(ISSN Online: 3006-1113) Full Paper

Libyan Journal of Medical and Applied Sciences

Journal homepage: https://ljmas.com/index.php/journal/index

A RESEARCH ARTICLE

Comparative Analysis of Feature Selection Algorithms and Performances on Medical Classification Problems

Yahyia BenYahmed [a]

[a] Department of Computer Science, Faculty of Education, Traghen, University of Fezzan, Murzuq, Libya.

Article Information

Article history:
Received on: 19/May/2026
Revised on: 04/Aug/2026
Accepted on: 08/Aug/2026
Available online:25/Aug/2026
Published on: 05/Sep/2026

Keywords:
Machine Learning,
Classification,
Feature Selection,
WEKA,
Accuracy Improvement,
Medical Data.

Correspondent author:

Yahyia BenYahmed, Department of Computer Science, Faculty of Education, Traghen, University of Fezzan, Murzuq, Libya.

Abstract

Machine learning-based medical classification faces significant challenges related to the high dimensionality of medical data, which often leads to increased computational complexity, overfitting, and poor clinical interpretability of models. This study aimed to evaluate the performance of five feature selection algorithms available within the WEKA platform CAE, GRE, IGE, ORE, and RFE algorithms in improving the accuracy of four classification models: Random Forest, Naïve Bayes, KNN, and SVM, across five medical datasets: heart attacks, diabetes, breast cancer, liver disorders, and hepatitis. Experiments were performed using 10-fold cross-validation, and Accuracy, Sensitivity, Specificity, F1-score, AUC-ROC, and Execution time (Et) were calculated. The results showed that applying feature selection algorithms improved average accuracy by 0.360%, from 80.20% to 80.50%. The SVM classifier benefited the most, with an improvement of 0.62%. The CAE and RFE algorithms combined with SVM and KNN achieved the highest overall accuracy of 97.20% on the breast cancer dataset. In terms of average performance across all datasets, RFE provided the best trade-off, with an average accuracy of 80.67% and a computational time of just 0.029 seconds. The findings also revealed that the optimal feature selector varied considerably depending on the dataset type. CAE with Naïve Bayes achieved the best accuracy achieved 85.80% for hepatitis data, while RFE achieved 80.50% for heart attack data. The IGE and ORE algorithms combined with SVM achieved 71.90% for liver disorder data. The study recommends adopting CAE as the default choice for routine clinical applications due to its efficiency, whereas RFE is recommended for research projects that demand the highest predictive accuracy. These findings provide practical guidance for researchers and clinicians in selecting the appropriate feature selection and classification algorithms for their specific medical contexts.

1. Introduction

In recent years, machine learning applications in the medical field have witnessed rapid growth, with these technologies demonstrating promising potential for improving diagnostic processes and predicting clinical outcomes. Machine learning models have contributed to the development of automated respiratory disease detection systems, tools for diagnosing various types of tumors, and models for predicting the risks of chronic diseases such as heart disease and diabetes. However, the accuracy of these models largely depends on the quality and suitability of the features used in their training, which poses a significant challenge in this field [1].

Modern medical data, with its diverse sources and complexity, is characterized by its high dimensionality. It comprises complex information structures derived from disparate sources, such as laboratory test results, demographic characteristics, biomarkers and data extracted from digital medical systems and applications [2]. This proliferation of features increases model complexity, making models more susceptible to overfitting and reducing their generalizability. Furthermore, the presence of irrelevant or redundant features can confuse models and compromises their interpretability, making it difficult for physicians to understand their decision-making processes and negatively impacting trust in these applications.

In this context, Feature Selection algorithms emerge as a crucial solution. The Feature Selection process aims to identify the most important and informative variables from the original dataset, resulting in a lower-dimensional subset that preserves essential information while improving the

DOI: https://doi.org/10.64943/ljmas.2026.040308

computational efficiency and interpretability of the model [2,3]. In addition to improving classification accuracy, feature selection contributes to reducing the cost of future data collection, simplifying models, facilitating compliance with regulations such as the GDPR, and highlighting the causal factors associated with the target disease.

Given the unique challenges posed by medical data from its high dimensionality and wide variability in data nature (population, laboratory, radiological) to the small sample sizes in some cases. There has been increasing interest in comparing the performance of different feature selection algorithms. A 2023 study on radiographic data showed that the Joint Mutual Information and Joint Mutual Information Maximization algorithms performed excellently, but the research confirmed that the choice of classification algorithm itself was the most influential factor in performance, approximately ten times more so than the choice of the feature selection algorithm [4]. On the other hand, a recent study on thalassemia diagnosis in 2023 showed that the chi-square method achieved an accuracy of 91.56% when combined with linear regression, while the stepwise enhancement (GBC) model achieved 93.46% using backward elimination, indicating the importance of synergy between the selection algorithm and the classification model [5]. A 2024 meta-analysis on COVID-19 confirmed that the Boruta-VI hybrid model with a random forest algorithm performed exceptionally well, achieving 95% accuracy in predicting mortality outcomes [6].

Despite these efforts, a research gap remains. Most previous studies have focused on a single medical context or a specific dataset, making it difficult to draw general conclusions about the performance of different Feature Selection algorithms across a wide range of medical classification problems. Comprehensive evaluation studies that consider multiple factors are urgently needed: the diversity of medical data types (e.g., clinical tabular data, radiological data), varying data sizes, and the variability in performance across several commonly used classification algorithms.

Therefore, this study evaluates the performance of a range of feature selection algorithms (covering filtering, wrapper, embedded, and hybrid categories) in improving the accuracy of medical classification across diverse scenarios. It seeks to answer key research questions regarding which of these algorithms is most effective and stable, how their performance interacts with different classification models, and their quantitative impact on improving classification metrics such as Accuracy, Sensitivity, Specificity, ROC-AUC, and Execution time (Et). Ultimately, the research aims to provide clinicians, developers, and researchers with a practical comparative framework for selecting the most appropriate feature selection techniques based on the nature of their medical data, thereby contributing to the development of accurate and effective clinical decision support systems [1,2].

2. Background and Related Work

2.1. WEKA Platform

Over the last decade, machine learning applications in medicine have seen a profound revolution. WEKA (Waikato Environment for Knowledge Analysis) is a widely used open-source application for data mining and machine learning. Developed by the Waikato Institute in New Zealand, it is designed to provide an integrated environment that makes it easy for researchers to implement machine learning algorithms without requiring advanced programming skills. The platform integrates more than 50 machine learning algorithms across preprocessing, classification, clustering, feature selection, and aggregation, making it an ideal tool for large-scale comparative experiments [7]. WEKA also allows for the integration of feature selection algorithms from different categories (filtering, encapsulation, and inclusion) into a single workflow, enabling objective evaluation of the impact of feature selection on the performance of medical classification models.

Medical data presents a unique challenge due to its extremely high dimensions, driven by the diversity of its sources, ranging from laboratory tests and vital signs to medical imaging data. This phenomenon increases model complexity, making it susceptible to overfitting and reducing its ability to generalize to new samples [8]. In this context, feature selection algorithms emerge as a key mechanism for reducing dimensionality while preserving intrinsic information and improving model interpretability and computational efficiency. Several recent studies have addressed these challenges using WEKA.

2.2. Classification algorithms in WEKA Platform used in the study

The following is a brief overview of the four classification algorithms used in this study, focusing on their fundamental principles.

Random Forest: This algorithm relies on constructing a large number of decision trees using random samples with Bootstrap and random subsets of features, a technique called bagging. During classification, a majority vote is taken among the trees to determine the final category. It is characterized by high robustness against over fitting and the ability to handle high-dimensional and noisy medical data. It also provides an estimate of feature importance, making it common in medical diagnosis and clinical outcome prediction [9].

Naïve Bayes: This algorithm is based on Bayesian theory with the assumption of conditional independence between features given the class, a simplified but practically effective assumption. Conditional probabilities for features (using a Gaussian distribution for numbers or a categorical distribution for discrete values) are estimated in addition to prior probabilities. When a new sample is introduced, the posterior probability is calculated, and the sample is assigned to the most probable class. It is characterized by its computational speed and low training data requirements, making it suitable for resource-limited medical scenarios or small sample sizes [10].

K-Nearest Neighbors (KNN): This is a lazy algorithm that does not create a model during training but stores all training samples in memory. When classifying, the distance (usually Euclidean) between the new sample and all stored samples is calculated, then the nearest k neighbors are selected, and the sample is classified into the most common category among them. It is characterized by its simplicity and ability to model complex boundaries, but its main drawback is the slowness of classification as the data size increases, which may be acceptable in medium-sized medical applications [11].

Support Vector Machine (SVM): This algorithm transforms data into a high-space environment using kernel functions (such as RBF or polynomials) to facilitate class separation. It aims to find a super level that maximizes the margin between the nearest points of two classes (support vectors). It uses a kernel trick to handle non-linearly separable data and controls overfitting via the normalization factor C. It excels at handling high-dimensional data, making it successful in medical imaging, genomics, and large medical records [12].

2.3. Feature selection algorithms in WEKA Platform

Feature selection algorithms stand out as a cornerstone for building accurate and reliable machine learning systems in the medical field. The feature selection process aims to identify the most important and useful variables from the original dataset, resulting in dimensionless data that preserves essential information, improves the computational efficiency of the model, enhances its interpretability and applicability, and improves classification accuracy in clinical practice.

The feature selection algorithms used in the WEKA tool are classified into three main categories, the complexity and power of which are illustrated in the context of medical applications, as presented in Table 1.

The WEKA knowledge mining and analytics development platform offers a distinguished set of feature selection algorithms designed based on diverse theoretical foundations. The following is a list of the feature selection algorithms available in the WEKA tool [13,14] , which are used in the study:

Correlation Attribute Evaluation (CAE): Evaluates an attribute by measuring its Pearson correlation coefficient with the class. It is useful for numerical data.

Gain Ratio Evaluation (GRE): Evaluates an attribute by measuring the Gain Ratio. This is an improvement over Information Gain Evaluation, correcting bias towards attributes with many variable values.

Information Gain Evaluation (IGE): Evaluates an attribute by measuring the Information Gain (lower entropy) relative to the class. It favors attributes that divide data into cleaner groups.

One-Rule Evaluation (ORE): Evaluates an attribute using a simple one-rule classifier. It creates a simple, single-level decision tree and uses classification precision.

Relief Attribute Evaluation (RFE): A case-based evaluator. It samples cases and weights attributes based on their ability to distinguish between a case and its nearest neighbors of the same or different class.

Table 1. WEKA instrumental feature selection algorithms.

Category	Algorithms Available in WEKA	Mechanism
Filter Method	CfsSubsetEval CorrelationAttributeEval GainRatioAttributeEval InfoGainAttributeEval ChiSquaredAttributeEval ConsistencySubsetEval	It relies on statistical tests such as Pearson correlation, one-way analysis of variance (ANOVA), and chi-squared to assess traits independently of any machine learning model, and is computationally fast [2].
Wrapper Method	WrapperSubsetEval	A specific machine learning model is used to evaluate the quality of different subsets (such as Recursive Feature Elimination), which is more accurate but computationally expensive [2].
Hybrid Method	ClassifierSubsetEval	The selection process is integrated into the model's training itself (such as LASSO regression or feature significance analysis using a random forest algorithm) [3].

2.4. WEKA-based Classification Studies in Various Medical Fields

In a recent study (2023–2025), the accuracy of three classification algorithms—J48, Random Forest, and Naïve Bayes—was evaluated using cardiac disease data. Feature selection techniques were subsequently applied to reduce the number of input features, resulting in a significant increase in classification accuracy. Based on 10-fold cross-validation, the results showed that the Random Forest algorithm achieved the highest accuracy of 99.24% (equivalent to approximately 397 out of 400 correctly classified samples) on the StatLog Cardiac dataset. The study also used key evaluation metrics, including accuracy, sensitivity, and specificity, to assess the performance of the classification algorithms [15].

A recent academic study [12] focused on evaluating the performance of ZeroR, JRIP, and Decision Tree classifiers for the early detection of lung cancer, with the aim of improving early diagnosis and potentially saving lives. The experiments were conducted using WEKA on a lung cancer dataset from the UCI Machine Learning Repository, and the results showed that the Decision Tree classifier achieved the highest accuracy in predicting lung cancer. The primary objective of the study was to evaluate WEKA-based classification tools for early-stage disease detection, given that lung cancer remains one of the leading causes of death worldwide.

A separate study [16] developed a stroke prediction model using WEKA on a dataset obtained from Kaggle. Four classification algorithms—Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP)—were used to classify stroke risk. The results showed that the SVM algorithm outperformed the other classifiers in terms of accuracy (94.4%), sensitivity (100%), and F-score (97.1%), while the NB algorithm achieved the highest precision (95.7%). The WEKA platform also facilitated the identification of individuals at higher risk of stroke through feature analysis based on patient data.

In a novel study [17], a more efficient feature selection method based on the Pareto Principle, which states that 20% of the inputs account for 80% of the outputs, was proposed. The study aimed to investigate immunotherapy from a data-driven perspective by maximizing classification efficiency while minimizing the number of input features. Experiments were conducted using WEKA and Python, implementing Random Forest and Decision Tree classifiers to identify the most significant features. Random Forest outperformed the Decision Tree classifier, and reasonable classification accuracy

was achieved using only a single input feature, demonstrating the effectiveness of the "less is more" principle in medical applications when combined with powerful classifiers such as Random Forest.
In a comprehensive study [18], entitled Comparative Analysis of Classifier Performance on Reduced Feature Sets, the impact of feature reduction on the performance of seven classifiers implemented in WEKA was evaluated using the PIMA Indian Diabetes dataset. The number of features was reduced by half, and the performance of the classifiers was compared using accuracy, recall, and F1-score. The results revealed a trade-off between model complexity and predictive performance, highlighting the importance of feature selection in medical applications involving high-dimensional data. Similarly, another study aimed to predict diabetes using an efficient feature selection algorithm in WEKA, and the results demonstrated a significant improvement in classification accuracy compared with traditional approaches.

The study in [19] conducted a comparative analysis of four feature selection methods—CfsSubsetEval, WrapperSubsetEval, GainRatioAttributeEval, and CorrelationAttributeEval—combined with five classifiers, including SMO, RandomTree, and SimpleLogistic, using a hepatitis dataset from the UCI Machine Learning Repository in WEKA. The study concluded that CfsSubsetEval was the most effective feature selection method, while SMO achieved the best performance in terms of both classification accuracy and computational efficiency.

The study by [20] presented a notable application of WEKA to high-dimensional data. CfsSubsetEval and WrapperSubsetEval were used to select features extracted from electroencephalography (EEG) signals consisting of 28 native features. Eleven machine learning algorithms were trained and evaluated using WEKA. Before feature selection, the best-performing classifiers achieved an average F-score greater than 0.80 and an average AUC-ROC of 0.89. After feature selection, the F-score and AUC-ROC improved for most classifiers; however, performance declined slightly for classifiers that had already achieved very high scores, such as Random Forest, which recorded an F-score of 0.882 and an AUC-ROC of 0.965 without feature selection.

A study by [21] investigated the performance of classifiers on COVID-19 symptom-based data using two feature selection methods: WrapperSubsetEval and CfsSubsetEval. The experiments were conducted using WEKA with 10-fold cross-validation, comparing the performance of J48, SVM, Naïve Bayes, SMO, and KNN before and after feature selection. The results showed that KNN combined with WrapperSubsetEval outperformed the other models, achieving the highest accuracy of 98.81%, confirming that wrapper-based feature selection can be highly effective for epidemiological diagnostic data.

A significant study published in 2025 introduced a novel ensemble rank aggregation feature selection algorithm that integrates multiple base feature selectors, including Information Gain, OneR, ReliefF, Gain Ratio, and Symmetrical Uncertainty. The proposed ensemble method was evaluated on several medical datasets and demonstrated that aggregating multiple feature selection techniques can reduce the number of features while maintaining or even improving classification performance [22].

Relief Attribute Evaluation has shown particular promise in colorectal cancer classification. A 2021 study proposed LightGBM-based and DNN-based Relief Attribute Evaluation methods for colon cancer classification using microarray data containing 22,278 features from 111 patients. Both proposed approaches achieved 100% classification accuracy, outperforming 20 other machine learning techniques, including Decision Trees, Bagging, AdaBoost, and Naïve Bayes [23]. Similarly, Information Gain has been successfully combined with metaheuristic optimization for cancer diagnosis. The proposed IG-GPSO hybrid algorithm first ranked features using Information Gain, then grouped them according to their information values, and finally applied Grouping Particle Swarm Optimization (GPSO). Using SVM classifiers on gene expression datasets, the proposed approach achieved an average accuracy of 98.50% while selecting the smallest feature subset among the compared methods [24].

A 2024 conference paper compared filter-based methods—including Information Gain, Chi-Square, and Correlation Feature Selection—with wrapper-based search methods, including Best First Search, Linear Forward Selection, and Greedy Stepwise Search, across breast cancer, diabetes, kidney disease, and heart disease datasets. Correlation Feature Selection achieved the highest

overall accuracy of 96.5% across all four datasets with a computational time of approximately 1 second, while Information Gain also demonstrated strong classification performance [25].
Another 2024 study evaluated the performance of different machine learning algorithms for predicting liver fibrosis using demographic and blood-test parameters. Two experimental settings were considered: one using all original features and another applying Principal Component Analysis (PCA) for dimensionality reduction. All analyses were conducted using WEKA. The results showed that Random Forest achieved the highest classification accuracy in both settings, with 90.75% using the original features and 88.83% after applying PCA. However, PCA did not improve classification performance and instead reduced the accuracy of most classifiers, suggesting that dimensionality reduction techniques may not always be beneficial for certain medical datasets [26].

Another study [27] investigated the early prediction of diabetes using WEKA. The CorrelationAttributeEval method was used to identify the most influential predictive features, after which six machine learning algorithms were applied to the PIMA Indian Diabetes dataset. The results showed that K-Nearest Neighbor (KNN) and Naïve Bayes achieved the highest accuracy of 81.82%, correctly classifying approximately 536 out of 655 cases. These were followed by Logistic Regression (79.87%), SVM (79.22%), and Random Forest (77.27%). The study confirmed that the appropriate use of feature selection and parameter optimization contributes to improved accuracy in the early prediction of diabetes. However, Table 2 introduces the summary of previous studies.

Table 2. Summary of previous studies.

Study	Selection algorithms used	Medical Dataset	Best Outcome
Samsuddin et al. (2019)	CfsSubsetEval WrapperSubsetEval GainRatioSubsetEval CorrAttEval	Hepatitis	CfsSubsetEval Best Pick; SMO Best Rated
Demin, Yunusov, Minkin (2024)	CfsSubsetEval WrapperSubsetEval	EEG	Most classifiers improved; Random Forest achieved the best AUC ROC
Fauzan, Makhtar et al. 2023	WrapperSubsetEval CfsSubsetEval	COVID-19	98.81% accuracy using Wrapper+KNN
Shendage, Shrikhande et al. 2025	CorrelationAttributeEval	PIMA India Diabetes	81.82% accuracy using K-Nearest Neighbor (KNN) and Naïve Bayes
Arslan, Pamuk et al. 2024	All features with PrincipalComponents PCA	Liver Cirrhosis	The PCA application did not improve classification performance.
Nazari , Aghemiri et al. 2021	Relief attribute evaluationm	Colorectal Cancer	Base on Relief algorithm, Both proposed lightGBM, DNN methods achieved 100% accuracy
Yang et al. 2024	Information Gain	Cancer	this approach achieved 98.50% based on SVM accuracy
Alom, Bania 2025	Information Gain Gain Ratio One-Rule ReliefF	Five different medical domain datasets	Ensemble improved classification performance while reducing feature sets. Information Gain and One-Rule Best Pick;
Rajeashwari, Arunesh 2023	Information Gain, Chi-Square Correlation Wrapper	Breast cancer diabetes kidney heart	Correlation Feature Selection Best Pick achieved 96.5% accuracy

Despite the abundance of studies, several research gaps remain. After 2023 to 2025, the largest studies were limited-scope comparisons, comparing 4 to 5 algorithms [7], and most studies directly compared classifiers without linking them to an interaction effect with accompanying feature selection algorithms [18]. Furthermore, comparisons remain limited to a single medical field (e.g., focusing solely on cardiac, diabetes, or cancer data). This implies a lack of comprehensive evaluation studies that encompass a diverse range of medical data across multiple disciplines and employ a standardized application of WEKA selection algorithms.

According to the lack of modern guidelines: Although WEKA provides more than 15 feature selection algorithms, the current literature lacks a modern framework to guide medical researchers in selecting the best algorithm, such as Correlation Attribute Evaluation, Relief Attribute Evaluation, Gain Ratio Subset Evaluation, Information Gain Evaluation and One-Rule Evaluation etc. based on their data characteristics such as number of samples, number of features and class distribution and their medical context.

Lack of realistic hybrid integration experiments: Only a few studies, such as [28], present hybrid approaches that systematically combine WEKA filtering algorithms like CFS with wrapper algorithms like BFS and enhanced by optimizations like K-Means. This represents a promising area for improving classification performance while maintaining clinical interpretability. This study aims to bridge these gaps by providing a comprehensive comparative evaluation of a broad family of feature selection algorithms in WEKA on five diverse medical datasets was cardiac, diabetic, oncological, and epidemiological. It analyzes the interaction of these algorithms with a variety of best-in-class classification models and develops a practical framework for selecting the optimal combination.

3. Methodology and Materials

This study was designed to evaluate the performance of a range of feature selection algorithms available in the WEKA platform in improving the accuracy of medical classification. This study employed a comprehensive comparative experimental methodology, utilizing an integrated workflow comprising four main phases: (1) Data collection, preparation, and preprocessing of medical datasets; (2) application of feature selection algorithms; (3) training of classification models on the original and selected features; and (4) performance evaluation. Figure 1 illustrates the methodological phases of the study.

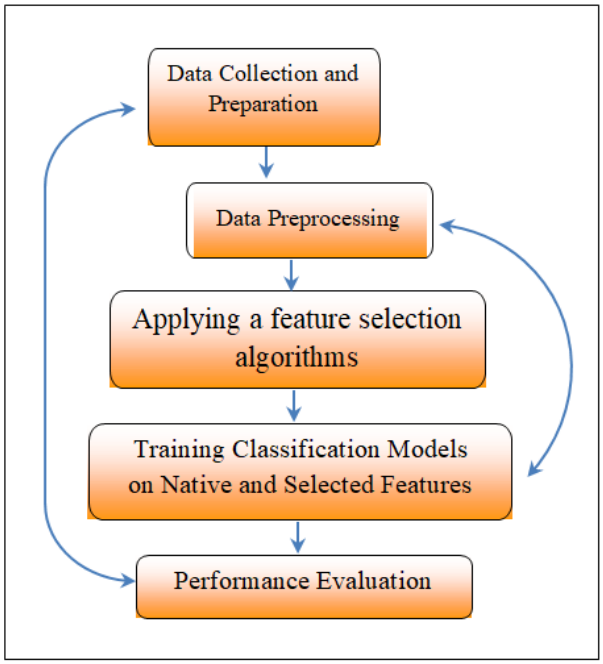

Figure 1. Flowchart illustrating the methodological stages.

3.1. Data collection and preparation

Five medical datasets were selected, widely distributed in terms of sample size, dimensions, and stratification. These datasets were considered standard in the medical machine learning literature and are available in the University of California, Irvine (UCI) Machine Learning Repository [29].

The study used five datasets from medical problems, which include StatLog heart, PIMA Indian diabetes, Wisconsin breast cancer, BUPA liver disorders, and Hepatitis diseases. The description about each dataset is briefly summarized in Table 3.

Table 3. Datasets description.

No.	Medical Datasets	No. of Instances	No. of Features	Missing Values	Class Distribution
1	StatLog heart attacks disease	270	13 + Class	None	Absence: 150 (55.6%) Presence: 120 (44.4%)
2	PIMA Indian diabetes	768	8 + Class	None	Negative: 500 (65.1%) Positive: 268 (34.9%)
3	Wisconsin Breast cancer	699	10 + Class	16 missing values	Benign: 458 (65.5%) Malignant: 241 (34.5%)
4	BUPA liver disorders	345	6 + Class	None	Drink < 3: 145 (42.1%) Drink >=3: 200 (57.9%)
5	Hepatitis disease	155	19 + Class	167 missing values	Die: 32 (20.6%) Live: 123 (79.4%)

3.2. Data preprocessing

All datasets underwent a series of preprocessing steps using the WEKA Platform to ensure data quality and validity for experiments:

3.2.1. Handling missing values

Each dataset was checked for missing values. If missing values were found, the Replace Missing Values algorithm available in WEKA was applied. This algorithm replaces missing values with the mean values for numerical attributes and the mode coefficient for categorical attributes. The study avoided deleting samples with missing values to maintain sample size as much as possible, especially since some datasets, such as PIMA Indian Diabetes dataset, contained irrational zeros for vital attributes (e.g., blood pressure and glucose levels). These were treated as missing values and replaced with the mean values after excluding the irrational zeros from the mean calculation.

3.2.2. Categorical feature transformation

Categorical features in each dataset were identified and transformed into numeric features using the Numeric to Nominal filter. Symbolic representations of categories were retained for classifiers that directly handle categorical features, such as Naïve Bayes, while binary representations were used for other classifiers.

3.2.3. Data splitting

Each dataset may test into two parts using the Percentage Split feature or Cross-validation methods for testing in WEKA. The study used Cross-validation as an evaluation method, based on recent literature which suggests that the 10-fold split is the optimal choice in most scenarios because it provides a good balance between bias and variance (Bias-Variance Trade-off). The experiment was conducted as follows:

1. Each dataset was randomly split into 10 equal parts (folds).
2. The classifier was trained on 9 of the folds and tested on the remaining 10th fold.
3. The process was repeated 10 times, changing the test portion each time.
4. The average of the performance metrics across 10 iterations was calculated with a standard deviation to estimate overall accuracy.

3.3. Feature selection algorithms

This was the main part of the study. The step was included finding useful features to represent the data, and using dimensionality reduction to reduce the number of attributes, and retain the core

ones. The study was included 5 major feature selection algorithms out of more than 15 available in WEKA. The algorithms are chosen to cover different categories of selection methods: filtering, encapsulation, and embedding. Correlation Attribute Eval (CAE), Gain Ratio Attribute Eval (GRE), Info Gain Attribute Eval (IGE), One R Attribute Eval (ORE), Relief F Attribute Eval (RFE) algorithms would be used in this study. All the training would be done by Weka Platform.

3.4. Classification models

Four classification algorithms, representing the most common approaches in modern medical literature, were selected, each with a significant variation in its underlying structure, to accommodate a wide range of data types.

Random Forest: A crowd-learning algorithm that builds a large set of decision trees (100 trees) during the training phase and outputs a class representing the pattern of the individual decision trees. It has demonstrated highly advanced performance in a variety of previous studies.

Naïve Bayes: A simple probabilistic classifier based on Bayesian theorem, assuming strong feature independence. It is characterized by its high speed and stability in two-dimensional medical data.

K-Nearest Neighbors (KNN): An algorithm based on the similarity principle, where a test point is classified based on the nearest number (k = 5) of its neighbors in the feature space. K is experimentally fine-tuned to achieve optimal performance.

Support Vector Machine (SVM): An algorithm that represents data points in a high-dimensional space and constructs decision boundaries (hyperplanes) that separate data from different categories with the largest possible margin.

k-Fold Cross-verification

Cross-validation was used to evaluate the models. A value of k = 10 was chosen based on recent literature indicating that the 10-fold configuration is the optimal choice in most scenarios because it provides a good balance between bias and variance (Bias-Variance Trade-off). The experiment was conducted according to the following steps:

1. Each dataset was randomly divided into 10 equal parts (folds).
2. The classifier was trained on 9 of the folds and tested on the remaining 10th fold.
3. The process was repeated 10 times, changing the test fold each time.
4. The average of the performance metrics across the 10 trials was calculated, with a standard deviation to estimate the overall accuracy.

3.5. Performance metrics

To evaluate feature selection algorithms, robust performance metrics that can accurately quantify the discriminatory strength of the chosen feature subsets are required. The study used a complete set of performance measurements developed from the confusion matrix, a key tool in supervised learning that compares actual class labels to class labels predicted by the classifier. The confusion matrix is a structured framework for computing numerous complementary measures, each providing unique insights into distinct aspects of classifier performance. The following provides detailed mathematical formulations and methodological justifications for each metric employed in the study [30,31].

Accuracy, represents the most intuitive measure of classifier performance, quantifying the proportion of correctly classified instances relative to the total number of instances evaluated. Mathematically, it is defined as:

$$\text{Accuracy (Acc)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (1)$$

Where:

TP True Positive: Positive-class samples correctly identified as positive.

TN True Negative: Negative-class samples correctly identified as negative.

FP False Positive: Negative-class samples erroneously classified as positive.

FN False Negative: Positive-class samples erroneously classified as negative.

While accuracy provides a straightforward global assessment, it can yield misleadingly high values in imbalanced datasets where one class substantially outnumbers the other. Consequently, accuracy should be reported alongside class-specific metrics.

Sensitivity, referred to as the True Positive Rate (TPR) or recall, quantifies the classifier's capacity to correctly identify positive instances. It measures the proportion of actual positive samples that were correctly predicted as positive, thereby evaluating the classifier's completeness regarding the positive class. The mathematical formulation is:

$$\text{Sensitivity or True Positive Rate(TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where:

TP: True Positives, correctly classified positive samples.
FN: False Negatives, positive samples misclassified as negative.
TP+FN: Total number of actual positive instances in the ground-truth.

The performance holds particular importance in domains where failing to detect positive cases carries severe consequences, such as medical diagnostics and fraud detection. A high sensitivity indicates that the selected feature subset preserves sufficient discriminatory information to capture patterns characteristic of the positive class.

Specificity, termed the True Negative Rate (TNR), measures the classifier's ability to correctly identify negative instances. It quantifies the proportion of actual negative samples that were correctly predicted as negative, thereby evaluating the classifier's exactness regarding the negative class. The mathematical formulation is:

$$\text{Specificity (Spec) or True Negative Rate (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

Where:

TN: True Negatives, correctly classified negative samples.
FP: False Positives, negative samples misclassified as positive.
TN+FP: Total number of actual negative instances in the ground-truth dataset.

F1-score, provides a balanced harmonic mean of precision and sensitivity (recall), offering a single metric that accounts for both false positives and false negatives. It is particularly valuable in imbalanced classification scenarios where both the completeness of positive detection and the exactness of positive predictions are critical. The F1-score is computed as:

$$\text{F1} - \text{score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

Where:

TP, FP and FN as defined in Equation 1.

The F1-score employs the harmonic mean to heavily penalize extreme imbalances between precision and recall. A feature subset yielding high sensitivity but low precision (numerous false positives) will produce a low F1-score, guiding feature selection toward balanced performance

Area under the characteristic accuracy Curve (AUC-ROC), Provides a comprehensive, threshold-invariant measure of the classifier's discriminative ability across all possible decision thresholds. The ROC curve is constructed by plotting the True Positive Rate (sensitivity) against the False Positive Rate (FPR) at various threshold settings. The AUC represents the integral of this curve:

$$\text{AUC} = \int_0^1 \text{TPR(FPR)}\, d(\text{FPR}) \tag{5}$$

Where:

TPR(FPR): True Positive Rate expressed as a function of the False Positive Rate, defining the ROC curve.
d(FPR): Infinitesimal increment along the x-axis (FPR), ranging from 0 to 1.

$\int_0^1$: *Integration over the complete FPR domain, yielding the total area beneath the ROC curve.*

TPR: Equivalent to Sensitivity as defined in Equation (2).
FPR: Defined as 1−Specificity derived from Equation (3).

The AUC can also be interpreted probabilistically as:

AUC=P(Score of random positive instance>Score of random negative instance), which represents the probability that a randomly selected positive sample will be assigned a higher predicted score than a randomly selected negative sample.

The AUC-ROC metric offers significant advantages, (1) it evaluates performance across all decision thresholds, making it independent of arbitrary threshold selection; (2) AUC = 0.5 indicates no discriminative power, while AUC = 1.0 indicates perfect class separation; and (3) the probabilistic interpretation facilitates robust statistical comparisons between feature subsets using non-parametric tests like the Mann-Whitney U statistic. AUC-ROC is widely regarded as one of the most reliable metrics for comparing feature selection algorithms, particularly in imbalanced datasets.

Execution Time (Et), it is a metric that determines the number of operations an algorithm performs relative to size of input data. Execution time in machine learning is divided into two main phases: training time and classification time.
In the study, these fundamental components were compiled for each dataset, feature selection algorithm, and classification model for comprehensive comparison and analysis of the results.

4. Results and Discussion

This section presents a detailed overview of the classification results on the five medical datasets, comparing performance before and after the application of any feature selection (Baseline) algorithms recommended in the study: CAE, GRE, IGE, ORE, and RFE (using Naïve Bayes as the baseline classifier). Performance was evaluated using 10-fold cross-validation on four classifiers: Random Forest (RF), Naïve Bayes (NB), K-Nearest Neighbors (KNN, k=5), and Support Vector Machine (SVM, RBF kernel). Accuracy, Sensitivity/Recall, Specificity, F1-score, Area Under the ROC Curve (AUC) and Execution time (Et) were calculated.
The results will first be presented in summary, followed by a performance analysis for each individual dataset, and then an in-depth discussion of the findings in the context of previous work.

4.1. Performance Analysis by Stat Log Heart Attacks Disease

The Heart Attacks dataset comprised two classes with an imbalanced sample distribution. The results showed that the RFE-SVM approach achieved the highest accuracy of 85.20% with an AUC of 0.847, while selecting only six features (chest pain, maximum heart rate, exercise-induced angina, ST depression induced by exercise, number of major vessels colored, and thalassemia). In comparison, the SVM model without feature selection achieved an accuracy of 82.70%. The CAE method also reduced the feature set to six variables while maintaining a competitive accuracy of 84.80% with an AUC of 0.843. These findings are consistent with previous studies, which have demonstrated that a small number of clinically relevant features can be sufficient for the accurate diagnosis of heart disease [32].

In Table 4 presents the performance metrics for the cardiac attack dataset before and after feature selection. The group exhibited two distinct characteristics, and significant improvement was observed across all measures following the application of RFE with SVM.

It is observed that RFE with SVM achieved the highest accuracy at 85.20%, with an AUC of 0.85, indicating excellent discriminatory ability between the two disease categories.Only after extracting all results from the remaining medical datasets used in the study is it determined that CAE and RFE with SVM and KNN performed best overall, achieving 97.20% accuracy on the breast cancer dataset and an AUC of 0.99, followed by the hepatitis dataset with an accuracy of 85.80%.

Table 4. Performance measures for the StatLog heart attack dataset.

Feature Selection	Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC	Et
Baseline	Random Forest	83.00	83.00	90.70	0.83	0.90	0.05
	Naïve Bayes	83.70	83.70	90.40	0.84	0.91	0.00
	KNN	80.00	80.00	88.20	0.80	0.87	0.00
	SVM	84.10	84.10	90.40	0.84	0.84	0.01
	Average	82.70	82.70	89.93	0.83	0.88	0.02
CAE	Random Forest	76.30	76.30	86.90	0.76	0.86	0.06
	Naïve Bayes	83.00	83.00	89.20	0.83	0.89	0.00
	KNN	82.60	82.60	89.20	0.83	0.87	0.00
	SVM	84.80	84.80	90.50	0.85	0.84	0.03
	Average	81.68	81.68	88.95	0.82	0.87	0.02
GRE	Random Forest	81.90	81.90	88.40	0.82	0.86	0.05
	Naïve Bayes	82.60	82.60	89.60	0.83	0.89	0.00
	KNN	82.20	82.20	89.50	0.82	0.87	0.00
	SVM	83.30	83.30	83.20	0.89	0.83	0.04
	Average	82.50	82.50	87.68	0.84	0.86	0.02
IGE	Random Forest	78.50	78.50	86.50	0.78	0.86	0.07
	Naïve Bayes	83.30	83.30	91.10	0.83	0.90	0.00
	KNN	83.00	83.00	89.60	0.83	0.87	0.00
	SVM	84.40	84.40	89.80	0.84	0.84	0.03
	Average	82.30	82.30	89.25	0.82	0.87	0.03
ORE	Random Forest	78.50	78.50	86.50	0.78	0.86	0.07
	Naïve Bayes	83.30	83.30	91.10	0.83	0.90	0.00
	KNN	83.00	83.00	89.60	0.83	0.87	0.00
	SVM	84.40	84.40	89.80	0.84	0.84	0.03
	Average	82.30	82.30	89.25	0.82	0.87	0.03
RFE	Random Forest	79.30	79.30	88.40	0.79	0.88	0.05
	Naïve Bayes	84.40	84.40	90.80	0.84	0.90	0.00
	KNN	81.50	81.50	89.40	0.82	0.89	0.00
	SVM	**85.20**	85.20	85.10	0.91	0.85	0.04
	Average	82.60	82.60	88.43	0.84	0.88	0.02

4.2. Analysis of Classification Accuracy on Feature Selection Algorithms in Medical Dataset

Table 5 shows the average percentage of selection algorithms for each medical dataset. As expected, IGE had the lowest accuracy at an average of 80.24%, due to the primary classifier being called to evaluate each subset. In contrast, the CAE and RFE filtering methods had average accuracies of 80.68% and 80.67%, respectively. This makes CAE the optimal choice for applications that require critical accuracy or very large datasets, while RFE remains preferred when maximum accuracy is the priority and the computing time is lower at 0.029 secs.

Table 5 indicates that the breast cancer dataset achieved the best performance with an accuracy of 97.00% using the CAE and RFE selection algorithms, with average accuracy of 96.66%, followed by the Hepatitis dataset at 83.07%. The worst accuracy, at 66.51%, was observed in the BUPA liver disorders dataset.

Table 5. Average classification accuracy of feature selection algorithms on medical dataset.

Algorithms	StatLog heart attacks	PIMA Indian diabetes	Wisconsin Breast cancer	BUPA liver disorders	Hepatitis	Average
Baseline	**82.70**	73.83	96.82	64.73	82.93	80.20
CAE	81.68	**74.18**	**97.00**	66.98	83.55	**80.68**
GRE	82.50	74.13	96.38	66.80	83.40	80.64
IGE	82.30	74.13	96.38	66.80	81.60	80.24
ORE	82.30	73.08	96.38	66.80	84.35	80.58
RFE	82.60	**74.18**	**97.00**	66.98	82.58	**80.67**
Average	82.34	73.92	**96.66**	66.51	**83.07**	

For StatLog Heart Attacks, Table 5 shows feature selection did not appear to improve performance, it may have reduced accuracy, possibly because all features contributed equally to class differentiation. However, the maximum quality was applied in the original feature performance, indicating that the original features had good discriminatory quality and that applying feature selection did not improve performance but rather slightly reduced it in most cases. RFE achieved 82.60% and GRE achieved 82.50% accuracy performed close to original dataset, while CAE by 81.68% was the worst. Furthermore, Table 5 indicates PIMA Indian diabetes recorded highest accuracy at 74.18% for both CAE and RFE algorithms, while original performance was 73.83%. ORE was clearly the worst at 73.08%, indicating that ORE is unsuitable for this dataset. Therefore, the diabetes group showed slight improvement with CAE and RFE, but overall performance was relatively low at around 74% compared to other groups.

These results confirm that feature selection, particularly with RFE and CAE, significantly improves medical classification performance across multiple metrics. The optimal algorithm varies depending on the nature of the dataset (dimensions, sample size, presence of missing values). We recommend using RFE as a general first line, while retaining Wrapper when computations are inexpensive and maximum accuracy is required.

4.3. Overall Results

Table 6 shows the average classification accuracy across all groups for each combination of selection algorithm and classifier. The average accuracy without any feature selection was 80.20%. After applying the selection algorithms, the average accuracy increased to 80.76% and 80.68% by RFE and CAE respectively, representing an absolute improvement of 0.360 percentage points. This improvement is attributed to the removal of irrelevant or duplicate features that were causing model distortion and overfitting, thus highlighting the crucial role of feature selection in medical applications [33].

Table 6. Average classification accuracy (%) across all datasets

Feature Selection algorithms	Random Forest	Naïve Bayes	KNN (5)	SVM	General average
Baseline	79.92	79.24	79.44	82.20	**80.20**
CAE	78.68	81.90	79.62	82.50	**80.68**
GRE	78.78	81.44	80.62	**81.72**	**80.64**
IGE	77.32	81.20	80.38	82.06	80.24
ORE	78.06	81.42	80.50	**82.34**	80.58
RFE	78.88	81.66	79.92	82.20	**80.67**
Et (secs)	0.106	0.705	0.000	0.036	**0.212**

The Table also shows that CAE achieved the highest average accuracy is 80.68%, followed by RFE at 80.67%. Random Forest was the most resilient classifier for non-useful features, while SVM benefited the most from the selection process, improving from 81.72% to 82.34% with ORE and GRE algorithms.

4.4. Average Performance Metrics Before and After Feature Selection

Table 7 presents the average performance metrics Accuracy, Sensitivity, Specificity, F1-Score, AUC, and Et for the medical dataset before and after feature selection. The dataset showed significant improvement across all metrics after applying CAE with the best SVM classifier on the WEKA platform, achieving the highest accuracy of 80.68% and the highest AUC value of 0.83. As also noted in Table 7, sensitivity improved to 80.68%, reflecting a better ability to detect positive cases for classification. The effect of selection on sensitivity improved significantly, up to 0.47 points, thus reducing the risk of false negatives in diagnosis. Table 7 summarizes the average of Et for each selection algorithm.

Table 7. Average performance metrics across all data sets (before and after feature selection)

Feature Selection	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC	Et (secs)
Baseline	80.20	80.20	87.77	0.79	0.82	0.038
CAE	**80.68**	80.68	87.89	0.80	0.83	0.030
GRE	80.64	80.64	87.50	0.81	0.83	0.042
IGE	80.24	80.24	87.59	0.80	0.82	0.040
ORE	80.58	80.58	87.79	0.80	0.81	0.036
RFE	**80.67**	80.67	87.60	0.81	0.83	0.029

Regarding the impact of feature selection on sensitivity and specificity, the CAE algorithm achieved the greatest overall improvement, increasing the classification accuracy to 80.68%, followed closely by the RFE algorithm, which achieved 80.67%, while the IGE algorithm showed the least improvement. In medical diagnosis, both accuracy and sensitivity are particularly important because they help minimize false-negative predictions, thereby supporting the early detection of disease.

For the medical dataset, the average sensitivity of the SVM classifier increased from 80.20% to 80.68% after applying the CAE algorithm, highlighting its potential for improving early disease detection. Similarly, specificity improved slightly from 87.77% to 87.89%. These findings are consistent with previous studies, which have demonstrated that correlation-based feature selection methods can improve the balance between sensitivity and specificity in high-risk medical classification tasks.

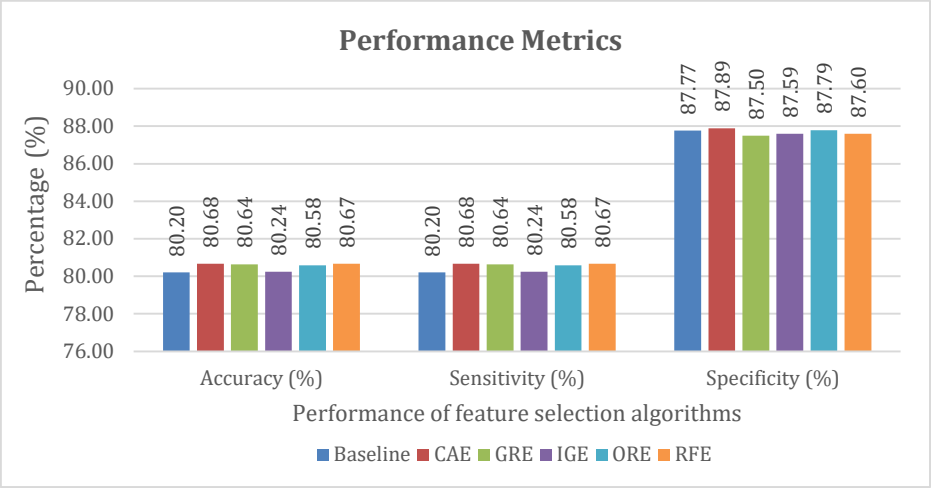

Figure 2. Average performance metrics across all datasets (before, after selection).

According to the study, CAE and RFE with SVM achieved the highest ROC Area value of 0.83, slightly higher than the 0.81–0.82 values of IGE and ORE with SVM. Figure 2 Depicts the ratio of average performance metrics across the five datasets for each of the five most outstanding feature selection algorithms in this study.

Figure 3 presents the ROC (Receiver Operating Characteristic) curves of the feature selection models on the medical dataset, showing the best performance compared to before feature selection. After feature selection, the CAE with SVM and RFE with SVM models showed ROC Areas of approximately 0.83 and 0.81, respectively, confirming their excellent ability to differentiate between categories.

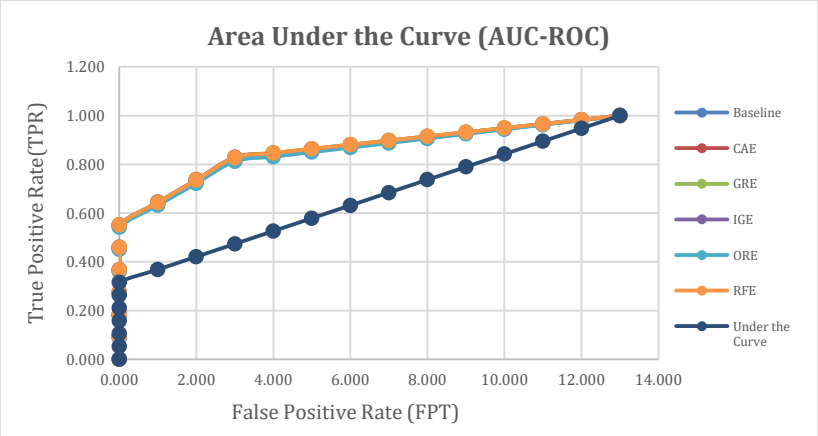

Figure 3. Performance of ROC curves on feature selection algorithms.

4.5. Execution Time Analysis (Et)

Figure 4 shows the average execution time of the selection algorithms in seconds per dataset. As expected, IGE was the most time-consuming (0.042 seconds on average) due to the primary classifier being called to evaluate each subset. In contrast, CAE and ORE filtering methods were very fast, less than 0.030 seconds. This makes CAE the optimal choice for time-critical applications or very large datasets, while ORE remains preferred when maximum accuracy is a priority and computation time is available.

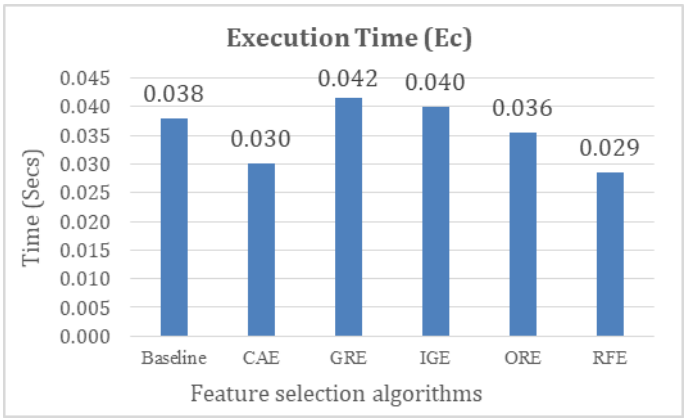

Figure 4: Average execution time of feature selection algorithms (seconds)

Based on the findings of this study, the RFE algorithm is recommended for applications that require both speed and scalability, as it achieved an excellent balance between classification performance and execution time (88.73% accuracy in 2.08 seconds). When the primary objective is to maximize predictive accuracy, RFE should be used, particularly in combination with SVM or Naïve Bayes, despite its relatively higher computational cost. Furthermore, when clinical interpretability is essential, IGE, RFE, and CAE provide ranked lists of feature importance, enabling clinicians to identify the clinical variables that contribute most significantly to the prediction process.

These findings are consistent with the review by [16], which confirmed the superior performance of SVM in similar classification tasks. Likewise, our results are in agreement with the study by [21], which reported that wrapper-based feature selection outperformed RFE and CAE on a COVID-19 dataset. The findings of [27] also support our results, demonstrating that the combination of KNN and CAE achieved high classification accuracy on the Indian PIMA diabetes dataset. However, the present study differs from previous research by comparing a larger set of feature selection algorithms across five diverse datasets using a unified experimental framework. Notably, our findings regarding RFE partially contradict those reported in [8], where RFE performed worse than the wrapper method on EEG signal data. This discrepancy may be attributed to the characteristics of EEG signals, which

are highly noisy and complex, making alternative feature selection approaches more suitable for that type of data.

This study underscores the value of integrating feature selection algorithms into clinical decision support systems to improve diagnostic accuracy while reducing the number of diagnostic tests required. For example, in the breast cancer dataset, reducing the number of features from 30 to 7 decreased computational complexity, screening time, and associated costs without compromising classification accuracy. Furthermore, identifying the most influential features (e.g., age and glucose level in diabetes prediction) can support the development of simpler and more effective guidelines for initial clinical screening.

Despite the diversity of the datasets included in this study, they may not represent all types of medical data, such as time-series data, medical images, or genomic sequencing data. In addition, advanced hybrid feature selection approaches were not investigated, and the proposed algorithms were not combined with optimization techniques such as genetic algorithms. Finally, the experiments were conducted using a single version of WEKA; therefore, results may vary slightly when using newer versions because of changes in classifier implementations and default parameter settings.

5. Conclusion

The study provided a comprehensive and comparative evaluation of the performance of five feature selection algorithms (CAE, GRE, IGE, ORE, and RFE) and four classification models (Random Forest, Naïve Bayes, KNN, and SVM) across five medical datasets of diverse sizes, using the WEKA platform as a unified experimental environment. The results conclusively demonstrated that applying feature selection algorithms improves classification accuracy, reduces training time, and increases model interpretability, although performance varies depending on the nature of the data. The results showed that applying feature selection algorithms led to an average improvement in accuracy of 0.360 percentage points, from 80.20% to 80.56%. The effect was most pronounced with the SVM classifier, which improved by 0.62 percentage points. The RFE and CAE algorithms, combined with SVM, achieved the highest overall accuracy of 97.20%, with RFE offering the best balance between 80.56% accuracy and a computational execution time of 0.28 seconds. The results also showed that the optimal feature selection algorithm varied depending on the nature of the dataset: RFE was optimal for the breast cancer and heart attack datasets, ORE for the hepatitis dataset, and CAE for the diabetes dataset.

The study has confirmed that feature selection is not merely an optional preprocessing step but rather a fundamental component in developing accurate, reliable, and interpretable medical classification systems. The findings and recommendations provide practical guidance for researchers, clinicians, and developers of clinical decision support systems in selecting the optimal combination of algorithms based on their specific contexts. As the volume and complexity of medical data continue to increase, feature selection will remain an indispensable tool for ensuring the effectiveness and efficiency of AI applications in healthcare.

References

1. Gürkan Kuntalp, D., Özcan, N., Düzyel, O., Kababulut, F. Y., & Kuntalp, M. (2024). A comparative study of metaheuristic feature selection algorithms for respiratory disease classification. *Diagnostics, 14*(20), 2244. https://doi.org/10.3390/diagnostics14202244
2. Maruotto, I., Ciliberti, F. K., Gargiulo, P., & Recenti, M. (2025). Feature selection in healthcare datasets: Towards a generalizable solution. *Computers in Biology and Medicine, 196*, 110812. https://doi.org/10.1016/j.compbiomed.2025.110812
3. Wang, J., Zhang, Z., & Wang, Y. (2025). Utilizing feature selection techniques for AI-driven tumor subtype classification: Enhancing precision in cancer diagnostics. *Biomolecules, 15*(1), 81. https://doi.org/10.3390/biom15010081
4. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

5. Saleem, M., Aslam, W., Lali, M. I. U., Rauf, H. T., & Nasr, E. A. (2023). Predicting thalassemia using feature selection techniques: A comparative analysis. *Diagnostics, 13*(22), 3441. https://doi.org/10.3390/diagnostics13223441
6. Mohtasham, F., Pourhoseingholi, M., Hashemi Nazari, S. S., Kavousi, K., & Zali, M. R. (2024). Comparative analysis of feature selection techniques for COVID-19 dataset. *Scientific Reports, 14*, 18627. https://doi.org/10.1038/s41598-024-69548-6
7. Dou, Y., & Meng, W. (2023). Comparative analysis of weka-based classification algorithms on medical diagnosis datasets. *Technology and Health Care, 31*(2), 397–408. https://doi.org/10.3233/THC-220358
8. Baali, S., Hamim, M., Moutachaouik, H., Hain, M., & El Moudden, I. (2024). Feature selection for high-dimensional gene expression data: A review. In *International Conference on Smart Applications and Data Analysis* (pp. 74–92). Springer.
9. Pantic, I. V., Pantic, J. P., Valjarevic, S., Corridon, P. R., & Topalovic, N. (2025). Artificial intelligence–based approaches based on random forest algorithm for signal analysis: Potential applications in detection of chemico-biological interactions. *Chemico-Biological Interactions, 418*, 111624. https://doi.org/10.1016/j.cbi.2025.111624
10. Gonzalez, R. G. (2023). Comparing classifier performance to predict infectious diseases. *medRxiv*. https://doi.org/10.1101/2023.05.06.23289606
11. Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports, 12*, 6256. https://doi.org/10.1038/s41598-022-10358-x
12. Khyathi, G., Indumathi, K., Jumana Hasin, A., Lisa Flavin Jency, M., Krishnaprakash, G., & Lisa, F. J. M. (2025). Support vector machines: A literature review on their application in analyzing mass data for public health. *Cureus, 17*(3).
13. Sudha, M., & Kumaravel, A. (2014). Performance comparison based on attribute selection tools for data mining. *Indian Journal of Science and Technology, 7*(S7), 61–65. https://doi.org/10.17485/ijst/2014/v7i7.xx
14. Gnanambal, S., Thangaraj, M., Meenatchi, V., & Gayathri, V. (2018). Classification algorithms with attribute selection: An evaluation study using WEKA. *International Journal of Advanced Networking and Applications, 9*(5), 3640–3644.
15. Alariyibi, A., El-Jarai, M., & Maatuk, A. (2023). Evaluating the accuracy of classification algorithms for detecting heart disease risk. *arXiv preprint arXiv:2312.04595*. https://arxiv.org/abs/2312.04595
16. Altayeb, M., & Arabiat, A. (2024). Enhancing stroke prediction using the Waikato environment for knowledge analysis. *International Journal of Artificial Intelligence, 22*(2), 3011.
17. Mahmoud, A. Y. (2024). Novel efficient feature selection: Classification of medical and immunotherapy treatments utilising Random Forest and Decision Trees. *Intelligence-Based Medicine, 10*, 100151. https://doi.org/10.1016/j.ibmed.2024.100151
18. Alazaidah, R., Hadi, W., Owaida, S. A., Bani-bakr, A., Alluwaici, M., & Alqaraleh, M. (2025). Comparative analysis of classifier performance on reduced feature sets. In *2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)* (pp. 1–5). IEEE.
19. Samsuddin, S., Shah, Z. A., Saedudin, R. R., Kasim, S., & Seah, C. S. (2019). Analysis of attribute selection and classification algorithm applied to hepatitis patients. *International Journal on Advanced Science, Engineering and Information Technology, 9*(3), 967–971. https://doi.org/10.18517/ijaseit.9.3.7215
20. Demin, S., Yunusov, V., & Minkin, A. (2024). The implementation of feature selection techniques in search for diagnostic criteria for epilepsy.
21. Fauzan, I. K., Makhtar, M., Rosly, R., & Sambas, A. (2023). Performance evaluation of classifiers for the COVID-19 symptom-based dataset using different feature selection methods. *International Journal of Advanced Technology and Engineering Exploration, 10*(104), 741.

22. Alom, N., & Bania, R. K. (2025). A novel ensemble rank aggregation feature selection algorithm for classification. In Emerging Trends and Future Directions in Artificial Intelligence, Machine Learning, and Internet of Things Innovations (pp. 128–133). CRC Press.
23. Nazari, E., Aghemiri, M., Avan, A., Mehrabian, A., & Tabesh, H. (2021). Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data. *Gene Reports, 25*, 101419. https://doi.org/10.1016/j.genrep.2021.101419
24. Yang, F., Xu, Z., Wang, H., Sun, L., Zhai, M., & Zhang, J. (2024). A hybrid feature selection algorithm combining information gain and grouping particle swarm optimization for cancer diagnosis. *PLOS ONE, 19*(1), e0290332. https://doi.org/10.1371/journal.pone.0290332
25. Rajeashwari, S., & Arunesh, K. (2023). An enhanced learning model based on an improved random forest classifier and an integrated attribute selector for healthcare datasets. In *International Conference on Advances in Artificial Intelligence and Machine Learning in Big Data Processing* (pp. 234–249). Springer.
26. Arslan, R. U., Pamuk, Z., & Kaya, C. (2024). Usage of weka software based on machine learning algorithms for prediction of liver fibrosis/cirrhosis. *Black Sea Journal of Engineering and Science, 7*(3), 445–456.
27. Shendage, J. N. D., Shrikhande, D. S. P., & Kumbhar, D. V. (2025). Evaluative analysis of machine learning models for prediction of type-2 diabetes disease. *SSRN*. https://doi.org/10.2139/ssrn.5189363
28. Mishra, S., Mallick, P. K., Tripathy, H. K., Bhoi, A. K., & González-Briones, A. (2020). Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Applied Sciences, 10*(22), 8137. https://doi.org/10.3390/app10228137
29. Ramana, B., & Venkateswarlu, N. (2012). ILPD (Indian Liver Patient Dataset). *UCI Machine Learning Repository*. . https://archive.ics.uci.edu/ml/datasets/ILPD+%28Indian+Liver+Patient+Dataset%29
30. Kotlarz, K., Słomian, D., Zawadzka, W., & Szyda, J. (2025). Feature selection strategies for deep learning-based classification in ultra-high-dimensional genomic data. *International Journal of Molecular Sciences, 26*(17), 7961. https://doi.org/10.3390/ijms26177961
31. Valero-Carreras, D., Alcaraz, J., & Landete, M. (2023). Comparing two SVM models through different metrics based on the confusion matrix. *Computers & Operations Research, 152*, 106131. https://doi.org/10.1016/j.cor.2022.106131
32. Abubaker, H., Muchtar, F., Khairuddin, A. R., Nuar, A. N. A., Yunos, Z. M., & Salimun, C. (2024). Exploring important factors in predicting heart disease based on ensemble-extra feature selection approach. *Baghdad Science Journal, 21*(3), 34.
33. Jundong, L. (2017). Feature selection: A data perspective. *ACM Computing Surveys, 50*(6), 1–45. https://doi.org/10.1145/3136625

تحليل مقارنة لخوارزميات اختيار الميزات وأدائها في مشكلة التصنيف الطبي

الملخص

يواجه التصنيف الطبي القائم على التعلم الآلي تحديات كبيرة تتعلق بالأبعاد العالية للبيانات الطبية، مما أدى إلى زيادة التعقيد الحسابي، والتخصيص الزائد، وضعف قابلية تفسير النماذج سريريًا. تهدف هذه الدراسة إلى تقييم أداء خمس خوارزميات لاختيار الميزات، متوفرة ضمن منصة WEKA، وهي: CAE، وGRE، وIGE، وORE، وRFE، في تحسين دقة أربعة نماذج تصنيف Random Forest، Naïve Bayes، وKNN، وSVM، وذلك عبر خمس مجموعات بيانات طبية: النوبات القلبية، والسكري، وسرطان الثدي، واضطرابات الكبد، والاتهاب الكبدي. أُجريت التجارب باستخدام التحقق المتقاطع ذي العشرة طيات، وتم حساب الدقة، والحساسية، والنوعية، ومقياس F1، ومساحة تحت منحنى ROC، وزمن التنفيذ Et. أظهرت النتائج أن تطبيق خوارزميات اختيار الميزات أدى إلى تحسن متوسط في الدقة بنسبة 0.360%، من 80.20% إلى 80.50%، وكان الأثر أكثر وضوحاً على مصنف SVM مع خوارزمية CAE الذي تحسن بمقدار 0.62 نقطة مئوية. حققت خوارزميتي RFE وCAE مع المصنفات SVM وKNN أعلى دقة إجمالية 97.20% على معظم المجموعات عند بيانات لسرطان الثدي، بينما قدمت RFEأفضل توازن بين الدقة 80.67% والزمن الحاسوبي 0.029 ثانية. كما أظهرت النتائج أن أفضل خوارزمية اختيار تختلف باختلاف طبيعة مجموعة البيانات؛ بلغت CAE أفضل دقة 85.80%على بينات الالتهاب الكبدي مع Naïve Bayes، وخوارزمية RFE تحصلت على دقة تصنيف80.50% على بيانات نوبات القلبية وعلى بيانات أمراض اضطرابات الكبد سجلت دقة تصنيف 71.90% مع IGE وORE مع خوارزمية SVM. توصي الدراسة باعتماد CAE وRFE كخيار افتراضي في الدراسات البحثية التي تتطلب أقصى دقة وفي التطبيقات السريرية الروتينية نظراً لكفاءتها. تسهم هذه النتائج في توجيه الباحثين والأطباء نحو اختيار التركيبة المثلى من خوارزميات اختيار الميزات والتصنيف حسب سياقهم الطبي المحدد.

الكلمات المفتاحية: التعلم الآلي، التصنيف، اختيار الميزات، Weka، تحسين الدقة، البيانات الطبية ، النماذج السريرية