Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, No. 2, No.01 (2023),
ISSN : **1906-9685**

OPTIMAL GENE CLASSIFICATION BASED ON HYBRID OPTIMIZATION ALGORITHMS AND ADAPTIVE BOOSTING USING ANN MODEL

Dr. G. Kalpana Associate Professor, Department of Computer Science,
Sri Ramakrishna College of Arts & Science for Women, Coimbatore kalpanacs@srcw.ac.in
M. Divyavani Ph.D. Research Scholar, Department of Computer Science,
Sri Ramakrishna College of Arts & Science for Women, Coimbatore divicse07@gmail.com

Abstract
Gene expression profiles have attracted a lot of attention for tumor classification in DNA microarray datasets, and gene selection is a key factor in enhancing the microarray data's classification performance. In this proposed research for hybrid feature selection and adaptive boosting on an Artificial Neural Network (ANN) as a classifier, and to test the colon microarray dataset. This approach is highly effective, as evidenced by the preliminary results obtained under a number of publicly available gene expression datasets. It can greatly reduce the dimensionality of gene expression datasets and accurately identify the most relevant genes. With respect to the current algorithm, the experimental results indicate that the proposed model better achieving 97 %. accuracy.

Keywords__ Colon Microarray, Feature selection, Biomarker, Artificial Neural Network, Firefly Optimization Algorithm

INTRODUCTION

Bioinformatics is a growing field that enables biologists to analyze organism data at various levels. However, predicting and classifying microarray data in biomedicine is challenging due to the rapid advancement of DNA microarray technology. This is particularly important for tumor classification, which is crucial for accurate cancer diagnosis and subtype identification. Current computational methods struggle to identify important genes, leading to higher learning costs and poor performance. The goal is to improve classification accuracy by developing effective gene selection techniques [1-2].

DNA microarray technology enables researchers to monitor thousands of genes simultaneously in experiments. However, the high-dimension, small sample size, and high-noise nature of gene expression data present analytical challenges. To improve classification accuracy, researchers use a decision tree algorithm and 10-fold cross-validation on three benchmark gene expression data sets. Further tests on trained ANN models with Adaboosting Classifier are conducted on various samples and cancer classification. The suggested strategy improves classification accuracy by using best parameters, reducing data set dimension, and validating the gene subset with the most information. Gene selection and classification in DNA microarray data are successfully achieved using ANN, supervised learning, and evolutionary algorithms [3-4].

The simultaneous monitoring of thousands of genes in a single experiment is made possible by DNA microarray technology. Nevertheless, the high-dimension, small sample size, and high noise level of gene expression data make analysis challenging. Finding a small subset of pertinent genes to increase classification accuracy while maintaining robustness is the major challenge. The colon microarray gene

expression data set is used to test the suggested method, which uses the decision tree algorithm and 10-fold cross-validation to achieve accuracy. Additionally, the trained ANN models with an adapting Adaboosting Classifier are tested for sample analysis and cancer classification. Investigational results demonstrate that the suggested method can improve classification accuracy with optimal parameters based on datasets, reduce the dimension of the data set, and validate the most informative gene subset. While supervised learning algorithms like K-nearest neighbor, decision trees, support vector machines, linear discriminant analysis, and artificial neural networks (ANN) have proven effective in classifying microarray data, evolutionary algorithms like genetic algorithms, hybrid PSO/GA, and particle swarm optimization have been used for gene selection [5-7].

The majority of technical and scientific fields have used artificial intelligence (AI) research more extensively and more popularly to build models for resolving a variety of issues. Intelligent systems that mimic or replicate human problem-solving abilities are included in artificial intelligence. ANNs are among these intellectual mechanisms, and they offer several benefits, including the capacity to learn and process vast amounts of irrelevant data. These mechanisms are derived from active and nonlinear techniques, where nonlinearities and variable interactions play a crucial role. Furthermore, ANNs' ability to solve a variety of issues, including classification issues, has been noticed by a few studies [7-9]. This research proposes a hybrid firefly optimization algorithm for tumor classification in DNA microarray datasets. The algorithm uses feature selection and adaptive boosting on an artificial neural network, outperforming the current algorithm by 97%. The approach reduces dimensionality and accurately identifies relevant genes in publicly available gene expression datasets.

In this paper, the background of the study is covered in Section II, methods and materials on clear view are covered in Section III, and the results are discussed in Section IV. The followed by next section V is conclusion.

BACKGROUND OF THE STUDY

In recent times, bioinformatics has emerged as a noteworthy area of study, providing biologists with the ability to examine organism data at the genomic, transcriptomics, and proteome levels. Microarray data prediction and classification is a major task in biomedicine [10-12]. This work is difficult due to the quick advancement of DNA microarray technology, since gene expression datasets frequently comprise thousands of genes but few samples [13]. A common issue with microarray gene expression data is tumor classification, which includes rare disease prediction and tumor detection. For a precise cancer diagnosis and subtype identification, these investigations are essential. However, many computational methods fail to identify important genes, leading to higher learning costs and worse performance [14-16]. This is because of the limited availability of samples and the abundance of genes in microarray data. The objective is to enhance the classification accuracy of cancer gene expression datasets by creating effective gene selection techniques that lower the dimensionality of microarray data [17-19].

The process of gene selection is used to improve classification accuracy and decrease the dimensionality of microarray data. In addition to selecting related genes and lowering computational costs, it eliminates noisy and unnecessary genes [20-23]. Filter, wrapper, embedded, and hybrid approaches are the four categories into which feature selection methods fall. The high speed and large dataset handling capabilities of filter methods make them popular, but they are susceptible to local optimum trapping [23-25]. Here described some of the key investigation about the model as follows,

Sun et.al., [26] presented a multi-filter ensemble method for gene selection based on cross-entropy; however, wrapper methods are computationally expensive, particularly when dealing with high-dimensional microarray datasets. Although this wrapper approach required a lot more processing power, it achieved superior classification accuracy [28]. In general, the framework can be created to combine the benefits of both the wrapper and the filter methods for feature selection in order to obtain a method that is accurate and efficient [32]. Wang et al.,[27] introduced a wrapper-based gene selection method that

improved classification accuracy but required more processing power. It used the Markov blanket technique to shorten evaluation times. embedded techniques, like those of Lopes et al., [29] is used for ensemble classification, and Li et al., [30] embedded feature selection algorithm, need more parameter adjustments and take a long time. hybrid strategies, such as Mav et al., [31] hybrid gene selection strategy can combine the benefits of feature selection using both filter and wrapper methods for an effective and precise approach. Algamal et al., [33] created a method for classifying gene expression that uses Bayesian Lasso quantile regression to account for outliers in the gene data. Lin et al., [34] suggested a method for gene selection to produce several subsets with varying gene combinations to aid in classification tasks. Intuitively sound artificial intelligence algorithms are computationally expensive, which has reduced the quantity and caliber of research in the area. Higher-speed processors are now being used to train and develop artificial intelligence algorithms, which has resulted in a notable rise in the number of successful studies. Maximizing the potential requires figuring out the optimal set of parameters [35].

In bioinformatics, wrapper models—which select feature subsets with higher prediction accuracy using search techniques like genetic algorithms (GA), particle swarm optimization (PSO), and ant colony optimization (ACO)—have become more and more popular. These techniques do, however, have certain drawbacks, including high computational costs and local optimum. Certain techniques exhibit sluggish convergence rates, are susceptible to local optimum phenomena, and eventually become overloaded [36-38]. Researchers are looking into adapting boosting on ANN classifiers and hybrid firefly wrapper-based feature selection as solutions to these problems. Hybrid techniques have been developed to combine the advantages of filter and wrapper approaches, balancing accuracy and efficiency when choosing the best feature subset. The purpose of this paper is to examine and enhance these hybrid approaches [39].

METHODS AND MATERIALS

This section presents the modified hybrid algorithm as well as the gene expression datasets used in this work. We test our proposed algorithm on three datasets. The dataset for colon cancer categories was created by Alon U. (1999) [22]. There are 40 samples for the tumor class and 22 samples for the normal class in these data. There are 2000 genes in this dataset. The data used in this paper are summarized in Table 1.

TABLE 1: MICROARRAY DATASET DESCRIPTION

Dataset	Class	Features	Instances	Samples
Colon	Normal, Cancer	1000/2000	62	40,22

Proposed Model: Hybrid FFF and Adaptive boosting on ANN as Classifier

Neural networks (NNs) are employed in real-world applications such as data mining and pattern recognition. As machine learning (ML) can select pertinent features from datasets, feature selection (FS) is an essential task in bioinformatics. There exist three categories of FS techniques: wrapper, embedded, and filter. Filter methods employ FS as a preliminary step prior to utilizing a classification algorithm, whereas wrapper methods integrate a classifier for the purpose of choosing pertinent features within an array. The filter approach uses FS as a preprocessing step prior to applying a classification algorithm, whereas the wrapper approach depends on several data evaluations. Aside from being less computationally demanding than wrapping approaches, embedded approaches are specific to an assumed learning algorithm and offer benefits like communication with the classification model. In general, ANNs work well for real-time classification tasks and are efficient at deriving meaning from complicated issues.

The process of classifying dataset elements into predefined groups is known as cancer classification, and it is based on machine learning. It can be used in medicine to diagnose and detect conditions like prostate, ovarian, colon, heart, and breast cancer, among other illnesses. The gathering and testing of data, data

cleaning, and the classification of sensitivity and specificity present difficulties, though. The suggested model uses an ANN as a classifier and combines hybrid FFF and adaptive boosting.

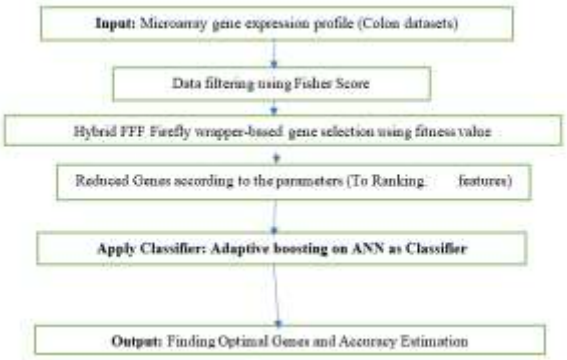


Figure 1: Architecture of Proposed Model: Hybrid FFF and Adaptive boosting on ANN as Classifier

A. Feature Selection: Hybrid FFF (Firefly wrapper-based feature selection)

A data pre-processing technique called feature selection (FS) is used to choose the most pertinent subset of the primary feature dataset and to reduce the dimensionality of the data. The Firefly Optimization Algorithm (Hybrid FFF) was utilized in this study to choose pertinent features. The process determines each feature's fitness value and ranks them. For every dataset, the 25 most prominent features were chosen. The filtered dataset was used to find predictive genes that maximize adaptive boosting on ANN classification performance using the Firefly Wrapper feature selection method, which is based on an evolutionary bio-inspired algorithm. Only the brightest firefly in the swarm is selected as the solution by the method, which compares each firefly to the others based on brightness.

B. Optimization of Hybrid FFF

The firefly FS method's methodology compares each firefly based on brightness. The best firefly is then selected based on their fitness function, which returns the solution. The feature selection method selected the top 25 features for each dataset1. It calculates the values of each feature based on its fitness value and checks the rank (modifying the fitness function). The higher rank features are sorted according to the rank. Formulated light intensity: (as the goal function i. e. I=Io e-Yr2 ij (1) Io= is the light intensity value obtained from the maximum distance were calculated, and this is referred to as the fitness of each firefly. (two fireflies) The attractiveness variation is represented by Y, whose value ranges from 0 to 200. βo is always set to 1, and rand is a random number between [1,1] that is set in ∝ order to allow the variation in the solution.

C. Classification: Adaptive boosting on ANN as Classifier

A network of tiny processors connected to solve problems is called an artificial neural network (ANN), which is an information processing system inspired by the human brain. Input, hidden, and output are its three layers. Weighted vectors and inputs determine the performance of the hidden layers, while raw data and feature vectors are sent to the input layer. The ratio of hidden to output units determines how well the output layer performs. Reducing the error between the intended output and the network's output is the aim of locating and estimating the mapping function between input and output spaces. An artificial neural network's weight matrices, which are contingent on the quantity of neurons in its hidden and output layers, comprise intermediate layers and output layer weight matrices.

When neural networks are being trained, the optimal model is chosen based on factors such as weights and neuron count. Assigning an input pattern as a gene expression profile to one of the introduced classes—such as normal or cancer—is the task of pattern classification in Adapting Boosting on ANN. On the basis of training, the model can forecast the class of fresh data. First, the data must be prepared by filtering, normalizing, and removing genes with low expression values, low information randomness, low

expression value changes, or noise. The hybrid Firefly Optimization algorithm system is used to apply the top-ranked genes. Ninety percent of the data is used for training and assessment, and ten percent is used for blind testing. Considerations include population size, chromosome length, inertia coefficient, mutation rate, training factors, and maximum rate.

D. Training based Adaptive boosting on ANN as Classifier

In order to maximize classification accuracy while minimizing the number of chosen predictive biomarker genes, the phase computes the fitness function for the hybrid FFF-Adaptive boosting on ANN classifier. K-nearest neighbor, Euclidian distance measures, and incremental weighting values are used to generate the accuracy.

To start, follow these steps: Step 1: Upload the colon microarray expression dataset; Step 2: Identify Input and Output; Step 3: Partition the Dataset for Training, Testing, and Validation; Step 4: Set the Neuron (start at 10 by default); Step 5: Train the Network Based on Adaptive Boosting on ANN as Classifier; Step 6: Record Error Performance; Step 7: Check the Error Performance. If performance is low, repeat steps 4–7 until high performance is achieved. Stop training if performance meets your expectations. Step 8: Test the network. Step 9: The network is ready to operate. End.

The process involves uploading a colon microarray expression dataset, identifying input and output, partitioning the dataset for training, testing, and validation, setting the default neuron, training a network using adaptive boosting on an ANN as a classifier, recording error performance, checking performance, repeating steps if necessary, stopping training if performance meets expectations, testing the network, and ensuring it is ready to operate.

RESULTS AND DISCUSSION

To improve colon cancer classification accuracy, the study used a neural network architecture (ANN). Data preprocessing, which removed redundancy and noise, came after data collection. MATLAB (2016a) was used for data partitioning. After that, the model underwent testing, validation, and training to produce an accurate estimate for the classification of colon cancer. Using the following formula for calculating accurate classification,

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \times 100.$$

A model for classifying colon cancer was implemented by the research using MATLAB Tool, colon dataset hybrid FFF, and adaptive boosting on ANN Algorithm. The Firefly Wrapper technique was used to carefully select neurons for the model's single hidden layer. The confusion matrix was used to determine the model's accuracy. The 70% was the ideal allocation for training, 15% for validation, and 15% for testing. MATLAB (2016a) software and the cancer dataset were used to simulate the experiments. Because the parameters for Adaptive boosting on ANN were appropriate, the average test accuracy was 97.97 percent, indicating good stability.

TABLE 2: RESULTS OF OPTIMAL GENE SELECTION FOR COLON CANCER DATASET

Dataset	**Colon Cancer Dataset**
No of Genes	7, 4, 4, 4,5
Optimal Gene Selected	0248, 0465,0642, 1311, 1313, 1671, 1772 1311, 1761, 1762, 1817 066, 1571, 1762, 1817 1311, 1761, 1762, 1817 1077, 1311, 1671, 1872, 1917
Best Accuracy	97.97%
Method	Hybrid FFF-Adaptive boosting on ANN as Classifier

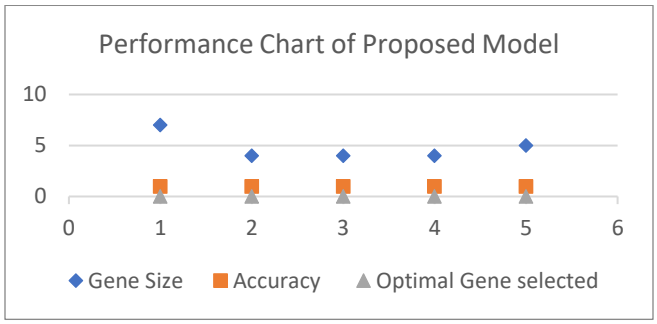
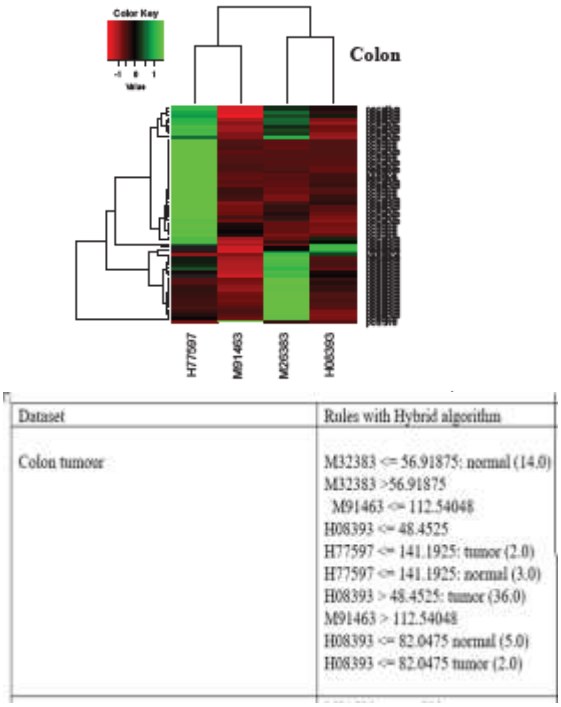


Figure 2 : Performance Chart of Proposed Model



Dataset	Rules with Hybrid algorithm
Colon tumour	M32383 <= 56.91875: normal (14.0) M32383 >56.91875 M91463 <= 112.54048 H08393 <= 48.4525 H77597 <= 141.1925: tumor (2.0) H77597 <= 141.1925: normal (3.0) H08393 > 48.4525: tumor (36.0) M91463 > 112.54048 H08393 <= 82.0475 normal (5.0) H08393 <= 82.0475 tumor (2.0)

Figure 3: Heat Map of Colon Microarray and prediction ules of colon least biomarkers genes

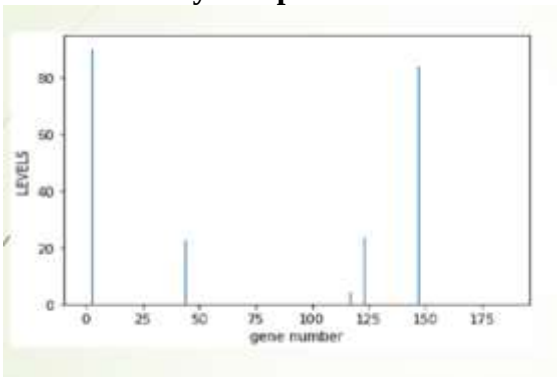


Figure 4: gene selection representation chart in proposed model

CONCLUSION

Artificial Intelligence (AI) research has become increasingly popular in various technical and scientific fields, including artificial neural networks (ANNs). ANNs are intelligent systems that mimic human problem-solving abilities and offer benefits such as learning and processing vast amounts of irrelevant data. These mechanisms are derived from active and nonlinear techniques, where nonlinearities and variable interactions play a crucial role. ANNs have been shown to solve various issues, including

classification problems, as examined in some studies. This work suggests a novel approach for better classification accuracy and biomarker discovery on ANNs called Hybrid FFF-Adaptive boosting. Adaboost with cross-validation is used for sample classification, and hybrid FFF is used for gene selection. Experiments conducted on colon microarray gene expression profiles validate that this method has the better classification accuracy.

REFERENCES

1] Peng, S., Xu, Q., Ling, X.B., et al., "Molecular Classification of Cancer Types from Microarray Data Using the Combination of Genetic Algorithm and Support Vector Machines," FEBS Letter, issue 552, pp. 358-362, 2003.

[2] Yang S, F. Han, J. Guan, "A Hybrid Gene Selection and Classification Approach for Microarray Data Based on Clustering and PSO," Springer, pp. 88-93, 2013.

[3] Sahua B, D. Mishrab, "A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data", International Conference on Modelling Optimization and Computing (ICMOC-2012), pp. 27-31, 2012.

[4] Shutao L, et al., "Gene selection using hybrid particle swarm optimization and genetic algorithm," Soft Computing, pp. 39–48, 2008.

[5] Moteghaed N, et al., "Biomarker Discovery Based on Hybrid Optimization Algorithm and Artificial Neural Networks on Microarray Data for Cancer Classification," Journal of Medical and Signals Sensors, pp.88-96, 2015.

[6] Shen Q, et al., "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," Computational Biology and Chemistry, Volume 32, Issue 1, pp. 53-60, 2008.

[7] Shen Q, et al., "A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification," Talanta, pp. 79–83, 2007.

[8] Kun-Huang Chen, et al., "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm," BMC Bioinformatics, 2014.

[9] Li-Yeh Chuang, et al., "A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data," Journal of Computational Biology, vol. 19, pp.68-82, 2012.

[10] Greenman, C. D. Haplo insufficient gene selection in cancer. Science **337,** 47–48, 2012.

[11] Li, Z. J., Liao, B., Cai, L. J., Chen, M. & Liu, W. H. Semi-supervised maximum discriminative local margin for gene selection. Scientific reports **8**, 2018.

[12] Sun, L. et al. Joint neighbourhood entropy-based gene selection method with fisher score for tumor classification. Applied Intelligence **49**(4), 1245–1259 (2019).

[13] Cao, J., Zhang, L., Wang, B. J., Li, F. & Yang, J. A fast gene selection method for multi-cancer classification using multiple support vector data description. Journal of Biomedical Informatics **53**, 381–389 (2015).

[14] Sun, L., Zhang, X. Y., Xu, J. C., Wang, W. & Liu, R. N. A gene selection approach based on the fisher linear discriminant and the neighbourhood rough set. Bioengineered **9**(1), 144–151 (2018).

[15] Liu, J., Cheng, Y. H., Wang, X. S., Zhang, L. & Wang, Z. J. Cancer characteristic gene selection via sample learning based on deep sparse filtering. Scientific Reports **8**, 8270 (2018).

[16] Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D. & Maulik, U. Recursive Memetic algorithm for gene selection in microarray data. Expert Systems with Applications **116**, 172–185 (2019).

[17] Sun, L. & Xu, J. C. Feature selection using mutual information-based uncertainty measures for tumor classification. Bio-Medical Materials and Engineering **24**, 763–770 (2014).

[18] Alanni, R., Hou, J. Y., Azzawi, H. & Xiang, Y. A novel gene selection algorithm for cancer classification using microarray datasets. BMC Medical Genomics **12**, 10 (2019).

[19] Sun, L., Xu, J. C. & Tian, Y. Feature selection using rough entropy-based uncertainty measures in incomplete decision systems. Knowledge-Based Systems **36**, 206–216 (2012).
[20] Sun, L., Xu, J. C. & Yin, Y. Principal component-based feature selection for tumor classification. Bio-Medical Materials and Engineering **26**, S2011–S2017 (2015).
[21] Sun, L., Wang, L. Y., Xu, J. C. & Zhang, S. G. A neighborhood rough sets-based attribute reduction method using Lebesgue and entropy measures. Entropy **21**(2), Article ID: 138 (2019).
[22] Wang, C. Z., Shi, Y. P., Fan, X. D. & Shao, M. W. Attribute reduction based on k-nearest neighborhood rough sets. International Journal of Approximate Reasoning **106**, 18–31 (2019).
[23] Sun, L., Zhang, X. Y., Xu, J. C. & Zhang, S. G. An attribute reduction method using neighborhood entropy measures in neighborhood rough sets. Entropy **21**(2), Article ID: 155 (2019).
[24] Sun, L., Liu, R. N., Xu, J. C., Zhang, S. G. & Tian, Y. An affinity propagation clustering method using hybrid kernel function with LLE. IEEE Access **6**, 68892–68909 (2018).
[25] Sina, T., Ali, N., Reza, R. & Parham, M. Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing **168**, 1024–1036 (2015).
[26] Sun, Y. Q., Lu, C. B. & Li, X. B. The cross-entropy based multi-filter ensemble method for gene selection. Genes **9**(**5**), (2018).
[27] Wang, A. G. et al. Wrapper-based gene selection with Markov blanket. Computers in Biology and Medicine **81**, 11–23 (2017).
[28] Chen, G. & Chen, J. A novel wrapper method for feature selection and its applications. Neurocomputing **159**, 219–226 (2015).
[29] Lopes, M. B. et al. Ensemble outlier detection and gene selection in triple-negative breast cancer data. BMC Bioinformatics **19**(1), 168–182 (2018).
[30] Li, J. T., Jia, Y. M. & Li, W. L. Adaptive huberized support vector machine and its application to microarray classification. Neural Computing and Applications **20**, 123–132 (2011).
[31] Mav, D. et al. A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. PloS One **13**(**2**), Article ID: e0191105 (2018).
[32] Rodriguez-Galiano, V. F., Luque-Espinar, J. A., Chica-Olmo, M. & Mendes, M. P. Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods. Science of the Total Environment **624**, 661–672 (2018).
[33] Algamal, Z. Y., Alhamzawi, R. & Ali, H. T. M. Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression. Computers in Biology and Medicine **97**, 145–152 (2018).
[34] Lin, H. Y. Reduced gene subset selection based on discrimination power boosting for molecular classification. Knowledge-Based Systems **142**, 181–191 (2018).
[35] Markid, H. Y., Dadaneh, B. Z. & Moghaddam, M. E. Bidirectional ant colony optimization for feature selection. IEEE International
Symposium on Artificial Intelligence and Signal Processing 53–58 (2015).
[36] Shah, S. & Kusiak, A. Cancer gene search with data-mining and genetic algorithms. Computers in Biology and Medicine **37**(2), 251–261 (2007).
[37] Jain, I., Jain, V. K. & Jain, R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Applied Soft Computing **62**, 203–215 (2018).
[38] Yu, H. L., Gu, G. C., Liu, H. B., Shen, J. & Zhao, J. A modified ant colony optimization algorithm for tumor marker gene selection. Genomics Proteomics & Bioinformatics **7**, 200–208 (2009).
[39] Shukla, A. K., Singh, P. & Vardhan, M. A hybrid gene selection method for microarray recognition. Biocybernetics and Biomedical Engineering **38**(4), 975–991 (2018).