Journal of Biological Regulators and Homeostatic Agents 2026, 40(3), 126.
https://doi.org/10.65746/jbrha126

Article

Balancing dimensionality reduction and information retention in medical datasets using adaptive relevance scoring and trifold evolutionary optimization

Asha Shanmugham*, Parvathy Meenakshi Sundaram

Department of Computer Science and Engineering, Sethu Institute of Technology, Pulloor, Kariapatti, Virudhunagar, Tamil Nadu 626115, India
*** Corresponding author:** Asha Shanmugham, asharithum@gmail.com

CITATION

Shanmugham A, Sundaram PM. Balancing dimensionality reduction and information retention in medical datasets using adaptive relevance scoring and trifold evolutionary optimization. Journal of Biological Regulators and Homeostatic Agents. 2026; 40(3): 126. https://doi.org/10.65746/jbrha126

ARTICLE INFO

Received: 4 June 2026
Revised: 15 June 2026
Accepted: 6 July 2026
Available online: 27 July 2026

Abstract: High-dimensional medical datasets often contain redundant, irrelevant, and noisy features that increase computational complexity. Existing approaches for dimensionality face several critical challenges: (i) aggressive feature reduction leading to loss of critical clinical information, (ii) selection unfairness in model-dependent methods such as Recursive Feature Elimination (RFE), (iii) reliance on variance in Principal Component Analysis (PCA) (iv) poor generalization of existing feature selection methods across diverse medical datasets and (v) uncertainty about whether retained features truly capture essential clinical information after DR. To address these limitations, this study proposes an adaptive feature selection framework that balances DR with effective information retention. The proposed method introduces a Gene Noise Filtering Layer along with entropy-based filtering, Mutual Information, and variance filtering to remove low-informative features. An Adaptive Relevance Score (ARS) is developed to identify clinically significant features. A Trifold Evolutionary Optimization (TEO) approach, integrating Particle Swarm Optimization, Quantum-inspired search, and genetic mutation, is employed to determine optimal feature subsets. Experimental evaluation achieves a Feature Stability Index (FSI) of 0.9819, indicating excellent preservation of informative features and high stability of the selected feature subsets. The proposed method improved classification accuracy, reduced dimensionality, and attained better preservation of clinically meaningful information across multiple medical datasets.

Keywords: adaptive relevance score; dimensionality reduction; gene noise filtering layer; mutual information; trifold evolutionary optimization

1. Introduction

Dimensionality Reduction (DR) refers to a set of techniques that transform high-dimensional data into a lower-dimensional representation without changing the most significant information [1]. The high-dimensional data, particularly those encountered in gene expression analysis has a variety of challenges to consider, such as the curse of dimensionality, where more features create inefficient levels of computation, more complex models, and a higher risk of overfitting [2,3]. Machine learning models can be ineffective in generalizing when the size of the feature space increases, leading to lower performance on unseen data [4]. DR methods are used to remove redundancy and irrelevant characteristics, to reduce noise and storage requirements, and to enhance computational efficiency [5,6]. The techniques convert the large datasets into smaller sizes that should retain the original format and necessary features of the datasets [7]. Consequently, DR is very vital in improving data visualization, interpretation, and predictive model performance in areas like bioinformatics, signal processing,

neuroinformatics, and other areas where the datasets usually involve thousands of features [8,9]. Medical data, especially those based on gene expression information, clinical data, and medical images, are generally high-dimensional, noisy, and small [10]. These behaviors greatly enhance the complexity of computations, memory usage, and possible erroneous classification of any disease, which may be detrimental to clinical decision-making [11,12]. The datasets for these applications may contain a large number of features but relatively few patient samples, leading to the well-known challenges of dimensionality and data sparsity [13]. The DR techniques like Principal Component Analysis (PCA) [14] and Recursive Feature Elimination (RFE) [15] are common in reducing the feature space, although feature elimination can be aggressive, potentially leading to loss of important medical data, feature selection bias, and loss of prediction accuracy [16]. Therefore, there is a growing need for advanced feature selection techniques that can effectively reduce dimensionality while preserving clinically significant information [17]. It is possible to strike a balance between the DR and the information retention to enhance the classification accuracy, decrease the cost of computing, and create more reliable and efficient medical diagnosis systems [18]. Traditionally, DR and feature selection methods such as PCA and RFE have been widely used in medical datasets. However, these approaches face several critical challenges (i) High-dimensional medical data increases computational complexity, memory consumption, and may reduce classification accuracy due to inefficient feature representation [19]. (ii) Aggressive feature reduction can result in the loss of clinically significant information, which is essential for accurate diagnosis [20]. (iii) model-dependent techniques such as RFE introduce feature selection bias due to their reliance on specific learning models and training data [21]. (iv) Variance-based methods such as PCA fail to ensure that selected features are predictive or clinically relevant, as they ignore class labels [22]. (v) Existing feature selection approaches are sensitive to noise and outliers and often lack robustness and generalizability across heterogeneous medical datasets, leading to inconsistent performance [23]. (vi) After DR, the interpretability of the selected feature set is reduced, and it becomes unclear whether the retained features preserve clinically meaningful information, thereby affecting model reliability and clinical trust [24].

To overcome the limitations of traditional optimization algorithms such as PSO, GA, and QPSO, we propose a hybrid feature selection framework that integrates Adaptive Relevance Scoring (ARS), entropy-based filtering, and Trifold Evolutionary Optimization (TEO). The proposed TEO combines swarm intelligence-based search, quantum-inspired position updating, and evolutionary mutation mechanisms to improve exploration capability, avoid premature convergence, and maintain population diversity in high-dimensional feature spaces. This hybrid optimization strategy enables effective selection of optimal gene subsets while preserving clinically significant information, improving robustness, and maintaining high predictive performance.

1.1. Research contributions

The study makes the following key contributions to balance DR and information retention in medical datasets.

- Information-Preserving Feature Selection Framework: This study develops a novel feature selection framework that can achieve the desired balance between DR and the maintenance of important information in high-dimensional medical data. It also solves the shortcomings of the conventional approaches that usually cause loss of vital features. Thus, enhancing accuracy in classification without compromising the efficiency of the computation.
- Adaptive Relevance Score (ARS) to Feature Evaluation: The study developed an ARS to measure feature relevance by combining predictive relevance, stability, and inter-feature interaction. This multidimensional scoring system guarantees strong and sound identification of informative features that are significant from traditional selection methods.
- Interaction-Aware Hybrid Trifold Evolutionary Optimization (TEO): We introduce a hybrid TEO approach that combines Particle Swarm Optimization, quantum-inspired search, and genetic mutation. The methodology improves the search in high-dimensional feature spaces and performs effective choice of optimal feature subsets.
- Increased Classification with less complexity: The suggested framework attains a high level of DR with significant predictive information and will consequently lead to a better classification. The model uses a small and informative subset of features to minimise the computational complexity without loss.

The manuscript is organized as follows. Section 2 reviews existing feature selection and optimization techniques for medical datasets. Section 3 proposes the hybrid framework of combining ARS and TEO. Section 4 explains the experimental design, data, and measurements. Finally, conclusions and future research directions are provided in Section 5.

2. Related work

This section analyses the existing research studies on feature selection techniques and optimization algorithms used on medical data to enhance DR and predictive accuracy. Kabir et al. [25] analyzed the performance of DR algorithms for cancer prediction using the Support Vector Machine (SVM) model and improved the prediction accuracy by reducing high-dimensional features. However, this work does not use adaptive gene relevance analysis and advanced hybrid optimization methods to select the most informative genes from high-dimensional datasets. Sanju et al. [26] developed a Hybrid Feature Selection and Deep Learning Framework (HFSDLF) for the prediction of thyroid disease using Random Forests, PCA, and L1 regularization, improving the efficiency of feature selection. However, this framework does not incorporate entropy-based gene noise filtering and adaptive relevance scoring mechanisms to preserve predictive and stable gene interactions during the feature selection process. Ashika et al. [27] developed an Optimized Rough Set Theory-Machine Learning (RST-ML) framework integrated with Multi-Criteria Decision-Making (MCDM) techniques for heart disease prediction using XGBoost, improving feature selection and diagnostic accuracy. However, the method does not address high-

dimensional gene optimization using hybrid swarm, quantum-inspired, and evolutionary search strategies for efficient feature subset selection.

Kumar et al. [28] proposed a hybrid framework combining classical and quantum-inspired machine learning techniques with the SVM model for heart disease prediction, improving diagnostic accuracy and predictive performance. However, the adaptive relevance-based gene ranking and trifold evolutionary optimization to enhance feature selection efficiency and reduce dimensionality. Razzaque et al. [29] proposed a PCA-based feature extraction and Modified Particle Swarm Optimization (MPSO) based feature selection framework for medical data classification using SVM, which achieved higher classification accuracy on high-dimensional gene expression datasets. However, future work can focus on improving feature selection strategies, enhancing optimization techniques for high-dimensional data, and developing more efficient methods for selecting compact and informative gene subsets.

By analysing these existing models, the medical data is difficult because it is high-dimensional, heterogeneous, and the interactions among features are complicated. Most of the methods are detrimental due to loss of vital information, selection bias, and lack of generalization because they are based on fixed or single criteria. Although optimization algorithms such as PSO, GA, and QPSO have been widely used for feature selection, individual methods still have limitations. PSO may suffer from premature convergence and reduced diversity during high-dimensional search, GA requires higher computational cost due to evolutionary operations, and QPSO may face challenges in maintaining a proper balance between exploration and exploitation. The feature count before and after DR is shown in **Table 1**. To overcome these problems, the present framework suggests a hybrid architecture that combines Adaptive Relevance Scoring and Trifold Evolutionary Optimization (TEO), which integrates swarm intelligence, quantum-inspired updating, and evolutionary mutation strategies to improve search efficiency, avoid local optima, and effectively select informative features.

Table 1. DR consequences for high-dimensional medical datasets.

Dataset	No. of features before	Relevant features (after)	Reduced (%)	Accuracy
RNA sequencing data [25]	20,531	250	98.78 %	95 %
microarray medical datasets [29]	12,600	70	99.44 %	88 %
ARS + TEO + SVM (Proposed)	54,675	80	99.85 %	99.54 %

3. Proposed methodology

The study proposes a novel dimensionality reduction and feature optimization mechanism for high-dimensional medical data as shown in **Figure 1**. The dataset is preprocessed by a Gene Noise Filtering Layer, which filters out genes that have very low Shannon entropy, which eliminates features that contain little information about the biological process, while keeping the significant features. Then, a hybrid filtering method using Mutual Information (MI), Information Gain and variance thresholding is employed to keep the most relevant genes related to the class labels to minimize the computational cost.

A gene is then assigned an Adaptive Relevance Score (ARS) that combines

predictive relevance, data variation stability and preservation of interactions between different genes. The genes with the greatest rankings in the ARS are chosen for further optimization. To explore high-dimensional feature space effectively, the proposed Trifold Evolutionary Optimization (TEO) framework integrates the Particle Swarm Optimization [30], quantum inspired probabilistic search and evolutionary mutation. Finally, a multi-objective fitness function is developed to maximize classification accuracy and minimize the subset size, and the selected gene subset is classified by SVM classifier for the better performance in disease prediction.

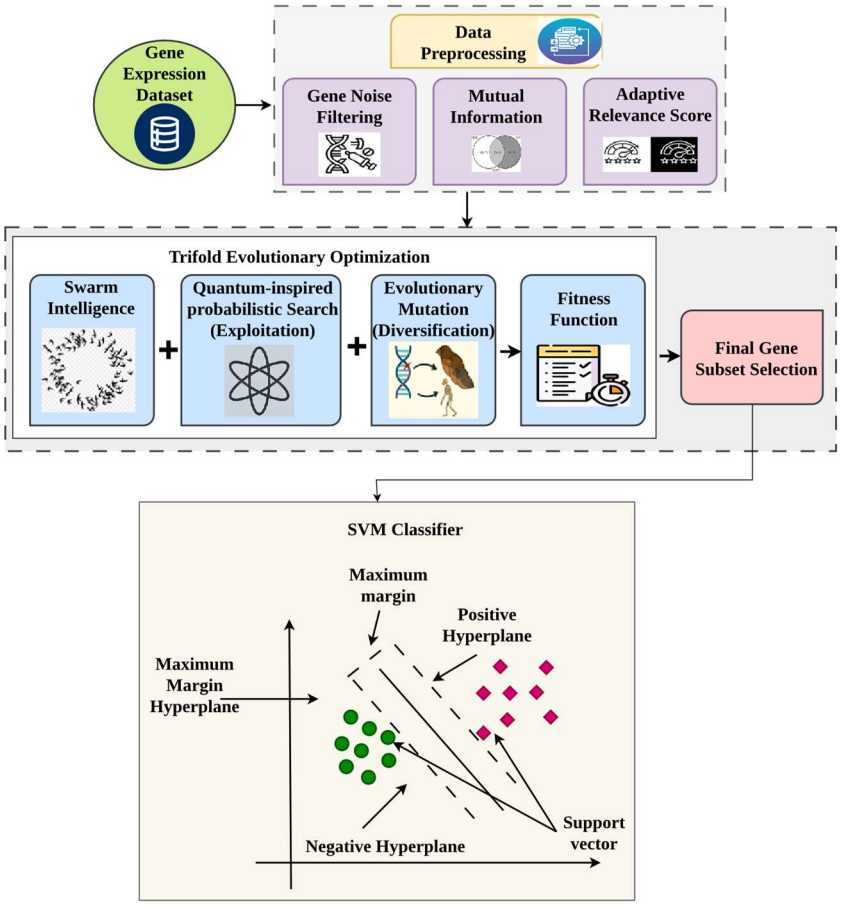


Figure 1. Overview of the proposed hybrid feature selection framework for gene expression data analysis.

3.1. Dataset description

The Breast Cancer Gene Expression (CuMiDa) dataset [31] is a high-dimensional microarray dataset containing gene expression profiles of 151 breast cancer samples with 54,676 gene features. There are 5 different types of breast cancer (plus healthy tissue) represented in this dataset (column “type”), making it suitable for multi-class classification tasks. Curated and preprocessed from Gene Expression Omnibus (GEO) as part of the CuMiDa database, it is commonly used in bioinformatics and ML research for cancer classification, feature selection, and biomarker discovery, though it presents challenges due to its large number of features and small sample size. The proposed approach incorporates a Gene Noise Filtering Layer, Adaptive Relevance Score (ARS), and Trifold Evolutionary Optimization (TEO) to ensure efficient and

meaningful feature selection. A strict separation between training and testing phases is maintained during preprocessing and optimization to avoid data leakage. In addition, a multi-objective fitness function is used to balance classification accuracy with feature subset compactness, helping to reduce both overfitting and underfitting issues.

3.2. Data preprocessing

Data preprocessing is used to reduce the high-dimensional Breast Cancer Gene Expression - CuMiDa data, to keep the most informative genes for the disease prediction [32]. Initially, entropy-based noise filtering is applied to identify and remove genes with very low variability across samples. These low-information genes exhibit minimal expression changes and therefore contribute little to classification performance.

The filtered noise is then used to calculate the relation of each gene to target disease classes using Mutual Information (MI) [33]. Genes with stronger associations with the class labels are retained for further analysis with more information. This will compress the feature space while maintaining the most important predictive data.

Next, the selected genes are ranked by a global score called the Adaptive Relevance Score (ARS) [34]. The ARS combines multiple evaluation factors such as predictive relevance, stability, and interaction preservation to find informative and stable genes. Lastly, the genes ranking in the top positions in the ARS results are chosen for the optimization step.

3.3. Trifold evolutionary optimization (TEO)

TEO is suggested to optimize gene subset selection of high-dimensional medical data [35], where the search space is enormous because of the existing large number of genes, by integrating the capabilities of Particle Swarm Optimization, quantum probability position updates, and evolutionary mutation methods to effectively search the solution space and find the most informative gene subsets that will maximize the classification performance and minimize redundancy, and the evolutionary mutation component adds variety to the population. In the proposed optimization process, the new position of the individual particles is calculated by a quantum-inspired update rule given by Equation (1).

Which X_i^{t+1} denotes the updated position of particle i at iteration $t+1$ and P_i^t denotes the best position of the particle i at iteration t. The term $Mbest^t$ represents the average best position of the swarm members, to direct the search to promising areas of the solution space.

$$X_i^{t+1} = P_i^t \pm \beta \mid Mbest^t - X_i^t \mid ln\left(\frac{1}{u}\right) \tag{1}$$

Where β is initialized within the range of $[0.5, 1.0]$ and is adaptively reduced during the optimization process to gradually shift the search behavior from global exploration to local exploitation. The updating strategy of β is defined as Equation (2).

$$\beta(t) = \beta_{max} - \frac{(\beta_{max} - \beta_{min}) \times t}{T_{max}} \tag{2}$$

Where β_{max} and β_{min} represent the maximum and minimum values of β, respectively, t denotes the current iteration, and T_{max} represents the maximum number of iterations. At the beginning of optimization, a higher β value increases the search diversity and helps particles explore different gene subsets. As iterations progress, β decreases gradually, enabling particles to focus on exploiting the most informative feature regions and improving convergence.

The logarithm term enables probabilistic movement of particles, allowing the optimizer to explore the high-dimensional feature space efficiently and avoid local optima. This mechanism is applied to select the optimal gene subset for disease prediction.

3.3.1. Fitness function

The fitness function is used to assign a numerical value to each of the subsets to balance between the accuracy of classification and the reduction of features [36]. It ensures that the selected subset achieves high predictive performance while minimizing the number of genes. This multi-objective formulation enhances the efficiency of computation and reduces redundancy with high-dimensional data. The fitness value can be calculated using Equation (3).

$$Fitness = \alpha \times Accuracy + (1 - \alpha) \times \left(\frac{N_{selected}}{N_{total}}\right) \quad (3)$$

Where Accuracy represents the classification performance of the selected gene subset, $N_{selected}$ denotes the number of selected genes determined by the optimizer, and N_{total} represents the total number of genes in the dataset. The parameter α controls the trade-off between classification accuracy and feature reduction. In this study, α is set to 0.8 based on experimental evaluation, assigning higher importance to classification accuracy while still encouraging the selection of a compact gene subset. A higher α value prioritizes prediction accuracy, whereas a lower α value increases the emphasis on reducing the number of selected features.

3.3.2. Final gene subset selection

Based on the fitness evaluation, the gene subset that has the highest fitness level is chosen as the final optimum one. The subset is the most informative and most relevant in prediction of the disease, and it narrows down on redundancy in the dataset. The optimized gene subset is then given as an input to the classification model to be further processed.

3.4. Classification model: Support vector machine (SVM)

The classification stage of the proposed framework uses an SVM [37] to classify samples based on the optimized gene subset obtained from the previous stages.

SVM constructs an optimal decision boundary that separates different disease classes based on maximum margin. Hence, better classification is achieved in high-dimensional datasets as shown in **Figure 2**. During training, the classifier learns the relationship between gene expression patterns and disease labels, and the trained model is used to predict the class of unseen samples. The optimized gene subset generated by the proposed optimization framework is used as input features for the

SVM classifier, which is defined as in Equation (4).

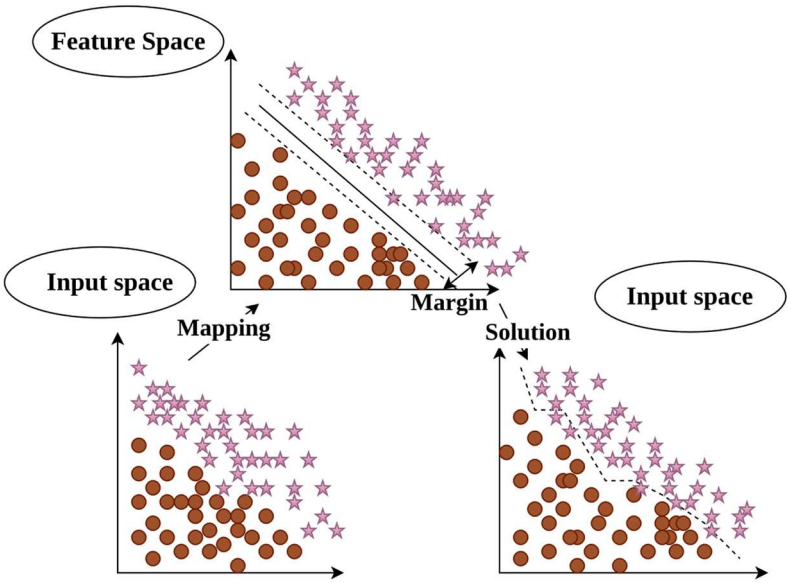


Figure 2. Architecture of support vector machine (SVM) showing data mapping from input space to feature space for maximum-margin classification.

$$f(x) = w^T x + b \tag{4}$$

Where $f(x)$ is the decision function and w is the weight vector that indicates the orientation of the separating hyperplane in the feature space. The parameter b shifts the hyperplane to give optimum separation between the classes. x is the input feature vector of the chosen subset of genes. This is to find the best values of w and b such that overall, the margin between the two classes is the greatest, and the classification errors are the lowest.

3.4.1. Optimization of the objective function

The SVM determines the optimal decision boundary by solving an optimization problem that maximizes the separation margin between different classes while minimizing classification errors. This objective ensures robust classification performance and improved generalization for high-dimensional medical datasets. The optimized gene subset selected by the proposed framework is used to train the classifier for accurate disease prediction as in Equations (5) and (6).

$$min_{w.b} = \frac{1}{2} \| W \|^2 \tag{5}$$

$$y_i(w^T x_i + b) \geq 1 \tag{6}$$

Where y_i is the label of the i^{th} training sample, which is typically +1 or -1 in binary classification. The constraint assures every training sample is properly categorized and at the same time, the largest possible separation margin among classes is maintained. The SVM attains greater generalization to unseen data by maximizing this margin.

4. Results and discussion

In this section, the proposed framework to evaluate its performance across high-dimensional datasets.

4.1. Hyperparameter and computational complexity analysis

The dataset is split into 70 % for training, 20 % for testing, and 10 % for validation. The preprocessing includes Normal Scaler for feature normalization, entropy-based filtering with a threshold defined as shown in Equation (7) and Adaptive Relevance Score (ARS)-based feature selection with optimized parameters.

$$Th = \mu(H) + 1.5\sigma(H) \quad (7)$$

Here, $\mu(H)$ and $\sigma(H)$ represent the mean and standard deviation of gene entropy distribution, respectively, while the multiplier 1.5 acts as a sensitivity control parameter to balance noise removal and information retention by regulating the aggressiveness of filtering. The Trifold Evolutionary Optimization (TEO) optimizer and Support Vector Machine (SVM) classifier are then applied with detailed hyperparameter settings, as summarized in **Table 2**.

Table 2. Hyperparameters of the proposed model.

Model	Hyperparameter	Value
Data preprocessing	scaler	StandardScaler
Entropy filtering	threshold	1.5
	top_pre	2000
	compute_stability folds	5
ARS	interaction_similarity k	5
	ARS weights R:S:I	0.4:0.3:0.3
	top_idx	800
	n_particles	40
	iterations	80
TEO optimizer	max_genes	80
	lambda_param	0.5
	penalty	0.9
	param_grid C	[0.1, 1, 10, 50, 100, 200, 500]
	param_grid gamma	[1, 0.1, 0.01, 0.001, 0.0001]
	kernel	‘rbf’
SVM classifier	probability	True
	class_weight	‘balanced’
	cv	5
	scoring	‘balanced_accuracy’
	Train	70 %
Split	Test	20 %
	Val	10 %

4.2. Performance analysis of the proposed model

This section analyses the performance metrics of the proposed model, which demonstrates better performance in all measures of evaluation. The model attains 99.54 % accuracy, 99.21 % precision, 99.54 % recall, specificity of 99.62 %, and 99.32 % of F1-score, indicating good performance of the proposed model. Also, the MCC of 98.95 %, AUC of 99.81 %, and NPV of .58 % shows the strong performance of the proposed model. The specificity of 99.62 % is also high, with very low False Positive Rates (FPR) of 0.0038 % and False Negative Rates (FNR) of 0.0057 % indicates minimal misclassification errors. Furthermore, the feature selection stability is validated using Feature Stability Index (FSI = 0.9819), demonstrating highly stable and consistent feature subsets across multiple runs. Generally, these results confirm that the proposed model is robust, stable, and highly effective for classification tasks.

Figure 3 displays the confusion matrix of the proposed framework using six classes: HER, basal, cell-line, luminal-A, luminal-B, and normal. Each row denotes the predicted labels, and each column denotes the true labels. The diagonal elements represent correctly classified samples, while the off-diagonal elements indicate misclassification cases. The proposed framework demonstrates strong classification ability across all biological classes. Some minor confusion is observed between biologically similar categories, such as luminal-A and luminal-B, due to their similar gene expression patterns. These misclassifications may occur because of overlapping molecular characteristics among related breast cancer subtypes. Overall, the results confirm the effectiveness of the selected gene subsets in improving multi-class classification performance.

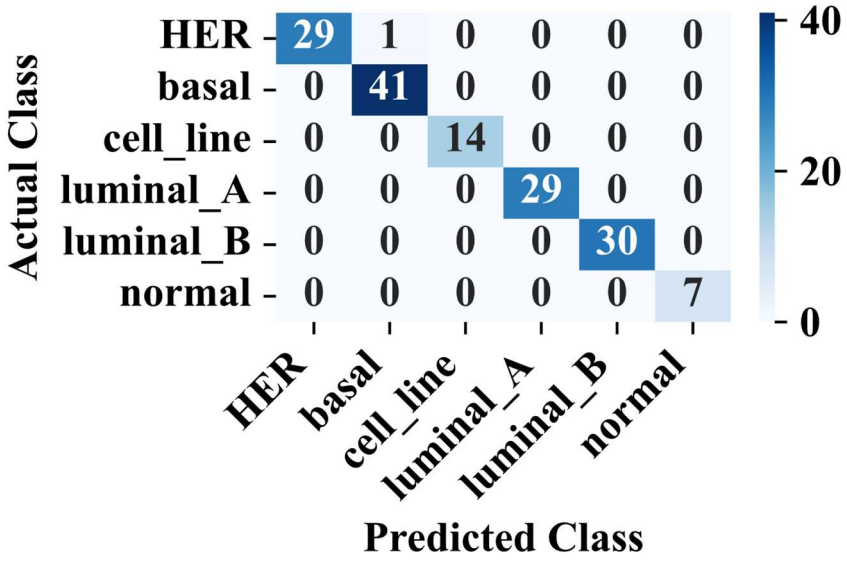


Figure 3. Confusion matrix of the proposed framework.

Figure 4 demonstrates the progressive reduction of features across five selection stages on a logarithmic scale on the y-axis. Initially, the dataset contains a very large number of original features that are slightly reduced after entropy filtering. There is a major reduction after preselection of MI, reducing the number of features to thousands of orders. Further refinement using ARS ranking reduces the feature. Finally, the TEO optimizer drastically minimizes the feature set, indicating highly efficient DR while

retaining the most relevant features for the model.

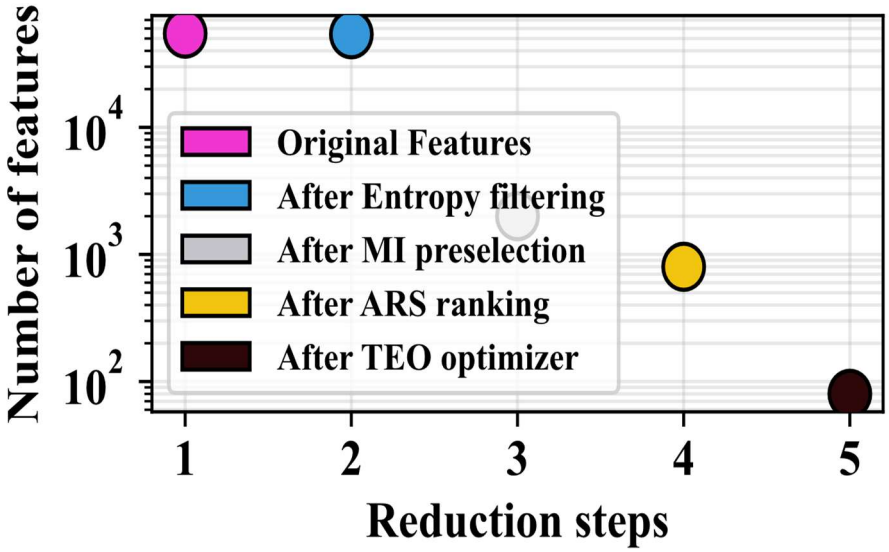
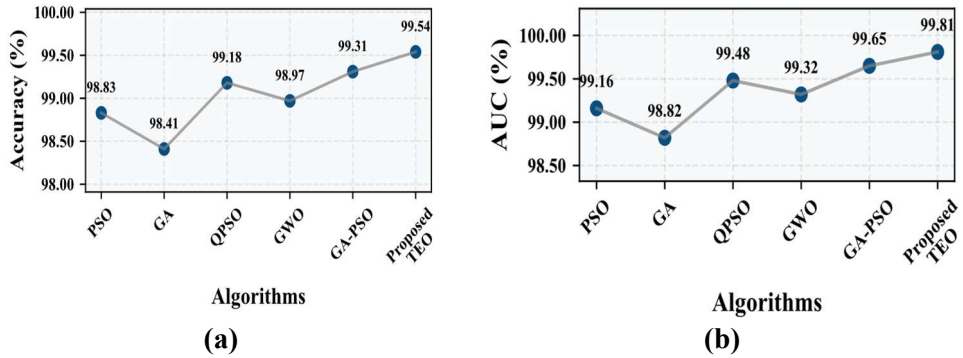


Figure 4. Number of features after each reduction stag.

Figure 5 compares the performance of different optimization algorithms, namely Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Quantum-behaved Particle Swarm Optimization (QPSO), Grey Wolf Optimization (GWO), the hybrid Genetic Algorithm–Particle Swarm Optimization (GA_PSO), and the proposed Thermal Exchange Optimization (TEO). As shown in the figure, the proposed TEO consistently achieves the best performance across all evaluation metrics. It attains the highest **Figure 5a** Accuracy, **Figure 5b** AUC, **Figure 5c** F1-score, **Figure 5d** Precision, **Figure 5e** Recall, **Figure 5f** Specificity, **Figure 5g** NPV and **Figure 5h** MCC, while producing the lowest **Figure 5i** FPR and **Figure 5j** FNR, Although GWO and GA_PSO provide competitive results with accuracy values of 99.31 % and 98.57 %, respectively, their F1-score, precision, recall, and MCC values remain lower than those of TEO. PSO and GA exhibit comparatively lower performance, whereas QPSO achieves moderate improvements but still falls short of the proposed approach. Overall, the consistent superiority of TEO across all metrics demonstrates its enhanced optimization capability, resulting in more reliable feature selection and improved classification performance compared with the existing optimization algorithms.



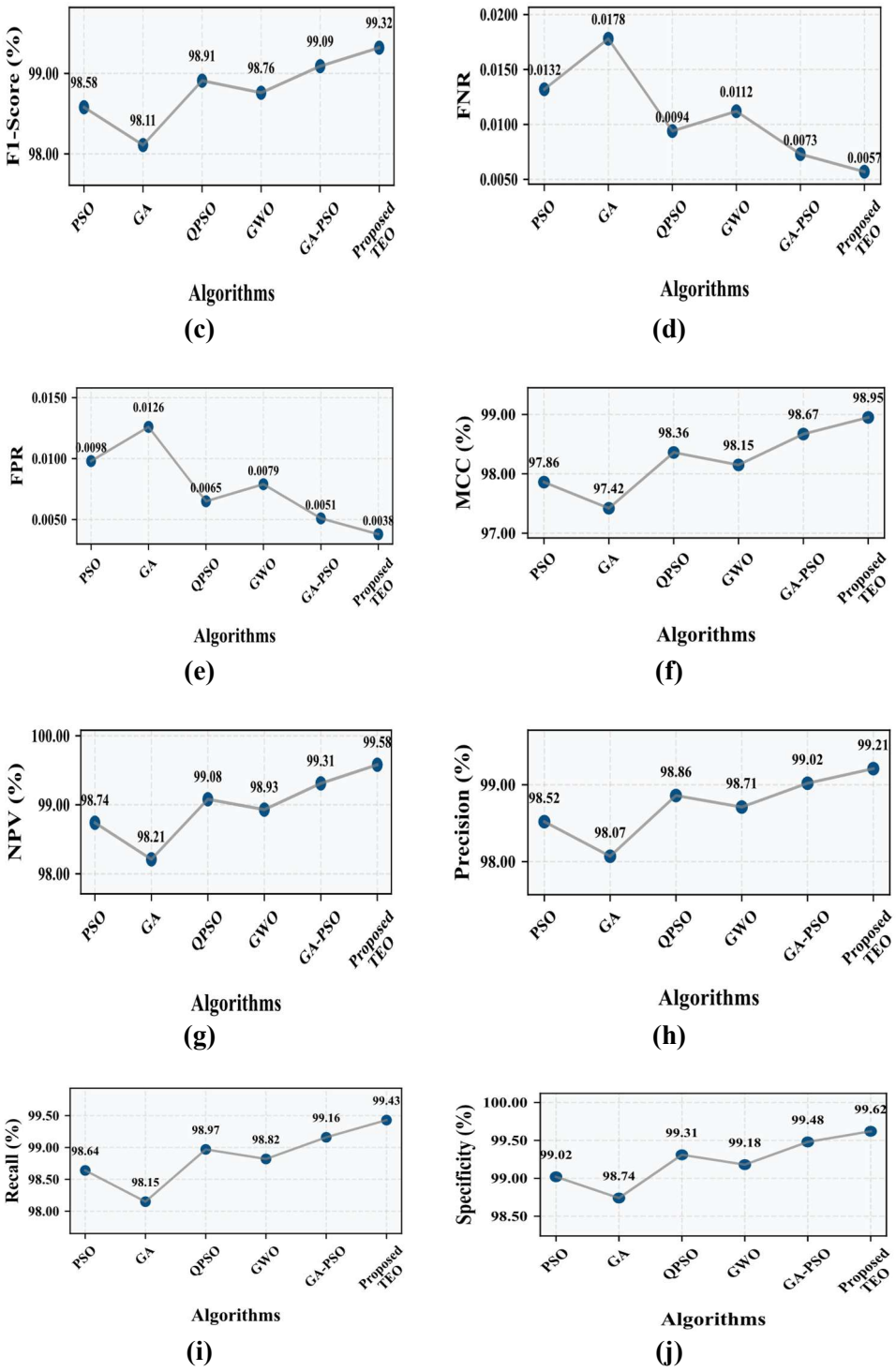


Figure 5. Comparison of optimization algorithms using multiple performance metrics, including **(a)** Accuracy, **(b)** AUC, **(c)** F1-score, **(d)** Precision, **(e)** Recall, **(f)** Specificity, **(g)** NPV, **(h)** MCC, **(i)** FPR, and **(j)** FNR.

Figure 6 presents the visualization of feature distribution comparing original features and features after filtering with entropy. The density of the points indicates that the majority of features are condensed to a range of optimized dimensions of features. The similarity between original and filtered features implies that entropy filtering does not remove informative or redundant features, but it does not change the

overall data structure. Although there is reduction, the fundamental distribution pattern is maintained, which means that there is slight loss of information. This confirms that entropy filtering effectively refines the feature space without distorting the intrinsic characteristics of the dataset.

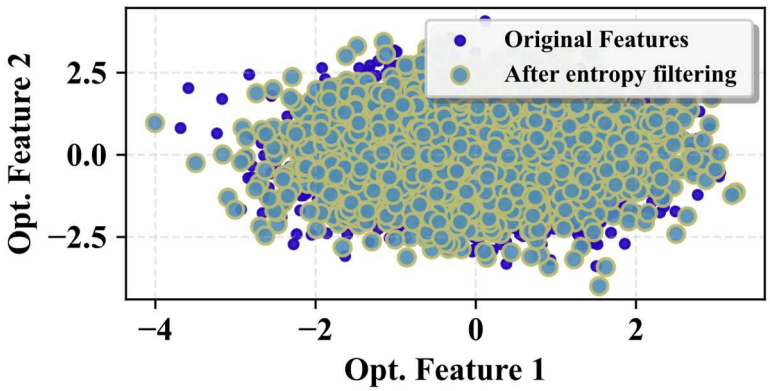


Figure 6. Visualization of feature distribution before and after entropy filtering.

Figure 7 shows the distribution of original features and the features that are retained after preselection of MI. The original features are very scattered over the feature space, which implies that there is a lot of dimensional redundancy. After MI preselection is more concentrated in the center, which corresponds to the features being most informative and relevant. This great decrease in spread shows that MI is certainly a good filter to eliminate the not relevant features of the data but leaves the fundamental model integral. The step enhances the quality of the features and sets the dataset ready to be optimized and classified further.

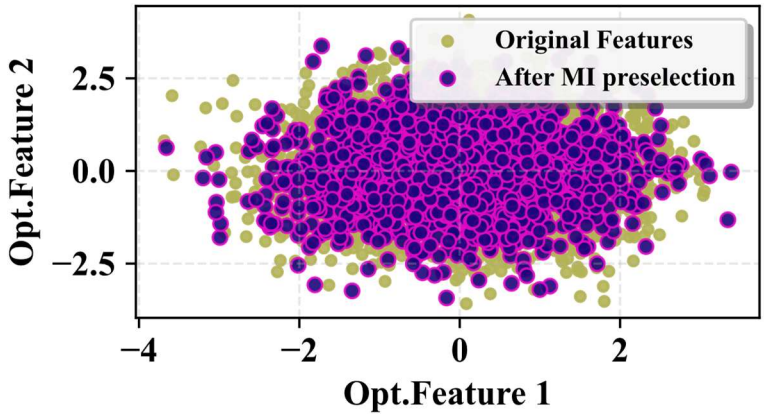


Figure 7. Feature space distribution before and after MI preselection.

Figure 8 presents the original feature set and the refined features after positioning by ARS. The original features are widely spread, which means that there are redundancy and irrelevant information. On the contrary, the ARS-ranked features are less and more centralized, which emphasizes the choice of the most important features. The lower density and closer clustering reveal that ARS is effective in standing high-impact features and removing noise. This action maximizes the ability of the feature space to be discriminatory and aid enhanced model performance at the later stages.

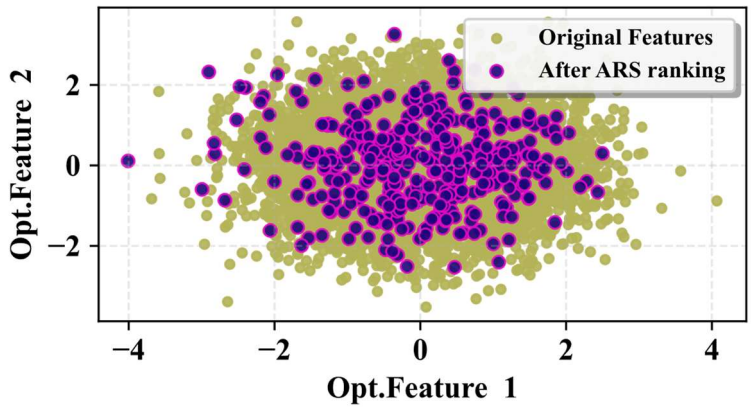


Figure 8. Feature space distribution before and after ARS ranking.

Figure 9 proves the reduction of features by comparing the initial feature distribution with the optimized features using the TEO algorithm. The features are over-spread on the feature space, meaning they are highly dimensional and redundant. The features left after the TEO optimization are extremely few, and it clearly shows aggressive and effective dimensionality reduction. The chosen features are not densely located and, on the contrary, are located in strategic positions, being the most informative and discriminative features. Finally, the TEO can remove unnecessary data but keep the key tendencies, resulting in enhanced computational work and model precision.

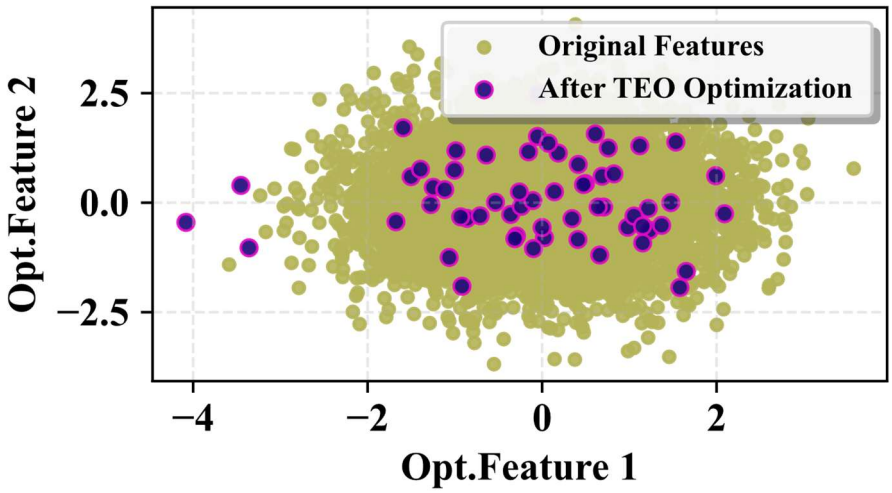


Figure 9. Feature space distribution before and after TEO optimization.

Figure 10 presents the Precision–Recall (PR) curves for the six different classes such as HER, basal, cell_line, luminal_A, luminal_B, and normal, are presented to measure the performance of the model. All curves remain close to the top-right corner, indicating consistently high precision across varying recall levels. The PR scores are very high for all the classes. Among these, basal has the highest performance of 0.996, followed by luminal_B at 0.994. Then cell_line at 0.993, HER at 0.989, luminal_A at 0.988, and normal at 0.984. Overall, the curves confirm excellent discriminative capability and reliability across all classes.

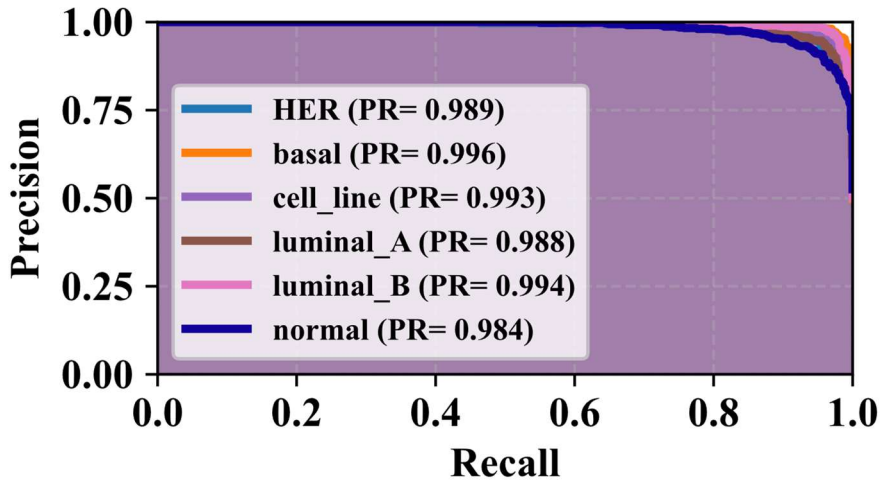


Figure 10. Precision–recall (PR) curves for multi-class classification performance.

Figure 11 shows the ROC curves for the six different classes, namely HER, basal, cell_line, luminal_A, luminal_B, and normal. It is evident that the model is trying to achieve the best trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). All curves are positioned close to the top-left corner, indicating excellent classification performance with high sensitivity and low false positives. The Area Under the Curve (AUC) values are very high for all classes, with luminal_B (0.996) and cell_line (0.994) achieving the best results, followed by basal (0.993), luminal_A (0.991), normal (0.988), and HER (0.986). The steep rise of the curves at low FPR further confirms strong discriminative ability. Overall, the model is trying to achieve high discriminative performance with the steeply rising curves at low FPR.

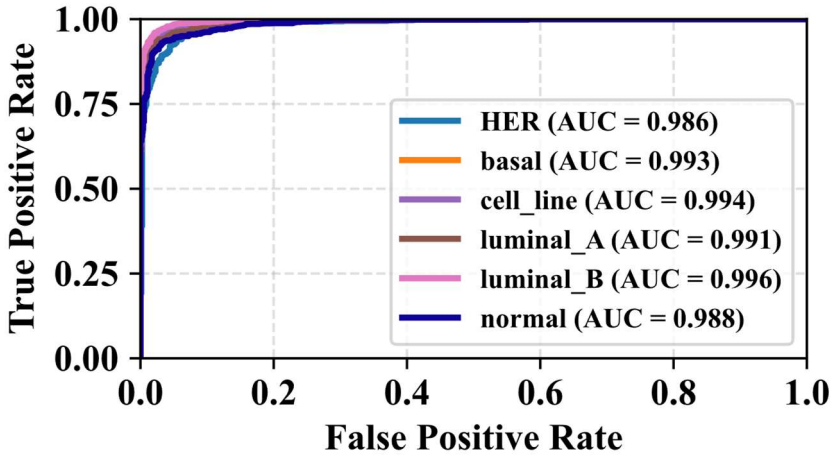


Figure 11. Receiver operating characteristic (ROC) curves for multi-class classification.

Table 3 presents the progressive reduction of features across the different selection processes and the percentage of reduction obtained during each stage. Initially, the entropy filter achieves a very low level of reduction, about 1.18 %, removing only irrelevant noise from the data. A major level of reduction is obtained during the preselection stage of the MI method, reducing the number of features from 54,031 to 2000, which is about 96.34 %, showing the effectiveness of the relevance filter. The ARS ranking process further reduces the number of features to 800

(98.54 %), concentrating on very important features. Finally, the most aggressive level of reduction is obtained by the TEO optimizer by reducing the number of features, ensuring only the most optimum and discriminative features are retained for better performance.

Table 3. Before and after feature reduction and percentage improvement.

Step	Before	After	Percentage (%)
Entropy filtering	54,675	54,031	1.18
MI preselection	54,031	2000	96.34
ARS ranking	2000	800	98.54
TEO optimizer	800	80	99.85

The Breast cancer gene expression - CuMiDa dataset [38] is a curated microarray gene expression dataset based on the GSE45827 experiment, designed for breast cancer classification and machine learning research. It contains 151 samples, 54,676 gene expression features, and 6 different breast cancer classes, making it suitable for high-dimensional data analysis and predictive modeling. The dataset is part of the CuMiDa (Curated Microarray Database) project, which provides carefully preprocessed and normalized cancer datasets for benchmarking bioinformatics and AI models.

The Leukemia Classification dataset [39] contains 15,135 microscopic blood cell images collected from 118 patients for detecting Acute Lymphoblastic Leukemia (ALL), one of the most common childhood cancers. It includes two labeled classes - normal cells and leukemia blast cells - making it suitable for binary image classification and deep learning tasks. The dataset is widely used for medical imaging research, computer vision, and AI-based cancer diagnosis because the images include real-world staining noise and expert-annotated labels. The Colorectal Cancer Global Dataset [40] contains worldwide colorectal cancer patient data, including demographics, lifestyle risk factors, medical history, cancer stages, treatment methods, survival outcomes, and healthcare costs. It is designed for machine learning tasks such as cancer risk prediction, survival analysis, healthcare analytics, and treatment outcome modeling. The dataset reflects global colorectal cancer trends and provides structured features like smoking history, obesity, genetic mutations, screening history, and mortality rates for predictive analysis.

Table 4 demonstrates that the proposed framework can be scaled for higher-dimensional medical data. All datasets had a feature reduction of over 99 %. Breast cancer gene expression - Curated Microarray Database (CuMiDa) dropped to 80 features out of 54,675, and the high Accuracy 99.54 %, Precision 99.21 %, Recall 99.43 %, and *F1*-Score 99.32 %. Such findings indicate that the framework is an effective way to dimensionally reduce and retain important clinical data and produce robust and reliable classification of a wide range and a large population of medical data.

Table 4. Scalability and performance analysis with diverse datasets.

Dataset	Before	After	Reduction rate (%)	Accuracy (%)	Precision (%)	Recall (%)	*F1*-score (%)
Breast cancer gene expression -CuMiDa	54,675	80	99.85	99.54	99.21	99.43	99.32
RNA-Seq Cancer [38]	20,531	86	99.58	99.49	99.17	99.33	99.26
Leukemia [39]	20,530	54	99.74	99.47	99.13	99.35	99.26
Colon Cancer [40]	2000	19	99.01	99.51	99.17	99.41	99.28

Table 5 summarizes the computational performance of the proposed framework using various classifiers. The most effective is ARS + TEO + SVM it provides quick inference and cost of few computations and alternative classifiers consume more resources and are slower. In general, ARS + TEO + SVM is a lightweight and effective solution to the problem of high-dimensional medical data.

Table 5. Computational performance of the proposed ARS and TEO framework with different classifiers.

Method	Params (M)	GFLOPs	Memory (MB)	FPS	Inference time (ms)
ARS + TEO + SVM	0.012	0.005	34	550	1.82
ARS + TEO + XGBoost	0.52	0.12	124	150	6.67
ARS + TEO + ResNet-based Models	25.6	48.3	427	25	51
ARS + TEO + Random Forest	2.3	1.6	207	90	11.1
ARS + TEO + (FS + SVM)	0.015	0.007	42	505	2.0

4.3. Ablation study

Table 6 presents the ablation study of the proposed framework by evaluating the contribution of MI, ARS, TEO, and EF. The results show that removing any component leads to a decline in performance, confirming the importance of each module. Combined ablation cases further reduce performance, indicating strong interaction among components. The full proposed model achieves the best results, demonstrating the effectiveness of the integrated framework. The EF layer performs minimal feature reduction 1.18 %, removing only irrelevant features while preserving informative data

Table 6. Performance comparison of various configurations.

Configuration	Accuracy	Precision	Recall	*F1*-score	Specificity	MCC	FPR	FNR	AUC	NPV
MI + ARS + TEO (without EF)	99.38	99.02	99.21	99.11	99.45	98.67	0.0055	0.0079	99.67	99.39
EF + ARS + TEO (without MI)	99.12	98.75	98.90	98.82	99.23	98.25	0.0073	0.0110	99.42	99.10
EF + MI + TEO (without ARS)	98.82	98.54	98.65	98.57	99.01	97.83	0.0100	0.0135	99.15	98.78
EF + MI + ARS (without TEO)	99.21	98.87	99.01	98.94	99.37	98.41	0.0063	0.0099	99.54	99.18
MI + TEO (without EF + ARS)	98.46	98.18	98.30	98.24	98.72	97.34	0.0128	0.0168	98.91	98.43

Table 6. (*Continued*).

Configuration	Accuracy	Precision	Recall	*F1*-score	Specificity	MCC	FPR	FNR	AUC	NPV
EF + TEO (without MI + ARS)	98.23	97.96	98.05	98.00	98.51	97.02	0.0149	0.0194	98.67	98.18
EF + MI (without ARS + TEO)	97.91	97.61	97.74	97.67	98.28	96.71	0.0172	0.0225	98.34	97.86
Full Proposed	99.54	99.21	99.43	99.32	99.62	98.95	0.0038	0.0057	99.81	99.58

4.4. Comparative analysis of the proposed model with the existing approaches

This segment compares the performance of existing classification methods with the proposed adaptive feature selection framework using SVM for high-dimensional medical dataset prediction. **Table 7** shows the comparative analysis of various classification methods used with the proposed framework. The existing methods, such as XGBoost, ResNet-based methods, Random Forest, and FS + SVM, provided accuracy between 98.12 % and 98.97 %. The accuracy provided by the suggested SVM model is high, i.e., 99.54 %, while the Precision is 99.21 %, Recall is 99.43 %, and *F1*-Score is 99.32 %, indicating the high effectiveness of the SVM model for accurate classification of reduced medical data without the loss of important information.

Table 7. Performance analysis of the proposed model with the existing models.

References	Method	Accuracy	Precision	Recall	*F1*-Score
[41]	XGBoost	98.82	98.50	98.67	98.58
[42]	ResNet-based Models Gomez-	98.97	98.75	98.81	98.78
[43]	Random Forest	98.61	98.34	98.42	98.38
[44]	FS + SVM	98.12	97.80	97.95	97.87
Proposed	SVM	99.54	99.21	99.43	99.32

4.5. Discussion

The experimental results demonstrate that the proposed adaptive feature selection framework effectively addresses the critical challenge of balancing DR and information retention in high-dimensional medical datasets. Medical datasets, especially the information of Breast cancer gene expression – CuMiDa datasets, have thousands of features and small samples, which may lead to overfitting problems of high computational costs and poor generalization. The proposed framework effectively addresses these issues by combining the use of entropy-based filtering, ARS, and TEO, leading to a better performance of classification and an effective list of features. The results show that the number of features was reduced from 54,675 to 80, which is a 99.85 % reduction rate, significantly lowering the level of computational work and the amount of memory consumed. Despite this aggressive reduction, the classification was exceptionally high, about 99.54 %, the suggested framework managed to retain the most informative features that were required to make sure that the disease was properly predicted. All classifiers were evaluated under the same hardware and software environment to ensure a fair comparison of inference time and

computational resource consumption. The experiments were conducted on a workstation equipped with an Intel Xeon E5-2680 v4 @ 2.40 GHz CPU, 16 GB DDR4 RAM, and an NVIDIA Quadro M2000 GPU, running Microsoft Windows 10 (64-bit). The models were implemented in Python 3.10.0 using Anaconda Distribution 2023.03, with Scikit-learn 1.7.2, NumPy 2.2.6, SciPy 1.15.3, Pandas 2.3.3, and Matplotlib 3.10.9.

This result shows the efficiency of the combination of several feature selection techniques in a step-by-step way to obtain efficiency and reliability.

The Gene Noise Filtering Layer was significant improving the quality of data in the preprocessing preliminary stage. The low-variability and redundant features were eliminated by the entropy-based on the filtering framework without any impact on the overall dataset distribution. Only these significant characteristics were sent to the next selection phases, therefore, minimizing noise and increasing the performance of the classification model. The results of the visualization process proved that the entropy filtering did not distort the intrinsic data structure but improved the quality of features, which indicates that it is appropriate to process high-dimensional medical data.

The MI step selects the features that are strongly related to the disease labels. This step has resulted in reducing the feature space considerably to a smaller number of highly relevant predictors, thereby enhancing the computer resourcefulness and efficiency. The trends in the clustering of the features, which followed the selection, were much focused and informative, which is evidence of the efficiency of the relevance-focused filtering techniques in the high-dimensional classification task. The introduction of the ARS is one of the major contributions of this research. Unlike the usual feature selection techniques, which consider only statistical relevance, the ARS mechanism incorporates the assessment of more than two criteria, which include predictive relevance, stability of selection, and similarity of interaction. This is a multi-dimensional assessment methodology where selected features not only contribute to a positive assessment on their own, but they also contribute to a positive assessment in terms of informing and providing consistency in the evaluation of different data samples. The findings indicate that ARS ranking resulted in a far better discrimination of features and robustness of the model compared to other ranking methods. The other significant aspect of the proposed framework is the TEO algorithm, which combines the benefits of Particle Swarm Optimization, quantum-inspired search, and genetic mutation.

This combination method of optimization enhances the exploration levels in the feature selection stage of the model. Consequently, the model can achieve the best feature subsets at minimal costs. The results of the comparative analysis indicate that the TEO algorithm has better classification accuracy compared to the traditional optimization algorithms, which are GA and PSO. This improvement indicates that the hybrid methods of optimization are effective in feature selection problems in high-dimensional data sets. The results of the classification performance also confirm the effectiveness of the proposed framework. The model recorded an accuracy of 99.54, precision of 99.21, recall of 99.43 and *F1*-score of 99.32, which is good classification performance by all measures. The value of specificity, i.e. 99.62 % and false positive and false negative rates indicate that the model can identify between the disease classes with minimum misidentification. Also, the strong discriminative power of the

proposed model is also supported by the high value of the AUC 99.81 %. All these performance indicators prove the stability and trust worthiness of the medical diagnosis applications based on the framework.

The ablation study is another indication of the significance of every element of the proposed framework. The performance of the classification significantly dropped when either of individual components, including entropy filtering, MI, ARS, and TEO, was excluded. This observation proves that the combination of various methods of feature selection and optimization is fundamental towards attaining optimal performance. The complete proposed model out-performed those with partial configurations always and indicated the significance of a multi-stage and multi-feature selection method. The efficiency of the proposed framework in comparison to other classification frameworks is also demonstrated by analyzing the computational performance of the proposed framework. The computational control of ARS, TEO, and SVM configurations is significantly low in terms of memory usage and time compared to other models like Random Forest and deep learning-based configurations. The framework is proposed in a method that it can be used in real-time clinical environments. In addition, the scalability analysis established the fact that the proposed framework was consistent in its performance using various medical datasets, such as RNA-seq Cancer, Leukaemia, and Colon Cancer datasets. In both situations, the framework had feature reduction rates of more than 99 % with a high classification accuracy. This finding illustrates the ability of the proposed model to generalize and the fact that it may be applicable when handling different medical diagnosis problems. Comprehensively, the results of this study point at the efficiency of the adaptive scoring-based and hybrid optimization approaches to the high-dimensional medical data analysis. The proposed framework not only improves prediction accuracy but also enhances computational efficiency and model stability, making it a valuable tool for modern healthcare analytics and decision support systems.

Although the proposed framework achieves high classification performance and significant feature reduction, some limitations remain. The current approach is mainly evaluated on publicly available breast cancer gene expression datasets, where the sample size is relatively limited compared to real-world clinical scenarios. The optimization process of TEO may also increase computational complexity when applied to extremely large-scale datasets. In addition, the selected gene subsets require further biological validation to confirm their clinical relevance. Furthermore, although the proposed framework demonstrates promising scalability across multiple medical datasets, detailed class-wise classification analyses were conducted only on the Breast Cancer CuMiDa dataset. Future work will extend the experimental evaluation to additional medical datasets by reporting class-wise performance metrics and confusion matrices to provide a more comprehensive assessment of the model's generalization capability. Future research will also focus on extending the framework to larger multi-centre medical datasets, incorporating deep learning-based feature representation methods, and developing adaptive optimization strategies to further improve scalability, interpretability, and real-time clinical applicability.

5. Conclusion

This study presented a feature selection framework to balance DR and information retention in medical datasets. Gene Noise Filtering Layer removed low information features. ARS is the combination of the relevance of the features, stability, and interaction to identify the significant features. TEO is the combination of Swarm Optimization, Quantum Search, and Mutation to optimize the feature sets. The multi-objective optimization functions to ensure the accuracy of the classification using the features. It improved the performance of the classification and reduced the complexity of the computation. It also stored crucial clinical data, which improved the usability of the model. In general, the proposed method is effective, scalable, and applicable to various medical data. The proposed framework is highly generalizable to different medical datasets. This made it a methodology in aiding the correct and trustworthy decision-making in real-world healthcare applications.

Author contributions: Conceptualization, AS and PMS; methodology, AS; software, AS; validation, AS and PMS; formal analysis, AS; investigation, AS; resources, AS; data curation, AS; writing—original draft preparation, AS; writing—review and editing, AS; visualization, AS; supervision, AS; project administration, AS; funding acquisition, PMS. All authors have read and agreed to the published version of the manuscript.

Funding: None.

Ethical approval: Not applicable.

Informed consent statement: Not applicable.

Data availability: The dataset is available in Kaggle repository.

Conflict of interest: The authors declare no conflict of interest.

References

1. Serani A, Diez M. A survey on design-space dimensionality reduction methods for shape optimization. Archives of Computational Methods in Engineering. 2026; 33(2): 1671–1698. doi: 10.1007/s11831-025-10349-x
2. Azarhoosh Z, Ghazaan MI. A review of recent advances in surrogate models for uncertainty quantification of high-dimensional engineering applications. Computer Methods in Applied Mechanics and Engineering. 2025; 433: 117508. doi: 10.1016/j.cma.2024.117508
3. Morabito A, De Simone G, Pastorelli R, et al. Algorithms and tools for data-driven omics integration to achieve multilayer biological insights: a narrative review. Journal of Translational Medicine. 2025; 23(1): 425. doi: 10.1186/s12967-025-06446-x
4. Javanmard Z, Zarean Shahraki S, Safari K, et al. Artificial intelligence in breast cancer survival prediction: A comprehensive systematic review and meta-analysis. Frontiers in Oncology. 2025; 14: 1420328. doi: 10.3389/fonc.2024.1420328
5. Xu W, Wang J, Zhang Y, et al. An optimized decomposition integration framework for carbon price prediction based on multi-factor two-stage feature dimension reduction. Annals of Operations Research. 2025; 345(2): 1229–1266. doi: 10.1007/s10479-022-04858-2
6. Prabaharan G, Sankar SU, Anusuya V, et al. Optimized disease prediction in healthcare systems using HDBN and CAEN framework. MethodsX. 2025; 14: 103338. doi: 10.1016/j.mex.2025.103338
7. Pantanowitz L, Pearce T, Abukhiran I, et al. Nongenerative artificial intelligence in medicine: Advancements and applications in supervised and unsupervised machine learning. Modern Pathology. 2025, 38(3):100680. doi: 10.1016/j.modpat.2024.100680

8. Naghib A, Gharehchopogh FS, Zamanifar A. A comprehensive and systematic literature review on intrusion detection systems in the internet of medical things: current status, challenges, and opportunities. Artificial Intelligence Review. 2025; 58(4): 114. doi: 10.1007/s10462-024-11101-w
9. Katib I, Albassam E, Sharaf SA, et al. Safeguarding IoT consumer devices: Deep learning with TinyML driven real-time anomaly detection for predictive maintenance. Ain Shams Engineering Journal. 2025; 16(2): 103281. doi: 10.1016/j.asej.2025.103281
10. Das S, Sultana M, Bhattacharya S, et al. XAI–reduct: Accuracy preservation despite dimensionality reduction for heart disease classification using explainable AI. The Journal of Supercomputing. 2023; 79(16): 18167–18197. doi: 10.1007/s11227-023-05356-3
11. Gouiaa F, Vomo-Donfack KL, Tran-Dinh A, et al. Novel dimensionality reduction method, Taelcore, enhances lung transplantation risk prediction. Computers in Biology and Medicine. 2024; 169: 107969. doi: 10.1016/j.compbiomed.2024.107969
12. Dhanka S, Sharma A, Kumar A, et al. Advancements in hybrid machine learning models for biomedical disease classification using integration of hyperparameter-tuning and feature selection methodologies: A comprehensive review. Archives of Computational Methods in Engineering. 2026; 33(1): 289–324. doi: 10.1007/s11831-025-10309-5
13. Ambrosino M, Palarea-Albaladejo J, Albanese S, et al. Assessing natural background concentrations of chemical elements in urban soils: A case study in Benevento (Italy). Science of The Total Environment. 2025; 975: 179298. doi: 10.1016/j.scitotenv.2025.179298
14. Ignat A. Experiments with dimensionality reduction on chest x-ray images. Procedia Computer Science. 2023; 225: 3415–3424. doi: 10.1016/j.procs.2023.10.336
15. Chin FY, Lem KH, Wong KM. Improving handwritten digit recognition using hybrid feature selection algorithm. Applied Computing and Informatics. 2026; 22(1–2): 105–114. doi: 10.1108/ACI-02-2022-0054
16. Anuragi A, Sisodia DS, Pachori RB. Mitigating the curse of dimensionality using feature projection techniques on electroencephalography datasets: an empirical review. Artificial Intelligence Review. 2024; 57(3): 75. doi: 10.1007/s10462-024-10711-8
17. Amin R, Yasmin R, Ruhi S, et al. Prediction of chronic liver disease patients using integrated projection based statistical feature extraction with machine learning algorithms. Informatics in Medicine Unlocked. 2023; 36: 101155. doi: 10.1016/j.imu.2022.101155
18. Wang Z, Gao D, Lu Y, et al. A mutual cross-attention fusion network for surface roughness prediction in robotic machining process using internal and external signals. Journal of Manufacturing Systems. 2025; 82: 284–300. 10.1016/j.jmsy.2025.06.018
19. Dhinakaran D, Srinivasan L, Raja SE, et al. Synergistic feature selection and distributed classification framework for high-dimensional medical data analysis. MethodsX. 2025; 14: 103219. doi: 10.1016/j.mex.2025.103219
20. Gildenblat J, Pahnke J. Dimensionality reduction with strong global structure preservation. Pattern Analysis and Applications. 2026; 29(1): 38. doi: 10.1007/s10044-025-01585-9
21. Xiong L, An J, Hou Y, et al. Improved support vector regression recursive feature elimination based on intragroup representative feature sampling (IRFS-SVR-RFE) for processing correlated gas sensor data. Sensors and Actuators B: Chemical. 2024; 419: 136395. doi: 10.1016/j.snb.2024.136395
22. Vinutha MR, Chandrika J, Krishnan B, et al. EPCA—enhanced principal component analysis for medical data dimensionality reduction. SN Computer Science. 2023; 4(3): 243. doi: 10.1007/s42979-023-01677-5
23. Maruotto I, Ciliberti FK, Gargiulo P, et al. Feature selection in healthcare datasets: Towards a generalizable solution. Computers in Biology and Medicine. 2025; 196: 110812. doi: 10.1016/j.compbiomed.2025.110812
24. Khan S, Mazhar T, Naz NS, et al. Advanced Feature Selection Techniques in Medical Imaging—A Systematic Literature Review. Computers, Materials, & Continua. 2025; 85(2): 2347. doi: 10.32604/cmc.2025.066932
25. Kabir MF, Chen T, Ludwig SA. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthcare Analytics. 2023, 3:100125. doi: 10.1016/j.health.2022.100125
26. Sanju P, Ahmed NS, Ramachandran P, et al. Enhancing thyroid disease prediction and comorbidity management through advanced machine learning frameworks. Clinical eHealth. 2025; 8: 7–16. doi: 10.1016/j.ceh.2025.01.002
27. Ashika T, Hannah Grace G. Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: An optimized RST-ML approach. Frontiers in Digital Health. 2025; 7: 1609308. doi: 10.3389/fdgth.2025.1609308

28. Kumar A, Dhanka S, Sharma A, et al. A hybrid framework for heart disease prediction using classical and quantum-inspired machine learning techniques. Scientific reports. 2025; 15(1): 25040. doi: 10.1038/s41598-025-09957-1
29. Razzaque A, Badholia A. PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification. Measurement: Sensors. 2024; 31: 100945. doi: 10.1016/j.measen.2023.100945
30. Jin J, Xu Y, He H, et al. A swin transformer-based hybrid reconstruction discriminative network for image anomaly detection. Scientific Reports. 2025; 15(1): 33929. doi: 10.1038/s41598-025-10303-8
31. Breast Cancer Gene Expression (CuMiDa) dataset: https://www.kaggle.com/datasets/brunogrisci/breast-cancer-gene-expression-cumida (accessed on 10 April 2026).
32. Xu H, Deng Q, Zhang Z, et al. A hybrid differential evolution particle swarm optimization algorithm based on dynamic strategies. Scientific reports. 2025; 15(1): 4518. doi: 10.1038/s41598-024-82648-5
33. Islam MR, Ahmed B, Hossain MA, et al. Mutual information-driven feature reduction for hyperspectral image classification. Sensors. 2023; 23(2): 657. doi: 10.3390/s23020657
34. Hoang VT, Dinh N, Le VT, et al. Detecting Anomalies in FinTech: A Graph Neural Network and Feature Selection Perspective. Computers, Materials, & Continua. 2026; 86(1): 1. doi: 10.32604/cmc.2025.068733
35. Li J, Sheng X, Luo X. Chaos–Quantum Particle Swarm Optimized Kriging for Symmetric Response Modeling and Multi-Objective Marketing Optimization in E-Commerce Systems. Symmetry. 2026; 18(5): 770. doi: 10.3390/sym18050770
36. Serani A, Diez M. A survey on design-space dimensionality reduction methods for shape optimization. Archives of Computational Methods in Engineering. 2026; 33(2): 1671–1698. doi: 10.1007/s11831-025-10349-x
37. Valero-Carreras D, Alcaraz J, Landete M. Comparing two SVM models through different metrics based on the confusion matrix. Computers & Operations Research. 2023; 152: 106131. doi: 10.1016/j.cor.2022.106131
38. CuMiDa dataset: https://www.kaggle.com/datasets/orvile/gene-expression-profiles-of-breast-cancer (accessed on 10 April 2026).
39. Leukemia Classification dataset: https://www.kaggle.com/datasets/andrewmvd/leukemia-classification (accessed on 10 April 2026).
40. Colorectal Cancer Global Dataset: https://www.kaggle.com/datasets/ankushpanday2/colorectal-cancer-global-dataset-and-predictions (accessed on 10 April 2026).
41. Islam MR, Chowdhury AI, Shama S, et al. Enhanced thyroid disease prediction using ensemble machine learning: A high-accuracy approach with feature selection and class balancing. Discover Artificial Intelligence. 2025; 5(1): 9. doi: 10.1007/s44163-025-00225-9
42. Gomez-Martinez V, Chushig-Muzo D, Veierød MB, et al. Ensemble feature selection and tabular data augmentation with generative adversarial networks to enhance cutaneous melanoma identification and interpretability. BioData Mining. 2024; 17(1): 46. doi: 10.1186/s13040-024-00397-7
43. Senyer Yapici I, Uzun Arslan R. Predictive analytics for thyroid cancer recurrence: A feature selection and data balancing approach. The European Physical Journal Special Topics. 2025; 234(15): 4751–4771. doi: 10.1140/epjs/s11734-025-01720-x
44. Reduwan NH, Aziz AA, Mohd Razi R, et al. Application of deep learning and feature selection technique on external root resorption identification on CBCT images. BMC Oral Health. 2024; 24(1): 252. doi: 10.1186/s12903-024-03910-w