International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Survey of Feature Selection Methods for the Analysis of Microarrays Data in Cancer

Dr. Shemim Begum[1], Ebne Saud Khan[2], Dr. Debasis Chakraborty[3]

Submitted: 27/05/2023 **Revised**: 05/07/2023 **Accepted**: 24/07/2023

Abstract: Cancerous gene selection and cancer identification are of great concern to biologists in interpreting the movements of genes in tissues at the molecular level. A Huge number of genes compared to fewer samples in microarray data pose a great difficulty in designing an appropriate machine learning model. To diagnose cancer and classify its types, obtaining significant genes analogous to cancer is crucial. Hence, it is a feature selection (FS) from gene expression data. Microarray datasets are noisy. Hence, significant FS algorithms are essential to select significant genes for classification. This paper depicts a review of FS methods, that have been reported in many journals to make use of microarray data-based cancer diagnosis. We hope this review will guide researchers to upgrade algorithmic developments in cancerous gene identification.

Index Terms: *cancer identification, feature selection, microarray data, Support vector machines*

1. Introduction

Microarray is an emerging technology that acknowledged the expression profiles of thousands of biomolecules to be examined simultaneously over distinct experimental conditions. Today, the evaluation of gene expression data accelerates novel insights into biological approaches through clustering [1 ,2], cancerous gene identification [3, 4], classification [5, 6], cancer subtype prediction [7-9], and so on. For example, the classification of microarray data lets the discovery of unlabeled data in expression profiles and identifies many diseases precisely. Clinical decision support in terms of diagnosis of cancer and the prognosis of clinical outcomes in response to treatment is a clear example of the medical application of microarray expression profiles. The medically successful, microarray-based diagnostic model solely rests on the benefits and pitfalls of the available FS methods. Although earlier research exhibits the usefulness of developing the right model for the diagnosis of cancer, corresponding research has reported finite experiments.

Concerning the FS algorithm, few datasets and types of cancer are involved. However, the report concluded from these studies cannot be concluded as a comparative study of analysis, because each report is made on account of different experimental methods, and each study incorporates learning algorithms distinctly. The best FS method, among the many possible methods, is, therefore, not perceptible from the literature. The integration of classification and gene selection in studying cancer datasets is not properly understood. The dataset is immersed in many genes with very few samples in microarray expression analysis.

The article [10, 11], prevailing a great challenge in identifying the most significant genes in each problem. This problem can be resolved by identifying significant genes using the FS method. Gene selection is an important task in improving the efficiency of the classifier [12] and, this is an essential part of the exploration of microarray data [13].

An effective FS method has several advantages in such a situation, where thousands of features are held by the datasets. Initially, dimensions are minimized to reduce the computing cost. Second, noises are removed to improve the efficiency of the classifier. Eventually, features that can be inferred in a better manner and can be usable in diagnosing the target disease are obtained. Generally, FS methods can be classified into three types, filter, wrapper, and embedded methods. A huge number of literatures regarding gene selection methods in the designing of an efficient classification model has been published. Often, samples of tumor for microarray data come from well-defined groups, from normal and diseased prognostic patients. The model to diagnose patients to the normal or diseased prognostic class. Because of the microarray data and of its specific tumor, in this paper, we demonstrate a review of the feature selection method for gene selection from the microarray dataset. Our intention in the survey is to grasp the awareness of the advantage of the different FS methods to the researchers in microarray data analysis.

2. Feature Selection and Classification

With increasing complexities of problems in every domain, huge amounts of data are generated day by day. These data

[1]*Dr. S. Begum is with the govt. College of Engg. and Textile Technology, Berhampore, MSD. She is with the Department of CSE.*
[2]*Ebne Saud Khan, Ph. D Research Scholar, Ram Krishana Dharmarth Foundation University, Ranchi*
[3]*Dr. Debasis Chakraborty is with the Asansol Engg. college, Asansol, W.B., India. He is with the Department of CSE.*
Correspondence: shemim_begum@yahoo.com

include imminent information, and hence, more effective models need to be evolved for the abstraction of hidden knowledge. This deluge of data solicits automated techniques of data analysis. Machine learning methods automatically detect features in data, and then use the features to predict the unlabeled data or to carry out other kinds of determination under uncertainty. In machine learning, FS is an essential part of the analysis of large amounts of data in various fields.

A. Feature Selection

Feature selection is the technique of finding a set of correlated features that are advantageous in designing a model. Not only are datasets ever increasing, but new datatypes such as datasets on the web, microarrays, proteomics, genomics, and system biology are becoming conventional, since many pattern classification methods were originally not evolved to manage huge amounts of irrelevant features, combining them with FS methods is essential in many applications [14-6]. These inconsistent features need to be irradicated before any machine-learning technique is applied. Thus, finding optimal numbers of features includes an additional layer of complexity in the modeling. Instead of recognizing optimal features from the entire dataset, first, the best possible subset is to be obtained and parallelly the model parameter is to be upgraded. During the process of FS, the training samples are usually labeled, unlabeled, or partly labeled, leading to the evolution of supervised, unsupervised, and semi-supervised FS methods. In supervised FS, all the instances are labeled, and the significance of a feature is measured by its correlation with the class label [17, 18]. On the other hand, an unsupervised FS algorithm exhibits data variance in its analysis of the relevance of the attributes [19], [20]. Whereas the semi-supervised feature relevance benchmark takes into consideration both the labeled and unlabeled samples [21, 22] to achieve an efficient feature subset selection. Based on the time and manner in which the efficiency of selected features is evaluated, different methods are elaborated, which are broadly classified into three classes: filter, wrapper, and embedded models. Table 1 depicts the topology of FS methods, exhibiting the noticeable merits and demerits of each method.

Moreover, some conventional FS algorithm of the filter method depends on the analysis of the relevance of features by perceiving the intrinsic properties of the data without considering any learning algorithm. The extracted features are presented as input data to the classification algorithm. Filter FS algorithms are computationally fast and simple and can easily scale to higher dimensional datasets. As the filter approaches are self-reliant i.e., not associated with any mining algorithm, features need to be evaluated only once, then those features are useful for computational purposes. The shortcoming of the filter method is that they do not act jointly with the classifier. Mostly, they are univariate in nature, which leads to ignoring feature reliance, and this gives rise to poor classification performance. This problem can be alleviated by incorporating multivariate filter techniques to give feature dependencies to some extent. Whereas the wrapper FS algorithm requires a predetermined learning algorithm and observes its performance on the extracted features to achieve relevant features. The wrapper approach comparatively slower than the filter approach, on account of the data mining algorithm, is applied to each feature subset. Moreover, if different mining algorithms have experimented on the data, the wrapper FS method becomes more computationally expensive [23]. Two types of search methods are normally considered: randomized and deterministic search algorithms. Wrapper FS methods are assisted by collaboration between feature subset search, feature reliance, and selecting the model. However, they are highly prone to overfit together with high computational burden.

Lastly, in the embedded model, the feature subsets are selected into the classifier during its formation and can be considered as a search that is explained by the space of feature subsets. The embedded method is fruitful as it interacts with the classification method and is less computationally intensive in comparison to wrapper methods [24]. Lastly, integrated methods are preferable to achieve a better subset of features.

TABLE.1: CATEGORIZATION OF FEATURE SELECTION METHODS. THE MERITS AND DEMERITS OF THE MODELS WILL HELP THE RESEARCHERS IN THIS DOMAIN TO CHOOSE A CERTAIN METHOD SUITABLE FOR THE UNDERLYING PROBLEM.

Models	Merits	Demerits
Filter models	Univariate	
	Scalable to high-dimensional datasets Computationally fast Independent of the learning algorithm Model free	Dependence among features are ignored Avoids interaction with the classifier Poor classification performance
	Multivariate	
	Feature dependencies Independent of the classifier Better computational complexity than wrapper methods Model free	Slower than univariate techniques Less scalable than univariate techniques Avoids interaction with the classifier
Wrapper models	Deterministic	
	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection
	Randomized	
	Less prone to local optima Interacts with the classifier Models feature dependencies Better classification performance	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms
Embedded models	Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection

B. Classification

Classification maps a sample into one of several built-in classes. [25], [26]. The task of pattern classification leads to pattern recognition, classification, description, and binding of patterns that have analogous natures in divergent engineering and technological disciplines. An unknown pattern class can be achieved by using any one of the machine learning techniques: 1) Supervised Learning; 2)

Unsupervised Learning; 3) Semi-Supervised Learning; and ensemble Learning. Unsupervised learning or clustering analyzes a set of gene expression profiles in search of finding the subset of genes, such that genes in the same cluster are equivalent in the same sense, and genes in the different cluster are unlike in the same sense. Supervised learning involves mapping between a set of input variables A, and the output variable, B, and applying the mapping to predict the unseen data. Supervised learning, which is the most important machine learning algorithm, is mostly used in many real-life problems. In the recent past, there has been a growing importance in the application of unlabeled data along with labeled data in machine learning [27]. The inspiration is very clear: in many areas, unlabeled data can be more economical and abundant in comparison to labeled data.

If relevant information can be captured from unlabeled data, learning from labeled samples can be substantial assistance in machine learning. Many semi-supervised learning algorithms have been developed for better performance, with experimental output acquired from different learning paradigms. The semi-supervised learning algorithm includes finding labels for word-sense disambiguation [28], co-training to distinguish webpages [29] and to enhance visual detectors [30], transudative support vector machine [7], in diagnosing cancer, Emerging Method [31] in classification text, graph-based methods [32], and many other domains. Furthermore, soft computing is very useful in pattern recognition.

3. Examination with Gene Microarray Expression

All living organisms are composed of trillions of cells each consisting of an entire copy of the genome. A genome is the complete set of genes present in a cell, and it consists of a sequence of DNA. The nucleus genome has protein-coding genes and non-coding genes. Normally, the genome is made of 23 pairs of chromosomes. The genome comprises all the information necessary for an individual to grow and function. The human genome project depicts that there are probably 20,000-25,000 protein-coding genes. By the process of gene expression, information that is enabled in a gene is converted into an observable phenotype and thereafter synthesized protein, which forms the structure of the cell. The gene expression is analogous to two main phases: transcription and translation. In the transcription process, segments of DNA are substituted into RNA (mRNA) by RNA polymerase. While the translation process uses mRNA for protein synthesis. The process of gene expression is carefully regulated, altering substantially under different situations and cell types. The RNA and protein of genes help to regulate other gene expression. The amount of gene expression can also be measured by observing a phenotype in connection with a gene.

A DNA microarray is a glass slide, onto which DNA molecules are chemically bonded at a special location in the array. Now, the glass slide is placed under a scanner, and an image with color is obtained. Each dot represents the expression level of a gene under experimental conditions. Each array location acts as a probe and holds many similar copies of the same molecule. Each probe specifies the measurement for a single gene, and an array is the measure for multiple genes.

Fig. 1. depicts a gene expression matrix and coding of microarray data. A microarray dataset is a 2D matrix $P=m_{ij}$, which comprises samples and biomolecules. Each element is the record of the expression level of the j^{th} microarray array for the i^{th} sample. Gene expression profiling has empowered the measurement of thousands of genes in a single RNA sample [33].

This method is used for the diagnosis and prognosis of cancer. Effectual microarray experiments need specific planning goals [34]. The intention of microarray studies is to achieve biologically meaningful insights from the microarray data and make use of this gained knowledge in a significant way. The essential steps to gain significant information from microarray data are skillful selection, classification, and class determination. Gene selection is the method of extracting a small subset of providential genes that are the most prognostic following the class label. This helps the learning model to maximize the classification performance. The extracted information can be considered as the pattern in the microarray data. Pattern analysis targets to find out relationships in data. In machine learning data is supposed to be in a vectorized form, and the relations are presented as classification rules, cluster structures, and regression functions. Microarrays are being used to identify worthy genes, types of gene activity, sites of transcription factors, modification in DNA copy number, genes that are affected in connection to treatment, time series (with and without providing treatment), classification of tumors, finding target genes, recognition of cancer biomarkers, providing antibiotic treatment, heart failure of human beings, and SNP linkage studies. Microarray analysis yields a huge number of facts on disease pathology. Progression, reaction to cellular -microenvironments, and finally give rise to early diagnosis and innovative therapy for diseased cells. However, there are many orientations of microarray studies, we limit our reviews on cancer studies, and we limit our review on cancer studies using FS techniques.

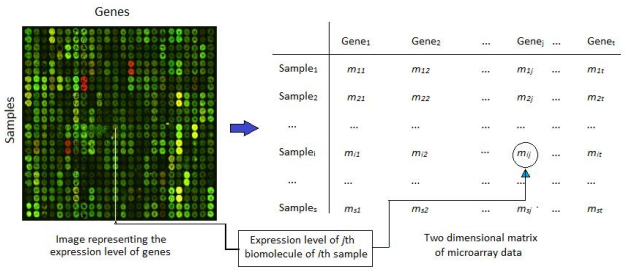

Fig. 1. Coding of the generated colored image to microarray data.

4. Gene Selection

Feature selection provides many advantages. In cancer diagnosis, it is simpler to study the expression level of very few genes than thousands of genes. Moreover, gene selection decreases the dimension of microarray data so that contributing to the reduction in computational hazards. In addition, FS yields a concise gene subset [35]. The FS problem can be considered an optimization problem in which the goal is to achieve a subset of features for which feature-subset selection metrics are optimized. The forward selection–backward elimination algorithms are used to select or remove features to improve the efficiency of the model. The three measures in the FS process are search, evaluation, and stop. The large dimensionality in microarray data renders a great challenge in computational methods. Over the past few years, many gene selection methods have been presented to determine significant genes for cancer identification and diagnosis [36-41]. Zhou et al. [42] used the Bayesian approach for choosing the useful genes from microarray data and the logistic regression model has been used for classifying the data. A comprehensive review of the feature selection methods has been depicted by saeys et al. [24]. For a wide review of FS techniques of microarray datasets, the perusal is required to follow [43]. The following module review different FS method for gene selection from gene expression data. Few of the heuristics for the selection of useful genes incorporate threshold on the detected fold-change difference in gene expression between the states, and the identification of the threshold point in each gene that reduces the training sample misclassification (Ben-Dor et al.,2000).

A. Filter-Based Methods

Filter-based approaches are fast and very efficient and have comparative assessment among different FS methods with microarray datasets [44, 45]. The Signal to Noise Ratio (SNR) method was first applied by Golub et al. in [46] to rank the genes for selection. The features are assessed independently instead of as a subset. The classification is performed on ALL with the application of the Weighted Voting (WV) method, which created two errors on the test dataset. Total Principal Component Regression (TPCR) was applied on ALL datasets and produced one classification error on 38 training samples and 34 test samples. Ben-Dor et al. [37] used heuristics on the fold-change difference in gene expression among the states. SVM was explored by Ramaswamy et al. to classify leukemia data [47]. The feature was extracted by the Signal to Noise Ratio (SNR) method. Shannon et al. applied mantel statistics to grade genes in [48]. Mantel statistics evaluate the correlation between two distance measures on the same data. This method avoids the application of a two-stage analysis of brain tumor tissue samples that were present in [49]. For the comprehensive study of hierarchical clustering, the reader is requested to follow W. Shannon et al. [50]. In [51] [52] the authors applied Emerging Patterns (EP) to identify genes from Leukemia [46] and Colon [53] datasets. EP takes out significant differences between classes. The model demonstrates that the extracted pattern can be used to classify the cancer data with higher efficiency.

The filter method, gene pair ranking was suggested in [54] to measure pairs of genes to discriminate two classes perfectly. Here ranking was done by applying two sample t-statistics on the gene expression value. The presented method outperformed the other methods in connection with Cross-Validation (CV) accuracy on cancer datasets. The ANOVA and two-sample t-test are the most widely used approaches in microarray studies. For the observational study of the different statistical methods such as t-statistics, mean difference, Significance Analysis of Microarray (SAM) with fudge factor, SNR, Kolmogorov-Smirnov statistics, Wilcoxon rank sum, p-value with maximal logistic regression for gene selection from four cancerous data, the reader is asked for the research paper [55]. Hudge considered two algorithms in [56] to extract significant genes from the leukemia dataset. Substantially, Diagonal Quadratic Discriminant Analysis (DQDA), was utilized to distinguish healthy cells from diseased cells. Leung et al. [57] employed many filters and wrappers FS methods for extracting significant genes from microarray data.

In [58], the author sorted out pertinent genes before classification. They considered the proportion of the sum of the squares in between classes to the sum of squares within the class for every gene and identified the genes with the maximum ratio [44]. Then classification was carried out on **Small Round Blue Cell Tumors** (SRBCT) and leukemia datasets with multi-class SVM. Experimental results exhibited that the method could classify cancer with predictive gene subsets. Meanwhile, many developed computational methods have evolved for microarray experiments. Mutual Information (MI) was applied in [59] to find significant genes for the diagnosis of cancer. A clustering algorithm was first deployed to irradicate redundant genes from gene expression datasets. Afterward, MI was employed on the remaining subsets. Their method minimizes the computational cost.

The decision tree-based FS method for evaluating attribute weight performs well on different gene expression datasets

[60]. Here, Naïve Bayes (NB) classifier has been used to find out the efficiency of the model with its simplicity. In [61], Bayes error was utilized in finding the optimum gene for the classification problem. Bayes error, which is a filter approach rests on the gene space only. The Wilcoxon test was preferred to evaluate the efficiency of each gene. The efficiency of this model is exhibited on five microarray datasets. Wang et al. in [62] applied class separability measure [44] and t-score [63], to ordered genes for four cancer datasets. These extracted gene subsets were then estimated with SVM and Fuzzy Neural Network (FNN) [64]. Wong and HSU [65] applied four filter methods to extract the optimum gene initially. Later, two classifiers were used to evaluate the efficiency of the FS methods on eight microarray datasets. However, the results are not so optimistic. Maj and pal [66] suggested fuzzy-rough sets to predict information on fuzzy approximation spaces. Promising genes are determined by increasing the relevance and decreasing the redundancy of the genes. The efficiency of the model, along with comparative results are presented on three microarray cancer datasets and two arthritis datasets. The dependent degree method, which comprises rough set-based FS is depicted in [67]. The proposed method exhibits its proficiency in comparison to other FS methods, such as chi-square, Relief, information gain, and Symmetrical Uncertainty (SU) in terms of average and best classification accuracy. Effective Range-Based Gene Selection (ERGS), which incorporates class discriminating power, was depicted in [68]. The experiment was conducted on colon, leukemia, lung, prostate cancer, and Diffuse large B-cell lymphoma (DLBCL) datasets. Fuzzy rough set-based feature selection was proposed in [69] on account of the approximation of the gain ratio. Mutual information between genes and classes is evaluated to extract informative genes for DNA microarray analysis. The performance was measured on two microarray datasets. In [70], Cilia et al. proposed five filter methods namely, chi-square, Relief, gain ratio, information gain, and symmetrical uncertainty. The effectiveness of the proposed approaches is measured using four classification methods namely, decision -tree, random forest, K nearest neighbor, and multilayer perceptron. The experiment used six datasets of breast cancer, colon, leukemia, lymphoma, lung, and ovarian. It suggests that random forest gives the best result without the FS method. Lee et al. propose Markov Blanket (MB) feature ranking method for classifying six microarray datasets. The new MB-based feature ranking method surpasses other univariate feature ranking methods and other multivariate feature selection methods [71].

B. Wrapper-Based Approaches

Wrapper models have been effectually used for significant gene selection. We have studied several research papers on this FS method. In the research paper [72], the authors suggested the perturbation method, which incorporates a very less amount of noise to the gene expression data, and re-clusters the data. Now, it differentiates the results within the real clustering. They disclosed a subset of melanomas, which are accountable for human skin cancer.

A magnificent gene-selection method, gene shaving [73] was reported to recognize gene subsets with huge variations within conditions and coherent gene expression patterns. This method deviates from traditional hierarchical clustering methods as the gene belongs to more than one cluster. Li et al. [74] applied Genetic Algorithm (GA) for the gene selection of colon and lymphoma datasets. The Principal Component Analysis (PCA) was depicted in [75] to achieve a major variation in gene expression data. PCA reduces the dimensionality of the dataset by transforming it into a new set of attributes. There is no clear correlation between the number of principal components and the quality of the cluster. The Partial Least Square (PLS) method determines genes by sequentially expanding gene signatures and the covariance between the vector of different class labels [76]. It is expected that PLS subsides PCA in terms of efficiency, because PLS utilizes the information on class labels, while PCA neglects it. Wrapper approaches were presented [77] for the subset of gene selection. The authors evolved supervised learning algorithms, which were blended with the FS algorithm for informative gene selection. The method eradicates insignificant genes. Evolutionary algorithms were introduced in [78] to find out the optimal set of significant genes from NC160 and leukemia datasets.

A magnificent FS was depicted in [79] to find biomarkers with the help of Gaussian processes, which efficiently measure the uncertainty in data because of a powerful probabilistic framework. The covariance function blended with the Automatic Relevance Determination (ARU) attribute depicts the correlation between samples to track the assistance from respective features. The novelty of the suggested method was experienced in the prostate cancer dataset. In [80], the author depicted a wrapper-oriented heuristic approach for gene subset selection. This method includes a gene on account of statistical significance and finds out an optimal subset of genes. This method is efficient in terms of computational cost also.

In [81], the author presented a files-based Multiple Kernel SVM (M-K-SVM) FS method for gene subset selection. This method is applied to the colon and ALL-AML datasets to measure the performance of the model. In [82], Pashaei et al. proposes a Chomp optimization Algorithm (ChoA) for biomedical data classification. Two binary variants of ChoA are introduced for the FS method. Initially S-shaped, V-shaped transition functions are applied to convert continuous ChoA to binary. Secondly, the crossover operator is used to enhance Chaos's exploration. The proposed approach is experienced with five microarray datasets and compared with six wrapper-based FS methods namely, Particle Swarm Optimization (PSO), Genetic Algorithm (GA), Ant Colony

Optimization (ACO), Batch Algorithm (BA), Firefly Algorithm (FA) and flower pollination. In [83], Balakrishnan depicted Salp Swarm Algorithm (SSA), which is a population-based optimization algorithm. The existing SSA failed to converge the initial random solution to the global optimization. Hence, the depicted improved SSA (iSSA), which succeeds in increasing the efficiency of salps to flourish in divergent areas by updating its location. The performance of iSSA is measured over six different microarray datasets. In terms of convergence, the proposed model surpasses the SSA by exploring .1033% more confident with the selected features.

C. Embedded Technique

Embedded approaches are less computationally intensive in comparison to wrapper methods. The extensive review of the application of the embedded method is as follows: Guyan et al. [84] depicted Support Vector Machine Recursive Feature Elimination (SVM-RFE) as a classifier and extracted genes by their weight in the SVM classifier. SVM-RFE was also applied in [85] to develop gene identification. A few authors further increased the recursive SVM approach by suggesting different evaluation criteria and optimization methods. Duan Raaja Pakse suggested a new FS [86] method using a backward elimination method to extract promising genes. The proposed approach evaluates each feature's score with the help of statistical analysis of the weight vectors of Linear SVM, which is trained on a sub-sample of the original train dataset. Recursive feature elimination (RFE) was proposed for iterative gene selection, which finds the genes that are pertinent to cancer. The method was exhibited on seven cancer datasets, including breast cancer, colon, central nervous system, ALL, lung, and ovarian cancer. Zhang et al. suggested a new estimation criterion for significant gene selection for breast cancer and spectrometry datasets [87]. The suggested recursive SVM is considered a voting scheme for biomarker selection. The limitation of the filter method is that they do not relate to the classifier. Most of the FS algorithms are univariate, which select the genes that have the strongest correlation. This accelerates poor classification performance, which can be alleviated by introducing a multivariate filter method to measure feature dependencies to some extent. Whereas the FS algorithm of the wrapper model needs a prearranged learning algorithm and evaluate its performance to pick out the significant gene for multi-class cancer classification. Their review depicts the suggested Fuzzy Support Vector Machine-Recursive Feature Elimination (FSVM-RFE) and finds out the effective genes, in comparison to other gene selection techniques, such as F-test and SVM-RFE. SVM-RFE is further developed in [88] by combining a **minimum Redundancy Maximum Relevance** (mRMR) filter. Mutual information between genes and classes is used to find the significant gene set, whereas mutual information among the genes renders redundancy. The method chooses a smaller number of genes in comparison to SVM-RFE and MRMR. Curious researchers are appealed to go through other approaches of RFE techniques [89

, 90] for cancer research. A Bayesian generalization of SVM was introduced in [91] to figure out optimal feature subsets by optimizing a single posterior objective function. The method picks up a set of significant genes that are applied for tumor identification. The literature in [92] presented a non-convex penalty for the hinge loss function in SVM to eradicate redundant genes to evolve a compact classifier. Shah et al. [93] illustrated the postulate of the learning of conjunction and disjunction of decision stumps in sample compression, Occam's Razor, and the PAC-Bayes learning model to achieve a small gene subset. Shafi et al. propose an embedded approach [94] by combining the "Mean Decrease Gini" and "Mean Decrease Accuracy" FS algorithm and random forest as a classifier to increase the efficiency of the model's accuracy. The approach is experienced in colon cancer datasets. Ding et al. propose the Optimized Extreme Learning Machine-Based Recursive Feature Elimination (OPELM -RFE) model as an embedded method. It also explores the alpha seeding algorithm to solve successive quadratic programming-problem in OELM. By fine-tuning the parameter penalty cost C, the model exhibits its proficiency more than SVM-RFE. Moreover, it requires fewer model selection parameters than SVM-RFE. Hence, the novel embedded RFE model based on the OELM classifier, suggests its proficiency more than other embedded methods [95].

D. Hybrid Feature Selection Technique

Researchers are highly motivated to hybrid methods, in which two or more FS methods identify significant feature subsets. In [96], the authors integrated the mRMR with other wrapper feature selection methods. The proposed method is differentiated from other methods. The result depicted that mRMR explores more promising results than other methods. mRMR FS method has emerged in a few articles [97-99] for prominent gene subset selection in cancer research. The hybrid method naming Fuzzy granular support vector machine-recursive feature elimination algorithm (F-GSVM -RFE) was depicted in [100] to take off unnecessary or noisy genes from the dataset. The empirical studies show that FGSVM-RFE surpasses other existing methods. The network constrained SVM (net SVM) was proposed in [101] to find promising genes from the datasets. The method was experimented on the breast cancer dataset. A hybrid FS Ensemble correlation-based gene selection algorithm (ECBGS), which is built on SU, was engineered by Piao et al. [102]. The hybrid method finds proficient genes from microarray data efficiently. The efficiency of the proposed method was manifested contrary to other FS methods. Genetic Algorithms (GA) and SVM are blended in [103] to find promising genes. The proposed method picks our promising genes through an iterative process, which renders

good classification accuracy in comparison to other methods [104]. A Multi-objective GA and SVM was blended as wrapper method to encode significant genes from the microRNA dataset. This model provides promising results and can be experienced in other domains also. Rohit et al. used the Altruistic Whale Optimization (AWO) algorithm for feature selection on microarray data [105]. First, it applies Pasi Lukka's filter-based feature ranking algorithm, which selects the top 300 genes from the dataset. Then applied the AltWOA algorithm to obtain a reduced feature set. The algorithm was applied to eight microarray datasets to show the efficiency of the proposed approach. In [106] Esraa Alhenawi suggested a hybrid method for microarray data processing. Here is an ensemble filter method blended with Intelligent Water Drop (IWD) algorithm by adding a local Search (LS) algorithm: Novel LS algorithm, Hill Climbing, or Tabu Search (TS) algorithm. Naïve Bayes classifier and five microarray datasets have been employed to measure the efficiency of the model. The proposed method improves classification accuracy with an average of 8.92% in three datasets out of five datasets and reduces the number of genes by 58.5% in all five datasets.

5. Conclusion

In the research paper, we analyzed the enormous contribution of the identification of gene markers and cancer prediction in biometric applications. In the field of biometrics, two main complaints are found by the researchers, very few labeled samples and large dimensions of the dataset. To address these problems, plenty of feature selection methods have been developed by the research community to extract the promising gene. Greater dedication has been offered in the recent few years for the development of various univariate filter-based FS methods. By the way, the Multivariate FS method can be the future orientation of research for bioinformatics researchers. Other, encouraging paths of future research are the evolution of ensemble, semi-supervised, and integrated approaches for feature extraction to obtain the soundness of selected features. In order to get rid of the small sample size of the microarray dataset in the field of bioinformatics, the model can be further developed by integrating with accurate evaluation criteria, which comprise an exciting orientation for the future.

References

[1] U. Maulik, A. Mukhopadhyay, and S. Bandyopadhyay, "Combining Pareto-optimal clusters using supervised learning for identifying co-expressed genes," *BMC Bioinformatics*, vol. 10, no. 27, 2009.

[2] U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, "Multi-class clustering of cancer subtypes through SVM based ensemble of Pareto-optimal solutions for gene marker identification," *PLoS One*, vol. 5, no. 11, p. e13803, 2010.

[3] Wu, "Differential gene expression detection and sample classification using penalized linear regression models," *Bioinformatics*, vol. 22, pp. 472-476, 2006.

[4] U. Maulik and A. Mukhopadhyay, "Simulated annealing based automatic fuzzy clustering combined with ANN classification for analyzing microarray data,"' *Computers & Operations Research*, vol. 37, no. 8, pp. 1369-1380, 2010.

[5] S. Mallik, A. Mukhopadhyay, and U. Maulik, "Integrated statistical and rule-mining techniques for DNA methylation and gene expression data analysis," *Journal of Artificial Intelligence and Soft Computing Research*, 2013.

[6] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, "Gene-expression based cancer subtypes prediction through feature selection and transductive SVM," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1111–1117, 2013.

[7] U. Maulik and D. Chakraborty, "Fuzzy preference-based feature selection and semi-supervised SVM for cancer classification," *IEEE Transactions on Nano Bioscience*, vol. 13, no. 2, pp. 152–160, 2014.

[8] Chakraborty and U. Maulik, "Identifying cancer biomarkers from microarray data using feature selection and semi-supervised learning," *IEEE Journal of Translational Engineering in Health and Medicine*, vol 2, 2014.

[9] G. Piatetsky-Shapiro and P. Tamayo, "Microarray data mining: facing the challenges," *SIGKDD Explorations Newsletter*, vol. 5, pp. 1–5, December 2003.

[10] M.Rocha, R.Mendas, P.Maria, et al., "A platform for the selection of genes in DNA Microarray data using evolutionary algorithms," *in Proceedings of 8th Annual Conf. on Genetic and Evolutionary Computation*, London, England, 2007, pp 415–423.

[11] Q. Shen and C. Shang, "Aiding classification of gene expression data with feature selection: A comparative study," *Journal of Computational Intelligence Research*, vol. 1, pp. 68–76, 2006.

[12] J. C. Rajapakse and P. A. Mundra, "Multiclass gene selection using Pareto-fronts," *IEEE Trans. Comput. Biol. Bioinformatics.*, vol.10, no. 1, pp. 87–97, Jan./Feb. 2013.

[13] Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[14] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*, Boston: Kluwer Academic Publishers, 1998.

[15] H. Liu and H. Motoda, (Eds.), *Computational Methods of Feature Selection*, Chapman, and Hall/CRC Press, 2007.

[16] M. R. Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[17] J. Weston, A. Elisseeff, B. Schoelkopf, et al., "Use of the zero norm with linear models and kernel methods," *Journal of Machine Learning Research,* vol. 3, pp. 1439–1461, 2003.

[18] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, vol. 5, pp. 845–889, 2004.

[19] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *Advances in Neural Information Processing Systems 18*, Cambridge, MA, 2005. MIT Press.

[20] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," *In Proceedings of the 24th International Conf. on Machine Learning (ICML),* pp. 1151–1157, 2007.

[21] Z. Xu, R. Jin, J. Ye, et al., "Discriminative semi-supervised feature selection via manifold regularization," *In IJCAI-09: Proceedings of the 21st International Joint Conf. on Artificial Intelligence*, 2009.

[22] M. L. Raymer, W. F. Punch, E. D. Goodman, et al., "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, Vol. 4, No. 2, pp. 164–171, 2000.

[23] Y. Saeys, I. Inza and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol.23, no. 19, pp. 2507– 2517, 2007.

[24] J. Hand, *Discrimination and Classification*, Chichester, U.K.: Wiley, 1981.

[25] S. M. Weiss and C. A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, San Francisco, Calif.: Morgan Kaufmann, 1991.

[26] O. Chapelle, B. Scholkopf, and A. Zien, (Eds.), Semi-Supervised Learning, MIT Press, Cambridge, MA, USA, 2006.

[27] Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," *In Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, pp. 189–196, 1995.

[28] Blum and T. M. Mitchell, "Combining labeled and unlabeled data with co-training," *COLT: In Proceedings of the Eleventh Annual Conf. on Computational Learning Theory*, pp. 92–100, 1998.

[29] Levin, P. Viola, and Y. Freund, "Unsupervised improvement of visual detectors using co-training," *In Proceedings of the Ninth IEEE International Conf. on Computer Vision*, pp. 626–633, 2003.

[30] K. Nigam, A. K. McCallum, S. Thrun, et al., "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2-3, pp. 103-134, 2000.

[31] X. Zhu, Z. Ghahramani and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," *In the 20th International Conference on Machine Learning (ICML),* 2003.

[32] Abudayur and O.U. Nalbantoglu, "A survey of feature selection strategies for DNA microarray classification", Computer Engineering and Intelligent Systems, vol. 14, No. 2, 2023.

[33] Tjaden and J. Cohen, "A survey of computational methods used in microarray data interpretation," *Applied Mycology and Biotechnology*, *Bioinformatics*, vol. 6, pp. 7–18, 2006.

[34] H. Yu and S. Xu, "Simple rule-based ensemble classifiers for cancer DNA microarray data classification," *International Conference on Computer Science and Service System (CSSS)*, pp. 2555–2558, 2011.

[35] 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," *Proc Natl Acad Sci USA*, vol.96, pp. 6745–6750, 1999.

[36] Ben-Dor, L. Bruhn, N. Friedman, et al., "Tissue classification with gene expression proles," *J Computational Biol*, vol. 7, pp. 559–584,2000.

[37] Y. Lu and J. Han "Cancer classification using gene expression data," *Inform Syst* vol. 28, no. 4, pp. 243–268, 2003.

[38] Ooiand P. Tan "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," *Bioinformatics*, vol. 19, pp.37–44, 2003.

[39] R. Fox and M. Dimmic, "A two-sample Bayesian *t*-test for microarray data," *BMC Bioinformatics*, vol. 7, pp. 126, 2006.

[40] Q. Shen, W. M. Shi, and W. Kong, "New gene selection method for multiclass tumor classification by class centroid," *J Biomed Inform* vol. vol. 42, pp. 59–65, 2009.

[41] X. Zhou, K. Y. liu and S. T. C.Wong, "Cancer classification and prediction using logistic regression with Bayesian gene selection," *J Biomed Inform*, vol. 37, pp. 249–259, 2004.

[42] Boln-Canedo, N. Sanchez-Maroo, A. Alonso-Betanzos, et al., "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111–135, 2014.

[43] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *J. Am. Statistical Assoc.*, vol. 97, pp. 77-87, 2002.

[44] J. Lee, J. B. Lee, M. Park, et al. "An extensive comparison of recent classification tools applied to microarray data," *Computational. Stat. and Data Anal.*, vol. 48, pp. 869885, 2005.

[45] 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, pp. 531–537, 1999.

[46] S. Ramaswamy, P. Tamayo, R. Rifkin, et al., "Multiclass cancer diagnosis using tumor gene expression signatures," Proceedings *of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[47] W. D. Shannon, M. A. Watson, A. Perry, et al., "Mantel statistics to correlate gene expression levels from microarrays with clinical covariates," *Genetic Epidemiology* vol. 23, no. 1, pp. 87–96, 2002.

[48] M. Watson, A. Perry, V. Budhjara, et al., "Gene expression profiling with oligonucleotide microarrays distinguish WHO grade of oligodendrogliomas," *Cancer Research*, vol. 61, no. 5, pp. 1825–1829, 2001.

[49] W. Shannon, R. Culverhouse, and J. Duncan, "Analyzing microarray data using cluster analysis," *Pharmacogenomics*, vol. 4, no. 1, pp. 41–52, 2003.

[50] J. Li and L. Wong, "Identifying good diagnostic genes groups from gene expression profiles using the concept of emerging patterns, *Bioinformatics*, vol. 18, no. 5, pp. 725–734, 2002.

[51] Dong and J. Li, "Efficient mining of emerging patterns: discovering trends and differences," *Proceedings of the fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, ACM Press, San Diego, CA, pp. 43–52, 1999.

[52] 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," *Proc. Natl Acad. Sci., USA*, pp. 6745–6750, 1999.

[53] T. B and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biology*, vol. 3, no. 4, pp. research0017.1-research0017.11, 2002.

[54] S. Ma, "Empirical study of supervised gene screening," *BMC Bioinformatics*, vol. 7, no. 537, 2006.

[55] Huang, "An integrated method for cancer classification and rule extraction from microarray data," J Biomed Sci., vol.16, no.1, 2009.

[56] Y. Leung and Y. Hung, "A multi-filter-multi-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vo1. 7, no .1, pp.108-117,2010.

[57] Y. Lee and C. K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, no. 9, 2003.

[58] Huang, T. W. S. Chow, E. W. M. Ma, et al., "Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis," *IEEE Transactions on Circuits and Systems*, vol. 52, no. 9, pp. 1909–1918, 2005.

[59] M. Hall, "A decision tree-based attribute weighting filter for naïve Bayes," Knowledge-Based Systems, Vol. 20, no. 2, pp. 120–126, 2007.

[60] J. G. Zhang and H. W. Deng, "Gene selection for classification of microarray data based on the Bayes error," *BMC Bioinformatics*, vol. 8, no. 370, 2007.

[61] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40–53, 2007.

[62] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data*, third ed. Duxbury Press, 1997.

[63] Y. Fray man and L. Wang, "Data mining using dynamically constructed fuzzy neural networks," *Lecture Notes in Artificial Intelligence*, vol. 1394, pp. 122-131, 1998.

[64] T. T. Wong and C. H. Hsu, "Two-stage classification methods for microarray data," *Expert Systems with Applications*, vol. 34, no. 1, pp.375-383, 2008.

[65] P. Maji and S. K. Pal, "Fuzzy Rough sets for information measures and selection of relevant genes from microarray data," *IEEE Transactions on systems, man, and cybernetics part b: cybernetics*, vol. 40, no. 3, pp. 741–752, 2010.

[66] X. Wang and O. Gotoh, "A Robust Gene selection Method for microarray-based cancer classification," *Cancer Informatics*, vol. 9, pp. 15–30, 2010.

[67] Chandra and Manish Gupta, "An efficient statistical feature selection approach for classification of gene expression data, *"Journal of Biomedical Informatics*, vol. 44, no. 4, pp. 529–535, 2011.

[68] J. Dai and Q Ju, "Attribute selection based on information ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, pp. 211–221, 2013.

[69] N. D. Cilia, De Stefano, C., F. Fontanella, S. Raimondo, & A. cotto di Freca, A., "An experimental comparison of feature-selection and classification methods for microarray datasets", *Information*, vol. 10, no. 3, pp. 109, 2019.

[70] J. Lee, I. Choi, & C. H Jun, "An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data", *Expert Systems with Applications*, 113971, 2020.

[71] M. Bittner, P. Meltzer, Y. Chen, et al., "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 6795, pp. 536–540, 2000.

[72] T. Hastie, R. Tibshirani, M. B. Eisen, et al., "Gene shaving as a method for identifying distinct sets of genes with similar expression patterns," Biology, vol. 1, no. 2, pp. research0003.10003.21, 2000.

[73] L. Li, C. R. Weinberg, T. A. Darden, et al., "Gene selection for sample classification based on gene expression data: study of sensitivity to choose of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

[74] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no.9, pp. 763- 774, 2001.

[75] K. E. Basford, G. J. McLachlan, and S. I. Rathnayake, "On the classification of microarray gene-expression data," *Briefings in Bioinformatics*, vol. 14, no. 4, pp. 402–410, 2012.

[76] R. Blanco, P. Larranaga, I. Inza, et al., "Gene selection for cancer classification using wrapper approaches," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 18, no. 8, pp. 1373–1390, 2004.

[77] U.T. Jirapech and S. Aitken, "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes," *BMC Bioinformatics*, vol. 6, 148, 2005.

[78] Z. Chu, Z. Ghahramani, F. Falciani, et al., "Biomarker discovery in microarray gene expression data with Gaussian processes," *Bioinformatics*, vol. 21. no. 16, pp. 3385–3393, 2005.

[79] R. Ruiz, Jos C. Riquelme, and Jess S. Aguilar-Ruizb, "Incremental wrapper-based gene selection from microarray data for cancer classification," *Pattern Recognition*, vol. 39, no. 12, pp. 2383–2392, 2006.

[80] Z. Chen, J. Li, and L. Wei, "A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue," *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 161–175, 2007.

[81] Pashaei, E., & Pashaei, E., "An efficient binary chimp optimization algorithm for feature selection in biomedical data classification", *Neural Computing and Applications*, vol. 34, no.8, pp. 6427–6451, 2022.

[82] Balakrishnan, K., Dhanalakshmi, R., & Khaire, U. M., "Improved salp swarm algorithm based on the levy flight for feature selection", *The Journal of Supercomputing*, vol. 7, no. 11, pp. 12399–12419,2021.

[83] Guyon, J. Weston, S. Barnhill, et al., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.

[84] Marchiori and M. Sebag, "Bayesian learning with Local support vector machines for cancer classification with gene expression data," *Applications of Evolutionary Computing*, Lecture Notes in Computer Science, vol. 3449, pp. 74-83, 2005.

[85] Kai-Bo Duan, and J. C. Rajapakse, "Multiple SVM-RFE for gene selection in cancer classification with expression data, *IEEE Transactions on Nano bioscience*, vol. 4, no. 3, pp. 228–234, 2005.

[86] X. Zhang, X. Lu, Q. Shi, et al. "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data," *BMC Bioinformatics*, vol. 7, no. 197, 2006.

[87] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Transactions on Nano bioscience*, vol. 9, no.1, pp. 31–37, 2010. [144] S. Niijima and S. Kuhara, "Recursive gene selection based on maximum margin criterion: a comparison with

[88] SVM-RFE," *BMC Bioinformatics*, vol. 7, no. 543, 2006.

[89] S. Niijiima and S. Kuhara, "Recursive gene selection based on maximum margin criterion: A comparison with SVM-RFE", *BMC Bioinformatics*, vol. 7, no. 543, 2006.

[90] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[91] B. Krishnapuram, L. Carin and A. Hartemink, "Joint classifier and feature optimization for comprehensive cancer diagnosis using gene expression data," *J Comput Biol.*, vol. 11, no. 23, pp.227–242, 2004.

[92] H. Zhang, J. Y. Ahn, X. D. Lin, et al., "Gene selection using support vector machines with a non-convex penalty," *Bioinformatics*, vol.22, no. 1, pp. 88-96, 2006.

[93] M. Shah, M. Marchand, and J. Corbeil, "Feature selection with conjunctions of decision stumps and learning from microarray data," IEEE *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 174–186, 2012.

[94] S. M. Shafi, M. I. Molla, J. J. Jui, & M. M. Rahman, "Detection of colon cancer based on microarray dataset using machine learning as a feature selection and classification techniques", *SN Applied Sciences*, vol. 2, no. 7, pp. 1–8, 2020.

[95] Ding, X., Yang, F., Jin, S., & Cao, J. (2021), "An Efficient Alpha Seeding Method for Optimized Extreme Learning Machine-based Feature Selection Algorithm", *Computers in Biology and Medicine*, 104505, 2021.

[96] Peng, F. Long and C. Ding, "Feature Selection on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[97] Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinformatics and Computational Biol.*, vol. 3, no. 2, pp. 185-205, 2005.
[98] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," *BMC Bioinformatics*, vol. 6, no. 76, 2005.
[99] Y. Zhang, C. Ding, and T. Li. "Gene selection algorithm by combining relief and mRMR," *BMC Genomics*, 9(Suppl 2): S27, 2008.
[100] Tang, Y. Zhang, Z. Huang, et al., "Recursive fuzzy granulation for gene subsets extraction and cancer classification," *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, no. 6, pp. 723–730, 2008.
[101] L. Chen, J. Xuan, R. B. Riggins, et al., "Identifying cancer biomarkers by network constrained support vector machines," *BMC Systems Biology*, vol. 5, no. 161, 2011.
[102] Y. Piao, M. Piao, K. Park, et al., "An Ensemble Correlation-Based Gene Selection Algorithm for Cancer Classification with Gene Expression Data," *Bioinformatics*, vol. 28, no. 24, pp. 3306–3315, 2012.
[103] J. Liu, G. Cutler, W. Li, et al., "Multiclass cancer classification and biomarker discovery using GA-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.
[104] Mukhopadhyay and U. Maulik, "An SVM-wrapped multi-objective evolutionary feature selection approach for identifying cancer-microRNA markers, "*IEEE Transactions on Nano Bioscience*, vol. 12, no. 4, pp. 275– 281, 2013.
[105] R. Kundu , S.Chattopadhyay a, E. Cuevas b, R. Sarka r, "AltWOA: Altruistic Whale Optimization Algorithm for feature selection on microarray datasets", Vol. 144, no. 105349, 2022.
[106] Alhenawi, R. AlSayyed, A. Hudaib, S. Mirjalili, "Improved intelligent water drop-based hybrid feature selection method for microarray data processing", Computational Biology and Chemistry, Vol. 103, no. 107809, 2023.
[107] S, D. A. (2021). CCT Analysis and Effectiveness in e-Business Environment. International Journal of New Practices in Management and Engineering, 10(01), 16–18. https://doi.org/10.17762/ijnpme.v10i01.97
[108] Agrawal, S. A., Umbarkar, A. M., Sherie, N. P., Dharme, A. M., & Dhabliya, D. (2021).Statistical study of mechanical properties for corn fiber with reinforced of polypropylene fiber matrix composite. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1072