TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

ENHANCED FEATURE SELECTION AND CLASSIFICATION OF BREAST CANCER SUBTYPES USING HEURISTIC OPTIMIZATION AND ENSEMBLE MODELS ON MICROARRAY DATA

[1]PREMALATHA KANDHASAMY, [2]MR. WASIM RAJA A, [3]M. VIGNESH, [4]DR.D.MADESWARAN, [5]JANANI S, [6]SIVAKUMAR KANDHASAMY

[1]DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
BANNARI AMMAN INSTITUTE OF TECHNOLOGY, ERODE, TAMIL NADU, INDIA
kpl_barath@yahoo.co.in

[2]ASSISTANT PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
SRI KRISHNA COLLEGE OF ENGINEERING AND TECHNOLOGY, COIMBATORE
wasimrajaa@skcet.ac.in

[3]ASSISTANT PROFESSOR, DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE
KARPAGAM INSTITUTE OF TECHNOLOGY, COIMBATORE - 641 105
TAMIL NADU, INDIA
vignesh.ai@karpagamtech.ac.in

[4]PROFESSOR, DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING,
SSM COLLEGE OF ENGINEERING, KOMARAPALAYAM.
madeshphd@gmail.com

[5]ASSISTANT PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,
KGISL INSTITUTE OF TECHNOLOGY, COIMBATORE,
janani441992@gmail.com

DEPARTMENT OF BIOMEDICAL ENGINEERING,
[6]KARPAGA VINAYAGA COLLEGE OF ENGINEERING AND TECHNOLOGY, CHENGALPATTU,
TAMIL NADU, INDIA
ksivakumar76@gmail.com

Abstract

Feature selection plays a crucial role in analyzing high-dimensional gene expression datasets, such as the GSE45827 breast cancer dataset, which contains numerous genes but a limited number of samples. The presence of irrelevant or redundant genes can negatively impact classification accuracy and biological interpretation. This study enhances classification performance by selecting the most informative genes using three optimization techniques: Self-Organizing Migrating Algorithm (SOMA), Particle Swarm Optimization (PSO), and Stellar Mass Black Hole Optimization (SMBO). To further refine the selected genes, ElasticNet is employed as a second-level feature selection method. The optimized gene subsets are then used in ensemble learning models, including Random Forest, Extreme Randomized Trees (ERT), and XGBoost, for breast cancer classification. Performance is evaluated using accuracy, precision, recall, F1-score, and the kappa constant. Results show that Random Forest achieves 100% accuracy with PSO, 90% with Cuckoo Search, and 97% with SOMA, while ERT reaches 100% accuracy using SMBO. Additionally, differentially expressed genes, pathway analysis, fold change of genes, and Kaplan-Meier survival analysis provide valuable biological insights into breast cancer biomarkers. These findings highlight the importance of feature selection in improving classification accuracy and biomarker discovery, supporting early detection and personalized oncology treatment strategies.

Keywords : Feature Selection, Gene Expression, Breast Cancer, Optimization Techniques, Ensemble Learning, Biomarker Identification

[1] Corresponding Author

1. INTRODUCTION

Gene expression data plays a fundamental role in understanding the molecular mechanisms underlying various diseases, including cancer [1]. It provides insights into gene activity levels across different biological conditions, enabling researchers to identify potential biomarkers for diagnosis, prognosis, and treatment response. Among various cancers, breast cancer remains one of the most prevalent and life-threatening diseases worldwide. Accurate classification of breast cancer subtypes using gene expression data can facilitate early detection, personalized treatment strategies, and improved patient outcomes. However, analyzing gene expression data poses significant challenges due to its high dimensionality, where thousands of genes are measured in a limited number of samples [2]. This creates issues related to computational complexity, overfitting, and reduced model generalization. Hence, feature selection is crucial to identifying the most relevant genes while eliminating redundant and irrelevant ones, thereby enhancing classification accuracy and biological interpretability.

Feature selection serves as a key preprocessing step in machine learning-based classification models, helping to improve performance by reducing the feature space. It enhances computational efficiency, prevents overfitting, and highlights genes that are biologically significant in disease progression. In the case of breast cancer, identifying a subset of genes that are highly informative can contribute to better diagnostic models and facilitate precision medicine [3]. Traditional statistical methods for feature selection, such as differential expression analysis, often fail to capture complex interactions between genes. Thus, advanced optimization techniques are needed to efficiently extract the most informative features from large-scale gene expression datasets.

High-dimensional gene expression datasets, such as the GSE45827 breast cancer dataset [4] contain thousands of genes, but only a small fraction contributes significantly to classification. Many genes are either irrelevant or redundant, negatively impacting classification performance and making it difficult to identify meaningful biomarkers. Conventional feature selection techniques often struggle with large datasets, leading to overfitting and suboptimal model performance. Additionally, most feature selection approaches do not leverage advanced optimization techniques to refine gene selection systematically. There is a need for a robust and efficient feature selection framework that integrates optimization algorithms and machine learning to improve breast cancer classification accuracy.

The presence of metastatic breast cancer in axillary lymph nodes is a critical factor in overall survival [5]. While lymph node status determination is routine, the invasive nature of the procedure and potential biases in node selection may lead to false negatives. Accurately predicting lymph node status using the primary tumor's gene expression profile could reduce the need for axillary lymph node dissection and its associated morbidity. Moreover, identifying patients with negative nodes but a high metastatic potential remains crucial. Further data are needed to refine predictive accuracy, but gene expression profiling holds promise in assessing metastatic risk.

Sotiriou et al (2003)[6] analyzed gene expression patterns from cDNA microarrays in 99 node-negative and node-positive breast cancer patients, correlating them with clinico-pathological characteristics and clinical outcomes. The results showed a strong association with estrogen receptor (ER) status and moderate association with tumor grade. Hierarchical clustering based on ER status identified two distinct tumor groups, correlating with basal and luminal subtypes. Cox regression identified 16 genes linked to relapse-free survival, with glutathione S-transferase M3 emerging as a key survival marker. These findings reinforce the clinical relevance of gene expression profiles for breast cancer prognosis.

Gene expression profiling is advancing the understanding of breast cancer heterogeneity at the genomic level. Validated assays like OncotypeDx and Mammaprint have shown promise in predicting patient outcomes and are being refined through clinical trials [7]. Ongoing research aims to identify host factors influencing prognosis, therapy response, and toxicity. As these analyses become more accessible, rigorous validation will be essential to ensure their clinical relevance and impact on personalized treatment strategies.

Breast cancer incidence has risen significantly in India, particularly among younger women, highlighting the need for molecular insights. A gene expression analysis of 29 tumors and 9 controls identified 2,413 differentially expressed genes, with notable overexpression of COL10A1, COL11A1, MMP1, MMP13, and underexpression of PLIN1, FABP4, and LEP. Key deregulated pathways include cell cycle regulation, metastasis, and lipid metabolism [8]. PAM50 classification confirmed molecular subtypes, and qPCR validation supported microarray findings, notably linking ADAMTS5 downregulation with older patients. This study provides valuable molecular insights into breast cancer in Indian women.

Breast cancer remains the leading cancer among women worldwide, with treatment strategies influenced by factors such as age, menstrual status, genetic variations, and immunological response. Recent advancements in cancer diagnosis have highlighted the importance of gene expression patterns in understanding tumor behavior. By combining gene expression signatures with traditional clinicopathological features, the accuracy of disease prognosis, early diagnosis, and therapy can be enhanced. This review explores the evolution of gene expression signatures for breast cancer, their advantages, future prospects, and provides an overview of available gene expression analysis tools and their specific benefits for breast cancer diagnosis [9].

The e Weighted gene co-expression network analysis WGCNA method [10] for a significant blue gene co-expression module was identified, strongly linked to various breast cancer subtypes. Additionally, eight key hub genes—CCNE1, CENPN, CHEK1, PLK1, DSCC1, FAM64A, UBE2C, and UBE2T—were recognized and validated as potential prognostic and therapeutic biomarkers for breast cancer.

A two-step transfer learning pipeline introduced for predicting breast cancer molecular subtypes using unannotated pathological images [11]. Pretrained models, including VGG16, ResNet50, ResNet101, and Xception, were trained on in-house and TCGA-BRCA datasets, with ResNet101 achieving the highest accuracy of 0.78 and slide-wise prediction accuracy of 0.913. The models outperformed the Genefu tool, demonstrating their potential for cost-effective and accurate breast cancer classification without region annotation, enhancing clinical applicability.

Differentially expressed genes (DEGs) involved in breast cancer progression and metastasis are identified with highlighting their interaction with target proteins and signaling pathways [12]. By analyzing 50 DEGs, 8 potential gene signatures were shortlisted, offering insights into their role in breast cancer. Pathway and miRNA target analyses revealed critical targets for therapeutic intervention, with several FDA-approved drug options identified. These findings emphasize the importance of these genes and their networks in understanding breast cancer and improving treatment strategies.

a breast cancer detection system [13] was proposed using a metaheuristic optimization algorithm inspired by humpback whale bubble-net hunting. The algorithm selects and weighs key features from breast cytology images to optimize a support vector machine classifier. The model achieved 98.82% accuracy, outperforming genetic algorithm and particle swarm optimization in feature selection and classification speed. The results demonstrate the system's strong potential for reliable automatic breast cancer detection.

The effects of glucose starvation (GS), metformin (MET), and 2-DG on MDA-MB-231 and MCF-7 breast cancer cells using gene expression profiling. MDA-MB-231 cells showed higher sensitivity to glucose deprivation, with apoptosis, DNA replication inhibition, and oxidative stress responses observed across treatments [14]. GS had the most significant impact, inducing autophagy and DNA methylation inhibition. Combining MET and GS could enhance cancer cell susceptibility to conventional chemotherapy.

Artificial neural networks (ANNs) are increasingly used for effective cancer detection, addressing the rising global risk of cancer. Accurate detection remains challenging due to limited data availability. While various cancer classification methods exist, improving classification accuracy remains a priority. This research proposes a two-step feature selection (FS) technique combined with a 15-neuron neural network (NN) to enhance cancer classification. The FS method reduces feature attributes, while the 15-neuron NN classifies the cancer with high accuracy. Using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, the proposed method achieves up to 99.4% classification accuracy, significantly outperforming existing technique [15].

Breast cancer remains the most prevalent cancer among women, with gene expression microarray studies playing a key role in classification and prognosis [16]. However, these datasets face challenges such as high-dimensionality, limited samples, and class imbalance, complicating feature selection. Various accurate feature selection techniques, particularly hybrid bio-inspired metaheuristic and wrapper methods, have been widely applied. This review explores the latest advancements in these hybrid methods for improving cancer classification accuracy.

Tumor heterogeneity and unclear metastasis mechanisms hinder effective targeted therapies for Triple-negative breast cancer (TNBC), which is characterized by high mortality and frequent distant metastasis. This study analyzed gene expression datasets from TNBC and Non-TNBC breast cancer (BrCa) cases, focusing on cancer driver genes. Recursive Feature Elimination (RFE) identified gene signatures differentiating TNBC, and machine learning models revealed XGBoost as the best-performing algorithm for a 25-gene subset. Among 34 differentially regulated genes, four were found to be potential prognostic biomarkers, with two novel genes (POU2AF1 and S100B) showing promise. Further pathway analysis highlighted key signaling pathways like MAPK, PI3-AkT, Wnt, and TGF-β, crucial in metastasis [17].

Through PPI network analysis, seven top-ranked upregulated key genes (BUB1, ASPM, TTK, CCNA2, CENPF, RFC4, and CCNB1) are identified associated with breast cancer from 127 differentially expressed genes (DEGs) between breast cancer and control samples [18]. This review highlighted metaheuristic-based hybrid approaches,

combining filter and wrapper methods for improved gene selection. The genetic algorithm (GA) is the most widely used wrapper method, while SVM is the most frequently applied classifier. Recent studies emphasize the Whale Optimization Algorithm (WOA) for feature selection, with future research expected to focus on hybrid WOA-GA methods for enhancing breast cancer classification accuracy

Gene expression profiles from the TCGA database utilized to identify differentially expressed genes (DEGs) in breast cancer using the edgeR R package [19]. Functional and pathway enrichment analysis revealed 28 significantly enriched pathways, which were downloaded from KEGG to construct a gene interaction network. From this, 154 key genes were identified, 23 of which overlapped with DEGs and were considered potential diagnostic markers. An SVM classification model built using these markers demonstrated strong predictive performance (AUC = 0.960 in training, 0.907 in validation). These findings provide valuable insights for breast cancer diagnosis and treatment research.

Breast cancer remains a global health concern, necessitating reliable biomarkers for early detection and treatment. Machine learning is integrated to analyze gene expression data, identifying key predictive genes such as TOP2A, AKR1C3, and EZH2 using the BGWO_SA_Ens algorithm [20]. From over 10,000 genes, 1404 were selected in the merged dataset (F1: 0.981, PR-AUC: 0.998, ROC-AUC: 0.995) and 1710 in GSE45827 (F1: 0.965, PR-AUC: 0.986, ROC-AUC: 0.972). Pathway enrichment revealed involvement in AMPK, Adipocytokine, and PPAR signaling. Drug-gene interaction analysis highlighted potential therapeutic targets, emphasizing computational approaches in biomarker discovery and precision oncology.

The XGBoost model was demonstrated strong multi-class classification performance on breast cancer gene expression data, achieving an overall F1-score of 78% [21]. Notably, Class 5 showed perfect precision, recall, and F1-score, while other classes exhibited varying degrees of precision and recall, indicating areas for improvement.

This study aims to address the challenge of high-dimensional gene expression data by leveraging advanced feature selection techniques. By integrating metaheuristic optimization algorithms with ElasticNet and ensemble learning, the study seeks to improve classification accuracy while ensuring biological relevance. The rationale behind using SOMA, PSO, and SMBO is their effectiveness in global optimization and feature reduction, enabling the extraction of highly relevant genes. ElasticNet is further incorporated as a secondary feature selection mechanism to enhance robustness. The selected features are then employed in ensemble learning models, which are known for their superior predictive capabilities in complex datasets.

This research aims to enhance breast cancer classification accuracy by identifying the most significant genes from the GSE45827 gene expression dataset. To achieve this, three optimization-based feature selection techniques—Self-Organizing Migrating Algorithm (SOMA), Particle Swarm Optimization (PSO), and Stellar Mass Black Hole Optimization (SMBO)—will be implemented and compared. The selected features will then be refined using ElasticNet to improve classification performance by eliminating redundant genes. Additionally, the study will evaluate the effectiveness of ensemble learning models, including Random Forest, Extreme Randomized Trees (ERT), and XGBoost, in classifying breast cancer samples. The proposed approach will be assessed using key performance metrics such as accuracy, precision, recall, F1-score, and the kappa constant. Furthermore, this research seeks to identify critical biomarkers that play a significant role in breast cancer classification, contributing to early detection and potential targeted treatment strategies.

Research Contributions

1. Development of an optimized feature selection framework using SOMA, PSO, and SMBO for gene selection in high-dimensional breast cancer data.
2. Implementation of ElasticNet as a secondary refinement method to enhance the interpretability and reliability of selected genes.
3. Performance evaluation of ensemble learning classifiers, demonstrating their efficacy in breast cancer classification.
4. Identification of potential gene biomarkers that could contribute to breast cancer diagnosis and treatment.

2 METHODS AND MATERIALS

The GSE45827 dataset is preprocessed using label encoding, KNN imputation for missing values (Gustavo Enrique Batista & Maria-Carolina Monard, 2002) and Min-Max normalization (Peshawa J. &Muhammad Ali, 2022) for feature scaling. Feature selection employs SOMA, PSO, Cuckoo Search, and SMBO, with ElasticNet refining selected genes by removing redundancies. Ensemble models—Random Forest [22], Extremely

Randomized Trees (ERT) [23], and XGBoost [24] are used for classification. Performance is assessed using accuracy, precision, recall, F1-score, and the kappa constant. Additionally, differentially expressed genes identified through ElasticNet are analyzed for biological significance, with pathway identification providing insights into their role in breast cancer progression, aiding early diagnosis and targeted treatment strategies.

2.1 Preprocessing

2.1.1 Replacing the Missing Values

K-Nearest Neighbors (KNN) Imputer is a method used to handle missing values by estimating them based on the nearest available data points. It calculates missing values by identifying the 'k' most similar instances using distance metrics like Euclidean distance. It imputes the missing data with the average or majority value from these neighbors. This approach maintains data consistency and captures relationships between features effectively.

2.1.2 Convert Categorical Values into Numerical Values

Label encoding is a technique used to convert categorical data into numerical values. Each unique category is assigned a distinct integer, making it suitable for machine learning models. For example, a feature with values ["normal," "basal," "cell-line"] would be encoded as [0, 1, 2]. This method is efficient but may introduce an unintended ordinal relationship among categories.

2.1.3 Normalization

Min-Max normalization is a technique used to scale numerical data within a specific range, typically between 0 and 1. It transforms values while preserving relationships between data points. The formula is:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where X is the original value, X_{min} is the minimum value in the dataset, and X_{max} is the maximum value in the dataset. This method ensures all values are scaled within [0,1] making them suitable for machine learning algorithms.

2.2 Heuristic Optimization Techniques

2.2.1 Self-Organizing Migrating Algorithm

The Self-Organizing Migrating Algorithm (SOMA) is a population-based optimization technique inspired by the social behavior of migrating individuals in a group [25]. In this method, a group of agents explores the search space collectively, with the agent having the best fitness value designated as the leader. Each agent migrates toward the leader iteratively, updated using a formula that incorporates the difference between its position and the leader's, scaled by a step size and a random factor. This stochastic nature ensures diversity in exploration and reduces the chances of getting trapped in local minima. SOMA is computationally efficient and suitable for solving complex, multimodal optimization problems.

Mathematical Formulation

1. Initialization: Define the population P with N individuals X_i in the D-dimensional search space:
$$P = \{X_1, X_2, X_3, \dots X_N\}$$
$$X_i = \{x_{i1}, x_{i2}, x_{i3}, \dots, x_{iD}\}$$
where x_{ij} is the j-th dimension of the i-th indivdual
2. Fitness Evaluation: Compute the fitness of each individual based on the objective function $f(X_i)$. Select the leader X_{leader} with the best fitness:
$X_{leader} = \underset{i}{argmin}\ f(X_i)$, for minimization problems
3. Migration Step: For each individual X_i (excluding the leader), update its position iteratively:
$$X_i^{(k+1)} = X_i^k + Step.rand.\left(X_{leader} - X_i^k\right)$$

where X_i^k is the current position of the k-th agent, Step is a migration parameter controlling the movement toward the leader, $rand$ is a random number $\in$[0,1]ensuring stochastic behaviour and $X_{leader} - X_i^k$ is the direction vector pointing toward the leader

4. Termination: Repeat the migration process for all agents over a predefined number of iterations or until a stopping criterion (e.g., fitness threshold or maximum iterations) is met.

2.2.2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a population-based optimization algorithm inspired by the social behavior of birds flocking or fish schooling. It was introduced by Kennedy and Eberhart in 1995 [26] and is widely used

for solving continuous and discrete optimization problems. PSO works by iteratively improving a population of candidate solutions, called particles, by moving them through the search space based on their own experiences and those of their neighbors.

1. Initialization: A swarm of N particles is initialized with random positions X_i and velocities V_i in the search space. Each particle represents a potential solution.
2. Fitness Evaluation: The fitness of each particle is calculated using the objective function $f(X_i)$. and the best position visited by each particle (personal best, p_i) is recorded. The global best position p_g, which is the best among all particles, is also identified.
3. Velocity Update: The velocity of each particle is updated using the formula:
   $V_i^{(k+1)} = \omega V_i^k + c_1.r_1.(p_i - X_i^k) + c_2.r_2.(p_g - X_i^k)$

where ω Inertia weight controlling exploration vs. exploitation. c_1, c_2 are the acceleration coefficients controlling the influence of personal and global bests. and r_1 & r_2 are random numbers in [0,1]

Position Update: The position of each particle is updated based on its new velocity:

$$X_i^{(k+1)} = X_i^k + V_i^{(k+1)}$$

Stopping Criteria: The algorithm repeats the velocity and position updates until a stopping condition, such as a maximum number of iterations or acceptable fitness value, is met.

2.2.3 Cuckoo Search

The Cuckoo Search is a metaheuristic optimization technique inspired by the behavior of some cuckoo species in laying their eggs in the nests of other host birds. It was developed by Xin-She Yang and Suash Deb in 2009 [27] . The algorithm combines concepts of Levy flights and the brood parasitism behavior of cuckoos.

Key Features of Cuckoo Search

1. Cuckoo Behavior:
   - Cuckoos lay their eggs in other birds' nests.
   - Some host birds discover these foreign eggs and may throw them out or abandon the nest.
2. Levy Flights:
   - Cuckoos search for new nests using Levy flights, a type of random walk with steps drawn from a Levy distribution.
   - Levy flights ensure global exploration of the search space.
3. Host-Parasite Interaction:
   - If a host detects a cuckoo egg, the nest may be abandoned, forcing the cuckoo to search for another.

Steps in the Cuckoo Search Algorithm

1. Initialization:
   - Generate an initial population of nests (solutions).
   - Define a fitness function to evaluate solutions.
2. Generate New Solutions:
   - Use Levy flights to generate new candidate solutions.
   - Replace an existing solution if the new one is better.
3. Discovery Probability:
   - A fraction of the worst solutions (nests) are replaced with new random solutions based on a "discovery probability" (pa).
4. Iterate:
   - Repeat the steps of generating new solutions, evaluating fitness, and replacing poor solutions until the stopping criterion (e.g., max iterations) is met.

Cuckoo Search Pseudocode

1. Initialize:
   - Number of nests (n), maximum iterations (max_iter), and discovery probability (pa).
2. Evaluate Fitness:
   - Assess each solution in the population.
3. Main Loop:
   - For each cuckoo:
     - Generate a new solution using Levy flights.
     - Compare it with a randomly selected solution.
     - Replace the solution if the new one is better.

- Abandon a fraction (pa) of the worst nests and generate new ones.
- Keep the best solutions.

4. Return the Best Solution.

Levy Flight Formula

Levy flight steps are calculated as:

$$x_{i+1} = x_i + \alpha . Levy(\beta)$$

x_i is the current position, α is the step size scaling factor and β is the Levy distribution parameter ($1 < \beta \leq 2$)

$$Levy \sim u = \frac{s}{\gamma^{\frac{1}{\beta}}}$$

where s and γ are random variables from normal distributions.

2.2.4 Stellar-Mass Black Hole Optimization

The Stellar-Mass Black Hole Optimization Algorithm (SMBO) is a novel optimization technique inspired by the gravitational dynamics of stellar-mass black holes. It was introduced by Premalatha and Balamurugan, 2015 [28], and is based on the astrophysical phenomena associated with black holes' gravitational attraction and their interaction with surrounding stars.

Key Concepts of Stellar-Mass Black Hole Optimization

1. Stellar-Mass Black Holes:
   - Represented as the best solutions in the search space.
   - These black holes exert a gravitational pull on stars (candidate solutions) around them.
2. Stars:
   - Represent potential solutions to the optimization problem.
   - They move toward black holes (best solutions) based on gravitational forces.
3. Gravitational Force:
   - Determines the influence of black holes on nearby stars, guiding the search process toward optimal solutions.
   - A stronger gravitational force leads to more significant updates in positions.
4. Event Horizon:
   - Represents the boundary around the black hole where solutions are absorbed if they fall within it.
   - Ensures convergence by pulling stars (solutions) close to the black hole.

Steps in the SMBO Algorithm

1. Initialization:
   - Randomly initialize a population of stars (solutions) in the search space.
   - Evaluate their fitness based on the objective function.
2. Gravitational Attraction:
   - Calculate the gravitational force exerted by black holes (best solutions) on the stars.
   - Update the positions of stars based on this force.
3. Update Black Holes:
   - Identify the best-performing stars as black holes for the next iteration.
   - Adjust the gravitational pull and event horizon radius to focus the search.
4. Event Horizon Check:
   - If a star enters the event horizon of a black hole, it is absorbed, and its position is updated.
5. Convergence:
   - Repeat the process until a stopping criterion is met (e.g., maximum iterations or desired fitness).

Pseudocode of SMBO

1. Initialize:
   - Define the population size, number of black holes, and other parameters.
   - Randomly generate a population of stars (solutions).
2. Fitness Evaluation:
   - Evaluate the fitness of all stars using the objective function.
3. Update:
   - Identify the best solutions as black holes.
   - Calculate gravitational force and update the positions of stars.
   - Check if stars enter the event horizon; if yes, reposition them.

4. Iterate:
    - Continue the process until convergence.
5. Output:
    - Return the best solution (global optimal).

2.3 Ensemble Learning Models

2.3.1 Extreme Randomized Trees (Extra Trees)

Extreme Randomized Trees (Extra Trees) is an ensemble learning method that extends the Random Forest algorithm. Unlike Random Forest, which selects optimal split points based on criteria like Gini impurity or entropy, Extra Trees randomly selects each feature. This approach increases model diversity and reduces variance, making it robust against overfitting. However, it may sacrifice some accuracy compared to Random Forest in certain datasets.

2.3.2 Random Forest

Random Forest is a popular ensemble learning algorithm that builds multiple decision trees and aggregates their outputs for classification or regression tasks. It introduces randomness by selecting random subsets of features and bootstrapping data samples for each tree. This reduces overfitting and improves generalization. The final prediction is determined by majority voting (for classification) or averaging (for regression). Random Forest is widely used due to its high accuracy, stability, and ability to handle missing values and feature importance ranking.

2.3.3 XGBoost (Extreme Gradient Boosting)

XGBoost is a powerful gradient-boosting framework optimized for speed and performance. It builds decision trees sequentially, where each new tree corrects errors made by the previous ones. XGBoost uses regularization techniques such as L1 and L2 penalties to prevent overfitting. Additionally, it incorporates parallelization, handling of missing values, and tree pruning to enhance efficiency. Due to its superior predictive accuracy and scalability, XGBoost is widely applied in machine learning competitions and real-world applications like finance, healthcare, and recommendation systems.

2.4 Performance Measures in Machine Learning

A confusion matrix is a performance evaluation tool for classification models, summarizing predictions into four categories: True Positives (TP), where the model correctly predicts positive cases; True Negatives (TN), where it correctly predicts negative cases; False Positives (FP), also known as Type I errors, where negative cases are incorrectly classified as positive; and False Negatives (FN), or Type II errors, where positive cases are wrongly classified as negative. It helps in calculating key metrics like accuracy, precision, recall, and F1-score, providing deeper insights into model effectiveness, especially in imbalanced datasets**.**

Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances in a dataset. It is a useful metric when the dataset is balanced but may be misleading in imbalanced datasets.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent classification outcomes.

Precision

Precision measures the proportion of correctly predicted positive instances out of all instances classified as positive. It is crucial when false positives need to be minimized, such as in medical diagnoses.

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

Recall indicates how well the model identifies actual positive instances. It is important to minimise false negatives, such as in disease detection.

$$Recall = \frac{TP}{TP + FN}$$

F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives are important.

$$F1 - SCORE = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

It is particularly useful in cases of imbalanced datasets.

Kappa Constant (Cohen's Kappa)

Cohen's Kappa evaluates the agreement between predicted and actual classifications while accounting for chance agreement. It is defined as:

$$\kappa = \frac{\text{observed agreement} - \text{expected agreement by chance.}}{1 - \text{expected agreement by chance.}}$$

A kappa value close to 1 indicates strong agreement, while a value near 0 suggests poor agreement.

2.5 Biological Insights

2.5.1 ElasticNet

ElasticNet [29] is an advanced regularization method that effectively combines the strengths of Lasso (L1) and Ridge (L2) regression, providing a robust solution for high-dimensional predictive modeling. ElasticNet regularization is an extension of ordinary least squares (OLS) regression that incorporates both L1 (Lasso) and L2 (Ridge) penalties to optimize feature selection and model stability.

$$Loss = RSS + \alpha\left(\lambda_1 \sum_{i=1}^{p} |\beta_i| + \frac{\lambda_2}{2}\sum_{i=1}^{p} \beta_i^2\right)$$

RSS is the residual sum of squares. α controls the overall strength of regularization, λ_1 and λ_2 are hyperparameters that control the contribution of Lasso and Ridge regularization respectively. βj represents the coefficients of the features.

- Lasso (L1) Regularization: Encourages sparsity, meaning it tends to drive the coefficients of less important features to zero.
- Ridge (L2) Regularization: Shrinks the coefficients but doesn't push them to zero, making it useful for handling collinearity among features.

By adjusting the combination of Lasso and Ridge, ElasticNet can be tailored to suit datasets with either a large number of features or highly correlated predictors.

2.5.2 Differentially Expressed Genes (DEGs)

DEGs are identified by analyzing variations in gene expression across different conditions, such as disease versus control groups. The process begins with preprocessing, where raw gene expression data undergo normalization using techniques like log transformation, Z-score scaling, or TPM normalization to ensure comparability. Next, filtering is performed to select genes with significant biological relevance or high variance. Statistical testing methods, including t-tests, ANOVA, or non-parametric tests, help determine significant expression differences, with advanced tools such as limma, DESeq2, and edgeR commonly employed [30]. To reduce false discoveries, multiple testing correction methods like Benjamini-Hochberg FDR correction are applied. Finally, genes are classified as DEGs based on predefined thresholds, such as an adjusted p-value (e.g., FDR < 0.05) and a fold change cutoff (e.g., $\log_2$FC > ±1). Once identified, these DEGs undergo functional enrichment analysis, including Gene Ontology (GO) and KEGG pathway analysis, to explore their biological significance and potential roles in disease mechanisms.

1. Group Samples: Separates normal and cancer sample columns.
2. Log2 Fold Change (LFC): Computes expression change between cancer and normal samples.
3. t-Test: Checks if the difference is statistically significant.
4. FDR Correction: Adjusts p-values using Benjamini-Hochberg to control false positives.
5. Filter DEGs: Selects genes with |Log2 Fold Change| > 1 and adjusted p-value < 0.05.

2.5.3 Pathway identification

Pathway identification [31] plays a vital role in gene expression analysis by revealing the biological processes, molecular functions, and signaling networks associated with differentially expressed genes (DEGs). This process involves mapping genes to established biological pathways to determine their functional significance in cellular activities. Functional enrichment analysis, conducted using resources like Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG), classifies genes based on their roles in biological processes, molecular functions, and cellular components. Additionally, curated pathway databases such as Reactome, BioCarta, and WikiPathways offer comprehensive insights into gene interactions. A widely used computational approach, Gene Set Enrichment Analysis (GSEA), assesses whether a predefined gene set is significantly enriched within a specific pathway. Identifying these pathways aids in understanding disease mechanisms, discovering potential drug targets, and recognizing biomarkers, ultimately providing deeper insights into complex biological interactions.

2.5.4 Biomarker Gene Identification

Biomarker gene identification is a critical process in biomedical research for detecting genes associated with disease diagnosis, prognosis, and treatment response. It involves analyzing gene expression patterns to distinguish potential biomarkers that serve as indicators of normal or pathological conditions. The process begins with data preprocessing, including normalization and filtering of gene expression data to remove noise. Differential gene expression analysis is then performed using statistical methods like limma, DESeq2, or edgeR to identify significantly altered genes. Feature selection techniques, such as machine learning algorithms (random forest, support vector machines, or deep learning), are often used to refine biomarker candidates. Functional enrichment analysis through Gene Ontology (GO) and KEGG pathway databases helps in understanding the biological relevance of identified genes. Finally, validation is conducted through independent datasets, experimental techniques like qRT-PCR, and clinical studies to confirm the biomarker's reliability. Accurate biomarker identification facilitates early disease detection, personalized medicine, and targeted therapies. Figure 1 illustrates the block diagram of the proposed methodology.

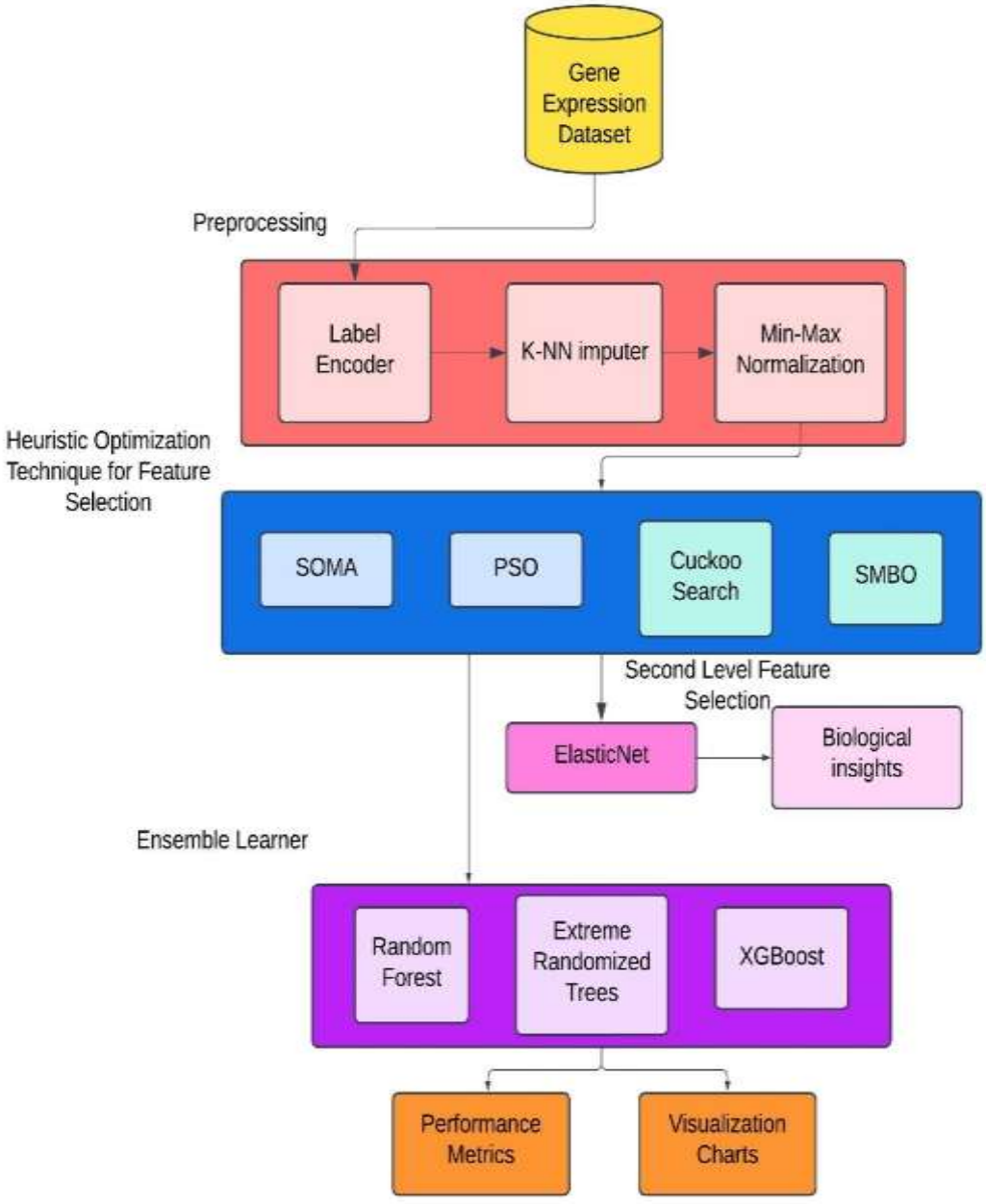


Figure 1 Block diagram of the proposed methodology

3. RESULTS AND DISCUSSIONS

Table 1 provides the description of the GSE-45827 dataset. Table 2 illustrates the parameters and their corresponding values. Table 3 lists the number of features selected by heuristic optimization techniques.

Table 1 Dataset Description

S.No.	Dataset	Number of Genes	Number of Samples	Reference
1	GSE-45827	16384	41 – Basal 30 - HER2, 29 - Luminal A 30 - Luminal B 14 - cell lines 11 - normal 44 BC samples and a subset of 11 normal samples	[4]

Table 2 Parameters and their corresponding values for the optimization techniques

Model	Parameter	Value
	Candidate Solution Representation	A binary vector is used, where '1' indicates the presence of a gene, and '0' denotes its absence in the feature selection process.
	Population Size	20
	Number of iterations	20
SOMA	Objective Function	1 - Accuracy of SVM
	Validation	Cross validation = 5
	step	0.11
	prt	0.3
	path_length	3
PSO	Objective Function	Accuracy of SVM
	Validation	Cross validation = 5
	c_1	1.5
	c_2	1.5
	ω	0.9
Cuckoo Search	Objective Function	1 - Accuracy of SVM
	Validation	Cross validation = 5
	β	1.5
	γ	0.01
	pa	0.25
	α	1
SMBO	Objective Function	1 - Accuracy of SVM

	Validation	Cross validation = 5
	β	0.5
	α	0.5

Table 3 :Number of features selected and the corresponding best fitness values

Model	Number of Significant features selected by the model	Best Fitness (1 - Accuracy)
SOMA	8234	0.046
PSO	683	0
Cuckoo Search	8135	0.046
SMBO	8222	0.034

Figure 2 illustrates the distribution of classes within the dataset.

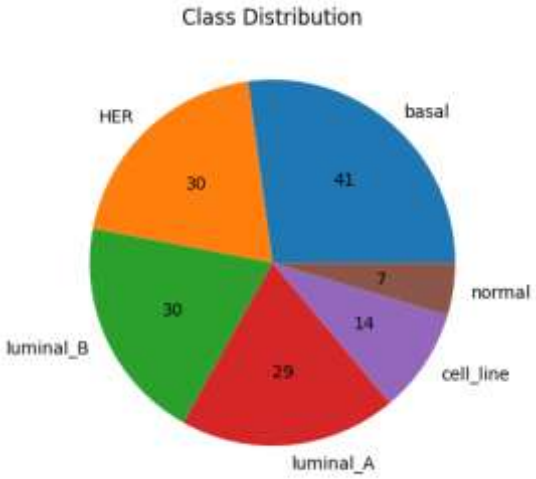


Figure 2 Class Distribution

3.1 SOMA

Figure 3 represents the fitness value obtained by SOMA. The graph begins with a relatively high fitness value of approximately 0.053, indicating that the initial feature subsets selected by SOMA resulted in lower accuracy, as fitness is inversely related to accuracy. A sharp decline in the fitness value occurs during the initial iterations, dropping from 0.053 to around 0.046, suggesting that SOMA quickly identifies better feature subsets early in the optimization process. After this rapid improvement, the fitness value stabilizes at approximately 0.0462, indicating that the algorithm has converged to a good solution and further iterations are unlikely to yield significant gains. The final fitness value of 0.0462 corresponds to an accuracy of around 95.38%. This demonstrates that SOMA effectively selected features that led to a high classification accuracy. The algorithm appears to have reached convergence, implying that extending the iterations further may not result in substantial improvements. The features selected during the process are crucial for the classification task, though further analysis would be needed to pinpoint which specific features were chosen.

TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

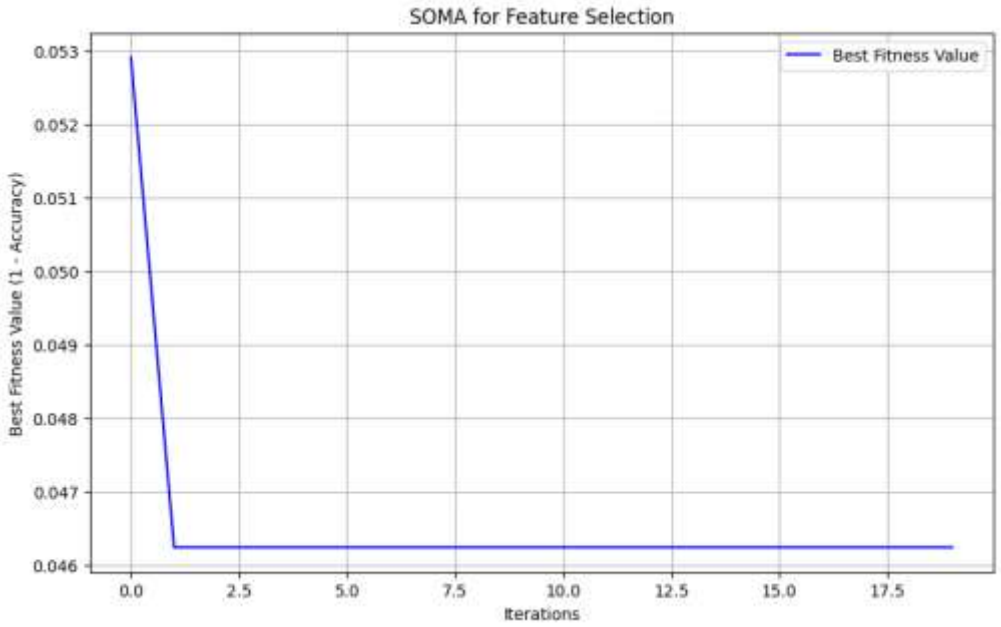
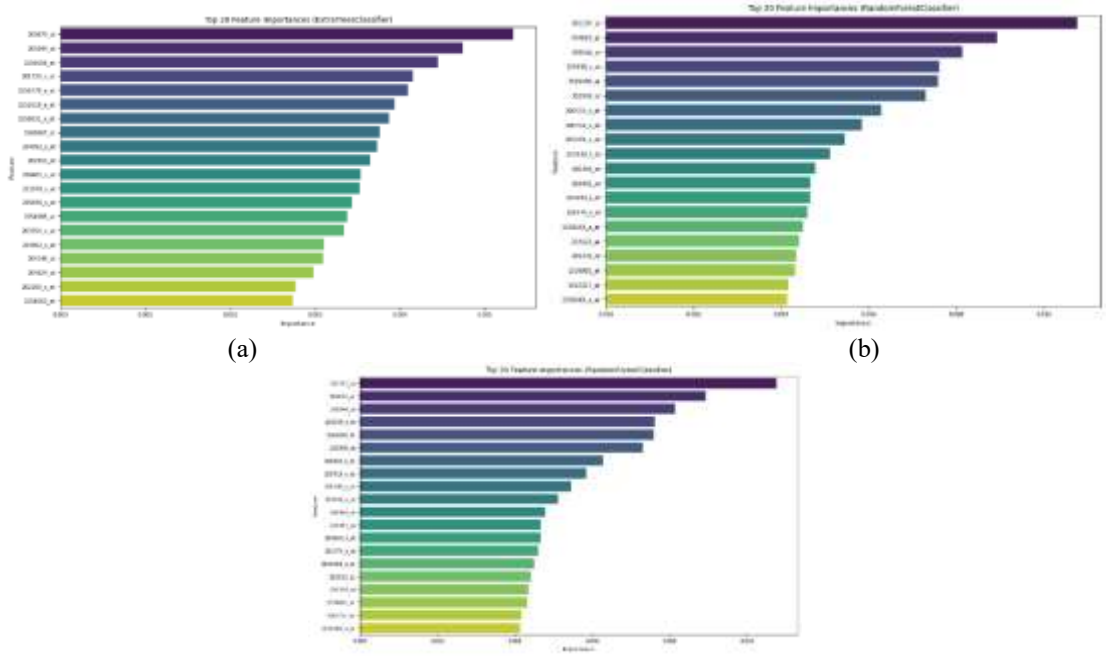


Figure 3 Fitness values obtained from SOMA

Figures 4(a), 4(b), and 4(c) showcase the top 20 features selected by Extremely Randomized Trees (ERT), Random Forest, and XGBoost, respectively. These features are chosen based on their importance in classifying the dataset and significantly contribute to model accuracy. The purpose of displaying these figures is to highlight how each model selects relevant features for prediction tasks. By comparing the feature selection methods, it becomes evident which features are consistently prioritized, reflecting their strong correlation with the target variable. This information is crucial for understanding the models' decision-making process and ensuring the interpretability of the results. Additionally, it provides insights into the most influential factors in the dataset that drive predictions.



(a) (b)

(c)

Figure 4 Top 20 Features Selected by Extremely Randomized Trees with SOMA features

Figure 5 presents the performance metrics of ERT, Random Forest, and XGBoost using the features selected by SOMA. In most metrics, Random Forest slightly outperforms both ERT and XGBoost. For accuracy, Random Forest and XGBoost both achieve 0.87, while ERT achieves 0.90, reflecting their strong classification capabilities.

Precision and Recall are high for all models, with Random Forest reaching 0.97, XGBoost 0.87, and ERT 0.90, showing that their predictions are generally reliable. The F1-scores align with the accuracy results, demonstrating a well-balanced performance in precision and recall. Both Random Forest and XGBoost attain scores of 0.96 and 0.87, respectively, while ERT scores 0.90. Lastly, the Kappa values for all models are high, indicating a strong level of agreement between predicted and actual classifications, with Random Forest at 0.96, ERT at 0.88, and XGBoost at 0.84.

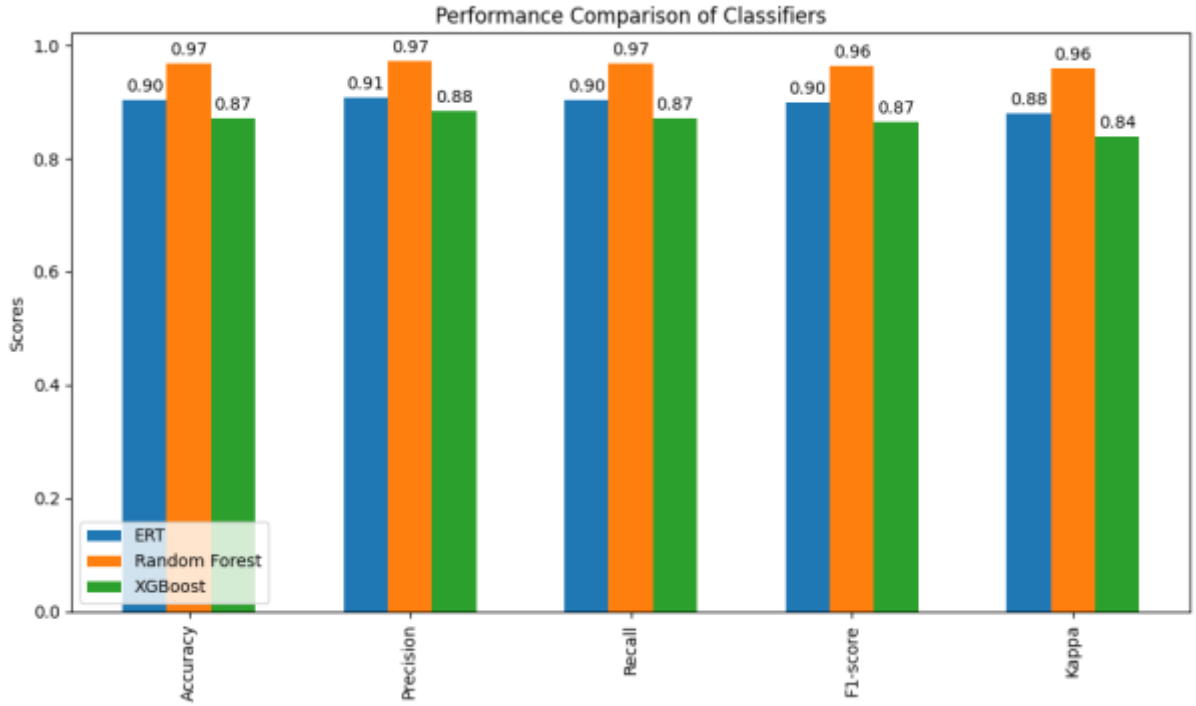


Figure 5 Performance metrics

3.2 PSO

Figure 6 illustrates the fitness function values achieved using PSO. The graph begins with relatively low fitness values, around 0.935, indicating that the initial random solutions explored by PSO resulted in lower accuracy. In the early iterations, particularly between the second and fifth iterations, there is a significant improvement in fitness, suggesting that PSO quickly identified more promising solutions by efficiently navigating the search space. Following this rapid enhancement, the fitness value reaches a plateau at 1.00 and remains constant for the remaining iterations, signifying that PSO has likely converged to an optimal or near-optimal solution with no further improvements. The final fitness value of 1.00 indicates that the algorithm achieved 100% accuracy in classification. This demonstrates PSO's effectiveness in optimizing feature selection and improving accuracy. The algorithm's ability to stabilize quickly suggests that further iterations would not yield additional benefits, confirming its reliability in identifying an optimal solution.

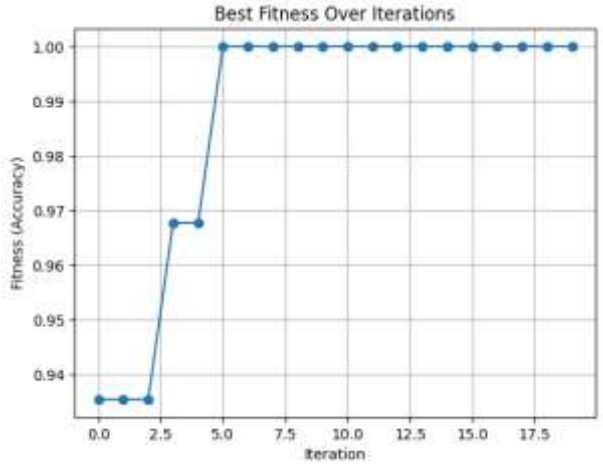
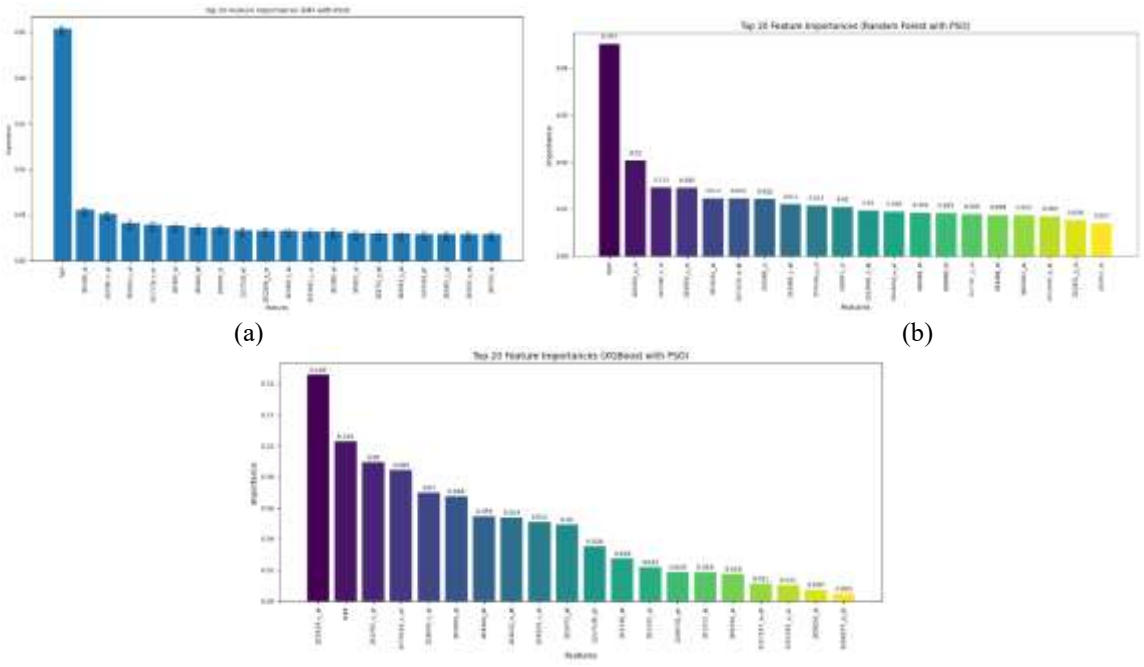


Figure 6 Fitness function values achieved using PSO

Figures 7(a), 7(b), and 7(c) showcase the top 20 features selected by ERT, Random Forest, and XGBoost, respectively.



(a) (b)

(c)

Figure 7 Top 20 Features Selected by Extremely Randomized Trees with PSO features

Figure 8 compares the performance of three machine learning classifiers—Extremely Randomized Trees (ERT), Random Forest, and XGBoost—across various evaluation metrics. The analysis reveals that all three models achieve near-perfect performance across all metrics, highlighting their effectiveness in classification. However, Random Forest outperforms ERT and XGBoost in most cases. In terms of accuracy, Random Forest attain a perfect score of 1.0, while ERT and XGBoost reached 0.97, demonstrating their ability to correctly classify the majority of instances.

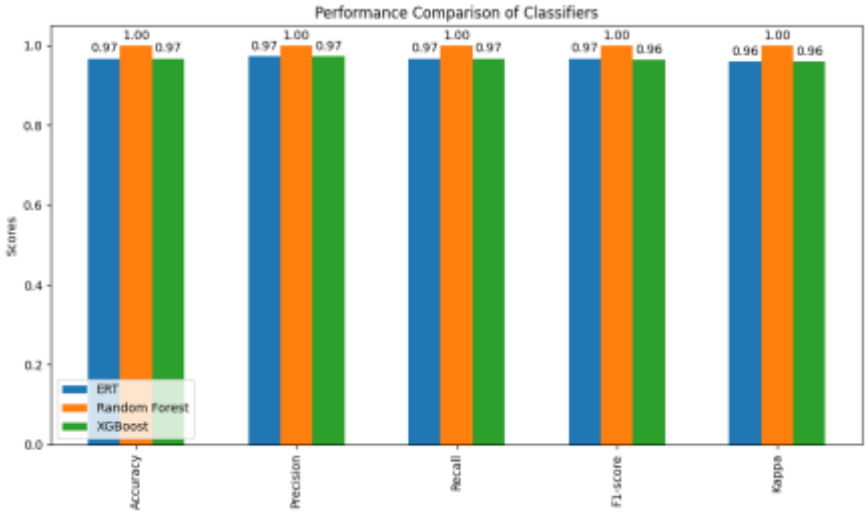


Figure 8 Performance metrics

3.3 CUCKOO SEARCH

Figure 9 shows the fitness function obtained from Cuckoo search. It begins with a relatively high fitness value of around 0.053, indicating that the initial feature subsets in Cuckoo Search likely had lower accuracy. However, there is a sharp drop in the fitness value around iteration 16, reducing to 0.046, signifying a significant

improvement in the solution as a better feature subset was found. Following this, the fitness value stabilizes at approximately 0.0462, suggesting that Cuckoo Search converged to an optimal solution, with no further significant improvements. The final fitness value corresponds to an accuracy of about 95.38%, demonstrating the effectiveness of the selected features in achieving high classification accuracy. Overall, the algorithm showed fast convergence, with the major improvement occurring early in the process, and the selected features proved to be highly accurate for the classification task.

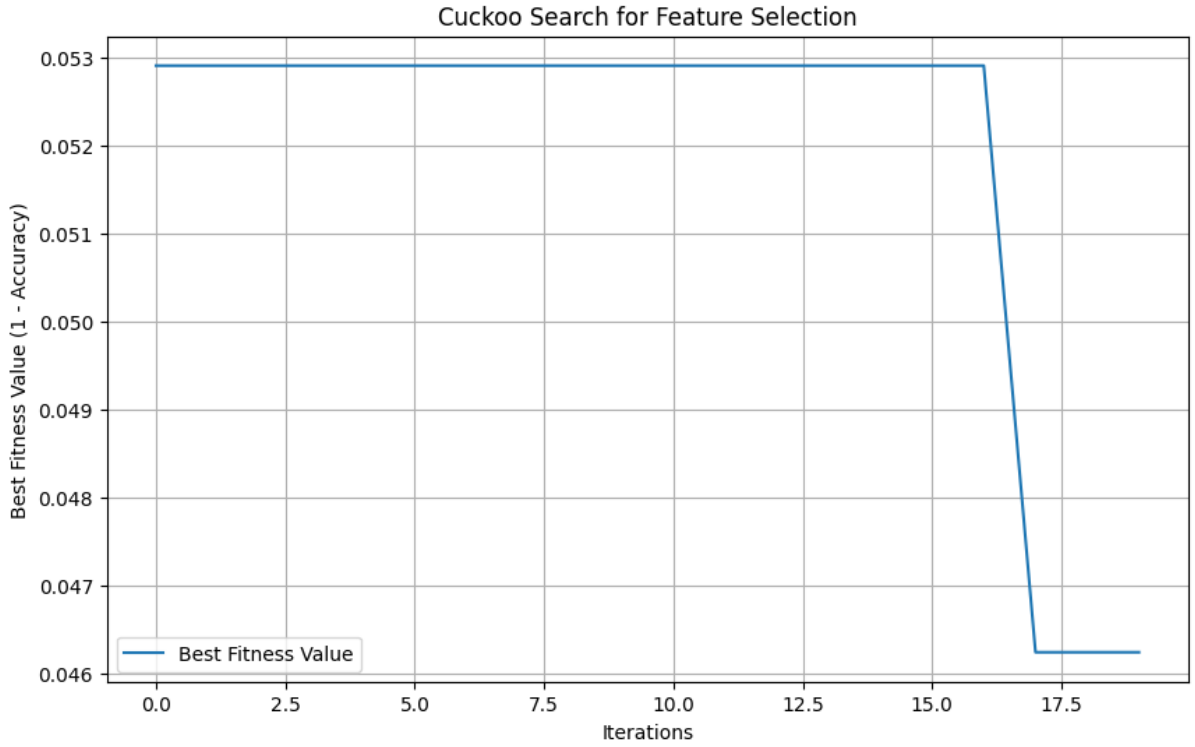
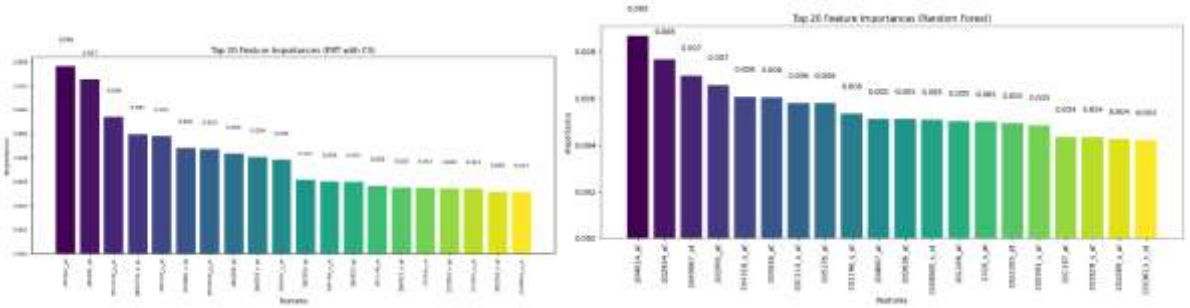
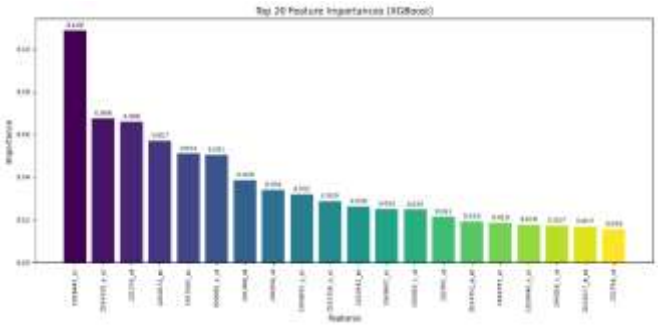


Figure 9 Fitness function values achieved using PSO

Figures 10(a), 10(b) and 10(c) represents the top 20 features selected from ERT, Random Forest, and XGBoost, respectively.



(a) (b)



(c)

Figure 10 Top 20 Features Selected by Extremely Randomized Trees with Cuckoo Search features

Figure 11 illustrates the performance metrics of classifiers using the selected features from the Cuckoo Search algorithm. Both ERT and Random Forest achieve an accuracy of 90%, while XGB attains an accuracy of 84%.

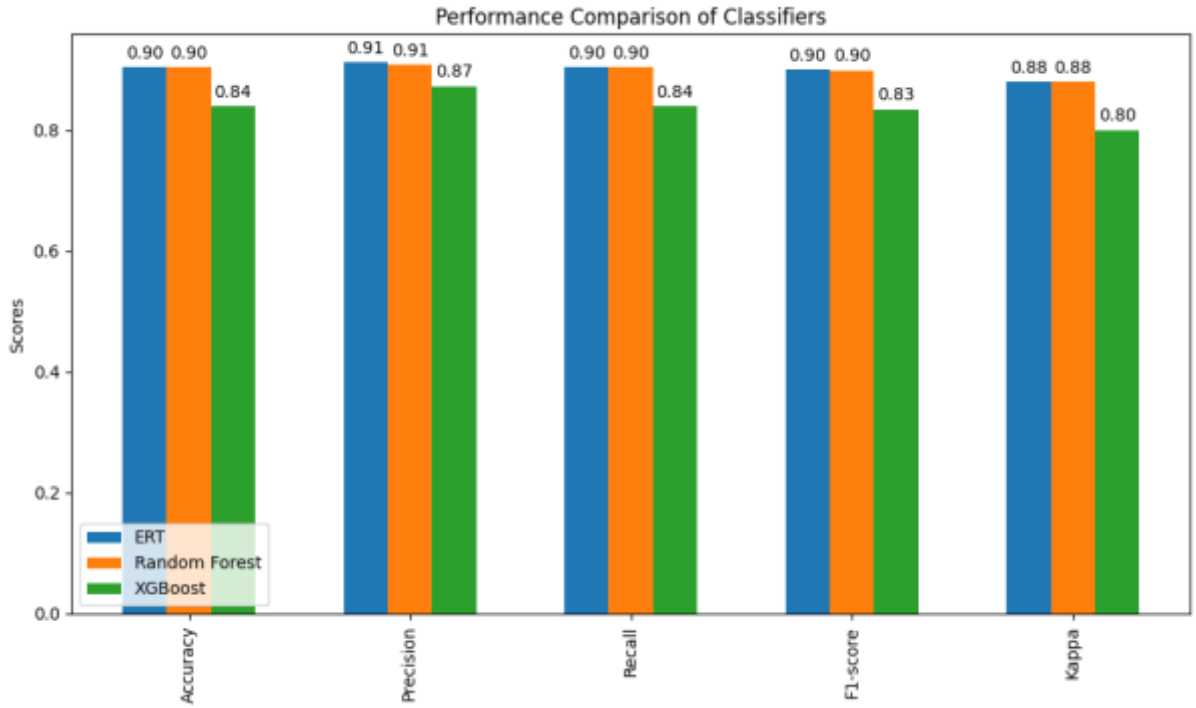


Figure 11 Performance metrics

3.4 SMBO

Figure 12 depicts the fitness values obtained from SMBO. The graph begins with a relatively high fitness value of around 0.053, indicating that the initial feature subsets selected by SMBO led to lower accuracy, as fitness is calculated as 1 minus accuracy. A significant drop in fitness occurs at iteration 7, from 0.053 to 0.046, suggesting that SMBO identified a notably better feature subset at this point. Following this sharp improvement, the fitness value stabilizes around 0.0462 and remains constant throughout the remaining iterations, signaling that the algorithm has likely converged to an optimal solution. This final fitness value corresponds to an accuracy of approximately 95.38%, demonstrating that the chosen features were highly effective for the classification task. The process highlights SMBO's efficiency in feature selection, with rapid convergence and an effective feature subset contributing to high classification accuracy.

Figure 13(a), 13(b), and 13(c) present the top 20 features selected by ERT, Random Forest, and XGBoost, respectively.

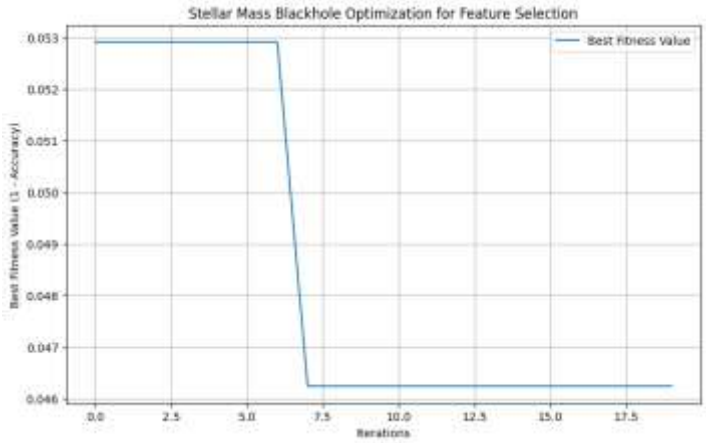
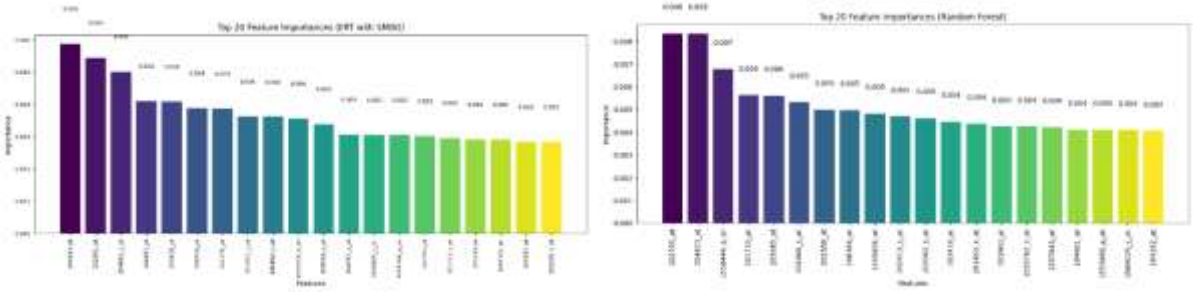
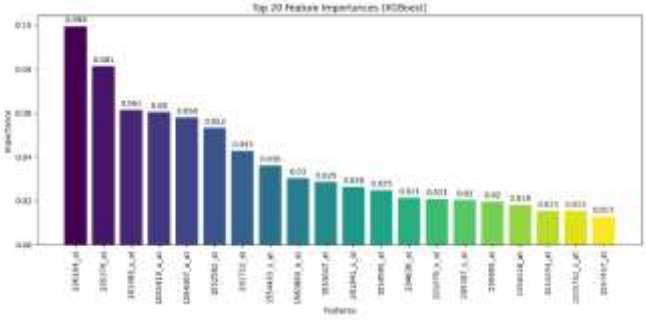


Figure 12 Fitness values obtained from SMBO



(a) (b)



(c)

Figure 13 Top 20 Features Selected by Extremely Randomized Trees with SMBO features

Figure 14 displays the performance metrics derived from SMBO. ERT achieved an accuracy of 100%, while Random Forest recorded 93% and XGBoost attained 81%.

TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access

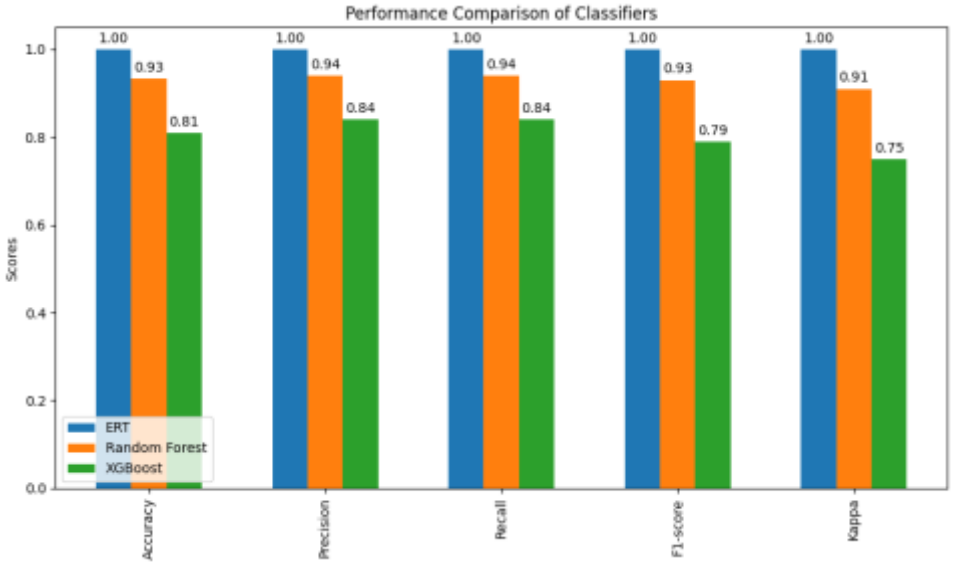


Figure 14 Performance metrics

3.5 BIOLOGICAL INSIGHTS

Figure 15 visualizes the coefficients, p-values, and adjusted p-values for the top 20 genes. The x-axis represents the gene names, while the y-axis shows the values of the coefficients and p-values. The blue bars represent the regression coefficients, indicating the effect size of each gene in the model. The orange bars correspond to the p-values, which assess the statistical significance of each gene's contribution. Higher p-values suggest that the corresponding gene is less significant in the model, while lower p-values (typically below 0.05) indicate statistical significance. The presence of adjusted p-values (not explicitly differentiated in the graph) suggests that multiple testing correction, such as the Bonferroni or Benjamini-Hochberg method, was applied to control for false positives. Genes with very low p-values (below 0.05) are likely to be important predictors, whereas genes with high p-values may not have a strong association with the outcome. The adjusted p-values further refine this significance by accounting for multiple comparisons.

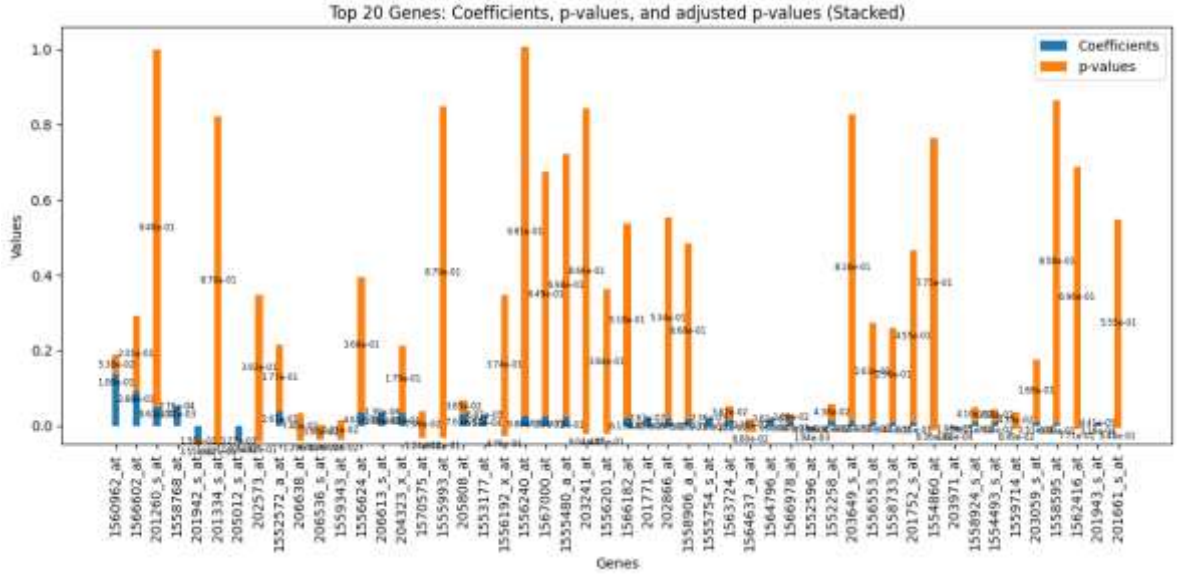


Figure 15 Top 20 genes p-values and adjusted p-values

Table 4 displays the top 20 biomarker genes along with their corresponding values.

Table 4 Top Biomarker Genes Descriptions

index	gene	t_statistic	p_value	p_adjusted	adjusted_p_value	coefficients_x	coefficients_y	coefficients	fold_change
131	1554866_at	-6.43762	2.06E-07	3.96E-06	3.96E-06	0	0	0	2.858211
210	1562687_x_at	-6.37622	2.47E-07	4.62E-06	4.62E-06	0	0	0	2.839925
192	1557578_at	-6.75254	7.99E-08	1.81E-06	1.81E-06	0	0	0	2.817224
158	1556182_x_at	-12.3837	2.39E-14	1.26E-11	1.26E-11	0	0	0	2.790487
488	206453_s_at	-11.987	6.04E-14	2.60E-11	2.60E-11	0	0	0	2.722198
652	205743_at	-6.90044	5.14E-08	1.26E-06	1.26E-06	0	0	0	2.670208
75	1554418_s_at	-6.19876	4.23E-07	7.07E-06	7.07E-06	0	0	0	2.630459
648	205200_at	-11.2937	3.20E-13	9.53E-11	9.53E-11	0	0	0	2.613967
277	1561692_at	-4.02228	0.000293	0.001436	0.001436	0	0	0	2.556815
405	1569818_at	-8.73803	2.56E-10	1.81E-08	1.81E-08	0	0	0	2.536128
613	204537_s_at	-3.79861	0.000557	0.002408	0.002408	0	0	0	2.534147
425	1570270_at	-5.7841	1.49E-06	1.98E-05	1.98E-05	0	0	0	2.492883
477	200908_s_at	-4.80415	2.89E-05	0.000225	0.000225	0	0	0	2.398546
479	202242_at	-6.30855	3.03E-07	5.47E-06	5.47E-06	0	0	0	2.388456
420	206747_at	-8.74539	2.50E-10	1.80E-08	1.80E-08	0	0	0	2.334788
56	1561574_at	-5.90596	1.03E-06	1.47E-05	1.47E-05	0.003048	0.003048	0.003048	2.330235
335	1565483_at	-5.54213	3.10E-06	3.60E-05	3.60E-05	0	0	0	2.285406
77	1552773_at	-5.80589	1.39E-06	1.88E-05	1.88E-05	0	0	0	2.225822

index	gene	t_statistic	p_value	p_adjusted	adjusted_p_value	coefficients_x	coefficients_y	coefficients	fold_change
63	1552825_at	-4.69891	3.97E-05	0.000292	0.000292	0	0	0	2.208701
275	1561668_at	-4.65752	4.49E-05	0.000321	0.000321	0	0	0	2.199601

3.5.1 Validation of Hub Genes Using an Independent Dataset and qRT-PCR

The differential expression of hub genes was validated using the independent GSE45827 dataset, which consists of 130 primary invasive breast cancer samples, categorized into 41 Basal-like, 30 HER2, 29 Luminal A, and 30 Luminal B subtypes, alongside 11 normal tissue samples. Furthermore, survival analysis of these hub genes was performed using the Kaplan-Meier Plotter [32], an online tool for discovering and validating survival biomarkers. A Kaplan-Meier plot is a statistical tool used to estimate survival functions from lifetime data, commonly applied in gene expression studies to evaluate survival differences based on gene expression levels. In the context of breast cancer, it is used to group patients based on high and low expression of specific genes, then plot survival curves for these groups over time. The plot visualizes the proportion of patients surviving at different time points for each group, and a log-rank test is often performed to assess if the survival difference between groups is statistically significant. If a gene correlates with survival outcomes, the Kaplan-Meier plot will show distinct survival curves for high and low expression groups, identifying it as a potential biomarker for prognosis.

205225_at, 201171_at, 204913_s_at, 205200_at, 204914_s_at, 201446_s_at, 203685_at, 204840_s_at, 206165_s_at, 201668_x_at, 203249_at, 204915_s_at, 205967_at, 203196_at, 200070_at, 203678_at, 206861_s_at, 206378_at, 203380_x_at, 202984_s_at, 204716_at, 201695_s_at, 204134_at, 203971_at, 203571_s_at, 201299_s_at, 202801_at, 204313_s_at '201299_s_at'

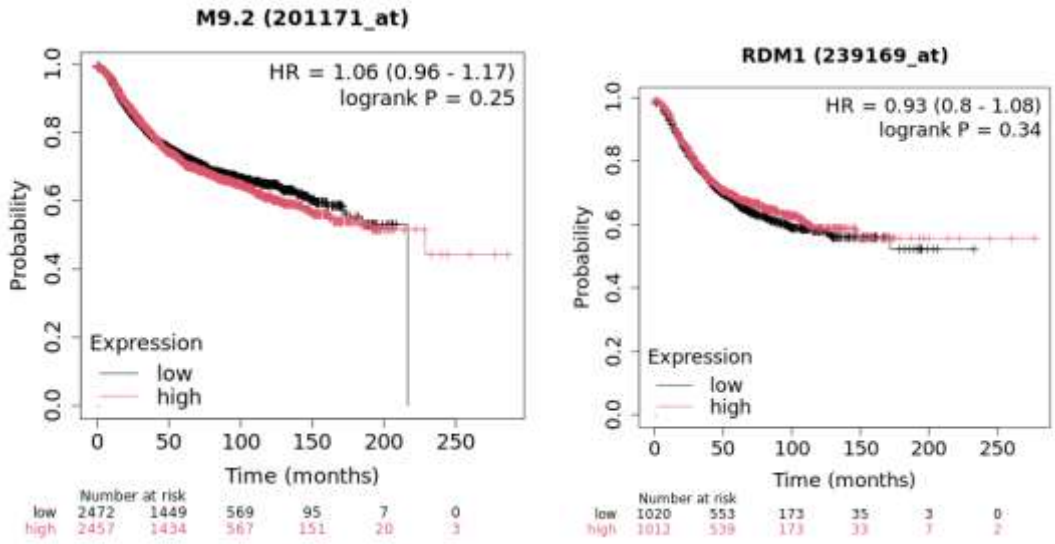
The sample '201299_s_at' -RDM1 (RAD52 Motif Containing 1) plays a crucial role in breast cancer by facilitating DNA damage repair, particularly in the repair of double-strand breaks. Its abnormal expression or mutations have been associated with tumor progression and resistance to chemotherapy. As a key player in the homologous recombination repair pathway, RDM1 helps maintain genomic stability in breast cancer cells. Given its significance in DNA repair mechanisms, it is considered a potential therapeutic target, with DNA repair inhibitors being explored to enhance the effectiveness of chemotherapeutic treatments.

3.5.2 Kaplan-Meier Survival Plot

The Kaplan-Meier (KM) survival plot illustrates the survival probabilities of two groups of patients based on the expression levels of the gene 201171_at over time. The x-axis represents time in months, while the y-axis shows the probability of survival. Two survival curves are plotted: one for patients with low expression (black line) and another for those with high expression (red line). The hazard ratio (HR) of 1.06 with a 95% confidence interval (0.96 - 1.17) indicates that patients in the high-expression group have a slightly higher risk of death compared to those in the low-expression group. However, the log-rank p-value of 0.25 suggests that this difference is not statistically significant. The number-at-risk table at the bottom provides the count of patients still being tracked at various time points, starting with roughly 2500 patients in each group and decreasing over time. Since the survival curves remain close to each other and the p-value is above 0.05, the analysis suggests that the expression level of this gene does not have a significant impact on patient survival. Figures 16(a) and 16(b) illustrate the Kaplan-Meier (KM) survival plots for GeneID 201171_at and 239169_at, respectively.

TPM Vol. 32, No. S4, 2025
ISSN: 1972-6325
https://www.tpmap.org/

Open Access



(a) (b)

Figure 16 Kaplan-Meier survival plot

Figure 17 represents a gene interaction network, likely generated using GeneMANIA [33] tool. In this network, nodes (circles) represent genes or proteins, while edges (lines) indicate interactions such as co-expression, genetic interactions, pathway involvement, physical protein-protein interactions, and predicted associations. The central gene, RDM1 (RAD52 motif 1), acts as a hub, displaying multiple interactions with other genes, suggesting its significant role in cellular functions. Notably, RDM1 is strongly linked to RAD52, implying a potential role in DNA repair or genomic stability. Other interacting genes, including PLK1, PLK4, PI4KB, and PARPBP, suggest associations with cell cycle regulation, kinase signaling, and DNA damage response. The network also includes several other genes with lighter edges, indicating weaker or indirect associations. This visualization provides insights into potential functional relationships between these genes, which could be further explored for their biological significance.

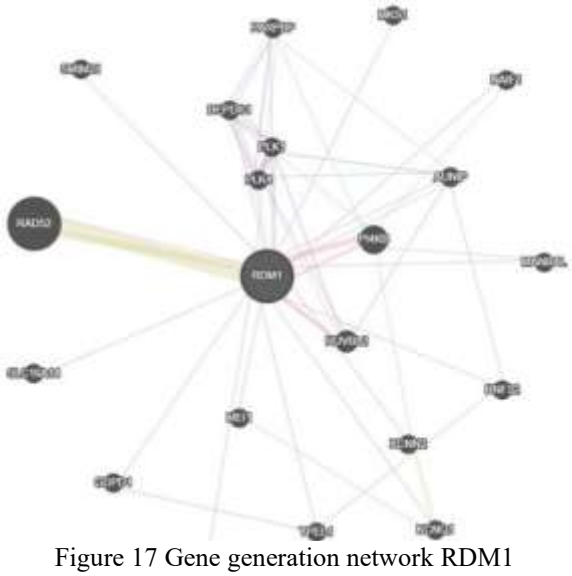


Figure 17 Gene generation network RDM1

Table 5 displays the gene set with differential expression for a p-value less than 0.05.

Table 5 Differentially Expressed Genes Count

Differentially Expressed Gene Set	Differentially Expressed Genes (p-value < 0.05)
HER and basal	4372
HER and cell_line	10074

HER and luminal_A	2460
HER and luminal_B	6885
HER and normal	5752
basal and cell_line	9462
basal and luminal_A	8265
basal and luminal_B	7725
basal and normal	9925
cell_line and luminal_A	10310
cell_line and luminal_B	9274
cell_line and normal	8112
luminal_A and luminal_B	4102
luminal_A and normal	9257
luminal_B and normal	8688

4. DISCUSSIONS

Feature selection or dimensionality reduction is of paramount importance in the analysis of gene expression data, particularly in high-dimensional datasets such as the GSE45827 breast cancer dataset. These datasets often contain a large number of genes, but the number of samples may be limited, making the identification of relevant features critical for achieving high classification accuracy. Without proper feature selection, the presence of irrelevant or redundant genes can lead to overfitting, reduced accuracy, and poor generalization of machine learning models. Moreover, the biological interpretation of the results becomes challenging due to the overwhelming number of features. Thus, the effective selection of informative genes can improve both the performance of predictive models and the biological understanding of the underlying disease mechanisms.

In this work, three heuristic optimization techniques—Self-Organizing Migrating Algorithm (SOMA), Particle Swarm Optimization (PSO), and Stellar Mass Black Hole Optimization (SMBO)—were employed for feature selection. These optimization algorithms are particularly suitable for gene expression data due to their ability to handle the complexity and high dimensionality of the dataset. SOMA, PSO, and SMBO explore the search space efficiently and identify the most relevant genes for classification tasks, improving model performance. By incorporating these techniques, we were able to significantly reduce the dimensionality of the data while retaining the most informative features that contribute to the classification of breast cancer.

Further refinement of the selected genes was achieved by applying ElasticNet, a second-level feature selection method. ElasticNet combines the strengths of both L1 and L2 regularization, making it particularly effective in handling multicollinearity and selecting a smaller number of highly informative genes. This additional layer of feature selection helped to ensure that only the most relevant and non-redundant genes were retained, thereby enhancing the interpretability of the results and reducing noise in the data.

The selected gene subsets were then used in ensemble learning models, including Random Forest, Extreme Randomized Trees (ERT), and XGBoost. Ensemble learners combine multiple weak models to produce a stronger model, improving accuracy and robustness. In this study, Random Forest achieved 100% accuracy with PSO-selected features, 90% accuracy with Cuckoo Search, and 97% accuracy with SOMA-selected features, demonstrating the potential of these optimization techniques to improve classification performance. ERT reached 100% accuracy using SMBO-selected features, highlighting the power of SMBO in optimizing the gene subset for classification tasks. These results underscore the importance of combining feature selection with ensemble learning to achieve high accuracy in breast cancer classification.

In addition to classification performance, this study also provided valuable biological insights. Differentially expressed genes (DEGs) were identified, and pathway analysis revealed key biological processes and pathways that may be implicated in breast cancer progression. The identification of biomarker genes is critical for developing personalized treatment strategies and improving early detection methods. The use of ElasticNet for feature selection further refined the list of relevant genes, enabling the identification of key biomarkers that could serve as targets for therapeutic intervention.

The proposed work highlights the significance of feature selection in enhancing the performance of machine learning models for gene expression data analysis. By incorporating heuristic optimization techniques such as SOMA, PSO, and SMBO, and using ensemble learning models, this study demonstrates the potential for improving classification accuracy while gaining biological insights into breast cancer. The identification of differentially expressed genes and key pathways not only aids in understanding the underlying biology of the disease but also contributes to the discovery of novel biomarkers for early detection and personalized oncology

treatments. The findings emphasize the importance of combining computational techniques with biological knowledge to develop more reliable predictive models and improve patient outcomes.

Data Availability

The GSE45827 dataset was retrieved from the **Gene Expression Omnibus (GEO)** database.

Conflict of Interests

The authors declare that there are no conflicts of interest related to this research, authorship, or publication of this paper

Funding

The authors declare that no external funding was received for this research

Author Contributions

Premalatha and Sivakumar contributed to the conception and design of the study, carried out computational experiments, analyzed and interpreted the data, and drafted the manuscript. Ramya revised the manuscript for clarity and coherence. Additionally, Premalatha played a key role in conceptualizing the project, assisting with data interpretation, and contributing to both drafting and revising the manuscript. All authors have read and approved the final version of the manuscript.

REFERENCES

1. Alberts, B.; Johnson, A.; Lewis, J.; et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Studying Gene Expression and Function. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26818/
2. Saheed, Y.K. Chapter 9 - Effective dimensionality reduction model with machine learning classification for microarray gene expression data. In Data Science for Genomics; Tyagi, A.K., Abraham, A., Eds.; Academic Press: Cambridge, MA, USA, 2023; pp. 153–164. https://doi.org/10.1016/B978-0-323-98352-5.00006-9
3. Bao, S.; He, G. Identification of Key Genes and Key Pathways in Breast Cancer Based on Machine Learning. Med. Sci. Monit. 2022, 28, e935515. https://doi.org/10.12659/MSM.935515
4. Gruosso, T.; Mieulet, V.; Cardon, M.; Bourachot, B.; Kieffer, Y.; Devun, F.; Dubois, T.; Dutreix, M.; Vincent-Salomon, A.; Miller, K.M.; et al. Chronic oxidative stress promotes H2AX protein degradation and enhances chemosensitivity in breast cancer patients. EMBO Mol. Med. 2016, 8, 527–549. https://doi.org/10.15252/emmm.201505891
5. West, M.; Blanchette, C.; Dressman, H.; Huang, E.; Ishida, S.; Spang, R.; Zuzan, H.; Olson, J.A., Jr.; Marks, J.R.; Nevins, J.R. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA 2001, 98, 11462–11467. https://doi.org/10.1073/pnas.20116299
6. Sotiriou, C.; Neo, S.; McShane, L.M.; Korn, E.L.; Long, P.M.; Jazaeri, A.; Martiat, P.; Fox, S.B.; Harris, A.L.; Liu, E.T. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc. Natl. Acad. Sci. USA 2003, 100, 10393–10398.
7. Bao, T.; Davidson, N.E. Gene expression profiling of breast cancer. Adv. Surg. 2008, 42, 249–260. https://doi.org/10.1016/j.yasu.2008.03.002
8. Malvia, S.; Bagadi, S.A.R.; Pradhan, D.; et al. Study of Gene Expression Profiles of Breast Cancers in Indian Women. Sci. Rep. 2019, 9, 10018. https://doi.org/10.1038/s41598-019-46261-1
9. Latha, N.R.; Rajan, A.; Nadhan, R.; Achyutuni, S.; Sengodan, S.K.; Hemalatha, S.K.; Varghese, G.R.; Thankappan, R.; Krishnan, N.; Patra, D.; et al. Gene expression signatures: A tool for analysis of breast cancer prognosis and therapy. Crit. Rev. Oncol. Hematol. 2020, 151, 102964. https://doi.org/10.1016/j.critrevonc.2020.102964
10. Wang, Y.; Li, Y.; Liu, B.; Song, A. Identifying breast cancer subtypes associated modules and biomarkers by integrated bioinformatics analysis. Biosci. Rep. 2021, 41, BSR20203200. https://doi.org/10.1042/BSR20203200
11. Nhut, P.N.; Chi-Cheng, H.; Tseng, L.M.; Chuang, E.Y. Predicting Breast Cancer Gene Expression Signature by Applying Deep Convolutional Neural Networks From Unannotated Pathological Images. Front. Oncol. 2021, 11, 769447. https://doi.org/10.3389/fonc.2021.769447
12. Altaf, R.; Nadeem, H.; Babar, M.M.; Ilyas, U.; Muhammad, S.A. Genome-scale meta-analysis of breast cancer datasets identifies promising targets for drug development. J. Biol. Res.-Thessalon. 2021, 28, 5. https://doi.org/10.1186/s40709-021-00136-7
13. Raiesdana, S. Breast Cancer Detection Using Optimization-Based Feature Pruning and Classification Algorithms. Middle East J. Cancer 2021, 12, 48–68. https://doi.org/10.30476/mejc.2020.85601.1294

14. Aoun, R.; El Hadi, C.; Tahtouh, R.; El Habre; Hilal, D. Microarray analysis of breast cancer gene expression profiling in response to 2-deoxyglucose, metformin, and glucose starvation. Cancer Cell Int. 2022, 22, 123. https://doi.org/10.1186/s12935-022-02542-w
15. Rahman, M.A.; Muniyandi, R.C. An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons. Symmetry 2020, 12, 271. https://doi.org/10.3390/sym12020271
16. Ali, P.J.M. Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements. ARO-The Sci. J. Koya Univ. 2022. https://doi.org/10.14500/aro.10955
17. Thalor, A.; Joon, H.K.; Singh, G.; Roy, S.; Gupta, D. Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. Comput. Struct. Biotechnol. J. 2022, 20, 1618–1631. https://doi.org/10.1016/j.csbj.2022.03.019
18. Ali, N.M.; Besar, R.; Aziz, N.A. Hybrid Feature Selection of Breast Cancer Gene Expression Microarray Data Based on Metaheuristic Methods: A Comprehensive Review. Symmetry 2022, 14, 1955. https://doi.org/10.3390/sym14101955
19. Zhang, S.; Jiang, H.; Gao, B.; Yang, W.; Wang, G. Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network. Front. Cell Dev. Biol. 2022, 9. https://doi.org/10.3389/fcell.2021.769447
20. Rakhshaninejad, M.; Fathian, M.; Shirkoohi, R.; Barzinpour, F.; Gandomi, A.H. Refining breast cancer biomarker discovery and drug targeting through an advanced data-driven approach. BMC Bioinformatics 2024, 25, 33. https://doi.org/10.1186/s12859-024-05657-1
21. Wu, X.; Sun, Y.; Liu, X. Multi-Class Classification of Breast Cancer Gene Expression Using PCA and XGBoost. Preprints 2024. https://doi.org/10.20944/preprints202410.1775.v1
22. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324
23. Batista, G.E.; Monard, M.C. A Study of K-Nearest Neighbour as an Imputation Method. In Soft Computing Systems - Design, Management and Applications; HIS: Santiago, Chile, 1–4 December 2002.
24. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. https://doi.org/10.1145/2939672.2939785
25. Diep, Q.B.; Truong, T.C.; Zelinka, I. Training artificial neural networks using self-organizing migrating algorithm for skin segmentation. Sci. Rep. 2024, 14, 22651. https://doi.org/10.1038/s41598-024-72884-0
26. Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In Proceedings of the IEEE International Conference on Neural Networks, Perth, Australia, 27 November–1 December 1995; Volume IV, pp. 1942–1948. https://doi.org/10.1109/ICNN.1995.488968
27. Yang, X.-S.; Deb, S. Cuckoo search via Lévy flights. In Proceedings of the World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), Coimbatore, India, 9–11 December 2009; pp. 210–214. arXiv:1003.1594v1
28. Premalatha, K.; Balamurugan, R. A nature inspired swarm based stellar-mass black hole for engineering optimization. In Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, India, 5–7 March 2015; pp. 1–8. https://doi.org/10.1109/ICECCT.2015.7225975
29. Zou, H.; Hastie, T. Regularization and Variable Selection Via the Elastic Net. J. R. Stat. Soc. Ser. B Stat. Methodol. 2005, 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
30. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010, 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616
31. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000, 28, 27–30. https://doi.org/10.1093/nar/28.1.27
32. Győrffy, B.; Lanczky, A.; Eklund, A.C.; Denkert, C.; Budczies, J.; Li, Q.; Szallasi, Z. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res. Treat. 2010, 123, 725–731. https://doi.org/10.1007/s10549-009-0674-9
33. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Sander, C. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010, 38, W214–W220. https://doi.org/10.1093/nar/gkq537