

RESEARCH

Open Access



Refining breast cancer biomarker discovery and drug targeting through an advanced data-driven approach

Morteza Rakhshaninejad^{1†}, Mohammad Fathian^{1*†}, Reza Shirkoohi^{2†}, Farnaz Barzinpour^{1†} and Amir H. Gandomi^{3,4†}

[†]Morteza Rakhshaninejad, Mohammad Fathian, Reza Shirkoohi, Farnaz Barzinpour and Amir H. Gandomi have contributed equally to this work.

*Correspondence: fathian@iust.ac.ir

¹ Industrial Engineering Department, Iran University of Science and Technology, Hengam Street, Tehran 1684613114, Tehran, Iran

² Cancer Biology Research Center, Cancer Institute, Imam Khomeini Hospital Complex, Tehran University of Medical Sciences, Keshavarz Boulevard, Tehran 1419733141, Tehran, Iran

³ Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo 2007, NSW, Australia

⁴ University Research and Innovation Center (EKIK), Óbuda University, Budapest 1034, Hungary

Abstract

Breast cancer remains a major public health challenge worldwide. The identification of accurate biomarkers is critical for the early detection and effective treatment of breast cancer. This study utilizes an integrative machine learning approach to analyze breast cancer gene expression data for superior biomarker and drug target discovery. Gene expression datasets, obtained from the GEO database, were merged post-preprocessing. From the merged dataset, differential expression analysis between breast cancer and normal samples revealed 164 differentially expressed genes. Meanwhile, a separate gene expression dataset revealed 350 differentially expressed genes. Additionally, the BGWO_SA_Ens algorithm, integrating binary grey wolf optimization and simulated annealing with an ensemble classifier, was employed on gene expression datasets to identify predictive genes including TOP2A, AKR1C3, EZH2, MMP1, EDNRB, S100B, and SPP1. From over 10,000 genes, BGWO_SA_Ens identified 1404 in the merged dataset (F1 score: 0.981, PR-AUC: 0.998, ROC-AUC: 0.995) and 1710 in the GSE45827 dataset (F1 score: 0.965, PR-AUC: 0.986, ROC-AUC: 0.972). The intersection of DEGs and BGWO_SA_Ens selected genes revealed 35 superior genes that were consistently significant across methods. Enrichment analyses uncovered the involvement of these superior genes in key pathways such as AMPK, Adipocytokine, and PPAR signaling. Protein-protein interaction network analysis highlighted subnetworks and central nodes. Finally, a drug-gene interaction investigation revealed connections between superior genes and anticancer drugs. Collectively, the machine learning workflow identified a robust gene signature for breast cancer, illuminated their biological roles, interactions and therapeutic associations, and underscored the potential of computational approaches in biomarker discovery and precision oncology.

Keywords: Breast cancer, Machine learning, Biomarker discovery, Ensemble method, Hybrid metaheuristic algorithm



Introduction

Breast cancer remains one of the foremost health challenges globally, consistently accounting for a significant fraction of cancer diagnoses and mortalities among women. The World Health Organization (WHO) statistics underscore the escalating prevalence of breast cancer (BC) cases and the associated mortalities [1]. The pivotal role of early detection cannot be overstated: it not only dramatically improves the prognosis but also potentially reduces mortality rates [2]. In this light, crafting precise and accurate predictive models for BC detection is of paramount importance, bridging the gap between early diagnosis and effective therapeutic interventions.

Despite technological advancements, the field of BC research grapples with numerous challenges, primarily in the identification of accurate biomarkers [3]. The high rate of false positives in BC detection is a considerable impediment, necessitating more reliable predictive models [4]. Another significant barrier is the scarcity of comprehensive datasets. To ameliorate this issue, integrating existing datasets can be invaluable [5]. Machine learning (ML) offers a transformative solution to these challenges by facilitating the use of combined gene expression datasets to enhance the accuracy and reliability of BC prediction.

Embarking on an interdisciplinary approach, this study melds ML and optimization algorithms with systems biology to identify genes intrinsically linked to BC. Through an exhaustive meta-analysis of the Gene Expression Omnibus (GEO) datasets, we harnessed techniques to tease out genes that starkly demarcate BC from normal samples. The curated gene signatures were then juxtaposed with differentially expressed genes and subjected to rigorous Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. This culminated in the formulation of a state-of-the-art feature selection (FS) algorithm, enhancing the efficiency and accuracy of biomarker identification. This interdisciplinary effort extends beyond mere prediction, delving into the identification of superior genes that offer promising insights into the mechanisms of the disease and potential therapeutic targets.

The primary aim of this paper is to develop and validate a novel predictive model for BC detection that surpasses current methods' limitations. This research, to the best of our knowledge, is the first to apply a unique combination of hybrid metaheuristic algorithms and ensemble models to gene expression datasets, targeting the discovery of biomarkers with unparalleled precision. The successful identification of key genes in this study marks a significant advancement in comprehending BC, opening up new possibilities for drug-gene interaction analysis and the development of personalized medicine strategies.

The organization of this paper is meticulously structured to guide the reader through the intricate processes and findings of the study. Following this introduction, in section "[Literature review](#)" conducts a critical review of the pertinent literature. In section "[Materials and methods](#)" delineates the materials and methods, encapsulating data procurement, FS protocols, ensemble method formulation, and drug-gene interaction analysis. In section "[Results](#)" unveils the empirical results, including differential gene expression analysis and the identification of superior genes. In section

“Discussion” interprets these findings within the broader context of biomarker discovery and precision oncology. The paper reaches its culmination in section “Conclusion”, where the implications of the study for the future of BC detection and treatment are synthesized.

Literature review

Breast cancer is a multifaceted disease with a genomic footprint that can be deciphered for early detection and treatment strategies. In the quest for precision medicine, ML algorithms have become indispensable tools in analyzing complex biomedical data, such as gene expression profiles, to identify potential biomarkers for diseases such as BC [6].

Unlike imaging or clinical data, gene expression data offer a dynamic picture of the cellular mechanisms at play during oncogenesis [7]. High-throughput technologies such as microarrays and RNA sequencing facilitate a holistic view of gene activity across the genome [8]. The GEO and ArrayExpress databases exemplify resources that have democratized access to such data, enabling extensive research into gene function and regulation [9].

The discovery of gene expression biomarkers is paramount for diagnostic and prognostic advancements in BC [10]. Differential gene expression analysis serves as a cornerstone methodology, distinguishing between normal and diseased states by identifying genes with significantly altered expression [11].

As the volume and complexity of gene expression data burgeon, ML offers a suite of algorithms capable of deciphering intricate patterns within the data [12]. Studies have utilized various ML approaches, such as decision trees (DT) [13], neural networks (NN) [14], support vector machines (SVM) [15], logistic regression (LR) [16], and random forests (RF) [17], alongside more recent innovations in deep learning (DL) [18, 19] and ensemble learning [20, 21] methods such as extreme gradient boosting (XGBoost) [5] and adaptive boosting (AdaBoost) [22], to identify significant biomarkers in BC.

The preprocessing of gene expression data, including normalization [23] and FS [24], is vital to ensure the reliability of subsequent analyses. The Robust Multiarray Average (RMA) is often employed for its robustness in normalization and expression calculation [25]. Feature selection is a critical step to highlight relevant genes and can be addressed through various methods, including filter, wrapper, embedded, and dimensionality reduction techniques [26].

Merging and combining gene expression datasets enhance the power and scope of analysis, enabling a comprehensive view across varied conditions, treatments, and populations, thus improving the generalizability and robustness of the findings [27]. Such integration is essential to address batch effects through normalization, which is crucial when datasets from different sources are combined [5]. Moreover, a larger dataset provides a more substantial basis for FS methods to uncover informative genes, leading to more accurate models of gene expression [28]. This process not only corroborates the findings across different studies, strengthening the reproducibility of research but also significantly advances the field of personalized medicine and our understanding of complex diseases.

The combinatorial challenge of FS has been approached with metaheuristic algorithms, which offer powerful solutions to NP-Hard problems through intelligent search strategies [29]. Algorithms such as genetic algorithms (GA) [30], simulated annealing (SA) [31], whale optimization algorithm (WOA) [32], grey wolf optimization (GWO) [33], and particle swarm optimization (PSO) [20] have been utilized across various high-dimensional datasets, demonstrating their efficacy in identifying relevant features.

Recent advancements in FS have seen a growing interest in hybrid methods that combine the strengths of different metaheuristic algorithms. These hybrid approaches aim to overcome the limitations of individual algorithms by integrating their unique search mechanisms. For instance, the fusion of GA with artificial bee colony (ABC) has shown promising results in navigating complex search spaces more effectively [34]. Another notable hybrid method involves combining the spotted hyena optimizer (MOSHO) with the salp swarm algorithm (SSA), which enhances the ability of FS in high-dimensional datasets [35]. A novel framework, introducing CS-GA and ABC-CS, combines the cuckoo search (CS) algorithm with ABC and GA to enhance both exploitation and exploration, thus achieving a balanced search process [36]. These hybrid methods are not only capable of handling the high dimensionality of datasets but also exhibit improved performance in terms of accuracy and computational efficiency in FS tasks.

In the realm of FS, the convergence of hybrid optimization techniques with advanced ML methodologies is forging innovative pathways. Notably, in these hybrid models, the use of ensemble learning frameworks [20] as objective functions for metaheuristic algorithms is an emerging and novel approach. This method adeptly refines the search strategy of metaheuristic algorithms to focus on enhancing the predictive accuracy and generalization of ensemble models [37], which are particularly effective in addressing the challenges of imbalanced and high-dimensional data. The result is a more refined FS that is inherently aligned with the core strengths and requirements of the ensemble methods [38].

Following the identification of biomarkers, their functional interpretation is enabled through enrichment analyses, utilizing resources such as GO [39] and KEGG [40]. These tools provide insights into gene functions and pathways, allowing for a deeper understanding of the biological significance behind the differential gene expression observed.

Upon analyzing biomarkers, the investigation of protein-protein interactions sheds light on cellular pathways integral to disease pathophysiology, guiding the creation of targeted therapies [41]. Moreover, comprehending drug-gene interactions becomes essential in formulating personalized medicine approaches, especially in the context of BC treatment [42].

While gene expression data offer tremendous potential in advancing BC research, several challenges persist that can affect the integrity and applicability of the results. Inconsistencies in data preprocessing, such as normalization and FS, can lead to

variable and sometimes unreliable outcomes. These issues underscore the need for rigorous and standardized methodologies in handling gene expression data [43]. Additionally, ML models, pivotal in analyzing these complex data, face their own set of challenges. Overfitting remains a primary concern, where models may perform excellently on training data but fail to generalize to new data [5]. This is further compounded by the often intricate nature of these models, leading to issues with interpretability and understanding of the underlying biological mechanisms. Furthermore, the effectiveness of these models is frequently contingent on the availability of large, high-quality training datasets, which are not always accessible or feasible to obtain [44]. These challenges highlight the need for ongoing refinement of ML techniques and data processing methods to enhance the reliability and applicability of gene expression analysis in BC research.

In addressing the complexities and challenges of gene expression data analysis in BC research, our study adopts a holistic and innovative approach. We merged two gene expression datasets to create a robust and diverse data foundation, which is essential for overcoming the limitations of data quality and quantity. This merged dataset is then rigorously processed using standardized methodologies for normalization and FS, ensuring reliable and consistent analysis. Recognizing that reliance on a single, merged dataset might not be fully representative, we have also prepared an additional gene expression dataset, named GSE45827, to test and validate the developed method. To specifically counter issues such as overfitting and to improve model interpretability, our research introduces a novel ensemble learning model, incorporating a specialized weighted voting method. This model is adeptly designed to evaluate the outputs from a hybrid metaheuristic algorithm that synergizes Binary Grey Wolf Optimization (BGWO) and SA. By combining these advanced techniques, our study not only enhances predictive accuracy and FS efficacy but also fills a critical gap in the realm of BC biomarker discovery, paving the way for more precise diagnostic and therapeutic advancements.

Materials and methods

Download datasets and preprocessing

The datasets pivotal to this study, specifically GSE10810 and GSE42568, were combined to form a merged dataset. Additionally, the separate dataset GSE45827 was sourced from the NCBI GEO database, all of which are available at www.ncbi.nlm.nih.gov/geo. Corresponding clinical data for these distinct GEO datasets were procured in tandem with the gene expression profiles. The merged dataset comprises a collective total of 179 BC samples and 44 samples earmarked as normal, combining the GSE10810 dataset, which spans 58 samples (31 BC and 27 normal), with the GSE42568 dataset, encompassing 121 samples (104 BC and 17 normal). Separately, the GSE45827 dataset includes 155 samples, segregated into 144 BC samples and a subset of 11 normal samples.

Data preprocessing is pivotal in the realm of microarray data mining. Proper preprocessing not only refines the raw data but also ensures that the resulting data serve as a

valid and optimal input for subsequent modeling. Raw datasets were sourced from the GEO repository using the 'GEOquery' package. Subsequent preprocessing, including background correction, normalization, and gene expression calculation, was undertaken using the RMA algorithm.

Furthermore, given that not all probes manifest across all samples, those unexpressed across the entirety of samples were excised. Initial identification was conducted using probe IDs, and gene symbols were subsequently appended based on these IDs. It was noted that certain probe IDs could not be mapped to any gene symbol, warning of their removal. Conversely, instances were observed where multiple probe IDs corresponded to a singular gene symbol. In such situations, to maintain a unified gene expression value, the mean of the expression values of probe IDs sharing the same gene symbol was computed and deemed the final expression value.

All datasets exceed the sample size of 50, ensuring substantial data for rigorous analysis. A consistent platform, GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array), was selected for all three datasets to minimize discrepancies from different platforms and maintain uniformity in gene set examination. Data ingestion was performed using the `ReadAffy` function from the `affy` package, followed by normalization using the RMA technique.

Merge datasets

Merging gene expression datasets, especially when they originate from different sources, presents challenges due to the introduction of nonbiological variations, often referred to as batch effects. In this study, the integration of the two datasets was accomplished using the `cbind` function in R. However, merging inherently brings forth the aforementioned batch effects.

Adjustments were carried out using the empirical Bayes algorithm to address and rectify these batch effects. This algorithm has been proficiently implemented in the `ComBat` function, which is part of the `SVA` package in R. The efficacy of the batch-effect correction achieved by the `ComBat` transformation was ascertained through Principal Component Analysis (PCA). Specifically, the `prcomp` function in R was employed to conduct the PCA, which endeavors to encapsulate high-dimensional data into its first two principal components. This dimensional reduction facilitates visualization, for which the `ggbiplot` package in R was utilized.

After the rigorous phases of preprocessing, batch-effect elimination, and validation, the culmination is a unified dataset. This merged dataset is composed of 179 samples and a comprehensive 10,629 features. Throughout the remainder of this research endeavor, this dataset will be referenced as the "merged dataset".

Differential gene expression analysis

Differential Gene Expression (DGE) analysis was pivotal in identifying the differentially expressed genes (DEGs) within the datasets. The analysis was carried out using the renowned `limma` package in R. The core principle behind the DGE analysis is to

pinpoint genes that manifest considerable alterations in their expression levels between contrasting groups, such as diseased versus healthy.

For the purpose of this research, the determination of DEGs was guided by rigorous statistical criteria. A gene was earmarked as differentially expressed if its log-fold change ($\log FC$) exceeded 2, indicating significant upregulation or downregulation of its expression. Furthermore, the significance of this differential expression was validated with a stringent p -value threshold set at less than 0.05. Only genes satisfying both of these criteria were classified as DEGs in both the merged and GSE45827 datasets.

Feature selection

As datasets grow in size and complexity, ensuring the relevance of features becomes crucial for the efficient application of ML algorithms [45]. Unnecessary features can not only slow down algorithms but can also compromise their accuracy [46]. Proper FS can significantly enhance algorithm speed and performance [24]. In this study, we focused on identifying a subset of genes from the merged and GSE45827 datasets that are most pertinent to BC prediction. The efficacy of our results hinges on the judicious selection of these genes.

To achieve this, we utilized a fusion of the GWO [47] and SA techniques, employing the BGWO_SA_Ens method. This method stands out due to its incorporation of an ensemble method in its objective function. Essentially, within the BGWO_SA_Ens framework, our goal is to determine a set of genes that maximize the classification accuracy of samples. Further insights into this method can be found in section “BGWO_SA_Ens feature selection algorithm” We also experimented with BGWO_Ens, GA_Ens, LASSO, MCFS_IFS, and mRMR_IFS [48] feature selection methods to compare their results with BGWO_SA_Ens, as detailed in section “BGWO_Ens, GA_Ens, and LASSO feature selection algorithms”.

BGWO_SA_Ens feature selection algorithm

BGWO

The Grey Wolf Algorithm takes inspiration from how wolves hunt. There are four main types of grey wolves: alpha (α), beta (β), delta (δ), and omega (ω) wolves. Grey wolves live in packs and adhere to strict social rules that rank them in the wolf hierarchy. Alpha wolves lead the pack and make decisions. Betas provide assistance to the alphas in making decisions. While deltas obey alphas and betas, they dominate omega wolves. In an effort to reach their prey, wolves attempt to update their position by following the top three wolves. The wolf tries to reach the prey position in the most efficient way possible using this algorithm.

In this algorithm, the location of prey represents the optimal solution, while the location of each wolf represents a candidate solution. In terms of grey wolf solutions, α is the best, β is the second best, δ is the third best, and ω represents the rest. After the wolves are positioned randomly, their fitness values are calculated. Then, all wolves must update their position relative to those of the top three wolves. Next, the algorithm repeatedly

updates the positions of the wolves. This is done by considering the top three performing wolves at each step. During this updating process, the best-performing wolf’s position and its performance score, termed the fitness value, are recorded in every cycle. The main goal of this algorithm is, by the end, to pinpoint the best position that is closest to the prey.

At iteration t , \vec{X}_α , \vec{X}_β , and \vec{X}_δ are the three best solutions. Equation (1) defines \vec{A}_1 , \vec{A}_2 , and \vec{A}_3 as coefficient vectors.

$$\vec{A} = 2a \times \vec{r}_1 - a \tag{1}$$

In Eq. (1), vector \vec{r}_1 is generated by random numbers between 0 and 1. The distance control factor a is used to balance the tradeoff between exploration and exploitation, beginning at a high value equal to 2 and decreasing linearly until it reaches 0. The factor a is calculated using Eq. (2), where t represents the current iteration and t_{max} represents the total number of optimization iterations.

$$a = 2 - t \times \frac{2}{t_{max}} \tag{2}$$

Equations (3), (4), and (5) determine \vec{D}_α , \vec{D}_β , and \vec{D}_δ as the distance vectors between α , β , δ and i .

$$\vec{D}_\alpha = |\vec{C}_1 \times \vec{X}_\alpha - \vec{X}_i| \tag{3}$$

$$\vec{D}_\beta = |\vec{C}_2 \times \vec{X}_\beta - \vec{X}_i| \tag{4}$$

$$\vec{D}_\delta = |\vec{C}_3 \times \vec{X}_\delta - \vec{X}_i| \tag{5}$$

Equation (6) defines \vec{C}_1 , \vec{C}_2 , and \vec{C}_3 as coefficient vectors, where vector \vec{r}_2 is generated by random numbers between 0 and 1.

$$\vec{C} = 2 \times \vec{r}_2 \tag{6}$$

The GWO locates wolves’ positions in continuous space, whereas the BGWO locates them in a hypercube search space with 0 or 1. We have to set some equations to update the positions of the wolves to move them closer or further away from the hypercube.

We can compute s_1^d , s_2^d , and s_3^d as the continuous step size values using Sigmoid function by Eqs. (7), (8), and (9). Here, d is the dimension of the search space.

$$s_1^d = \frac{1}{1 + \exp(-10(A^d \times D_\alpha^d - 0.5))} \tag{7}$$

$$s_2^d = \frac{1}{1 + \exp(-10(A^d \times D_\beta^d - 0.5))} \quad (8)$$

$$s_3^d = \frac{1}{1 + \exp(-10(A^d \times D_\delta^d - 0.5))} \quad (9)$$

A binary step is represented by Eqs. (10), (11), and (12), where $r3$ is a random number between 0 and 1.

$$bstep_1^d = \begin{cases} 1 & \text{if } s_1^d \geq r3 \\ 0 & \text{else} \end{cases} \quad (10)$$

$$bstep_2^d = \begin{cases} 1 & \text{if } s_2^d \geq r3 \\ 0 & \text{else} \end{cases} \quad (11)$$

$$bstep_3^d = \begin{cases} 1 & \text{if } s_3^d \geq r3 \\ 0 & \text{else} \end{cases} \quad (12)$$

The distances that i will move relative to α , β , and δ are known as $bstep_1$, $bstep_2$, and $bstep_3$. X_1 , X_2 , and X_3 , which are calculated in Eqs. (13), (14), and (15), are binary vectors affected by the movement of α , β and δ wolves, respectively.

$$X_1^d = \begin{cases} 1 & \text{if } X_\alpha^d + bstep_1^d \geq 1 \\ 0 & \text{else} \end{cases} \quad (13)$$

$$X_2^d = \begin{cases} 1 & \text{if } X_\beta^d + bstep_2^d \geq 1 \\ 0 & \text{else} \end{cases} \quad (14)$$

$$X_3^d = \begin{cases} 1 & \text{if } X_\delta^d + bstep_3^d \geq 1 \\ 0 & \text{else} \end{cases} \quad (15)$$

After obtaining X_1^d , X_2^d , and X_3^d , the position of X_i in the next iteration ($t + 1$) is updated using a simple stochastic crossover as Eq. (16). Where $r4$ represents a random number chosen from the uniform distribution $\in [0, 1]$

$$X_i^d(t + 1) = \begin{cases} X_1^d & \text{if } r4 < \frac{1}{3} \\ X_2^d & \text{elseif } \frac{1}{3} \leq r4 < \frac{2}{3} \\ X_3^d & \text{else} \end{cases} \quad (16)$$

Algorithm 1 Discrete Grey Wolf Optimizer (BGWO)

```

1: Initialize step;
2: Initialize  $a$ ,  $A$ , and  $C$  ;
3: Initialize a population of  $n = 50$  grey wolves positions randomly;
4: Compute the fitness value for each grey wolf;
5: Select the best three grey wolves as  $X_\alpha$ ,  $X_\beta$  and  $X_\delta$ ;
6:  $t \leftarrow 0$ ;
7:  $t_{max} \leftarrow 100$ ;
8: while  $t \leq t_{max}$  do
9:   Update  $a$ ,  $A$ , and  $C$  by Eqs. (2), (1), and (6);
10:  Update the position of each grey wolf by equations (7-16);
11:  Calculate the fitness value of each grey wolf;
12:  Update  $X_\alpha$ ,  $X_\beta$  and  $X_\delta$ ;
13:   $T_0 \leftarrow 30$ ,  $S_I \leftarrow X_\alpha$ ,  $S_{Best} \leftarrow X_\alpha$ , and  $T_F \leftarrow 0.01$ ;
14:  while  $T_0 > T_F$  do
15:    Create a new solution  $S_n$  around  $S_I$  using the roulette wheel operator;
16:    Calculate the fitness value of  $S_n$ ;
17:     $\Delta f \leftarrow f(S_n) - f(S_I)$ ;
18:    if  $\Delta f > 0$  then
19:       $S_{Best} \leftarrow S_n$ ;
20:    else
21:       $P \leftarrow \exp\left(\frac{-\Delta f}{T}\right)$ ;
22:       $r_1 \leftarrow$  a random number between 0 and 1;
23:      if  $P > r_1$  then
24:         $S_{Best} \leftarrow S_n$ ;
25:      end if
26:    end if
27:     $T_0 \leftarrow T_0 \times 0.5$ ;
28:  end while
29:   $X_\alpha \leftarrow S_{Best}$ ;
30:   $t \leftarrow t + 1$ ;
31: end while
32: Return  $X_\alpha$ ;

```

SA

SA is an optimization algorithm based on annealing that mimics how the material is heated and slowly cooled to achieve a uniform structure [49]. The algorithm generates an initial random solution. Then, in subsequent iterations, a new neighborhood solution is generated and evaluated using the fitness function. Whenever a new solution is better, it is accepted, whereas when it is the worst, it is accepted based on a Boltzmann probability.

$$P = \exp\left(\frac{-\Delta f}{T}\right) \quad (17)$$

In Eq. (17), Δf represents the difference between the previous and new solutions' fitness values, and T represents the temperature parameter.

At the end of each iteration of the grey wolf algorithm, we enter the value of X_α as the initial value into the SA algorithm and select a neighborhood solution using the roulette wheel strategy. To generate the neighborhood solution, the roulette wheel operator chooses one

of the Swap, Insertion, Inversion, R2L (Rotate to Left), or R2R (Rotate to Right) operators. After a neighborhood is created, its fitness value is measured and compared with the fitness value of the input X_α . If we conclude that the neighbor solution is accepted due to the higher fitness value or by using the Boltzmann probability, we will replace it with the initial value. Otherwise, we will create another neighborhood until we reach the predetermined temperature. Therefore, we go to the next step of the grey wolf and use the new value of X_α . This process is shown in pseudocode 1.

Objective function

Our methodology incorporates an ensemble learning approach, utilizing four distinct ML algorithms, XGBoost (clf_1), SVM (clf_2), RF (clf_3), and DT (clf_4), to determine the fitness of potential solutions.

We commence by partitioning our merged dataset into three subsets: 60% for training, 20% for validation, and the remaining 20% for testing. The process initiates with selecting the features present in the potential solution, followed by training the ensemble using 60% of the samples from the training set. We implemented Synthetic Minority Over-sampling Technique (SMOTE) on the training set to address class imbalance, while the validation and test sets were kept in their original, untouched state.

Posttraining, we employ the validation set to assess the performance of the base classifiers. This evaluation paves the way for the computation of their respective weights, followed by a weighted voting procedure to derive the results of the ensemble strategy. This entire workflow is meticulously repeated as necessary to compute the fitness value for varying solutions, as guided by Eqs. (18) and (19). In addition, the final 20% of the dataset, designated as the testing set, is utilized for gauging the model's performance and facilitating comparative analyses.

$$Avg_{Eclf} = \frac{Ac_{Eclf} + Pr_{Eclf} + Re_{Eclf} + F1_{Eclf} + F2_{Eclf}}{5} \quad (18)$$

$$Fit(Sol) = (\varphi \times Avg_{Eclf}) + \left((1 - \varphi) \times \frac{NF}{AF} \right) \quad (19)$$

In the above equations, the fitness value of a solution, denoted as $Fit(Sol)$, is computed by averaging the evaluation metrics of the Ensemble classifier ($Eclf$), with $\varphi = 0.8$ representing the significance attached to the number of selected features. Here, NF is the number of selected features, and AF is the total number of features in the dataset.

The symbols Ac , Pr , Re , $F1$, and $F2$ represent the evaluation metrics: Accuracy, Precision, Recall, F1 score, and F2 score, respectively. These metrics are detailed in Eqs. (20), (21), (22), (23), and (24).

$$Ac = \frac{TP + TN}{TP + FP + FN + TN} \quad (20)$$

$$Pr = \frac{TP}{TP + FP} \quad (21)$$

$$Re = \frac{TP}{TP + FN} \quad (22)$$

$$F1 = 2 \times \left(\frac{Pr \times Re}{Pr + Re} \right) \quad (23)$$

$$F2 = \frac{5 \times Pr \times Re}{(4 \times Pr) + Re} \quad (24)$$

Given that cancer samples are considered positive and normal samples are considered negative, the values of TP, TN, FP, and FN in the above equations are defined as follows:

- *True Positive (TP)* Correctly predicted cancer instances.
- *True Negative (TN)* Correctly predicted normal instances.
- *False Positive (FP)* Normal instances wrongly predicted as cancer.
- *False Negative (FN)* Cancer instances wrongly predicted as normal.

To implement weighted voting, we calculate a weight for each base classifier post-training, utilizing the validation set, as depicted in Eq. 25.

$$fw_{clf_i} = \frac{Ac_{clf_i} \times Pr_{clf_i} \times Re_{clf_i} \times F1_{clf_i} \times F2_{clf_i}}{\sum_{k=1}^4 (Ac_{clf_k} \times Pr_{clf_k} \times Re_{clf_k} \times F1_{clf_k} \times F2_{clf_k})} \quad (25)$$

In the weighted voting procedure, we amalgamate the results from each base classifier employing their respective weights, fw_{clf_i} , which ensures that classifiers demonstrating superior performance on the validation set exercise a more substantial influence on the ensemble's final prediction.

The ensemble classifier's prediction for each sample can be expressed as:

$$Pred_{Eclf}(x) = \arg \max_{c \in C} \sum_{i=1}^4 fw_{clf_i} \times Pred_{clf_i}(x) \quad (26)$$

In Eq. (26), $Pred_{Eclf}(x)$ is the class predicted for sample x by the ensemble classifier, C is the set of possible class labels, and $Pred_{clf_i}(x)$ is the class predicted for sample x by base classifier clf_i .

The performance metrics Ac_{Eclf} , Pr_{Eclf} , Re_{Eclf} , $F1_{Eclf}$, and $F2_{Eclf}$ of the ensemble classifier are then computed using the predictions from the ensemble classifier $Eclf$, as defined in Eqs. (20), (21), (22), (23), and (24).

Importantly, the ensemble's performance typically surpasses the performance of its individual components, especially when the base classifiers are diverse. This is because the ensemble leverages each classifier's strengths while mitigating their weaknesses.

BGWO_Ens, GA_Ens, and LASSO feature selection algorithms

BGWO_Ens

BGWO_Ens is akin to BGWO_SA_Ens, with the primary distinction being the absence of the SA process. It harnesses the BGWO algorithm, which is a variant of the GWO tailored for binary-encoded problems, making it suitable for FS. After determining the

optimal subset of features using BGWO, these selected features are fed into an ensemble of classifiers, structured similarly to the ensemble method delineated in section “[Objective function](#)”. By integrating the search capabilities of BGWO with the predictive power of ensemble models, this method aims to achieve high predictive accuracy while minimizing overfitting, particularly in high-dimensional datasets.

GA_Ens

The GA_Ens method synergizes GA with ensemble learning. Genetic Algorithms are heuristic search algorithms inspired by natural selection. Potential feature subsets are encoded as chromosomes, which evolve over generations to identify the most effective subsets. Once the optimal features are selected using GA, they are then used to train an ensemble of ML models, following the ensemble structure detailed in section “[Objective function](#)”. The melding of GA’s evolutionary search with ensemble modeling offers a robust method that adeptly handles FS and prediction tasks.

LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression-based method incorporating L1 regularization. This regularization can push some feature coefficients to be precisely zero, allowing LASSO to inherently perform FS by excluding features with zero coefficients. Unlike BGWO_Ens and GA_Ens, LASSO does not naturally integrate an ensemble approach. However, its prowess lies in its ability to manage multicollinearity, select a feature subset while concurrently performing regression, and offer ease of interpretation.

To remove the repeated information and streamline the content in the two sections, here is a revised version:

MCFS-IFS

In this study, the Monte Carlo Feature Selection (MCFS) method [48] was employed for feature selection from gene expression datasets. The datasets were prepared and divided into training and test sets, with the training set balanced using the SMOTE. The MCFS process, conducted over 100 iterations, involved selecting and evaluating a random subset of 5000 features from the training dataset in each iteration. These features were assessed using a RF Classifier and a 10-fold cross-validation approach. The key aspect was scoring each feature based on its presence in high-performing subsets, thereby determining its importance in predicting the target variable. Post-iteration, the top 5000 features were ranked according to their scores for further analysis.

Incremental Feature Selection (IFS) [50] was then applied to refine this feature set, involving iterative evaluation of feature subsets, beginning with the top 5 features and increasing in size by 5 features at each step. A range of classifiers, including DT, KNN, RF, SVM, and NN, were used to assess the performance of each subset using 10-fold cross-validation. Various metrics such as F1 score, PR-AUC, ROC-AUC, Mathews correlation coefficient (MCC), and Balanced Accuracy (BAc) where MCC and BAc can be computed by Eqs. (27) and (28) were employed for a comprehensive assessment. The subset that achieved the highest F1 score was identified as the optimal feature set and was further tested on a separate test set to validate its generalizability and effectiveness.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (27)$$

$$BAc = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (28)$$

mRMR-IFS

Similarly, the mRMR (Minimum redundancy maximum relevance) feature selection method [51] was integrated with IFS to enhance the robustness and accuracy of our gene expression data analysis in BC research. Using the ‘pymrmr’ Python library, the mRMR algorithm selected the top 5000 features from over 10,000 potential genes based on their mutual information quotient with the target variable. This selection ensured that each feature contributed uniquely to the predictive model.

The subsequent IFS process evaluated these mRMR-selected features in smaller subsets to determine the most impactful set for predictive modeling. Starting with the top 5 features and incrementally increasing by 5 in each iteration, the process assessed feature subsets up to the top 5000. Performance evaluation through a 10-fold cross-validation with classifiers like DT, KNN, RF, and SVM, along with metrics including F1 score, ROC-AUC, PR-AUC, MCC, and BAc, helped identify the best feature set. This rigorous approach pinpointed the most influential genes for breast cancer detection and diagnosis, significantly contributing to the advancement of breast cancer biomarker discovery.

Superior genes selection

In the intricate landscape of genomics, different analytical methods can yield varying gene sets of interest. Our approach involved conducting Differential Expression (DE) analysis and the BGWO_SA_Ens feature selection method as two independent processes across two distinct datasets. While DE analysis identifies genes that manifest a significant change in expression levels between conditions, the BGWO_SA_Ens feature selection is applied to assess that the genes are also impactful in predictive modeling. As a result, the intersection of these methods (two sets from DE analysis and two sets from BGWO_SA_Ens across the datasets) provides us with a set of “superior genes” that are both differentially expressed and integral in ML predictions. This dual analysis approach allows for a comprehensive and unbiased identification of key genes, ensuring a robust selection of candidates for further biological investigation.

To further elucidate the roles and implications of these superior genes, three advanced analyses were undertaken. First, an enrichment analysis was performed to uncover the underlying biological pathways and processes in which these genes are predominantly involved, providing insights into their functional significance. Subsequently, network analysis furnished a clearer picture of the interrelations and interactions among these genes, offering a systems biology perspective on their collective behavior and potential regulatory patterns. Last, by investigating potential drug interactions, we probed the therapeutic avenues and implications these genes might open up, potentially paving the way for targeted treatments or pharmacological interventions. Collectively, these

analyses not only amplify our understanding of the identified superior genes but also fortify their relevance in both research and therapeutic contexts.

Enrichment analysis of superior genes

Enrichment analysis deciphers the functional roles of superior genes highlighted through DE analysis and ML. By statistically assessing the association between these genes and specific diseases, the analysis furnishes gene-set profiles. This finding sheds light on the involvement of these genes in cancer progression. KEGG pathway enrichment analysis and GO [52, 53] further enhance this understanding, offering classifications such as molecular functions (MF), biological processes (BP), and cellular components (CC).

PPI network analysis

Analyzing protein-protein interactions (PPI) bridges the knowledge gap between superior genes and cellular functionality [54]. Utilizing the STRING database [55], we mapped interactions with a confidence threshold of 0.4, visualized through Cytoscape 3.9.1. Key regions in this PPI network were pinpointed using the MCODE plugin, while the cytoHubba plugin highlighted influential genes within the network. This holistic approach unravels the intricacies of gene interactions, potentially directing us towards novel therapeutic avenues.

Exploration of drug interactions with superior genes

Understanding drug-gene dynamics is pivotal for crafting effective BC treatments. Researchers can sculpt more individualized, effective therapeutic strategies by studying the interplay between drugs and disease-related genes. The DGIdb database [56] served as a resource for unearthing drugs that potentially interact with our superior genes. Visual representations, crafted using Cytoscape, further clarify these interactions, propelling insights into disease biology and therapeutic innovation.

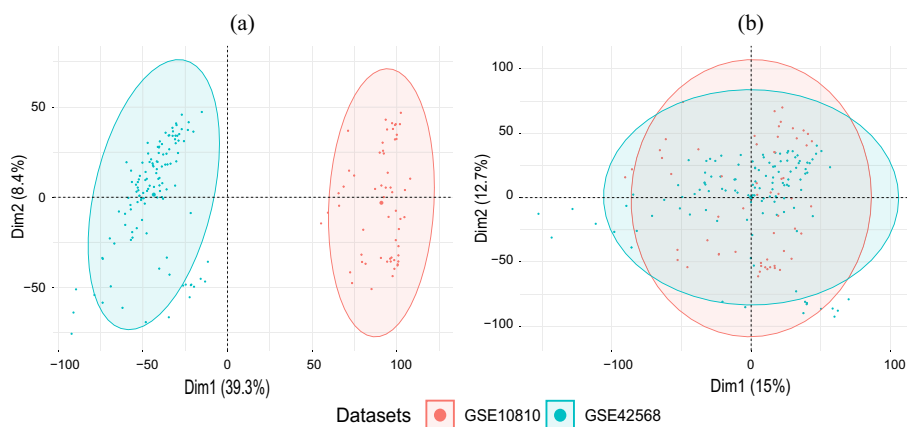


Fig. 1 PCA visualization of the merged dataset **a** before and **b** after batch effect removal using the ComBat function

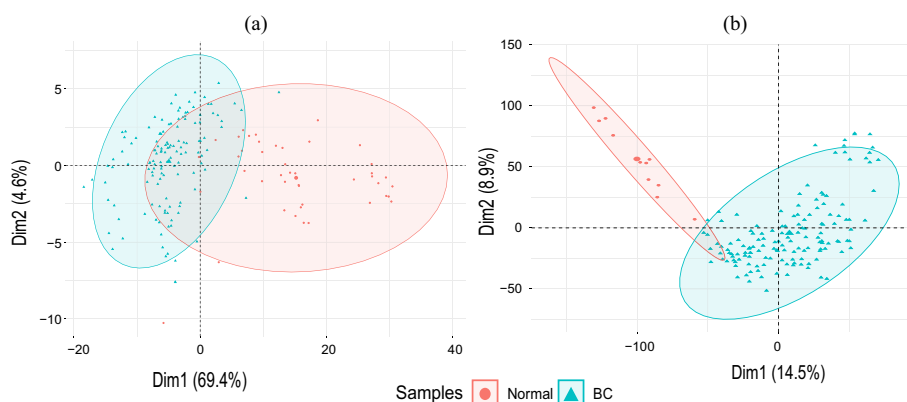


Fig. 2 PCA plots of BC and normal samples for **a** merged dataset and **b** GSE45827 dataset

Results

Data preprocessing and merging

Following the meticulous preprocessing of the GSE10810, GSE42568, and GSE45827 datasets, refined gene expression matrices were obtained for each of these datasets. This processing consisted of 179 samples in the merged dataset and 155 samples in the GSE45827 dataset, encompassing 135 BC samples and 44 normal samples in the merged dataset, and 144 BC samples with 11 normal samples in the GSE45827 dataset. The preprocessing steps resulted in a comprehensive set of 10,629 genes for analysis in the merged dataset, and 11,731 genes in the GSE45827 dataset.

After the preprocessing phase, a PCA plot was generated to visualize the batch effects inherent in the merged dataset. The evident clustering of samples from different datasets illustrated the presence of these batch effects as shown in Fig. 1a. However, after applying the empirical Bayes algorithm through the ComBat function, a subsequent PCA plot showcased the mitigation of these batch effects. The samples

Table 1 Quantitative distribution of differentially expressed genes, upregulated and downregulated entities

Dataset	Total genes	DEGs	Upregulated	Downregulated
Merged Dataset	10629	164	34	130
GSE 45827	11731	350	208	142

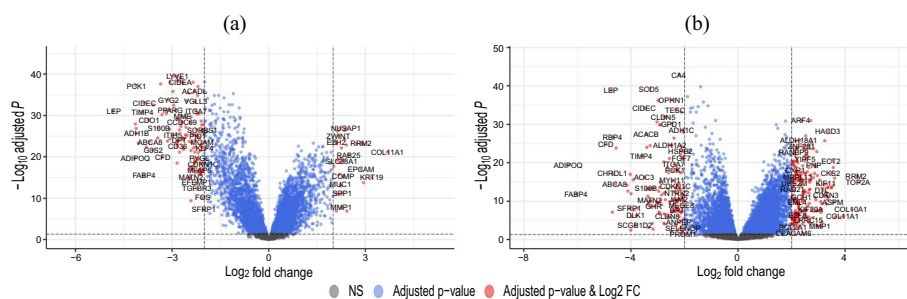


Fig. 3 Graphical Representation of DEGs via Volcano Plot for **a** merged dataset and **b** GSE45827 dataset



Fig. 4 Clustering heatmap for the merged dataset distinguishing BC from normal profiles

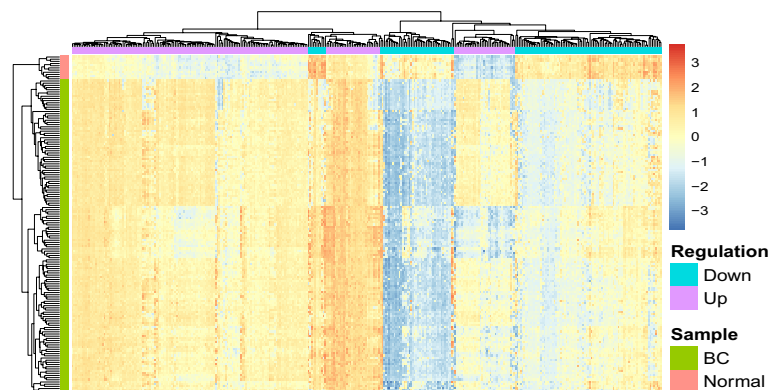


Fig. 5 Clustering heatmap for the GSE45827 dataset distinguishing BC from normal profiles

were no longer clustered based on their source dataset, as shown in Fig. 1b, which indicated the successful removal of the batch effects.

Additionally, PCA plots elucidating the clustering of samples were constructed as displayed in Fig. 2. BC samples distinctly grouped together, apart from the normal samples, revealing the inherent differences in their gene expression patterns. This demarcation in the PCA space affirms the significance of the gene expression features in distinguishing between BC and normal samples in both the merged dataset 2a and the GSE45827 dataset 2b.

Differentially expressed genes identification

From the comprehensive analysis of the 10,629 genes in the merged dataset, 164 genes were discerned as differentially expressed. These were delineated based on the following criteria: Adjusted p -value < 0.05 and $|\log_2 \text{foldchange}(FC)| > 2$. Consequently, from the identified DEGs, 34 genes exhibited upregulation while 130 genes were found to be downregulated. Additionally, analysis of the GSE45827 dataset revealed 350 DEGs among 11731 genes, with 208 upregulated and 142 downregulated genes. Table 1 summarizes the count of genes after DEG analysis for both datasets.

Table 2 Performance metrics for each feature set obtained using BGWO_SA_Ens in Merged Dataset and GSE45827

Fset	Dataset	NF	FV ¹	F1	PR-AUC	ROC-AUC
Fset1	Merged Dataset	1610	0.059	0.942	0.984	0.953
Fset2	Merged Dataset	2027	0.063	0.936	0.981	0.946
Fset3	Merged Dataset	1864	0.059	0.931	0.964	0.873
Fset4	Merged Dataset	3055	0.072	0.791	0.981	0.955
Fset5	Merged Dataset	1900	0.058	0.931	0.963	0.872
Fset6	Merged Dataset	3108	0.064	0.794	0.978	0.938
Fset7 ²	Merged Dataset	1404	0.044	0.981	0.998	0.995
Fset8	Merged Dataset	3255	0.061	0.791	0.967	0.879
Fset9	Merged Dataset	1763	0.051	0.769	0.958	0.857
Fset10	Merged Dataset	3134	0.061	0.791	0.980	0.948
Fset1	GSE45827	3018	0.059	0.910	0.978	0.956
Fset2 ²	GSE45827	1710	0.046	0.965	0.986	0.972
Fset3	GSE45827	2100	0.061	0.914	0.968	0.939
Fset4	GSE45827	3914	0.063	0.879	0.969	0.954
Fset5	GSE45827	1636	0.051	0.901	0.976	0.959
Fset6	GSE45827	2126	0.047	0.931	0.977	0.956
Fset7	GSE45827	3222	0.061	0.910	0.973	0.948
Fset8	GSE45827	2219	0.047	0.936	0.978	0.956
Fset9	GSE45827	2670	0.049	0.958	0.977	0.955
Fset10	GSE45827	3899	0.060	0.892	0.973	0.947

¹ Fitness Value

² Best performing feature set for the dataset

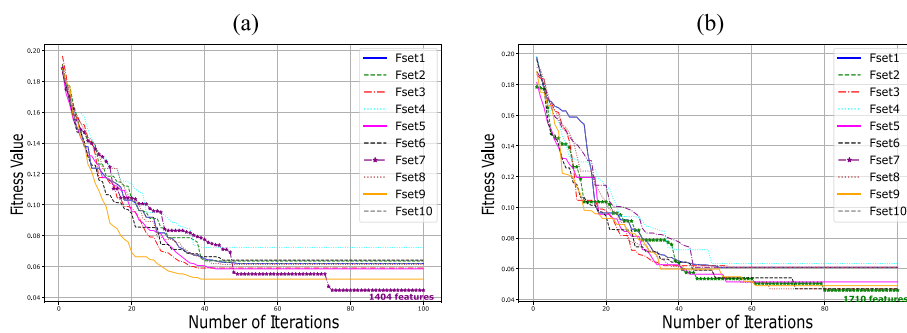


Fig. 6 Comparative Analysis of Feature Selection Convergence: **a** Merged Dataset and **b** GSE45827 Dataset using BGWO_SA_Ens Method

These identified DEGs significantly contributed to the statistical demarcation between the BC and normal samples within the merged dataset, as well as the GSE 45827 dataset. Figure 3 illustrates the volcano plot of these DEGs. The plot’s X-axis signifies the fold change, while the Y-axis represents the \log_{10} transformed Adjusted p -value for every gene. The DEGs are prominently marked in red and annotated accordingly.

Furthermore, heatmaps were generated to provide a visual representation of the distinction between samples based on their group classification as illustrated in Figs. 4 and 5. The hierarchical clustering depicted in each heatmap efficiently segregated the BC samples from the normal samples, emphasizing the effectiveness of the selected gene

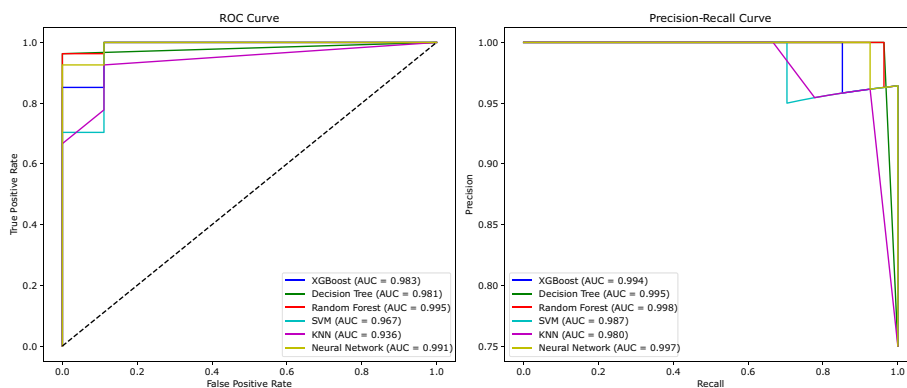


Fig. 7 ROC-AUC and PR-AUC performance of Fset7 in the Merged Dataset

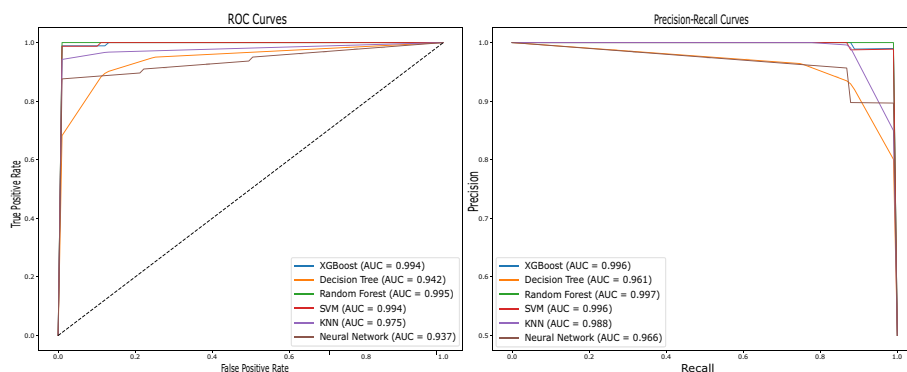


Fig. 8 ROC-AUC and PR-AUC performance of Fset2 in the GSE45827 Dataset

expression features in discriminating between the two groups. Notably, the gradient in the heatmaps, marked by distinct color patterns for BC and normal samples, underlines the pronounced differences in their gene expression profiles for both the merged dataset and the GSE45827 dataset.

Feature selection results

BGWO_SA_Ens performance

We executed the BGWO_SA_Ens method ten times, yielding ten different feature sets, designated Fset1 through Fset10, for each dataset. Each of these feature sets was then used to test six base classifiers, XGBoost, DT, RF, SVM, KNN, and NN, using a test set that only contained features selected by BGWO_SA_Ens. This was carried out for all ten selected feature sets for each dataset. Subsequently, we computed the F1, PR-AUC, and ROC-AUC evaluation metrics to assess the performance of the selected features. The results, reflecting the mean values derived from the six base classifier methods, are presented below.

From Table 2, it is evident that Fset7 exhibited superior performance compared to the other sets in the Merged Dataset, and Fset2 was the best performer in the GSE45827 dataset, as indicated by the rows in bold text. These feature sets encompassed a selection of 1404 genes for the merged dataset and 1710 genes for the GSE45827

Table 3 Comparison of the performance metrics of different feature selection methods

Algorithm	Dataset	NF	F1	PR-AUC	ROC-AUC	MCC	BAc
All Features	Merged Dataset	10629	0.862	0.870	0.858	0.814	0.905
	GSE45827	11731	0.873	.978	0.961	0.895	0.795
BGWO_SA_Ens	Merged Dataset	1404	0.984	0.986	0.977	0.967	0.977
	GSE45827	1710	0.965	0.986	0.972	0.953	0.965
BGWO_Ens	Merged Dataset	1941	0.898	0.972	0.965	0.839	0.916
	GSE45827	2114	0.865	0.965	0.931	0.829	0.914
GA_Ens	Merged Dataset	2557	0.954	0.986	0.977	0.943	0.965
	GSE45827	2488	0.908	0.986	0.972	0.894	0.947
LASSO	Merged Dataset	10	0.885	0.874	0.852	0.831	0.913
	GSE45827	8	0.729	0.904	0.872	0.571	0.772
MCFS-IFS	Merged Dataset	1850	0.940	0.979	0.973	0.945	0.898
	GSE45827	475	0.951	0.975	0.963	0.930	0.866
mRMR-IFS	Merged Dataset	355	0.945	0.971	0.954	0.916	0.953
	GSE45827	1660	0.926	0.940	0.964	0.897	0.934

dataset. Figures 6a and 6b display the optimization plots for the merged dataset and the GSE45827 dataset, respectively. In Fig. 6a, it is demonstrated that Fset7 attained a better fitness value (FV) than other feature sets at iteration 100 for the merged dataset. Similarly, Fig. 6b shows that Fset2 outperformed other feature sets at iteration 100 in the GSE45827 dataset. Additionally, Figures 7 and 8 illustrate the PR-AUC and ROC-AUC performance of Fset7 in the merged dataset and Fset2 in the GSE45827 dataset, respectively.

Comparison to other methods (BGWO_Ens, GA_Ens, LASSO, MCFS-IFS, mRMR-IFS)

In our comprehensive analysis, we conducted a comparative evaluation of various FS techniques, including BGWO_Ens, GA_Ens, LASSO, MCFS-IFS, mRMR-IFS, applied to both the merged dataset and the GSE45827 dataset for breast cancer biomarker identification. These FS methods yielded significantly different sets of genes, with BGWO_Ens identifying 1941 in the merged dataset and 2114 in the GSE45827 dataset, GA_Ens 2557 in the merged dataset and 2488 in the GSE45827 dataset, LASSO 10 in the merged dataset and 8 in the GSE45827 dataset, MCFS-IFS 1850 in the merged dataset and 475 in the GSE45827 dataset, and mRMR-IFS 355 in the merged dataset and 1660 in the GSE45827 dataset, indicating a wide variance in feature selection granularity.

To assess the performance and effectiveness of these FS strategies, we incorporated the selected features within our test dataset. We proceeded to employ six distinct base classifiers, XGBoost, DT, RF, SVM, KNN, and NN, to evaluate the predictive power of the features identified by each FS method. Classifiers were rigorously tested using 10-fold cross-validation to ensure stability and reliability in the feature evaluation process. The final feature set was determined for each FS technique, comparable to the best feature set (Fset) identified by our BGWO_SA_Ens method.

The results are meticulously summarized in Table 3, presenting a side-by-side comparison of the most effective outcomes derived from the five FS techniques. In this table, the bold text highlights the best performing algorithm that most accurately predict

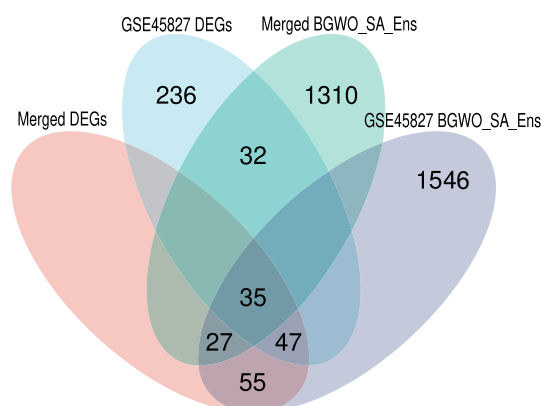


Fig. 9 Venn diagram illustrating the intersection of DEGs and BGWO_SA_Ens selected genes. The overlapping region denotes the identified superior genes. Venn diagram illustrating the intersection of DEGs and BGWO_SA_Ens selected genes from the merged and GSE45827 datasets. The overlapping region denotes the identified superior genes

BC, as evidenced by the metrics given. This comparative analysis extends to the results obtained from the features selected by BGWO_SA_Ens and contrasts them against the performance when the full spectrum of features is employed. Notably, the ROC and Precision-Recall curves, as illustrated in the Figs. 7 and 8, provide a visual and quantitative representation of the classifiers’ performance, with AUC scores serving as a benchmark for comparison.

The Random Forest classifier, for instance, achieved AUC scores of 0.995 in ROC and 0.998 in Precision-Recall metrics for the merged dataset, and similarly, 0.995 in ROC and 0.998 in Precision-Recall metrics for the GSE45827 dataset. These results indicate exceptional classification accuracy. Such consistency across different datasets highlights the potential of our hybrid BGWO_SA_Ens approach in identifying the most predictive biomarkers for BC, potentially paving the way for more targeted and effective therapeutic strategies.

Superior genes identification

In our analysis across two distinct datasets, we identified different sets of differentially expressed genes (DEGs) and selected genes using the BGWO_SA_Ens algorithm. For the merged dataset, a total of 164 genes were identified as DEGs, and the BGWO_SA_Ens algorithm selected 1404 genes. In contrast, for the GSE45827 dataset, 350 genes were identified as DEGs, with the BGWO_SA_Ens algorithm selecting 1710 genes. An intersection of these four sets (DEGs and BGWO_SA_Ens selected genes from both datasets) revealed a commonality of 35 genes. These overlapping genes, which are present in both the DEGs and the BGWO_SA_Ens selected genes from each dataset, are referred to as the “superior genes.” This subset represents genes that are consistently significant across different analytical methods and datasets, highlighting their potential importance in the context of breast cancer.

The relationship and overlap between the DEGs and the genes selected through the BGWO_SA_Ens algorithm across both datasets are schematically represented in Fig. 9,

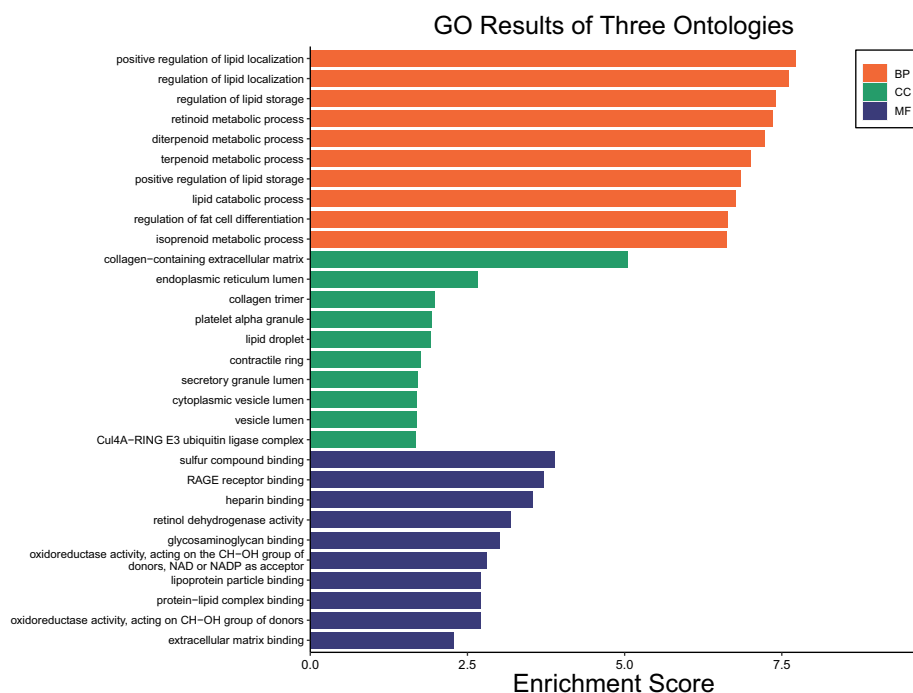


Fig. 10 GO enrichment analysis results categorized into MF, BP, and CC. The EnrichmentScore demonstrates the importance of each category in relation to the 35 superior genes. Distinct color codings for MF, BP, and CC allow for easy differentiation and visualization of the results

a Venn diagram. This visualization emphasizes the significance of the identified superior genes and provides a clear depiction of their derivation from the intersection of the two methodologies and datasets.

Enrichment analysis results

The comprehensive enrichment analysis was segmented into two pivotal sections, focusing on GO enrichment and KEGG pathway enrichment respectively.

GO enrichment

Using GO analysis, the molecular characteristics of the 35 superior genes were delineated across three ontological categories. In terms of MF, the most significant functionalities identified were *sulfur compound binding*, *RAGE receptor binding*, and *heparin binding*, each possessing a noteworthy enrichment score. Within the BP category, processes such as *positive regulation of lipid localization*, *regulation of lipid localization*, and *regulation of lipid storage* were prominently enriched. In the context of CCs, the key enriched components comprised *collagen containing extracellular matrix*, *endoplasmic reticulum lumen*, and *collagen trimer*. The significance of these functionalities, processes, and components is underscored by their respective Enrichment Scores, highlighting their potential pivotal roles in the physiological manifestations observed. The comprehensive visual representation of these GO enrichments is illustrated in Fig. 10.

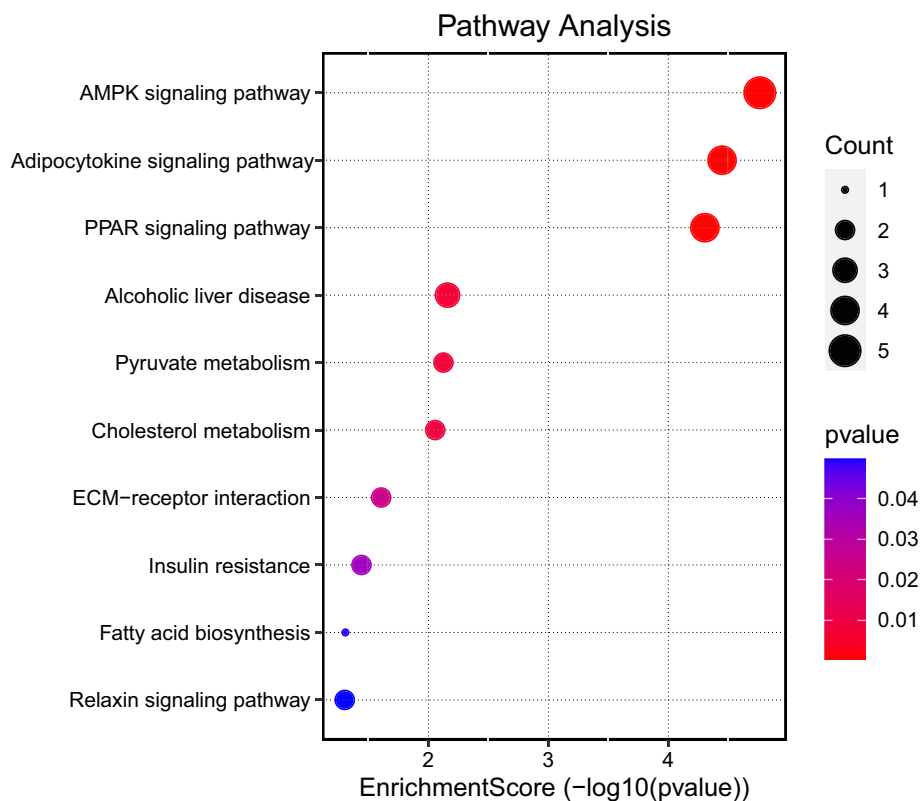


Fig. 11 KEGG pathway analysis displaying the enrichment score of various pathways related to the 35 superior genes. The pathways shown are ranked based on their significance and relevance. A higher EnrichmentScore indicates greater significance. The count and *p*-value further provide insight into the number of genes involved and the statistical significance of each pathway, respectively

KEGG pathway enrichment

The KEGG analysis spotlighted several pathways potentially intertwined with the biological activities of the superior genes. Preeminent among these pathways were the *AMPK signaling pathway*, *Adipocytokine signaling pathway*, and *PPAR signaling pathway*. The *AMPK signaling pathway*, with the highest EnrichmentScore, is particularly pivotal, suggesting a potentially significant involvement in the processes under investigation. Other pathways such as *Alcoholic liver disease* and *Pyruvate metabolism* also exhibited consequential EnrichmentScores, indicative of their importance in the broader molecular schema. A comprehensive visual overview of the KEGG pathway enrichments is presented in Fig. 11.

Analysis of protein-protein interaction networks

Within this section, we probe the nuanced realm of protein-protein interactions (PPI) pertaining to our designated cohort of superior genes. Our objective is to elucidate the concealed links and underscores pivotal subnetworks that might elucidate BC at a molecular level.

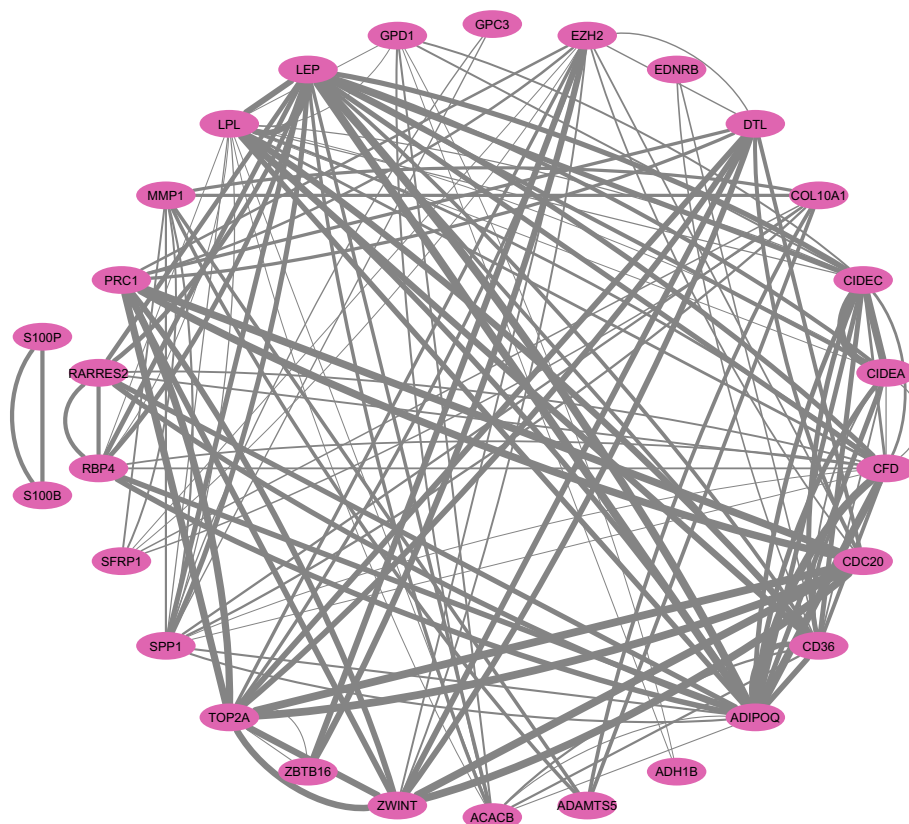


Fig. 12 Integrated Protein-Protein Interaction (PPI) network schematic of 28 superior genes, representing a subset of the total 35 identified superior genes with known interactions. The thickness of the lines between nodes corresponds to their combined score

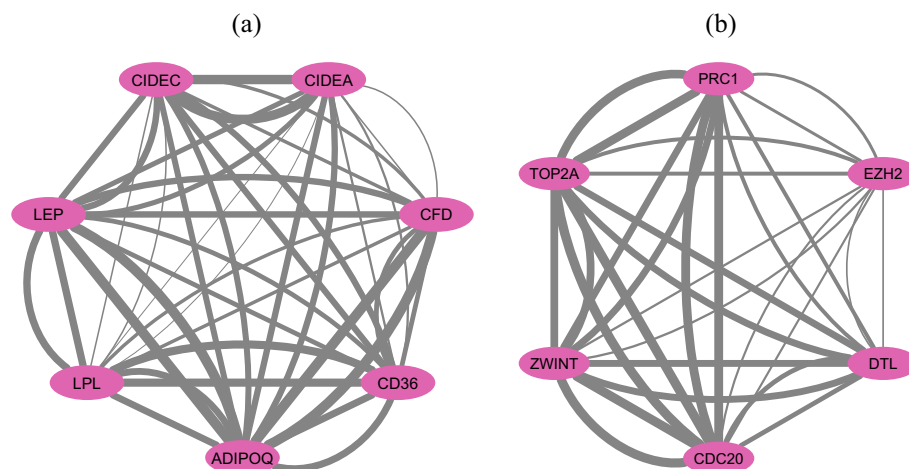


Fig. 13 Principal subnetworks: **a** MCODE metric of 7 **b** MCODE metric of 6. The thickness of the lines between nodes corresponds to their combined score

As illustrated in Fig. 12, the PPI network manifests as a complex matrix of interactions encompassing the delineated superior genes. Individual nodes in this construct signify the superior genes, and the connecting edges depict the probable affiliations

and mutual interactions between them. This all-encompassing illustration provides a succinct perspective on the molecular dynamics inherent to BC.

Embedded within this multifaceted PPI network, the MCODE tool has identified regions with high connectivity, illustrated as subnetworks in Figure 13. Such subnetworks typically signify proteins that synergistically partake in analogous biological functionalities or pathways. They possess the capability to unveil seminal insights into BC’s molecular mechanics.

The inaugural subnetwork, depicted in Fig. 13a, is conferred an MCODE metric of 7, insinuating elevated interconnectivity among its constituent proteins. This conglomerate potentially typifies a cadre of proteins synergistically operating within a distinct cellular trajectory, thereby hinting at their implications in BC molecular dynamics.

The sequential subnetwork, represented in Fig. 13b, while being less expansive and having diminished connectivity in comparison to the former, remain of paramount relevance. The second registers an MCODE metric of 6, insinuating its representation of separate functional assemblies within BC’s multifarious molecular puzzle.

To discern the paramount nodes within these subnetworks, our approach was augmented by the cytoHubba tool. Nodes characterized by elevated centrality metrics occupy cardinal positions within the construct, potentially modulating its overarching framework and operational dynamics. Such quintessential proteins might exert pivotal functions in the pathogenesis, rendering them prospective foci for therapeutic endeavors.

These revelations furnish profound insights into the intricate molecular orchestration linked to BC. By emphasizing these superior genes and their interrelationships, potential avenues for therapeutic strategies may surface. Further experimental verification of these networks and their pivotal nodes will strengthen their relevance in BC and could accelerate the development of innovative therapeutic approaches.

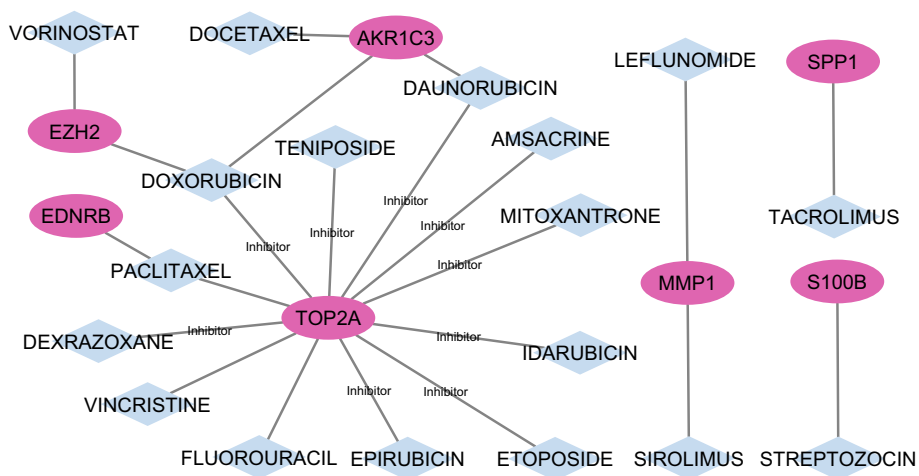


Fig. 14 Graphical elucidation of drug-gene dynamics, highlighting the multifaceted therapeutic frontier

Investigation of drug-gene interactions in oncology

Delineating the nuanced interactions between pharmacological agents and genetic elements in BC is crucial for the evolution of therapeutic modalities. The discernment and systematic study of these drug-gene interactions pave the way for the formulation of targeted therapeutic regimens, enhancement of treatment precision, reduction of adverse reactions, and the optimization of patient prognosis. The prominence of genes in oncogenic predisposition and disease trajectory emphasizes the imperative of comprehending the molecular dynamics between drugs and the specific genes implicated in oncogenesis.

For this study, we leveraged the DGIdb (v4.2.0) database, an exhaustive repository encompassing drug-gene associations and potential drug target genes. This database integrates data spanning diverse origins, including peer-reviewed literature, solidifying its authority on drug-gene synergies.

To decipher and represent the aforementioned drug-gene associations, we employed Cytoscape, a sophisticated tool dedicated to network synthesis and visualization. Figure 14 showcases the derived network, with pharmacological agents delineated as blue nodes and genetic entities as pink nodes. This graphical abstraction elucidates the dense meshwork of interactions between pivotal genes and their corresponding pharmacological modulators. Importantly, this networked representation imparts a nuanced view of associations that might remain obscured in a mere tabulated format.

To encapsulate, the drug-gene interconnection illustrated in Fig. 14 provides invaluable insight into the labyrinthine domain of drug-gene dynamics. Such a perspective augments our comprehension of prospective therapeutic trajectories and propels the formulation of bespoke therapeutic solutions for BC, epitomizing the potential of precision medicine within the realm of oncology.

Discussion

The genetic intricacies of BC call for advanced methods in biomarker discovery. Confronting the high-dimensionality of gene expression data, our BGWO_SA_Ens algorithm effectively discerns crucial biomarkers indicative of BC. We addressed the constraints of sample size by merging datasets, thus broadening the scope of our meta-analysis.

The incorporation of ensemble learning, which leverages multiple classifiers, is pivotal to the BGWO_SA_Ens algorithm's success in honing gene subset selection. The adeptness of ensemble learning in managing complex data bolsters the feature selection strength of BGWO_SA_Ens. It results in more accurate predictions and versatile model applications. Moreover, the combination of BGWO and SA's FS capabilities ensures robust and reliable results.

While BGWO_SA_Ens, BGWO_Ens, and GA_Ens represent advanced metaheuristic algorithms that leverage the strengths of ML to find optimal solutions, their comparison with methods like LASSO, MCFS-IFS, and mRMR-IFS offers a broader perspective on FS approaches. LASSO, known for its simplicity and effectiveness in feature reduction, tends to select a smaller set of features but may oversimplify complex biological data, as indicated by its lower performance metrics in our study. On the other hand, MCFS-IFS and mRMR-IFS, while offering more comprehensive feature selection, may not always provide the optimal balance between feature reduction

and predictive accuracy. In contrast, BGWO_SA_Ens excels in identifying a moderate yet highly effective gene set, balancing feature reduction with high predictive performance. This underscores the advantage of BGWO_SA_Ens in handling high-dimensional, complex datasets typical in biological research. However, it's worth noting that the increased complexity and computational demand of BGWO_SA_Ens could be seen as a drawback compared to more straightforward methods like LASSO. Overall, our findings suggest that while simpler methods have their merits, especially in less complex datasets, the sophisticated nature of BGWO_SA_Ens offers significant advantages in navigating the intricacies of genomic data for breast cancer research.

In the comparative analysis of different FS methods, the BGWO_SA_Ens algorithm demonstrated superior performance in terms of F1, PR-AUC, ROC-AUC, MCC, and BAC compared to BGWO_Ens, GA_Ens, LASSO, MCFS-IFS, and mRMR-IFS. BGWO_Ens and GA_Ens, despite selecting larger sets of features, did not match the performance of BGWO_SA_Ens. Additionally, the inclusion of MCFS-IFS and mRMR-IFS in the evaluation also provided a broader comparison basis. Notably, BGWO_SA_Ens excelled with an F1 score of 0.984, PR-AUC of 0.986, ROC-AUC of 0.977, MCC of 0.967, and BAC of 0.977 for the merged dataset, and similar high scores for GSE45827, as shown in Table 3. In contrast, LASSO's performance was the lowest, and while MCFS-IFS and mRMR-IFS had varying results, they did not outperform BGWO_SA_Ens. This comparison underscores BGWO_SA_Ens's ability to select a gene set that is not only moderate in size but also highly effective, as shown by the robust performance metrics across both datasets.

In our study, through meticulous methodology, the BGWO_SA_Ens algorithm selected a set of 1404 genes from the merged dataset and 1710 from the GSE45827 dataset. In parallel, we identified 164 differentially expressed genes (DEGs) from the merged dataset and 350 from the GSE45827 dataset. Crucially, the intersection of these four feature sets yielded 35 genes that were common to both datasets, which we designated as superior genes. This subset represents a highly significant group, as these genes are validated both through differential expression analysis and the advanced FS capabilities of the BGWO_SA_Ens algorithm. This intersection reinforces the validity of the selected genes and underscores the robustness of our approach in identifying key biomarkers for BC.

Additionally, our exploration of GO and KEGG enrichment analyses has revealed significant insights into the biological relevance of genes identified by our FS method. In the realm of biological processes, the enrichment results highlighted several key areas, with the "positive regulation of lipid localization" pathway being particularly notable. This pathway includes CD36, ADIPOQ, ACACB, CIDEA, LPL, and SPP1 genes, underscoring their potential role in BC biology. Similarly, in cellular components, the "collagen-containing extracellular matrix" component was identified as significant. This component features COL10A1, ITIH5, ADIPOQ, ADAMTS5, GPC3, SFRP1, and RARRES2 genes, pointing to their involvement in cancer-related cellular structures. In Molecular Functions, "sulfur compound binding" was determined as significant which implicating PCOLCE2, ADAMTS5, ACACB, SFRP1, and LPL genes.

Our KEGG pathway analysis further illuminated the importance of the "AMPK signaling pathway" in BC, with a notable association of genes such as CD36, ADIPOQ,

ACACB, CIDEA, and LEP. This pathway is critical in regulating metabolic processes, including lipid metabolism and energy homeostasis, and has been linked to cancer development and progression [40]. The association of these genes with the AMPK signaling pathway emphasizes their potential role in modulating tumor growth, metabolism, and response to therapy in breast cancer biology.

In our comprehensive analysis, we identified several key genes that are potentially pivotal in the pathogenesis and progression of BC. Among them, CIDEA (Cell Death-Inducing DFFA-Like Effector A) is downregulated in breast tumors, suggesting its significant role in lipid metabolism and energy balance with implications for cancer cell survival and proliferation [57]. LEP (Leptin), identified as a superior downregulated gene in differential expression analysis, is recognized for its influence on energy regulation and a notable role in tumorigenesis, especially in BC. It contributes to cell proliferation and angiogenesis, key factors in cancer progression [58].

Another significant gene, acetyl-CoA carboxylase beta (ACACB), which is crucial for fatty acid oxidation and is downregulated in BC, is associated with increased survival and reduced drug resistance, highlighting its potential as a target for inhibiting tumor cell proliferation and metabolic reprogramming [59]. The Lipoprotein Lipase (LPL) gene stands out for its role in lipid metabolism and its potential involvement in providing fatty acids to cancer cells, thereby supporting their growth and survival [60]. ADIPOQ (Adiponectin), primarily known for its role in glucose regulation and fatty acid breakdown, has been observed to have anti-proliferative effects on BC cells, suggesting a protective role [61].

Our research has unveiled a range of additional genes with promising roles in BC, as determined through GO and KEGG enrichment analyses. These include RARRES2 (Retinoic Acid Receptor Responder 2) [62], AKR1C3 (Aldo-Keto Reductase Family 1 Member C3) [57], SPP1 (Secreted Phosphoprotein 1) [63], CIDEA (Cell Death-Inducing DFFA-Like Effector C) [64], CD36 (Cluster of Differentiation 36) [59], and MMP1 (Matrix Metalloproteinase 1) [65]. These genes contribute to the intricate molecular framework of BC, offering new avenues for exploration and potential therapeutic intervention.

Our findings underscore the crucial role of lipid metabolism disruptions in BC, particularly through the AMPK signaling pathway, indicating their importance in BC's development and progression. The genes involved in these processes offer insights into BC's mechanisms and pave the way for new therapeutic and biomarker development strategies. The comprehensive integration of GO and KEGG enrichment analyses with gene functions highlights BC's complexity, advocating for a holistic understanding and targeted approach to this disease.

The detailed PPI network analysis has revealed key genes and functional modules, enhancing our understanding of BC's intricate molecular interactions. This analysis, illustrated in figures like 12 and subnetworks in 13, identifies genes involved in lipid metabolism, cell cycle regulation, and EMT. These findings not only uncover underlying protein interactions but also suggest potential biomarkers and therapeutic targets, forming a foundation for future research in unraveling BC's molecular complexities and advancing precision medicine.

Our study's focus on drug-gene interactions has shed light on potential BC treatments, particularly highlighting topoisomerase II alpha (TOP2A) [66] inhibitors and their role in BC's development and progression. Drugs such as vincristine, teniposide, mitoxantrone, etoposide, epirubicin, idarubicin, daunorubicin, amsacrine, paclitaxel, dexrazoxane, doxorubicin, and fluorouracil all target TOP2A, indicating its crucial role in either the development or progression of BC. This examination reveals complex interplays between genes and drugs, suggesting pathways for therapeutic exploitation and strategies to hinder cancer progression. It points towards a personalized treatment approach in BC, targeting specific genes with multiple drugs to tailor treatments. Overall, this study demonstrates the effectiveness of computational techniques in identifying significant BC biomarkers, emphasizing the need for further experimental validation and the potential to extend these methods to other cancers for precision medicine advancements.

Conclusion

In conclusion, this study demonstrates the efficacy of an innovative ML workflow in identifying robust biomarkers and therapeutic targets for BC. By synergizing metaheuristic optimization, ensemble learning, differential expression analysis, and network biology, a panel of 35 superior genes was revealed. These genes were not only differentially expressed in BC but also integral to predictive modeling, affirming their significance.

Collectively, this multipronged informatics approach overcomes key challenges in biomarker discovery. It enhances model accuracy, generalizability, and biological relevance compared to prevailing techniques. The findings propel novel possibilities for diagnostic, prognostic and therapeutic innovation in BC. Moving forward, experimental validation of the identified genes and drug targets will further cement this study's contributions to precision oncology.

The framework developed is widely applicable for mining biomarkers from multifaceted biomedical data. Furthermore, the proposed BGWO_SA_Ens algorithm sets a new standard for efficient FS in ML pipelines. By synergizing the strengths of optimization, ensembles, and differential expression, more refined and biologically insightful gene signatures can be derived. This advances personalized medicine across diverse diseases. In summary, this research underscores the translational potential of informatics in augmenting biomarker discovery and augmenting our comprehension of complex diseases.

Author contributions

MR was primarily responsible for composing the main text of the manuscript and preparing the figures and dataset. MF supervised the project team and reviewed both the methodology and the structure of the manuscript. MR and MF conducted the experiments. RS reviewed the datasets, results, and text, providing medical consultation to confirm the findings. AHG reviewed and revised both the text and methodology. FB reviewed the text and methodology. All authors read and approved the final version.

Funding

Not applicable.

Availability of data and materials

The datasets examined in this study can be found in the GEO repository, under the NCBI GEO accession numbers GSE42568 [67] and GSE10810 [68]. All codes associated with this paper are available at https://github.com/mornejd/Refining-Breast-Cancer-Biomarker-Discovery-BGWO_SA_Ens. Also, for any additional inquiries contact the corresponding author, Mohammad Fathian (fathian@iust.ac.ir).

Declarations

Ethics approval and consent to participate

The ethic approval is not required since we used publicly available datasets.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 3 December 2023 Accepted: 15 January 2024

Published online: 22 January 2024

References

1. WHO: Breast cancer. <https://www.who.int/news-room/fact-sheets/detail/breast-cancer> Accessed 2023-07-12, 2023.
2. Ferroni P, Zanzotto FM, Riondino S, Scarpato N, Guadagni F, Roselli M. Breast cancer prognosis using a machine learning approach. *Cancers*. 2019;11(3):328.
3. Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Front Genet*. 2019;10:256.
4. Osareh A, Shadgar B. Machine learning techniques to diagnose breast cancer. In: 2010 5th International Symposium on Health Informatics and Bioinformatics, pp. 114–120 2010. IEEE
5. Thalor A, Joon HK, 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.
6. Sugimoto M, Hikichi S, Takada M, Toi M. Machine learning techniques for breast cancer diagnosis and treatment: a narrative review. *Annals Breast Surg*. 2023.
7. Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: a review. *Bioengineering*. 2023;10(2):173.
8. Foltz SM, Greene CS, Taroni JN. Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously. *Commun Biol*. 2023;6(1):222.
9. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, et al. NCBI GEO: Archive for functional genomics data sets—10 years on. *Nucleic Acids Res*. 2010;39(suppl-1):1005–10.
10. Sahu A, Qazi S, Raza K, Singh A, Verma S. Machine learning-based approach for early diagnosis of breast cancer using biomarkers and gene expression profiles. In: *Computational Intelligence in Oncology: Applications in Diagnosis, Prognosis and Therapeutics of Cancers*, pp. 285–306. Springer, Singapore 2022. https://doi.org/10.1007/978-981-16-9221-5_17
11. 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 Develop Biol*. 2022;9: 811585.
12. Mazlan AU, Sahabudin NA, Remli MA, Ismail NSN, Mohamad MS, Nies HW, Abd Warif NB. A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data. *Processes*. 2021;9(8):1466.
13. Peng Y, Li W, Liu Y. A hybrid approach for biomarker discovery from microarray gene expression data for cancer classification. *Cancer Inf*. 2006;2:117693510600200030.
14. Mohamed TI, Ezugwu AE, Fonou-Dombeu JV, Ikotun AM, Mohammed M. A bio-inspired convolution neural network architecture for automatic breast cancer detection and classification using rna-seq gene expression data. *Sci Rep*. 2023;13(1):14644.
15. Majid A, Ali S, Iqbal M, Kausar N. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Comput Methods Programs Biomed*. 2014;113(3):792–808.
16. Taghizadeh E, Heydarheydari S, Saberi A, JafarpourNesheli S, Rezaeijo SM. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinf*. 2022;23(1):1–9.
17. Kim B-C, Kim J, Lim I, Kim DH, Lim SM, Woo S-K. Machine learning model for lymph node metastasis prediction in breast cancer using random forest algorithm and mitochondrial metabolism hub genes. *Appl Sci*. 2021;11(7):2897.
18. Mathema VB, Sen P, Lamichhane S, Orešič M, Khoomrung S. Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. *Computat Struct Biotechnol J* 2023.
19. Nasser M, Yusof UK. Deep learning based methods for breast cancer diagnosis: a systematic review and future direction. *Diagnostics*. 2023;13(1):161.
20. Alrefai N, Ibrahim O. Optimized feature selection method using particle swarm intelligence with ensemble learning for cancer classification based on microarray datasets. *Neural Comput Appl*. 2022;34(16):13513–28.
21. Mirsadeghi L, Haji Hosseini R, Banaei-Moghaddam AM, Kavousi K. Earn: an ensemble machine learning algorithm to predict driver genes in metastatic breast cancer. *BMC Med Genom*. 2021;14(1):122.
22. Zheng J, Lin D, Gao Z, Wang S, He M, Fan J. Deep learning assisted efficient adaboost algorithm for breast cancer detection and early diagnosis. *IEEE Access*. 2020;8:96946–54.
23. Kenn M, Cacsire Castillo-Tong D, Singer CF, Cibena M, Kölbl H, Schreiner W. Microarray normalization revisited for reproducible breast cancer biomarkers. *BioMed Res Int* 2020.
24. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electric Eng*. 2014;40(1):16–28.
25. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.

26. Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, Fradet Y, Droit A. Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. *Front Genet.* 2019;10:452.
27. Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput Biol Med.* 2021;134: 104481.
28. Bommert A, Welchowski T, Schmid M, Rahnenführer J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings Bioinf.* 2022;23(1):354.
29. Alromema N, Syed AH, Khan T. A hybrid machine learning approach to screen optimal predictors for the classification of primary breast tumors from gene expression microarray data. *Diagnostics.* 2023;13(4):708.
30. Farid AA, Selim G, Khater H. A composite hybrid feature selection learning-based optimization of genetic algorithm for breast cancer detection. Preprints 2020.
31. Haznedar B, Arslan MT, Kalinli A. Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. *Med Biol Eng Comput.* 2021;59:497–509.
32. Kundu R, Chattopadhyay S, Cuevas E, Sarkar R. Altwoa: altruistic whale optimization algorithm for feature selection on microarray datasets. *Comput Biol Med.* 2022;144: 105349.
33. Hajieskandar A, Mohammadzadeh J, Khalilian M, Najafi A. Molecular cancer classification method on microarrays gene expression data using hybrid deep neural network and grey wolf algorithm. *J Ambient Intell Hum Comput.* 1–11 2020.
34. Alshamlan HM, Badr GH, Alohalı YA. Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification. *Comput Biol Chem.* 2015;56:49–60.
35. Sharma A, Rani R. C-hmshssa: gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. *Comput Methods Programs Biomed.* 2019;178:219–35.
36. Aziz RM. Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data. *Med Biol Eng Comput.* 2022;60(6):1627–46.
37. Wu L, Ye X, Zhang Y, Gao J, Lin Z, Sui B, Wen Y, Wu Q, Liu K, He S, et al. A genetic algorithm-based ensemble learning framework for drug combination prediction. *J Chem Inf Model* 2023.
38. Budhreja S, Dohorjeh M, Singh B, Tan S, Dohorjeh Z, Lai E, Merkin A, Lee J, Goh W, Kasabov N. Filter and wrapper stacking ensemble (FWSE): a robust approach for reliable biomarker discovery in high-dimensional omics data. *Brief Bioinf.* 2023;24(6):382.
39. Tan W, Liu M, Wang L, Guo Y, Wei C, Zhang S, Luo C, Liu N. Novel immune-related genes in the tumor microenvironment with prognostic value in breast cancer. *BMC Cancer.* 2021;21(1):1–16.
40. Liu L, Zhai W, Wang F, Yu L, Zhou F, Xiang Y, Huang S, Zheng C, Yuan Z, He Y, et al. Using machine learning to identify gene interaction networks associated with breast cancer. *BMC Cancer.* 2022;22(1):1070.
41. Elbashir MK, Mohammed M, Mwambi H, Omolo B. Identification of hub genes associated with breast cancer using integrated gene expression data with protein-protein interaction network. *Appl Sci.* 2023;13(4):2403.
42. MotieGhader H, Tabrizi-Nezhadi P, Deldar Abad Paskeh M, Baradaran B, Mokhtarzadeh A, Hashemi M, Lanjanian H, Jazayeri SM, Maleki M, Khodadadi E, et al. Drug repositioning in non-small cell lung cancer (NSCLC) using gene co-expression and drug-gene interaction networks analysis. *Sci Rep.* 2022;12(1):9417.
43. Ahmed S, Kabir M, Ali Z, Arif M, Ali F, Yu D-J. An integrated feature selection algorithm for cancer classification using gene expression data. *Comb Chem High Throughput Screen.* 2018;21(9):631–45.
44. Zhang X, Yang L, Zhang D, Wang X, Bu X, Zhang X, Cui L. Prognostic assessment capability of a five-gene signature in pancreatic cancer: a machine learning based-study. *BMC Gastroenterol.* 2023;23(1):68.
45. Punitha S, Stephan T, Gandomi AH. A novel breast cancer diagnosis scheme with intelligent feature and parameter selections. *Comput Methods Programs Biomed.* 2022;214: 106432. <https://doi.org/10.1016/j.cmpb.2021.106432>.
46. Chen JW, Dhahbi J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci Rep.* 2021;11(1):1–15.
47. Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. *Adv Eng Softw.* 2014;69:46–61.
48. Lin X, Ma Q, Chen L, Guo W, Huang Z, Huang T, Cai Y-D. Identifying genes associated with resistance to KRAS G12C inhibitors via machine learning methods. *Biochimica et Biophysica Acta (BBA)-General Subjects.* 2023;1867(12): 130484.
49. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. *Science.* 1983;220(4598):671–80.
50. Zhang Y-H, Li ZD, Zeng T, Chen L, Huang T, Cai Y-D. Screening gene signatures for clinical response subtypes of lung transplantation. *Mol Genet Genom.* 2022;297(5):1301–13.
51. Huang F, Fu M, Li J, Chen L, Feng K, Huang T, Cai Y-D. Analysis and prediction of protein stability based on interaction network, gene ontology, and KEGG pathway enrichment scores. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics.* 2023;3: 140889.
52. Kanehisa M, Goto S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28(1):27–30.
53. Consortium GO. The gene ontology (go) project in 2006. *Nucleic Acids Res.* 2006;34(suppl-1):322–6.
54. Rao VS, Srinivas K, Sujini G, Kumar G. Protein-protein interaction detection: methods and analysis. *Int J Proteom* 2014.
55. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47(D1):607–13.
56. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, Griffith M, Griffith OL, Wagner AH. Integration of the drug-gene interaction database (DGIDB 4.0) with open crowdsourcing efforts. *Nucleic Acids Res.* 2021;49(D1):1144–51.
57. Wang L, Yu L, Shi J, Li F, Zhang C, Xu H, Yin X, Wang L, Lin S, Litvinova A, et al. Functional regulations between genetic alteration-driven genes and drug target genes acting as prognostic biomarkers in breast cancer. *Sci Rep.* 2022;12(1):10641.
58. Wang Y, Du L, Jing J, Zhao X, Wang X, Hou S. Leptin and leptin receptor expression as biomarkers for breast cancer: a retrospective study. *BMC Cancer.* 2023;23(1):1–8.

59. Bai J, Zhang X, Kang X, Jin L, Wang P, Wang Z. Screening of core genes and pathways in breast cancer development via comprehensive analysis of multi gene expression datasets. *Oncol Lett.* 2019;18(6):5821–30.
60. Bavis MM, Nicholas AM, Tobin AJ, Christian SL, Brown RJ. The breast cancer microenvironment and lipoprotein lipase: another negative notch for a beneficial enzyme? *FEBS Open Bio.* 2023;13(4):586–96.
61. Pasha HF, Mohamed RH, Toam MM, Yehia AM. Genetic and epigenetic modifications of adiponectin gene: potential association with breast cancer risk. *J Gene Med.* 2019;21(10):3120.
62. Li YQ, Sun FZ, Li CX, Mo HN, Zhou YT, Lv D, Zhai JT, Qian HL, Ma F. *Rarres2* regulates lipid metabolic reprogramming to mediate the development of brain metastasis in triple negative breast cancer. *Mil Med Res.* 2023;10(1):34.
63. Zeng P, Zhang X, Xiang T, Ling Z, Lin C, Diao H. Secreted phosphoprotein 1 as a potential prognostic and immunotherapy biomarker in multiple human cancers. *Bioengineered.* 2022;13(2):3221–39.
64. Dogan S, Leopold J, Hoffmann D, Kubitschke H, Blauth E, Ficorella C, Zschau A, Schiller J, Käs J. Identification of lipid droplet-associated genes in breast cancer patients. *Res Square* 2022. <https://doi.org/10.21203/rs.3.rs-2157661/v1>.
65. Jafarpour M, Moeini A, Maryami N, Nahvijou A, Mohammadian A. Classification of potential breast/colorectal cancer cases using machine learning methods. *Int J Cancer Manage.* 2023;16(1): 135724. <https://doi.org/10.5812/ijcm-135724>.
66. Wu Y, Han Y, Li Q, Zhang P, Yuan P, Luo Y, Fan Y, Chen S, Cai R, Li Q, et al. Predictive value of topoisomerase ii alpha protein for clinicopathological characteristics and prognosis in early breast cancer. *Breast Cancer Res Treat.* 2022;193(2):381–92.
67. Clarke C, Madden SF, Doolan P, Aherne ST, Joyce H, O'driscoll L, Gallagher WM, Hennessy BT, Moriarty M, Crown J, et al. Correlating transcriptional networks to breast cancer survival: a large-scale co-expression analysis. *Carcinogenesis.* 2013;34(10):2300–8.
68. Pedraza V, Gomez-Capilla JA, Escaramis G, Gomez C, Torné P, Rivera JM, Gil A, Araque P, Olea N, Estivill X, et al. Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. *Cancer Interdiscip Int J Am Cancer Soc.* 2010;116(2):486–96.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.