AIMS Mathematics, 11(1): 2682–2701.
DOI: 10.3934/math.2026108
Received: 28 September 2025
Revised: 07 January 2026
Accepted: 14 January 2026
Published: 27 January 2026

http://www.aimspress.com/journal/Math

Research article

Improved chi-square feature selection for robust heart disease data classification

Heba Nayl[1], Elkhateeb S. Aly[2,3,*], Amira Rezk[4] and M. E. Fares[1]

[1] Department of Mathematics, Faculty of Science, Mansoura University, Mansoura, Egypt

[2] Department of Mathematics, College of Science, Jazan University, P.O. Box 114 Jazan 45142, Kingdom Saudi Arabia

[3] Nanotechnology research unit, College of Science, Jazan University, P.O. Box 114 Jazan 45142, Kingdom Saudi Arabia

[4] Department of Information Systems, Faculty of Computer and Information Sciences, Mansoura University, Mansoura, Egypt

* **Correspondence:** Email: elkhateeb@jazanu.edu.sa, Tel: +966541663540.

Abstract: Early diagnosis of heart disease is vital for reducing mortality and improving patient outcomes; yet, accurate prediction remains a significant challenge owing to the complexity and high dimensionality of medical data. Data preprocessing is essential for overcoming these issues by cleaning, transforming, reducing, and balancing data to provide reliable inputs for feature selection and classification. This study introduces an improved chi-square (χ^2) feature selection framework combined with multiple classifiers to enhance predictive performance. Our method was applied to Cleveland heart disease and diabetes datasets, where numeric attributes were discretized into categorical values, enabling χ^2 to select the most informative features while eliminating redundancy. Several classifiers, including support vector machine (SVM), logistic regression (LR), K-nearest neighbors (KNN), and naive Bayes (NB), were trained using both the reduced subset and the complete feature set. Results show that the preprocessing include χ^2 feature selection, achieved the highest performance. On the Cleveland dataset, the model attained a mean accuracy of 93.72%, precision of 94.01%, recall of 93.72%, F1-score of 93.74%, and an area under the curve(AUC) of 97.87%, while on the diabetes dataset, it achieved mean values of 93.55% accuracy, 94.23% precision, 93.55% recall, 93.48% F1-score, and an AUC 93.53%. The main contribution of this work lies in integrating discretization with χ^2 based selection to produce a compact and discriminative feature subset. With a minimal number of selected features, the proposed approach delivers robust, accurate, and computationally efficient heart disease prediction, outperforming existing methods.

Keywords: heart disease prediction; feature selection; chi-square statistical test; naive Bayes;

classification performance
Mathematics Subject Classification: 97R40, 97P50, 62F07

1. Introduction

Accurate prognosis of cardiac disease is both essential and challenging. Heart disease, often linked to coronary artery dysfunction, weakens the body and disrupts vascular function, particularly in adults and the elderly. Cardiovascular diseases cause over 18 million deaths worldwide every year, according to the World Health Organization (WHO) [1], cardiovascular diseases claim more than 18 million lives annually [2], and in the United States alone, approximately one billion dollars are spent daily on heart disease treatment [3]. Early prediction is therefore critical for improving patient outcomes. With the rapid growth in the volume, variety, velocity, and veracity of healthcare data, effective computational models are needed to identify disease risk factors, support personalized and cost-effective treatment, and enhance overall quality of care [4]. Early warning signs of myocardial ischemia, a major cause of heart attacks and strokes, frequently include chest pain, shortness of breath, an irregular heartbeat, fatigue, and fainting. Timely detection of these symptoms can significantly improve patient survival. Diagnosis typically relies on clinical tests such as angiography, electrocardiograms (ECG), and blood analyses, along with expert medical evaluation. However limited access to specialists and the potential for human error pose challenges to accurate diagnosis. Consequently, there is growing research interest in developing intelligent, machine-assisted systems to automate this process and reduce diagnostic errors. Artificial intelligence (AI), particularly machine learning [5] and deep learning [6], has shown promise in supporting the diagnosis and prognosis of cardiac diseases, as well as in the analysis and interpretation of medical images [7]. Support vector machine (SVM), K-nearest neighbors (KNN), logistic regression (LR), and decision tree (DT) are examples of machine learning strategies that are commonly used [8]. But the effective implementation of these methods might be significantly influenced by the availability of high-quality data. Therefore, appropriate data handling during the preprocessing phase, which also includes the (FS) stage, is essential to guarantee success. Choosing the best feature subset can result in increased classification accuracy, improved generalization by avoiding overfitting, and less computational complexity. Algorithms for FS may be divided into two main categories: filter methods and wrapper methods. Filter methods are independent of the classifier and use statistical or probabilistic techniques to choose important features [9, 10]. Filter-based approaches are further divided into two types: uni-variate methods, which evaluate each feature individually with respect to the target variable, and multivariate methods, which account for dependencies among multiple features [11]. Filter-based FS methods have also gained significant attention due to their simplicity and scalability. These approaches evaluate the relevance of features using statistical or probabilistic measures independently of classifiers, making them efficient and less prone to overfitting. They are broadly categorized into uni-variate and multivariate methods [11]. Uni-variate filters assess each feature individually against the target class using tests such as Mann-Whitney U test, t-test, information gain, Pearson correlation, Fisher's exact test, and χ^2 test. Although computationally efficient and suitable for high-dimensional data, uni-variate methods overlook feature interactions. Multivariate filters, such as Relief-based algorithms, minimal redundancy-maximal relevance (mRMR)

and fast correlation-based filter (FCBF), evaluate groups of features jointly, capturing dependencies and reducing redundancy. However, they are computationally more demanding and less scalable compared to uni-variate filters. Contrarily, wrapper techniques make use of the performance of classification algorithms to determine the optimal feature subsets [12]. Prior to applying FS, raw data must be preprocessed to address issues such as incompleteness, inconsistency, noise, or redundancy, as these challenges can otherwise result in unreliable outcomes. Data preprocessing encompasses selection, cleaning, and transformation steps that refine datasets, resolve quality issues, and convert raw data into formats suitable for mining algorithms, thereby enhancing interpretability and overall model performance [13]. After preprocessing and feature selection, data mining techniques are applied to extract meaningful patterns and insights [14]. Prepared datasets can be analyzed through clustering approaches [15] or classification methods [16, 17], implemented using supervised or unsupervised machine learning algorithms. These techniques enable the identification of hidden structures, the development of predictive models, and the generation of actionable knowledge. The extracted results are commonly communicated through data visualization methods [18, 19] and knowledge representation techniques [20, 21], which facilitate understanding and decision making. In this study, the dataset is first preprocessed to ensure reliability, consistency, and suitability for analysis. Feature selection is then applied as a critical step to enhance classification performance. A uni-variate filter based approach is adopted, employing χ^2 test to select the most relevant features. The χ^2 method is selected due to its efficiency, scalability, ability to mitigate overfitting, and suitability for medical datasets that contain categorical features. To assess the effectiveness of the resulting feature subsets, multiple machine learning algorithms are applied, including SVM, LR, KNN, and NB.

2. Related work

To develop an intelligent healthcare framework for predicting heart disease, Ali et al. [22] proposed combining feature fusion with ensemble deep learning approaches. Their system integrates electronic health records with sensor data to provide insightful health information. To decrease computational burden and increase system efficiency, feature selection was performed using Information Gain, while the conditional probability approach gave each class its own unique weights to improve the accuracy of predictions. The proposed system was benchmarked against traditional classifiers using feature fusion, FS, and weighting methods on heart disease data. Using Orange, Weka, Rapid Miner, Knime, MATLAB, and Scikit-learn [23] compared six common data mining approaches for identifying heart illness. The measures of accuracy, sensitivity, and specificity were used to assess performance. The best results were obtained with MATLAB, especially with its Artificial Neural Network (ANN) model. Based on user experience and receiver operating characteristic (ROC) curve analysis, their research ended with advice on how to choose tools, emphasizing MATLAB's advantages. The PIMA Indian diabetes dataset was analyzed using machine learning approaches by Kalagotla et al. [24]. They utilized a correlation-based FS technique, followed by AdaBoost, and they went on to present a novel stacking strategy that combines MLP, SVM, and LR models. Their stacking framework consistently outperformed AdaBoost and produced good results in a variety of diagnostic datasets, such as the Wisconsin breast cancer dataset and the Cleveland heart disease dataset. Additionally, the study highlighted the significance of data protection in healthcare research, pointing out the dangers of centralized data storage. They emphasized federated learning (FL) as a potential approach for

distributed model training that preserves the privacy of healthcare information. Latha and Jeeva [25] investigated ensemble-based classification methods to improve heart disease prediction using the Cleveland dataset. They evaluated bagging, boosting, stacking, and majority voting approaches, reporting accuracy improvements ranging from 5–7%, with majority voting providing the best gain of 7.26%. Moreover, when ensemble methods were combined with FS, the predictive accuracy increased further, reaching a maximum of 85.48% with majority voting applied to the optimized feature subset. [26] applied principal component analysis (PCA) for feature reduction before classification, transforming correlated features into principal components for improved model efficiency. Sarra et al. [27] proposed an SVM-based classification model combined with χ^2 feature selection, which reduced dataset dimensionality while maintaining key attributes. Their approach improved accuracy from 85.29% to 89.7% and cut computational load by half, outperforming several traditional methods. By combining FL with feature selection and extraction, Kapila et al. [28] extended this study. Utilizing both the Cleveland heart disease and diabetes datasets, they employed linear discriminant analysis (LDA) for feature extraction, as well as χ^2 and ANOVA for FS. Unlike other methods, this one uses LDA to extract features. Their FL framework, which was built around a centralized strategy, assured data security by keeping sensitive information local and just sharing model updates. The integration of FS and FE in the FL environment resulted in better dimensionality reduction, increased class separability, and higher prediction accuracy across benchmark datasets. Based on the preceding discussion, it is evident that many studies have employed uni-variate filter based methods particularly χ^2 test for feature selection in medical datasets. χ^2 method is well regarded for its simplicity, scalability, and ability to mitigate overfitting [11, 25]. Its main limitation, however, is that it evaluates features independently, thereby overlooking potential interactions between attributes [9]. Despite this drawback, χ^2 remains highly effective for categorical medical data due to its robustness and low computational cost. Building on these strengths, the present study proposes a tailored preprocessing pipeline that integrates discretization with χ^2 test to generate a compact and discriminative feature subset. This subset is then evaluated using multiple classifiers, yielding a computationally effective, accurate, and reliable approach to predicting heart disease.

3. Research approach

3.1. Data collection

One of the most popular datasets for forecasting heart disease, notably in machine learning research, is the Cleveland heart disease dataset. There are 14 features in the dataset, including demographic, medical, and diagnostic variables, that were obtained from 303 individuals. The UCI Machine Learning Repository makes the dataset available to the public. This dataset includes a range of features that are classified as binary, nominal, or numerical, all of which provide essential information about the patient's features that are important for determining cardiovascular risk. Table 1 summarizes key descriptive statistics, including standard deviation (SD), range, mean, and information on missing values or potential outliers. This dataset is frequently used because it contains a diverse set of variables that capture the complex nature of heart disease. The target variable is binary, where 0 denotes the absence of heart disease and 1 denotes the presence with severity levels ranging from 1–4. This study also employs the diabetes UCI dataset*, which contains 520 patient records described by 17 attributes.

*https://www.kaggle.com/datasets/alakaaay/diabetes-uci-dataset

Among these, *age* (ranging from 20 to 65 years) is the only numerical feature, while the remaining are categorical, including sex, alopecia, muscle stiffness, partial paresis, delayed healing, irritability, itching, visual blurring, genital thrush, polyphagia, weakness, sudden weight loss, polydipsia, polyuria, and obesity. The target variable (*class*) is binary, indicating either a positive or negative diabetes diagnosis. Given that nearly all attributes are categorical, this dataset is particularly well-suited for our method χ^2, which measures the dependency between the class target and categorical features.

Table 1. Descriptive statistics of features in the Cleveland heart disease dataset.

Feature	Description	Range	Mean	Sd	Missing	Outlier
	Binary Features					
Exang	Indicates whether angina occurred during exercise (yes=1).	[0, 1]	0.3267	0.4698	No	Yes
Sex	Patient's sex (female=0, male=1).	[0, 1]	0.6799	0.4673	No	Yes
FBS	Fasting blood sugar greater than 120 mg/dl (true=1).	[0, 1]	0.1485	0.3562	No	Yes
	Nominal Features					
CP	Type of chest pain (values 1-4).	[1, 4]	3.1584	0.9601	No	No
RestECG	Resting electrocardiogram outcomes (0, 1, 2).	[0, 2]	0.9901	0.9950	No	No
Slope	Slope of the peak exercise ST segment (values 1–3).	[1, 3]	1.6007	0.6162	No	No
Thal	Thalassemia status (normal=3, fixed defect =6 , reversible defect =7).	[3, 7]	4.7342	1.9710	Yes	Yes
CA	Number of major blood vessels visualized via fluoroscopy (range 0–3).	[0, 3]	0.6722	0.9344	Yes	Yes
	Numeric Features					
Thalach	Maximum heart rate achieved.	[71, 202]	149.6073	22.8750	No	Yes
Oldpeak	ST depression relative to rest, induced by exercise.	[0, 6.2]	1.0396	1.1611	No	Yes
Age	Age of the patient (in years).	[29, 77]	54.4389	9.0387	No	Yes
Chol	Serum cholesterol concentration (mg/dl).	[126, 564]	246.6931	51.7769	No	Yes
Trestbps	Resting blood pressure (mm Hg).	[94, 200]	131.6898	17.5997	No	Yes

Abbreviations: Sd = Standard Deviation

3.2. Preprocessing steps

The process of preparing data is essential in data mining applications, as raw data often contains errors, inconsistencies, and omissions that can degrade the performance of learning and mining algorithms [29]. To address these issues, preprocessing techniques are applied to improve data quality before model building. Recent studies broadly classify preprocessing methods into three main categories: data cleaning (e.g., noise filtering and missing value imputation), data reduction (e.g., feature and instance selection), and data transformation (e.g., normalization and aggregation).

3.2.1. Data cleaning

- **Missing Values:** Missing values are a major problem that must be addressed during preprocessing, before applying machine learning algorithms. It can result from human error, equipment failure, withheld information, or inconsistent data. Anomalies may also cause missing values when removed. These values are typically handled through imputation or repair techniques to maintain data integrity. Features with over 45% missing values were excluded from the analysis. For the remaining features, binary features were imputed with a constant value, categorical variables with the mode, and numerical variables with either the mean [30] or median.
- **Outliers:** Outliers differ from noise: while noise is generally meaningless and should be removed, outliers may contain both irrelevant and valuable (exceptional) information. Outliers are data points that significantly deviate from expected patterns. They may result from measurement errors, misreporting, sampling issues, or reflect rare but true values. In medical datasets, outliers are often retained rather than removed, as they may represent rare but clinically significant cases

rather than errors [31]. Eliminating these data points could lead to a loss of important information, particularly when modeling conditions with high variability or rare presentations. Approaches such as Z-score and inter-quartile range (IQR) are used to identify it [32].

3.2.2. Discretization

Discretization transforms continuous numerical data into nominal or categorical attributes by dividing the data range into non-overlapping intervals, each mapped to a discrete value. This enables the data to be treated as nominal, facilitates statistical tests such as χ^2 test, and enhances interpretability. Equal-frequency and equal-width binning are two discretization techniques. Common discretization techniques include equal-frequency and equal-width binning [29].

3.2.3. Bin selection rules

- Sturges' rule [33]: We used the equal-width binning method, where the number of bins is determined using Sturges' rule [33], defined as

$$n = 1 + 3.322 \log_{10}(k), \tag{1}$$

  where k and n are the number of observations and bins, respectively.
- Freedman-Diaconis rule [34]: Determines bin width based on IQR, making it more robust to skewed distributions and outliers.
- Scott's rule [35]: Computes bin width using the standard deviation of the data and is suitable for data that is approximately normally distributed but more sensitive to outliers than the Freedman-Diaconis rule.

Table 2 reports the number of bins obtained for each numeric feature using the three rules.

Table 2. Number of bins obtained using different discretization rules.

Numeric feature	Sturges	Freedman-Diaconis	Scott
oldpeak	10	9	8
age	10	13	11
chol	10	13	11
trestbps	10	13	10
thalach	10	13	10

Importantly, altering the discretization scheme directly influences χ^2 feature selection outcome. Splitting numeric attributes into ten bins resulted in a compact set of highly informative features and consistently yielded superior classification performance compared with alternative binning configurations. In this study, numeric attributes were discretized into ten bins following Sturges' rule. Although this rule is optimal for approximately normally distributed data and several numeric attributes in the datasets exhibit deviations from normality, the adopted discretization strategy yielded a more stable and discriminative feature subset, resulting in consistently higher classification performance compared with alternative binning configurations.

3.2.4. Data normalization

Numerical features are rescaled to a common scale using a preprocessing technique known as normalization, without distorting the differences in their value ranges [36]. It is essential in algorithms

sensitive to scale, as it ensures that each feature contributes equally to the analysis. In this study, normalization was applied after imputation to maintain consistency across numeric features [37]. We applied min-max normalization, which scales each feature to a specified range $[n_{min}, n_{max}]$ using the following formula:

$$v' = \frac{v - v_{\min}}{v_{\max} - v_{\min}} \times (n_{\max} - n_{\min}) + n_{\min}, \tag{2}$$

where

- v denotes the original feature value,
- v' denotes the normalized value,
- $v_{\min}$ and $v_{\max}$ represent the minimum and maximum values of the original feature,
- $n_{\min}$ and $n_{\max}$ represent the lower and upper bounds of the target range.

We introduce χ^2 test for feature selection (see Section 3.3). Since the χ^2 test requires non-negative input values, the data were normalized to the range $[0, 1]$ instead of $[-1, 1]$, ensuring compatibility with the test.

3.2.5. Balanced dataset

The synthetic minority over-sampling technique (SMOTE) is a popular way to handle class imbalance in classification problems. SMOTE creates synthetic samples by interpolating between existing minority instances and their closest neighbors, rather than simply replicating instances of the minority class. This helps create a more balanced training dataset, allowing machine learning algorithms to learn decision boundaries better and avoid bias toward the majority class. When the minority class is underrepresented, SMOTE works especially well as it improves model generalization without increasing the risk of overfitting associated with traditional oversampling methods [38]. Importantly, SMOTE is applied only on the training data, and the evaluation remains unbiased.

Figure 1 illustrates the proposed preprocessing and classification pipeline. Numeric features are first discretized and then all data normalized, followed by χ^2 based feature selection to retain the most relevant attributes. The dataset is then split into training and testing sets, with SMOTE applied to the training set only to address class imbalance. Classifiers are subsequently trained on the selected features, and performance is evaluated using standard metrics.

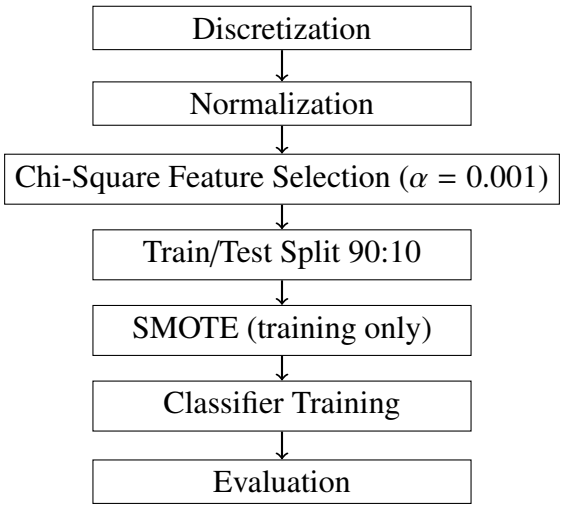


Figure 1. Schematic diagram of the improved χ^2 based feature selection algorithm.

3.3. Chi-square test for feature selection

Overfitting often results from having too many features, which can reduce a model's ability to generalize. Therefore, selecting the most predictive features for both testing and training datasets is essential for improving model performance [39]. This involves retaining features that are meaningful to the machine learning classifier while discarding those that are noisy or irrelevant [40]. In our study, before applying the machine learning classifier, relevant features were selected using χ^2 statistical test [41]. χ^2 test is a correlation-based feature selection method evaluates the relationship between the target class and each feature. It calculates the χ^2 statistic to determine whether a feature is dependent on the predicted attribute. A higher χ^2 score indicates a stronger dependency, suggesting that the feature is more relevant to the prediction task. It should be noted that the χ^2 test applies only to categorical variables. Therefore, before applying the χ^2 test [42], continuous features must be discretized. In this process, each feature is divided into discrete intervals based on specific binning boundaries. These boundaries are determined using Sturges' rule [33], which aims to minimize discretization error while preserving the informational content of the original data.

In a binary classification context for heart disease, let the dataset consist of t instances categorized into two classes: negative (absence of disease) and positive (presence of disease), where p and $t - p$ are the number of positive and negative class instances, m represents the instances where feature F is present, and $t - m$ where it is absent see Table 3. χ^2 test assesses the deviation between observed and expected frequencies under the assumption of independence between a feature and the class label. Let the observed frequencies be denoted by $f_{11}, f_{10}, f_{01}, f_{00}$, and the corresponding expected frequencies by $E_{11}, E_{10}, E_{01}, E_{00}$. The expected values, assuming independence, are calculated as follows:

$$E_{11} = \frac{(f_{11} + f_{10})(f_{11} + f_{01})}{t}. \tag{3}$$

Similarly, E_{10}, E_{01}, and E_{00} can be calculated. The general χ^2 test statistic is given by

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{4}$$

In the case of a 2×2 contingency table, χ^2 can be computed as

$$\chi^2 = \frac{(f_{11} - E_{11})^2}{E_{11}} + \frac{(f_{10} - E_{10})^2}{E_{10}} + \frac{(f_{01} - E_{01})^2}{E_{01}} + \frac{(f_{00} - E_{00})^2}{E_{00}}. \tag{5}$$

Table 3. Table for calculating χ^2 score for feature X_i.

	Positive class	**Negative class**	**Total**
Feature X_i occurs	f_{11}	f_{10}	$f_{11} + f_{10} = m$
Feature X_i does not occur	f_{01}	f_{00}	$f_{01} + f_{00} = t - m$
Total	$f_{11} + f_{01} = p$	$f_{10} + f_{00} = t - p$	t

3.4. Feature ranking and hypothesis testing using chi-square test

Feature selection begins by computing the χ^2 score for each feature using Eq (5). This score quantifies the strength of association between the target class and a feature. A higher χ^2 value shows

a stronger statistical relationship with the class label (positive or negative). Features are prioritized based on their χ^2 statistic, arranged in descending order. To ensure that the selected features are not statistically significant merely by chance, each is further evaluated through hypothesis testing using the chi-squared distribution.

The hypotheses for the χ^2 test are defined as follows: the null hypothesis (H_0) assumes that a feature is independent of the class label (no association, not predictive), whereas the alternative hypothesis (H_1) assumes dependence (an association exists). A significance level of $\alpha = 0.001$ is used. For each feature, the p-value is computed from its corresponding χ^2 statistic. If $p \leq \alpha$, the null hypothesis is rejected in favor of the alternative, indicating that the feature is statistically significant. This stricter threshold ($\alpha = 0.001$ vs the more common $\alpha = 0.01$) selects a smaller subset of features with stronger statistical significance, improving feature relevance, reducing redundancy, and enhancing interpretability, which contributed to improved classification performance see [43].

According to the chi-squared scores and the predefined significance level, the top five features selected for classification in the heart disease dataset are: thal, exang, ca, oldpeak, and slope. Similarly, for the diabetes dataset, the top eight features identified as most influential for classification are: polydipsia, polyuria, gender, polyphagia, sudden weight loss, partial paresis, visual blurring,and weakness.

4. Validation strategy and evaluation metrics

4.1. Validation strategy

To ensure robust performance assessment, we employed 30 repeated stratified train-test splits, preserving class distributions and reducing sampling bias. Model performance was evaluated using accuracy, precision, recall, F1-score, and AUC reported as mean ± 99.9% confidence intervals to provide a conservative, variance-aware estimate of stability. AUC curves were aggregated across splits, with mean curves and shaded variability bands illustrating the robustness of discriminative performance.

4.2. Evaluation metrics

To thoroughly assess the proposed model's performance, we employ a range of standard evaluation metrics, including F1-score, accuracy, recall, precision, and the receiver operating characteristic (ROC) curve with its corresponding AUC [44]. These metrics capture complementary aspects of classification effectiveness and are defined as follows:

- **Accuracy:** The fraction of correctly classified instances out of the total, reflecting the model's overall correctness.
- **Precision:** The proportion of true positive predictions among all positive predictions, indicating how well the model avoids false positives.
- **Recall (sensitivity):** The proportion of true positive predictions among all actual positives, measuring the model's ability to identify relevant cases.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced evaluation metric, especially useful for imbalanced datasets.
- **ROC curve and AUC:** ROC curve illustrates the trade-off between true positive rate and false

positive rate across different thresholds. AUC condenses this performance into a single value, with higher values indicating better discriminative capability.

Together, these metrics provide both an overall assessment of model accuracy and a detailed evaluation of its ability to handle imbalanced classes and detect positive cases reliably.

5. Classifiers

In the machine learning technique known as supervised learning, labeled input-output pairings are used to train models so they can learn the relationship between features and target variables. Once trained, these models can generalize to unseen data and provide accurate predictions. In this study, four supervised classification algorithms are employed, each widely used in medical data analysis and optimized with appropriate parameters to enhance predictive performance.

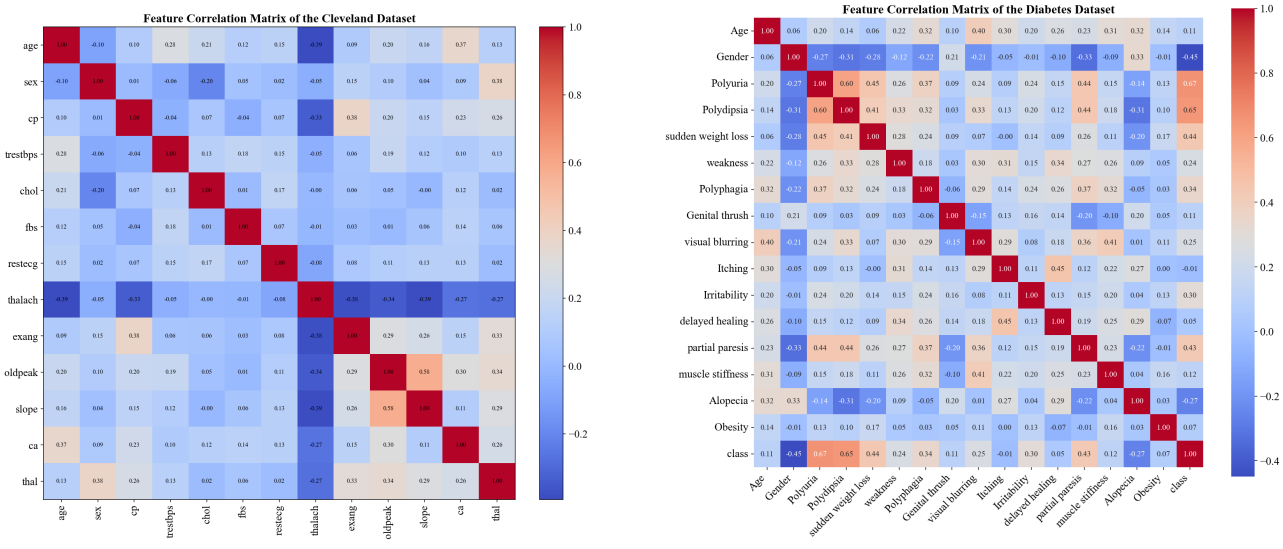
5.1. Naive Bayes (NB)

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, which assumes that features are conditionally independent given the class, as illustrated in Figure 2. Despite its simplicity, it performs well in many real-world scenarios [45]. Among its merits, the computation process is easier compared to many other classifiers, and it is particularly well suited to continuous-valued attributes [46].

The model estimates the posterior probability $P(C_t \mid F)$ as

$$P(C_t \mid F) = \frac{P(F \mid C_t) \cdot P(C_t)}{P(F)}, \tag{5.1}$$

where C_t is the class target(0 or 1), $P(F)$ the marginal likelihood, $P(C_t)$ the prior probability, $P(F \mid C_t)$ is the likelihood, and, the feature vector $F = (f_1, f_2, \ldots, f_n)$.



(a) **(b)**

Figure 2. Feature correlation matrices of the datasets: (a) Cleveland dataset and (b) diabetes dataset.

5.2. K-nearest neighbor (KNN)

It assigns an unknown data point based on the majority class of its nearest neighbors [47]. It is widely applied in medical datasets, pattern recognition, cluster analysis, and image processing. In healthcare, KNN has been used with LDA to build a warning system for hypertension and cardiovascular disease [48], to identify predictive features of chronic disease [49], and to support cardiac patient analysis. In our study, the model is configured with $K = 10$, and Euclidean distance is applied to evaluate:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}, \tag{5.2}$$

where

- $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is the feature vector of the query (test) instance,
- $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ is the feature vector of a training instance,
- N is the total number of features (dimensions),
- $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$.

5.3. Support vector machine (SVM)

It is a statistical, kernel-based method that constructs an optimal hyperplane to separate data points of different classes, where the closest instances to the hyperplane, called support vectors, determine the decision boundary [50]. By mapping data into higher-dimensional spaces, SVM can handle nonlinear classification using kernels such as polynomial, radial basis function or linear, typically employing Euclidean distance with parameters controlling the flexibility of the RBF function [51, 52]. In our study, the model was configured with probability estimation enabled ($probability = True$), allowing probabilistic outputs in addition to class predictions. SVM offers several merits, including high accuracy compared to many classifiers, effective handling of complex nonlinear data, and reduced susceptibility to overfitting.

5.4. Logistic regression (LR)

It is a linear model for binary classification that employs the logistic (sigmoid) function to estimate the probability that an input instance belongs to the positive class [53]. In the presence of arbitrary outliers in the covariate matrix, robust approaches such as Robust Logistic Regression(RoLR) have been proposed, which estimate parameters via a simple linear programming procedure. RoLR is the first logistic regression approach that offers performance guarantees in the presence of corrupted covariates, demonstrating robustness against a consistent proportion of adversarial outliers. Beyond regression, it can also be applied to binary classification when a fraction of training samples are corrupted [54]. The model is configured with $max\_iter = 1000$ to ensure convergence during training, and predicts the positive class with probability:

$$p = P(y = 1 \mid f) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 f_1 + \cdots + \beta_n f_n)}}, \tag{5.3}$$

where $f_1, f_2, \ldots, f_n$ are input features and $\beta_0, \ldots, \beta_n$ are model coefficients.

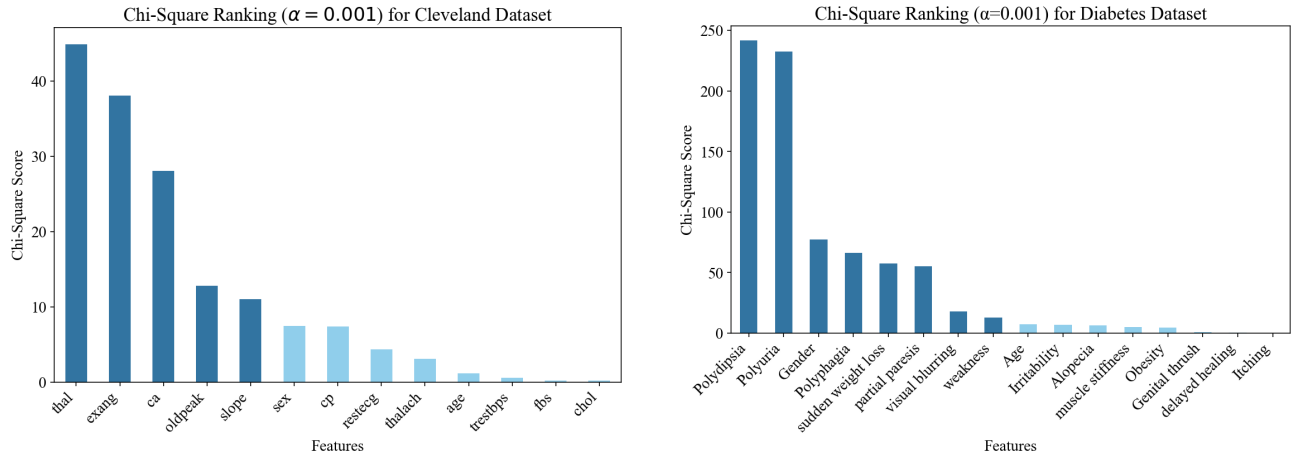
6. Results and discussion

6.1. Chi-square feature ranking

Figure 3 presents χ^2 feature ranking for both datasets at a significance level of α=0.001.

For the Cleveland heart disease dataset (Figure 3(a)), the features *thal*, *exang*, *ca*, *oldpeak*, and *slope* achieved the highest χ^2 scores, indicating a strong statistical association with the target variable. These attributes were therefore the most relevant for predicting heart disease, whereas features such as *trestbps*, *fbs*, and *chol* exhibited limited significance.

For the diabetes dataset (Figure 3(b)), the most discriminative features were *polydipsia*, *polyuria*, *gender*, *polyphagia*, *sudden weight loss*, *partial paresis*, *visual blurring*, and *weakness*. These variables attained substantially higher χ^2 scores than the rest, underscoring their critical role in differentiating diabetic from non-diabetic cases. A significance level of $\alpha = 0.001$ is used see Table 4. The smaller p-values do not automatically reduce false positives, using $\alpha = 0.001$ prioritizes features with stronger associations, complementing effect size and domain knowledge in the selection process.



(a) **(b)**

Figure 3. Chi-square feature ranking for the (a) Cleveland dataset and (b) diabetes dataset.

Table 4. Chi-square test results for feature selection: (a) Cleveland dataset and (b) diabetes dataset.

(a)

Feature	Chi-square	p-value	Sig. ($p \leq 0.001$)
thal	44.878	2.10×10^{-11}	**Yes**
exang	38.053	6.89×10^{-10}	**Yes**
ca	28.038	1.19×10^{-7}	**Yes**
oldpeak	13.857	1.97×10^{-4}	**Yes**
slope	10.984	9.19×10^{-4}	**Yes**
sex	7.433	6.40×10^{-3}	No
cp	7.385	6.58×10^{-3}	No
restecg	4.322	3.76×10^{-2}	No
thalach	3.149	7.60×10^{-2}	No
age	1.260	2.62×10^{-1}	No
trestbps	0.592	4.42×10^{-1}	No
fbs	0.165	6.85×10^{-1}	No
chol	0.105	7.46×10^{-1}	No

(b)

Feature	Chi-square	p-value	Sig. ($p \leq 0.001$)
Polydipsia	241.5710	1.79×10^{-54}	**Yes**
Polyuria	232.3692	1.81×10^{-52}	**Yes**
Gender	77.4953	1.33×10^{-18}	**Yes**
Polyphagia	66.3968	3.69×10^{-16}	**Yes**
Sudden weight loss	57.7493	2.98×10^{-14}	**Yes**
Partial paresis	55.3143	1.03×10^{-13}	**Yes**
Visual blurring	18.1246	2.07×10^{-5}	**Yes**
Weakness	12.7243	3.61×10^{-4}	**Yes**
Age	7.4429	6.37×10^{-3}	No
Irritability	6.8918	8.66×10^{-3}	No
Alopecia	6.2491	1.24×10^{-2}	No
Muscle stiffness	4.8750	2.72×10^{-2}	No
Obesity	4.5006	3.39×10^{-2}	No
Genital thrush	0.8963	3.44×10^{-1}	No
Delayed healing	0.6202	4.31×10^{-1}	No
Itching	0.0157	9.00×10^{-1}	No

6.2. Performance metrics results

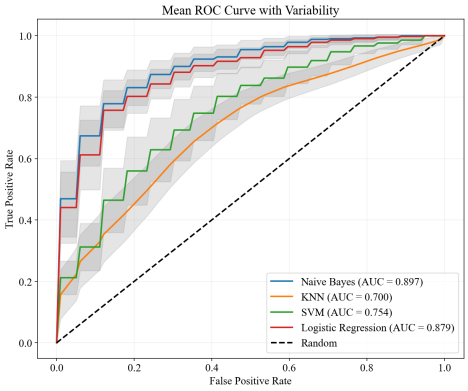
The classification models were evaluated using 30 repeated stratified train-test splits to ensure reliable performance estimates. All metrics accuracy, precision, recall, F1-score, and AUC are reported

as mean ± 99.9% confidence intervals, providing a conservative, variance-aware assessment that emphasizes stability rather than overfitting. Table 5 summarizes the results. Across models, high mean accuracy, precision, recall, F1-score, and AUC values were observed, with narrow confidence intervals indicating consistent performance and minimal variance.

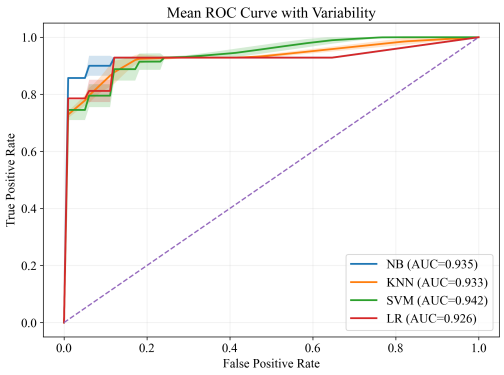
Table 5. Performance metrics of the evaluated models (mean ± 99.9% CI).

Data	Accuracy	Precision	Recall	F1-score	AUC
(a) Cleveland	93.5% ± 0.010	93.7% ± 0.012	93.4% ± 0.011	93.5% ± 0.011	96.8% ± 0.009
(b) Diabetes	93.7% ± 0.019	94% ± 0.018	93.7% ± 0.019	93.7% ± 0.019	97.9% ± 0.010

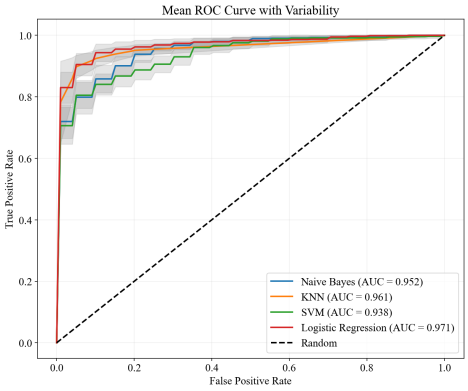
The results indicate strong and stable classification performance across both datasets. High values of accuracy, precision, recall, F1-score, and AUC, together with narrow 99.9% confidence intervals, demonstrate the robustness and consistency of the evaluated models. In particular, the high AUC values suggest excellent discriminative capability, while the tight confidence bounds confirm low variability across repeated stratified evaluations. Aggregated ROC curves Figure 4 further illustrate model robustness, showing mean curves with shaded bands representing variability across splits. The combination of repeated stratification, conservative confidence intervals, and visual ROC assessment demonstrates that the models achieve stable, generalizable performance rather than overfitting to individual splits.



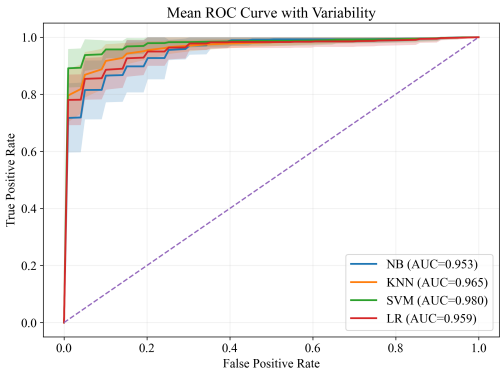
(a) All Features



(b) Selected Features



(c) All Features



(d) Selected Features

Figure 4. Comparison of ROC curves before and after applying chi-square feature selection: (a) and (b) Cleveland dataset, (c) and (d) diabetes dataset.

(a) Cleveland dataset

- **All features:** Using all features, model performance varied across classifiers. NB achieved the highest AUC of 0.897, indicating strong discriminative ability despite its simplicity. LR also performed well with an AUC of 0.879. SVM and KNN showed moderate performance with AUCs of 0.754 and 0.700, respectively. These results suggest that while some models can capture the underlying patterns with all features, others may be affected by noise or redundant information, limiting their predictive accuracy.
- **Selected features:** After feature selection, all models showed substantial improvements in AUC, highlighting the benefit of removing irrelevant feature. SVM achieved the highest AUC of 0.942, followed closely by KNN (0.933), NB (0.935), and LG (0.926). The consistent improvement across classifiers demonstrates that the selected subset of features preserves the most informative attributes for heart disease prediction, resulting in more robust and discriminative models.

(b) Diabetes dataset

- **All features:** Using all features, all classifiers demonstrated strong predictive performance. LR achieved the highest AUC of 0.971, closely followed by KNN (0.961) and NB (0.952), while SVM obtained an AUC of 0.938. These high values indicate that most features contribute meaningful information for diabetes prediction, allowing models to achieve excellent discriminative ability even without feature selection.
- **Selected features:** After feature selection, model performance improved slightly for some classifiers and markedly for SVM. SVM achieved the highest AUC of 0.980, KNN improved to 0.965, NB reached 0.953, and LG obtained 0.959. The results suggest that selecting the most informative features enhances the discriminative power of the models, particularly benefiting SVM, and confirms that careful feature selection can further refine predictive accuracy even when overall model performance is already high.

The results presented in Table 6 demonstrate that the proposed method significantly outperforms existing models applied to both the Cleveland heart disease and diabetes datasets. [27] employed the traditional χ^2 test to select six features, achieving an accuracy of 89.47%. Among recent studies, the highest reported accuracy on the Cleveland dataset was 91.8%, achieved by [55] using Pearson correlation combined with the grey wolf optimizer. Results with the 10 best selected features are shown in (a) for the Cleveland heart disease dataset and (b) for the diabetes dataset. Reported accuracies are 92.3% for diabetes and 88.52% for Cleveland [28], using ANOVA combined with LDA or χ^2. Our proposed method achieved a superior mean accuracy of 93.72% on the Cleveland dataset using and 93.55% on diabetes datasets. The improvement stems from optimized discretization of numeric features using 10 bins, which yields a more stable and discriminative feature subset, resulting in consistently higher classification performance compared with alternative binning configurations, along with a stricter significance threshold ($\alpha = 0.001$) that selects fewer but more statistically relevant features and yields superior performance compared with $\alpha = 0.01$.

Tables 7 and 8 report the paired t-test results comparing the proposed method's accuracy to previous literature on the Cleveland and diabetes datasets. For both datasets, the proposed method consistently outperformed prior results, achieving 0.9354 on Cleveland and 0.9372 on diabetes. The observed improvements are statistically significant at the 5% significance level ($\alpha = 0.05$), with t-statistics

of 3.298 and 3.991 and p-values of 0.0458 and 0.0160, respectively. These results confirm that the proposed approach provides a significant performance gain over existing methods.

Table 6. Comparison between proposed model with existing models.

S.No	Author(s)	Year	Dataset	Best Model	Accuracy	Precision	Recall	F1-Score
1	[56]	2021	Diabetes dataset	Ensemble model	79.22	78.3	78.6	78.3
2	[57]	2021	Diabetes dataset	Voting Classifier	79.04	73.48	71.45	80.6
3	[58]	2021	Diabetes dataset	Stacking Classifier	79.04	73.48	71.45	80.6
4	[27]	2022	Cleveland heart	χ^2 feature selection	89.47	89.40	89.40	89.40
5	[24]	2023	Disease dataset	CART Classification	87.25	88.24	84.51	-
6	[59]	2022	Cleveland heart	MLP + PSO Hybrid Algorithm	84.61	80.08	88.3	84.4
7	[55]	2024	Cleveland heart	Pearson Correlation + Grey Wolf Optimizer	91.8	94.1	88.9	92.8
8	[28]	2025	Cleveland heart	Anv + LDA or Chi + LDA	88.52	87.87	90.62	89.23
			Diabetes	Anv + LDA or Chi + LDA	92.3	94.36	94.36	94.36
9	Proposed method	2025	Cleveland heart	Improved Chi + NB Classifier	**93.72**	**94.01**	**93.72**	**93.74**
			Diabetes dataset	Improved Chi + SVM Classifier	**93.55**	**94.23**	**93.55**	**93.48**

Table 7. Paired t-test results for Cleveland datasets.

Dataset	[27]	[59]	[55]	[28]	our result	t-statistic	p-value	Significant
CL	0.8947	0.8461	0.9180	0.8852	0.9354	3.298	0.0458	Yes

Table 8. Paired t-test results for diabetes datasets.

Dataset	[56]	[57]	[28]	[24]	our result	t-statistic	p-value	Significant
Diabetes	0.7922	0.7904	0.9230	0.8725	0.9372	3.991	0.0160	Yes

7. Conclusions

Our study underscores the critical role of data preprocessing and feature selection in enhancing the prediction of heart disease and diabetes. By integrating χ^2 feature selection with optimized discretization producing a stable and discriminative feature subset and a stricter significance threshold selects fewer and more relevant features, leading to consistently higher classification performance. We achieved significant improvements in key performance metrics. Training with the selected features consistently demonstrated superior performance relative to those utilizing the full feature set, with Naive Bayes and KNN showing particularly strong gains in accuracy and AUC. For both the Cleveland and diabetes datasets, χ^2 method effectively removed irrelevant features, leading to more efficient and accurate classification. ROC curves and AUC values further confirm the superior generalization ability of models built on the selected features. Importantly, this approach facilitates the identification of an optimal feature subset, contributing to faster and more reliable diagnosis of heart disease. Overall, these findings highlight the value of targeted feature reduction in heart dataset analysis and demonstrate that the proposed preprocessing pipeline provides a robust, interpretable framework for building high-performance diagnostic systems.

Author contributions

Heba Nayl: Conceptualization, Data analysis, Investigation, Data curation, Methodology, Software, Formal analysis, Visualization, Project administration, Writing–original draft, Writing–review & editing; Elkhateeb S. Aly: Project administration, Validation, Funding acquisition, Writing–review & editing; Amira Rezk: Supervision, Validation, Methodology, Formal analysis, Project administration, Writing–review & editing; M. E. Fares: Supervision, Validation, Project administration, Writing–review & editing.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors gratefully acknowledge the funding of the Deanship of Graduate Studies and Scientific Research, Jazan University, Saudi Arabia, through Project No. JU-202503229-DGSSR-ORA-2025.

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. *World Health Organization (WHO), Cardiovascular diseases*, World Health Organization office, 2025. Available from: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.
2. H. Ahmed, E. M. Younis, A. Hendawi, A. A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, *Future Gener. Comp. Sy.*, **111**, (2020), 714–722. https://doi.org/10.1016/j.future.2019.09.056
3. Y. Hao, M. Usama, J. Yang, M. S. Hossain, A. Ghoneim, Recurrent convolutional neural network based multimodal disease risk prediction, *Future Gener. Comp. Sy.*, **92** (2019), 76–83. https://doi.org/10.1016/j.future.2018.09.031
4. T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, W. Shi, Federated learning of predictive models from federated electronic health records, *Int. J. Med. Inform.*, **112** (2018), 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007
5. S. Wadhawan, R. Maini, A systematic review on prediction techniques for cardiac disease, *Int. J. Inf. Technol. Sy.*, **15** (2022), 1–33. https://doi.org/10.4018/IJITSA.290001
6. D. Han, X. Yang, G. Li, S. Wang, Z. Wang, J. Zhao, Highway traffic speed prediction in rainy environment based on APSO-GRU, *J. Adv. Transport.*, **2021** (2021), 4060740. https://doi.org/10.1155/2021/4060740
7. K. A. Lara Hernandez, T. Rienmüller, D. Baumgartner, C. Baumgartner, Deep learning in spatiotemporal cardiac imaging: a review of methodologies and clinical usability, *Comput. Biol. Med.*, **130** (2021), 104200. https://doi.org/10.1016/j.compbiomed.2020.104200

8. M. Juhola, H. Joutsijoki, K. Penttinen, D. Shah, R. P. Pölönen, K. Aalto-Setälä, Data analytics for cardiac diseases, *Comput. Biol. Med.*, **142** (2022), 105218. https://doi.org/10.1016/j.compbiomed.2022.105218
9. A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang, Benchmark for filter methods for feature selection in high-dimensional classification data, *Comput. Stat. Data Anal.*, **143** (2020), 106839. https://doi.org/10.1016/j.csda.2019.106839
10. M. Labani, P. Moradi, F. Ahmadizar, M. Jalili, A novel multivariate filter method for feature selection in text classification problems, *Eng. Appl. Artif. Intel.*, **70** (2018), 25–37. https://doi.org/10.1016/j.engappai.2017.12.014
11. N. Pudjihartono, T. Fadason, A. Kempa-Liehr, J. O'Sullivan, A review of feature selection methods for machine learning-based disease risk prediction, *Front. Bioinform.*, **2** (2022), 927312. https://doi.org/10.3389/fbinf.2022.927312
12. A. Mustaqeem, S. M. Anwar, M. Majid, Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants, *Comput. Math. Method. Med.*, **2018** (2018), 7310496. https://doi.org/10.1155/2018/7310496
13. V. Çetin, O. Yıldız, A comprehensive review on data preprocessing techniques in data analysis, *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, **28** (2022), 299–312.
14. P. Yıldırım, D. Birant, Application of data mining techniques in cloud computing: a literature review, *Pamukkale Univ. J. Eng. Sci.*, **24** (2018), 336–343. https://doi.org/10.5505/pajes.2017.65642
15. I. A. Venkatkumar, S. J. K. Shardaben, Comparative study of data mining clustering algorithms, *Proceedings of International Conference on Data Science and Engineering (ICDSE)*, 2016, 1–7. https://doi.org/10.1109/ICDSE.2016.7823946
16. B. Çığşar, D. Ünal, Comparison of data mining classification algorithms determining the default risk, *Scientific Programming*, **2019** (2019), 8706505. https://doi.org/10.1155/2019/8706505
17. S. Umadevi, K. S. J. Marseline, A survey on data mining classification algorithms, *Proceedings of International Conference on Signal Processing and Communication (ICSPC)*, 2017, 264–268. https://doi.org/10.1109/CSPC.2017.8305851
18. S. Ajibade, A. Adediran, An overview of big data visualization techniques in data mining, *International Journal of Computer Science and Information Technology Research*, **4** (2016), 105–113.
19. A. Kunjir, H. Sawant, N. Shaikh, Data mining and visualization for prediction of multiple diseases in healthcare, *Proceedings of International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, 2017, 329–334. https://doi.org/10.1109/ICBDACI.2017.8070858
20. X. Zhou, C. Yang, N. Meng, Method of knowledge representation on spatial classification, *Proceedings of Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, 237–240. https://doi.org/10.1109/FSKD.2009.775
21. Y. Guowei, L. Xinghua, T. Xuyan, A new knowledge representation based matter element system and the related extension reasoning, *Proceedings of International Conference on Natural Language Processing and Knowledge Engineering*, 2003, 89–94. https://doi.org/10.1109/NLPKE.2003.1275874

22. F. Ali, S. El-Sappagh, S. R. Islam, D. Kwak, A. Ali, M. Imran, et al., A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion, *Inform. Fusion*, **63** (2020), 208–222. https://doi.org/10.1016/j.inffus.2020.06.008
23. I. Tougui, A. Jilbab, J. El Mhamdi, Heart disease classification using data mining tools and machine learning techniques, *Health Technol.*, **10** (2020), 1137–1144. https://doi.org/10.1007/s12553-020-00438-1
24. S. Kalagotla, S. Gangashetty, K. Giridhar, *A novel stacking technique for prediction of diabetes*, *Comput. Biol. Med.*, **135** (2021), 104554. https://doi.org/10.1016/j.compbiomed.2021.104554
25. C. Latha, S. Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, *Informatics in Medicine Unlocked*, **16** (2019), 100203. https://doi.org/10.1016/j.imu.2019.100203
26. T. Parhizkar, E. Rafieipour, A. Parhizkar, Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction, *J. Clean. Prod.*, **279** (2021), 123866. https://doi.org/10.1016/j.jclepro.2020.123866
27. R. R. Sarra, A. M. Dinar, M. A. Mohammed, K. H. Abdulkareem, Enhanced heart disease prediction based on machine learning and χ^2 statistical optimal feature selection model, *Designs*, **6** (2022), 87. https://doi.org/10.3390/designs6050087
28. R. Kapila, S. Saleti, Federated learning-based disease prediction: a fusion approach with feature selection and extraction, *Biomed. Signal Proces.*, **100** (2025), 106961. https://doi.org/10.1016/j.bspc.2024.106961
29. S. García, J. Luengo, F. Herrera, *Data preprocessing in data mining*, Cham: Springer, 2015. https://doi.org/10.1007/978-3-319-10247-4
30. J. Luengo, S. García, F. Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods, *Knowl. Inf. Syst.*, **32** (2012), 77–108. https://doi.org/10.1007/s10115-011-0424-2
31. A. Smiti, A critical overview of outlier detection methods, *Comput. Sci. Rev.*, **38** (2020), 100306. https://doi.org/10.1016/j.cosrev.2020.100306
32. A. Boukerche, L. Zheng, O. Alfandi, Outlier detection: methods, models, and classification, *ACM Comput. Surv.*, **53** (2020), 55. https://doi.org/10.1145/3381028
33. H. A. Sturges, The choice of a class interval, *J. Amer. Stat. Assoc.*, **21** (1926), 65–66. https://doi.org/10.1080/01621459.1926.10502161
34. Z. Chuan, W. Yusoff, M. Aziz, A. Senawi, T. Ken, A comparison study between Doane's and Freedman-Diaconis' binning rule in characterizing potential water resources availability, *J. Phys.: Conf. Ser.*, **1366** (2019), 012103. https://doi.org/10.1088/1742-6596/1366/1/012103
35. D. Scott, Scott's rule, *Wiley Comput. Stat.*, **2** (2010), 497–502. https://doi.org/10.1002/wics.103
36. P. Muhammad Ali, R. Faraj, Data normalization and standardization: a technical report, *Machine Learning Technical Report*, **1** (2014), 1–6.
37. A. Alizadeh Naeini, M. Babadi, S. Homayouni, Assessment of normalization techniques on the accuracy of hyperspectral data clustering, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, **XLII-4/W4** (2017), 27–30. https://doi.org/10.5194/isprs-archives-XLII-4-W4-27-2017
38. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.*, **16** (2002), 321–357. https://doi.org/10.1613/jair.953

39. W. Yue, Z. Wang, H. Chen, A. Payne, X. Liu, Machine learning with applications in breast cancer diagnosis and prognosis, *Designs*, **2** (2018), 13. https://doi.org/10.3390/designs2020013
40. F. Aliyar Vellameeran, T. Brindha, A new variant of deep belief network assisted with optimal feature selection for heart disease diagnosis using IoT wearable medical devices, *Comput. Method. Biomec.*, **25** (2022), 387–411. https://doi.org/10.1080/10255842.2021.1955360
41. L. Ali, C. Zhu, M. Zhou, Y. Liu, Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection, *Expert Syst. Appl.*, **137** (2019), 22–28. https://doi.org/10.1016/j.eswa.2019.06.052
42. H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, *Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence*, 1995, 388–391. https://doi.org/10.1109/TAI.1995.479783
43. D. Colquhoun, The reproducibility of research and the misinterpretation of p-values, *R. Soc. Open Sci.*, **4** (2017), 171085. https://doi.org/10.1098/rsos.171085
44. Ž. Vujović, Classification model evaluation metrics, *Int. J. Adv. Comput. Sci.*, **12** (2021), 599–606. https://doi.org/10.14569/IJACSA.2021.0120670
45. N. Boyko, K. Boksho, Application of the naive Bayesian classifier in work on sentimental analysis of medical data, *Proceedings of 3rd International Conference on Informatics and Data-Driven Medicine*, 2020, 1–12.
46. B. Tarle, R. Tajanpure, S. Jena, Medical data classification using different optimization techniques: a survey, *IJRET*, **5** (2016), 101–108.
47. M. Brameier, W. Banzhaf, A comparison of linear genetic programming and neural networks in medical data mining, *IEEE Trans. Evolut. Comput.*, **5** (2001), 17–26. https://doi.org/10.1109/4235.910462
48. P. Tang, M. Tseng, Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification, *Proceedings of International Conference on Machine Learning and Cybernetics*, 2009, 3070–3075. https://doi.org/10.1109/ICMLC.2009.5212633
49. V. Sneha Latha, P. Y. L. Swetha, M. Bhavya, G. Geetha, D. K. Suhasini, Combined methodology of the classification rules for medical data-sets, *International Journal of Engineering Trends and Technology*, **3** (2012), 32–36.
50. M. Islam, M. Chowdhury, S. Khan, Medical image classification using an efficient data mining technique, *Proceedings of 7th Asia-Pacific Complex Systems Conference*, 2004, 34–42.
51. Y. Xing, J. Wang, Z. Zhao, Combination data mining methods with new medical data to predicting outcome of coronary heart disease, *Proceedings of International Conference on Convergence Information Technology*, 2007, 868–872. https://doi.org/10.1109/ICCIT.2007.204
52. M. Samb, F. Camara, S. Ndiaye, Y. Slimani, M. Esseghir, A novel RFE-SVM-based feature selection approach for classification, *International Journal of Advanced Science and Technology*, **43** (2012), 27–36.
53. T. Rymarczyk, E. Kozłowski, G. Kłosowski, K. Niderla, Logistic regression for machine learning in process tomography, *Sensors*, **19** (2019), 3400. https://doi.org/10.3390/s19153400
54. P. Komarek, Logistic regression for data mining and high-dimensional classification, Ph.D Thesis, Carnegie Mellon University, 2004.

55. M. Shokouhifar, M. Hasanvand, E. Moharamkhani, F. Werner, Ensemble heuristic-metaheuristic feature fusion learning for heart disease diagnosis using tabular data, *Algorithms*, **17** (2024), 34. https://doi.org/10.3390/a17010034
56. A. Prakash, O. Vignesh, R. Rani, S. Abinayaa, An ensemble technique for early prediction of type 2 diabetes mellitus–a normalization approach, *Turkish Journal of Computer and Mathematics Education*, **12** (2021), 2136–2143.
57. S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, *International Journal of Cognitive Computing in Engineering*, **2** (2021), 40–46. https://doi.org/10.1016/j.ijcce.2021.01.001
58. P. Rajendra, S. Latifi, Prediction of diabetes using logistic regression and ensemble techniques, *Computer Methods and Programs in Biomedicine Update*, **1** (2021), 100032. https://doi.org/10.1016/j.cmpbup.2021.100032
59. R. Aggrawal, S. Pal, Sequential feature selection and machine learning algorithm-based patient's death events prediction and diagnosis in heart disease, *SN Comput. Sci.*, **1** (2020), 344. https://doi.org/10.1007/s42979-020-00370-1

AIMS Press