International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 1693-1706
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Enhanced Diagnostic Precision for Cardiovascular Diseases through the Synergistic Application of GDE_Lasso Feature Selection and Random Forest Classification Techniques

B. Kalaivani[1]*, A. Ranichitra[2]

[1]Research Scholar Department of Computer Science Sri S.Ramasamy Naidu Memorial College (Affiliated to Madurai Kamaraj University) Sattur-626 203,Tamilnadu,India
* **Corresponding Author Email:** kalaivanisrnm2008@gmail.com - **ORCID:** 0009-0000-1802-4586

[2]Assistant Professor Department of Computer Science Sri S.Ramasamy Naidu Memorial College Sattur-626 203,Tamilnadu,India
Email: ranichitra117@gmail.com - **ORCID:** 0000-0001-6071-0635

Article Info:

DOI: 10.22399/ijcesen.736
Received : 05 January 2025
Accepted : 17 March 2025

Keywords :

Gaussian based Differential Entropy,
Information gain,
BIC,
Modified LASSO,
Random Forest.

Abstract:

Cardiovascular diseases (CVD) pose a significant global health challenge, contributing substantially to mortality rates worldwide. Early detection and diagnosis of CVD are critical, and machine learning techniques offer promising avenues for analyzing risk factors and implementing preventive measures. Feature selection methods can also help reduce diagnostic costs. Hence, in this work, Gaussian-based differential entropy for information gain with the Lasso (GDE_Lasso) feature selection model is proposed. The goal is to optimize diagnostics by streamlining processes, minimizing tests, and enabling targeted interventions. The proposed model is evaluated on Cleveland Datasets 1 and 2, respectively. This work compares the performance of Logistic Regression, Naïve Bayes, SVM, KNN, Decision Tree, XG Boost, and Random Forest for the considered datasets by applying the Z-score method. It was found that Random Forest performs well among the considered classifiers. Therefore, this study evaluates the performance of Random Forest with and without applying the GDE_Lasso feature selection algorithm.

1. Introduction

Cardiovascular disease (CVD) manifests as a medical condition marked by the obstruction of blood vessels, resulting in heart attacks accompanied by chest discomfort. Furthermore, it encompasses a range of other heart-related disorders and carries the risk of heart failure, potentially leading to severe complications, including fatal outcomes [1]. Posing a life-threatening risk, cardiovascular disease (CVD) increases the likelihood of secondary diseases that can damage various organs, including the heart and circulatory system, consequently leading to CVD. Malfunctioning blood vessels and the heart also contribute to disorders like rheumatism, disorders of the brain's vascular system, and heart artery diseases. Individuals suffering from CVD are at higher risk of developing obesity, abnormal lipid levels, increased glucose levels, and high blood pressure [2]. Cardiovascular Disease remains a significant global issue, accounting for 31% of deaths total of 17.7 million in 2015. Projections indicate that this number will rise by 2023, cementing its status as a primary cause of mortality, impacting nearly 20 million individuals annually [3]. The amalgamation of progressive technology in computers and information systems is ready to transform and simplify the management of crucial daily data essential for effective decision-making in the field of medicine [4].

Machine learning methods have received extensive attention in the healthcare area, mostly in the context of evidence-based decision-making within clinical environments. Researchers usually utilize these strategies to forecast the occurrence of heart disease. An essential area of study involves creating innovative systems personalized for using artificial intelligence to forecast heart disease [5]. Effective application of machine learning pivots on the use of reliable training and testing data. In a research that explores heart failure within a clinical decision

support system, various classifier models like Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree, XG Boost, and Random Forest were assessed using a classification approach [6].

Healthcare datasets are derived from a variety of medical sources like lab results, medications, diagnoses, and procedures. These data are grouped based on specific characteristics. It is vital in research applications, particularly in datasets containing numerous variables. The method highlighted in many studies significantly improves the precision of classifier models. Additionally, it aids in pinpointing the essential predictive features [7]. Physicians can successfully determine the severity of a problem by classifying it based on predetermined criteria. The optimization of feature selection contributes to enhancing prediction accuracy [8]. By utilizing advanced selection techniques such as data mining, Information Gain-driven differential entropy, and LASSO, data preparation is optimized for precise forecasts. Our main goal is to use LASSO and entropy-based Information Gain feature selection to address the issues with the considered cardiovascular datasets. Our objective is to address challenges with two distinct cardiovascular datasets, utilizing entropy-based Information Gain ratio feature selection and LASSO. This method elevates the dependability of predicting cardiovascular diseases while addressing problem of overfitting and underfitting commonly encountered in machine learning [9].

Measurement of the ambiguity surrounding continuous variables, especially in the context of feature selection, is a fundamental application of differential entropy in the area of machine learning. Differential entropy is used to rank the features according to their significance and dimensions to discriminate between groups or categories in datasets. This quantitative evaluation is important for regression and classification tasks because it directs the choice of features with greater information gain, which reflects lower entropy and enhances the predictive modelling effectiveness.

The effectiveness of the work lies in its emphasis on the significance of efficient feature selection for enhancing classification accuracy and predictive precision. It leverages machine learning techniques that have demonstrated success in predicting cardiac disease, drawing inspiration from prior research. However, it innovates by integrating data from two different datasets and employing several feature selection techniques. The goal is to identify the best predictive models for heart disease prediction, potentially benefiting the medical community, through comparison and evaluation based on performance measures. This research highlights the importance of feature selection and presents a three-phase approach involving outlier removal, Gaussian-based differential entropy for information gain based feature selection, and the application of the Modified LASSO algorithm. Predicting cardiovascular disease is significantly improved when the chosen features are incorporated into a Random Forest model. Evaluation metrics such as Accuracy, Precision, Recall, and F1-score are utilized to assess the model performance and compare it with existing methods.

The paper is structured in the following manner: Section II offers a literature review of existing studies, specifically focusing on heart disease prediction models. This review pinpoints research gaps, which are then addressed in the complexity detection part. In Section III, the paper details the proposed flow, and methodology, and presents the pseudo-code of the study. Subsequently, Section IV explores the dataset explanation, conducts performance investigation, and comparative analysis, and provides an in-depth discussion of the results. Finally, Section V concludes the paper and puts forward suggestions for future research.

2. Literature review

Machine learning faces challenges in effectively diagnosing heart disease due to the complexity of managing datasets with numerous features, leading to issues such as overfitting and computational challenges. Feature selection is a crucial technique employed to identify and use the most relevant features, improving prediction accuracy, reducing noise, and enabling more informed decisions in personalized treatment plans. The discussion in this section also covers prediction models based on ML classifiers.

Salman Pathan, M., et al. [10] obtained an overall stroke prediction accuracy of 80% by utilizing the Cardiovascular Health Study (CHS) dataset. They employed principle component analysis to lessen the dimensionality of the feature space, the DT technique to perform feature selection, and the MLP network to build a classification model. After being trained on the best feature set, the model beat previous methods, recognizing strokes with an outstanding accuracy of 97.7%.

Bsoul, M.A., et al. [11] using UCL clinical datasets, logistic regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Multi-Layer Perceptron (MLP) artificial neural network were employed to predict coronary heart disease. Data pre-processing was conducted using the Synthetic Minority Oversampling Technique (SMOTE) and the K-means approach. SMOTE was

combined with recursive feature selection techniques and the Genetic Algorithm (GA). Among these methods, random forest (RF) exhibited an accuracy of 86.6%.

Sumwiza, K., et.al. [12] suggest an ensemble approach that combines various feature selection approaches with the Random Forest (RF) algorithm. The strategy comprises using training datasets to develop a cardiovascular disease prediction model, handling missing data, and using data mining and correlation coefficients to remove outliers. With an astounding accuracy of 99%, this model surpasses K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Logistic Regression (LR) models.

Jabbar, M.A., et al. [13] formulated a predictive model for heart disease utilizing statlog heart disease dataset by integrating feature selection methods with the random forest model. This dataset consists of 270 instances with features similar to the Cleveland dataset. The authors utilized backward elimination with Chi-squared feature assessment. This approach builds and tests models until the accuracy stops improving by applying the Chi-squared test to rank the features. The lowest-ranked feature is then deleted repeatedly. Their best-performing model had an accuracy of 83.7%.

Iscra, K.., et al. [14] investigate the complex early-stage differential diagnosis between ischemic heart disease (IHD) and Dilated Cardiomyopathy (DCM) by utilizing Left Ventricular Ejection Fraction (LVEF) and Heart Rate Variability (HRV) analysis. Despite the potential of machine learning, clinicians are faced with challenges stemming from the lack of transparency in black-box models. The research compares the performance of interpretable models, including classification tree, logistic regression, and naïve Bayes algorithms, on a sample of 196 IHD and 117 DCM subjects. Notably, the naïve Bayes model achieves the highest accuracy at 73.5%, providing interpretable results crucial for clinical decision-making.

Saw, M., et al. [15] provide to the discourse surrounding heart disease-related mortality, emphasizing the importance of leveraging data analysis to convert collected data into actionable insights. To increase the accuracy of heart disease prognosis, the study makes use of a healthcare dataset and machine learning's Logistic Regression technique. As evidenced by its 87.02 percent accuracy rate when applied to the Framingham datasets, the literature highlights the effectiveness of the logistic regression model in classifying heart diseases.

Mehmood, A., et al. [16] suggests a novel CVD probability prediction model called CardioHelp, based on DL-CNN has been developed. Utilizing a feed-forward CNN model with a single input and single output, it focuses on temporal modelling for early heart disease diagnosis. The incorporation of the Least Absolute Shrinkage (LASSO) technique in feature selection enhances dataset interpretability and prediction accuracy. The results demonstrate an impressive accuracy of 97%, outperforming existing models.

Sharawi, M., et al. [17] which was impacted by humpback whale behaviour, offers a feature selection technique based on the Whale Optimization Algorithm (WOA). The model uses a wrapper-based method to determine which feature subset is best for maximizing classification accuracy while decreasing features. The study demonstrates the superiority of WOA in achieving optimum feature combinations when compared to PSO and GA across 16 datasets. This illustrates the balanced approach that WOA takes between exploration and exploitation.

The examination of current literature on the diagnosis of cardiovascular disease has revealed the following deficiencies:

1. Deficiencies encompass constrained generalization in small datasets with limited samples. Additionally, the model is confined to forecasting two categories—cardiovascular disease and non-cardiovascular disease—highlighting constraints in predicting disease severity and adjusting to diminutive datasets, serving as pivotal limitations [18].
2. Challenges in parameter identification, high computational complexity, and default training limits impede improvements to feature selection with wrapper-based approaches. In the context of cardiovascular disease, these issues impact effective feature exploration, parameter tuning, computational demands, and generalization performance [16].

3. Research methodology

The roadmap outline of the proposed model is presented in Figure 1, illustrating the workflow, which involves two distinct datasets undergoing pre-processing using the Z-score imputation methodology to detect and address missing values. To enhance efficiency and prevent overfitting, the methodology incorporates hybrid feature selection techniques, including Gaussian-based differential entropy for information gain with the Lasso (GDE_Lasso) feature selection model. The comparative analysis involves various classifiers, assessing both the selected and original features. The datasets are divided into training and testing sets, with 80% used for training and 20% for testing, depending on the models' learning rates.

The proposed method effectively identifies cardiovascular diseases with high accuracy [19].

Z-score analysis is a popular method for locating outliers. It reduces noise and highlights notable departures from the typical patterns of data. This technique, which is commonly applied in many fields, computes Z-scores to identify data points that deviate significantly from the mean, usually more than two or three standard deviations. As a result, removing these outliers improves the data's general integrity [20].

$$\text{ZScore} = \frac{(\mathbf{x}-\bar{\mathbf{x}})}{\boldsymbol{\sigma}} \qquad (1)$$

The relationship between x (the standardized random variable), $\bar{x}$ (the mean), and σ (the standard deviation) is expressed in equation (1).

The Z-score assessment was conducted on Dataset 1, consisting of 592 instances and 14 features sourced from the heart disease dataset. These instances were segregated into training and testing subsets, with approximately 80% (roughly 448 cases) allocated to the train set and the remaining 20% (around 112 cases) reserved for the test set. After removing 32 outliers, the total instances available for further analysis amounted to 560.

A similar methodology was applied to the second dataset, which included 1190 cases and 12 attributes from the heart disease dataset. Approximately 80% of the instances (or 929 cases) in the dataset were used as train sets, while the remaining 20% (or 233 cases) were used as test sets. 1162 occurrences in total could be examined further after 28 outlier cases were removed [21]. Table 1is comparative summary of model performance across different datasets.

Initially, the comparison of Z-score analysis for the Cleveland datasets is revealed in Table 2, along with its corresponding graphical representation in Figure 2. Feature selection plays a crucial role in increasing the accuracy of the model by removing associated or unnecessary information [22]. The selection process included several methods, including the filter method, wrapper method, and embedded techniques. Filter techniques are utilized to rapidly rank features, wrapper methodologies aim to enhance model performance, and embedded algorithms handle both feature selection and training of the model simultaneously. The effects of choosing certain features and correlated features on model performance include heightened computational complexity and diminished interpretability [23]. Among various statistical methods, differential entropy for information gain and Lasso highlighted the identification of the crucial characteristics [24]. To determine the most important features in the dataset, this work used a unique hybrid approach that combined the Modified Lasso technique with information gain based on Sequential Model Gaussian-based differential entropy for information gain.

Information Gain, an essential technique employed in decision tree partitioning, is determined by the difference in entropy, a widely used concept in ML. Entropy measures a dataset's uncertainty in a Decision Tree and is often used as a criterion to determine the best division at each node. In the fields of information theory and statistical inference, this measure is useful for feature selection because it denotes a decrease in entropy [25].Substituting traditional entropy with differential entropy is beneficial as it directly handles continuous variables without information loss, aligns with the natural Gaussian distribution of many datasets for more meaningful results, and offers computational efficiency and improved model performance by providing precise measures of uncertainty for continuous features, enhancing feature selection and tree splitting in Random Forests.

A continuous random variable, like the typical normal distribution, can have its normal differential entropy determined using these methods [26].

LASSO (Least Absolute Shrinkage and Selection Operator) is a powerful technique for feature selection and dimensionality reduction, crucial for handling high-dimensional data. It imposes a penalty on the absolute values of regression coefficients, effectively shrinks some coefficients to zero and removes less important features from the model. This characteristic makes LASSO particularly effective for identifying and excluding superfluous features.

The standard LASSO regression equation (5) involves a regularization parameter α, as follows [27]:

$$L(\beta) = \left(\frac{1}{(2*n)}\right) * \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j * X_{ij}))^2 + \alpha * \sum_{j=1}^{p}|\beta_j| \quad \text{------ (5)}$$

The optimal value of α is crucial for LASSO's performance. It balances the trade-off between model difficulty and integrity of fit. A lower α results in a less sparse model that better fits the training data, while a higher α leads to a sparser model with potentially better generalization. To determine the optimal α, and to minimize the prediction error for the validation data, Bayesian Information Criterion (BIC) is replaced with α in equation (5). BIC is a powerful tool for model selection that considers both model fit and complexity [28]. The BIC is represented as:

Differential Entropy(S) = $-\int_{\infty}^{-\infty} f(x) * [log(f(x))]dx$---- (2)

In equation (2) substitute the following steps:

Step 1: Begin by expressing the standard Gaussian distribution as follows:

Where, f(x) = $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Step2: Next, calculate the logarithmic expression and substitute the Differential Entropy formula 2

$$S = -\int_{\infty}^{-\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} [-log(\sigma\sqrt{2\pi}) - \frac{(x-\mu)^2}{2\sigma^2})]dx$$

Step4: Integral distribute the terms,

$$\mathbf{S} = \int_{\infty}^{-\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathbf{log}\,(\sigma\sqrt{2\pi})]dx - \int_{\infty}^{-\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}]dx$$

Step5: Apply Integration for the equation in step 4 and simplify it as

Gaussian Based Differential Entropy(S) = $0.5 * \sigma\sqrt{2e\pi}$ ---------- (3)

Finally, substituting equation (3) in (4) for Information-Gain (IG):

GDE for IG = **GDE** $(S) - \sum_{\infty}^{-\infty}[\left(\frac{|S_V|}{|S|}\right) * \mathbf{GDE}\ (S_V)]$---------- (4)

$$BIC = -2\log(L) + k * \log(n) \quad \text{----- (6)}$$

In equation (6) ,Where L is the probability of the model, k represent the number of parameters, and n is the number of observations [29]. Substituting the alpha value in equation (5) into the BIC in equation (6) yields equation (7).

$$L(\beta) = \left(\frac{1}{(2*n)}\right) * \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j * X_{ij}))^2 + (-2\log(L) + k * \log(n)) * \sum_{j=1}^{p}|\beta_j| \text{ ------ (7)}$$

In equation (7), integrates BIC into the LASSO framework, enhancing the model selection process by balancing fit and complexity, thus preventing overfitting while ensuring effective feature selection. The model Modified Lasso is the combination of LASSO with BIC leverages the strengths of both methods, making it a robust choice in various statistical modelling tasks. An inventive method for creating a classification model for intelligent cardiovascular risk prediction. It investigates the application of machine learning methods in this field. One such method is Random Forest, a supervised classification system that uses random selection and bagging to create a forest of trees. This ensemble model efficiently handles missing data, minimizes the danger of overfitting, shortens training times, and provides estimates for significant classification variables, all of which improve accuracy [30]. The proposed hybrid approach combines Random Forest with GDE_Lasso feature selection to provide a robust and adaptable method for classification tasks. The inclusion of Gaussian-based differential entropy for calculating Information Gain in Random Forest classification aims to optimize feature selection, improve decision tree splitting, and enhance overall model performance. By leveraging the Gaussian distribution's natural fit and mathematical properties, this approach ensures accurate, efficient, and theoretically justified measures of uncertainty and information, ultimately leading to more effective and robust machine learning models. This integration aims to improve accuracy by preventing overfitting, reducing training time, anticipating significant features, and effectively handling missing data. The key purpose of the method is to enhance the efficiency of machine learning techniques by Gaussian-based differential entropy for information gain, specifically in the context of Random Forest classification. Pseudo code 1 outlines the procedures required to implement the recommended technique.

4. Results and Discussions

The results of applying the suggested system are shown in this section. It covers topics including dataset descriptions, experimental results, and comparative analyses. The research focuses on two primary datasets sourced from the UCI repository, providing detailed descriptions for each of these datasets.

Pseudo code 1: Proposed Method of GDE_Lasso Feature Selection

step1.: Compute the differential entropy S using the standard normal distribution function

$$S = -\int [f(x) * \log(f(x))]\, dx \text{ from } -\infty \text{ to } \infty$$

Step 2: Gaussian distribution properties can be used to simplify S:

$$S = \sigma * \text{sqrt}(2e\pi) * 0.5$$

Step 3: Calculate Information Gain (IG) for all subsets SV:

$$\textbf{Information Gain (IG)} = \textbf{Gaussian Based Differential Entropy } (\boldsymbol{S}) - \sum_{\infty}^{-\infty}\left[\left(\frac{|S_V|}{|S|}\right) * \textbf{Gaussian Based Differential Entropy}(\boldsymbol{S_V})\right]$$

Step 4: The feature selection is performed using the formula

$$L(\beta) = \left(\frac{1}{(2*n)}\right) * \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j * X_{ij}))^2 + (-2\log(L) + k * \log(n)) * \sum_{j=1}^{p}|\beta_j|$$

Step 5: Finally, apply the Machine Learning Classifier Random Forest with the feature selected in step 4.

Step 6: Evaluate the Performance Metrics (Accuracy, Precision, Recall, F1- Score) of the Proposed Model

The evaluation metrics extensively employ a widely recognized cardiovascular dataset, which comprises a total of 302 patient records, with each record containing 76 features. However, only 14 distinct variables are used for the evaluation to contrast our findings with those of earlier research. There are also 290 patient records in another dataset, and each record has 14 characteristics. After these datasets are combined, a consolidated dataset with 592 instances total—each with 14 attributes—is produced. The datasets last characteristic functions as the prediction target, showing whether a patient has heart disease (1) or not (0) [31].

A combined total of 1190 cases is obtained by combining five different datasets from different sources. Contributions to these datasets come from Switzerland (123 instances), Cleveland (303 instances), Hungarian (294 instances), VA Long Beach (200 instances), and the Statlog project (270 instances). Age, sex, type of chest pain, blood sugar, ECG, maximum heart rate, angina, old-peak, ST, and target are among the twelve unique features that make up the combined dataset. Among these attributes, the "Output" property serves as the predicted attribute, while the remaining eleven attributes serve as input attributes. In this context, an "Output" value of 0 denotes no heart disease, while a value of 1 signifies the presence of heart disease [32].

This section explores the results obtained by using the suggested strategy on the two datasets. The assessment includes a thorough analysis of Random Forest and other classifiers' performance before and after the suggested approach's deployment.

The confusion matrix that shows the overall results of feature selection from the cardiovascular disease dataset 1 is shown in Figure 3. It provides insight into the model's capacity to distinguish between individuals with heart disease and those without it based on actual results, which show 45 true negatives, 0 false positives, 2 false negatives, and 65 true positives.

A ROC curve, shown in Figure 4, has been used for an extensive evaluation of the model. This curve illustrates the proportion between True Positive Rate (TPR) and False Positive Rate (FPR), presenting a graphical illustration of the model's performance. The area under the ROC curve, which runs from 0 to 1, shows that the model has an outstanding ability to distinguish between classes, with an accuracy of 98.2% for dataset 1.The closer the ROC curve approaches 1, the more robust the model's classification capabilities, highlighting its efficacy in making precise predictions.

The confusion matrix presented in Figure 5 shows the overall results of feature selection from the cardiovascular disease dataset2. Based on actual outcomes, it indicates 102 true negatives, 3 false positives, 5 false negatives, and 128 true positives, providing insight into the model's effectiveness in distinguishing between individuals who have heart disease and those who do not.

A ROC (Receiver Operating Characteristic) curve, illustrated in Figure 6, is utilized for a comprehensive evaluation of the model. This curve illustrates the balance between True Positive Rate (TPR) and False Positive Rate (FPR), providing a visual representation of the model's performance. The area under the ROC curve, which runs from 0 to 1, shows that the model has a great capacity to differentiate between classes, with an accuracy of 96.5% for dataset 2. The model's capacity for classification improves when the ROC curve approaches 1, demonstrating how well it can produce precise predictions. The closer the ROC curve is to 1, the stronger the model's classification abilities, underscoring its effectiveness in making accurate predictions. The internal comparison involved evaluating the proposed system with two distinct datasets, namely Cleveland Disease. The

results corresponding to the execution of various classification models are presented. Initially, the RF model exhibited higher accuracy compared to other classifiers in dataset 1. In dataset 1, the RF model first showed better accuracy than other classifiers. Likewise, Table 3 illustrates that in dataset 2, the RF outcomes produced the maximum accuracy in comparison with the other classifier.

Table 3 and Figure 7 reveal that the RF achieved a notable accuracy rate of 95.8% in Dataset 1, outperforming other models. A comparison of the accuracy results for Logistic Regression, Naïve Bayes, SVM, KNN, Decision Tree, and XG Boost revealed 84.03 %, 80.67 %, 91.6 %, and 93.28 %, respectively. The analysis unequivocally highlights RF's superior accuracy. Despite the success of the current algorithms, additional comparisons were conducted using key metrics such as accuracy, precision, recall, and F1 score.

Table 3 shows an internal comparison of various classifiers before the proposed model, while Figure 8 provides an overview of the two datasets. Notably, Random Forest (RF) demonstrated a significant accuracy rate of 94.96%, surpassing its competitors. XG Boost scored 92.44%, KNN revealed 68.49%, Decision Tree reached 88.24%, Naïve Bayes 85.29%, and SVM 80.25%. Logistic regression received an 80 point67 percent score. The analysis highlights the exceptional performance of RF, which is notable for its remarkable accuracy. Even though the current algorithms performed admirably, further comparisons were made while taking significant metrics into account. The evaluation uses accuracy as a primary metric to gauge the performance of machine learning models. A hybrid approach is applied across two datasets, starting with 12 input features, which are subsequently reduced to 9 using the Modified LASSO method. Table 4 presents a comparative analysis for Dataset 1, illustrating the performance of the proposed model that combines hybrid methods with Random Forest classifiers. A similar assessment is conducted for Dataset 2, as shown in Table 5.

The findings indicate that for Dataset 1, the GDE_Lasso method achieves the highest accuracy at 98.21%, surpassing the Random Forest model at 95.8% and the Modified Differential Entropy-Based Information Gain + RF model at 96.43%. Additionally, the GDE_Lasso method secures the highest F1-score of 98.48%. For Dataset 2, the GDE_Lasso method also leads with an accuracy of 96.64%, followed by the Random Forest model at 94.96% and the Modified Differential Entropy-Based Information Gain + RF model at 95.8%. Moreover, the GDE_Lasso method achieves an F1-score of 96.97%, outperforming the other models. Graphical representations of the method's performance are available in Figures. 9 and 10 for Dataset 2. Figure 11 illustrates the graphical representation of feature ranking for the proposed algorithms, emphasizing their relative importance. Along with ranking the attributes, this visual also highlights the key risk factors associated with cardiovascular disease. It provides insightful information on the importance of certain features within the parameters of the proposed algorithm. To effectively design risk assessment models, it is crucial to comprehend the relative relevance of traits in predicting the outcomes of cardiovascular disease. This information is essential for understanding the relative importance of features in predicting cardiovascular disease outcomes, aiding in the development of effective risk assessment models.

Table 1. *Comparative Summary of Model Performance across Different Datasets*

Authors	Dataset	Methods & Techniques	Accuracy
Salman Pathan, M., et al. [10]	Cardiovascular Health Study (CHS)	PCA, DT, MLP	97.7%
Bsoul, M.A., et al. [11]	UCL Clinical datasets	LR, SVM, KNN, MLP, RF, SMOTE, K-means, recursive feature selection	86.6%
Sumwiza, K., et.al. [12]	Kaggle repository	Ensemble method, RF, data mining, correlation coefficients	99%
Jabbar, M.A., et al. [13]	Statlog heart disease dataset	Random forest, Chi-squared feature assessment, backward elimination	83.7%
Iscra, K., et al. [14]	clinical data	LVEF, HRV analysis, interpretable models	73.5%
Saw, M., et al. [15]	Healthcare dataset	Logistic Regression	87.02%
Mehmood, A., et al. [16]	Heart Disease Dataset	DL-CNN, LASSO	97%
Sharawi, M., et al. [17]	UCI data repository	Whale Optimization Algorithm, wrapper-based feature selection	Superiority over PSO and GA

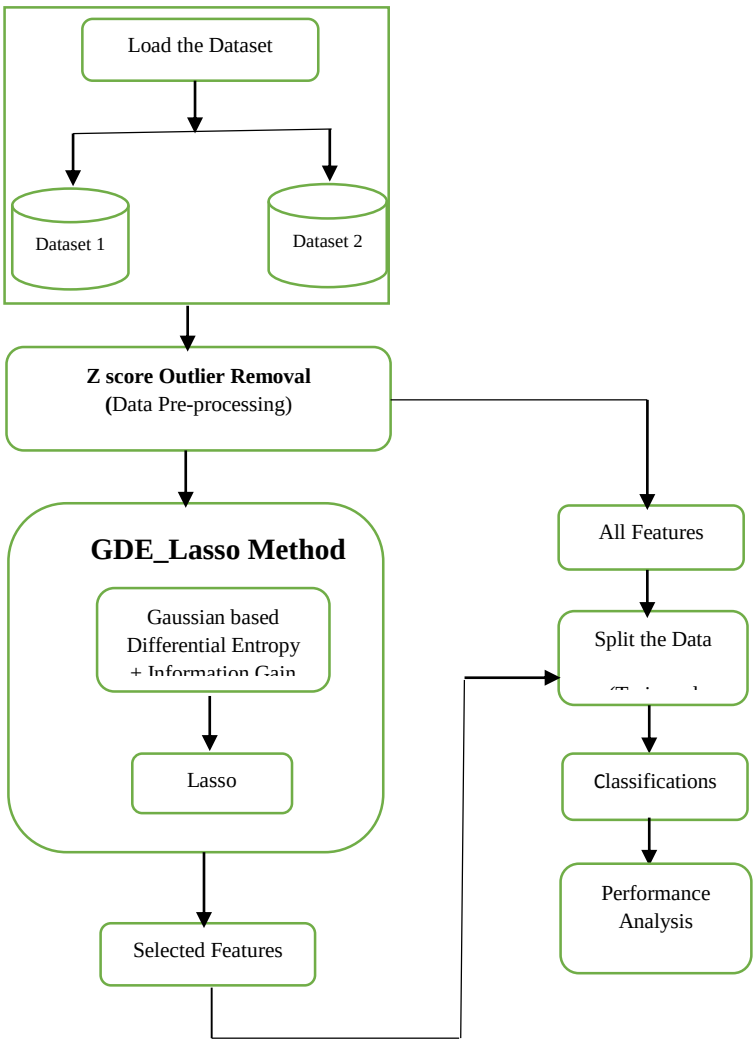

Figure 1. *Workflow of the GDE_Lasso Method*

Table 2. *Comparison of Z-Score Analysis*

Datasets	Z score Analysis (Instances)				
	Original	Outlier Removal	Total (Z Score)	Training (80%)	Testing (20%)
Dataset 1	592	32	560	448	112
Dataset 2	1190	28	1162	929	233

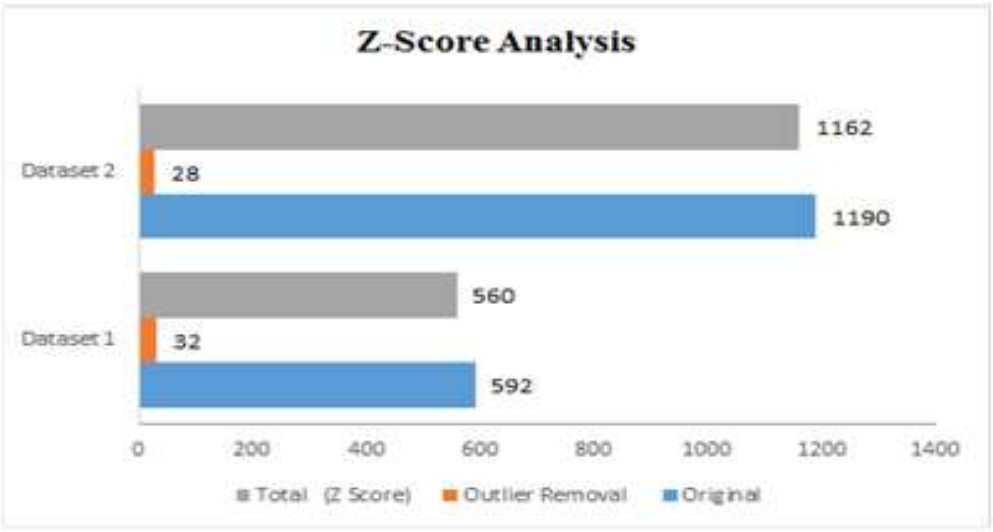

Figure 2. *Z-Score Analysis*

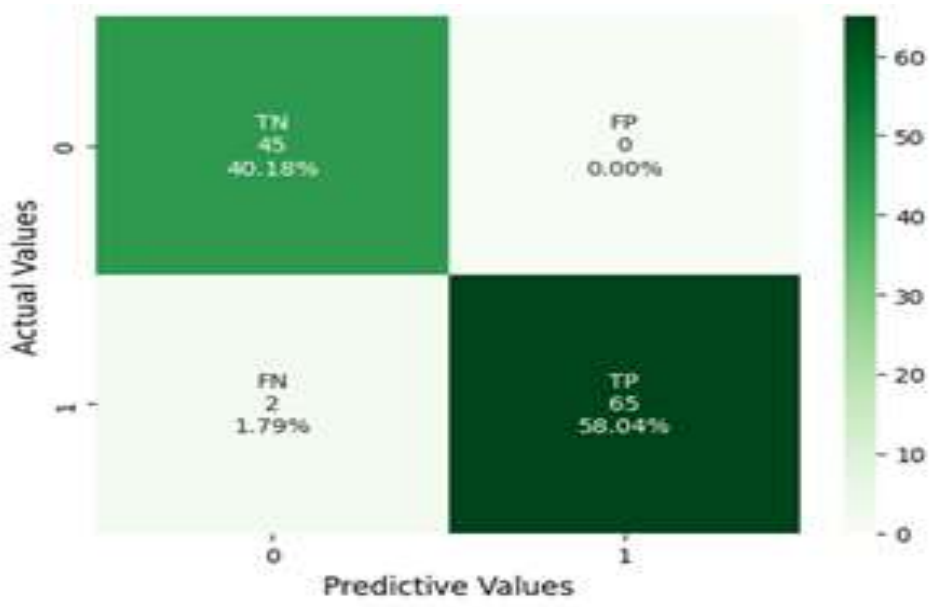

Figure 3. *Confusion Matrix for Dataset 1*

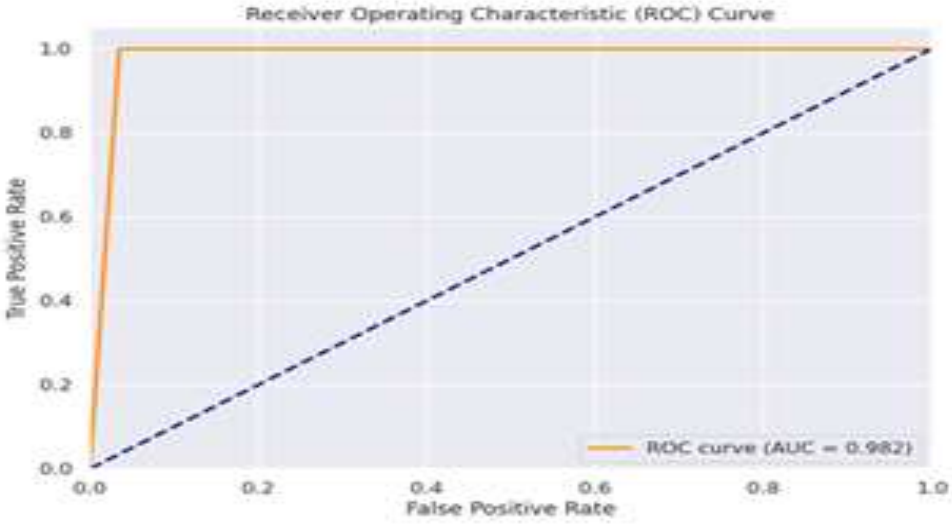

Figure 4. *ROC Curve for Dataset 1*

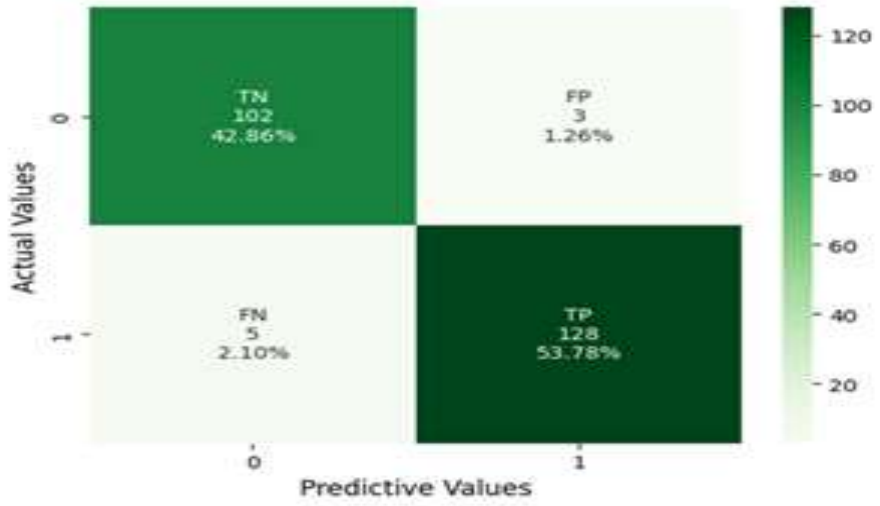

Figure 5. *Confusion Matrix for Dataset 2*

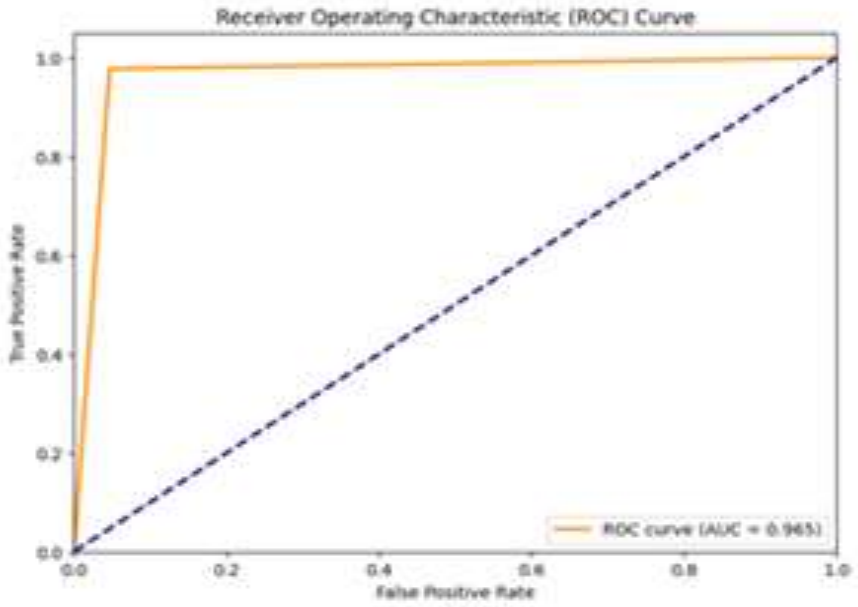

Figure 6. *ROC Curve for Dataset 2*

Table 3. *Performance of various Machine Learning Models for the considered Datasets*

Datasets	Different Models	Performance Metrics (%)			
		Accuracy	Precision	Recall	F1-Score
Dataset 1 (592 instances)	Logistic Regression	84.03	92.06	80.56	85.93
	Naïve Bayes	80.67	88.89	80.67	82.96
	SVM	84.03	92.06	80.56	85.93
	KNN	70.59	80.95	68.92	74.45
	Decision Tree	91.6	96.83	88.41	92.42
	XG Boost	93.28	100	88.73	94.03
	Random Forest	**95.8**	**100**	**92.65**	**96.18**
Dataset 2 (1190 instances)	Logistic Regression	80.67	81.68	82.95	82.31
	Naïve Bayes	85.29	85.29	85.27	86.27
	SVM	80.25	81.54	82.17	81.85
	KNN	68.49	71.43	69.77	70.59
	Decision Tree	88.24	89.76	88.37	89.06
	XG Boost	92.44	93.7	92.25	92.97
	Random Forest	**94.96**	**94.66**	**96.12**	**95.38**

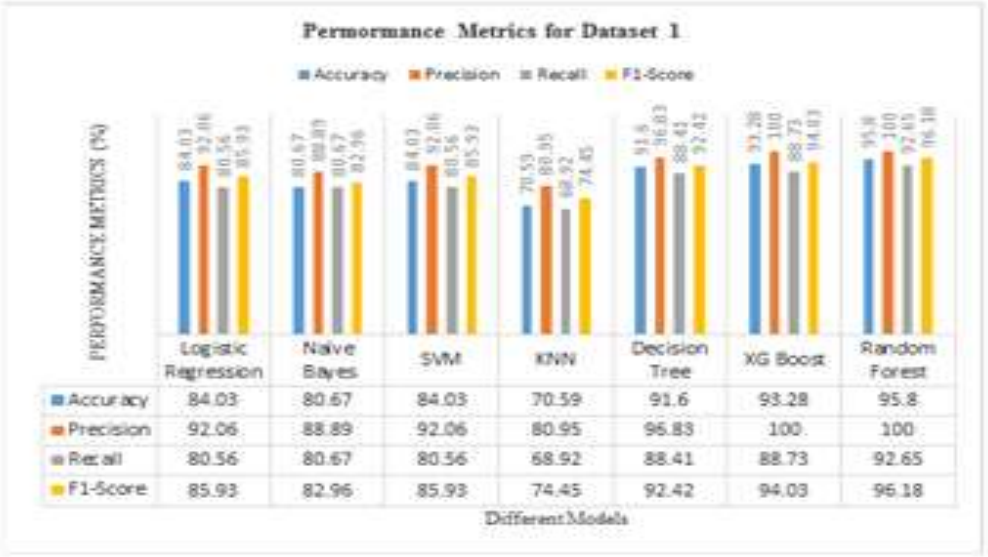

Figure 7. *Performance Metrics for Dataset 1*

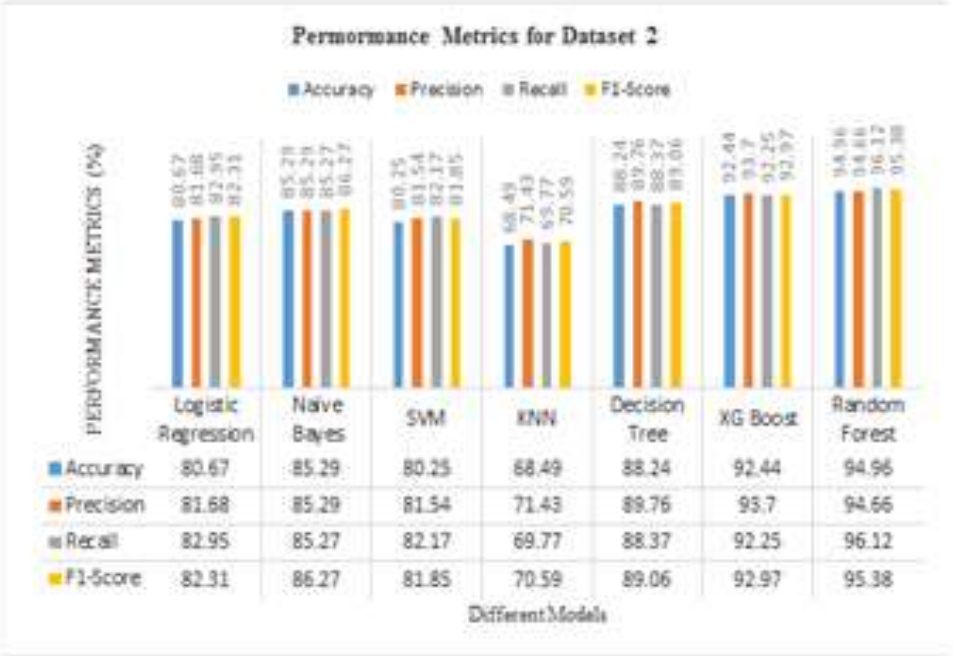

Figure 8. *Performance Metrics for Dataset 2*

Table 4. *Performance Metrics of Comparative Models for Dataset 1*

Models	Accuracy	Precision	Recall	F1-Score
Random Forest	95.8	100	92.65	96.18
Modified Differential Entropy Based Information Gain + RF [30]	96.43	96.92	96.92	96.92
GDE_Lasso Method	**98.21**	**97.01**	**100**	**98.48**

Table 5. *Performance Metrics of Comparative Models for Dataset 2*

Models	Accuracy	Precision	Recall	F1-Score
Random Forest	94.96	94.66	96.12	95.38
Modified Differential Entropy Based Information Gain + RF [30]	95.8	97.67	94.74	96.18
GDE_Lasso Method	**96.64**	**97.71**	**96.24**	**96.97**

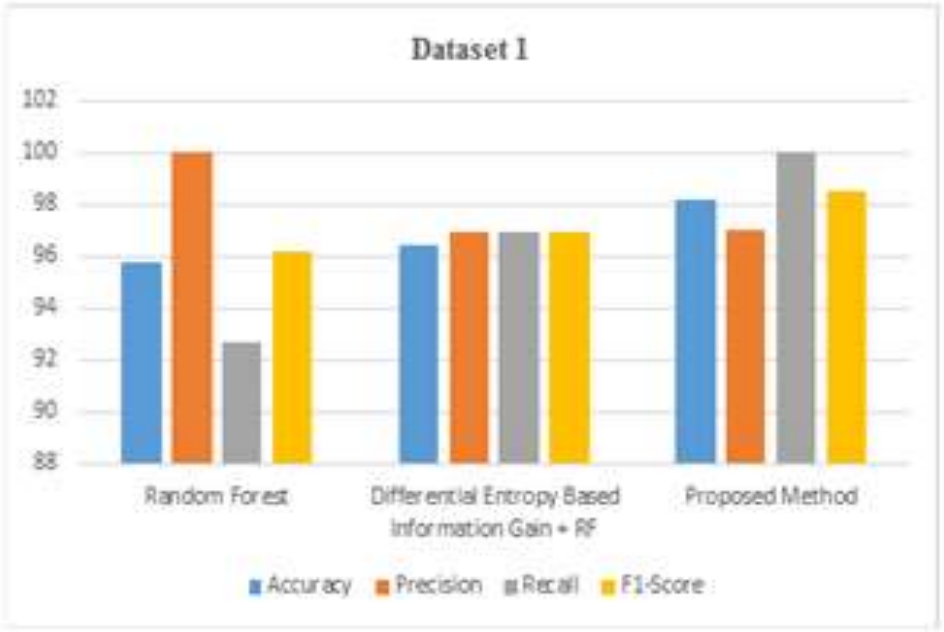

Figure 9. *Analysis of the Proposed Method for Dataset 1*

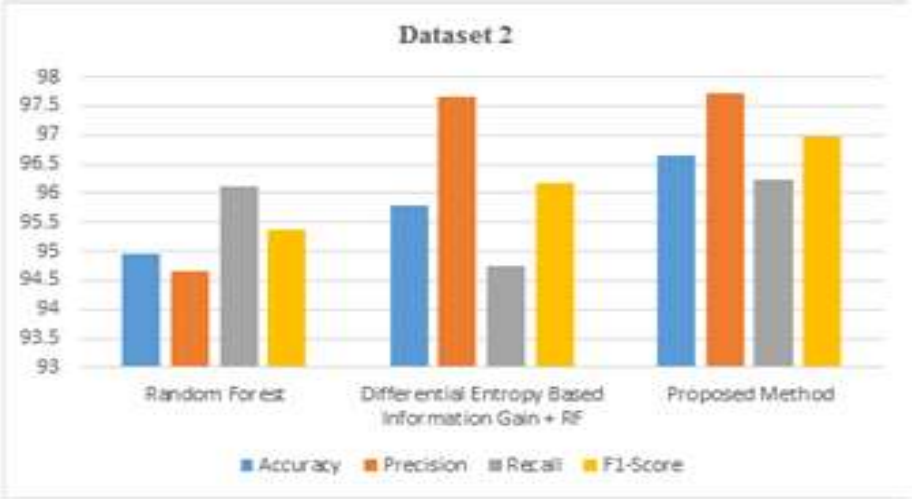

Figure 10. *Analysis of the Proposed Method for Dataset 2*

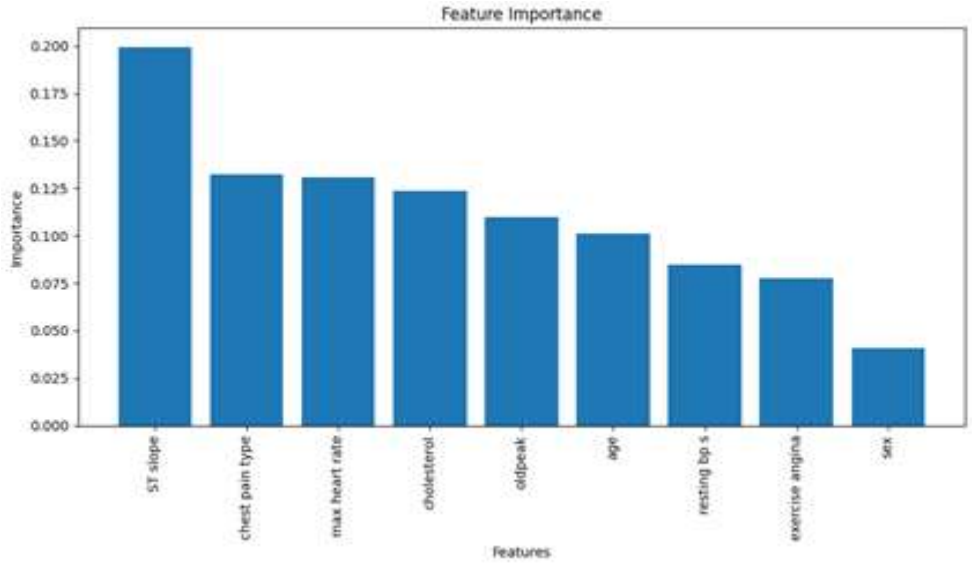

Figure 11. *Feature Importance for Selected Features (K=9)*

4. Conclusions

Cardiovascular research plays a crucial role in addressing the severity of heart disease, underscoring the importance of early detection for effective risk management. Feature selection methods are valuable for pinpointing essential attributes, which helps reduce diagnostic costs. This study highlights the significant potential of machine learning technologies, particularly the GDE_Lasso feature selection model, for early detection and diagnosis of cardiovascular diseases. By combining Gaussian-based differential entropy for information gain with Lasso and Random Forest, the proposed hybrid model shows exceptional performance in accuracy, specificity, sensitivity, and precision compared to traditional methods. The integration of this model with the Random Forest classifier achieved remarkable accuracy rates of 98.21% and 96.64% on two different Cleveland datasets, surpassing conventional methods and demonstrating its effectiveness through confusion matrices. The impressive accuracy rates on the Cleveland datasets confirm the model's effectiveness. This approach not only streamlines diagnostic processes and reduces costs but also facilitates targeted interventions, making it a valuable tool for intelligent cardiovascular risk prediction. Future research could explore further improvements and applications of this model in various clinical environments to enhance cardiovascular health outcomes.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** B.Kalaivani: Conception study and Design, Data analytics, Methodology, draft manuscript preparation. A.Ranichitra: Conceptualization, Methodology, Analysis and interpretation of results, Validation, Review & Editing.
- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]Kedia, V., Regmi, S.R., Jha, K., Bhatia, A., Dugar, S. and Shah, B.K., (2021). Time Efficient IOS Application For CardioVascular Disease Prediction Using Machine Learning. *In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 869-874). IEEE.

[2]Yoshimura, R., Nakagami, T., Hasegawa, Y., Oya, J. and Babazono, T., (2022). Association between changes in body weight and cardiovascular disease risk factors among obese Japanese patients with type 2 diabetes. *Journal of Diabetes Investigation*, 13(9),1560-1566.

[3]Reddy, N.S.C., Nee, S.S., Min, L.Z. and Ying, C.X., (2019). Classification and feature selection approaches by machine learning techniques: Heart disease prediction. International Journal of Innovative Computing, 9(1).

[4]Djerioui, M., Brik, Y., Ladjal, M. and Attallah, B., (2020), September. Heart Disease prediction using MLP and LSTM models. *In 2020 International Conference on Electrical Engineering (ICEE)* (pp. 1-5). IEEE

[5]Wankhede, J., Sambandam, P. and Kumar, M., (2022). Effective prediction of heart disease using hybrid ensemble deep learning and tunicate swarm algorithm. *Journal of Biomolecular Structure and Dynamics*, 40(23),13334-13345.

[6]Vivekanandan, T. and Iyengar, N.C.S.N., (2017). Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Computers in biology and medicine*, 90,125-136.

[7]Elavarasi, D., Kavitha, R. and Aanjankumar, S., (2023), December. Navigating Heart Health with an Elephantine Approach in Clinical Decision Support Systems. *In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)* (pp. 1416-1423). IEEE.

[8]Udhan, S. and Patil, B., 2023. Novel Deep Neural Network for Early Prediction and Prevention of Cardiovascular Disease. DOI: 10.21203/rs.3.rs-3294920/v1

[9]Singh, M.S. and Choudhary, P., (2017), August. Stroke prediction using artificial intelligence. *In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 158-161). IEEE.

[10]Salman Pathan, M., Nag, A., Mohisn Pathan, M. and Dev, S., (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *arXiv e-prints*, pp.arXiv-2206.

[11]Bsoul, M.A., Qusef, A. and Abu-Soud, S., (2022). Building an optimal dataset for arabic fake news detection. *Procedia Computer Science*, 201,665-672.

[12]Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P. and Bamurigire, P., (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. *Informatics in Medicine Unlocked*, 41;101316

[13]Jabbar, M.A., Deekshatulu, B.L. and Chandra, P., (2016). Prediction of heart disease using random forest and feature subset selection. In Innovations in Bio-Inspired Computing and Applications: Proceedings of the *6th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2015)* held in Kochi, India during December 16-18, 2015 (pp. 187-196). Springer International Publishing.

[14]Iscra, K., Miladinović, A., Ajčević, M., Starita, S., Restivo, L., Merlo, M. and Accardo, A., (2022). Interpretable machine learning models to support differential diagnosis between Ischemic Heart Disease and Dilated Cardiomyopathy. *Procedia Computer Science*, 207;1378-1387.

[15]Saw, M., Saxena, T., Kaithwas, S., Yadav, R. and Lal, N., (2020), January. Estimation of prediction for getting heart disease using logistic regression model of machine learning. *In 2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-6). IEEE.

[16]Mehmood, A., Iqbal, M., Mehmood, Z., Irtaza, A., Nawaz, M., Nazir, T. and Masood, M., (2021). Prediction of heart disease using deep convolutional neural networks. *Arabian Journal for Science and Engineering*, 46(4),3409-3422.

[17]Sharawi, M., Zawbaa, H.M. and Emary, E., (2017), February. Feature selection approach based on whale optimization algorithm. *In 2017 Ninth international conference on advanced computational intelligence (ICACI)* (pp. 163-168). IEEE

[18]Saqlain, S.M., Sher, M., Shah, F.A., Khan, I., Ashraf, M.U., Awais, M. and Ghani, A., (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. *Knowledge and Information Systems*, 58;139-167.

[19]Aggarwal, V., Gupta, V., Singh, P., Sharma, K. and Sharma, N., (2019), April. *Detection of spatial outlier by using improved Z-score test. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 788-790). IEEE

[20]Mohamed, S.M., Malhat, M.G. and Elhady, G.F., (2022). prediction of cardiovascular disease using machine learning techniques. IJCI. International *Journal of Computers and Information*, 9(2);25-44.

[21]Kalaivani, B. and Ranichitra, A., (2022). A comparative study of machine learning approaches for proactive cardiovascular disease prediction. *Int J Health Sci*, 6(S8),5390-5400.

[22]Liang, Q., Zhao, S., Zhang, J., Deng, H., Damm, W., Hess, D., Schweda, M., Sztipanovits, J., Bengler, K., Biebl, B. and Fränzle, M., (2024). Cyber-physical systems. *ACM Transactions on*, 8(1).

[23]Gupta, A. and Singh, A., (2023). EDL-NSGA-II: Ensemble deep learning framework with NSGA-II feature selection for heart disease prediction. *Expert Systems*, 40(7);e13254.

[24]Cai, T.T., Liang, T. and Zhou, H.H., (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *Journal of Multivariate Analysis*, 137;161-172.

[25]ThanhNoi, P. and Kappas, M., (2017). Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. *Sensors*, 18(1);18

[26]Kalaivani, B. and Ranichitra, A., (2024). Optimizing Cardiovascular Disease Prediction: Harnessing Random Forest Algorithm with Advanced Feature Selection. DOI: 10.21203/rs.3.rs-3834700/v1

[27]Emmert-Streib, F. and Dehmer, M., (2019). High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*, 1(1), pp.359-383.

[28]Lorah, J. and Womack, A., (2019). Value of sample size for computation of the Bayesian information criterion (BIC) in multilevel modeling. *Behavior research methods*, 51, pp.440-450.

[29]Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F.J.M., Ignatious, E., Shultana, S., Beeravolu, A.R. and De Boer, F., (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*, 9,19304-19326.

[30]Djerioui, M., Brik, Y., Ladjal, M. and Attallah, B., (2020), September. Heart Disease prediction using MLP and LSTM models. *In 2020 International Conference on Electrical Engineering (ICEE)* (pp. 1-5). IEEE.

[31]Bhuyan, M.K., (2019). Computer vision and image processing: Fundamentals and applications. *CRC Press.*

[32]Kalaivani, B. and Ranichitra, A., (2023). Unveiling the Impact of Outliers: An Improved Feature Engineering Technique for Heart Disease Prediction. *In International Conference on IoT Based Control Networks and Intelligent Systems (pp. 469-478).* Singapore: Springer Nature Singapore.