12] propose a density-based unsupervised approach using DBSCAN clustering for identifying outliers in heart disease data. Their method adapts clustering parameters to achieve high accuracy in diagnosing cardiac abnormalities, showing an improvement over traditional clustering methods that struggle with noise sensitivity and initial point selection.

Feature selection, a critical preprocessing step, has been the focus of several studies. Javeed et al.[13] introduce a novel feature selection method using a floating window with adaptive size, which, when combined with neural network classifiers, significantly enhances heart disease prediction accuracy. This method outperforms numerous existing techniques, demonstrating its effectiveness in refining feature sets for better model performance.

Dissanayake and Md Johar[14] conducted a comprehensive evaluation of feature selection techniques on heart disease prediction. Backward feature selection via decision trees attained the highest accuracy and precision, highlighting the significance of optimum feature subset selection.

Spencer et al.[15] investigated the relationship between feature selection and categorization by PCA, Chi-squared tests, and Relief. Their findings underscore the variability in performance depending on the combination of feature selection and machine learning algorithms, with the best results achieved using Chi-squared feature selection and BayesNet.

Kadhim and Radhi[16] emphasize the importance of early heart disease detection and propose a model that leverages optimized machine-learning algorithms. Their approach, which includes patient data processing, model training, and hyperparameter optimization, achieves high accuracy, particularly with the Random Forest algorithm, underscoring the potential of machine learning in enhancing clinical decision-making.

Angelpreethi[17] used a lexicon and machine learning-based comparison evaluation to classify students' perspectives on the Covid-19 epidemic, demonstrating the utility of hybrid techniques for handling large-scale opinion data. Similarly, Angelpreethi[18] used fuzzy linguistic hedges for sentiment categorization, demonstrating how fuzzy-based approaches can improve decision-making in uncertain situations.

Ganie et al.[19] combined ensemble learning and explainable AI across several datasets to improve prediction reliability, whereas Govindu et al.[20] tested various ensemble models for performance improvements. Patra et al.[21] proposed vision-based transformer ensembles, which demonstrated the role of deep learning in cardiovascular detection. Ashika and Hannah Grace[22] suggested a stacked ensemble with MCDM-based ranking to optimize predictions using RST-ML integration. Zaidi et al.[23] created HeartEnsembleNet, a hybrid ensemble architecture for evaluating cardiovascular risk. Kumar et al.[24] merged classical and quantum-inspired approaches in a hybrid framework to demonstrate cross-disciplinary innovation. Similarly, Naz et al.[25] and Bihri et al.[26] used stacking-based ensemble models with explainable AI to make robust and interpretable predictions. Finally, Sakyi-Yeboah et al.[27] used supervised ensemble tree methods to validate ensemble learning's efficacy in heart disease categorization.

Despite these advances, existing methods have numerous limitations, including biases proposed during imputation, ineffective management of intricate data distributions in outlier detection, and inadequate consideration of synergistic feature relationships in conventional feature selection techniques. The proposed HEART-PREP solution overcomes these limitations by using complex ensemble selection algorithms specifically designed for heart disease datasets. HEART-PREP effectively enhances data quality and predictive performance, and hence, yields more reliable clinical decision support systems (CDSS) and epidemiological studies for the prognosis and management of heart diseases by exploiting methods such as EMERALD to achieve robust imputation, RACE to achieve accurate outlier detection, and DYNAMIC to achieve optimum selection of features.

2 Methodology

2.1 HEART-PREP algorithm

The HEART-PREP algorithm (which is shown in Algorithm 1) is a collection of stages utilized to construct heart disease dataset from numerous sources therefore it can be evaluated. First, this algorithm combines five datasets ('cleveland.csv', 'hungarian.csv', 'longbeach.csv', 'switzerland.csv', and 'statlog.csv') into one, and imputes missing values, eliminates unwanted data instances, normalizes all values to a common range (0 to 1), and also this algorithm selects the very significant attributes for deeper exploration.

The Cleveland dataset (cleveland.csv) originally contained 303 records with 14 attributes. The Hungarian dataset (hungarian.csv) contains 294 records that share the same 14 commonly used attributes as Cleveland. The Long Beach dataset (longbeach.csv) contains 200 records which typically uses 14 attributes. The Switzerland dataset (switzerland.csv) includes 123 records with 14 commonly used attributes. Finally, the Statlog dataset (statlog.csv), also known as the Heart Disease Dataset - Statlog Project, is made up of 270 records, each with 13 attributes with one target variable, for a total of 14.

Algorithm 1: HEART-PREP algorithm

Input : Five heart disease datasets: 'cleveland.csv', 'hungarian.csv', 'longbeach.csv', 'switzerland.csv', and 'statlog.csv'.

Output : A dataset, containing the selected features, is unified and refined.

Step 1 : **Data Loading and Merging:**
- Load data from the five specified heart disease datasets.
- Eliminate duplicates and retain only one header row to maintain the dataset uniformed.

Step 2 : **Handling Missing Data:**
- Impute missing values employing the EMERALD Algorithm. // **Algorithm 2**

Step 3 : **Outlier Removal:**
- Use the RACE Algorithm to identify and remove outliers. // **Algorithm 3**

Step 4 : **Normalization:**
- Apply Min-Max normalization to convert dataset values within the range of 0 to 1.

Step 5 : **Feature Selection:**
- Utilize the DYNAMIC Algorithm to pick the most pertinent attributes. // **Algorithm 4**

Step 6 : Return the combined pre-processed dataset comprising the picked attributes.

The HEART-PREP algorithm starts by loading and merging the five heart disease data sets: 'cleveland.csv', 'hungarian.csv', 'longbeach.csv", switzerland.csv', and 'statlog.csv'. This will begin by reading the five data sets into memory and combining them into one overall data set. It is important to ensure that all of the header information is consistent and to remove any duplicate entries that may be the result of the data points that overlap between the data sets. This builds a well-prepared and refined dataset for accurate management in subsequent algorithm steps.

After the datasets are merged, the next problem is missing data, which is a common issue in medical datasets. Thus, the HEART-PREP algorithm fills the missing data with the EMERALD method (Ensemble Method for Regression-based Assessment of Lost Data Using Weighted Prediction). EMERALD starts by splitting the Data into complete ('train Data') and missing ('test Data'). For every feature that has missing values, three regression models (M5P Tree, Gaussian Processes, and Linear Regression) are trained on a part of the whole dataset. Each model's performance is measured by the Mean Squared Error (MSE), and the weighted prediction is calculated by inverse MSEs. The weighted ensemble prediction is used to fill in the missing values, which is more reliable than using only one model.

After handling the missing values, the HEART-PREP algorithm continues with the outlier detection and removal part with the help of the RACE (Robust Anomaly Detection and Cluster-based Ensemble) algorithm. RACE first finds the medians and Median Absolute Deviations (MADs) for each feature and then finds the enhanced Z-scores. The anomalies are then found by taking the 95th percentile of these z-scores. An improved k-means clustering method, which finds the distance of each data point and centroid of clusters using the Manhattan distance, is used by RACE in order to eliminate the outliers. The data points that appear farther away from their respective cluster centroids than a certain threshold are marked as anomalies. Thus, RACE unites statistical methods with clustering-based techniques and removes the outliers from the dataset.

Normalization: It is a part of the machine learning pipeline preprocessing. With the HEART-PREP algorithm, we apply Min-Max Normalization to all feature values, such that each feature value is included in the range between 0 and 1. Thus, we guarantee that features with larger scales do not dominate over others during the machine learning model training. DYNAMIC (Diverse Integrated Multi-criteria Ensemble Feature Selection): In the final step of preprocessing, we select the most pertinent

features from the data using DYNAMIC. DYNAMIC is a combination of multiple feature selection techniques to provide a robust set of features.

Features are assessed by classifier subset results, target correlation, and ranking in classifiers. As a finding, only the most significant and predictive features are kept. It shrinks dimensionality, reduces overfitting, as well as optimizes model clarity with findings. Finally, the HEART-PREP algorithm gets a refined dataset which has the selected features after the entire preprocessing stages are completed. For subsequent machine learning procedures, this new dataset can be used.

HEART-PREP resolves data quality problems based on missing values imputation, noise and outliers' removal to enable accurate as well as dependable heart disease prediction models. This flow diagram showed in Figure 1.

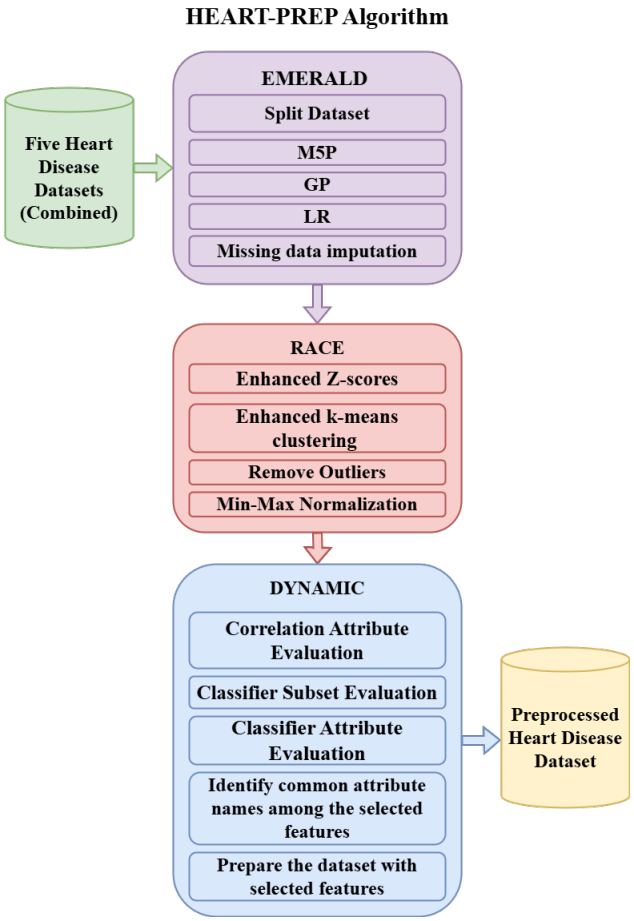

Fig 1. System Architecture of HEART-PREP Algorithm

2.2 EMERALD Algorithm

The EMERALD algorithm is the algorithm utilized to impute missing values which evaluates and fills gaps thus the dataset becomes usable and complete (Algorithm 2). This is especially important in fields such as medical research, where gaps in patient records can substantially impact the results of analyses and predictions.

The Emerald algorithm is demonstrated in Algorithm 2.

Algorithm 2: EMERALD Algorithm

Input : Dataset with missing values (‘data‘)
Output : Imputed dataset (‘imputed Data‘)
Step 1 : Divide the dataset into complete (train Data) and incomplete (test Data) data:
- ‘train Data‘: Rows with no missing values.
- ‘test Data‘: Rows with missing values.

Step 2 : Identify attributes with missing values (missing Attributes).
Step 3 : For each missing attribute in ‘missing Attributes‘:

a. Split ‘train Data‘ into ‘train Subset‘ (70%) and ‘validation Subset‘ (30%).
b. Train regression models on ‘train Subset‘:
- Train an M5P Tree (M5P) model.
- Train a Gaussian Process (GP) model.
- Train a Linear Regression (LR) model.

c. Evaluate model performance on ‘validation Subset‘:
- Calculate Mean Squared Error (MSE) for each model:

$$\text{MSE}_{\text{M5P}}, \text{MSE}_{\text{GP}}, \text{MSE}_{\text{LR}}$$

- Calculate the inverse MSE for each model:

$$\text{InvMSE}_{\text{M5P}} = \frac{1}{\text{MSE}_{\text{M5P}}}, \text{InvMSE}_{\text{GP}} = \frac{1}{\text{MSE}_{\text{GP}}}, \text{InvMSE}_{\text{LR}} = \frac{1}{\text{MSE}_{\text{LR}}}$$

- Calculate the sum of inverse MSEs:

$$\text{SumInvMSE} = \text{InvMSE}_{\text{M5P}} + \text{InvMSE}_{\text{GP}} + \text{InvMSE}_{\text{LR}}$$

- Calculate normalized weights for each model:

$$\text{Weight}_{\text{M5P}} = \frac{\text{InvMSE}_{\text{M5P}}}{\text{SumInvMSE}}, \text{Weight}_{\text{GP}} = \frac{\text{InvMSE}_{\text{GP}}}{\text{SumInvMSE}}, \text{Weight}_{\text{LR}} = \frac{\text{InvMSE}_{\text{LR}}}{\text{SumInvMSE}}$$

d. Predict missing values in ‘test Data‘ using the trained models:
- Get predictions from each model for the missing values:

$$\hat{y}_{M5P}, \hat{y}_{GP}, \hat{y}_{LR}$$

e. Calculate the weighted ensemble imputed value:
- Compute the weighted ensemble prediction for each missing value:

$$\hat{y}_{ensemble} = Weight_{M5P} \cdot \hat{y}_{M5P} + Weight_{GP} \cdot \hat{y}_{GP} + Weight_{LR} \cdot \hat{y}_{LR}$$

Step 4 : Replace missing values in ‘test Data‘ with the ensemble imputed values.
Step 5 : Combine ‘train Data‘ and imputed ‘test Data‘ to create the imputed dataset (‘imputed Data‘).
Step 6 : Return the imputed dataset (‘imputed Data‘).

The algorithm begins by segmenting the dataset into two parts: ‘train Data‘, which contains rows with no missing values, and ‘test Data‘, which includes rows with missing values. This way, the algorithm can train and validate the models on the entire dataset. In order to segment the dataset into two parts, the algorithm first finds all the features (or attributes) that have missing attributes. This ’missing attributes’ identification step facilitates imputation by forcing attention to certain attributes that need to be imputed. This way, the algorithm can focus on attributes that need attention. For every attribute that has missing values, the EMERALD algorithm follows a training and validation procedure. The entire data set (’train Data’) is split into two parts: ’train Subset’ (70%) and ’validation Subset’ (30%). ’train Subset’ is trained on three different regression models: M5P Tree (which is a combination of decision trees and linear regression functions), Gaussian Processes (a non-parametric model based on Gaussian distributions) and Linear Regression (variables relationship is modelled linearly). The performance of the three

models is evaluated by using the Mean Squared Error (MSE) on the 'validation Subset'. Mean Squared Error is a commonly used measure to evaluate the performance of regression models. It computes the average of the squared errors, the error being the difference between the observed and predicted results. The MSE equation is shown in **Eq. (1)**:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

Where:

- n is the number of data points.
- y_i is the actual value of the i-th data point.
- $\hat{y}_i$ is the predicted value of the i-th data point.

The lower the MSE value, the better the model's prediction. To show performance, the algorithm computes the inverse of each model's MSE. Here, maximum values indicating superior findings. For allocating ensemble weights which sum to one, these inverse MSEs are normalized, ensuring that each model's contribution to the end result is proportional to its precision. Following that, regression models forecast missing values from the 'test Data'. These values are imputed utilizing the weighted ensemble strategy, which takes advantage of each model's advantages to provide more accurate imputations. To create the ultimate complete dataset prepared for evaluation, the imputed 'test Data' is combined with the 'train Data'. As a result, the EMERALD algorithm offers a dependable, extensive approach to missing data imputation. This algorithm combines numerous models to generate precise and excellent forecasts.

2.3 RACE Algorithm

To identify and eliminate outliers from the dataset, the RACE algorithm is the technique utilized, guaranteeing that abnormal or extreme values don't distort evaluation or model efficiency.

Algorithm 3 showed the RACE algorithm.

Algorithm 3: RACE Algorithm
Input : Imputed Dataset
Output : Outlier Removed Dataset
Step 1 : Data Preparation:
- Parse the imputed dataset into a structured numeric data matrix.

Step 2 : Statistical Processing:
- Calculate medians and MADs for each attribute.
- Compute enhanced Z-scores ('calculate Z Scores')

$$EZ = \frac{x - median}{MAD}$$

Step 3 : Anomaly Detection with Enhanced Z-scores:
- Using the 95th percentile of absolute enhanced Z-scores threshold is calculated.
- Identify anomalies ('detect Anomalies') where any enhanced Z-score exceeds the threshold.

Step 4 : Enhanced Clustering-Based Anomaly Removal:
Apply enhanced k-means clustering ('cluster-based Based Anomaly Removal'):
- ◆ Set the number of clusters.
- ◆ Use Manhattan distance to calculate distances between instances and their cluster centroids.
- ◆ Define an anomaly threshold based on the 95th percentile of distances to centroids.
- ◆ Flag instances as anomalies if their distance exceeds the threshold.

Step 5 : Combine Anomaly Indices:
- Create a set 'Anomaly Indices' to store unique anomaly indices detected by both enhanced Z-score analysis and enhanced clustering.

Step 6 : Generate Output Dataset:
- Exclude instances identified as anomalies based on their indices in the 'Anomaly Indices' set.
- Format the cleaned dataset into 'outlier Removed Dataset', excluding identified anomalies.

Step 7 : Output Results:
- Return 'outlier Removed Dataset', representing the cleaned dataset with outliers removed.

This algorithm begins with data preparation, which involves parsing the dataset (typically in CSV format) into a structured numeric data matrix.

The parameters in the HEART-PREP model were deliberately kept simple and focused to ensure interpretability along with clinical relevance. Each parameter selected directly improves the quality of the unified dataset. Data loading and merging were required to combine multiple heart disease datasets into a consistent structure, eliminating redundancy and ensuring a solid foundation. Handling missing data with the EMERALD algorithm was critical, as incomplete medical records are common and can skew results if not addressed. Outlier removal with the RACE algorithm was used to remove abnormal or noisy entries that could distort classifier learning and reduce prediction reliability. Normalization using Min-Max scaling was chosen to standardize feature ranges, allowing for equal contribution from all attributes with improving classifier convergence. Finally, feature selection using the DYNAMIC algorithm was critical for reducing dimensionality, removing irrelevant attributes, and retaining only the most important predictors of heart disease, thereby improving classification efficiency and accuracy. These parameters were carefully chosen because they all address a critical preprocessing challenge in medical datasets, which ensures data integrity, consistency, and improved downstream performance.

At this point, you should be ready to apply statistical processing and clustering methods, but each feature/record must be in the correct column. In this step, the algorithm proceeds to statistical processing, computing medians and Median Absolute Deviations (MADs) of each feature. The median is a point estimate of central tendency that is less sensitive to outliers than the mean; MAD measures variation about the median and is also less sensitive to outliers than standard deviation. Based on these metrics, improved Z-scores are computed for each attribute. This enhancement over typical Z-scores, which typically employ mean and standard deviation, renders the computation more powerful by concentrating on characteristics that are less sensitive to outliers, the median and MAD, thus optimizing the trustworthiness of the anomaly identification.

The algorithm then applies these improved Z-scores to the anomaly detection process. The 95th percentile of absolute enhanced Z-scores determines the dynamic threshold for anomaly detection, obtaining variations from the median. Anomalies are flagged by Points that exceed this threshold, with the MAD and median used to ensure dependability. By first calculating cluster count, the RACE algorithm improves on the k-means procedure. And also utilizing Manhattan distance to measure data-centroid proximity for stronger detection.

The Manhattan distance measure, which computes absolute differences in each dimension and adds them up (as opposed to Euclidean distance, which computes the shortest distance in a straight line), is particularly useful for cases where it is important to emphasize overall structural deviations in data. The improved k-means clustering method is preferred over Euclidean distance because the absolute difference in any single dimension is less influential compared to the difference in any other dimension, making it more likely to reliably detect structural anomalies in the dataset.

After that, an anomaly threshold is calculated based on the 95th percentile of the distances. An anomaly is any instance in which the distance is greater than that threshold. To optimize detection accuracy, this clustering identifies anomalies utilizing Manhattan distance. To guarantee uniqueness as well as eliminating duplicates, the indices from Z-score evaluation and clustering are integrated utilizing a set structure. Anomalies are eliminated, as well as the refined dataset resulted in the similar valid format as the input and also prepared for subsequent evaluation. By combining clustering as well as statistical measures, the RACE algorithm offers these 3 benefits namely 1) eliminates outliers dependably, 2) dataset quality improvement, 3) accuracy analysis and machine learning efficiency.

2.4 DYNAMIC Algorithm

The DYNAMIC algorithm is the feature selection technique which is utilized to select the most important features in the dataset. This algorithm decreases dimensionality and enhances the precision and performance of model.

The DYNAMIC feature selection algorithm is used at the end of the HEART-PREP algorithm, not at the beginning. The reasoning is that feature selection works best with a clean with consistent dataset that is free of missing values, duplicates,

outliers, also including scale imbalances. If features are selected prior to preprocessing, they may be biased because of noise, incomplete records, or distorted value ranges, resulting in the discarding of relevant attributes with the retention of irrelevant ones. By first unifying the datasets, imputing missing data (using EMERALD), eliminating outliers (using RACE), and normalising values, we ensure that the DYNAMIC algorithm evaluates features on a refined dataset. This sequencing improves the reliability of the chosen features, thereby improving the robustness of the downstream classification models.

Algorithm 4 discussed the DYNAMIC algorithm.

Algorithm 4: DYNAMIC Algorithm
Input : Normalized Dataset
Output : Dataset with selected features.
Step 1 : Load and Prepare Data:
- Convert the normalized Dataset into an Instances object, facilitating programmatic manipulation of the dataset.

Step 2 : Set Class Index:
- Identify the column representing the class label.

Step 3 : Feature Selection Methods:
- Correlation Attribute Evaluation:
  - Evaluate attribute correlation with the class label.
  - Use a 'Ranker' to select top attributes by correlation.
- Classifier Subset Evaluation:
  - Train a classifier to evaluate attribute subsets.
  - Apply 'Greedy Stepwise' to optimize attribute selection.
- Classifier Attribute Evaluation:
  - Rank attributes using a classifier.
  - Select top attributes with a 'Ranker'.

Step 4 : Combine Selected Features:
- Create lists to store attributes selected by the different methods: correlation, classifier subset, and classifier attribute evaluations.
- Identify attributes that are common among the methods (common Attribute Names).

Step 5 : Modify Dataset:
- Retain only attributes in common Attribute Names.
- Generate a dataset With Selected Features from retained attributes.

This procedure begins with loading and preparing the dataset, which is typically presented in CSV format, into a programmatically manipulable object, like Weka's 'Instances' object. This first step is critical because it guarantees that the dataset is properly formatted for subsequent processing and evaluation, preserving the data's integrity throughout the feature selection procedure.

Once the dataset is loaded, the algorithm will discover the class index. The class index detects the dependent feature. It enables the algorithm to split features from class labels. As well as it concentrates feature selection on factors impacting the target. For feature selection, numerous methods used in the DYNAMIC algorithm: By correlation with the class label, Correlation Attribute Evaluation ranks attributes; To evaluate feature sets for cumulative influence Classifier Subset Evaluation uses a classifier with Greedy Stepwise; and utilizing classifier-based predictive value Classifier Attribute Evaluation ranks attributes. To ascertain that the selected features are significant on an individual basis and synergistically efficient, attributes chosen by various methods are preserved, guaranteeing their significance both independently and collectively. Then, the dataset is shrinked to these common features, building a pre-processed "selected Data" set. This Multiple standards method improves predictive capacity and data excellence based on the combination of correlation evaluation, subset assessment, and classifier-based ranking. For excellent heart disease datasets appropriate for machine learning, the HEART-PREP algorithm constructs on this by including novel and sophisticated imputation technique, efficient outlier detection technique, normalization technique, and excellent feature selection technique.

3 Experimental Results and Discussions

This section compares the heart disease classification outcomes of Kadhim et al.'s[16] enhanced algorithms with the HEART-PREP algorithm. As well as they are assessed on accuracy, precision, recall, and F1-score across SVM, KNN, DT, and RF. Both algorithms were executed in Java utilizing Weka. The results explicitly present classifier-specific metrics, such as accuracy, precision, recall, and F1-score, for both SVM, KNN, DT, along with RF. As shown in Figures 2 and 3, HEART-PREP consistently outperforms Kadhim et al.'s approach, achieving superior results across all evaluation metrics. HEART-PREP, in particular, achieves an average accuracy gain of +2.26% with noticeably improved precision, recall, and F1-score, with the most significant improvement observed in DT recall, which increased by +0.07. These consistent improvements demonstrate that HEART-PREP not only replicates but also strengthens the benefits of previous preprocessing strategies, providing more consistent with balanced performance across multiple classifiers. This direct and metric-driven comparison with existing research demonstrates the HEART-PREP pipeline's effectiveness and superiority in advancing heart disease classification.

Kadhim et al. attained high accuracy via sophisticated preprocessing. And also, they improved classifier performance, with SVM attaining 89.0% accuracy (0.87 precision, 0.93 recall, 0.90 F1) and KNN attaining 88.6% accuracy (0.87 precision, 0.92 recall, 0.89 F1), while DT and RF also showed efficient performance. DT got 89.9% accuracy, 0.93 precision, 0.86 recall, and 0.87 F1-score. RF got 94.9% accuracy, 0.94 precision, 0.96 recall, and 0.95 F1-score.

Figures 2 and 3 indicate that HEART-PREP did better than Kadhim et al.'s[16] approach on all metrics for SVM, KNN, DT, and RF. HEART-PREP achieved 92.26% accuracy, 0.92 precision, 0.92 recall, and an F1-score of 0.92 for SVM; 91.07% accuracy, 0.91 precision, 0.91 recall, and an F1-score of 0.91 for KNN; 92.86% accuracy, 0.93 precision, 0.93 recall, and an F1-score of 0.93 for DT; and 95.24% accuracy, 0.95 precision, 0.95 recall, and an F1-score of 0.95 for RF.

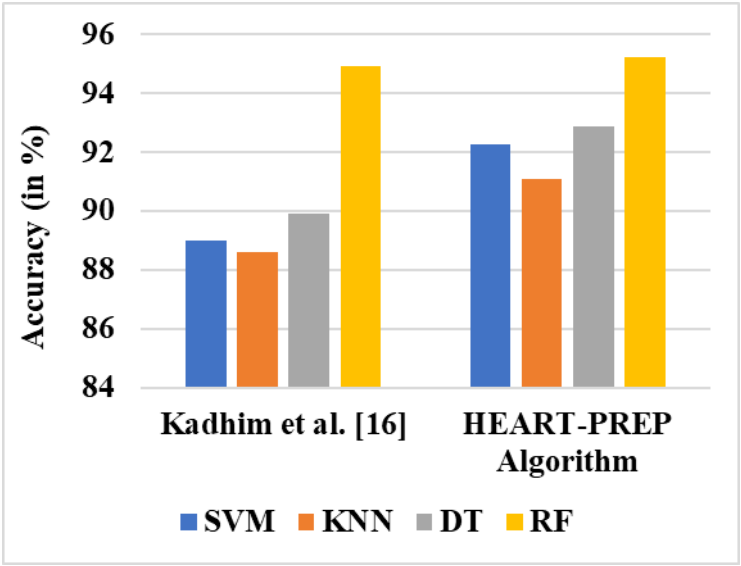

Fig 2. Accuracy Comparison

The superior performance of HEART-PREP can be attributed to its comprehensive preprocessing steps. HEART-PREP ensures data integrity by merging and deduplicating multiple datasets, effectively handles missing data using the EMERALD algorithm, removes outliers robustly with RACE, normalizes data through Min-Max scaling to a consistent range, and selects the most relevant features using the DYNAMIC algorithm. These steps collectively enhance the quality and relevance of the dataset used by classifiers, leading to improved accuracy and balanced precision-recall trade-offs. Compared to older methods by Kadhim et al.[16], this makes HEART-PREP a strong and effective way to classify heart disease.

Overall, HEART-PREP consistently enhances performance across all classifiers: SVM shows a +3.26% accuracy gain (89.00 → 92.26), KNN enhances by +2.47% (88.60 → 91.07), DT increases by +2.96% (89.90 → 92.86), and RF achieves a modest yet meaningful +0.34% improvement (94.90 → 95.24), resulting in an average accuracy gain of +2.26%. Beyond accuracy, HEART-PREP consistently improves precision (+0.025), recall (+0.01), and F1-score (+0.025). The biggest significant improvement is seen in DT recall, which increases from 0.86 to 0.93 (+0.07), showing improved sensitivity in recognizing true positive cases of heart disease. These consistent, model-agnostic, and reproducible benefits emphasize HEART-PREP's uniqueness as

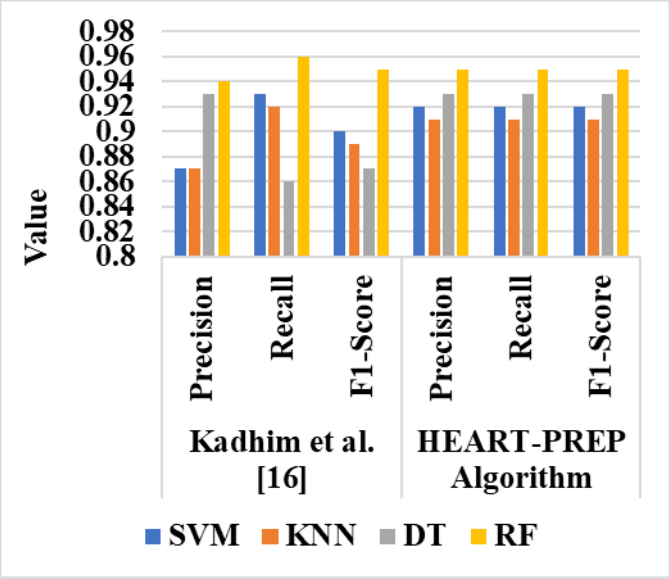

Fig 3. Precision, Recall, and F1-score Comparison

an ensemble-based preprocessing pipeline that improves data integrity and quality, hence increasing downstream classifier performance without changing the underlying architectures.

4 Conclusion

This study proposed HEART-PREP, an ensemble-based preprocessing framework that directly aligns with the study's goal of enhancing dataset quality and predictive accuracy in heart disease detection by combining RACE for outlier removal, EMERALD for missing value imputation, Min-Max scaling for normalization, and DYNAMIC feature selection for identifying influential attributes. Experimental findings on five benchmark datasets showed consistent gains across classifiers, with accuracy enhancements of +3.26% for SVM (92.26%), +2.47% for KNN (91.07%), +2.96% for DT (92.86%), and +0.34% for RF (95.24%), as well as increases in precision (+0.025), recall (+0.01), and F1-score (+0.025), with DT recall demonstrating the greatest increase (+0.07). The unique feature of HEART-PREP is its comprehensive ensemble-driven preprocessing pipeline, which greatly improves classification performance without adjusting model architectures, outperforming traditional preprocessing methods. These results reinforce HEART-PREP's dependability for heart disease prediction and suggest that it might be extended with deep learning-based preprocessing, adaptive feature selection, and hybrid clinical-data integration, paving the way for larger applications in both medical and non-medical areas.

In the future, the HEART-PREP algorithm can be improved by incorporating advanced data augmentation techniques to tackle class imbalance, investigating deep learning-based imputation along with outlier detection methods for increased robustness, and expanding the framework to include additional medical datasets for better generalization. Furthermore, adaptive feature selection tactics and automated hyperparameter optimization can be used to refine preprocessing and improve classifier performance in a variety of clinical applications.

5 Acknowledgement

The authors thank, DST-FIST, Government of India for funding towards infrastructure facilities at St. Joseph's College (Autonomous), Tiruchirappalli-620002.

References

1) Saboor A, Usman M, Ali S, Samad A, Abrar MF, Ullah N. A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*. 2022;2022(1):1410169. https://doi.org/10.1155/2022/1410169.
2) Tougui I, Jilbab A, Mhamdi JE. Heart disease classification using data mining tools and machine learning techniques. *Health and Technology*. 2020;10(5):1137–1144. https://doi.org/10.1007/s12553-020-00438-1.

3) Benhar H, Idri A, Fernández-Alemán JL. Data preprocessing for heart disease classification: A systematic literature review. *Computer Methods and Programs in Biomedicine*. 2020;195:105635. https://doi.org/10.1016/j.cmpb.2020.105635.
4) Hu Z, Du D. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. *PLoS One*. 2020;15(9):e0237724. https://doi.org/10.1371/journal.pone.0237724.
5) Ardeti VA, Kolluru VR, Varghese GT, Patjoshi RK. An outlier detection and feature ranking based ensemble learning for ECG analysis. *International Journal of Advanced Computer Science and Applications*. 2022;13(6). https://dx.doi.org/10.14569/IJACSA.2022.0130686.
6) Ay Ş, Ekinci E, Garip Z. A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases. *The Journal of Supercomputing*. 2023;79(11):11797–11826. https://doi.org/10.1007/s11227-023-05132-3.
7) Sorkhabi LB, Gharehchopogh FS, Shahamfar J. A systematic approach for pre-processing electronic health records for mining: Case study of heart disease. *International Journal of Data Mining and Bioinformatics*. 2020;24(2):97–120. https://doi.org/10.1504/IJDMB.2020.110154.
8) Zemlianyi O, Baibuz O. Methods for imputing missing data on coronary heart disease. *Системні технології*. 2024;2(151):33–49. https://doi.org/10.34185/1562-9945-2-151-2024-04.
9) Parimala S, Kumutha R, Rose MVS, Nanmaran R. Improved Moth Flame Optimization with Deep Convolutional Neural Network for Colorectal Cancer Classification using Biomedical Images;vol. 1. IEEE. 2024. https://doi.org/10.1109/icait61638.2024.10690631.
10) Berkelmans GF, Read SH, Gudbjörnsdottir S, Wild SH, Franzen S, Graaf YVD, et al. Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice. *Journal of Clinical Epidemiology*. 2022;145:70–80. https://doi.org/10.1016/j.jclinepi.2022.01.011.
11) Liu CH, Tsai CF, Sue KL, Huang MW. The feature selection effect on missing value imputation of medical datasets. *Applied Sciences*. 2020;10(7):2344. https://doi.org/10.3390/app10072344.
12) Nanehkaran YA, Licai Z, Chen J, Jamel AA, Shengnan Z, Navaei YD, et al. Anomaly Detection in Heart Disease Using a Density-Based Unsupervised Approach. *Wireless Communications and Mobile Computing*. 2022;2022(1):6913043. https://doi.org/10.1155/2022/6913043.
13) Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ. Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. *Mobile Information Systems*. 2020;2020(1):8843115. https://doi.org/10.1155/2020/8843115.
14) Dissanayake K, Johar MGM. Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Applied Computational Intelligence and Soft Computing*. 2021;2021(1):5581806. https://doi.org/10.1155/2021/5581806.
15) Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. *Digital Health*. 2020;6:2055207620914777. https://doi.org/10.1177/2055207620914777.
16) Kadhim MA, Radhi AM. Heart disease classification using optimized Machine learning algorithms. *Iraqi Journal For Computer Science and Mathematics*. 2023;4(2):3. https://doi.org/10.52866/ijcsm.2023.02.02.004.
17) Angelpreethi A. Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic. *IJIEMR Transactions*. 2023;12(2):380–387. Available from: https://ijiemr.org/public/uploads/paper/590451676720775.pdf. 10.48047/IJIEMR/V12/ISSUE02/60.
18) Angelpreethi A. Fuzzy based sentiment classification using fuzzy linguistic hedges for decision making. *Mapana Journal of Sciences*. 2023;22(Special Issue 2):63–79. https://doi.org/10.12723/mjs.sp2.4.
19) Ganie SM, Pramanik PK, Zhao Z. Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. *Scientific Reports*. 2025;15(1):13912. https://doi.org/10.1038/s41598-025-97547-6.
20) Govindu S, Aswini K, Reddy YA, Gopisankar T. Enhancing Heart Disease Prediction Using Different Ensemble Models. IEEE. 2025. https://doi.org/10.1109/ICICT64420.2025.11005288.
21) Patra R, Dutta S, Roy IK, Basak P, Ghosh A. Heart Disease Detection using Vision-Based Transformer Ensemble Models. *Procedia Computer Science*. 2025;258:3554–3569. https://doi.org/10.1016/j.procs.2025.04.611.
22) Ashika T, Grace GH. Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: an optimized RST-ML approach. *Frontiers in Digital Health*. 2025;7:1609308. https://doi.org/10.3389/fdgth.2025.1609308.
23) Zaidi SA, Ghafoor A, Kim J, Abbas Z, Lee SW. HeartEnsembleNet: An innovative hybrid ensemble learning approach for cardiovascular risk prediction. *Healthcare*. 2025;13(5):507. https://doi.org/10.3390/healthcare13050507.
24) Kumar A, Dhanka S, Sharma A, Bansal R, Fahlevi M, Rabby F, et al. A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques. *Scientific Reports*. 2025;15(1):25040. https://doi.org/10.1038/s41598-025-09957-1.
25) Naz M, Khalid A, Hameed A, Taj R, Mumtaz W, Alotaibi FA, et al. Meta-Ensemble Learning for Heart Disease Prediction: A Stacking-Based Approach with Explainable AI. *IEEE Access*. 2025. https://doi.org/10.1109/ACCESS.2025.3588683.
26) Bihri H, Charaf LA, Azzouzi S, Charaf ME. A Robust Stacking-Based Ensemble Model for Predicting Cardiovascular Diseases. *AI*. 2025;6(7):160. https://doi.org/10.3390/ai6070160.
27) Sakyi-Yeboah E, Agyemang EF, Agbenyeavu V, Osei-Nkwantabisa A, Kissi-Appiah P, Moshood L, et al. Heart Disease Prediction Using Ensemble Tree Algorithms: A Supervised Learning Perspective. *Applied Computational Intelligence and Soft Computing*. 2025;2025(1):1989813. https://doi.org/10.1155/acis/1989813.