TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 2, April 2021, pp. 491~498
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i2.15948

The importance of data classification using machine learning methods in microarray data

Aws Naser Jaber[1], Kohbalan Moorthy[2], Logenthiran Machap[3], Safaai Deris[4]
[1,2] College of Computing and Applied Sciences, Faculty of Computing, Universiti Malaysia Pahang, Malaysia
[3]Department of Computing and Information Technology, Tunku Abdul Rahman University College Malaysia, Malaysia
[4]Faculty of Bioengineering and Technology, Universiti Malaysia Kelantan, Malaysia

Article Info

Article history:

Received Feb 29, 2020
Revised Aug 13, 2020
Accepted Aug 29, 2020

Keywords:

Cancers
DNA
Gene expression
Machine learning
Microarrays
RNA

ABSTRACT

The detection of genetic mutations has attracted global attention. several methods have proposed to detect diseases such as cancers and tumours. One of them is microarrays, which is a type of representation for gene expression that is helpful in diagnosis. To unleash the full potential of microarrays, machine-learning algorithms and gene selection methods can be implemented to facilitate processing on microarrays and to overcome other potential challenges. One of these challenges involves high dimensional data that are redundant, irrelevant, and noisy. To alleviate this problem, this representation should be simplified. For example, the feature selection process can be implemented by reducing the number of features adopted in clustering and classification. A subset of genes can be selected from a pool of gene expression data recorded on DNA micro-arrays. This paper reviews existing classification techniques and gene selection methods. The effectiveness of emerging techniques, such as the swarm intelligence technique in feature selection and classification in microarrays, are reported as well. These emerging techniques can be used in detecting cancer. The swarm intelligence technique can be combined with other statistical methods for attaining better results.

Corresponding Author:

Kohbalan Moorthy
College of Computing and Applied Sciences, Faculty of Computing,
Universiti Malaysia Pahang
Kuantan, Pahang 26300, Malaysia
Email: kohbalan@umk.edu.my

1. INTRODUCTION

Bioinformatics involves the use of computers in managing biological information [1]. In general, Bioinformatics involves microarrays classification, organization, and interpretation. This technology can be employed to solve various problems related to the biological field. For example, the prediction process can be used to control and prevent many diseases such as Cancer. The process can also be used to discover new disease markers. Detection of mutations in gene expression patterns would essentially lead to the development of an efficient therapy method [2]. The gene controls cell development, and malfunction of genes leads to tumour formation or cancer. Deoxyribonucleic acid (DNA) microarray approach is a powerful approach that helps explore the genetic defects in the human body [3]. For example, microarray technology has led to successful cancer diagnosis [4, 5]. Gene expression studies that involve gene selection, classification and clustering have been carried out [6]. A suitable hybrid system in bioinformatics has been developed to detect cancer (gene mutation) and other diseases in a more accurate manner [7].

2. LITERATURE REVIEW

Cancer involves abnormal cell-growths, which is fatal if it is ignored as the developed tumour may spread to other body parts via bloodstream [8]. There are several common cancers: lung, prostate, breast, and oral cancers and colons [9]. An example of the cancer level shown in Figure 1, and for this, we need to know what the DNA microarray technology and what kind the machine is learning that applied to the microarray.

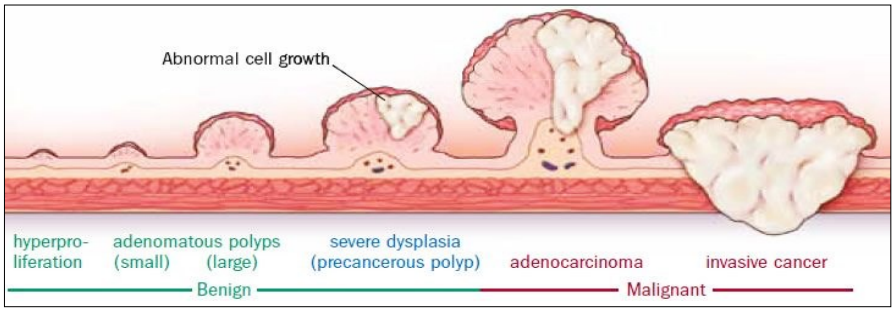

Figure 1. Cancer progression in the colon

2.1. Microarray technology

In Biology, DNA is the blueprint of an organism and therefore it has all the information necessary for the biological process. When it comes to computational biology, this information is required for prediction, analysis and so much more. Hence, this information is required to be present in a form where computationally relevant for processing and analysis. Therefore, gene expression profiles in the form of numerical representation are used to perform feature selection, and classification. Microarray technology is the solution to such a requirement. DNA microarray (also commonly known as DNA chip or biochip) can be adapted to reveal the expression levels of many genes in a single reaction concurrently [10]. Firstly, the structure of the protein is different from the structure of the gene and its analysis is still difficult. Therefore, the analysis of thousands of proteins will take a great deal of time. We also saw that one amino acid can be encoded by several codons, sequence of amino acids in the protein We will have several probabilities for the gene formula that produced this protein. The easiest way is to extract the mRNA in the cell and measure its percentage. Generally, the hybridization principle is used in Microarray technology to measure the gene expression levels in the human body [11].

Figure 2 shows the basic process involved in microarray technology. To conduct gene expression profiling, DNA sample and control sample from a patient is obtained. Then, DNA in the sample is denatured into single-stranded molecules. After that, the single-stranded molecules are cut into smaller fragments and then label it with a fluorescent dye. The green dye is for the control sample and red dye is for a normal sample. Both samples are inserted into the chip to hybridize or bind with the synthetic DNA on the chip. After the hybridization, the gene expression can be identified through the changes of colour on the chip. Therefore, this technology can be used in cancer diagnosis and drug response [12]. Thus, via machine learning, significant information about genes representing a disease state and those highly associated genes that shared biological features can be extracted [13].

An accurate cancer diagnosis can be attained by executing the microarray data classification by simply building classifiers to compare the gene expression profiles of tissues of known and unknown cancer status [14]. As a result, the classification process could be misleading due to the existence of noisy and irrelevant data. Therefore, a feature selection method should be devised to reduce the size of the feature set, or gene set [15]. In general, a microarray diagnosis process involves feature selection and classification [16]. To update, many machine learning algorithms have been developed for detecting mutations, e.g., ANN, SVM, clustering, and swarm intelligence algorithm. By using these methods, an optimal subset of genes can then be chosen to build a classification model.

2.2. Feature selection techniques

There are three feature selection techniques in classification, i.e., filter, wrapper, and embedded methods as shown in Figure 3. Filter based approaches are well known for data filtering or pre-processing to rank the genes and then the highly ranked genes will be used in further analysis. Then for the wrapper-based method, gene selection is done using the machine learning method and uses cross-validation to assess the feature subset score. Whereas embedded based. However, microarray data contain many non-significant features that would degrade the performance of most of the learning algorithms [17].

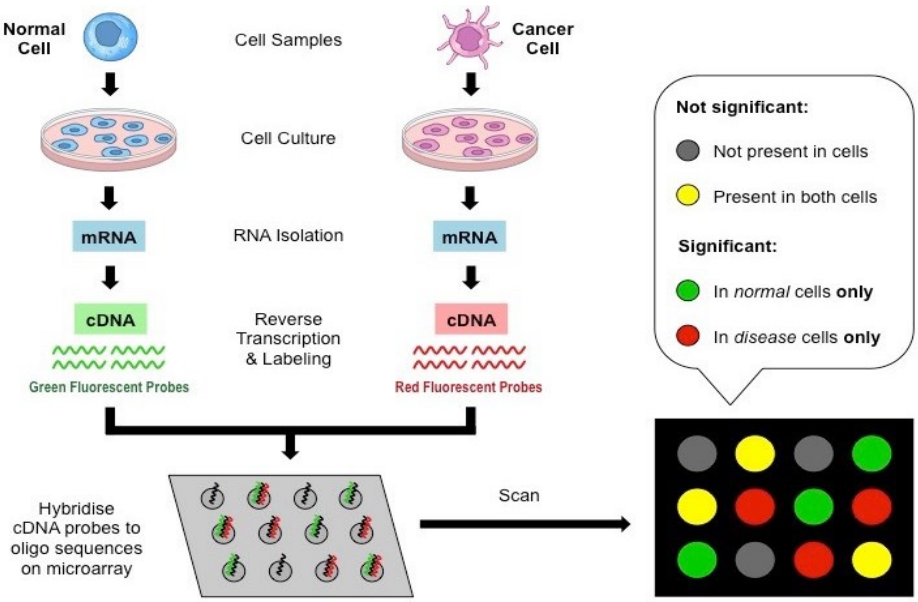

Figure 2. DNA microarrays

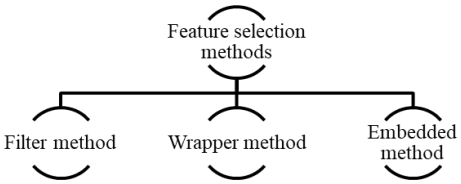

Figure 3. The feature selection methods

2.3. Different methods of feature selection

Normalization involves reducing unwanted variation within arrays. Typical assumptions made in some major normalization methods are:

- Only a small number of genes are differentially expressed in terms of condition.
- Annotation: This process involves gene characterization.
- Summarization: Performing only a single measurement after performing a combination in a certain manner
- Statistical Analysis: From a statistical point of view, the number of genes could be larger than the number of samples, thus leading to faulty classification. Feature selection should be performed by selecting the most informative gene to improve the accuracy and efficiency of the classification process and to address the problem of dimensionality.
- Biological Interpretation: To interpret microarray data, one must have an adequate number of replicate measurements to determine results that have real predictive value. Dimensionality reduction is therefore essential.

In microarray classification, samples are classified into both abnormal (cancer) and normal datasets based on microarray measurements [18, 19]. It is challenging to train the classifiers on such datasets of high dimensionality [20]. Preprocessing is an essential step to address this dimensionality problem, and then apply the classification algorithm for monitoring model complexity via regularization. Machine learning enables a system to automatically perform the learning process. It is not a real learning process; however, the system can recognize complex data patterns and make intelligent decisions based on computational methods. Classification is a procedure used to categorize sample data into a few classes. Some popular classification methods employed in data mining and other fields are artificial neural network (ANN), decision tree, support vector machine (SVM) and swarm intelligence.

Artificial neural network (ANN) or neural network (NN) is a method in artificial intelligence that mimics the complex processes as in the human brain. ANN requires a huge number of units' collection that is interconnected to permit communication between the units. The unit also denoted as nodes or neurons. They are simple processers function in parallel. Next is the decision tree method; this is a predictive modelling tool that falls under supervised learning. There are two main entities in decision tree called nodes. Besides, there

are two types of decision trees such as classification and regression trees. SVM is another popular supervised classification method. The basic principle of SVM is, creates hyperplane that separates the dataset into classes. Furthermore, the swarm intelligence method is to use numerous simple agents with no rule to interact locally and globally. Popular swarm intelligence algorithms are ant colony optimization (ACO), artificial bee colony optimization (ABC), and particle swarm optimization (PSO).

3. RELATED WORK

Several microarray applications have been reported in related review [21]. However, microarray can be hybridized with machine learning algorithms such as non-swarm intelligence and swarm intelligence algorithms. After detecting and filtering gene expression datasets, samples should be accurately classified into known groups by the features of gene expression. Hence, support vector machines (SVM), prediction analysis of microarrays (PAM), classification and regression trees (CART), K nearest-neighbor (K-NN) methods can be employed. Turgut, *et al.,* applied a machine learning classifier for microarray breast cancer. First, they perform the right types of machine learning algorithms without applying any feature selection, and then they used two different feature selections. Examples of machine learning algorithms KNN, SVM, decision trees, MLP, random forest, logistic regression, adaboost and gradient boosting machine [22]. They claimed that MLP did not improve accuracy.

Bharathi, A. M. Natarajan minimized the gene set for more accurate classification using ANOVA [23]. The ranking of a gene was computed using ANOVA. SVM was used as a classifier. The technique was compared with the T-test classifier. Interestingly, the hybridization technique of ANOVA and SVM was accurate even using a minimum number of genes. While, another research proposed an artificial immune recognition system to classify microarray data (cancer, disease or normal tissues) [24]. The result was then compared with those of other classifiers. In AIRS, a memory cell is used for training samples to build a classifier. The experiment was applied to colon cancer, brain tumour, and nine tumour datasets. AIRS performed better than other machine learning methods such as KNN, OneR, and Naïve Bayes.

Karayianni, *et al.,* employed the fuzzy clustering method with viewpoints to identify unlabeled samples [24]. The clusters are identified by calculating the expression mean of each feature with labelled samples. This technique was applied to breast cancer, brain cancer, AML, and MLL datasets. Sudip Mandal and Indrojit Banerjee applied ANN to diagnose and detect cancer [25]. A special kind of ANN called multilayer feed forward neural network (MLFF) was used. The performance of ANN is dependent on parameters such as the number of hidden layers, number of nodes and weights. Different datasets consisting of breast and lung cancerous cells were employed. Two analyses were performed: cross-validation and new dataset testing. Datasets were divided into training (80%) and testing (20%) datasets. Due to the noise in the dataset, the accuracy was 96% after cross-validation and 94% for new dataset testing. ANN was designed with a single hidden layer, but the structure of ANN can be tuned for better accuracy.

In [26], they used the α depended on the degree-based feature selection method to solve the imbalance problem between the feature number and the instance number in microarray data-based gene expression. The classification accuracy of smaller gene size was better than that of larger gene size. Nine datasets have been used in this study such as colon tumour, central nervous system tumour, diffuse large B cell lymphoma, leukemia 1, AML, lung cancer, prostate cancer, breast cancer, and leukaemia. The results were compared with other techniques such as NB (Naïve Bayes), DT (decision tree), SVM (support vector machine) and K-NN (K-nearest neighbour). As reported, the k-NN classifier had better performance under seven α values. Li and colleagues assessed five feature selection methods such as KNN, C4.5, Naïve Bayes and SVM with leukemia and ovarian cancer datasets [27, 28] has presented a comparative study on three feature selection methods with four data sets. They used prostate, colon tumour, and Leukemia and Hepato datasets. SVM performs better on all the datasets.

Chanho and Sung-Bae conducted an analysis of colon cancer and Lymphoma datasets by seven gene selection methods and six classifiers. Besides, Ji-Gang and Hong-Wen developed a gene selection method based on Bayes error filter (BBF) [29]. BBF can select significant genes while removing non-significant genes. This evaluated using datasets include colon, prostate, lymphoma, leukemia, and DSLBCL. They had used SVM and KNN for measuring accuracies. They observed that SVM performed well on all the datasets used. Xing, Jordan, and Karp studied different classifiers such as the Gaussian classifier, regression classifier, and KNN. Feature reduction by these three methods shows better results. They proposed a hybrid approach of filter and wrapper for feature selection in high dimensional data. Mainly they have used Markov Blanket filtering and then classified with the use of three different classifiers. Thus, these classifiers able to perform better with the reduced significant feature space compared to full feature space [30].

On top of that, Onskog and colleagues had presented microarrays classification on seven cancer-related data. Double cross-validation methods are applied to obtain a strong error rate. The results

show that SVM with a radial basis kernel and linear kernel performed steadily with these data sets. Moreover, based on the t-test there is a synergistic association between the methods and gene selection process [31]. Besides this, [32] proposed a machine learning study on prostate cancer data set. In particular, the t-test and interquartile range are combined for feature selection. The results produced show that Bayes Network is outperformed, Naïve Bayes. In [33] different discrimination methods are used for classification on three cancer gene expression data sets. The methods are nearest-neighbour classifiers, linear discriminant analysis, and classification trees. From the output, the nearest neighbour classifies better compared to the decision tree classifier.

Furthermore, Sung Bae and colleagues used three microarray data sets namely, Leukemia, colon, and Lymphoma with feature selection and classifiers. The investigation results show that the ensemble classifiers produced the best classification rate compared to other methods [34]. Abusamra has done an investigation on eight different feature selection methods and three classification methods. The feature selection methods include max minority, information gain, Gini index, t- statistics, the sum of variances and one-dimension support vector machine was compared. The classification methods are SVM, KNN, and random forest. Two types of glioma expression data sets are used in this experiment. The results show that the selection of significant genes had boosted classification accuracy. In both datasets, SVM performed better than other classification methods [35].

The maximum relevance minimum redundancy (mRMR) algorithm is a special group of filter-based approaches which able to select concurrently highly predictive but uncorrelated features. This algorithm mainly selects features subset having the maximum association with a class (relevance) and the minimum association between themselves (redundancy). The feature's ranking is given based on minimal-redundancy-maximal-relevance measures. Hence F-statistics is used to calculate the relevance and Pearson correlation coefficient is used to calculate the redundancy [36]. Besides this, [37] developed the Monte Carlo feature selection (MCFS) algorithm to identify informative features. The MCFS algorithm is integrating interdependencies among features. It has some similarity as in random forest methodology but differs in terms of feature ranking calculation [37].

Besides this, Alshamlan and colleagues proposed a new feature selection method called minimum redundancy maximum relevance (mRMR) hybrid with an artificial bee colony (ABC). This algorithm is specifically to select significant genes from the microarray. The experiment is conducted with six binary and multiclass data sets. The produced result shows that the proposed algorithm has achieved better classification accuracy compared to mRMR-GA and mRMR-PSO algorithms [38]. Jayger, Sengupta, and Ruzzo [39] study various gene selection methods for microarray data classification. They used various statistics test with gene selection methods. The statistics tests include Fisher, Golub, Wilcoxon, TNoM, and t-test.

Huawen, Lei and Huijie compared various gene selection methods [40]. They compared ensemble gene selection by grouping with the other three gene selection methods FCBF, mRMR, and ECRP. They used five datasets with these techniques. They used two classification methods Naïve Bayes and KNN. They compared and analyzed which classification method is effective. While, in [25], they employed the fuzzy clustering method with viewpoints to identify unlabeled samples. The viewpoints were constructed by computing the average expression for each feature (probe/gene) in the samples with a label. In their work, the previously available microarray expression data was introduced as viewpoints in the clustering process. The technique was applied to breast cancer, brain cancer, AML, and MLL datasets. The method was found to be better than other clustering algorithms such as K means, fuzzy c-means, affinity propagation, and the clustering method based on prior biological knowledge. However, Table 1 shows the most related works in microarray DNA.

Table 1. Most related work for the Microarray DNA

Dataset	References
Acute lymphoblastic leukemia (ALL), AML, MLL, DLBCL, Lymphoma, SRBCT	[26] [27, 29, 44]
Breast Cancer	[32]
Colon Cancer	[29, 44]
Prostate	[16, 29, 44]
Ovarian cancer	[27]
Lung Cancer	[16, 26]
Glioma, Brain tumour, CNS (Central Neural System)	[16, 35]
Hepatic	[44]
Diabetes	[32]

Furthermore, [41] hybrid cellular automata and ant colony optimization method to select the significant genes then used for classification. Thus, it has produced high accuracy compared to other selected methods as shown in the paper. Moreover in [42], an artificial neural network (ANN) is applied to ALL and

AML datasets. This research had generated 98% accuracy, where there is no error in ALL datasets and one error in AML dataset. The cancer genome atlas (TCGA) is a pilot project launched by the National Institute of Health (NIH). This is basically to create a comprehensive atlas of cancer genomic profiles. Hence, most of the gene expression data are publicly available at TCGA that are used in prognosis and diagnosis [43].

4. CONCLUSION

This paper reviews the existing classification techniques applied in microarrays that contain high dimensional data. The high dimensional data problem can be solved using feature selection methods. Many gene selection methods have been used to classify cancerously or any other disease datasets with multi or binary classes. The underlying challenge is the efficient detection of different infected genes with different characteristics such as mutated genes caused by viruses, radiation, mutagenic chemicals. Machine learning techniques have been proposed to analyze microarray data. Hybridized methods can eliminate noise, reduce the number of features and ease classification. Swarm intelligence algorithms such as ant colony optimization (ACO), artificial bee colony optimization (ABC), particle swarm optimization (PSO) are powerful in feature selection. Hybridization between the classical machine learning techniques and the emerging machine learning techniques such as swarm intelligence algorithms can yield better results in diagnosis and classification. Currently, researchers have developed hybridized computational methods with swarm intelligence (SI) methods and proven that these hybridized systems are more accurate. Nevertheless, a model that solely relies on swarm intelligence algorithms should be built and analyzed.

5. ACKNOWLEDGEMENTS

We would like to thank Universiti Malaysia Pahang for supporting this work under the RDU Grant, Grant number: RDU190373. We also appreciate the Ministry of Education for supporting this work under the Grant Number: RACER/1/2019/ICT02/UMP//4.

REFERENCES

[1] K. Lan, D. T. Wang, S. Fong, L. S. Liu, K. K. Wong, and N. Dey, "A survey of data mining and deep learning in bioinformatics," *Journal of medical systems,* vol. 42, no. 8, pp. 1-28, 2018.
[2] Y. Dor and H. Cedar, "Principles of DNA methylation and their implications for biology and medicine," *The Lancet,* vol. 392, pp. 777-786, 2018.
[3] H. Q. Truong, L. T. Ngo, and W. Pedrycz, "Granular fuzzy possibilistic C-means clustering approach to DNA microarray problem," *Knowledge-Based Systems,* vol. 133, pp. 53-65, 2017.
[4] S. Mittal, H. Kaur, N. Gautam, and A. K. Mantha, "Biosensors for breast cancer diagnosis: A review of bioreceptors, biotransducers and signal amplification strategies," *Biosensors and Bioelectronics,* vol. 88, pp. 217-231, 2017.
[5] C. Xu and S. A. Jackson, "Machine learning and complex biological data," *Genome Biology,* vol. 20, no.1, pp. 1-4, 2019.
[6] F. P. Barthel, W. Wei, M. Tang, E. Martinez-Ledesma, X. Hu, S. B. Amin*, et al.*, "Systematic analysis of telomere length and somatic alterations in 31 cancer types," *Nature genetics,* vol. 49, no. 3, pp. 349-357 2017.
[7] P. Mamoshina, A. Vieira, E. Putin, and A. Zhavoronkov, "Applications of deep learning in biomedicine," *Molecular pharmaceutics,* vol. 13, pp. 1445-1454, 2016.
[8] Y. Lin, J. C. Mauro, and G. Kaur, "Bioactive Glasses for Cancer Therapy," in *Biomedical, Therapeutic and Clinical Applications of Bioactive Glasses*, ed: Elsevier, 2019, pp. 273-312.
[9] V. Gupta, M. Sengupta, J. Prakash, and B. C. Tripathy, *Basic and applied aspects of biotechnology*: Springer, 2017.
[10] D. Ofengeim, N. Giagtzoglou, D. Huh, C. Zou, and J. Yuan, "Single-cell RNA sequencing: unraveling the brain one cell at a time," *Trends in molecular medicine,* vol. 23, no. 6, pp. 563-576, 2017.
[11] M. Afzal, I. Manzoor, and O. P. Kuipers, "A fast and reliable pipeline for bacterial transcriptome analysis case study: serine-dependent gene regulation in Streptococcus pneumoniae," *JoVE (Journal of Visualized Experiments),* 2015.
[12] S.-B. Liang and L.-W. Fu, "Application of single-cell technology in cancer research," *Biotechnology Advances,* vol. 35, pp. 443-449, 2017.
[13] R. C. Thompson, A. F. Seasholtz, J. O. Douglass, and E. Herbert, "Cloning and distribution of expression of the rat corticotropin-releasing factor (CRF) gene," in *Corticotropin-Releasing Factor*, ed: CRC Press, 2018, pp. 1-12.
[14] L. Chen, X. Pan, X. Hu, Y. H. Zhang, S. Wang, T. Huang*, et al.*, "Gene expression differences among different MSI statuses in colorectal cancer," *International journal of cancer,* vol. 143, pp. 1731-1740, 2018.
[15] G. Taşkın, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Transactions on Image Processing,* vol. 26, no. 6, pp. 2918-2928, 2017.
[16] R. Liu, X. Wang, K. Aihara, and L. Chen, "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers," *Medicinal research reviews,* vol. 34, no. 3, pp. 455-478, 2014.

[17] F. Li, D. Miao, and W. Pedrycz, "Granular multi-label feature selection based on mutual information," *Pattern Recognition,* vol. 67, pp. 410-423, 2017.
[18] J. Li, J. Wang, Y. Zheng, and H. Xiao, "Microarray classification with noise via weighted adaptive elastic net," in *2017 6th Data Driven Control and Learning Systems (DDCLS)*, 2017, pp. 416-420.
[19] C. Kang, Y. Huo, L. Xin, B. Tian, and B. Yu, "Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine," *Journal of Theoretical Biology,* vol. 463, pp. 77-91, pp. 77-91, 2019.
[20] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom, "Generative and discriminative text classification with recurrent neural networks," *arXiv preprint arXiv:1703.01898,* 2017.
[21] C. Huang, R. Mezencev, J. F. McDonald, and F. Vannberg, "Open source machine-learning algorithms for the prediction of optimal cancer drug therapies," *PLoS One,* vol. 12, 2017.
[22] S. Turgut, M. Dağtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pp. 1-3, 2018.
[23] A. Bharathi and A. Natarajan, "Cancer classification of bioinformatics data using anova," *International journal of computer theory and engineering,* vol. 2, no. 3, pp. 369-373, 2010.
[24] K. N. Karayianni, G. M. Spyrou, and K. S. Nikita, "Clustering microarray data using fuzzy clustering with viewpoints," in *2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE)*, 2012, pp. 362-367.
[25] S. Mandal and I. Banerjee, "Cancer classification using neural network," *International Journal of Emerging Engineering Research and Technology,* vol. 3, no. 7, pp. 172-178, 2015.
[26] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," *Genome informatics,* vol. 13, pp. 51-60, 2002.
[27] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics,* vol. 20, no. 15, pp. 2429-2437, 2004.
[28] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE transactions on nanobioscience,* vol. 9, no. 1, pp. 31-37, 2009.
[29] J.-G. Zhang and H.-W. Deng, "Gene selection for classification of microarray data based on the Bayes error," *BMC bioinformatics,* vol. 8, no. 1, pp. 1-9, 2007.
[30] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in *Icml*, 2001, pp. 601-608.
[31] J. Önskog, E. Freyhult, M. Landfors, P. Rydén, and T. R. Hvidsten, "Classification of microarrays; synergistic effects between normalization, gene selection and machine learning," *BMC bioinformatics,* vol. 12, 2011.
[32] K. Raza and A. N. Hasan, "A comprehensive evaluation of machine learning techniques for cancer class prediction based on microarray data," *arXiv preprint arXiv:1307.7050,* 2013.
[33] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association,* vol. 97, no. 457, pp. 77-87, 2002.
[34] S.-B. Cho and H.-H. Won, "Machine learning in DNA microarray analysis for cancer classification," in *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19*, 2003, pp. 189-198.
[35] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma," *Procedia Computer Science,* vol. 23, pp. 5-14, 2013.
[36] C. DING and H. PENG, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Journal of Bioinformatics and Computational Biology,* vol. 03, pp. 185-205, 2005.
[37] M. Dramiński, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics,* vol. 24, no. 1, pp. 110-117, 2007.
[38] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *Biomed research international,* vol. 2015, no. 9, pp. 1-15, 2015.
[39] J. Jäger, R. Sengupta, and W. L. Ruzzo, "Improved gene selection for classification of microarrays," in *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing,* vol. 8, 2003, pp. 53-64.
[40] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection for cancer classification," *Pattern Recognition,* vol. 43, no. 8, pp. 2763-2772, 2010.
[41] F. Vafaee Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization," *Genomics,* vol. 107, no. 6, pp. 231-238, 2016.
[42] A. K. Dwivedi, "Artificial neural network model for effective cancer classification using microarray gene expression data," *Neural Computing and Applications,* vol. 29, no. 12, pp. 1545-1554, 2018.
[43] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary oncology (Poznan, Poland),* vol. 19, no. 1A, pp. A68-A77, 2015.
[44] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience,* vol. 4, no. 3, pp. 228-234, 2005.

BIOGRAPHIES OF AUTHORS

Aws Naser Jaber is currently an Information Security Expert at EarthLink Telecommunications. He received his BSc from Al-Mustansiriya University, his MSc from Universiti Sains Malaysia, and his PhD from Universiti Malaysia Pahang. His expertise includes cryptography, cybersecurity, artificial intelligence, and bioinformatics.

Kohbalan Moorthy is currently a Senior Lecturer at the Faculty of Computing at Universiti Malaysia Pahang. He received his BSc and PhD from Universiti Teknologi Malaysia in the area of Computer Science. His area of expertise includes Bioinformatics, Medical Informatics, Computational Biology, Artificial Intelligence, IoT and other related fields in Computer Science. He has won many research awards on both the local and international levels. He is also a Microsoft Certified Professional (MCP) and Professional Technologist (P.Tech). He is currently active in Research, Management, Teaching & Learning in the University Malaysia Pahang.

Logenthiran Machap is currently working as a Lecturer at the Department of Computing and Information Technology in Tunku Abdul Rahman University College. He completed B.Sc. Bioinformatics from National University of Malaysia (UKM) and Master of Computer Science from University of Technical Malaysia Melaka (UTeM). He is a PhD candidate for computer science from the University of Technology Malaysia (UTM). He is working on the co-clustering algorithm and its application on cancer microarray gene expression data. His research interests include data mining, machine learning, artificial intelligence, and bioinformatics.

Safaai Deris is a Professor at the Faculty of Bioengineering and Technology, University Malaysia Kelantan. He was a Professor of Artificial Intelligence and Software Engineering at the Faculty of Computing, Universiti Teknologi Malaysia. He is the CIO and Director of the Center for Information and Communication Technology. His previous positions include Deputy Dean of the School of Graduate Studies and Head of Software Engineering Department. He is also Head of Artificial Intelligence and Bioinformatics Research Group, He received the M. Eng. degree in Industrial Engineering, and the D. Eng. degree in Computer and System Sciences, both from the Osaka Prefecture University, Japan. His research interests include software engineering, applications of intelligent techniques in planning, scheduling, bioinformatics, and system biology. He has published more than 200 journals and conference refereed papers.