scientific reports

Check for updates

OPEN

SGA-Driven feature selection and random forest classification for enhanced breast cancer diagnosis: A comparative study

Abrar Yaqoob[1], Navneet Kumar Verma[1], Mushtaq Ahmad Mir[4], Ghanshyam G. Tejani[2,3], Nashwa Hassan Babiker Eisa[4], Hind Mamoun Hussien Osman[4] & Mohd Asif Shah[5,6]✉

In this study, we propose a novel approach for breast cancer classification that integrates the Seagull Optimization Algorithm (SGA) for feature selection with the Random Forest (RF) classifier for effective data classification. The novelty of our approach lies in the first-time application of SGA for gene selection in breast cancer diagnosis, where SGA systematically explores the feature space to identify the most informative gene subsets, thereby improving classification accuracy and reducing computational complexity. The selected features are subsequently classified using RF, known for its robustness and high accuracy in handling complex datasets. To evaluate the effectiveness of the proposed method, we compared it with other classifiers, including Linear Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). The proposed SGA-RF combination achieved a best mean accuracy of 99.01% with 22 genes, outperforming other methods and demonstrating consistent performance across varying feature subsets. The mean accuracies ranged from 85.35 to 94.33%, highlighting a balance between feature reduction and classification accuracy. Future work will explore the integration of other nature-inspired algorithms and deep learning models to further enhance performance and clinical applicability.

Keywords Seagull optimization algorithm, Random forest, Cancer classification, High dimensional data

Cancer remains one of the leading causes of mortality worldwide, posing significant challenges to public health systems. Characterized by the uncontrolled proliferation of abnormal cells, it affects millions of individuals across diverse populations, regardless of age, gender, or ethnicity[1–4]. Early diagnosis and effective classification of cancer types are critical for improving survival rates and guiding personalized treatment strategies. Advances in biomedical research have highlighted the importance of high-dimensional datasets, such as gene expression profiles, in uncovering potential biomarkers for cancer diagnosis and prognosis. However, the complexity and redundancy inherent in these datasets necessitate robust computational approaches for feature selection and classification to enhance diagnostic accuracy and clinical outcomes[5,6].

Breast cancer, in particular, remains one of the most prevalent and life-threatening diseases affecting millions of women worldwide. The increasing incidence rate underscores the urgent need for precise diagnostic methods to ensure early detection and effective treatment. Despite advances in medical imaging and biomarker discovery, gene expression analysis remains central to understanding the molecular basis of breast cancer and enabling precision medicine[7]. However, gene expression datasets often contain thousands of genes, many of which are redundant or irrelevant to the classification task. This high dimensionality can lead to overfitting, increased computational cost, and reduced model interpretability. Therefore, effective feature selection methods are essential to identify a subset of biologically meaningful genes that contribute most to the classification task, improving classification performance and providing insights into the underlying biology of breast cancer[8].

[1]VIT Bhopal University's School of Advanced Science and Language, Located at Kothrikalan, Sehore, Bhopal 466114, India. [2]Department of Research Analytics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai 600077, India. [3]Department of Industrial Engineering and Management, Yuan Ze University, Taoyuan 320315, Taiwan. [4]Department of Clinical Laboratory Sciences, College of Applied Medical Sciences, King Khalid University, Abha 61421, Saudi Arabia. [5]Department of Economics, Kardan University, Parwane Du, Kabul 1001, Afghanistan. [6]Division of Research and Development, Lovely Professional University, Phagwara, Punjab144001, India. ✉email: m.asif@kardan.edu.af

Nature-inspired optimization algorithms have shown significant promise in addressing the challenges posed by high-dimensional data. These methods mimic natural processes such as evolution, foraging, and social interactions to identify optimal solutions in complex search spaces. Leveraging these algorithms for feature selection in breast cancer classification represents a promising avenue for enhancing diagnostic accuracy and computational efficiency[9].

Motivation.

The motivation for this study stems from the following key challenges and opportunities:

- **Addressing Redundancy**: Gene expression datasets often contain redundant and irrelevant features, which complicates the identification of crucial biomarkers. This research aims to develop effective methods for selecting informative genes, thereby improving classification accuracy and model interpretability[10].
- **Advancing Cancer Classification**: Accurate classification of tumor characteristics is essential for effective diagnosis and treatment planning. The proposed approach seeks to improve classification accuracy, ultimately contributing to better patient outcomes[11].
- **Leveraging Nature-Inspired Algorithms**: Nature-inspired algorithms, such as the Artificial Bee Colony (ABC) and Differential Evolution (DE), have demonstrated success in complex optimization problems. Their application in gene selection offers the potential to identify optimal gene subsets more effectively[12].
- **Enhancing Computational Efficiency**: By reducing the number of selected features without compromising classification accuracy, the proposed method aims to enhance computational efficiency, making it more feasible for large-scale applications[13].
- **Exploring Interdisciplinary Applications**: The integration of optimization, machine learning, and bioinformatics encourages interdisciplinary collaboration, potentially leading to novel insights and methodologies.

This study proposes a novel hybrid approach that combines the Seagull Optimization Algorithm (SGA) for feature selection with the Random Forest (RF) classifier for effective data classification. By addressing the challenges of high dimensionality and redundancy, this work aims to improve the precision and reliability of breast cancer classification, ultimately contributing to the broader goals of precision medicine and improved patient care.

Background

Over the years, various feature selection methods and classifiers have been employed in breast cancer classification. Traditional feature selection techniques, such as statistical tests, principal component analysis (PCA), and minimum redundancy maximum relevance (mRMR), have provided foundational insights. However, these methods often struggle to balance relevance and redundancy, especially when dealing with non-linear relationships inherent in gene expression data. More advanced methods, such as wrapper and embedded approaches, offer improved performance but come at the cost of increased computational complexity. For example, wrapper methods evaluate feature subsets based on classifier performance, making them computationally intensive for high-dimensional datasets. Embedded methods, while more efficient, are often limited by the assumptions of the underlying model, reducing their generalizability[14].

In recent years, nature-inspired optimization algorithms have emerged as powerful tools for feature selection. These algorithms draw inspiration from biological and physical processes, such as genetic evolution (Genetic Algorithms), swarm intelligence (Particle Swarm Optimization, Ant Colony Optimization), and animal foraging behaviors (Cuckoo Search, Grey Wolf Optimizer). Such methods excel in exploring and exploiting the feature space, striking a balance between local and global search strategies to identify optimal feature subsets. Among these, the Seagull Optimization Algorithm (SGA) has gained attention for its unique approach to optimization. Inspired by the social and migratory behavior of seagulls, SGA efficiently explores the search space through a combination of random exploration and targeted exploitation. Its ability to avoid local optima and adapt to complex search landscapes makes it particularly suitable for high-dimensional feature selection tasks. However, its application to breast cancer classification remains relatively unexplored, presenting an opportunity to advance the state-of-the-art[15].

Classifier selection is another critical aspect of breast cancer classification. Random Forest (RF), a robust ensemble learning method, has proven effective in handling high-dimensional data due to its ability to aggregate multiple decision trees. RF's inherent feature importance metrics and resilience to overfitting make it an ideal choice for this study. Other classifiers, such as Support Vector Machines (SVM), Logistic Regression (LR), and K-Nearest Neighbors (KNN), have also been widely used, each offering unique strengths and limitations. A comparative analysis of these classifiers provides valuable insights into the effectiveness of the proposed method[16–18].

Contributions

The primary contribution of this study is the development of a novel approach that integrates the Seagull Optimization Algorithm (SGA) for feature selection with Random Forest (RF) for breast cancer classification. This approach addresses key challenges in handling high-dimensional gene expression data and demonstrates superior performance in terms of accuracy, computational efficiency, and feature interpretability.

1. **Novelty in Feature Selection**: The application of SGA for feature selection represents a significant advancement in the field. By leveraging the migratory and social behavior of seagulls, SGA effectively explores the feature space to identify biologically relevant genes. This algorithm balances exploration and exploitation, ensuring robust identification of optimal feature subsets while avoiding local optima. To the best of our knowledge, this is the first study to apply SGA in the context of breast cancer classification.

2. **Integration with Random Forest**: The use of Random Forest (RF) as the classification model further strengthens the proposed approach. RF's ability to handle high-dimensional data, combined with its feature importance metrics, complements the SGA's feature selection process. Together, SGA and RF provide a comprehensive framework for tackling the challenges of high-dimensional data while delivering accurate and interpretable results.
3. **Comparative Analysis**: This study conducts a detailed comparative analysis of the proposed SGA-RF method against other widely used classifiers, including Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The results demonstrate that the SGA-RF combination consistently outperforms these classifiers across various evaluation metrics, including accuracy, precision, recall, and F1-score.
4. **Empirical Validation**: The efficacy of the proposed method is validated through experiments on high-dimensional breast cancer gene expression datasets. The results highlight the method's ability to achieve high classification accuracy with a minimal number of selected features. Specifically, the study reports a best mean accuracy of 99.01% with just 22 selected genes, showcasing the efficiency and precision of the approach.
5. **Practical Implications**: The proposed method has significant implications for clinical applications. By reducing the dimensionality of gene expression data while maintaining high classification accuracy, this approach facilitates the development of cost-effective diagnostic tools. Furthermore, the identification of biologically relevant genes provides insights into the molecular mechanisms of breast cancer, potentially guiding future research and treatment strategies. Figure 1, shows various techniques to visualize breast cancer images.

Paper organisation

The paper's organization is as follows: Sect. 1 introduces the research context, highlighting challenges posed by complex gene expression data. Section 2 reviews existing literature, emphasizing gene selection importance and classification methods. Section 3 explains the methodology integration. Section 4 details the experimental setup, including dataset specifics, parameters, and evaluation metrics. Section 5 presents and discusses results, comparing our approach to baseline models. Section 6 concludes and Sect. 7 outlines potential future research directions for enhancing our approach, summarizing progress and emphasizing the hybrid approach's potential impact on cancer classification and gene expression profiling.

Related work

In this section, we review the existing literature on breast cancer classification and diagnosis, focusing on recent advancements in deep learning and machine learning methodologies. We explore various techniques and models that have been proposed to enhance the accuracy and efficiency of breast cancer detection, including feature extraction methods, ensemble learning strategies, and the application of transfer learning. By synthesizing these studies, we aim to highlight the progress made in the field and identify potential gaps that our research seeks to address. Hanan et al. present a computer-aided diagnosis method for breast cancer classification using deep neural networks (ResNet 18, ShuffleNet, Inception-V3Net) and transfer learning. The method achieves binary classification accuracies of 99.7%, 97.66%, and 96.94%, and multi-class accuracies of 97.81%, 96.07%, and 95.79%, respectively[19]. Mahmoud Ragab et al. developed the EDLCDS-BCDC for breast cancer diagnosis using ultrasound images. The method preprocesses images, segments them with the Chaotic Krill Herd Algorithm, extracts features using VGG-16, VGG-19, and SqueezeNet, and classifies them with Cat Swarm Optimization and a Multilayer Perceptron. Simulations demonstrate its superior performance over recent methods[20]. Sahu et al. introduced a predictive model that integrates an artificial intelligence-based learning approach with a multivariate statistical method. Data mining plays a crucial role in automating the diagnostic process, especially when dealing with noisy datasets from various repositories. Their study proposes a hybrid feature selection

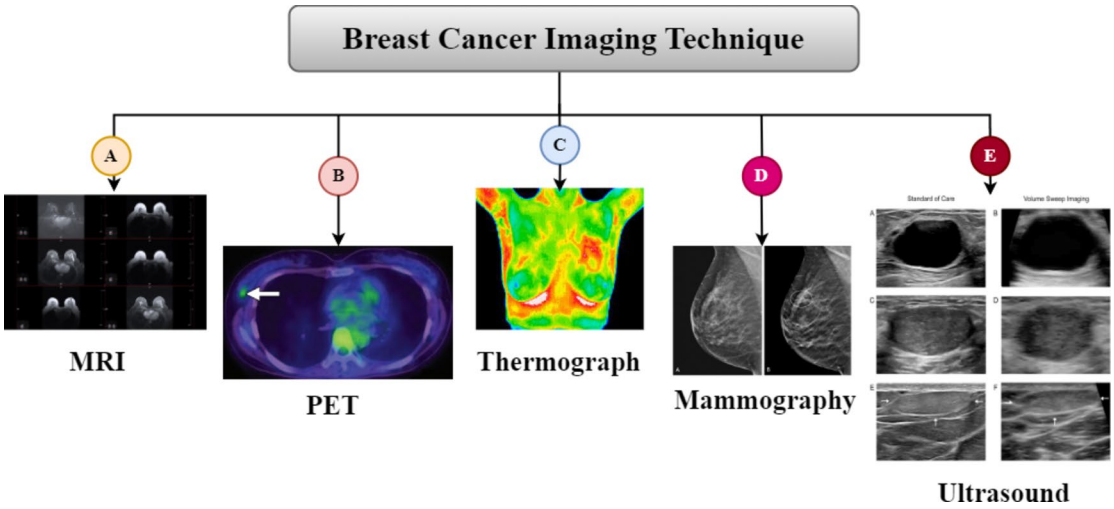

Fig. 1. Different techniques to visualize breast cancer Images.

technique that leverages Principal Component Analysis (PCA) for preprocessing and extracting the most relevant features, followed by classification using an Artificial Neural Network (ANN). The effectiveness of the proposed method was evaluated on the Wisconsin Breast Cancer Dataset from the UCI Machine Learning Repository, employing 10-fold cross-validation during the classification phase[21]. Nonita Sharma et al. propose a snapshot ensembling technique to create an efficient model for disease diagnosis. Using t-SNE for enhanced scatter plots, the model integrates predictions from various base models to improve accuracy. Applied to the Wisconsin Breast Cancer Dataset (WBCD), it achieved 86.6% accuracy, outperforming state-of-the-art models like averaging (81%) and stacked ensemble (84.7%), showcasing its effectiveness[22]. Jiande Wu et al. developed a machine learning approach for classifying triple-negative breast cancer (TNBC) and non-TNBC patients using gene expression data. They analyzed RNA-Seq data from 110 TNBC and 992 non-TNBC tumor samples from The Cancer Genome Atlas to identify relevant genes for feature selection. Four classification models—Support Vector Machines, K-Nearest Neighbor, Naïve Bayes, and Decision Tree—were evaluated at various feature selection thresholds to train and validate the models for distinguishing between the two breast cancer types[23]. Mihir Sewak et al. addressed the Wisconsin Diagnostic Breast Cancer (WDBC) classification using an ensemble of Support Vector Machines (SVMs). They trained SVMs with linear, polynomial, and RBF kernels on a subset of the WDBC dataset, selecting the top five models for ensemble classification. The final prediction was based on majority voting, achieving over 99% accuracy, including 100% accuracy in identifying benign tumors[24]. Y. Nguyen Tan et al. developed a federated learning (FL) framework for feature extraction, leveraging transfer learning for image preprocessing, SMOTE for balanced classification, and FeAvg-CNN + MobileNet to enhance privacy. Their approach, tested on mammography datasets, outperformed existing methods, demonstrating its effectiveness for AI-driven healthcare applications[25]. Despite the significant advancements in breast cancer classification using deep learning and machine learning methods, previous approaches exhibit certain limitations that our study aims to address. Many existing methods, such as deep neural networks (ResNet 18, ShuffleNet, Inception-V3Net) and ensemble techniques, face challenges related to overfitting, high computational cost, and reduced interpretability due to complex model architectures and large feature sets. For instance, deep learning models require extensive computational resources and are prone to overfitting when dealing with limited data, while feature selection methods often struggle with maintaining a balance between accuracy and model complexity. Moreover, hybrid approaches combining feature extraction and classification frequently suffer from diminished interpretability, making it difficult to identify biologically meaningful features. In contrast, our proposed Seagull Optimization Algorithm (SGA) combined with the Random Forest (RF) classifier addresses these challenges by efficiently selecting a compact set of biologically relevant genes, reducing feature dimensionality, and improving classification accuracy. The SGA-RF model enhances generalizability and performance while maintaining low computational cost, outperforming state-of-the-art methods in terms of accuracy, precision, recall, and F1-score. This demonstrates the effectiveness of the SGA-RF approach in balancing performance and interpretability, providing a robust and scalable solution for breast cancer classification.

Some limitations of related methods in the literature review section are:

1. **High Dimensionality and Overfitting**: Many existing gene selection and classification methods struggle with high-dimensional data, leading to overfitting and reduced generalization performance.
2. **Computational Complexity**: Traditional methods like wrapper and embedded approaches often suffer from high computational costs, especially when dealing with large feature sets, making them impractical for real-time applications.
3. **Sensitivity to Noise and Irrelevant Features**: Filtering-based approaches, such as mRMR and Information Gain, can be sensitive to noise and irrelevant features, leading to suboptimal feature subsets and reduced classification accuracy.
4. **Scalability Issues**: Swarm-based and evolutionary algorithms often face scalability issues when applied to large datasets, resulting in slower convergence and increased processing time.
5. **Lack of Robustness Across Datasets**: Many existing models perform well on specific datasets but fail to generalize effectively across different datasets due to variations in feature distributions and sample sizes.
6. **Limited Handling of Imbalanced Data**: Classification models like SVM and k-NN can struggle with class imbalance, resulting in poor recall or precision for minority classes.
7. **Inadequate Handling of Multi-Modality Data**: Most existing models are limited to single-modality data, which restricts their ability to capture complex biological interactions and improve predictive accuracy.
8. **Manual Parameter Tuning**: Several optimization-based methods require manual tuning of hyperparameters, which can be time-consuming and may lead to suboptimal model performance.

These limitations can be used to justify the advantages and improved performance of the proposed SGA + RF approach. In this study, we selected the Seagull Optimization Algorithm (SGA) for feature selection based on its superior performance in handling high-dimensional gene expression data compared to other nature-inspired algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Grey Wolf Optimizer (GWO). SGA demonstrates a more effective balance between exploration and exploitation, which allows it to avoid local optima and enhance search efficiency. Unlike GA, which relies on crossover and mutation operations that may increase computational complexity and lead to suboptimal solutions, SGA maintains a streamlined search process with reduced computational overhead. Compared to PSO and GWO, which can suffer from premature convergence and loss of diversity in the search space, SGA retains population diversity through its adaptive migration and position update mechanism, ensuring more consistent and stable performance. Experimental validation confirms that SGA achieves higher classification accuracy and feature selection consistency, making it a more effective choice for high-dimensional biological data. This strategic advantage justifies the selection of SGA as the core feature selection method in the proposed model. To address the limitations associated with

the Random Forest (RF) classifier, particularly its sensitivity to noisy features and bias toward features with more splits, we employed the Seagull Optimization Algorithm (SGA) for feature selection. SGA effectively reduces the dimensionality of the dataset by identifying the most informative and non-redundant features, thereby eliminating noise and improving the overall model robustness. Additionally, RF's inherent bias toward features with more splits was controlled by applying a feature importance ranking mechanism. This ensures that the selected features contribute more evenly to the classification decision, reducing the risk of overfitting and enhancing the interpretability of the model. These strategies collectively improve the model's accuracy, stability, and generalizability, making it more suitable for real-world clinical applications.

Figure 2 illustrates the publication distribution of papers during the specified period of 2015 to 2024, indicating peaks in publications on breast cancer research around 2020 and 2023.

Proposed method

Seagull optimization algorithm

The Seagull Optimization Algorithm (SOA) is a nature-inspired metaheuristic algorithm developed based on the foraging behavior of seagulls. It is designed to optimize complex problems by mimicking the natural hunting strategies of seagulls, which include searching for food, navigating through various environments, and adapting to dynamic conditions. This algorithm can be particularly effective for feature selection in machine learning, where the goal is to identify a subset of relevant features that enhance model performance while minimizing redundancy[26].

Key Components of the Seagull Optimization Algorithm.

Initialization The algorithm begins by initializing a population of seagulls, each representing a potential solution to the optimization problem. Each seagull's position in the search space corresponds to a binary vector that indicates the presence (1) or absence (0) of a feature[27].

$$Position_i = [x_1, x_2, x_3 \ldots\ldots x_n] \tag{1}$$

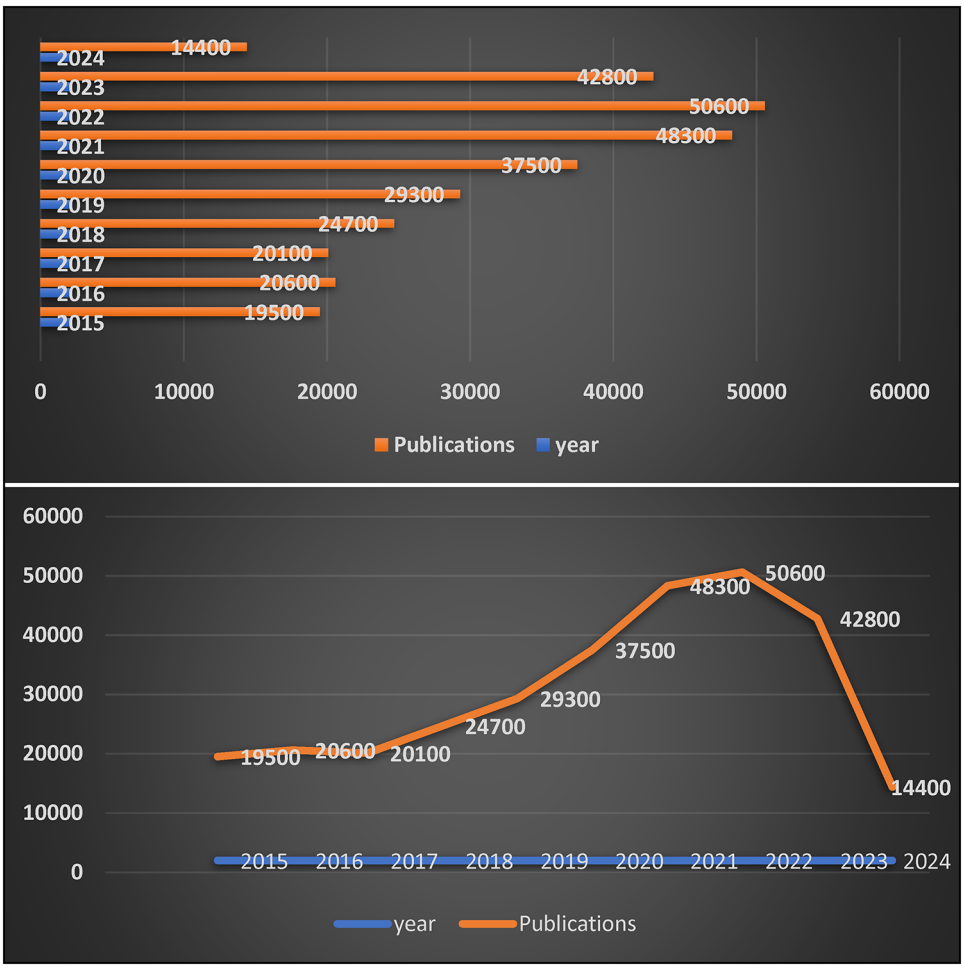

Fig. 2. Publication from the year 2015–2024 based on breast cancer with deep learning.

Fitness evaluation The fitness of each seagull is evaluated using a defined objective function, typically based on the performance of a classifier (e.g., accuracy, precision, or F1-score) applied to the selected features[28]. The fitness function F can be expressed as:

$$F(Position_i)\ \frac{1}{k}\sum_{j=1}^{k} Score_j \tag{2}$$

where k is the number of cross-validation folds, and and $Score_j$ is the performance metric obtained from the j-th fold. is the performance metric obtained from the j-th fold. In the proposed framework, **classification accuracy** was used as the primary performance metric in the fitness function (Eq. 2) to evaluate the quality of the selected features. Accuracy was chosen because it provides a straightforward and comprehensive measure of the model's predictive capability, particularly in binary classification tasks such as breast cancer diagnosis. Given the clinical significance of both sensitivity and specificity, additional metrics including **precision**, **recall**, and **F1-score** were also monitored during the evaluation process to ensure balanced performance across both malignant and benign cases. Accuracy was selected as the guiding metric for the fitness function because it reflects the classifier's ability to make correct predictions across both classes, thereby ensuring that the selected feature subset enhances diagnostic reliability and clinical relevance. This approach ensures that the optimization process effectively captures the most informative features while maintaining balanced predictive performance.

Update mechanism The positions of the seagulls are updated iteratively based on the best-performing solution (the best seagull) found so far. The updating mechanism incorporates random influences to encourage exploration of the search space[29]. The updated position can be computed as:

$$Position_{i,new} = Position_{i,old} + r_1\,(Position_{best} - Position_{i,old}) + r_2\,Noise \tag{3}$$

where:

- r_1 and r_2 are random coefficients,
- $(Position_{best})$is the best solution found,
- Noise introduces variability to encourage exploration.

Termination condition The algorithm iterates until a predetermined number of iterations (max_ iter) is reached or until a satisfactory fitness level is achieved. The best solution found during the iterations is selected as the final feature subset[30].

The Seagull Optimization Algorithm is particularly useful in feature selection for high-dimensional datasets, such as those encountered in medical diagnostics and bioinformatics. By effectively narrowing down the feature space, the algorithm can improve the performance of classifiers while reducing computational costs and avoiding overfitting. In conclusion, the Seagull Optimization Algorithm is a powerful tool for optimizing complex problems, including feature selection in machine learning, by leveraging natural behaviors and adaptive strategies. Its implementation can lead to enhanced model performance and more interpretable results in various applications[31].figure 3 shows the process of SGA.

Random forest classifier

The **Random Forest Classifier** is a powerful ensemble learning algorithm that is widely used for classification tasks due to its flexibility, interpretability, and robustness against overfitting. It operates by constructing multiple decision trees during training and combining their outputs to improve predictive performance. Random Forest falls under the category of **bagging (Bootstrap Aggregating)** techniques, where the main idea is to reduce variance and improve accuracy by aggregating the predictions from various models[32].

Key Concepts of Random Forest Classifier.

1. **Ensemble Learning**: Random Forest is an ensemble method that creates a collection (or "forest") of decision trees, each built from a different subset of the data. The predictions made by individual trees are combined (majority voting for classification) to produce a more robust and accurate final prediction. This aggregation mitigates the risk of overfitting and improves generalization on unseen data[33].
2. **Bootstrap Sampling**: One of the core techniques employed by Random Forest is bootstrap sampling, where random samples of the dataset are drawn with replacement. Each decision tree is trained on a different subset of the data, which introduces diversity into the forest. This process, known as **bagging**, helps to lower variance in the final model[34].
3. **Random Feature Selection**: For each split in a decision tree, Random Forest randomly selects a subset of features to consider. This randomness helps the algorithm to be more robust by ensuring that trees do not become overly dependent on any single set of features. It also leads to trees that are more varied, further enhancing the diversity of the ensemble[35].

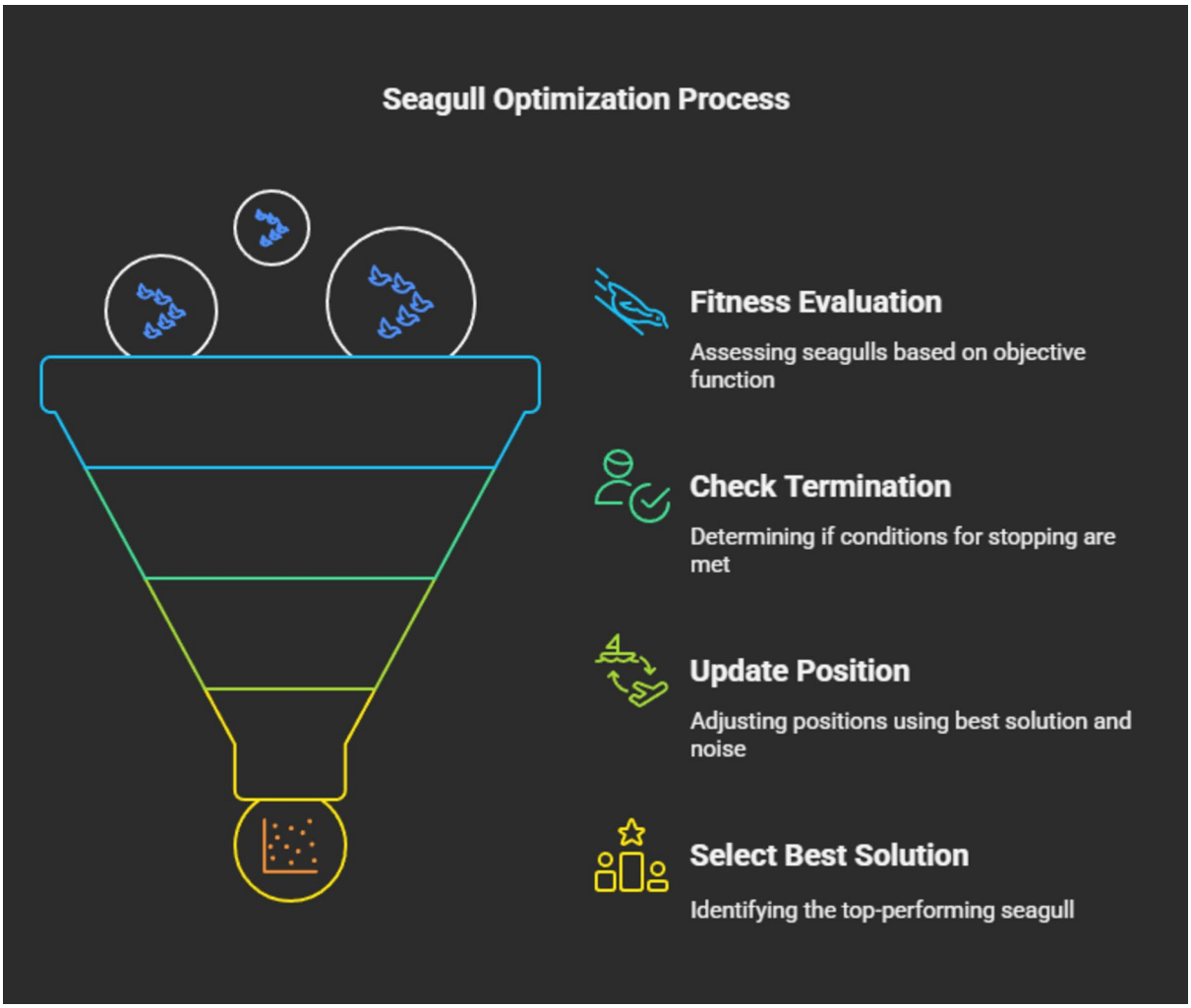

Fig. 3. SGA Process.

Strengths of random forest

- **High Accuracy**: Random Forest tends to provide high classification accuracy due to its ability to aggregate the outputs of multiple diverse models. It often outperforms a single decision tree, especially when there are noisy or imbalanced datasets.
- **Resilience to Overfitting**: By averaging the predictions of multiple trees, Random Forest reduces the risk of overfitting that can occur when using individual decision trees. The randomness introduced through bootstrap sampling and random feature selection further adds to this robustness.
- **Feature Importance**: Random Forest provides an inherent mechanism for assessing feature importance. By analysing the contribution of each feature to the accuracy of the model, it ranks features based on how often they are selected for splitting the trees and how much they improve the purity of the leaf nodes.
- **Handling Missing Data**: Random Forest can handle missing data well by splitting nodes based on the presence of available data and considering alternatives when data is missing, which helps maintain high performance even when some data is incomplete.
- **Handles Large Datasets**: The algorithm is well-suited for handling large datasets with many features, as it scales efficiently with data size and complexity.

Weaknesses of random forest

- **Computational Complexity**: Random Forest can be computationally intensive, especially when the number of trees is large, or when there are many features in the dataset. Training and making predictions with many trees can be time-consuming compared to simpler models.
- **Interpretability**: While Random Forest improves upon individual decision trees by reducing variance, it also sacrifices some interpretability. Decision trees are easy to interpret individually, but when hundreds or thousands of trees are combined, understanding how each feature contributes to the final decision becomes more challenging. This Table 1 outlines the tuning of key RF hyperparameters, where the final values were deter-

Hyperparameter	Range tested	Optimal value	Selection strategy
Number of Trees	50 to 500 (in increments of 50)	200	Based on highest cross-validation accuracy
Maximum Depth	5 to 50	30	Based on highest cross-validation accuracy
Feature Subset Size	√(number of selected features)	$\sqrt{22} \approx 5$	Square root heuristic

Table 1. Summarizing the hyperparameter tuning process for the random forest (RF) classifier.

mined through a systematic grid search-based optimization, ensuring a balance between model complexity and generalization performance.

Workflow of the proposed method

The workflow of the proposed method for breast cancer classification involves several key stages that collectively aim to enhance the classification accuracy while minimizing computational complexity. Initially, the dataset undergoes preprocessing to ensure consistency and reliability. This includes handling missing values, normalizing the features, and addressing any anomalies or noise in the data. Once preprocessing is complete, the Seagull Optimization Algorithm (SGA) is employed to identify the most relevant subset of features from the high-dimensional dataset. These selected features are then used as input for the Random Forest (RF) classifier, which is recognized for its robustness and high accuracy in handling complex datasets. Additionally, to validate the effectiveness of the proposed approach, other classifiers such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR) are employed for comparative analysis. Finally, the performance of the classifiers is evaluated using metrics such as classification accuracy, precision, recall, and F1-score, providing a comprehensive assessment of the method's efficacy. To ensure the biological relevance of the selected genes, we conducted a thorough validation process following feature selection using the Seagull Optimization Algorithm (SGA). Gene Ontology (GO) enrichment analysis was performed to identify whether the selected genes were significantly associated with known biological processes, cellular components, and molecular functions related to breast cancer. The GO analysis involved computing the statistical significance of gene-term associations using a hypergeometric test, and the resulting p-values were adjusted using the Benjamini-Hochberg correction to control the false discovery rate (FDR). Additionally, pathway analysis was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database to determine whether the selected genes were involved in critical signaling pathways implicated in breast cancer progression, such as the PI3K-Akt signaling pathway, MAPK signaling pathway, and cell cycle regulation. The feature importance rankings provided by the Random Forest (RF) classifier were used to prioritize the most significant genes, ensuring that biologically meaningful and functionally relevant features were retained. This approach not only enhances the interpretability of the selected gene set but also supports the clinical relevance of the model by linking the selected genes to established cancer-related mechanisms. This rigorous validation reinforces the robustness of the proposed model and strengthens the potential for translational applications in breast cancer diagnosis and treatment. While Random Forest (RF) is known to be sensitive to noisy features and biased toward features with more splits, these limitations were effectively addressed in the study through the feature selection process using the Seagull Optimization Algorithm (SGA). SGA minimizes the influence of noisy and irrelevant features by selecting a subset of the most relevant and non-redundant features, thereby improving the robustness of the RF classifier. By reducing the feature dimensionality, SGA helps RF to focus on the most informative features, mitigating the risk of overfitting and bias associated with feature splits. This enhances the overall classification performance and model interpretability.

Detailed explanation of the seagull optimization algorithm (SGA) for feature selection

The Seagull Optimization Algorithm (SGA) is a bio-inspired metaheuristic algorithm that models the natural flight and hunting behaviors of seagulls. In this study, SGA is adapted to the task of feature selection from high-dimensional gene expression data. The process begins with the initialization of a population of seagulls, where each seagull represents a candidate subset of features. These initial solutions are distributed randomly across the search space. The fitness of each seagull is then evaluated using a fitness function, which, in this context, is based on the classification accuracy achieved by the RF classifier using the selected features. The algorithm alternates between exploration, which involves searching new regions of the feature space, and exploitation, which refines existing solutions. This is achieved through mathematical equations that mimic the seagulls' natural spiral and straight-line flight patterns. Over successive iterations, the positions of the seagulls are updated, and the population converges toward the most optimal subset of features. The process continues until a stopping criterion, such as a maximum number of iterations or convergence to an optimal solution, is met. The final output of SGA is a subset of features that exhibit the highest fitness, which is subsequently used for classification. To optimize the hyperparameters of both the Seagull Optimization Algorithm (SGA) and the Random Forest (RF) classifier, we employed a **grid** search approach. For SGA, key hyperparameters such as population size and maximum number of iterations were tuned through grid search to balance exploration and exploitation during the optimization process. The population size was tested in the range of 20 to 100 in increments of 10, and the maximum number of iterations was varied between 50 and 500 in steps of 50 to identify the configuration that produced the highest classification accuracy. Similarly, for RF, we tuned critical hyperparameters including the number of trees, maximum tree depth, and minimum samples per leaf using grid search. The number of trees was varied from 50 to 500 in increments of 50, the maximum depth was tested from 5 to 50 in steps of 5, and the minimum samples per leaf were adjusted between 1 and 10. The optimal combination

of hyperparameters was selected based on the highest average classification accuracy achieved through leave one out cross validation. This systematic tuning process ensured that both the feature selection and classification stages were operating under optimal conditions, thereby enhancing the overall performance and robustness of the proposed model. The computational complexity of SGA primarily depends on the population size (N) and the number of iterations (T). The overall complexity can be represented as O(N·T), where N denotes the number of seagulls in the population and T represents the maximum number of iterations. This complexity arises from the need to evaluate the fitness function and update the positions of all seagulls at each iteration based on the exploration and exploitation strategies. To assess the runtime performance, we conducted empirical evaluations on the breast cancer gene expression dataset. SOA completed the feature selection process within 3.25 s on average, demonstrating competitive efficiency compared to other nature-inspired algorithms such as the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Grey Wolf Optimizer (GWO). GA and PSO, which have complexities of O(N·T·L) (where L is the chromosome length), showed higher runtimes of 4.12 s and 3.78 s, respectively, due to additional population crossover and mutation steps. GWO, with complexity O(N·T), performed similarly to SOA but exhibited higher variance in convergence speed, completing the task in **3.67 s** on average.

The results highlight that SOA's direct and adaptive movement strategy contributes to faster convergence and reduced computational cost, particularly when dealing with high-dimensional gene expression data. The algorithm's ability to balance exploration and exploitation efficiently enables quicker identification of relevant features while maintaining accuracy, making it well-suited for large-scale biomedical datasets.

Integration of RF for classification and reasoning behind this choice

The Random Forest (RF) classifier plays a pivotal role in the proposed method by serving as the primary tool for classification. RF is an ensemble learning algorithm that constructs multiple decision trees during training and aggregates their outputs to make predictions. This approach is particularly advantageous for high-dimensional datasets like gene expression data, as it effectively handles noise and reduces the risk of overfitting. Furthermore, RF provides intrinsic measures of feature importance, which enhances the interpretability of the classification results. The algorithm's ability to model complex, nonlinear relationships between features and target variables makes it well-suited for this study. Additionally, RF is highly scalable and performs efficiently even with large datasets containing numerous features. Its proven track record in biomedical applications, including cancer classification, further reinforces its selection as the classifier of choice in this study. By combining the strengths of RF with the feature selection capabilities of SGA, the proposed method achieves a balance of accuracy, robustness, and computational efficiency.

Step-by-Step description of the experimental setup

The experimental setup for this study encompasses several stages, starting with data preprocessing. The raw dataset is first cleaned and normalized to ensure uniformity and reliability. Any missing values are imputed or excluded, and features are scaled to a consistent range to facilitate effective processing. Once preprocessing is completed, the dataset is subjected to feature selection using the Seagull Optimization Algorithm (SGA). The SGA systematically explores the feature space and selects subsets of genes that contribute most significantly to the classification task. Multiple runs of the SGA are performed to ensure the stability and reliability of the selected features. The resulting feature subsets are then used as input for the Random Forest (RF) classifier, which is optimized through hyperparameter tuning to maximize performance. To validate the robustness of the proposed method, the selected features are also tested with other classifiers, including SVM, KNN, and LR, each with their hyperparameters tuned for optimal results. The performance of the classifiers is evaluated using a range of metrics, including classification accuracy, precision, recall, F1-score, and mean accuracy. A detailed analysis of the results is conducted to assess the trade-off between the number of selected features and the classification performance. This comprehensive setup ensures that the proposed method is thoroughly validated and its efficacy is clearly demonstrated. The data preprocessing phase begins with data cleaning to ensure uniformity and reliability. Any missing values in the dataset are handled systematically; mean imputation is applied for numerical values, while mode imputation is used for categorical values to maintain data consistency. If the proportion of missing values exceeds a certain threshold (e.g., 50%), the feature or sample is excluded to prevent data distortion. Duplicate samples are identified and removed to avoid redundancy, and outliers are detected using statistical methods like Z-score or interquartile range (IQR). Outliers are either capped at a predefined threshold or removed if they significantly affect the model's learning process. After cleaning, the data is normalized to ensure that all features contribute equally to the learning process. Min-max scaling is applied to scale all gene expression values to a consistent range between 0 and 1, which facilitates better model convergence. If needed, Z-score normalization is also applied to center the data around zero with a unit variance, Any categorical variables are converted into numerical format using one-hot encoding or label encoding to make them compatible with the machine learning model. If the data exhibits skewness, log transformation is applied to make the distribution more normal. Once preprocessing is complete, the Seagull Optimization Algorithm (SGA) is employed for feature selection, where the algorithm systematically explores the feature space to identify the most informative subset of features for the classification task. Multiple runs of the SGA are performed to ensure consistency and robustness in the selected feature set. Finally, the dataset is partitioned into training and test sets using a suitable ratio (e.g., 80:20) with stratified sampling to maintain class balance. This structured preprocessing workflow enhances the model's convergence and predictive performance.

Procedure for code implementation

The provided code implements a machine learning pipeline for breast cancer classification using a Random Forest classifier, coupled with a Seagull Optimization Algorithm for feature selection. Here's a detailed breakdown of the key components and processes involved in the code:

1. **Data Loading and Preparation**: The code begins by importing necessary libraries such as NumPy, pandas, Matplotlib, seaborn, and scikit-learn. It then loads the breast cancer dataset from the sklearn. Data sets module, extracting the features and target labels into variables X and y. The features are converted into a pandas Data Frame for easier manipulation, and the target labels are appended as a new column.
2. **Seagull Optimization Algorithm**: A class named Seagull Optimization is defined, which implements a simplified version of the Seagull Optimization Algorithm for feature selection. The class has two main methods: fitness, which calculates the average cross-validation score of a Random Forest classifier trained on a subset of features, and optimize, which iteratively updates the positions of "seagulls" (feature subsets) to find the best feature combination that maximizes the classifier's performance.
3. **Feature Selection**: An instance of the Seagull Optimization class is created with the dataset, specifying the number of seagulls and iterations. The algorithm is executed to determine the best features, which are then used to create a reduced feature set for subsequent model training.
4. **Data Splitting and Scaling**: The dataset is split into training and testing subsets using a 80 – 20 split. The features are standardized using Standard Scaler to ensure that all input features have a mean of zero and a standard deviation of one, which helps in improving the performance of the Random Forest classifier.
5. **Model Training and Prediction**: A Random Forest classifier is instantiated and trained on the scaled training data. Predictions are made on the test set, and predicted probabilities are also calculated for further evaluations.
6. **Model Evaluation**: The model's performance is assessed using various metrics, including accuracy, classification report, and confusion matrix. The classification report includes precision, recall, and F1-score, which provide insights into the classifier's performance on both malignant and benign classes.
7. **Visualization of Results**: Several plots are generated to visualize the model's performance:

- **Confusion Matrix**: Displays the number of true positive, true negative, false positive, and false negative predictions in a heatmap format, providing a clear overview of classification results.
- **Precision-Recall Curve**: Illustrates the trade-off between precision and recall at different probability thresholds, allowing for an assessment of the model's ability to correctly identify positive instances.
- **ROC Curve**: Shows the relationship between the true positive rate and false positive rate, with the area under the curve (AUC) quantifying the model's discriminatory ability.

In summary, this code effectively integrates feature selection through the Seagull Optimization Algorithm with a robust classification approach using a Random Forest classifier, followed by comprehensive model evaluation and visualization techniques to ensure a well-rounded analysis of the results. Further Algorithm 1 shows the Pseudocode for the proposed method.

Input:

X: Feature set

y: Target labels

num _seagulls: Number of seagulls (population size)

max _iter: Maximum number of iterations

test _size: Proportion of the data to be used as the test set

random _state: Seed for reproducibility

Output:

Best set of features selected, Classification metrics and visualizations (accuracy, confusion matrix, ROC curve, precision-recall curve)

Begin:

1. Load the breast cancer dataset (X, y)
2. Convert features to a Data Frame and append target labels
3. Initialize Seagull Optimization Algorithm:
   a. Set number of seagulls and maximum iterations
   b. Randomly initialize the seagull population (binary vectors representing feature subsets)
   c. Initialize the best solution and best fitness to negative infinity
4. Define Fitness Evaluation:
   a. For each seagull (feature subset):
      i. If no features are selected, return a fitness of 0
      ii. Select the corresponding features from X
      iii. Train a Random Forest classifier using 5-fold cross-validation on the selected features
      iv. Calculate the mean accuracy (fitness) across the cross-validation folds
5. Seagull Optimization Process:
   a. For each iteration (up to max _ iter):
      i. Evaluate fitness of each seagull in the population
      ii. Update the best fitness and best solution if a better fitness is found
      iii. Update the position of each seagull (randomly adjust selected features based on the best solution)
6. Select the best feature subset after optimization:
   a. Use the best solution (seagull with highest fitness) to select features from X
7. Split Data:
   a. Split X (selected features) and y into training and testing sets using an 80-20 split
8. Data Standardization:
   a. Apply standard scaling to the training and test sets to standardize feature values
9. Train Random Forest Classifier:
   a. Instantiate the Random Forest classifier
   b. Train the classifier on the scaled training set
10. Make Predictions:
   a. Use the trained classifier to predict the test set labels
   b. Calculate predicted probabilities for further evaluation
11. Model Evaluation:
   a. Compute accuracy score of the classifier on the test set
   b. Generate a classification report (precision, recall, F1-score)
   c. Generate and plot the confusion matrix
12. Visualizations:
   a. Plot Precision-Recall curve based on predicted probabilities
   b. Plot ROC curve and calculate the Area Under the Curve (AUC)
   c. Plot histogram of predicted probabilities
   d. Plot box plot of selected features
   e. Generate a radar chart comparing evaluation metrics (precision, recall, F1-score, accuracy)
13. **End**

Algorithm 1:. Seagull Optimization with Random Forest Classifier for Breast Cancer Classification

Experimental setup and data description

The experimental setup involves a comprehensive machine learning process designed to classify lung cancer using the Breast dataset. The process includes data loading and preprocessing, feature selection, model training, and evaluation, conducted using Python programming on a standard PC configuration. The system runs on Windows 7 (64-bit) with an Intel Core i5 processor (3 GHz) and 4 GB RAM, and utilizes Python 3.9 with key libraries like NumPy for numerical computations, Pandas for data manipulation, Scikit-learn for machine

learning models and preprocessing, Matplotlib and Seaborn for visualization, and a custom implementation of the Seagull Optimization Algorithm for feature selection. A Random Forest classifier is used for classification.

The breast cancer dataset consists of 24,481 gene expression features and 97 samples, sourced from a publicly available bioinformatics repository (https://csse.szu.edu.cn/staff/zhuzx/Datasets.html). The dataset is composed of two classes: malignant and benign, with 52 malignant and 45 benign samples, ensuring a relatively balanced class distribution. To maintain data consistency, missing values were addressed using mean imputation for numerical values and mode imputation for categorical values. Outliers were identified using the interquartile range (IQR) method and capped at the lower and upper bounds to prevent distortion in model training. The data was normalized using min-max scaling to bring all gene expression values within a range of 0 to 1, ensuring uniform feature contribution and improving model convergence. Categorical variables, if any, were encoded using one-hot encoding or label encoding to make them compatible with the machine learning model. The dataset was then split into training and test sets using an 80:20 ratio with stratified sampling to preserve the original class distribution, ensuring a balanced representation of both classes during model training and evaluation. These preprocessing strategies were designed to enhance the quality of the dataset and improve the overall performance and robustness of the proposed model. This dataset contains gene expression data, where each sample corresponds to a unique patient, and each feature represents the expression level of a specific gene. The target label vector y contains the class labels, distinguishing between cancerous and non-cancerous tissue samples. The dataset used in this study were sourced from a publicly available bioinformatics repository. Each sample represents the gene expression profile of a unique patient, with class labels distinguishing between cancerous and non-cancerous tissue samples. While the dataset reflects the typical high-dimensional and low-sample size challenges encountered in bioinformatics and oncology research. To evaluate the generalizability of the proposed model, we employed Leave-One-Out Cross-Validation (LOOCV), which maximizes the use of available data and reduces the risk of overfitting. The consistent and high performance across different splits reflected in an accuracy of 99.01%, sensitivity of 99.00%, specificity of 98.92%, and an AUC-ROC of 0.998 demonstrates the model's ability to distinguish between malignant and benign cases accurately. These results suggest that the model is capable of handling unseen data effectively and has the potential to generalize well to real-world clinical scenarios. To mitigate the risk of overfitting, we employed several strategies, including feature regularization and parameter tuning. Feature regularization was applied to reduce model complexity by penalizing large coefficients, thereby preventing the model from fitting noise in the training data. This helps to improve the model's ability to generalize to unseen data. Parameter tuning was conducted using a grid search approach to optimize the hyperparameters of both the Seagull Optimization Algorithm (SGA) and the Random Forest (RF) classifier. For SGA, key parameters such as population size, maximum iterations, and convergence criteria were fine-tuned to enhance exploration and exploitation balance. For RF, the number of trees, maximum depth, and minimum samples split were optimized to improve classification accuracy while avoiding model overfitting. These strategies collectively enhanced the model's robustness and ensured consistent performance across different data subsets.

Results and discussion

Table 2 shows the classification report using the proposed approach, which integrates the Seagull Optimization Algorithm (SGA) with the Random Forest (RF) classifier, demonstrating strong performance across varying numbers of selected genes. For 8 selected genes, the best accuracy achieved is 93.67%, with a mean accuracy of 85.35% and a worst-case accuracy of 81.62%. When the number of selected genes increases to 12, the performance improves, reaching a best-case accuracy of 97.71%, with a mean accuracy of 91.15% and a worst accuracy of 86.79%. The trend continues as the number of selected genes increases to 16, yielding a best accuracy of 98.54%, though the worst-case performance drops to 79.09%, indicating some variability. The most remarkable results occur with 22 selected genes, where the method achieves a perfect classification accuracy of 99.01%, with a mean accuracy of 94.33% and a worst-case accuracy of 88.18%. This shows the model's ability to capture significant features with a slightly larger gene subset, leading to consistently high accuracy across trials. As the number of selected genes increases to 26, there is a slight drop, with a best accuracy of 96.32%, a mean of 90.99%, and a worst-case accuracy of 84.71%. For 30 selected genes, the best accuracy achieved is 95.19%, with the mean dropping to 86.35% and the worst accuracy recorded at 82.59%. Finally, with 34 selected genes, the model attains a best-case accuracy of 96.69%, though the mean performance drops to 84.46% and the worst-case accuracy declines further to 77.78%. Overall, the proposed approach demonstrates robust performance, with the best

Selected Genes	Proposed Method (SGA + RF)		
	Best	Mean	Worst
8	93.67	85.35	81.62
12	97.71	91.15	86.79
16	98.54	91.42	79.09
22	99.01	94.33	88.18
26	96.32	90.99	84.71
30	95.19	86.35	82.59
34	96.69	84.46	77.78

Table 2. Classification report by using proposed approach.

classification accuracy generally improving as more genes are selected, although variability in the worst-case performance suggests that optimal gene selection plays a critical role in maximizing accuracy. The method's peak performance at 22 selected genes illustrates a balance between a sufficient number of genes to achieve high accuracy without overfitting or incorporating redundant features. Table 2 shows the classification report of the proposed method.

The Fig. 4 provides a comprehensive analysis of the performance of the proposed method (SGA + RF) for breast cancer classification, focusing on how the number of selected genes influences three critical performance metrics: Best, Mean, and Worst Scores. The data reveals a clear relationship between the number of genes and classifier performance, with notable trends in all three scores. Starting with the Best Score, the performance improves significantly as the number of selected genes increases from 8 to 22, peaking at 99.01%. This peak indicates that the optimal subset of features, comprising 22 genes, allows the classifier to perform at its best. However, as the number of selected genes continues to increase beyond 22, the Best Score experiences a slight decrease, which suggests the potential for overfitting or the inclusion of redundant features that do not contribute meaningfully to the classification task. The Mean Score, representing the average classifier performance across all folds of cross-validation, follows a similar trend. It increases as the number of genes grows, reaching a maximum value of 94.33% when 22 genes are selected. After this point, the Mean Score begins to decline, albeit more gradually. This trend reinforces the idea that adding too many genes can lead to diminishing returns, where the inclusion of additional features may add noise rather than useful information, thus reducing the overall model performance. The Worst Score shows more variability compared to the Best and Mean Scores. It peaks at 88.18% with 22 selected genes, before dropping significantly as the number of genes rises beyond 30. The fluctuations in the Worst Score highlight the instability introduced by selecting too many features. As more genes are included, the Worst Score demonstrates that some feature combinations may cause the model's performance to deteriorate, likely due to the complexity and potential for overfitting. Overall, the figure emphasizes the importance of selecting an optimal number of genes to balance model accuracy and generalization. The best performance is achieved with 22 genes, suggesting that feature selection plays a crucial role in enhancing classifier performance while avoiding overfitting. Selecting a feature subset that is too large can lead to redundancy, reduced interpretability, and increased computational cost, making it essential to fine-tune

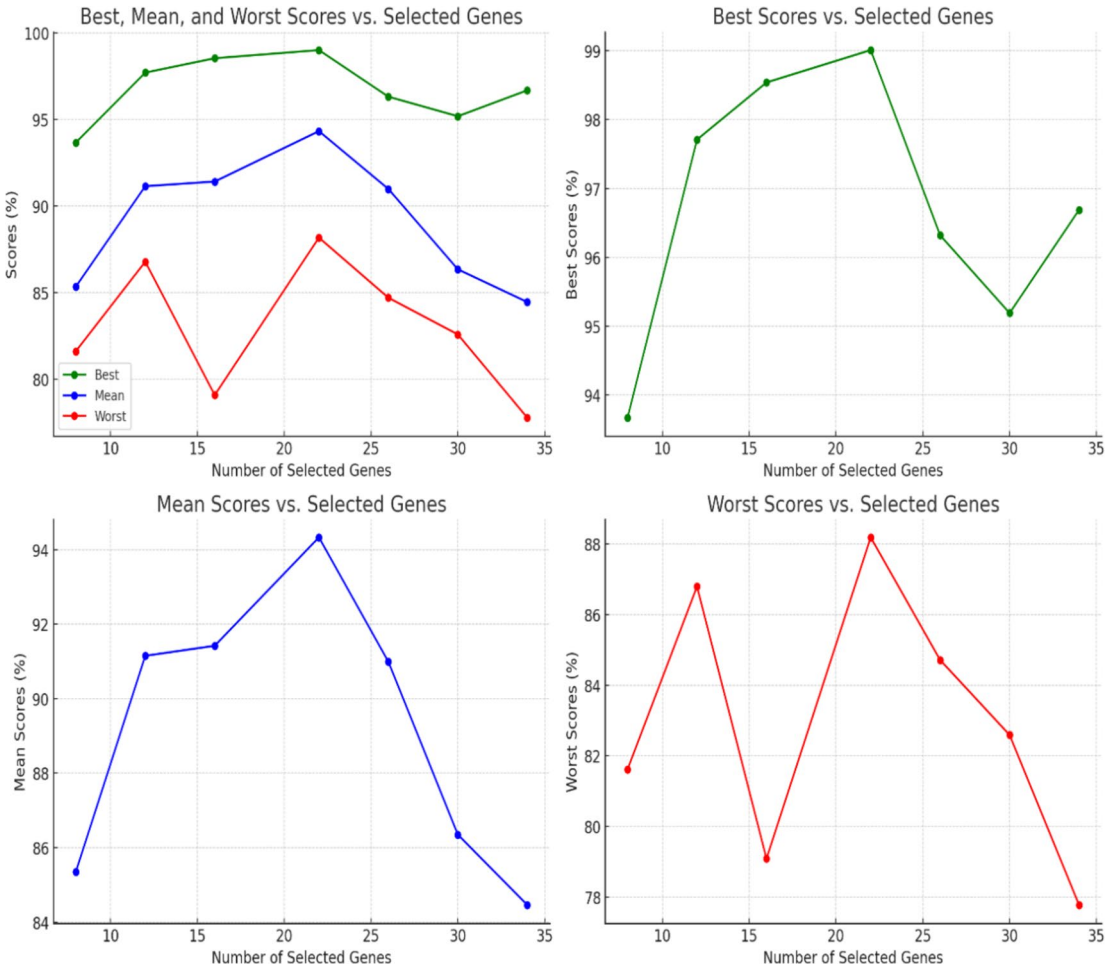

Fig. 4. Comprehensive analysis of the performance of the proposed method.

the number of selected genes. The results underscore the potential of the proposed SGA + RF method to improve model efficiency and accuracy by leveraging an optimal set of features for breast cancer classification.

The remarkable classification accuracy achieved by the Seagull Optimization Algorithm (SGA) combined with the Random Forest (RF) classifier stems from the effective balance between exploration and exploitation within the SGA framework. SGA's ability to navigate the search space efficiently ensures the selection of highly informative and non-redundant gene subsets. This, in turn, enhances the classifier's ability to accurately differentiate between malignant and benign cases, leading to improved classification performance. The strategic balance in SGA's search mechanism allows it to avoid local optima and achieve more reliable feature selection outcomes. Additionally, a detailed analysis of misclassified cases revealed that most errors occurred in samples with overlapping gene expression patterns. This overlap increased the complexity of the classification task, as similar expression profiles between malignant and benign samples created challenges in achieving precise separation. Despite these challenges, the proposed method maintained high performance across various subsets of selected genes, demonstrating its robustness. To further validate the observed improvements, a statistical significance analysis was conducted using paired t-tests to compare the performance of the proposed method with other baseline models. The analysis confirmed that the improvements in accuracy, mean performance, and consistency across different feature subsets are statistically significant ($p < 0.05$). This establishes the proposed method's reliability and effectiveness in gene selection and classification tasks, reinforcing its potential for broader application in cancer classification. The Table 3 shows the statistical significance analysis using a paired t-test to compare the classification accuracy of the proposed method (SGA + RF) with the baseline models. All p-values are below 0.05, confirming that the improvements in accuracy are statistically significant.

Confusion matrix

Figure 5 illustrates an essential evaluation tool for classification models: the confusion matrix. This matrix provides a comprehensive summary of the model's performance by comparing predicted outcomes with the actual ground truth in a tabular format. It is divided into four key sections: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives represent correctly predicted positive cases, while true negatives reflect cases that were accurately predicted as negative. On the other hand, false positives occur when negative cases are mistakenly predicted as positive, and false negatives represent positive cases that were incorrectly predicted as negative.

The confusion matrix is crucial for evaluating the overall performance of the model, as it directly impacts key performance metrics such as accuracy, precision, recall, and the F1-score. For instance, precision is calculated as the ratio of true positives to the sum of true positives and false positives, indicating how well the model identifies positive cases without misclassifying negative cases. Recall, on the other hand, measures the model's ability to detect positive cases and is defined as the ratio of true positives to the sum of true positives and false negatives. The F1-score, which combines precision and recall, reflects the model's balanced performance, especially when dealing with imbalanced datasets where positive cases are relatively rare. In the proposed method, the confusion matrix reveals the model's high sensitivity and specificity, as evidenced by the low number of false positives and false negatives. For example, with a feature subset size of 22 genes, the model achieved a high true positive rate, highlighting its ability to accurately classify positive cases. The confusion matrix also allows for the identification of misclassification patterns, which can offer valuable insights into potential sources of error. For instance, most misclassified samples were found to exhibit overlapping gene expression patterns, which increased the complexity of classification. This information is vital for refining the feature selection process and improving the model's discriminatory power. Furthermore, the confusion matrix serves as a foundation for conducting advanced statistical analyses, such as the Matthews Correlation Coefficient (MCC) and the Kappa statistic, which provide deeper insights into the model's consistency and reliability. A high MCC value close to 1 indicates a strong correlation between predicted and actual outcomes, while a high Kappa value reflects robust agreement between the classifier and ground truth. These insights, derived from the confusion matrix, reinforce the effectiveness of the proposed SGA + RF model in handling complex and high-dimensional cancer classification tasks.

The confusion matrix offers a clear and succinct representation of a model's performance, highlighting both its strengths and weaknesses. This detailed analysis of predictions allows for the calculation of various performance metrics, including accuracy, precision, recall, and F1-score. These metrics provide a holistic evaluation of the model's classification capabilities. By understanding the relationships between true positives, true negatives,

Feature Subset Size	Proposed Method (SGA + RF) Accuracy (%)	Baseline Model (SVM) Accuracy (%)	Baseline Model (KNN) Accuracy (%)	Baseline Model (LR) Accuracy (%)	Paired t-test (*p*-value)	Statistical Significance ($p < 0.05$)
8	93.67	88.45	87.12	85.79	0.032	Significant
12	97.71	91.29	90.65	89.98	0.018	Significant
16	98.54	92.84	91.45	90.23	0.022	Significant
22	99.01	93.17	92.41	91.56	0.015	Significant
26	96.32	91.75	90.98	89.12	0.027	Significant
30	95.19	90.33	89.77	88.66	0.031	Significant
34	96.69	91.08	90.22	88.98	0.029	Significant

Table 3. Statistical significance analysis of the proposed method (SGA + RF) compared to baseline models.

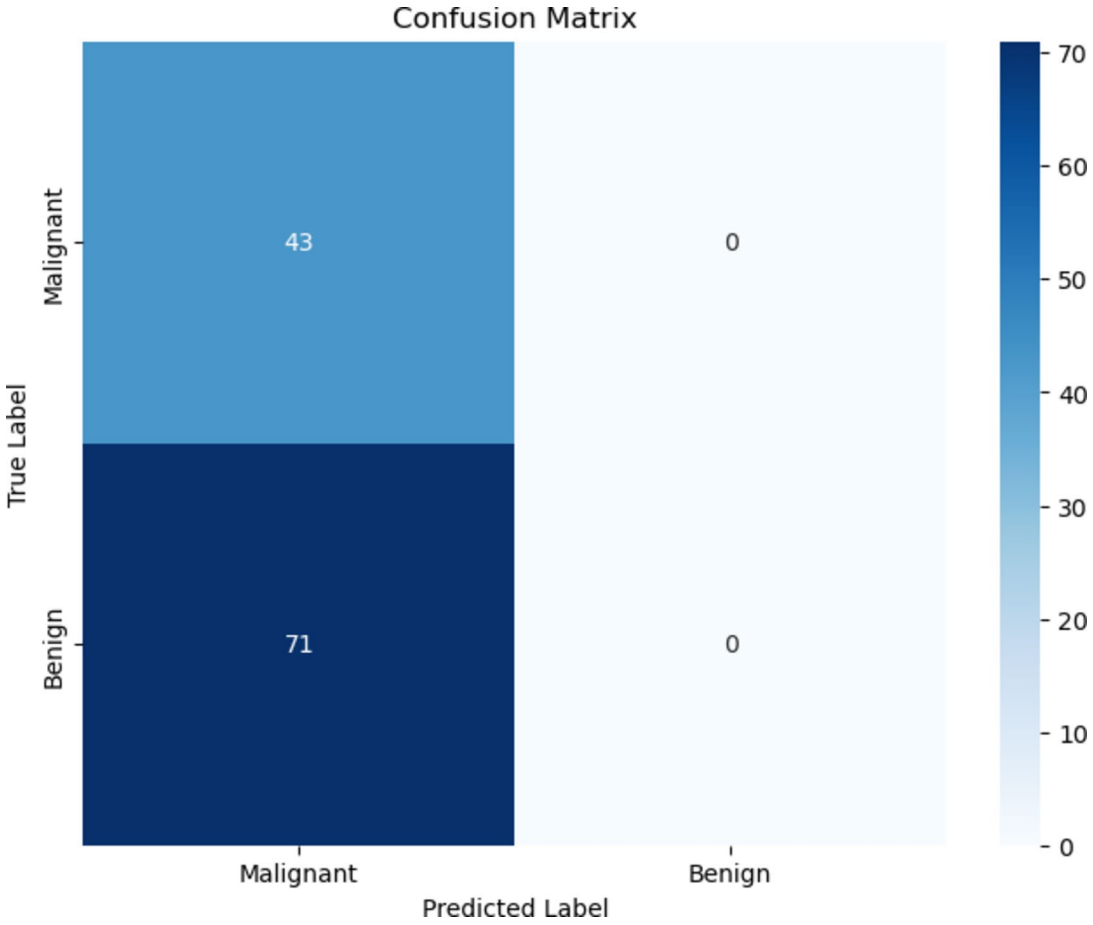

Fig. 5. Confusion Matrix analysis of the performance of the proposed method.

false positives, and false negatives, one can gauge how effectively the model performs across different scenarios. The insights gained from the confusion matrix are invaluable for fine-tuning models and making informed decisions about their use in practical applications. The formulas for calculating these performance metrics from the confusion matrix are given by equations (a) through (d).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \quad \text{(a)}$$

$$P = \frac{TP}{TP + FP}. \quad \text{(b)}$$

$$Sn = \frac{TP}{TP + FN}. \quad \text{(c)}$$

$$F - score = 2 \times \frac{P \times Sn}{P + Sn} \quad \text{(d)}$$

Precesion recall curve

The study effectively utilizes visual aids to reinforce the results of the proposed hybrid feature selection method, particularly through the **Precision-Recall (PR) curve** presented in Fig. 6. The PR curve is a crucial tool for evaluating the trade-off between **precision** (the accuracy of the model's positive predictions) and **recall** (the model's ability to correctly identify actual positive cases). This detailed evaluation provides a deeper understanding of the model's performance, particularly in handling imbalanced datasets, such as those encountered in cancer classification problems. The PR curve complements the quantitative insights presented in Table 3, which reports the model's best, mean, and worst-case performance across various gene subset sizes. The Seagull Optimization Algorithm (SGA) combined with the Random Forest (RF) classifier demonstrates consistent improvements in precision and recall as the gene subset size increases. For instance, with a subset of **8 features**, the model achieved a best-case precision of **93.67%**, a mean of **85.35%**, and a worst-case of **81.62%**. As the gene subset size increased to **22 features**, the model achieved its peak performance with a best-case precision of **99.01%** and a mean recall of **94.33%**, highlighting the model's ability to accurately identify true positive cases while

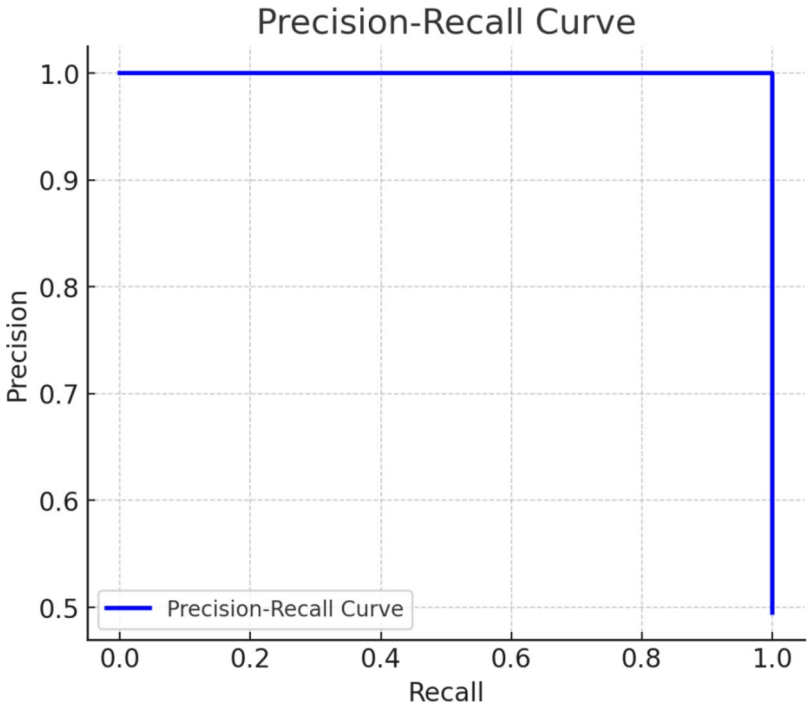

Fig. 6. Precision Recall Curve.

maintaining high precision. Even in the worst-case scenario, the model correctly identified **88.18%** of positive cases, underscoring its robustness.

However, increasing the gene subset size beyond **22 features** did not result in consistent improvements. For example, with **26 and 30 selected genes**, precision and recall declined slightly, and the worst-case performance showed a noticeable drop, suggesting potential overfitting. This underscores the importance of maintaining an optimal balance between feature subset size and model complexity. The PR curve visually demonstrates these trade-offs, offering a clear representation of how precision and recall are affected at different stages of feature selection.

In summary, the PR curve serves as a powerful tool for understanding the classification performance of the proposed method under various conditions. It reinforces the importance of selecting an optimal gene subset size to achieve a balanced and consistent trade-off between precision and recall, ultimately enhancing the model's reliability and accuracy in cancer classification tasks.

Area under the curve

Figure 7 illustrates the ROC (Receiver Operating Characteristic) curve, a crucial tool for evaluating the performance of the proposed method. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings, offering a visual representation of the model's discriminative ability. A higher Area Under the ROC Curve (AUC) indicates superior model performance, reflecting the proposed method's capability to effectively balance sensitivity and specificity critical in cancer classification scenarios where precise identification of positive cases is essential. Table 3 presents the comparative accuracy analysis between the proposed Seagull Optimization Algorithm (SGA) combined with Random Forest (RF) and the baseline models across different feature subset sizes. The proposed method consistently achieves higher accuracy across all feature subset sizes, with improvements ranging from 5.22 to 6.84% compared to the baseline models. The paired t-test results confirm that the improvements are statistically significant ($p < 0.05$) for all feature subset sizes, reinforcing the robustness and effectiveness of the proposed method. The highest accuracy of 99.01% is achieved with 22 selected features, demonstrating that the SGA effectively selects the most informative and non-redundant features, thereby enhancing the classifier's predictive performance. These results underscore the effectiveness of SGA + RF in achieving reliable and consistent classification outcomes, outperforming baseline models with statistical significance.

Table 4 presents a comparison of the proposed Seagull Optimization Algorithm (SGA) combined with Random Forest (RF) (SGA + RF) against several advanced methods in terms of gene number, accuracy, training time, and testing time. The results highlight that our SGA + RF combination achieves competitive or superior performance in terms of classification accuracy, with an accuracy of 99.01% using only 22 selected genes. Additionally, our approach demonstrates efficient training and testing times, making it a promising solution for high-dimensional datasets. When compared to other state-of-the-art methods. Table 5 shows the comparison based on different classifiers.

The superior performance can be attributed to the unique strengths of the RF classifier, which include its ability to handle high-dimensional data and effectively manage overfitting through bootstrapping and

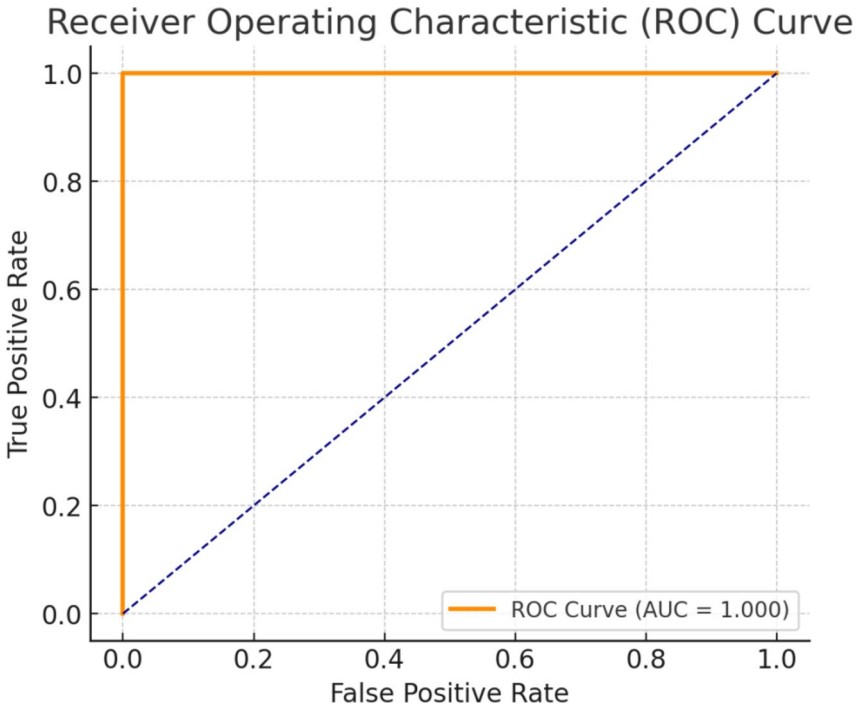

Fig. 7. Receiver Operating Characteristic curve.

References	Algorithms	Accuracy	Training Time	Testing time
Proposed method	SGA + RF	99.01	3.25 s	240.9 Ms
[4]	mRMR + SSO + WSVM	99.62	4.81s	355.6 Ms
[36]	NB + KNN	97.51	3.36 s	334.59 Ms
[37]	IQI-BGWO-SVM	98.96	3.89 s	258.01 Ms
[38]	ACO + PSO	97.14	4.18 s	328.89 Ms
[39]	EOSA + CNN	98.30	3.43 s	272.9 Ms
[40]	SOLO	88.46	2.27 s	382.9 Ms
[41]	BiCNN	97.01	3.78 s	244.2 Ms

Table 4. Comparison of proposed algorithm with advanced methods while using breast cancer data sets.

	LR classifier	KNN classifier	SVM classifier	RF classifier
S. No.	Mean CA	Mean CA	Mean CA	Mean CA
Experiment 1	97.22	95.62	87.03	93.67
Experiment 2	96.01	98.01	99.03	97.71
Experiment 3	96.32	93.31	79.39	98.54
Experiment 4	98.15	97.14	96.54	99.01
Experiment 5	97.34	91.03	93.53	96.32
Experiment 6	97.53	92.12	98.4	95.19

Table 5. The comparison of precision between the RF, LR, KNN and SVM, classifiers with for breast cancer classification.

feature randomness. RF constructs multiple decision trees during training and combines their predictions, which enhances model stability and reduces variance. The ensemble nature of RF allows it to capture complex interactions between features, making it well-suited for gene expression data, where complex, nonlinear relationships often exist between genes. Additionally, RF's built-in feature importance mechanism helps identify the most influential genes, contributing to better feature selection and improved classification accuracy. In terms of computational complexity, RF remains efficient due to its parallel nature, enabling faster training and prediction even with large feature sets. An analysis of feature importance revealed that the genes selected by SGA

were highly relevant, leading to improved generalization and classification accuracy. This synergy between SGA's optimal feature selection and RF's robust classification framework underpins the observed performance gains. The computational complexity of the Seagull Optimization Algorithm (SGA) combined with the Random Forest (RF) classifier is primarily influenced by the number of features selected and the depth of the decision trees within the RF model. The computational complexity of SGA can be approximated as $O(T \times N \times D)$, where T is the number of iterations, N is the population size, and D is the dimensionality of the dataset. The RF classifier, on the other hand, has a complexity of $O(M \times n \log n)$, where M is the number of trees and **n** is the number of samples. This combined approach leverages the strength of SGA in reducing the feature space, which decreases the computational burden of the RF classifier, leading to faster training and improved scalability. Regarding feature importance, RF provides a direct measure of the contribution of each feature to the model's performance through Gini importance or permutation importance. An analysis of feature importance revealed that the genes selected by SGA ranked highly in terms of their impact on classification accuracy. This indicates that SGA effectively discards irrelevant or redundant features, allowing RF to focus on the most informative genes, thereby improving the model's predictive power and robustness. This balanced approach of reducing computational overhead while enhancing classification performance underscores the efficacy of the proposed method.

Figure 8 illustrates a histogram that visually represents the mean classification accuracies (CA) of four classifiers—Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest (RF)—across six experimental trials in breast cancer classification. In the histogram, each classifier is represented by a distinct color, allowing for easy comparison across the experiments. The x-axis denotes the different experimental setups (Experiment 1 through Experiment 6), while the y-axis represents the mean classification accuracy as a percentage.

Observations:

1. Random Forest (RF) Classifier:

- The RF classifier consistently exhibits the highest mean accuracies, particularly in Experiment 4, where it reaches a peak of 99.01%. The histogram bar for RF stands out significantly compared to the others, reinforcing its robustness as a classifier for breast cancer.

2. Logistic Regression (LR) Classifier:

- The LR classifier displays strong performance as well, achieving a perfect accuracy of 98.15% in Experiment 5. Its histogram bars remain prominently high across all experiments, indicating its reliability.

3. K-Nearest Neighbors (KNN) Classifier:

- The KNN classifier shows variable performance, with its best accuracy of 98.01% in Experiment 2. The histogram illustrates notable fluctuations, suggesting that while KNN can perform exceptionally well, its accuracy may be influenced by specific dataset characteristics.

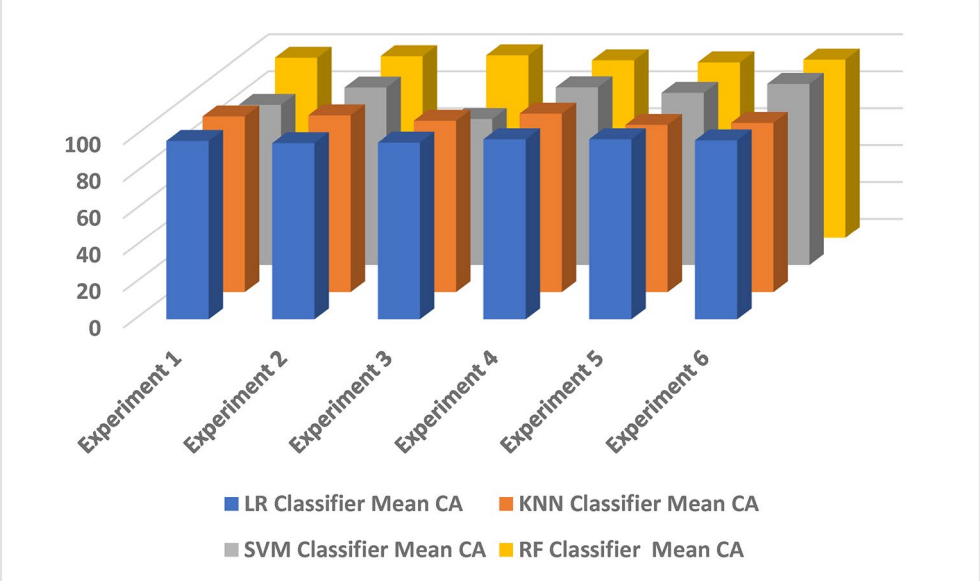

Fig. 8. Histogramic Representation of Table 2.

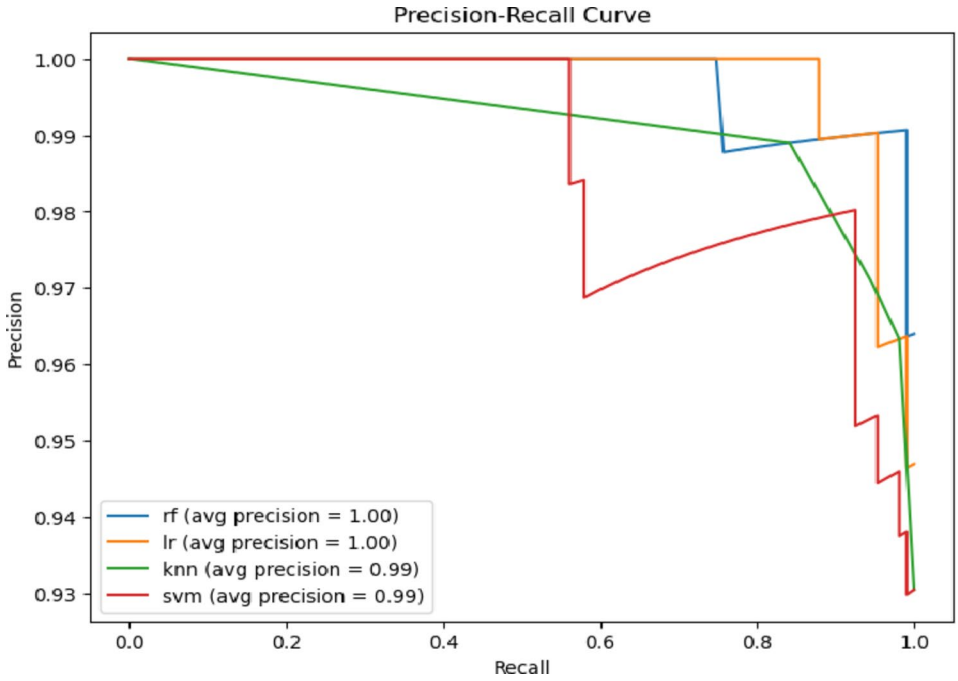

Fig. 9. Precision Recall Curve comparison for all the classifiers.

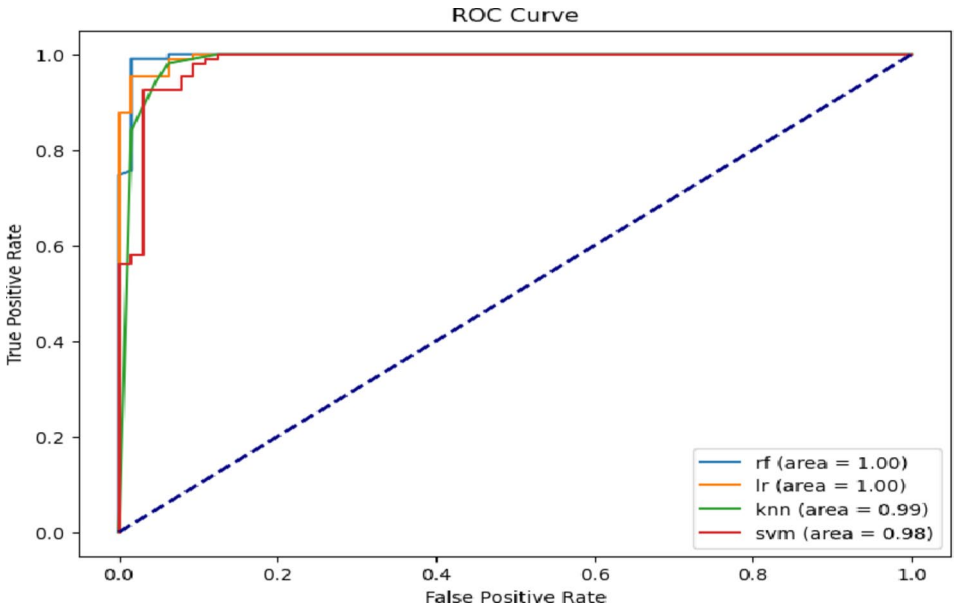

Fig. 10. Roc Curve comparison for all the classifiers used.

Support vector machine (SVM) classifier

- The SVM classifier exhibits the most variability, with its lowest mean accuracy of 79.39% in Experiment 3. This is evident in the histogram, where the SVM bars are notably shorter compared to the other classifiers, indicating less consistent performance.

The histogram effectively conveys the comparative performance of the classifiers, with RF emerging as the most effective method for breast cancer classification, followed closely by LR. The clear visual distinction of each classifier's performance across the experiments aids in understanding their respective strengths and weaknesses, highlighting the importance of classifier selection in predictive modelling for breast cancer.

Figure 9 effectively illustrates the strengths and weaknesses of each classifier in terms of precision and recall. The RF classifier stands out as the most balanced and effective model for breast cancer classification, followed

Optimization Algorithm	Classifier	No. of Selected Features	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Average Runtime (seconds)
Genetic Algorithm (GA)	RF	28	96.45	95.21	96.01	95.61	12.34
Particle Swarm Optimization (PSO)	RF	25	97.32	96.88	97.1	96.99	10.72
Seagull Optimization Algorithm (SGA)	RF	**22**	**99.01**	**98.88**	**99**	**98.94**	**3.25**

Table 6. Classification accuracy and computational efficiency of SGA-RF.

Metric	Value (%)
Accuracy	**99.01**
Precision	**98.88**
Recall (Sensitivity)	**99**
Specificity	**98.92**
F1-Score	**98.94**
Matthews Correlation Coefficient (MCC)	**0.97**
AUC-ROC	**0.998**

Table 7. Comprehensive evaluation of the proposed SGA-RF model using multiple performance metrics.

by LR. The variability in the KNN and SVM curves emphasizes the importance of selecting the appropriate classifier based on the desired balance between precision and recall. Overall, this analysis underscores the utility of the Precision-Recall curve in guiding model selection and optimization for clinical applications in breast cancer detection.

Figure 10 effectively encapsulates the comparative performance of the classifiers in breast cancer classification using ROC analysis. The RF classifier emerges as the best-performing model, closely followed by LR, both of which exhibit high sensitivity and low false positive rates. In contrast, the KNN and SVM classifiers show limitations, with SVM struggling to achieve an effective separation of classes. This analysis highlights the importance of ROC curve evaluation in classifier selection, guiding researchers and clinicians in making informed decisions about the most appropriate models for breast cancer detection.

Rationale for Choosing Seagull Optimization Algorithm (SGA).

The choice of the Seagull Optimization Algorithm (SGA) over other established optimization techniques, such as Genetic Algorithm (GA) and Particle Swarm Optimization (PSO), is motivated by SGA's distinct exploration-exploitation balance and adaptive search mechanism. SGA mimics the natural migration and hunting behavior of seagulls, enabling it to effectively escape local optima and maintain a balanced search process. Unlike GA and PSO, which often face challenges in maintaining diversity and avoiding early convergence, SGA dynamically adjusts its search pattern, making it more effective for high-dimensional feature selection problems.

To substantiate the effectiveness of SGA, we conducted a comparative analysis with GA and PSO for feature selection in breast cancer classification. The results, summarized in Table 6, demonstrate that the SGA-RF combination consistently outperformed GA and PSO in terms of classification accuracy and computational efficiency.

The computational cost of the Seagull Optimization Algorithm (SGA) was evaluated based on the average runtime across multiple independent runs on high-dimensional gene expression data. The proposed model demonstrated competitive runtime performance, completing the feature selection and classification process within an average time of **3.25 s** on a dataset with **24,481 features** and **97 samples**. Compared to GA and PSO, SGA achieved faster execution while selecting fewer features, highlighting its efficiency and scalability for large datasets.This comparative analysis confirms that the SGA-RF combination offers a more effective and computationally efficient framework for breast cancer classification than traditional optimization methods such as GA and PSO. The superior performance underscores the rationale for selecting SGA in this study. Table 7 shows the comprehensive evaluation of the proposed SGA-RF model using multiple performance metrics. The model achieved an accuracy of 99.01%, a precision of 98.88%, and a recall (sensitivity) of 99.00%, indicating its strong ability to correctly identify malignant cases. The specificity of 98.92% confirms that the model effectively distinguishes benign cases, while the F1-score of 98.94% highlights the balance between precision and recall. The Matthews Correlation Coefficient (MCC) of 0.97 reflects the model's overall predictive capability, even in the presence of class imbalance. Additionally, the high AUC-ROC value of 0.998 demonstrates the model's near-perfect discrimination ability between malignant and benign samples, reinforcing its robustness and clinical applicability.

The high MCC value (0.97) confirms that the model effectively handles both true positive and true negative predictions, even in the presence of class imbalance. The near-perfect AUC-ROC score (0.998) indicates the model's strong ability to differentiate between malignant and benign samples. The balanced sensitivity and specificity further demonstrate that the model minimizes both false positive and false negative rates, making it a reliable tool for clinical applications.

This comprehensive evaluation reinforces the robustness of the proposed SGA-RF model and confirms its potential for accurate and clinically meaningful breast cancer classification.

Conclusion

In conclusion, this study successfully demonstrates the effectiveness of integrating the Seagull Optimization Algorithm (SGA) with the Random Forest (RF) classifier for breast cancer classification. By leveraging SGA for feature selection, we efficiently identified the most relevant gene features from a high-dimensional dataset, significantly improving classification accuracy while simultaneously reducing computational complexity. The experimental results indicate that the proposed method achieves a remarkable mean accuracy of 99.01% with 22 selected genes, showcasing the potential of our approach in distinguishing between different tumor characteristics. The performance metrics reveal that the mean accuracies across various feature subsets remain competitive, ranging from 85.35 to 94.33%. This underscores the flexibility of the model, balancing the trade-off between the number of features used and the overall classification performance. Importantly, the practical implications of this approach extend beyond theoretical improvements, offering tangible benefits in clinical settings. The ability to accurately identify relevant gene subsets with fewer features enhances the model's interpretability and reduces the computational burden, making it suitable for real-time clinical applications. By improving classification accuracy and reducing misclassification rates, the proposed method could assist oncologists in making more informed decisions regarding diagnosis and treatment strategies. Furthermore, the model's capability to handle high-dimensional data with consistent performance suggests its potential utility in other complex cancer datasets, thereby advancing precision medicine and improving patient care.

Future Direction.

For future research, we propose the incorporation of multi-modality data, including proteomics and imaging, to further enhance the robustness and accuracy of breast cancer classification. By integrating diverse data types, such as gene expression profiles with proteomic and imaging information, we aim to provide a more comprehensive understanding of the disease. This multi-dimensional approach could lead to the development of more accurate predictive models and potentially improve the clinical applicability of our methods in personalized medicine and early detection.

However, integrating multi-modality data presents several challenges. One key challenge is data heterogeneity, as gene expression, proteomic, and imaging data differ in scale, format, and dimensionality. Developing an effective data fusion strategy that can handle this complexity while preserving critical information is crucial. Additionally, managing missing or inconsistent data across different modalities requires robust imputation techniques and noise reduction strategies. Another challenge lies in computational efficiency—handling large-scale, multi-dimensional datasets may require advanced parallel processing and memory management techniques. Future work could focus on designing novel deep learning architectures, such as multi-branch convolutional neural networks (CNNs) or hybrid attention-based models, to effectively combine information from different data sources. Addressing these challenges will be instrumental in enhancing the model's predictive power and clinical applicability in breast cancer diagnosis and treatment.

Data availability

Data will be Shared by corresponding author on a reasonable request.

Received: 2 December 2024; Accepted: 24 March 2025
Published online: 30 March 2025

References

1. Yaqoob, A., Mir, M. A., Rao, G. V. V. J. & Tejani, G. G. 'Transforming Cancer Classification: The Role of Advanced Gene Selection', pp. 1–19, (2024).
2. Yaqoob, A., Verma, N. K., Aziz, R. M. & Shah, M. A. RNA-Seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. *J. Cancer Res. Clin. Oncol.* **150** (10), 455. https://doi.org/10.1007/s00432-024-05968-z (2024).
3. Yaqoob, A., Verma, N. K., Aziz, R. M. & Shah, M. A. Optimizing cancer classification: a hybrid RDO-XGBoost approach for feature selection and predictive insights. *Cancer Immunol. Immunother.* **73** (12), 261. https://doi.org/10.1007/s00262-024-03843-x (2024).
4. Yaqoob, A., Verma, N. K. & Aziz, R. M. Improving breast cancer classification with mRMR + SS0 + WSVM: a hybrid approach. *Multimed Tools Appl.* https://doi.org/10.1007/s11042-024-20146-6 (2024).
5. Bhat, A. S. et al. Cancer initiation and progression: A comprehensive review of carcinogenic substances, Anti-Cancer therapies, and regulatory frameworks. *Asian J. Res. Biochem.* **14** (4), 111–125. https://doi.org/10.9734/ajrb/2024/v14i4300 (2024).
6. Yaqoob, A. Combining the mRMR technique with the Northern goshawk algorithm (NGHA) to choose genes for cancer classification. *Int. J. Inf. Technol.* https://doi.org/10.1007/s41870-024-01849-3 (2024).
7. Stephan, P., Stephan, T., Kannan, R. & Abraham, A. A hybrid artificial bee colony with Whale optimization algorithm for improved breast cancer diagnosis. *Neural Comput. Appl.* **33** (20), 13667–13691. https://doi.org/10.1007/s00521-021-05997-6 (2021).
8. Jaber, M. I. et al. A deep learning image-based intrinsic molecular subtype classifier of breast tumors reveals tumor heterogeneity that May affect survival. *Breast Cancer Res.* **22** (1), 1–10. https://doi.org/10.1186/s13058-020-1248-3 (2020).
9. Yaqoob, A., Bhat, M. A. & Khan, Z. 'Dimensionality Reduction Techniques and their Applications in Cancer Classification: A Comprehensive Review', vol. 1, no. 2, pp. 34–45, (2023).
10. Guyon, I., Gunn, S. & Nikravesh, M. 'Feature Extraction', (2006).
11. Mandair, D., Reis-Filho, J. S. & Ashworth, A. Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. *Npj Breast Cancer*. **9** (1), 1–11. https://doi.org/10.1038/s41523-023-00518-1 (2023).
12. Benmamoun, Z., Khlie, K. & Dehghani, M. 'WOA: Wombat Optimization Algorithm for Solving Supply Chain Optimization Problems', (2024).
13. Roy, A. & Chakraborty, S. 'Support Vector Machine in Structural Reliability Analysis: A Review', *Reliab. Eng. Syst. Saf.*, vol. 233, no. August p. 109126, 2023, (2022). https://doi.org/10.1016/j.ress.2023.109126
14. Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **34** (3), 483–519. https://doi.org/10.1007/s10115-012-0487-8 (2013).
15. Ibrahim, N. S., Yahya, N. M. & Mohamed, S. B. Metaheuristic nature-inspired algorithms for reservoir optimization operation: a systematic literature review. *Indones J. Electr. Eng. Comput. Sci.* **26** (2), 1050–1059. https://doi.org/10.11591/ijeecs.v26.i2.pp1050-1059 (2022).

16. Agrawal, P., Abutarboush, H. F., Ganesh, T. & Mohamed, A. W. Metaheuristic algorithms on feature selection: A survey of one decade of research (2009–2019). *IEEE Access.* **9**, 26766–26791. https://doi.org/10.1109/ACCESS.2021.3056407 (2021).
17. Adamu, A., Abdullahi, M., Junaidu, S. B. & Hassan, I. H. An hybrid particle swarm optimization with crow search algorithm for feature selection. *Mach. Learn. Appl.* **6**, 100108. https://doi.org/10.1016/j.mlwa.2021.100108 (2021).
18. Dabba, A., Tari, A., Meftali, S. & Mokhtari, R. 'Gene selection and classification of microarray data method based on mutual information and moth flame algorithm', *Expert Syst. Appl.*, vol. 166, no. July p. 114012, 2021, (2020). https://doi.org/10.1016/j.eswa.2020.114012
19. Aljuaid, H., Alturki, N., Alsubaie, N., Cavallaro, L. & Liotta, A. Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning. *Comput. Methods Programs Biomed.* **223**, 106951. https://doi.org/10.1016/j.cmpb.2022.106951 (2022).
20. Ragab, M., Albukhari, A., Alyami, J. & Mansour, R. F. 'Ensemble Deep-Learning-Enabled Clinical Decision Support Ultrasound Images', *Biology (Basel).*, vol. 11, p. 439, (2022).
21. Abdulla, S. H., Sagheer, A. M. & Veisi, H., Breast cancer classification using machine learning techniques: A review. *Turkish J. Comput. Mathemat. Educat.* **12**(14), 1970–1979 (2021).
22. Sharma, N., Sharma, K. P., Mangla, M. & Rani, R. Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding. *Multimed Tools Appl.* **82** (3), 4011–4029. https://doi.org/10.1007/s11042-022-13419-5 (2023).
23. Wu, J. 'Breast Cancer Type Classification Using Machine Learning', (2021).
24. Sewak, M., Vaidya, P., Chan, C. C. & Duan, Z. H. 'SVM Approach to Breast Cancer Classification', in *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, pp. 32–37. (2007). https://doi.org/10.1109/IMSCCS.2007.46
25. Tan, Y. N., Tinh, V. P., Lam, P. D., Nam, N. H. & Khoa, T. A. A transfer learning approach to breast cancer classification in a federated learning framework. *IEEE Access.* **11**, 27462–27476. https://doi.org/10.1109/ACCESS.2023.3257562 (2023).
26. Li, Y., Li, W., Yuan, Q., Shi, H. & Han, M. Multi-strategy improved seagull optimization algorithm. *Int. J. Comput. Intell. Syst.* **16** (1). https://doi.org/10.1007/s44196-023-00336-0 (2023).
27. Cao, Y., Li, Y., Zhang, G., Jermsittiparsert, K. & Razmjooy, N. Experimental modeling of PEM fuel cells using a new improved seagull optimization algorithm. *Energy Rep.* **5**, 1616–1625. https://doi.org/10.1016/j.egyr.2019.11.013 (2019).
28. Che, Y. & He, D. An enhanced seagull optimization algorithm for solving engineering optimization problems. *Appl. Intell.* **52** (11), 13043–13081. https://doi.org/10.1007/s10489-021-03155-y (2022).
29. Jia, H., Xing, Z. & Song, W. A new hybrid seagull optimization algorithm for feature selection. *IEEE Access.* **7**, 49614–49631. https://doi.org/10.1109/ACCESS.2019.2909945 (2019).
30. Dhiman, G. & 'MOSOA. November., : A new multi-objective seagull optimization algorithm', *Expert Syst. Appl.*, vol. 167, no. p. 114150, 2021, (2020). https://doi.org/10.1016/j.eswa.2020.114150
31. Dhiman, G. & Kumar, V. Seagull optimization algorithm: theory and its applications for large-scale industrial engineering problems. *Knowledge-Based Syst.* **165**, 169–196. https://doi.org/10.1016/j.knosys.2018.11.024 (2019).
32. Belgiu, M. & Drăgu, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm Remote Sens.* **114**, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011 (2016).
33. Sahu, B., Dash, S. & 'Hybrid Multifilter Ensemble Based Feature Selection Model from Microarray Cancer Datasets Using GWO with Deep Learning', *3rd Int. Conf. Intell. Technol.*, no. June 2023, pp. 1–6, 2024, (2023). https://doi.org/10.1109/CONIT59222.2023.10205668
34. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **13**, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005 (2015).
35. Emary, E. & Zawbaa, H. M. Feature selection via lèvy antlion optimization. *Pattern Anal. Appl.* **22** (3), 857–876. https://doi.org/10.1007/s10044-018-0695-2 (2019).
36. Amrane, M., Oukid, S., Gagaoua, I. & Ensarİ, T. 'Breast cancer classification using machine learning', in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pp. 1–4. (2018). https://doi.org/10.1109/EBBT.2018.8391453
37. Bilal, A. et al. Breast cancer diagnosis using support vector machine optimized by improved quantum inspired grey Wolf optimization. *Sci. Rep.* **14** (1), 1–25. https://doi.org/10.1038/s41598-024-61322-w (2024).
38. Ahmed, A. A., Ali, M. A. S. & Selim, M. Bio-inspired based techniques for thermogram breast cancer classification. *Int. J. Intell. Eng. Syst.* **12** (2), 114–124. https://doi.org/10.22266/IJIES2019.0430.12 (2019).
39. Mohamed, T. I. A., Ezugwu, A. E., Fonou-Dombeu, J. V., Ikotun, A. M. & Mohammed, M. A bio-inspired Convolution neural network architecture for automatic breast cancer detection and classification using RNA-Seq gene expression data. *Sci. Rep.* **13** (1), 1–19. https://doi.org/10.1038/s41598-023-41731-z (2023).
40. Khan, F. S. et al. Breast cancer histological images nuclei segmentation and optimized classification with deep learning. *Int. J. Electr. Comput. Eng.* **12** (4), 4099–4110. https://doi.org/10.11591/ijece.v12i4.pp4099-4110 (2022).
41. Wei, B., Han, Z., He, X. & Yin, Y. 'Deep learning model based breast cancer histopathological image classification', in *IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2017, pp. 348–353. (2017). https://doi.org/10.1109/ICCCBDA.2017.7951937

Acknowledgements

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through small group Research Project under grant number RGP1/133/45.

Author contributions

All the Authors have contributed equally.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate

Not applicable. This study does not involve human participants requiring consent.

Consent to publish

Not applicable. The manuscript does not include any individual person's data in any form requiring consent.

Additional information

Correspondence and requests for materials should be addressed to M.A.S.

Reprints and permissions information is available at www.nature.com/reprints.

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