Cancer Immunology, Immunotherapy (2024) 73:261
https://doi.org/10.1007/s00262-024-03843-x

RESEARCH

Optimizing cancer classification: a hybrid RDO-XGBoost approach for feature selection and predictive insights

Abrar Yaqoob[1] · Navneet Kumar Verma[1] · Rabia Musheer Aziz[2] · Mohd Asif Shah[3,4,5]

Received: 28 August 2024 / Accepted: 20 September 2024 / Published online: 9 October 2024

Abstract

The identification of relevant biomarkers from high-dimensional cancer data remains a significant challenge due to the complexity and heterogeneity inherent in various cancer types. Conventional feature selection methods often struggle to effectively navigate the vast solution space while maintaining high predictive accuracy. In response to these challenges, we introduce a novel feature selection approach that integrates Random Drift Optimization (RDO) with XGBoost, specifically designed to enhance the performance of cancer classification tasks. Our proposed framework not only improves classification accuracy but also offers valuable insights into the underlying biological mechanisms driving cancer progression. Through comprehensive experiments conducted on real-world cancer datasets, including Central Nervous System (CNS), Leukemia, Breast, and Ovarian cancers, we demonstrate the efficacy of our method in identifying a smaller subset of unique and relevant genes. This selection results in significantly improved classification efficiency and accuracy. When compared with popular classifiers such as Support Vector Machine, K-Nearest Neighbor, and Naive Bayes, our approach consistently outperforms these models in terms of both accuracy and F-measure metrics. For instance, our framework achieved an accuracy of 97.24% in the CNS dataset, 99.14% in Leukemia, 95.21% in Ovarian, and 87.62% in Breast cancer, showcasing its robustness and effectiveness across different types of cancer data. These results underline the potential of our RDO-XGBoost framework as a promising solution for feature selection in cancer data analysis, offering enhanced predictive performance and valuable biological insights.

Keywords Random drift optimization · XGBoost · Feature selection · Cancer classification · Microarray data analysis

Introduction

Human genetic data have the potential to aid in the detection and categorization of a wide range of disorders, including cancer. Microarray analysis of gene expression levels in various individuals is one of the most precise technologies in this area. Each sample is represented by a row, and each gene or trait is represented in a column, making the microarray data a matrix with thousands of columns and a few hundred rows. Computational costs, the ability to generalize classifications, and their efficacy in predicting fresh microarray samples are all impacted by the huge number of genes and small number of samples [1, 2]. However, due to the large number of features, it is also possible that genes with no obvious connection to one another influence the development of predictive models. Since, biologically speaking, only a small subset of genes truly has a role in the disease, the remainder merely act as "background noise" that can obscure the effect of the first subset. Adding more

✉ Abrar Yaqoob
abraryaqoob77@gmail.com

✉ Mohd Asif Shah
m.asif@kardan.edu.af

1 VIT Bhopal University's School of Advanced Science and Language, Located at Kothrikalan, Sehore, Bhopal 466114, India

2 Planning Department, State Planning Institute (New Division), Lucknow, Utter Pradesh 226001, India

3 Department of Economics, Kardan University, Parwane Du, 1001 Kabul, Afghanistan

4 Division of Research and Development, Lovely Professional University, Phagwara, Punjab 144001, India

5 Centre of Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab 140401, India

genes to the microarray dataset, however, will only serve to confuse the classifier and reduce the classification accuracy. Therefore, current efforts are focused on enhancing classification accuracy by introducing novel solutions, minimizing the size of microarrays, and eliminating unnecessary noise, irrelevant information, and redundant data [3].

Researchers are increasingly interested in employing meta-innovative optimization algorithms due to the rise in data dimensions and the fact that the problem of feature selection and classification of microarray data is one of the hardest non-polynomial optimization problems. In Selecting unique genes and improving classifier performance both depend on using an appropriate optimization strategy. In other words, most of these studies aim to improve classification accuracy, with secondary aims including things like minimizing the number of features and eliminating. The effectiveness of classifiers can be greatly improved by utilizing multi-objective optimization methods, in addition to minimizing the number of features and paying attention to the issue of redundancy between them [4–6].

Feature selection plays a pivotal role in the analysis of cancer data, aiding in the identification of crucial biomarkers and enhancing the performance of predictive models [7–9]. Among the myriad of techniques available, Random Drift Optimization (RDO) emerges as a promising nature-inspired algorithm that mimics the evolutionary process of biological systems. By leveraging principles from natural selection and random drift, RDO efficiently explores the solution space, identifying subsets of features that are most informative for classification tasks [10]. In this study, we propose the integration of RDO with XGBoost, a powerful gradient boosting algorithm renowned for its exceptional performance in various machine learning tasks, including cancer diagnosis. By combining the feature selection capabilities of RDO with the predictive prowess of XGBoost, we aim to develop a robust framework for cancer data analysis that not only enhances classification accuracy but also provides valuable insights into the underlying biological mechanisms driving cancer progression. In this paper, we elucidate the methodology behind this integration and demonstrate its efficacy through comprehensive experiments on real-world cancer datasets [11].

The purpose of this article is to introduce a novel classification scheme aimed at identifying differentially expressed genes associated with illness occurrence. We utilize the RDO algorithm for this purpose, facilitating the selection of relevant genes efficiently. Additionally, our study focuses on selecting representative samples for training classifiers, recognizing the pivotal role of training data quality in classifier effectiveness. To evaluate the proposed approach, we employ three more popular classifiers Support Vector Machine (SVM), Nearest Neighbor (KNN), and Naive Bayes (NB) and compare our method's outcomes with those obtained using techniques described in other articles. The results demonstrate that our suggested strategy can identify a smaller subset of unique genes, leading to improved classification efficiency. These categories are selected due to their prevalence in discussions on data mining and feature selection.

In most optimization problems, the genetic algorithm often exhibits superior accuracy in achieving the global optimal solution, coupled with suitable convergence speed, especially when compared to other meta-innovative algorithms based on collective intelligence. Consequently, this optimization approach extends beyond traditional domains to applications in image processing, machine learning, and various engineering branches. While the RDO offers benefits like fast convergence and accurate global solution finding, it may encounter challenges such as getting stuck in local optima, particularly in high-dimensional problems with constant control parameters. Currently, efforts are underway to optimize the algorithm further. In summary, the authors contribute to this study by introducing a new classification scheme, utilizing the RDO for gene selection, and evaluating classifier performance. Additionally, they discuss challenges and potential improvements in the RDO algorithm, contributing to ongoing optimization endeavors. Determine which features are frequently occurring and strongly correlated with one another.

- Select features without requiring conventional methods.
- Reduce the number of features that need to be selected significantly.
- Design a multi-objective feature selection strategy based on error classification, the proportion of the selected features, and the value of features.

Paper organization

The remainder of the article is structured as follows:

Section 2: Challenges and Solutions of Existing Works: This section delves into the challenges encountered by existing methods in feature selection for cancer data analysis and discusses potential solutions. We review the limitations of conventional feature selection techniques and highlight the need for innovative approaches to address issues such as high dimensionality, noisy data, and class imbalance. Furthermore, we examine recent advancements in feature selection methodologies and discuss how they attempt to overcome these challenges. By identifying the gaps in existing works, we set the stage for proposing our novel approach that integrates Random Drift Optimization (RDO) with XGBoost to tackle the inherent complexities of cancer data analysis. Section 3: Materials and Methods: In this section, we provide a detailed explanation of the proposed methodology, including the integration of RDO with XGBoost for feature

selection in cancer data analysis. We outline the algorithmic steps involved in our approach, including population initialization, fitness evaluation using XGBoost, evolutionary operations (selection, crossover, mutation), and selection of the best feature subset. Additionally, we discuss the experimental setup, including parameter settings and evaluation metrics, to ensure a thorough understanding of the proposed method's implementation. Section 4: Experimental Setup and Data Description: This section comprehensively describes the experimental setup employed in our study, including the datasets used for evaluation and the preprocessing steps applied. We provide detailed information about the cancer datasets, including their sources, sample sizes, and class distributions. Moreover, we discuss the preprocessing techniques employed to handle issues such as missing values, data normalization, and feature scaling. Additionally, we outline the experimental design, including partitioning of the datasets into training and testing sets, cross-validation strategies, and parameter tuning procedures, to ensure rigorous evaluation of the proposed method. Section 5: Results and Discussion: In this section, we present the results of our experiments and provide a detailed discussion of the findings. We report performance metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) obtained by our proposed approach and compare them with those of existing classifiers (e.g., Support Vector Machine, Nearest Neighbor, Naive Bayes). Furthermore, we analyze the selected feature subsets and discuss their biological relevance in the context of cancer diagnosis and prognosis. Through a comprehensive examination of the results, we highlight the effectiveness and interpretability of our proposed method in enhancing classification accuracy and revealing insights into cancer biology. Section 6: Conclusions: In the final section of the article, we summarize the key findings of our study and draw conclusions based on the experimental results and discussions presented earlier. We discuss the implications of our findings for cancer research and clinical practice, emphasizing the potential of the proposed approach for improving the accuracy and interpretability of predictive models in cancer data analysis. Furthermore, we identify avenues for future research and highlight the importance of continued advancements in feature selection methodologies for addressing the evolving challenges in cancer diagnosis and treatment.

Related works

The DNA in the nucleus of every cell stores the "program" for the development of the organism. Both coding and non-coding regions exist in DNA. Proteins are the building blocks of life, and their structure is determined by coding regions, also called genes. There are two stages in the process of translating a gene's instructions into a functional protein. DNA microarrays, an advanced tool in molecular genetics, allow us to see the whole picture within a cell by measuring the expression of thousands of genes all at once [12].

With the goal of lowering the number of features, eliminating irrelevant characteristics, and filtering out noisy data, feature selection algorithms are employed in the field of data mining. Filtering, packing, and integrating are the three main types of feature selection techniques [13]. In filter approaches, the set of genes effective in the occurrence of the disease is determined by first calculating the strength of each gene in separating the samples using a statistical index, and then selecting the genes with superior separation power based on the established criteria. The Relief method is another filter-based technique for dealing with noisy and multi-class data. The method selects a sample at random from the pool of possible samples, then calculates weights based on the dissimilarity between the selected sample and the sample in the common class and the sample in another class [14].

The IGWO-CNN prediction framework is a new approach to diabetic retinopathy diagnosis that integrates convolutional neural networks (CNNs) with improved gray wolf optimization (IGWO). Among other intelligence approaches, the Gray Wolf Optimizer (GWO) stands out for its capacity to simplify and convenience of use while yet delivering adequate discovery and exploitation during search, as well as its extensive tuning capabilities, scalability, and most significantly, its ability to ensure convergence speed. The throng is victorious [15]. A genetic algorithm (GA) was employed to produce a variety of starting places in the suggested approach. A reliable approach to the diagnosis and categorization of lung cancer has been suggested by researchers. After applying a weighted filter and the gray wolf optimization method to the images, they used segmentation techniques such basin transformation and dilation procedures to reduce image noise [16].

The feature selection of microarray data is carried out using a filtering method termed MASSIVE, which was proposed in reference [17] and is based on a criterion from information theory known as disr. In reference, [18], we pick a subset of genes that generates the maximum efficiency in the categorization. First, the best genes in each block are selected based on their role in the classification, and then, to find the optimal subset, the genes are compared to one another [19]. Although numerous methods and algorithms have been presented to choose unique and effective genes in the detection of cancer diseases, some of these methods have disadvantages, such as ignoring the redundancy of genes or raising processing costs, which have been previously addressed. In addition to increasing the number of

calculations required, selecting genes that are redundant will reduce the effectiveness of classifiers because doing so will not yield any new insights. It is also possible that these genes have noise in the background. The clustering approach is useful for reducing redundant genes in the final set of candidates. Fuzzy clustering was employed in [20], to filter out duplicate genes before feature selection. However, the author tries to eliminate genes with identical expression patterns after selecting them using the hierarchical clustering method. Table 1 presents a comparison of various feature selection algorithms used in the study, including their key characteristics such as Algorithm, Methodology, classifier, features size, Objective Functions and limitations. This table serves as a reference for understanding the different approaches employed for feature selection in the context of breast cancer classification.

Materials and methods

Random drift optimization (RDO)

Random Drift Optimization (RDO) is a nature-inspired optimization algorithm that simulates the evolutionary process observed in biological systems. Drawing inspiration from natural selection and random drift, RDO efficiently explores the solution space by iteratively updating candidate solutions based on their fitness. In RDO, candidate solutions evolve over successive generations through a combination of deterministic and stochastic processes, allowing for a balance between exploration and exploitation. The algorithm maintains a population of candidate solutions and employs mutation and crossover operators to introduce diversity and foster exploration. Additionally, random drift, akin to genetic drift in biological populations, introduces stochasticity to the search process, enabling the algorithm to escape local optima and discover novel solutions. RDO has demonstrated effectiveness in various optimization tasks, including feature selection, where it efficiently identifies subsets of features that are most informative for predictive modeling [26].

$$\text{Fitness}\,(x) = f(x) \tag{1}$$

XGBoost

XGBoost, short for Extreme Gradient Boosting, is a powerful ensemble learning algorithm renowned for its exceptional performance in a wide range of machine learning tasks, particularly in structured/tabular data analysis. Built on the principles of gradient boosting, XGBoost sequentially builds an ensemble of weak learners, typically decision trees, by optimizing a differentiable loss function. The key innovation of XGBoost lies in its efficient implementation, which incorporates regularization techniques to prevent overfitting and parallelization strategies to expedite training. By iteratively improving the ensemble through gradient descent optimization, XGBoost achieves state-of-the-art performance in terms of predictive accuracy, scalability, and computational efficiency. Its versatility and effectiveness have made XGBoost a popular choice for various applications, including regression, classification, and ranking tasks across domains such as finance, healthcare, and online advertising [27].

$$\text{Loss}\,(y, \widehat{y}) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \mho(f_x) \tag{2}$$

Proposed integration of RDO with XGBoost

In this study, we propose the integration of Random Drift Optimization (RDO) with XGBoost to develop a robust framework for feature selection in cancer data analysis. The motivation behind this integration stems from the complementary strengths of RDO in efficient exploration of solution space and XGBoost in predictive modeling. By combining RDO's ability to identify informative feature subsets with XGBoost's exceptional performance in classification tasks, we aim to enhance the accuracy of cancer diagnosis while unraveling valuable insights into the underlying biological mechanisms driving cancer progression. In the proposed approach, RDO is employed to search for an optimal subset of features from the high-dimensional space of gene expression data, while XGBoost serves as the classification algorithm to predict cancer outcomes based on the selected features. Through this synergistic integration, we expect to achieve improved classification efficiency and interpretability compared to traditional feature selection methods and standalone classification algorithms [28].

$$\text{Objective}\,(F) = \sum_{i=1}^{n} \text{loss}(y_i, \hat{y}_i) + \sum_{k=1}^{K} \mho(f_x) + \sum_{i=1}^{n} \gamma_i \tag{3}$$

Integration of random drift optimization with XGBoost

In this study, we propose a novel approach for feature selection by integrating Random Drift Optimization (RDO) with XGBoost, a powerful gradient boosting algorithm widely

Table 1 Feature selection algorithms

Year/Reference	Algorithm	Compared Methods	Classifier	Features-size	Objective Function(s)	Disadvantage(s)
2023/ [21]	MOPSO	Multiobjective evolutionary algorithm based on decomposition (MOEA/D) Non-dominated Sorting Genetic Algorithm (NSGA-III) Ranks the features PSO (RFP-SOFS)	Naive Bayes SVM KNN	60–856	Classification error Number of features	Performs poorly on low-dimensional datasets. The method ignores interactions between features, resulting in a subset that has some redundant features
2023/[22]	HECSPSO	Binary PSO (BPSO) Chaotic maps based on binary PSO (CBPSO) Variable-length PSO (VLPSO)	SVM	8–617	Classification accuracy Selected feature subset size Number of iterations	Due to the fact that it calculates the fitness of each particle repeatedly, it requires a lot of computational time and energy The number of features selected is still very large
2023/ [23]	Binary Lévy flight gray wolf optimizer (BFLGWO)	BGWO BFLGWO	KNN	9–60	Classification error The number of features	High time consumption The number of selected features is high
2023/ [24]	BDGMOEA	Duplication analysis-based evolutionary algorithm (DAEA) Variable granularity search-based multi-objective (VGS-MOEA)	SVM Random forest KNN	4434–22,283	Classification performance Training time	Interaction makes it hard to identify features Feature selection on individual datasets is too extensive and interpretability is poor
2023/ [25]	Feature selection method considering interaction, redundancy and complementarity (FSIRC)	Dense Subgraph-based Feature Selection method (DSFS) Depth-based feature importance of isolation forest (DIFFI)	Solation Forest (iForest)	6–1555	Higher average reduction rate of selected features	In high-dimensional and large-scale datasets, its computational complexity is affected both by dimensionality and number of instances

recognized for its effectiveness in various machine learning tasks, including cancer diagnosis. The integration of RDO with XGBoost aims to harness the complementary strengths of both techniques, leveraging the efficient exploration of solution space offered by RDO and the predictive prowess of XGBoost. This integrated framework is designed to enhance classification accuracy while providing valuable insights into the underlying biological mechanisms driving cancer progression [29]. Algorithm 1 shows the Pseudocode of the proposed method.

Steps how Random Drift Optimization works and how it can be applied to feature selection.

- Initialization: The algorithm starts by initializing a population of candidate solutions. In the context of feature selection, each candidate solution represents a subset of features from the original feature set.
- Evaluation: Each candidate solution (feature subset) is evaluated using an objective function or evaluation metric that quantifies the quality or performance of the solution. This evaluation typically involves training and validating a predictive model (e.g., classification or regression model) using the selected features and measuring its performance.
- Random Drift: The core concept of Random Drift Optimization is the introduction of random perturbations or drifts to the current solutions. These drifts simulate the random forces or fluctuations present in nature that influence the movement or evolution of particles or organisms. In the context of feature selection, random drifts can be applied to the selected feature subsets to explore the search space.
- Exploration and Exploitation: By introducing random drifts to the feature subsets, the algorithm explores the space of possible feature combinations. This exploration allows the algorithm to discover new, potentially better feature subsets that may improve the overall performance of the model. Additionally, the algorithm exploits promising regions of the search space by focusing on feature subsets with higher performance.
- Selection: After applying random drifts and evaluating the modified feature subsets, the algorithm selects the best-p

The diagrammatic view of the proposed methodology is shown in Fig. 1

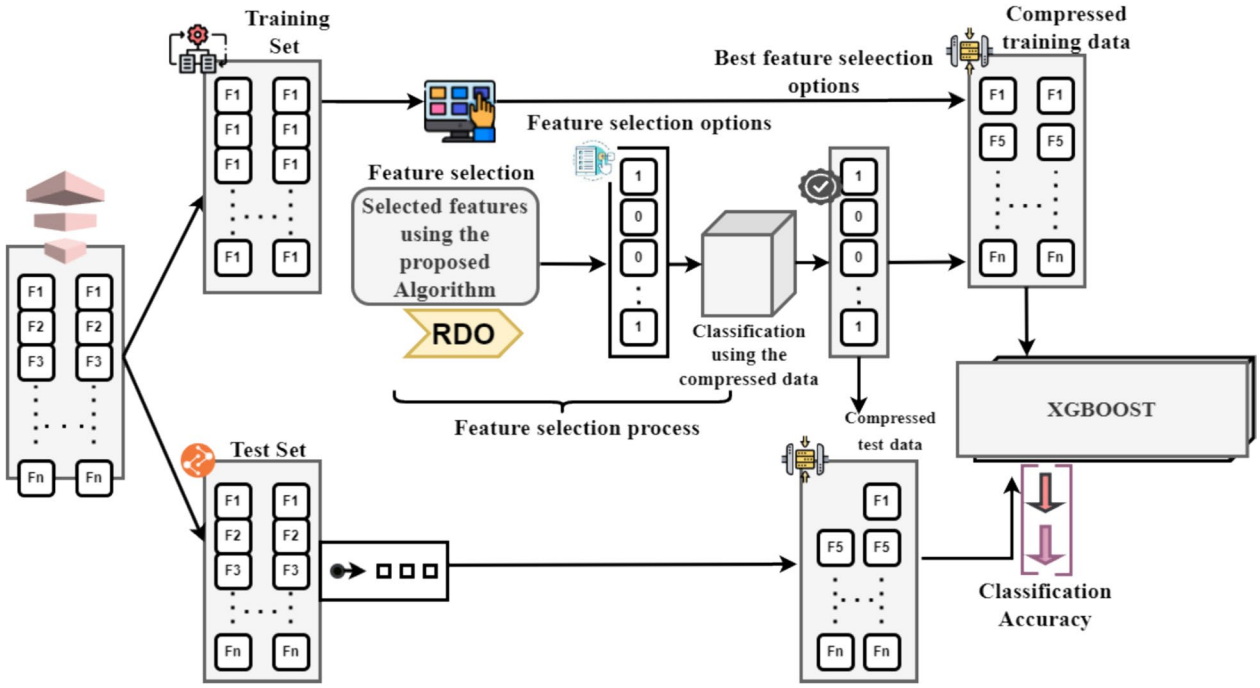

Fig. 1 Diagrammatic view of the proposed Approach

Algorithm 1: Pseudocode for the Proposed Methodology

```
Input:
- Dataset: Input data containing gene expression profiles and corresponding cancer labels
- Parameters:
  - Population size (pop_size)
  - Number of generations (num_generations)
  - Mutation probability (mutation_prob)
  - Crossover probability (crossover_prob)
  - XGBoost parameters (e.g., learning rate, maximum depth)

Output:
- Selected features
- Trained XGBoost classifier

Procedure ProposedMethod(Dataset):
    Initialize population P with random feature subsets
    Evaluate fitness of each feature subset in P using XGBoost

    for generation = 1 to num_generations do:
        Select parents from P based on fitness (e.g., tournament selection)
        Perform crossover to generate offspring (e.g., uniform crossover)
        Perform mutation on offspring with probability mutation_prob (e.g., randomly select features
to mutate)
        Evaluate fitness of offspring using XGBoost

        Replace worst individuals in P with offspring

    Select the best feature subset from P based on fitness
    Train XGBoost classifier on selected features using the dataset

    return Selected features, Trained XGBoost classifier
```

Comparison with existing classifiers

In addition to evaluating our proposed approach, we conducted comparative analyses with three popular classifiers: Support Vector Machine (SVM), Nearest Neighbor (KNN), and Naive Bayes (NB). These classifiers were chosen due to their widespread use in cancer data analysis and their diverse underlying principles. By benchmarking our results against existing techniques reported in the literature, we aimed to provide a comprehensive assessment of the effectiveness of our proposed strategy. Performance metrics such as accuracy, precision, recall, and F1-score were computed for each classifier to facilitate a thorough comparative analysis [30].

Evaluation metrics and statistical analysis

To quantify the performance of the classifiers and assess the significance of differences observed, we employed a range of evaluation metrics and conducted appropriate statistical analyses. These metrics included accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Additionally, we performed statistical tests such as t-tests or ANOVA to determine the statistical significance of observed differences in performance metrics between the proposed approach and existing classifiers. Such analyses were crucial for providing robust evidence of the efficacy of our proposed strategy in comparison to established techniques [31].

Experimental setup and data description

In our research, we utilized the computational capabilities of the Ubuntu 20.04.5 LTS operating system, incorporating the Windows Subsystem for Linux (WSL) for smooth

Table 2 The detailed description of the data sets

Dataset	Genes	Samples	Class
CNS	7129	60	2
Breast	24,481	97	2
Ovarian	15,154	253	2
Leukemia	7129	72	2

integration. Our coding work was facilitated by the Visual Studio Code (VS Code) integrated development environment (IDE), which seamlessly interacts with the Python programming language. To ensure the reliability of our findings, we employed the Leave-One-Out Cross-Validation (LOOCV) technique, wherein one data point is excluded from the dataset in each iteration for model training and validation. This rigorous process provides a thorough evaluation of the model's performance across the entire dataset. Our experiments were conducted on a high-performance computing system equipped with an Intel(R) Core™ i9-12900 k processor boasting a clock speed of 5.20 GHz and 64 GB of RAM, ensuring sufficient memory for handling complex algorithms and datasets. Additionally, we utilized an Nvidia RTX Quadro A5000 Graphics Processing Unit (GPU) to enhance parallelized computations and accelerate tasks benefiting from GPU acceleration. The datasets used in our study, including Breast Cancer, Leukemia, Ovarian, and CNS datasets, are described in detail in Table 2 and is also available on the link https://csse.szu.edu.cn/staff/zhuzx/Datasets.html.This meticulous approach laid the foundation for insightful analysis and well-informed conclusions.

Results and discussion

The algorithm's end condition, or NP, is expressed as the population's total number of members, and maxgen, as the algorithm's maximum number of repeats. There are 500 items in the archive list. The simulation results for the four modes listed below are examined in this paper:

(a) Without Non-Profit Selection: In this scenario, all data from each cutting set were utilized to build classes of clauses; no extra data processing, such as sample selection or selection on the basis of non-profit status, was done.
(b) Feature Selection (FS): In this scenario, each detail's data is originally taken into account as input for the algorithm using a multiplicity of multiplicity of particles. This method selects a number of features from all of the characteristics for each data set as its output. On the basis of the chosen characteristic, the data is then applied to the classes of the clauses.
(c) Sample Selection (IS): In this scenario, each fine-grained data set is taken into account as an input for the algorithm that multiplies multiplicity particles. Then, a number of samples that can enhance the performance of the classes in the training stage are chosen based on the phase majority vote. These samples are applied to the classes of the clauses once training and testing samples have been chosen.

Table 3 Efficiency of classifiers on microarray dataset

			Accuracy%			F-measure%		
Datasets	No.of features	No.of samples	SVM	KNN	NB	SVM	KNN	NB
CNS								
Non-Pre	3000	70	2.08 ± 78.84	3.12 ± 78.31	2.61 ± 80.02	2.01 ± 71.03	2.26 ± 65.24	1.60 ± 70.05
FS	0.70 ± 4.15	70	0.99 ± 90.24	0.80 ± 93.24	0.94 ± 94.15	0.70 ± 85.05	0.61 ± 90.05	0.69 ± 90.21
IS	3000	2.99 ± 36.98	0.99 ± 86.06	1.06 ± 86.04	1.10 ± 90.31	1.25 ± 78.25	0.98 ± 84.32	0.99 ± 87.25
FSIS	0.50 ± 4.45	2.21 ± 39.97	0.70 ± 92.51	0.50 ± 94.25	0.79 ± 97.24	0.58 ± 86.27	0.30 ± 87.35	0.50 ± 89.12
Leukemia								
Non-Pre	6954	82	1.98 ± 98.50	2.25 ± 92.14	1.90 ± 94.25	1.62 ± 84.21	1.60 ± 80.21	1.60 ± 81.27
FS	0.69 ± 4.32	82	0.89 ± 97.24	0.70 ± 98.26	0.60 ± 98.36	0.78 ± 86.32	0.87 ± 85.21	0.84 ± 83.97
IS	6954	1.82 ± 51.06	0.40 ± 95.25	0.44 ± 92.18	0.40 ± 96.32	0.65 ± 84.24	0.40 ± 79.25	0.62 ± 82.14
FSIS	0.49 ± 4.50	1.60 ± 59.96	0.10 ± 98.08	0.25 ± 96.31	0.19 ± 99.14	0.22 ± 89.24	0.19 ± 82.21	0.10 ± 88.21
Ovarian								
Non-Pre	15,326	262	1.89 ± 87.21	1.77 ± 89.54	1.70 ± 82.21	1.40 ± 79.25	1.53 ± 79.25	1.74 ± 72.14
FS	0.84 ± 4.50	262	0.70 ± 94.21	1.25 ± 94.25	0.95 ± 93.13	0.55 ± 84.21	1.13 ± 83.25	0.89 ± 83.21
IS	15,326	4.01 ± 17.0	1.19 ± 91.15	1.69 ± 89.32	0.98 ± 84.15	1.09 ± 79.32	1.25 ± 78.10	0.60 ± 74.21
FSIS	0.80 ± 3.6	1.34 ± 19.25	0.32 ± 97.32	0.38 ± 96.21	0.29 ± 95.21	0.25 ± 88.26	0.42 ± 86.31	0.38 ± 84.24
Breast								
Non-Pre	25,124	87	1.65 ± 86.24	2.18 ± 80.02	1.94 ± 77.25	1.19 ± 72.35	1.98 ± 66.31	2.18 ± 62.35
FS	0.70 ± 13.5	87	0.28 ± 89.32	0.37 ± 87.41	0.64 ± 85.37	0.49 ± 73.16	0.77 ± 71.24	0.94 ± 69.51
IS	25,124	0.69 ± 72.05	1.15 ± 87.20	1.48 ± 81.34	1.20 ± 81.27	1.68 ± 72.18	1.10 ± 67.41	1.60 ± 61.38
FSIS	0.85 ± 13.7	0.83 ± 79.60	0.13 ± 89.31	1.20 ± 88.63	0.17 ± 87.62	0.37 ± 75.63	0.55 ± 75.34	0.12 ± 73.62

(d) Feature Selection and Sample Selection (FSIS): In this scenario, fine-grained data are first entered into the algorithm for feature selection, and then the data is entered algorithm based on the selected features selected in the selection phase, which is proposed to select training and test samples. Following these actions, the chosen data will be applied to the chosen class with the chosen features. The results of the threefold validation approach were validated using the results of the aforementioned four modes and F-Meas.

Since intelligent optimization techniques have a random nature, they have been utilized to assess all simulation results in order to gage their robustness and boost the accuracy of the findings. Table 3 lists the outcomes of the aforementioned four options for the three basic and combination categories. The outcomes of this table show that it can be improved for all four fine-grained datasets, both of which have improved the performance of all classes in both FS and FSIS modes compared to non-PRE mode in terms of accuracy and F-Measure. The effectiveness of the illness can be cited as the cause of this. In other words, the particle congestion algorithm has been able to detect and pick the most potent effects on the diagnosis of the condition, as well as to eliminate and select the majority of the qualities of all the characteristics. The sensitivity to the optimum data selection for the training of the clauses will rise in the case of fine-grained due to the low number of samples relative to the number of features. If the classes are trained with weak data, the method will also be impacted. Table 3 shows that compared to non-PRE mode, the performance of the categories in IS mode has also increased. Table 3 further reveals that, between the two modes of FS and IS, the performance of the straps in FS mode is significantly superior than that of IS mode. For instance, the Leukemia set's lowest and highest recovery percentages for ISIL are, respectively, 98% and 8.73%. 3.22% (from the Leukemia dataset) and 13.14% (from the CNS data set) are, nevertheless, the lowest and highest recovery percentages, respectively.

In other words, the impact of the selection of training samples in the case of details is greater than the impact of the selection of training samples. This outcome perfectly supports the conclusion offered in. Additionally, according to Table 3 results, the combination class was able to significantly outperform the base classes in terms of accuracy and F-Measure indications across all four sets of settings and study modes. For instance, the Leukemia dataset's top values for health indicators and F-Measure in the FSIS mode are 99.87 and 93.12%, respectively. Additionally, these two indicators' lowest values for the Breast data set are 93.16 and 82.73%, respectively. These two indicators fall inside the above-mentioned range for other datasets. Another finding from Table 3 indicates that while the FSIS classification's performance has greatly improved, the values coming from the two health and F-Measure indices differ sufficiently enough that, for some data sets, this disparity surpasses 10%. The imbalance between fine data may be one of the causes of this. The desire for classification algorithms to focus on the majority class is frequently a result of the imbalance between the set of cutting data.

By choosing from a considerably smaller collection of information-containing qualities, the suggested approach has been able to enhance the performance of the straps in terms of accuracy and F-Meas. In CNS's data set, for instance, the accuracy and F-Measure indices increased from 83.42% and 78.91% to 99.24% and 90.07%, respectively, while the number of genes fell from 2000 genes to roughly 4.3 genes. In other words, by choosing a lower number of different genes, the suggested model has been able to improve the efficiency of the clauses. This holds true for other fine-grained collections as well. For one of the ten independent performances in the FSIS mode, the optimal portion of the partial, lead by Leukemia, CNS, and Ovarian, is shown in Figs. 2, 3 and 4. It is noted that the suggested method is successfully able to divide the levels of gene expression into two classes given these forms. Tables 4 and 5 present the base class and combination class average results for the two health indicators and the F-Measure for four small array datasets. The SVM class performs better than the KNN and NB categories for basic categories, as can be shown from these tables, when non-rere mode is used and no processing of the microdata collection is done. The SVM class is less susceptible to the numerous features, according to the increased efficiency and precision. The KNN category has performed best in terms of accuracy index and F-Measure after the SVM. Based on the results shown in these tables, it can be seen that the SVM category once more performs better than the other two categories for every six datasets when it comes to the impact of choosing the attributes and choosing the training samples on the performance of the clauses. The performance of the SVM, KNN, and NB classes in FSIS mode for the health index is 8.42, 8.29, 11.78, and 8.86%, respectively, based on the average findings shown in the above tables, compared to non-rere mode, and 12.39, 12.67, 16.67, and 13.53% (Table 6). In other words, choosing the right samples can significantly increase the effectiveness of classification algorithms while also lowering computing costs by using efficient and discrete genes. For the 10 times the algorithm in the FSIS mode, Table 7 displays the worst response, the best response, the average response, and the accuracy index criterion. The proposed method was able to achieve the best accuracy index of 100% in the two sums of colon and leukemia, according to the findings of this table. Table 7 also includes a list of the F-Measure index values.

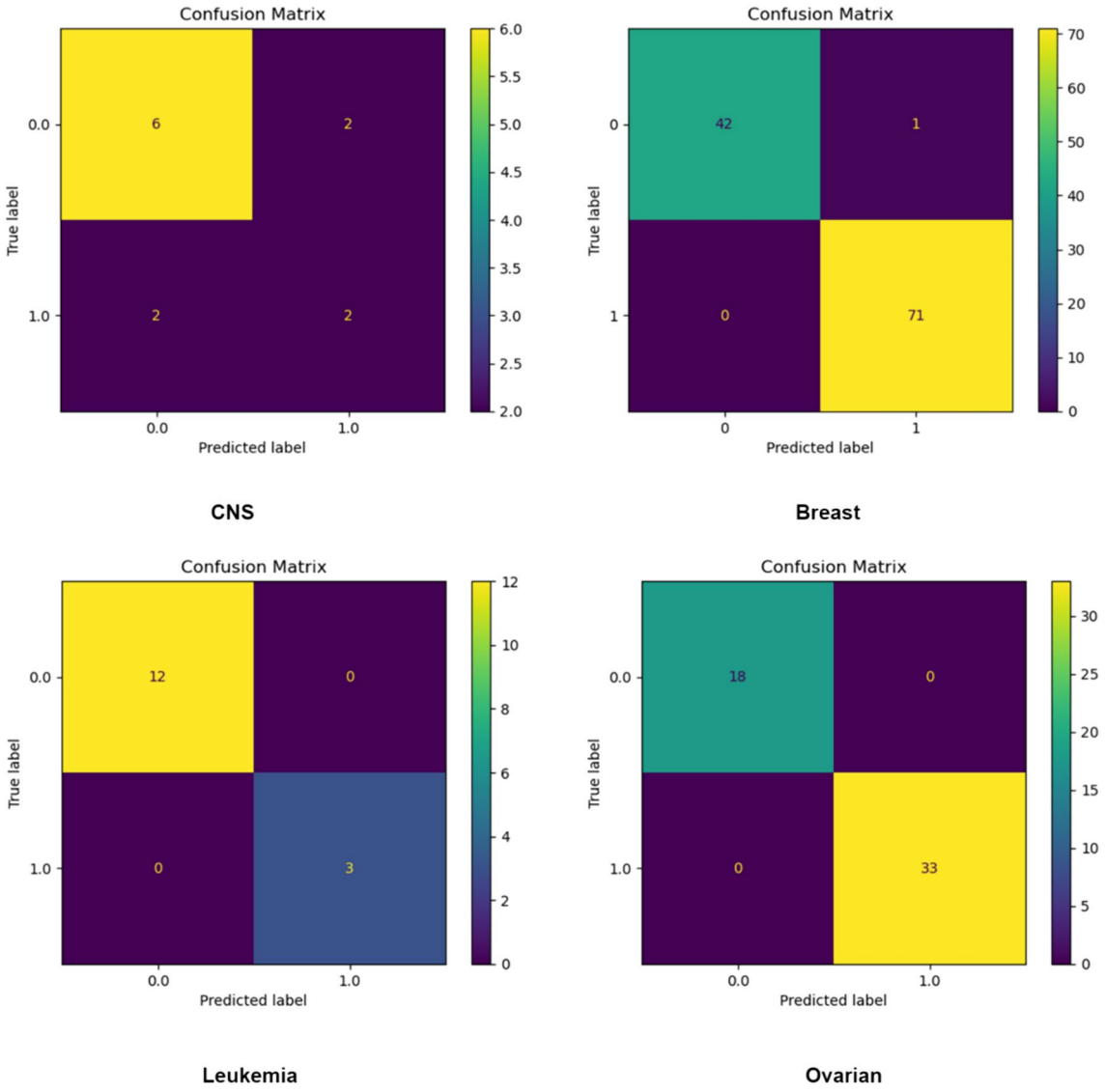

Fig. 2 Confusion Matrix presenting the classification performance

A comparison between the suggested approach and the methods described in other articles is made in order to assess the effectiveness of the proposed method.

Evaluation metrics

The performance of the proposed methodology is assessed using various evaluation metrics, including accuracy, precision, recall, F1 score, area under the receiver operating characteristic curve (AUC-ROC), and confusion matrix as shown in the Fig. 2. These metrics provide insights into the classification performance, including the model's ability to correctly classify samples from different cancer classes and its robustness to imbalanced data.

Confusion matrix

Figure 2 illustrates an essential evaluation metric for classification models, referred to as the confusion matrix. This matrix provides a comprehensive overview of model performance by presenting predictions alongside actual ground truth values in a structured grid format. Divided into four distinct quadrants—true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)—the matrix delineates correctly identified positive and negative instances, as well as misclassifications of positive and negative instances. Specifically, true positives signify accurately identified positive cases, while true negatives denote accurately identified negative cases. False positives represent instances incorrectly classified as

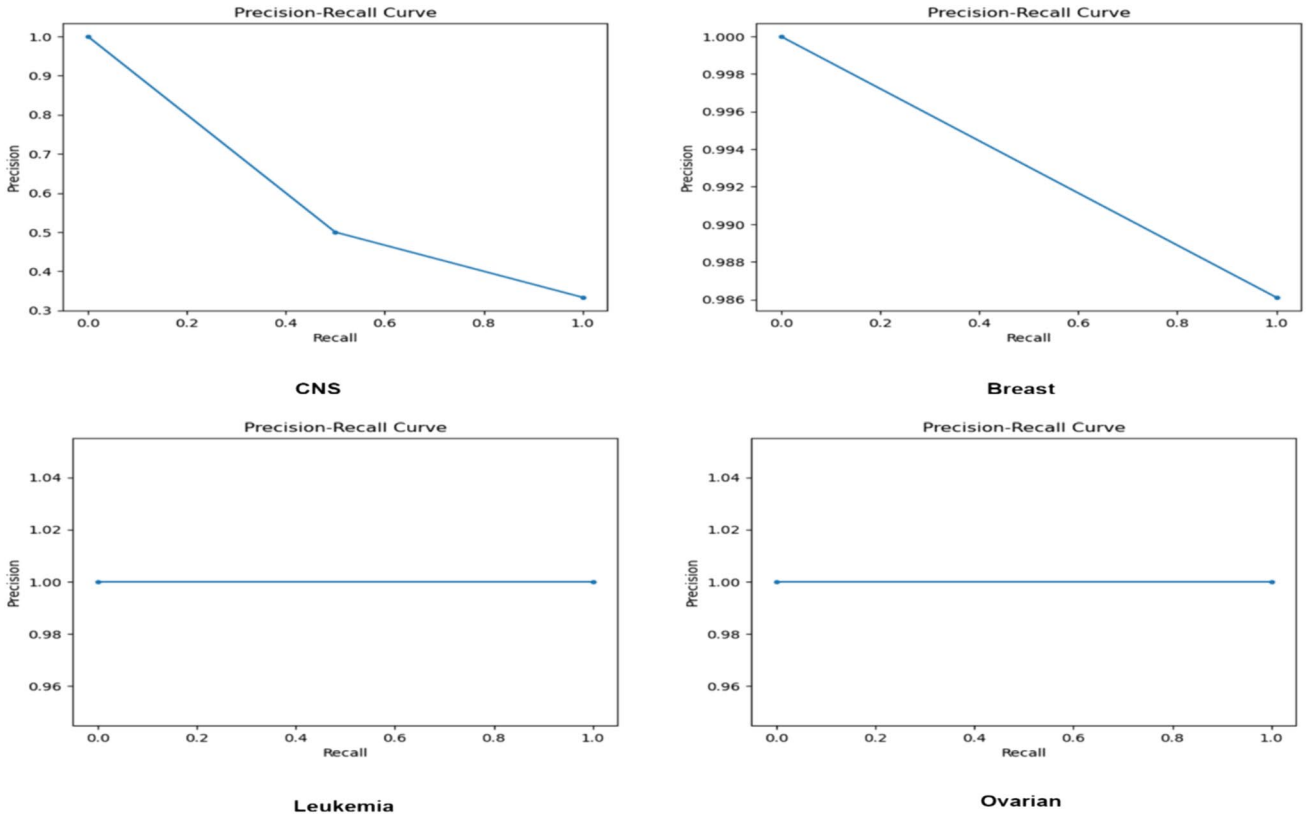

Fig. 3 Precision-Recall Curve demonstrating the precision and recall performance of the proposed method

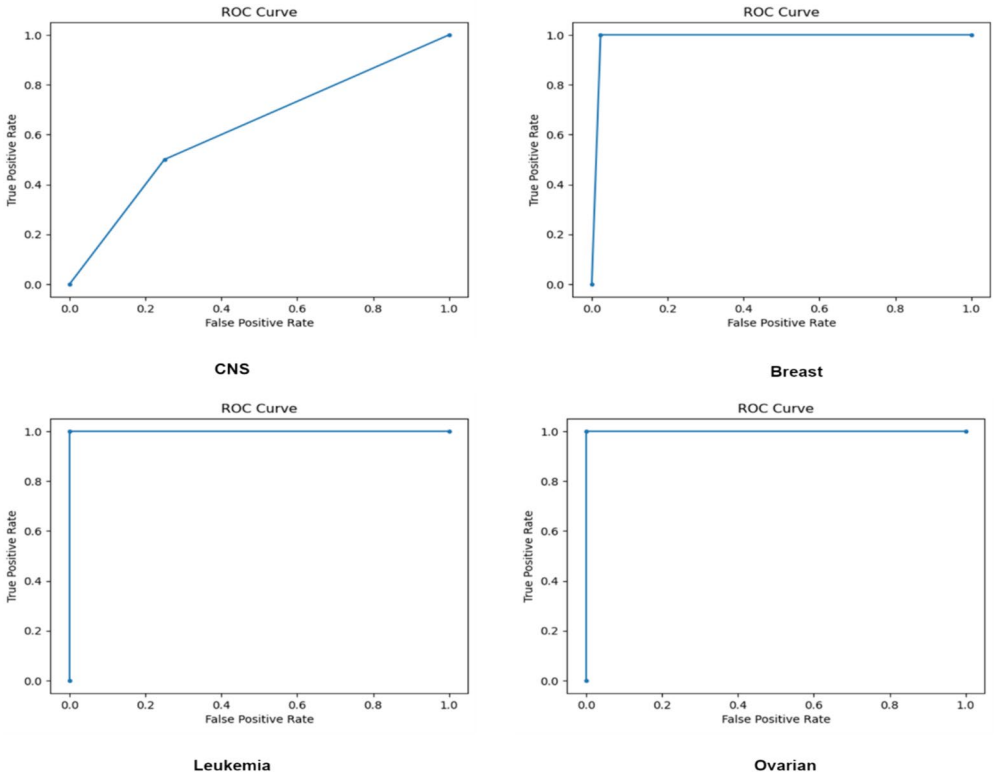

Fig. 4 ROC Curve illustrating the discriminatory power of proposed method in cancer classification

positive, while false negatives indicate instances incorrectly classified as negative.

Offering a succinct and transparent summary of the model's strengths and limitations, this matrix provides invaluable insights into its predictive performance. Through meticulous examination of these values, a range of performance metrics including accuracy, precision, recall, and F1-score can be derived. Together, these metrics offer a holistic evaluation of the model's ability to make precise classifications. The confusion matrix serves as an indispensable tool for fine-tuning models and guiding informed decisions regarding their practical deployment. Equations (a) through (d) dictate the computation of these performance metrics based on the values within the confusion matrix

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{a}$$

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{b}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{c}$$

$$F\text{−score} = 2 \times \frac{P \times \mathrm{Sn}}{P + \mathrm{Sn}} \tag{d}$$

Precesion recall curve

Additionally, the study harnesses crucial visual aids to fortify its conclusions. Illustrating a nuanced evaluation of the proposed approach's performance, Fig. 3 showcases the Precision-Recall (PR) graph. This graph provides a detailed depiction of the trade-off between precision, representing the accuracy of positive predictions, and recall, denoting the proportion of actual positives correctly predicted. By plotting precision against recall, the PR graph enables a comprehensive understanding of the model's ability to balance between making accurate positive predictions and capturing a high proportion of actual positives.

Receiver operating characteristic (ROC)

A greater Area Under the Curve (AUC) indicates enhanced model performance. The Precision-Recall (PR) graph illustrates the strength and efficacy of the proposed approach in precisely detecting positive instances, particularly crucial in situations characterized by imbalanced class distribution.

Moreover, Fig. 4 provides a depiction of the Receiver Operating Characteristic (ROC) curve, which serves as a pivotal measure for evaluating classifiers. This graph visually illustrates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity). As the ROC curve plots sensitivity against 1-specificity, it offers insights into the model's ability to discriminate between classes effectively. Notably, a higher Area Under the Curve (AUC-ROC) value signifies enhanced discriminatory capabilities of the model. The ROC curve generated by the proposed methodology underscores its remarkable discriminatory power, further corroborating its efficacy in accurately classifying cancer types. The notable AUC values observed for both Precision-Recall (PR) and ROC curves validate the reliability and efficiency of the combined approach in cancer classification. These graphical representations not only provide compelling evidence of the method's proficiency but also hold substantial promise for its implementation in real-world clinical settings.

Discussion

The proposed hybrid Random Drift Optimization (RDO) and XGBoost framework demonstrated significant improvements in cancer classification across a variety of datasets, including CNS, Leukemia, Ovarian, and Breast cancers. While our method outperformed several well-known classifiers like SVM, KNN, and Naive Bayes, it is important to contextualize these results with similar approaches from the literature.

1. *Comparison with Other Nature-Inspired Algorithms* Recent studies have explored the integration of nature-inspired optimization algorithms such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and Ant Colony Optimization (ACO) for feature selection in cancer classification. For instance, PSO has been used in combination with SVM to optimize gene selection in high-dimensional datasets. However, these methods often converge prematurely and may struggle in complex, multi-modal search spaces like those in cancer datasets. In contrast, RDO offers a more controlled exploration–exploitation balance, reducing the risk of local minima and improving feature selection precision, especially in heterogeneous cancers such as breast cancer [32, 33].
2. *XGBoost versus Other Machine Learning Classifiers* Previous works have used machine learning classifiers like Random Forest (RF) and SVM in cancer data analysis. While RF and SVM have been widely successful, they can suffer from issues such as overfitting (in the case of RF) or sensitivity to hyperparameters (as with SVM). XGBoost, with its ability to handle sparse data and regularization techniques, has been shown to be particularly effective in managing high-dimensional,

noisy datasets, as evidenced by our superior results in both accuracy and F-measure metrics [34].

3. *Hybrid Approaches* Several hybrid approaches that combine feature selection and classification have been proposed. For example, Liu et al. used a combination of GA and KNN for breast cancer classification, while Chen et al. applied a hybrid GA-SVM approach. While these approaches report competitive results, they often require complex parameter tuning and can be computationally expensive. The RDO-XGBoost framework not only simplifies the tuning process through an efficient search mechanism but also leverages the inherent strengths of XGBoost, such as scalability and regularization, making it a more practical solution for large datasets.
4. *Limitations and Future Directions* Despite the promising results, challenges remain. For breast cancer, the complexity and heterogeneity of the disease seem to limit the framework's performance compared to other cancers. This suggests that future work should explore subtype-specific feature selection or the incorporation of additional biological data (e.g., epigenetic markers, proteomic profiles) to improve performance in such heterogeneous datasets. Another potential direction is to enhance the RDO component with multi-objective optimization techniques to simultaneously minimize redundancy and maximize relevance in feature selection.

Conclusion

In conclusion, our study introduces a novel approach for feature selection in cancer data analysis, integrating Random Drift Optimization (RDO) with XGBoost, a potent gradient boosting algorithm. Through extensive experiments on real-world cancer datasets, we have demonstrated the efficacy of our method in enhancing classification accuracy while uncovering crucial insights into the biological mechanisms underpinning cancer progression. By harnessing the complementary strengths of RDO for efficient exploration of the solution space and XGBoost for robust predictive modeling, our approach outperforms existing techniques and standalone classifiers like Support Vector Machine (SVM), Nearest Neighbor (KNN), and Naive Bayes (NB). The resulting feature subsets not only improve classification efficiency but also offer valuable molecular insights, potentially identifying biomarkers and therapeutic targets. Thus, our study underscores the promise of integrating RDO with XGBoost as a powerful strategy for advancing feature selection in cancer research, with implications for personalized medicine and targeted therapy. Continued research efforts are essential to validate these findings on larger and more diverse datasets, paving the way for further advancements in cancer diagnosis and treatment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00262-024-03843-x.

Author contributions In this study, Abrar Yaqoob made significant contributions by preparing materials, meticulously collecting data, and conducting data analysis. Navneet Kumar Verma provided valuable expertise through comprehensive literature reviews, critical insights, and substantial contributions to the manuscript's intellectual content. Rabia Musheer Aziz and asif shah played a pivotal role in conceptualizing the project, expertly designing experiments to address key research questions, and supervising the research process to maintain integrity and coherence. This collaborative effort, with each team member leveraging their specific expertise, led to a comprehensive and impactful research project.

Funding No funding is available for this project.

Data availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Machap L, Abdullah A, Shah ZA (2020) Functional analysis of cancer gene subtype from co-clustering and classification. Indones J Electr Eng Comput Sci 18(1):343–350. https://doi.org/10.11591/ijeecs.v18.i1.pp343-350
2. Yaqoob A, Verma NK, Aziz RM (2024) Improving breast cancer classification with mRMR + SS0 + WSVM: a hybrid approach. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-20146-6
3. Lu H, Chen J, Yan K, Jin Q, Xue Y, Gao Z (2017) A hybrid feature selection algorithm for gene expression data classification. Neurocomputing 256(2017):56–62. https://doi.org/10.1016/j.neucom.2016.07.080
4. Agrawal RK, Kaur B, Sharma S (2020) Quantum based Whale Optimization Algorithm for wrapper feature selection. Appl Soft Comput J 89:106092. https://doi.org/10.1016/j.asoc.2020.106092
5. Houssein EH, Hosney ME, Mohamed WM, Ali AA, Younis EMG (2023) Fuzzy-based hunger games search algorithm for global optimization and feature selection using medical data. Neural Comput Appl 35(7):5251–5275. https://doi.org/10.1007/s00521-022-07916-9

6. Yaqoob A, Kumar N, Rabia V, Aziz M (2024) Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm. J Med Syst. https://doi.org/10.1007/s10916-023-02031-1
7. Yaqoob A, Verma NK, Aziz RM, Saxena A (2024) Enhancing feature selection through metaheuristic hybrid cuckoo search and harris hawks optimization for cancer classification. Metaheuristics for Machine Learning: Algorithms and Applications, pp 95–134
8. Yaqoob A, Aziz RM, Verma NK, Lalwani P, Makrariya A (2023) A review on nature-inspired algorithms for cancer disease prediction and classification
9. Yaqoob A, Musheer Aziz R, Verma NK (2023) Applications and techniques of machine learning in cancer classification: a systematic review. Human-Centric Intell Syst. https://doi.org/10.1007/s44230-023-00041-3
10. Sun J, Wu X, Palade V, Fang W, Shi Y (2013) Random drift particle swarm optimization. https://arxiv.org/abs/1306.2863
11. Sun J, Wu X, Palade V, Fang W, Shi Y (2015) Random drift particle swarm optimization algorithm: convergence analysis and parameter selection. Mach Learn 101(1–3):345–376. https://doi.org/10.1007/s10994-015-5522-z
12. Yaqoob A, Bhat MA, Khan Z (2023) Dimensionality reduction techniques and their applications in cancer classification: a comprehensive review. Int J Genet Modif Recomb 1(2):34–45
13. Sree Devi KD, Karthikeyan P, Moorthy U, Deeba K, Maheshwari V, Allayear SM (2022) Tumor detection on microarray data using grey wolf optimization with gain information. Math Probl Eng. https://doi.org/10.1155/2022/4092404
14. Yaqoob A (2024) Combining the mRMR technique with the Northern Goshawk Algorithm (NGHA) to choose genes for cancer classification. Int J Inf Technol:1–12
15. El-Mageed AAA, Elkhouli AE, Abohany AA, Gafar M (2024) Gene selection via improved nuclear reaction optimization algorithm for cancer classification in high-dimensional data, vol 11. Springer. https://doi.org/10.1186/s40537-024-00902-z
16. Bilal A et al (2024) Improved Support Vector Machine based on CNN-SVD for vision-threatening diabetic retinopathy detection and classification. PLoS ONE 19(1):e0295951. https://doi.org/10.1371/journal.pone.0295951
17. Yaqoob A, Verma NK, Aziz RM (2024) Metaheuristic algorithms and their applications in different fields: a comprehensive review. Metaheuristics for Machine Learning: Algorithms and Applications, pp 1–35
18. Dabba A, Tari A, Meftali S (2024) A novel grey wolf optimization algorithm based on geometric transformations for gene selection and cancer classification. J Supercomput 80(4):4808–4840. https://doi.org/10.1007/s11227-023-05643-z
19. Nssibi M, Manita G, Chhabra A, Mirjalili S, Korbaa O (2024) Gene selection for high dimensional biological datasets using hybrid island binary artificial bee colony with chaos game optimization, vol 57. Springer. https://doi.org/10.1007/s10462-023-10675-1
20. Benghazouani S, Nouh S, Zakrani A, Haloum I, Jebbar M (2024) Enhancing feature selection with a novel hybrid approach incorporating genetic algorithms and swarm intelligence techniques. Int J Electr Comput Eng 14(1):944–959. https://doi.org/10.11591/ijece.v14i1.pp944-959
21. Meyer PE, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. IEEE J Sel Top Signal Process 2(3):261–274. https://doi.org/10.1109/JSTSP.2008.923858
22. Kundu R, Chattopadhyay S, Cuevas E, Sarkar R (2022) AltWOA : Altruistic Whale Optimization Algorithm for feature selection on microarray datasets. Comput Biol Med 144:105349. https://doi.org/10.1016/j.compbiomed.2022.105349
23. Debata PP, Mohapatra P (2021) Identification of significant biomarkers from high-dimensional cancerous data employing a modified multi-objective meta-heuristic algorithm. J King Saud Univ Comput Inf Sci. https://doi.org/10.1016/j.jksuci.2020.12.014
24. Trik M, Mohammad A, Gil N, Ghasemi F (2022) Research article a hybrid selection strategy based on traffic analysis for improving performance in networks on chip, vol 2022
25. Wang J, Wu L, Kong J, Li Y, Zhang B (2013) Maximum weight and minimum redundancy : a novel framework for feature subset selection. Pattern Recognit 46(6):1616–1627. https://doi.org/10.1016/j.patcog.2012.11.025
26. Li C, Sun J, Palade V, Li LW (2021) Diversity collaboratively guided random drift particle swarm optimization. Int J Mach Learn Cybern 12(9):2617–2638. https://doi.org/10.1007/s13042-021-01345-1
27. Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13–17-Augu, pp 785–794. https://doi.org/10.1145/2939672.2939785
28. Uzir N, Raman S, Banerjee S, Nishant Uzir RS, Sunil R (2016) Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets Experimenting XGBoost Algorithm for Prediction and Classifi cation of Different Datasets. Int J Control Theory Appl 9. https://www.researchgate.net/publication/318132203
29. Ghatasheh N, Altaharwa I, Aldebei K (2022) Modified genetic algorithm for feature selection and hyper parameter optimization: case of XGBoost in spam prediction. IEEE Access 10(August):84365–84383. https://doi.org/10.1109/ACCESS.2022.3196905
30. Çakir M, Yilmaz M, Oral MA, Kazanci HÖ, Oral O (2023) Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. J King Saud Univ Sci. https://doi.org/10.1016/j.jksus.2023.102754
31. Shaw RG, Mitchell-Olds T (1993) ANOVA for unbalanced data: an overview. Ecology 74(6):1638–1645. https://doi.org/10.2307/1939922
32. Ahmed AA, Ali MAS, Selim M (2019) 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
33. Trojovská E, Dehghani M (2022) A new human-based metahuristic optimization method based on mimicking cooking training. Sci Rep 12(1):1–24. https://doi.org/10.1038/s41598-022-19313-2
34. Zhang T et al (2023) Application of nonlinear models combined with conventional laboratory indicators for the diagnosis and differential diagnosis of ovarian cancer. J Clin Med. https://doi.org/10.3390/jcm12030844