Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4092404, 10 pages
https://doi.org/10.1155/2022/4092404

Research Article

Tumor Detection on Microarray Data Using Grey Wolf Optimization with Gain Information

K. Dhana Sree Devi,[1] P. Karthikeyan,[2] Usha Moorthy ,[3] K. Deeba,[4] V. Maheshwari,[2] and Shaikh Muhammad Allayear[5]

[1]*Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India*
[2]*School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India*
[3]*School of Computer Science and Engineering, REVA University, Bangalore, India*
[4]*Department of Computer Applications, Auxilium College (Autonomous)-Vellore, India*
[5]*Department of Multimedia and Creative Technology, Daffodil International University, Daffodil Smart, Khagan, Ashulia, Dhaka, Bangladesh*

Correspondence should be addressed to Usha Moorthy; ushmitha@gmail.com

Received 17 March 2022; Revised 9 May 2022; Accepted 13 May 2022; Published 15 June 2022

Academic Editor: Shimin Wang

Microarray data are becoming a more essential source of gene expression data for interpretation and analysis. To improve the detection accuracy of tumors, the researchers try to use the lowest feasible collection of the most gene expression studies, and relevant gene expression patterns are found. The purpose of this article is to use a data mining strategy and an optimized feature selection method focused on a limited dense tree forest classifier to evaluate and forecast colon cancer data. More specifically, merging the "gain information" and "Grey wolf optimization" was incorporated as a feature selection approach into the random forest classifier, to improve the prediction model's accuracy. Our suggested technique can decrease the load of high-dimensional data, and it allows quicker computations. In this research, we provided a comparison of the analysis model with feature selection accuracy over model analysis without feature selection accuracy. The extensive experimental findings have shown that the suggested method with selecting features is beneficial, outperforming the good classification performances.

1. Introduction

Most colon cancers have become a considerable public health issue, and most of these cancers have expanded speedily worldwide. GLOBOCAN Database 2018 examined new 1,849,518 colorectal cancer (CRC) instances and 880,792 CRC-related deaths. The CRC is 0.33% the main cause of most cancer-associated deaths in the USA, 2019. The latest study by Wong Martin [1] suggests that about 25 percent of CRC instances contain a genetic propensity. At initial stage, generic cancers classification technique is build totally on DNA microarray gene expression monitor [2]. They additionally recommended similar microarray data would possibly give a classification technique for most cancers. The microarray technology, which largely constructs the expression of genes, has widely utilized in prognosis additionally assessment of colon-related malignancies. Timely identification of cancer is crucial in accurate detection and therapy. Microarray-based data contain hundreds of gene information, and subsample sizes are often smaller. They had a difficult time identifying the most significant genes using microarray data because not every gene had adequate check-out facts and many of them are redundant. The two current strategies, feature transformation and selection of acquiring feature genes for most cancer classification, were built totally on gene expression data [3].

Feature transformation is a technique for creating a unique set of modern datasets from existing ones to achieve feature reductions, even if the author requires strong

discriminating power, typically to not retain the biological data in the initial sequence. The loss of data is reflected in data transformation interpretability, and it is not possible to spot a list of cancer-related target genes unlike methods for feature transformation and selection that do not generate a new set of features. Ghazavi and Liao take off nonredundant and relevant functions and keep their exceptional classification performance accuracy [4]. Feature selection now no longer contains replacing original features hence lowering the dimensionality information trouble to create a trust model of the dataset used. Even though, the techniques with a characteristic selection have received a similar interest aspect. The selection techniques may be separated into 3 major categories: filters, wrappers, and embedded techniques [5]. The filter method is a way of selecting features that is dependent on any future machine learning techniques and is based on a few statistical feature performances. They were technically quick and completely dependent on dataset characteristics. One of the most significant dangers is the overlook function connections. Wrapping-based methods were primarily focused on finding algorithms, which iteratively evaluate data against a set of machine learning rules to get the best subset of features. For datasets with numerous properties, algorithms are not only slower than filters but also computationally costly. Because they interact with the classifiers for the selection of features, embedded methods are minimally computationally expensive and faster than the other types of feature selection algorithms. Different forms of random forests, decision trees, and artificial neural networks are popular embedded methods. Gain information (GI) and Grey Wolf optimization were presented as strategies for choosing variables (GWO). An classifier is then developed for analyzing colon cancer. GWO is the most efficient swarm intelligence-based metaheuristic method. GWO is used in several optimization methods such as clustering applications, design and tuning controllers, power dispatch problems, robotics and path planning, scheduling problems, wireless sensor networks, and medical and biomedical applications [6]. While researching the usage of GWO in various engineering problems, we found that GWO has an ability to handle huge variable numbers and to escape local solutions while solving a large-scale problem. Since GWO has the ability to handle a large amount of variable with better solution, we intend to apply GWO for microarray data as it is a more essential source of gene expression data for interpretation and analysis.

For different environments, scholars have proposed abundant selection algorithms. There are several optimization algorithms inspired from nature or purely mathematical-driven methods. Some of them include monarch butterfly optimization (MBO), which optimizes the search strategy by decreasing the local optima, which in turn decreases the premature convergence and reduces the number the local maxima. This behavior is inspired from the monarch butterflies [7]. The slime mould algorithm is another kind of optimization algorithm inspired from the slime mould mode in nature. This algorithm mimics its behavior from the morphological changes of slime mould physarum polycephalum in completing its lifecycle. The entire lifecycle is modeled into a mathematical one, and the authors found it can be useful for optimization problems [8]. There is an algorithm inspired from the orientation of moths called moth swarm algorithm (MSA) [9]. MSA is modeled by capturing the movements of moths in moonlight. This can be used to create a learning based on association yielding an immediate memory, which utilizes levy-based mutation ethics to increase the cross-population environment and movement in spiral. Hunger Games Search (HGS) [10] optimization is based on inheriting the characteristics from hunger behavior of animals. The hunger-driven behavior of wild animals is inherited and modeled to a mathematical model and applied to solve a range of optimization problems. The RUNge Kutta optimizer [11] is different from bio-inspired algorithms and is meant to solve a variety of optimization problems in future. RUN utilizes the slope variation logic that is computed by the Runge–Kutta method as a searching mechanism. The drawback in this method is that this optimization works on large datasets only. The colony predation algorithm (CPA) is an efficient optimization method that focuses on utilizing a mathematical method inspired from the hunting group of animals such as prey encircling, prey dispersing, targeting hunters, animal strategy, and adjusting strategy [12]. Weighted mean of vectors is the most commonly used optimization algorithm in the literature prior to the discovery of bio-inspired algorithms. It works on the simple logic of assigning weights to the elements present in the vectors using a normalized function and computing the solution for the question set [13]. However, scheduling problem with no-wait constraints widely exists in the real-life process of steel production, computer systems, food processing, chemical industry, pharmaceutical industry, concrete products, etc. Many experts and scholars have studied optimization problems with zero constraints. For instance, to overcome no-wait scheduling idea with m-machine, a hybrid algorithm is taken based on the genetic algorithm and simulated annealing. To minimize the makespan of Flow Shop scheduling idea, several variants of descending search and Tabu search algorithm were proposed, and a strategy based on a dynamic Tabu list was also proposed, which enhanced the algorithm's ability to jump out of local optimum to a certain extent. A hybrid optimization algorithm based on variable neighborhood descent and PSO was used to solve Flow Shop scheduling with two optimization goals. To minimize the weighted sum of maximum completion time and total completion time, the literature proposed a TOB (trade-off balancing) algorithm based on machine idle times. For Job Shop scheduling optimization in which each job has its optimization strategy, the literature proposed a hybrid genetic algorithm, in which the genetic operation is treated as a subproblem and transformed into asymmetric travelling salesman problem. In the abovementioned commonly used production scheduling algorithms, no consideration is given to the great product structure differences, processing parameter differences, and the need for further deep processing after assembly of jobs in the real-life manufacturing process of nonstandard products.

In fact, to quickly respond to the ever-changing market and alleviate the pressure of nonstandard products in research and trial production, some enterprises have established dedicated production workshops to improve production efficiency of less-than-truckload, personalized products and nonstandard products. However, some order-oriented SMEs organize production according to orders. During the production process, there are a large number of nonstandard products that demand scribing, hand lapping, scraping, and precision templates. Big differences exist in product structure and component parameters and jobs demand further deep processing after assembly, so parts cannot be predicted and prepared in advance, and production must be advanced according to BOM (bill of material) [14]. The problem of requiring further deep processing after jobs assembly is often referred to as integrated scheduling problem (ISP). For ISP, literatures discussed a hybrid optimization method of bottleneck shifting and genetic algorithm. The literature pointed out that common no-wait scheduling algorithms can only deal with the case where the number of no-wait child nodes is 1. However, in ISP, there are abundant cases in which further deep processing is required after jobs assembly; that is, the number of no-wait child nodes can be greater than 1 in ISP. Therefore, ISP with no-wait constraints is more complicated [15]. Some of the recent heuristic algorithms include monarch butterfly optimization (MBO), slime mould algorithm (SMA), moth search algorithm (MSA), hunger games search (HGS), Runge–Kutta method (RUN), and Harris hawks optimization (HHO). There are several optimization algorithms in the literature using various nature-inspired techniques for optimization but still GWO is not utilized in the field of microarray detection and so we proposed in this work. While researching the above published works, it is clear that optimization is still a major issue.

Our Contributions include the following:

(1) A data mining strategy and a feature selection method based on a threshold optimized forest classifier are proposed to optimize the feature selection

(2) Ensemble of "gain information" and "Grey wolf optimization" was incorporated as a feature selection approach into the random forest classifier, to improve the prediction model's accuracy

(3) Provided a comparison of model analysis with feature selection over model analysis without feature selection

2. Literature Survey

A detailed review on gene selection proves that feature selection is a significant thing for data mining procedure. Xian et al. proposed a particle deletion using a strategy with a computed fitness value through clustering, and the corresponding particles are generated using the importance of feature and this ensemble of the two algorithms is used to compute the crossover of both qualities of particles. Salem et al. published a study that used gene expression profiles to classify human cancers [16]. In this feature selection technique, from the initial microarray, the information gain (IG) was used to identify genes. In addition, the genetic algorithm (GA) was used for reducing the features utilizing the IG. For cancer categorization using genetic programming (GP) (or diagnosis), data mining algorithms are used. Seven cancer gene expression datasets were utilized to verify the technique. The author established accuracy of 85.48 percent for colon cancer and 88.87 percent for breast cancer (central nervous system *cancer*), 96.05 percent (leukaemia72), 72.3 percent (lung *cancer* in Ontario), and 100 percent for all cancers (lung cancer, Michigan)), 93.7 percent (DLBCL, Harvard), and 100 percent (lung cancer, Michigan) (prostate).

Bennett et al. provided an ensemble feature technique that combines the support vector machine feature selection reduction (SVM-RFE) technique with Bayes error filter (BBF) for accuracy improvement as a hybrid genetic algorithm strategy [17]. The attributes were sorted using SVM-RFE, and the superfluous sorted attributes were removed using BBF. After that, the dataset was classified using the SVM method. The best classification accuracy for the Leukaemia73 dataset was 96.1 percent.

On 10 datasets, the authors of Gunavathi et al. investigated the effects of GA combined with k-nearest-neighbors (KNN) and are analyzed using SVM classifiers [18]. Three filters were used to decrease the number of characteristics identified by the GA. The SVM ensembles with KNN algorithms are utilized to predict the data at the end. Accuracy of the SVM is nearly identical to that of the KNN classifier on most datasets; the only exception was the Leukaemia72 dataset, for which the authors used fivefold cross-validation.

Canedo et. al. developed a new method of bifurcation of classifiers. The researchers utilized a voting mechanism to classify the samples [19]. They used tenfold cross-validation to apply their methods to ten microarray datasets, and classification accuracy rates are well improved for several datasets, and moreover, the authors tried to focus on the level of possible parameter optimization in gene selection using the GA. This is considered the best performing area in optimization methodology.

Using a gene expression dataset, in a combination of four classifiers for cancer classification, a GA-based optimization was utilized for gene selection. As classifiers, naive Bayes feature selector, SVM hyperplane optimization, CUBIST, and decision tree forests classifiers were employed [20]. Lymphoma, Lung, CNS, Colon, Leukaemia38, and Leukaemia73 were the six datasets studied, with top prediction rates of 97.1 percent, 99.03 percent, 81.3 percent, 87.8%, 100 percent, and 97.06 percent, respectively.

Salem and Hanaa highlighted the results of a study that used gene expression to classify early breast cancer [21]. Their method uses an IG approach to identify important genes from both the initial microarray data and then gives it as an input to a GA-based arithmetic material to enable the optimization at a better speed. The specificity of classification was 100 percent.

Bouazza and Sara Haddou presented findings from a study that used SVM and KNN classifiers to classify cancer [22]. This study analyzed the data using various feature selection techniques (such as Fisher, Relief, SNR, and T-Statistics) on multiple gene regulation account sets of data (Prostate, Colon, and Leukaemia) for both KNN and SVM classifiers on 3 datasets of profiled gene expression. Merging the SNR feature selection with SVM yielded the best results. The best classification accuracy rates were 95 percent for the colon and breast data and 100 percent for leukaemia-based data utilizing SNR feature selection with the KNN classifier.

Wang et al. [23] proposed an ensemble feature selection method based on the sampling method. The feature vector is sampled using the threshold value set by a sample selector, and the sampled features are ranked using the bootstrapping method. Based on the ranking, the top-ranked features are aggregated using a data aggregation function. High-dimensional microarray data are utilized to test the proposed ensemble model, and the results show the efficiency of the proposed algorithm is better compared to benchmark optimization algorithms. A detailed review of feature selection methods for microarray data for cancer disease classification is detailed by Esra'a Alhenawi [24].

There are several classification mechanisms proposed by various researchers using different approaches. One such novel approach is carried out by Konstantina et al. [25]. The authors utilized the transcriptomics dataset for modeling the gene expression data. Deep learning-based time series modeling is used for time series analysis to provide inference on network regulatory. The authors achieved an accuracy of 98%. Several experiments were carried out to show the efficiency of the proposed algorithm in gene expression time series data.

A concept of altruism is proposed by Rohit Kundu et al. [26]. Altruism concept is embedded in the WOA to manipulate the candidate solutions to reach the local optima over the fitness value of the iterations. This method of alteration increased the fitness value of the WOA. The efficiency of altruistic WOA is tested on the microarray dataset for optimized feature selection. The authors used eight microarray datasets from cancer data repository to show the supremacy of the proposed algorithm in these datasets.

There are three phases to the proposed technique. After the data are provided in the feature learning process, the IG filter identifies one of the essential features. The GWO algorithm then minimizes the volume of features that have been selected. The system's final stage involves using the SVM classifier to generate cancer classification. The following is a summary of the technique shown in Figure 1.

3. Proposed System

The ensemble model of gain information and grey wolf optimization is the proposed system, and we utilized in this research. Since the ensemble model provides better performance than benchmark optimization algorithms, we used this ensemble technique for performance improvement. The

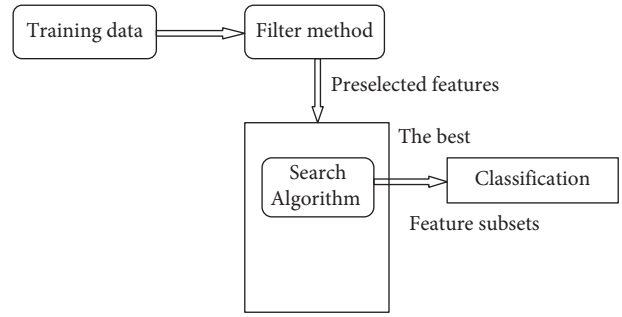

Figure 1: Information gain process.

methodology of this analysis is depicted in Figure 2. Firstly, the data collection plays a very crucial in the process. The data from the first stage were then moved to the second stage, where they were classified. We employed two techniques in the third stage to choose MDI and MDG features for data training and testing. A comparative architecture analysis was carried out.

3.1. Data Samples. During the collection process of the concerned data, data on colon cancer gene expression were collected from Alon et al. [27]. There are 63 samples (testing) and approximately 2000 genes (attributes) among colon cancer patients in the databases. There are 42 cancer cell samples among them, as well as 22 normal biopsies. The data from colon tumor samples are shown in Table 1.

3.2. Classification Performance Evaluation in the Absence of Feature Selection. With all of the attributes in this technique, an RF classification with tenfold cross-connection was utilized to evaluate the model's results.

3.3. Evaluation of Feature Selection with Classification. MDG and MDA were utilized as a feature selection approach to predict the more relevant to the meaningful feature. We then used selected features to build a robust model and followed the same process as defined in the previous phase.

3.4. Performance Measures. We compare the effectiveness of the system without feature selection with the model with feature selection in this procedure. To evaluate the classification's consistency, we used recalls, accuracy, correctness, and F1-score measures. The neural network, which is used to analyze the quality of classifiers, produces predictable results.

Tables 1 and 2 show the confusion matrix's interpretation and the algorithm for calculating performance indicators, evenly. Recalls, also known as sensitivity, are the proportion to properly predict positive instances of every discovery in the label class. The precision metric reveals which of the positive results are correct. This shows that the ratio of accurately predicted classes to total classes is the accuracy of a classifier. The F1-score is calculated using a weighted average of accuracy and recall. When there is an unbalanced class percentage, the F1-

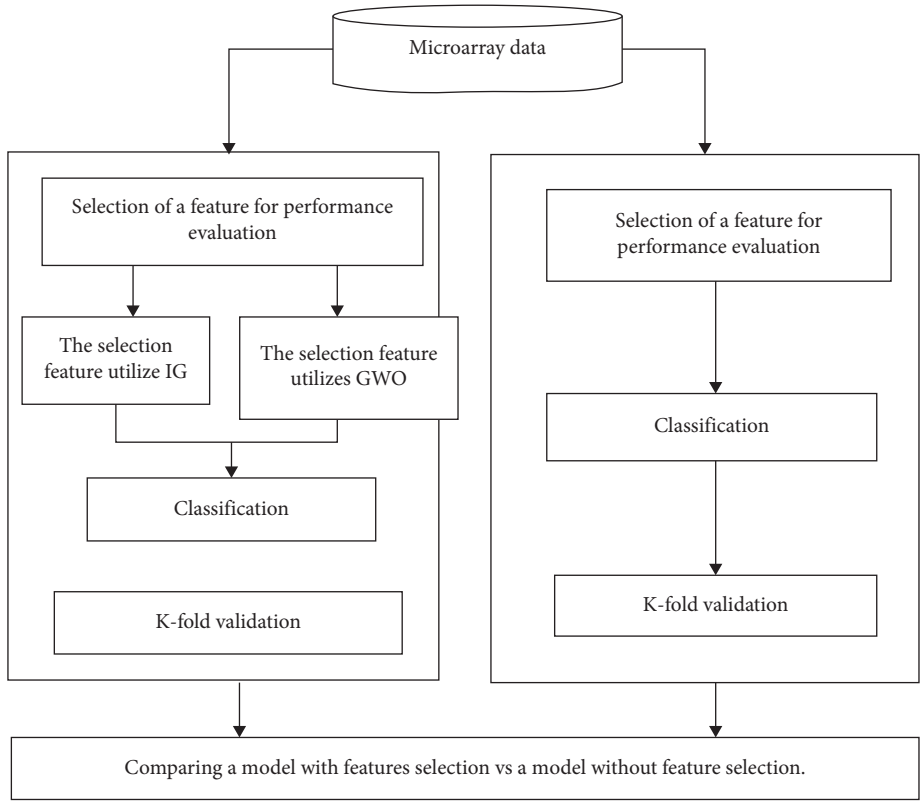

Figure 2: Framework of the proposed model.

Table 1: Confusion matrix.

Predicted	Actual	
	Positive	Negative
Predicted positive	True positives (TPs)	False negatives (FNs)
Predicted negative	False positive (FPs)	True negatives (TNs)

Table 2: Performance measure representation.

Performance metrics	Formula
Recall	$TP/TP + FN$
Precision	$TP/TP + FP$
F1-score	2 ∗ Recall ∗ Precision/Recall + Precision
Accuracy (%)	$TP + TN/TP + FP + FN + TN$

score is frequently the most important than the precision since it accounts for either false positives or false negatives:

$$VI(xj) = \frac{1}{n_{\text{tree}}} \sum_{t=1}^{n_{\text{tree}}} \cdot \frac{\sum_{i\epsilon \text{OOB}} I\left(y_i = f\left(X_i\right)\right) - \sum_{i\epsilon \text{OOB}} I\left(y_i = f\left(X_i^j\right)\right)}{|\text{OOB}|}. \quad (1)$$

where VI (x_j) denotes the value significance of x_j. If ignoring (or permuting) a variable decreases the random forest's accuracy, it is regarded as more significant. As a result, variables that have a considerable mean drop in precision are the most important for accuracy.

4. Feature Selection Algorithm Description

4.1. Gene Selection Using Entropy and Information Gain (Ig). Entropy is considered a fundamental term in the information theory that is used to calculate the homogeneity of features. For instance, homogeneous samples have an entropy of zero, while evenly split samples have one value for entropy. A high feature dimensionality data and a minimum size make data categorization exceedingly challenging. A minimal percentage of millions of gene characteristics analyzed are more relevant for given disease data. As a result, only the most important features should be kept. A thorough analysis of the gene profiles would aid in the selection of the gene, which is the most significant for problem identification.

$E(Z) = -D^+ log_2(D^+) - D^- log_2(D^-)$ for a sample with negative and positive attributes.

The entropy model can be summarized as the following equation [28]:

$$\text{Entropy}(Z) = \sum_{i=1}^{v} -\left(D_i \log_2 D_i\right). \quad (2)$$

where $D\,i$ is the probability of categorical variables a priori, and Z and k are indexes in the classification systems that indicate a certain category.

Consider the case of 2 classification issues in particular (V is known for classes). We consider gene with n potential values (j_1, j_2... j_n). The following is the entropy:

$$\text{Entropy}\left(\frac{k}{j}\right) = \sum_{j=1}^{n} p(j) \sum_{k=1}^{v} p\left(\frac{k}{j}\right) \log_2\left(p\left(\frac{k}{j}\right)\right), \quad (3)$$

where $p(k|j)$ is called probability distribution in variation K assuming variable J remains static and is computed across full variables with subclasses. In calculating IG, entropy is their most important factor [29]. The entropy across all variables in the sample of data is determined by the distribution of the features in the data sample. The information is then divided into feature sections. The entropy of each group to be measured independently, and the overall entropy can be computed by combining the entropy data of all groups. The entropy of particular groupings of sample data is then subtracted from their total entropy of the dataset distribution [30]:

$$IG(J) = \text{Entropy}(S) - \text{Entropy}\left(\frac{k}{j}\right). \quad (4)$$

When gene J and category K are unrelated, $IG(J) = Entropy(S)\ Entropy\ (k|j) = zero$, while when they are related, $Entropy(S) > Entropy\ (k|j)$, resulting in $IG(J) > 0$. A greater association between J and K is directly proportional to a larger discrepancy between K and J. For classification, a feature selection within a higher IG value is more relevant. As an outcome, genes with higher IG values are chosen first from the original set of high-dimension genes to serve as the sample for feature gene selection [31].

Those steps of the IG algorithm have illustrated the form of the IG flowchart in Figure 2. The suitable output is a subgroup Y of the real variable W, with a group of attributes W in the input data set. The attributes that will be used for classifiers are first analyzed. Second, for each class, the entropy of all subsamples is calculated using (1). The probability of every value of one attribute is then computed, but the conditional entropy for each attribute is calculated using (2). For all attributes, the IG is calculated using (3). The resulting IG values are sorted ascendingly, with all values over a particular threshold value being selected.

4.2. The Mathematical Model of GWO. Mirjalili and Lewis established the GWO method to discover proper, restricted, and uncontrolled objective functions [32]. Grey wolves' social hierarchy serves as inspiration for the GWO algorithm. Artificial wolves in a virtual environment imitate tracking, encircling, and attacking actions, among others. The GWO social hierarchy divides wolves into four groups: alpha (α), beta (β), delta (δ), and omega (ω). The best option is to think of wolves. The second- and third-placed options are known as β wolves and δ wolves, respectively. Wolves guide α, β, δ wolves' hunting activity, which reflects optimization procedures. The remaining population is regarded as ω, and their movements are conditioned by those three dominant wolves. Figure 3 depicts the power hierarchy of a wolf pack. The wolf commands all of the subordinates,

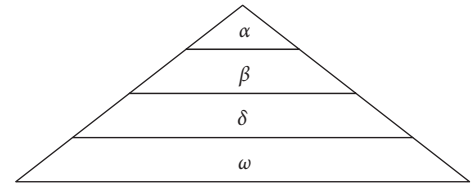

Figure 3: Hierarchical orders of grey wolves in nature.

according to the dominance hierarchy. Similarly, β and δ wolves have an influence on the wolves in their social roles.

Grey wolves begin their hunting behavior by enveloping their target. The following equations are used to represent the encircling behavior in mathematics:

$$\begin{aligned} \vec{D} &= \left|\vec{C}.\vec{X_p}(t) - \vec{X}(t)\right|, \\ \vec{X}(t+1) &= \vec{X_p}(t) - \vec{A}.\vec{D}. \end{aligned} \quad (5)$$

where t represents iterations, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $\vec{X}_p$ in X is the location vectors of a grey wolf, and Y is the position vector of the prey. The coefficients vector is donated in the following way:

$$\begin{aligned} \vec{A} &= 2\vec{a}.\vec{r} - \vec{a}, \\ \vec{C} &= 2.\vec{r_2}. \end{aligned} \quad (6)$$

The variable of the constant vectors $\vec{X}$ decreases linear from 2 to 0 as a $(t) = 2 - (t-1)$ (2/maxIter), where maxIter is the maximum number of iterations.

Hunting behavior follows their circling in the prey. The movements of grey wolves are regulated by donates wolves (α, β, δ):

$$\vec{D_\alpha} = \vec{C_1}\vec{X_\alpha} - \vec{X}, \quad (7)$$

$$\vec{D_\beta} = \vec{C_2}\vec{X_\beta} - \vec{X}, \quad (8)$$

$$\vec{D_\delta} = \vec{C_3}.\vec{X_\delta} - \vec{X}, \quad (9)$$

$$\vec{X_1} = \vec{X_\alpha} - \vec{A_1}.\vec{D_\alpha}, \quad (10)$$

$$\vec{X_2} = \vec{X_\beta} - \vec{A_2}.\vec{D_\beta}, \quad (11)$$

$$\vec{X_3} = \vec{X_\delta} - \vec{A_3}.\vec{D_\delta}, \quad (12)$$

$$\vec{X}(t+1) = \frac{\vec{X_1} + \vec{X_2} + \vec{X_3}}{3}, \quad (13)$$

where β, ξ represents the number of iterations, D and $\vec{C}$ are coefficients vectors, X is the space vectors of a grey wolf, and Y is the position vector of the prey. $\vec{X}$ $(t+1)$ is the resultant vector. The coefficient vector is donated in the following way.

The above equations were utilized to arithmetically model the wolf pack's hunting behavior in this regard. In the simulated environment, (7)–(13). Table 3 concludes with the

Table 3: The GWO's pseudocode.

(1). initialize GWO population of solution vectors
(2). define coefficient vectors $\overrightarrow{A}$, $\overrightarrow{C}$ and $\overrightarrow{a}$
(3). evaluate fitness value f(xi), $i = 1,\ldots,$ nWolf
(4). determine $\overrightarrow{X_\alpha}$, $\overrightarrow{X_\beta}$, and $\overrightarrow{X_\delta}$
(5). **while** iter < maxIter
(6). **for** $i = 1$:nWolf
(7). movement of grey wolves, according to Equation 18
(8). end for
(9) update coefficients $\overrightarrow{A}$, $\overrightarrow{C}$ and $\overrightarrow{a}$
(10). calculate fitness values f(xi), $i = 1,\ldots,$n
(11). update $\overrightarrow{X_\alpha}$, $\overrightarrow{X_\beta}$, and $\overrightarrow{X_\delta}$
(12). iter ←iter + i
(13). **end while**
(14). return $\overrightarrow{X_\alpha}$, f($\overrightarrow{X_\alpha}$)

GWO's pseudocode. The fitness value is calculated by diving the survival rate by the highest survival rate.

4.3. Classification Algorithm Description. In this study, the prediction of colon cancer was tested using random forest, a well-known classification method for prediction models. CART is a mixed classifier made up of unpruned decision trees, whereas RF is a collection of unpruned data sets (classification and regression trees). The CART method is described in great depth in this book [33]. When conducting classification research, the RF forecast is the intermediate majority of each tree category votes [34]. The architecture of an RF model for predicting colon class is shown in Figure 2.

5. Algorithm Description for Random Forest

Here, initial datasets are $D(X, Y)$ and RF creates a simple decision tree: where n is training observations, K is classes, and (x_i, y_i) collection of cases whose class membership is determined (X, Y). (x_i, y_i) can be used to represent the combined classifier (X, Y). Find the best classifier that minimizes error in comparison with the original dataset [35].

6. Description of K-Fold Cross-Validation

Cross-validation is a recreation sample technique in the test machine learning methods on a bounded set of data samples. The unique argument in the technique is k, which denotes the number of sample groups that should be conquered into a set of datasets. As a result, that approach is called as k-fold cross-validation. This technique is known as tenfold cross-validation when the values of k are set to 10 [36].

The steps for training K-fold cross-validation are as follows:

(i) Divided full dataset into k equivalent sections, each one of which is referred to as a fold. The names of the folds should be $f_1, f_2 \ldots f_k$.

(ii) For $i = 1$ to k, preserve the fi bend in the validated model and the subsequent k-1-fold in the classification model.

Table 4: Confusion matrix without feature selection.

Actual	Predicted class	
	Abnormal	Normal
Abnormal	35	5
Normal	5	12

Table 5: Performance analysis of the model without feature selection.

Classes	Recall	Precision	F1-score	Accuracy (%)
Abnormal	00.85724	00.89	00.90	84.870
Normal	00.81	00.7277	00.7618	
Weighted measure (%)	82.58	82.85	82.67	

Table 6: Confusion matrix with feature selection.

Actual	Predict classes	
	Abnormal	Normal
Abnormal	38	2
Normal	3	22

Table 7: Performance of feature selection with the analysis of the models.

	Recall	Precisions	F1-scores	Accuracy (%)
Abnormal	00.95123	00.976	00.9628	96.166
Normal	00.95240	00.90908	00.9702	
Weighted measure (%)	96.15	96.15	96.13	

(iii) Create a model given a dataset, and test its accuracy using the validation data.

(iv) The model's value is defined by the accuracy average of all k-fold cross-validation occurrences.

6.1. System Requirements. Anaconda Enterprise 4.
CPU: 2 × 64 bit 2.8 GHz 8.00 GT/s CPUs.
RAM: 32 GB (or 16 GB of 1600 MHz DDR3 RAM).
Storage: 300 GB.

These experimental data from the three phases are summarized in this section: classification assessment without feature selection, classifiers evaluations with feature selection, and comparison analyzed evaluations. The complete dataset is divided into two groups for experimental testing: normal and abnormal, using each of the 3000 genes. Table 4 shows the correlation coefficient, as well as the performance evaluation between the two groups in terms of recalls, clarity, F1-score, and accurate scores. Our random forest classification model can correctly classify 53 of 63 objects, as shown in Tables 5, yielding weighted recalls, accurate, and F1-scores of 82.78 percent, 82.77 percent, and 82.775 percent, respectively.

Table 8: Comparison of analysis of the models.

Model	Evaluation metric			
	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
Models without feature selection	84.87	84.68	84.68	84.871
Models with feature selection	96.15	96.15	96.13	96.166

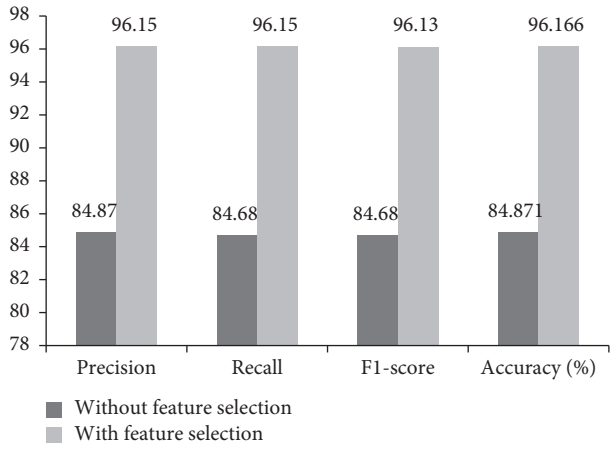

Figure 4: Comparison of analysis of the models.

Table 9: Comparison of the performance with different methods.

Publications	Methods	Number of attributes	Accuracy (%)	Time complexity	Time (ms)
Simone A et al. [13]	FDT	22	79.13	O(n log n)	0.54
Nguyen et al. [14]	MAPH + PNN	5	85.19	O(n ∗ n)	0.21
Lingyun Gao et al. [15]	FCBFS + SVM	14	90.45	O(n log n)	0.43
Salem H et al. [16]	GP + IG + GA	60	84.68	$O(n+k)$	0.34
Our proposed method	**IG + GWO**	**33**	**95.16**	O(log n)	**0.12**

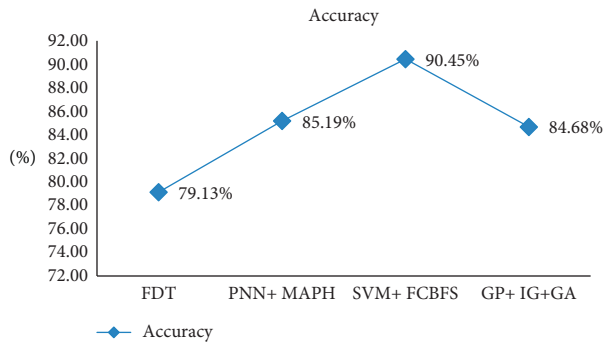

Figure 5: Graphical comparison of the model for various evaluation metric.

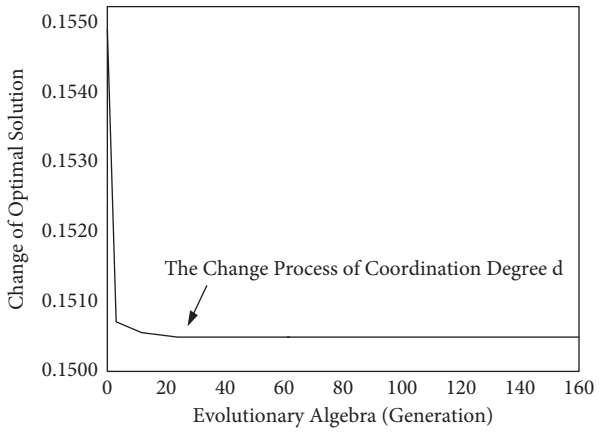

Figure 6: Process execution evolutionary timeline for the proposed system.

Using all genes, this model is 84.870 percent accurate. To eliminate the more relevant and redundant genes from every dataset, we used mean decreased with accuracy and means decreased Gini as feature selection methods. In Tables 6 and 7, the finally confusion matrix and efficiency measures depend upon the top 33 genes, respectively.

The models depending upon the top 35 identified genes could accurately recognize 57 samples out of 61 samples with a 95.161 percent accuracy. In addition, the models received 94.10 percent for weighted recalls, accuracy, and F1-scores, and 94.10 percent for weighted recalls, accuracy, and F1-scores. Table 7 shows a comparison of the model with and without feature selection [37–47].

When the models with feature selection were utilized, all of the measured parameters outperform their equivalents

when the model without feature selection was employed, as shown in Table 8. Figure 4 depicts the model's overall findings based on performance metrics in a graphical representation. Table 9 shows that the time complexity of our proposed system is the best.

Figure 5 shows a comparison of our suggested models and previous techniques. Table 9 illustrates that our technique outperforms all other procedures with less information about the expression of genes. Figure 6 describes the process execution evolutionary timeline for the proposed system.

7. Conclusion

A "gain information"- and "Grey wolf optimization"-based ensemble model was incorporated as an optimal feature selection approach into the random forest classifier to improve the prediction model's accuracy. Our suggested technique can decrease the load of high-dimensional data, and it allows quicker computations. Additionally, we provided the comparison of classification model analytics for feature selection over prediction analysis without feature selection. The extensive experimental findings have shown that the suggested method with selecting features is beneficial, outperforming the good classification performances. In fact, feature selection is an accurate method, and its time complexity increases exponentially with the problem scale. In order to alleviate the running time of large-scale instance, multilevel window technology can be adopted. That is, the preprocessing scheme could be subdivided for multiple levels of windows to reduce the number of jobs in a window. By making full use of the time period in which the tasks of the previous window are being processed, the tasks of the next window are solved by using constraint programming solver. Therefore, this work can serve as the research basis for feature selection problem. [47].

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] M. C. Wong, H. Ding, J. Wang, P. S. Chan, and J. Huang, "Prevalence and risk factors of colorectal cancer in Asia," *Intestinal research*, vol. 17, no. 3, pp. 317–329, 2019.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] M. Xi, J. Sun, L. Liu, F. Fan, and X. Wu, "Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine," *Computational and Mathematical Methods in Medicine*, vol. 20169 pages, Article ID 3572705, 2016.

[4] S. N. Ghazavi and T. W. Liao, "Medical data mining by fuzzy modeling with selected features," *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 195–206, 2008.

[5] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[6] M. S. Kumar and J. Prabhu, "A hybrid model collaborative movie recommendation system using K-means clustering with ant colony optimisation," *International Journal of Internet Technology and Secured Transactions*, vol. 10, no. 3, p. 337, 2020.

[7] M. S. Kumar and J. Prabhu, "Comparison of multi-criteria recommendation system for improving accurate prediction," *Journal of Advanced Research in Dynamical & Control Systems*, vol. 10, no. 8, 2018.

[8] G. G. Wang, S. Deb, and Z. Cui, "Monarch butterfly optimization," *Neural Computing & Applications*, vol. 31, no. 7, pp. 1995–2014, 2019.

[9] M. S. Kumar and J. Prabhu, "Hybrid model for movie recommendation system using fireflies and fuzzy c-means," *International Journal of Web Portals*, vol. 11, no. 2, pp. 1–13, 2019.

[10] D. Oliva, S. Esquivel-Torres, S. Hinojosa et al., "Opposition-based moth swarm algorithm," *Expert Systems with Applications*, vol. 184, Article ID 115481, 2021.

[11] M. Sandeep Kumar and J. Prabhu, "A case study on recommendation systems based on big data," in *Smart Intelligent Computing and Applications*, pp. 407–417, Springer, Singapore, 2019.

[12] I. Ahmadianfar, A. A. Heidari, A. H. Gandomi, X. Chu, and H. Chen, "RUN beyond the metaphor: an efficient optimization algorithm based on Runge Kutta method," *Expert Systems with Applications*, vol. 181, Article ID 115079, 2021.

[13] J. Tu, H. Chen, M. Wang, and A. H. Gandomi, "The colony predation algorithm," *Journal of Bionics Engineering*, vol. 18, no. 3, pp. 674–710, 2021.

[14] I. Ahmadianfar, A. A. Heidari, S. Noshadian, H. Chen, and A. H. Gandomi, "INFO: an efficient optimization algorithm based on weighted mean of vectors," *Expert Systems with Applications*, vol. 195, Article ID 116516, 2022.

[15] H. N. Nguyen, T. N. Vu, S. Y. Ohn, Y. M. Park, M. Y. Han, and C. W. Kim, "Feature elimination approach based on random forest for cancer diagnosis," in *Proceedings of the Mexican International Conference on Artificial Intelligence*, pp. 532–542, Springer, Berlin, Germany, November 2006.

[16] X. F. Song, Y. Zhang, Y. N. Guo, X. Y. Sun, and Y. L. Wang, "Variable-size cooperative coevolutionary particle swarm optimization for feature selection on high-dimensional data," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 5, pp. 882–895, 2020.

[17] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Applied Soft Computing*, vol. 50, pp. 124–134, 2017.

[18] J. Bennet, C. Ganaprakasam, and N. Kumar, "A hybrid approach for gene selection and classification using support vector machine," *The International Arab Journal of Information Technology*, vol. 12, pp. 695–700, 2015.

[19] C. Gunavathi and K. Premalatha, "Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification," *International Journal of Comput Electronic Automation Control and Information Engineering*, vol. 8, no. 8, pp. 1490–1497, 2014.

[20] V. B. Canedo, N. S. Maroño, and A. A. Betanzos, "An ensemble of filters and classifiers for microarray data classification," *Pattern Recognition*, vol. 45, no. 1, pp. 531–539, 2012.

[21] J. Y. Yeh, "Applying data mining techniques for cancer classification on gene expression data," *Cybernetics & Systems*, vol. 39, no. 6, pp. 583–602, 2008.

[22] H. Salem, G. Attiya, and N. El-Fishawy, "Early diagnosis of breast cancer by gene expression profiles," *Pattern Analysis & Applications*, vol. 20, no. 2, pp. 567–578, 2017.

[23] A. Wang, S. Ve, W. A. Hatamleh, K. D. Haouam, B. Venkatesh, and D. Sweidan, "Advanced lightweight feature interaction in deep neural networks for improving the prediction in click through rate," *Annals of Operations Research*, pp. 1–15, 2021.

[24] A. Wang, H. Liu, J. Yang, and G. Chen, "Ensemble feature selection for stable biomarker identification and cancer classification from microarray expression data," *Computers in Biology and Medicine*, vol. 142, Article ID 105208, 2022.

[25] E. A. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: a systematic review," *Computers in Biology and Medicine*, vol. 140, Article ID 105051, 2022.

[26] K. Kourou, G. Rigas, C. Papaloukas, M. Mitsis, and D. I. Fotiadis, "Cancer classification from time series microarray data through regulatory dynamic bayesian networks," *Computers in Biology and Medicine*, vol. 116, Article ID 103577, 2020.

[27] R. Kundu, S. Chattopadhyay, E. Cuevas, and R. Sarkar, "AltWOA: altruistic whale optimization algorithm for feature selection on microarray datasets," *Computers in Biology and Medicine*, vol. 144, Article ID 105349, 2022.

[28] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 201513 pages, Article ID 198363, 2015.

[29] U. Alon, N. Barkai, D. A. Notterman et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.

[30] J. C. Baez, T. Fritz, and T. Leinster, "A characterization of entropy in terms of information loss," *Entropy*, vol. 13, no. 11, pp. 1945–1957, 2011.

[31] M. Walowe Mwadulo, "A review on feature selection methods for classification tasks," *International Journal of Computer Applications Technology and Research*, vol. 5, no. 6, pp. 395–402, 2016.

[32] S. Ve, C. Shin, and Y. Cho, "Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city," *Building Research & Information*, vol. 49, no. 1, pp. 127–143, 2021.

[33] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[34] D. J. Hand, "Principles of data mining," *Drug Safety*, vol. 30, no. 7, pp. 621-622, 2007.

[35] R. Harb, X. Yan, E. Radwan, and X. Su, "Exploring precrash maneuvers using classification trees and random forests," *Accident Analysis & Prevention*, vol. 41, no. 1, pp. 98–107, 2009.

[36] L. Y. Chang and H. W. Wang, "Analysis of traffic injury severity: an application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019–1027, 2006.

[37] S. Sucharita, B. Sahu, and T. Swarnkar, "A comprehensive study on the application of grey wolf optimization for microarray data," *Data Analytics in Bioinformatics: A Machine Learning Perspective*, pp. 211–248, 2021.

[38] S. Ve and Y. Cho, "A rule-based model for Seoul Bike sharing demand prediction using weather data," *European Journal of Remote Sensing*, vol. 53, no. 1, pp. 166–183, 2020.

[39] V. E. Sathishkumar, J. Park, and Y. Cho, "Using data mining techniques for bike sharing demand prediction in metropolitan city," *Computer Communications*, vol. 153, pp. 353–366, 2020.

[40] R. M. Aziz, "Nature-inspired metaheuristics model for gene selection and classification of biomedical microarray data," *Medical & Biological Engineering & Computing*, vol. 60, pp. 1–20, 2022.

[41] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar et al., "Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators," *Knowledge-Based Systems*, vol. 223, Article ID 107034, 2021.

[42] R. M. Aziz, *Application of Nature Inspired Soft Computing Techniques for Gene Selection: A Novel Frame Work for Classification of Cancer*, Springer, Berlin, Germany, 2022.

[43] A. Hajieskandar, J. Mohammadzadeh, M. Khalilian, and A. Najafi, "Molecular cancer classification method on microarrays gene expression data using hybrid deep neural network and grey wolf algorithm," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2020.

[44] R. Aziz, C. K. Verma, M. Jha, and N. Srivastava, "Artificial neural network classification of microarray data using new hybrid gene selection method," *International Journal of Data Mining and Bioinformatics*, vol. 17, no. 1, p. 42, 2017.

[45] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," *Expert Systems with Applications*, vol. 166, Article ID 114012, 2021.

[46] N. Yuvaraj, K. Srihari, G. Dhiman et al., "Nature-inspired-based approach for automated cyberbullying classification on multimedia social networking," *Mathematical Problems in Engineering*, vol. 202112 pages, Article ID 6644652, 2021.

[47] R. A. Musheer, C. K. Verma, and N. Srivastava, "Novel machine learning approach for classification of high-dimensional microarray data," *Soft Computing*, vol. 23, no. 24, pp. 13409–13421, 2019.