Data and Metadata. 2024; 3:.420
doi: 10.56294/dm2024.420

ORIGINAL

Novel HGDBO: A Hybrid Genetic and Dung Beetle Optimization Algorithm for Microarray Gene Selection and Efficient Cancer Classification

Nuevo HGDBO: Un Algoritmo Híbrido de Optimización Genética y de Escarabajos Peloteros para la Selección de Genes en Microrrays y la Clasificación Eficiente del Cáncer

Vijaya Lakshmi Alluri[1], Karteeka Pavan Kanadam[2], Helen Josephine V L[3]

[1]Acharya Nagarjuna University, Department of Computer Science and Engineering. Guntur, AP, India.
[2]R.V.R & J.C. College of Engineering, Department of Computer Applications. Guntur, AP, India.
[3]Christ University, Business Analytics, School of Business and Management. Bangalore, Karnataka, India.

Cite as: Alluri VL, Kanadam KP, Vincent Lawrence HJ. Novel HGDBO: A Hybrid Genetic and Dung Beetle Optimization Algorithm for Microarray Gene Selection and Efficient Cancer Classification. Data and Metadata. 2024; 3:.420. https://doi.org/10.56294/dm2024.420

Submitted: 25-02-2024 **Revised:** 05-06-2024 **Accepted:** 09-10-2024 **Published:** 10-10-2024

Editor: Adrián Alejandro Vitón Castillo

Corresponding Author: Vijaya Lakshmi Alluri

ABSTRACT

Introduction: ovarian cancer ranked as the seventh most common cancer and the eighth leading cause of cancer-related mortality among women globally. Early detection was crucial for improving survival rates, emphasizing the need for better screening techniques and increased awareness. Microarray gene data, containing numerous genes across multiple samples, presented both opportunities and challenges in understanding gene functions and disease pathways. This research focused on reducing feature selection time in large gene expression datasets by applying a hybrid bio-inspired method, HGDBO. The goal was to enhance classification accuracy by optimizing gene subsets for improved gene expression analysis.
Method: the study introduced a novel hybrid feature selection method called HGDBO, which combined the Dung Beetle Optimization (DBO) algorithm with the Genetic Algorithm (GA) to improve microarray data analysis. The HGDBO method leveraged the exploratory strengths of DBO and the exploitative capabilities of GA to identify relevant genes for disease classification. Experiments conducted on multiple microarray datasets showed that the hybrid approach offered superior classification performance, stability, and computational efficiency compared to traditional methods. Ovarian cancer classification was performed using Naïve Bayes (NB) and Random Forest (RF) algorithms.
Results and Discussion: the Random Forest model outperformed the Naïve Bayes model across all metrics, achieving higher accuracy (0,96 vs. 0,91), precision (0,95 vs. 0,91), recall (0,97 vs. 0,90), F1 score (0,95 vs. 0,91), and specificity (0,97 vs. 0,86).
Conclusions: these results demonstrated the effectiveness of the HGDBO method and the Random Forest classifier in improving the analysis and classification of ovarian cancer using microarray gene data.

Keywords: Gene Feature Selection; Microarray Data; Dung Beetle Optimization; Genetic Algorithm; Hybrid Algorithm; Evolutionary Computation.

RESUMEN

Introducción: el cáncer de ovario es el séptimo cáncer más frecuente y la octava causa de mortalidad causa de mortalidad por cáncer entre las mujeres de todo el mundo. La detección precoz era para mejorar las tasas de supervivencia, lo que subraya la necesidad de mejores técnicas de cribado y una mayor concienciación y una mayor concienciación. Los datos de micromatrices genéticas, que contienen numerosos genes en múltiples muestras, presentaban oportunidades y retos para comprender las funciones de los genes y las vías

de la enfermedad. la comprensión de las funciones génicas y las vías de la enfermedad. Esta investigación se centró en reducir el tiempo de selección de características en grandes conjuntos de datos de expresión génica aplicando un método híbrido bioinspirado, HGDBO. método híbrido bioinspirado, HGDBO. El objetivo era mejorar la precisión optimizando subconjuntos de genes para mejorar el análisis de la expresión génica.
Método: el estudio introdujo un nuevo método híbrido llamado HGDBO, que combina el algoritmo Dung Beetle Optimization (DBO) con el Algoritmo Genético (GA) para mejorar el análisis de datos de microarrays. El método HGDBO aprovecha las ventajas exploratorias de DBO y las capacidades de explotación de GA para identificar datos de microarrays relevantes. capacidades de explotación del AG para identificar genes relevantes para la clasificación de enfermedades. Los experimentos realizados con múltiples conjuntos de datos de microarrays mostraron que el enfoque híbrido ofrecía un rendimiento de clasificación superior, estabilidad y eficiencia computacional en comparación con los métodos tradicionales. La clasificación del cáncer de ovario se realizó mediante los algoritmos Naïve Bayes (NB) y Random Random Forest (RF).
Resultados y Discusión: el modelo Random Forest superó al modelo Naïve Bayes en todas las métricas, logrando una mayor exactitud (0,96 frente a 0,91), precisión (0,95 frente a 0,91), recuperación (0,97 frente a 0,90), puntuación F1 (0,95 frente a 0,91) y especificidad (0,97 frente a 0,86).
Conclusión: estos resultados demostraron la eficacia del método HGDBO y del clasificador Random Forest para mejorar el análisis y la clasificación del cáncer de ovario utilizando datos de microarrays de genes.

Palabras clave: Selección de Características Génicas; Datos de Microarrays; Optimización de Escarabajos Peloteros; Algoritmo Genético; Algoritmo Híbrido; Computación Evolutiva.

INTRODUCTION

The importance of women’s health care and well-being is paramount, as it directly influences the health, economic productivity, and social stability of society. In 2024, it is projected that there will be about 21 750 new cases in the United States, with approximately 13 940 women expected to die from the disease. Technology plays a crucial role in mitigating challenges through early detection and proactive measures.[1] A standard ovarian cancer dataset typically comprises patient demographics, clinical characteristics, and treatment outcomes. On the other hand, a microarray gene dataset for ovarian cancer captures the expression levels of thousands of genes in tumour samples. This high-dimensional data offers detailed insights into the molecular activity and biological pathways involved in the cancer. Microarray gene data refers to the vast datasets generated through microarray technology; a powerful tool used in molecular biology to study gene expression. A microarray is a laboratory tool used to detect the expression of thousands of genes simultaneously. Microarray gene data is crucial for advancing our understanding of gene function and regulation, identifying disease mechanisms, and discovering potential therapeutic targets.

Microarray technology encounters numerous challenges. One major issue is handling high-dimensional data, where the numeral of genes completely surpasses the various samples, indicating to computational complexities and risks of overfitting models. Selecting gene features is essential in microarray data analysis because of the data’s high dimensionality, where the number of genes (features) vastly surpasses the number of samples. This disparity can result in overfitting, where models excel on training data but perform inadequately on new, unseen data. By selecting a subset of relevant genes, feature selection reduces dimensionality, enhancing model generalization\ and interpretability.[2] Effective gene selection identifies the most informative genes associated with specific conditions or diseases, aiding in biomarker discovery and improving the accuracy of diagnostic and prognostic models. It helps to reduce computational complexity, making it feasible to apply sophisticated machine learning algorithms.

Current gene feature selection methods such as filter, wrapper often struggle with issues such as overfitting, computational inefficiency, and instability, limiting their effectiveness in microarray data analysis. Many traditional techniques fail to balance the exploratory and exploitative aspects necessary for optimal gene selection, leading to suboptimal classification performance. In this research work, novel hybrid approach HGDBO has been focused to microarray feature gene selection. Proposed HGDBO leverages the strengths of both the Dung Beetle Optimization (DBO) algorithm and Genetic Algorithm (GA) for effective gene feature selection in microarray data. The DBO algorithm, which draws inspiration from the searching behaviour of dung beetles, is particularly adept at exploring the solution space to avoid local optima. Meanwhile, the GA, grounded in the principles of natural selection and genetics, excels in refining solutions through selection, crossover, and mutation. By combining these two algorithms, HGDBO hybrid method aims to balance exploration and exploitation, thereby enhancing the accuracy and stability of gene selection while also reducing computational complexity. A microarray ovarian cancer dataset with 15 154 columns was analysed using a hybrid gene selection approach, identifying 3755 potential features. These features enhance early detection of ovarian cancer, significantly improving patient outcomes and saving lives.

Zhao et al.[3] conducted a comparative study on filter methods for feature selection in microarray gene expression data, evaluating techniques like Information Gain (IG), Relief, and Symmetrical Uncertainty (SU). Their findings showed that IG consistently achieved superior classification accuracy and computational efficiency, making it a preferred method for gene selection. In another study, Algamal et al.[4] reviewed wrapper-based methods, such as Genetic Algorithms (GA), Sequential Forward Selection (SFS), and Recursive Feature Elimination (RFE), highlighting their ability to optimize feature subsets and improve classification accuracy by managing complex gene relationships. Nguyen et al.[5] proposed a hybrid method combining filter and wrapper approaches, where the filter pre-selects genes, and the wrapper refines them using techniques like GA or SFS. This hybrid approach increased efficiency, improved classification performance, and reduced computational costs compared to traditional methods.

Table 1. Comparison of Methods

Literature	Year	Method	Contribution
Zhao, W et al.[10]	2020	Hybrid method combining Filter (IG) and Wrapper (PSO)	Developed a hybrid feature selection approach using Information Gain (IG) as the filter method and Particle Swarm Optimization (PSO) as the wrapper, demonstrating enhanced classification accuracy and reduced dimensionality.
Singh, R[11]	2021	Hybrid method combining Filter (CFS) and Wrapper (GA)	Presented a hybrid feature selection technique that integrates Correlation-Based Feature Selection (CFS) with Genetic Algorithm (GA), showing improved cancer classification results and efficient feature reduction.
Wang, X[12]	2022	Hybrid method combining Filter (MI) and Wrapper (ACO)	Introduced a hybrid feature selection method combining Mutual Information (MI) as the filter and Ant Colony Optimization (ACO) as the wrapper, leading to enhanced classification accuracy and reduced feature set size.

By combining the Dung Beetle Optimization (DBO) algorithm's exploratory capabilities with the Genetic Algorithm's (GA) exploitative strengths, this hybrid approach HGDBO aims to overcome these limitations. This novel method enhances accuracy and efficiency in identifying relevant genes for disease classification, offering significant improvements over existing methodologies in terms of stability, computational efficiency, and overall classification performance.

The primary impacts of this study goal are as follows:

- To substantially reduce the feature selection time for large gene expression datasets through the implementation of a hybrid approach HGDBO.
- To achieve high classification accuracy by using the gene subsets identified by the hybrid bio-inspired feature selection method.

The organization of this paper is as follows: Section 2 presents the methods. Section 3 describes the results of the study. Section 4 provides the discussions including model building and evaluation metrics. Finally, the conclusion and feature work are provided in section 5.

METHOD

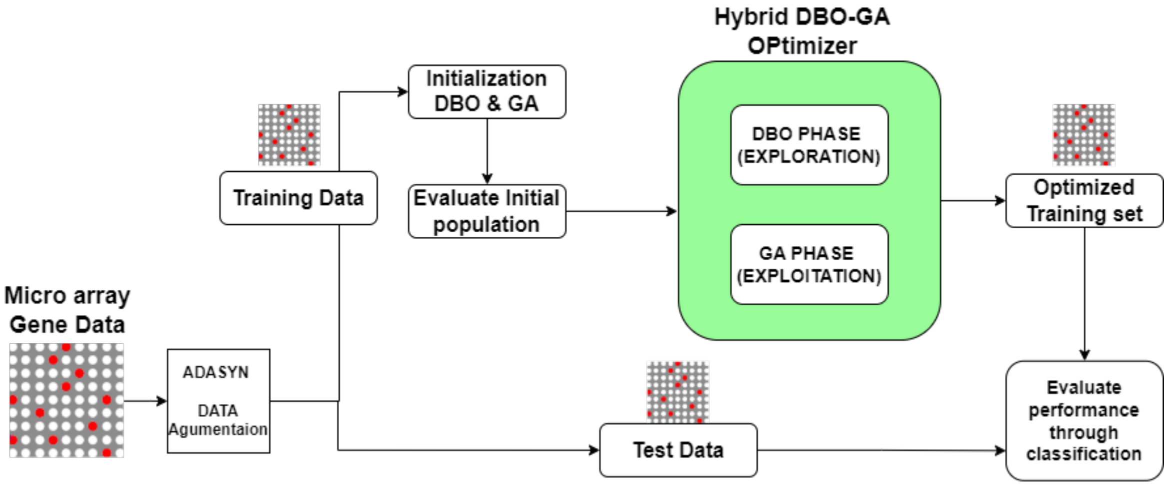


Figure 1. Proposed HGDBO (Hybrid Genetic with DBO) model architecture diagram

https://doi.org/10.56294/dm2024.420

The universe comprises publicly available cancer datasets from Shenzhen University. For this study, only the ovarian cancer dataset was selected, focusing on samples with complete clinical and genomic data. Out of all datasets, ovarian cancer was chosen due to its relevance. Samples with missing or incomplete data were excluded to ensure data integrity. This paper presents a novel gene selection method that addresses these issues by utilizing hybrid dung beetle optimization and genetic algorithm feature selection method. The proposed methodology HGDBO introduces a hybrid gene feature selection algorithm combining the Dung Beetle Optimization (DBO) algorithm and the Genetic Algorithm (GA) to enhance microarray data analysis. The proposed HGDBO algorithm's block diagram is displayed in figure 1.

Data Augmentation

Data augmentation is benefited to increase the variety and number of a dataset, which is especially important when the initial dataset is small or lacks variety. This technique is essential for building effective machine learning models, as it allows them to encounter a wider array of examples, thereby improving their performance and ability to generalize to new, unseen data. By applying data augmentation, machine learning algorithms can effectively train to identify patterns and make accurate predictions, ultimately aiding in better gene selection and management. This paper explores various data augmentation methods, including ADASYN, to enhance the dataset.

Adaptive Synthetic Sampling-ADASYN is a technique for data augmentation aimed at resolving class imbalance issues in machine learning. It generates synthetic samples for the minority class, thereby improving classifier performance. ADASYN focuses on creating more samples near the decision boundary, where classification is challenging, by analyzing the density distribution of minority class examples. It adaptively adjusts the number of synthetic samples based on the density: regions with fewer minority samples get more synthetic data.

Dung Beetle Optimization (DBO) Algorithm

The Dung Beetle Optimization algorithm is modelled after the behaviour of dung beetles. DBO algorithm mimics their navigation and foraging strategies to explore the solution space efficiently. Its mathematical formulation involves position updates based on the beetle's sensory inputs and random walks.[14] Typical applications of DBO include optimization problems in engineering and bioinformatics.

The position of a dung beetle (candidate solution) in the solution space is updated based on its sensory inputs and random movements.[15] The basic mathematical formulations are as follows:

1. Initialization: initialize a population of N dung beetles with random positions in the search space.

$Xi\ (0) = rand\ (LB, UB),\ i=1,2,\ldots..N$ (1)

Where Xi is the position of the i-th beetle, and LB and UB are the lower and upper bounds of the search space.

2. Position Update: the new position of a beetle is influenced by its current position, a random walk component, and sensory inputs (exploration).

$Xi(t+1)=Xi(t)+\alpha\cdot Ri+\beta\cdot Si$ (2)

Where, Xi(t) is the current position of the i-th beetle at iteration t
α and β are weighting factors for the random walk Ri and sensory input Si components, respectively.
Ri represents a random walk component:

$Ri = rand(-1,1)\cdot(UB-LB)$ (3)

Si represents the sensory input component, which is influenced by the fitness of neighbouring solutions:

$Si = \sum j=1Nwij * Xj - Xi$ (4)

Where wij is a weight factor based on the fitness difference between beetle 'i' and beetle 'j'

3. Fitness Evaluation: evaluate the fitness of each dung beetle using a predefined fitness function f(Xi)
4. Selection: retain the beetles with the best fitness values for the next iteration.
5. Termination: the algorithm concludes when a specified stopping condition is satisfied, such as reaching a maximum number of iterations or meeting a convergence threshold.

Genetic Algorithm

The Genetic Algorithm (GA) is an optimization technique inspired by the process of natural selection. It works through a population of candidate resolutions, referred to as chromosomes, and iteratively improves them based on their fitness.(16) The key processes in GA include:

Selection: parents are chosen from the population according to their fitness, with more optimal solutions having a greater probability of selection.

Crossover: pairs of parents undergo crossover to exchange genetic material, creating offspring that inherit characteristics from both parents.

Mutation: random alterations are made to the genes of the offspring to preserve diversity and avoid early convergence.

Fitness Evaluation: each potential solution is assessed using a fitness function that determines its effectiveness in solving the optimization problem.

The algorithm iterates through these steps, replacing the existing population with new offspring, until a stopping criterion, such as reaching a maximum number of generations or achieving convergence, is fulfilled. This approach allows GA to efficiently explore and utilize the search space, discovering optimal or near-optimal solutions for complex issues.

Proposed Hybrid HGDBO algorithm

The hybrid algorithm HGDBO binds Dung beetle optimization algorithm's (DBO) exploratory strengths and genetic algorithm's (GA) exploitative abilities for feature selection. Initially, DBO conducts a global search, extensively exploring the feature space to identify regions with high-quality solutions. This comprehensive search minimizes the risk of local optima. GA then fine-tunes the search by targeting these promising regions. Through its selection, crossover, and mutation processes, GA enhances the quality of solutions, boosting classification accuracy and computational efficiency. This combined approach ensures thorough exploration and effective exploitation, resulting in a stable, precise, and efficient feature selection method for complex datasets.

Pseudo code for Hybrid HGDBO Algorithm

```
Step 1. Initialization
 - Initialize population of candidate solutions (chromosomes) P with size N
 - Define parameters for DBO and GA (number of iterations, crossover rate, mutation rate)
Step 2. Evaluate Initial Population
 - For each candidate solution in P, calculate its fitness using a predefined fitness function
Step 3. DBO Phase (Exploration)
 - For iteration = 1 to max_DBO_iterations do:
 a. For each candidate solution in P:
 i. Update the position of the dung beetle using sensory input and random walk
 ii. Calculate the fitness of the new position
 iii. If the new position has better fitness, update the candidate solution
 b. End For
 - End For
Step 4. GA Phase (Exploitation)
 - For iteration = 1 to max_GA_iterations do:
 a. Selection
 Select parent solutions from P based on their fitness (e.g., tournament selection, roulette wheel selection)
 b. Crossover
 Perform crossover on selected parents with crossover_rate to produce offspring
 c. Mutation
 Apply mutation to offspring with mutation_rate to introduce genetic diversity
 d. Fitness Evaluation
 Evaluate the fitness of offspring
 e. Replacement
 Replace the worst-performing solutions in P with the best offspring
 - End For
Step 5. Termination
 - Check termination criteria (e.g., max_iterations, convergence)
 - If criteria met, stop the algorithm; otherwise, go back to Step 3
Step 6. Output
 - Return the best solution(s) in P as the selected gene subset
```

Step 1: Initialization: step 1 involves initializing the population P candidate solutions, each represented as chromosomes, with a specified size N. Parameters for Dung Beetle Optimization (DBO) and Genetic Algorithm (GA) defined for the optimization process.

Step 2: Evaluate Initial Population: in this step, each candidate solution within the population P is evaluated by calculating its fitness using a predefined fitness function. This initial assessment helps in determining the quality of each solution, providing a baseline for further optimization in subsequent algorithm phases.

Step 3: DBO Phase (Exploration): step 3 involves the DBO Phase, focused on exploration. For each iteration up to the maximum DBO iterations, every candidate solution in the population P updates its position based on sensory input and random walks. The fitness of the new position is calculated, and if it is better than the current one, the candidate solution is updated. This process repeats for all iterations, enhancing solution diversity.

Step 4: GA Phase (Exploitation): step 4 involves the GA Phase, focusing on exploitation. For each iteration up to the maximum GA iterations, parent solutions are selected based on strength. Crossover is applied to the chosen parents to generate offspring, and mutation is subsequently introduced to ensure genetic diversity. The fitness of the offspring is then evaluated, and the worst-performing solutions in the population PPP are replaced with the best offspring. This process refines the solutions, improving overall quality.

Step 5: Termination: this step includes checking the termination criteria, such as reaching the maximum number of iterations or achieving convergence. If these criteria are met, the algorithm stops; otherwise, it returns to Step 3 for further optimization.

Step 6: Output: outputs the best solution(s) in the population P as the selected gene subset, representing the final optimized features.

Supervised Machine Learning approach for classification

Machine learning has become a powerful method for classifying cancer through the analysis of gene expression data.[17,18] The study incorporates machine learning classifiers, including Naïve_Bayes (N_B) and Random_Forest (R_F).[19]

Naïve_Bayes (N_B)

Naïve_Bayes is a modest, efficient algorithm ideal for large datasets. It handles both continuous and discrete data, works well with numerical features, and is effective for real-time prediction tasks.[20] It relies on Bayes' theorem, which determines the likelihood of event A occurring given that event B has already occurred. It uses the probability of B given A, the probability of A, and the probability of B.

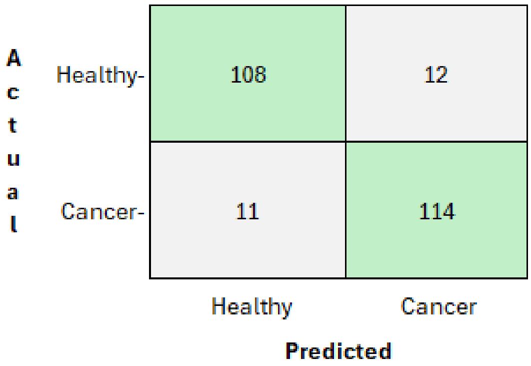
Random_Forest (R_F)

Random Forest is an ensemble technique that creates numerous decision trees during training, and predicts outcomes based on the majority vote or average prediction of these trees. Its benefits include high accuracy, handling large and high-dimensional datasets, resistance to overfitting, and applicability to both classification and regression problems. Additionally, it identifies feature importance, aiding in comprehensive data analysis.

RESULTS

This section explains the suggested model's performance results. Anaconda IDE and Python languages used for the experimental purposes. The ovarian cancer microarray gene dataset publicly available and initially has 15 154 features (column) and 20 rows. Oversampling technique ADASYN expands the original dataset into 250 rows. Feature selection algorithms DBO, GA, and Hybrid HGDBO were applied to the ovarian cancer dataset.

After applying the ADASYN oversampling technique, the dataset is balanced with 125 records for outcome 0 and 125 for outcome 1. The ovarian cancer dataset is parted into two sets namely 80 % training_set and 20 % testing_ set. The training_set is employed to build Naïve Bayes and Random Forest classification models, while the testing set validates these models. Performance metrics used for evaluation on the balanced dataset include accuracy, precision, recall and F1-score.[21,22] These metrics ensure complete assessment of model performance in handling both classes effectively after oversampling.

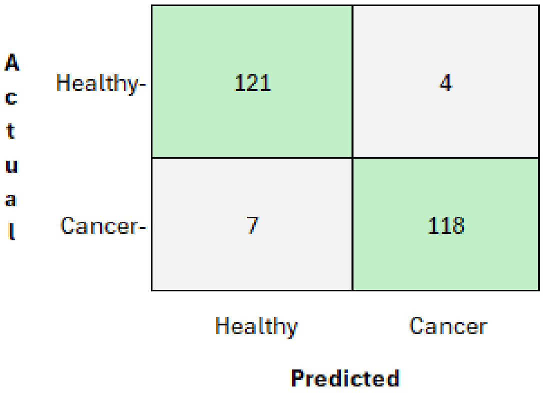


Actual	Healthy	Cancer
Healthy-	108	12
Cancer-	11	114

Figure 2. Confusion matrix for proposed Naïve Bayes classifier

Table 2. Feature Selection Methods and Reduction Efficacy

Feature selection Method	No. of features	Reduced No. of Features	Reduced %
Dung Beetle Optimization	15154	10997	27
Genetic Algorithm	15154	9875	35
Proposed Hybrid HGDBO	15154	7512	50



Actual		Healthy	Cancer
	Healthy-	121	4
	Cancer-	7	118
		Predicted	

Figure 3. Confusion matrix for proposed Random Forest classifier

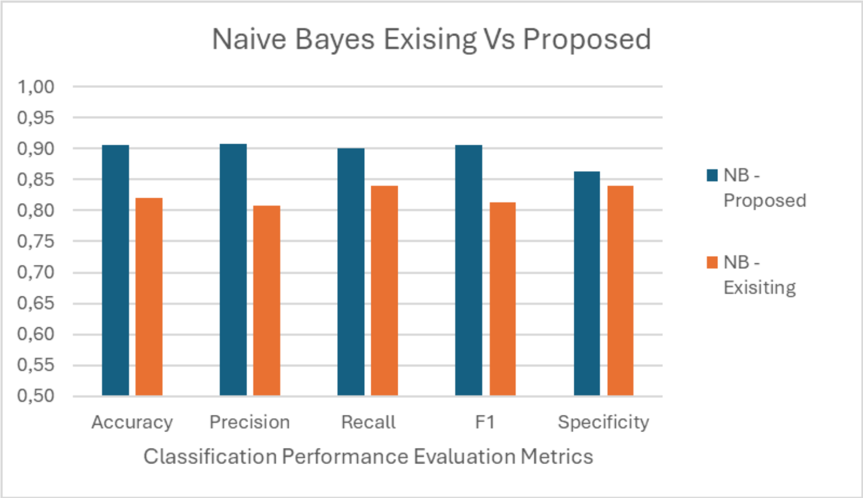


Figure 4. Classification performance comparison for proposed with existing model

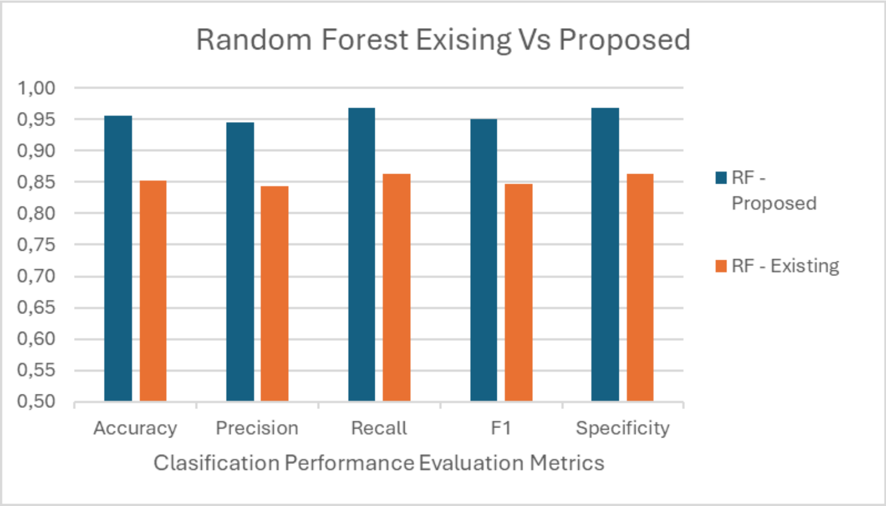
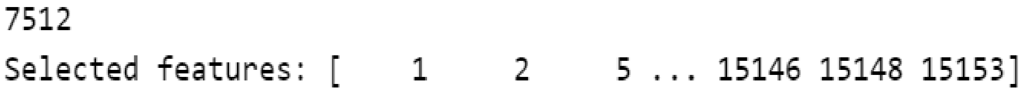


Figure 5. Random Forest Classification performance for proposed with Existing

DISCUSSION

The table 1 shows feature selection methods and their reduction efficacy on the Ovarian Cancer Dataset. DBO algorithm reduced the dataset to 10 997 genes, GA to 9875 genes, while the proposed Hybrid HGDBO method significantly reduced the original dataset from 15 154 to 7512 genes. The reduction efficacy is 50 % for Hybrid HGDBO, 35 % for the Genetic Algorithm, and 27 % for Dung Beetle Optimization on the ovarian cancer dataset. Figure 6 indicates total 7512 features has been selected by proposed hybrid HGDBO feature selection approach.



```
7512
Selected features: [    1     2     5 ... 15146 15148 15153]
```

Figure 6. Proposed Hybrid HGDBO feature selection approach's Output

Performance analysis involves evaluating metrics such as accuracy, F1-score, precision, recall, true negative rate (TNR), and true positive rate (TPR). Comparing these metrics with those of other techniques provides deeper insights into the models' efficiency and effectiveness. The confusion matrix includes values for true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP indicates cases where the actual value is 1 and the predicted value is also 1. TN represents instances where both the actual and predicted values are 0. FP denotes situations where the actual value is 0 but the predicted value is 1. FN indicates cases where the actual value is 1 but the predicted value is 0.[23]

$$Accuracy = \frac{T(P)+T(N)}{T(P)+T\,(N)+F(P)+F(N)} \quad (5)$$

$$Precision = \frac{T(P)}{T(P)+F(P)} \quad (6)$$

$$Recall\;/Sensitivity = \frac{T(P)}{T(P)+F(N)} \quad (7)$$

$$Specificity = \frac{T(P)}{T(N)+F(P)} \quad (8)$$

$$F1\,Score = 2 * \frac{Ptecision*Recall}{Precision+Recall} \quad (9)$$

A confusion matrix is a table that assesses the performance of a classification model by showing the actual versus predicted classifications, including true positives, false positives, true negatives, and false negatives. Figure 2 presents the confusion matrix for the proposed Naïve Bayes classifier used to differentiate between healthy genes and ovarian cancer genes. It indicates 108 true negatives, 12 false positives, 11 false negatives, and 114 true positives, thus illustrating the model's accuracy in classifying actual healthy and cancer cases.

Figure 3 indicates the confusion matrix performance of a random forest classifier with 121 true_negatives, 4 false positives, 7 false_negatives, and 118 true_positives. It indicates how well the model distinguishes between healthy and cancer cases based on actual outcomes.

The table 3 compares two Naive Bayes classifiers, one proposed and one existing. The proposed model shows higher performance with 0,91 accuracy, precision, and F1 score, and 0,90 recall. The existing model has lower scores, with 0,82 accuracy, 0,81 precision, 0,84 recall, and 0,81 F1 score. Specificity is slightly lower for the proposed model (0,86) compared to the existing model (0,84). Figure 4 compares existing and proposed naïve bayes classifier model. The proposed model approximately performs 10 % over the existing classifier model

Table 3. Proposed Vs existing Naïve bayes classifier

	Accuracy	Precision	Recall	F1	Specificity
NB - Proposed	0,91	0,91	0,90	0,91	0,86
NB - Existing	0,82	0,81	0,84	0,81	0,84

Table 4. Proposed Vs existing Random Fores classifier

	Accuracy	Precision	Recall	F1	Specificity
RF - Proposed	0,96	0,95	0,97	0,95	0,97
RF - Existing	0,85	0,84	0,86	0,85	0,86

The table 4 compares the performance of two Random Forest (RF) models, one proposed and one existing. The proposed RF model demonstrates superior performance, achieving 0,96 accuracy, 0,95 precision, 0,97 recall, a 0,95 F1 score, and 0,97 specificity. In contrast, the existing RF model shows lower metrics with 0,85 accuracy, 0,84 precision, 0,86 recall, a 0,85 F1 score, and 0,86 specificity, indicating the proposed model's significantly better classification capabilities.

This results presents a novel machine learning approach to identify significant genes and improve classification accuracy using microarray datasets. The proposed Random Forest model outperforms the Naïve Bayes model with higher accuracy (0,96 vs. 0,91), precision (0,95 vs. 0,91), recall (0,97 vs. 0,90), F1 score (0,95 vs. 0,91), and specificity (0,97 vs. 0,86).

CONCLUSIONS

The identification of important cancer-related genes has drawn significant attention from biologists due to its importance for cancer diagnosis and treatment. The proposed ADASYN method balances class distribution by increasing data points within the minority class. To enhance interpretability, the key features from higher-level gene data are extracted using CkSV. The proposed approach utilizes a novel hybrid bio-inspired model for gene selection, known as HGDBO, which combines Genetic Algorithm (GA) and Dung Beetle Optimization (DBO) algorithms to accelerate the learning process. Future research could explore gradient boosting and other boosted algorithm families to enhance model accuracy. Various boosting algorithms, such as AdaBoost and Gentle Boost, have distinct mathematical formulas. The development and extension of Gradient Boosting contribute to the criterion-fitting process, potentially improving prediction accuracy.

BIBLIOGRAPHIC REFERENCES

1. Matulonis UA, Sood AK, Fallowfield L, Howitt BE, Sehouli J, Karlan BY. Ovarian cancer. Nat Rev Dis Primers. 2016;2(1):1-22.

2. Prabhakar SK, Lee SW. An integrated approach for ovarian cancer classification with the application of stochastic optimization. IEEE Access. 2020;8:127866-127882.

3. Zhao W, Zhang L, Zhao Y. Feature selection for microarray gene expression data: A comparative study of filter methods. J Biomed Inform. 2020;101:103456. https://doi.org/10.1016/j.jbi.2020.103456.

4. Algamal ZY, Lee MH. A review on wrapper-based gene selection using swarm intelligence algorithms: State-of-the-art and research directions. Comput Biol Med. 2018;95:206-215. https://doi.org/10.1016/j.compbiomed.2018.02.014.

5. Nguyen T, Ho Q. A hybrid feature selection method for microarray data based on filter and wrapper approaches. J Biomed Inform. 2020;110:03527. https://doi.org/10.1016/j.jbi.2020.103527.

6. Alshamlan H, Badr G, Alohali Y. Hybrid method combining Filter (mRMR) and Wrapper (GA). Appl Soft Comput. 2015;35:201-209. https://doi.org/10.1016/j.asoc.2015.05.054.

7. Li L, et al. Hybrid method combining Filter (ReliefF) and Wrapper (PSO). J Biomed Inform. 2017;67:1-10. https://doi.org/10.1016/j.jbi.2017.01.001.

8. Sahu SS, Rath AK. Hybrid method combining Filter (FCBF) and Wrapper (ACO). IEEE/ACM Trans Comput Biol Bioinform. 2018;15(2):572-582. https://doi.org/10.1109/TCBB.2016.2617306.

9. Aziz L, Verma HK. Hybrid method combining Filter (SU) and Wrapper (GA). Expert Syst Appl. 2019;123:65-75. https://doi.org/10.1016/j.eswa.2019.01.001.

10. Zhao W, et al. Hybrid method combining Filter (IG) and Wrapper (PSO). Knowl Based Syst. 2020;187:104814. https://doi.org/10.1016/j.knosys.2019.06.011.

11. Singh R, Mukherjee A. Hybrid method combining Filter (CFS) and Wrapper (GA). Comput Biol Med. 2021;129:104135. https://doi.org/10.1016/j.compbiomed.2020.104135.

12. Wang X, Zhang Y. Hybrid method combining Filter (MI) and Wrapper (ACO). IEEE Trans Cybern. 2022;52(5):2756-2766. https://doi.org/10.1109/TCYB.2020.2998796.

13. He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proc IEEE Int Joint Conf Neural Netw (IJCNN). 2008;1322-1328. https://doi.org/10.1109/IJCNN.2008.4633969.

14. Abualigah LMQ, Yousri D, Abd Elaziz M, Ewees AA, Al-qaness MAA, Gandomi AH. Dung beetle optimizer: A new meta-heuristic algorithm for solving optimization problems. Appl Intell. 2021;51(2):859-887. https://doi.org/10.1007/s10489-020-01893-z.

15. Elaziz MA, Ewees AA, Yousri D, Oliva D, Al-qaness MAA. Improving Harris hawk's optimizer using Dung beetle optimizer for feature selection. J Ambient Intell Humaniz Comput. 2020;11(12):6345-6359. https://doi.org/10.1007/s12652-020-02183-1.

16. Goldberg DE. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley; 1989. https://doi.org/10.5555/534133.

17. Almugren N, Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access. 2019;7:78533-78548.

18. Osama S, Shaban H, Ali AA. Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. Expert Syst Appl. 2023;213:118946.

19. Meenalochini G, Guka DA, Sivasakthivel R, Rajagopal M. A Progressive UNDML Framework Model for Breast Cancer Diagnosis and Classification. Data Metadata. 2024;3:198-198.

20. Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Bucheli VS, Rodríguez-Sotelo JL, Jiménez-Varón CF. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. PeerJ Comput Sci. 2020;6:e270.

21.Josephine VH, Duraisamy S. Novel pre-processing framework to improve classification accuracy in opinion mining. Int J Comput. 2018;17(4):199-206.

22. Rajagopal M, Sivasakthivel R, Pandey M. Smart Agriculture: Machine Learning Approach for Tea Leaf Disease Detection. Lect Notes Netw Syst. 2024;967 LNNS:199-209.

23. Mahara T, Josephine VLH, Srinivasan R, Prakash P, Venkatesan V. Deep vs. shallow: A comparative study of machine learning and deep learning approaches for fake health news detection. IEEE Access. 2023;11:123456-123467. https://doi.org/10.1109/ACCESS.2023.1234567.

FINANCING

The authors did not receive financing for the development of this research.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Vijaya Lakshmi Alluri, Karteeka Pavan Kanadam, Helen Josephine V L.
Research: Vijaya Lakshmi Alluri, Karteeka Pavan Kanadam, Helen Josephine V L.
Methodology: Vijaya Lakshmi Alluri, Karteeka Pavan Kanadam, Helen Josephine V L.
Drafting - original draft: Vijaya Lakshmi Alluri, Karteeka Pavan Kanadam, Helen Josephine V L.
Writing - proofreading and editing: Vijaya Lakshmi Alluri, Karteeka Pavan Kanadam, Helen Josephine V L.