Feature selection with ensemble learning for prostate cancer diagnosis from microarray gene expression

Abdu Gumaei
Research Chair of Pervasive and Mobile Computing, King Saud University, Saudi Arabia
Taiz University, Yemen

Rachid Sammouda
King Saud University, Saudi Arabia

Mabrook Al-Rakhami
Research Chair of Pervasive and Mobile Computing, King Saud University, Saudi Arabia

Hussain Al Salman
King Saud University, Saudi Arabia

Ali El-Zaart
Beirut Arab University, Lebanon

Abstract
Cancer diagnosis using machine learning algorithms is one of the main topics of research in computer-based medical science. Prostate cancer is considered one of the reasons that are leading to deaths worldwide. Data analysis of gene expression from microarray using machine learning and soft computing algorithms is a useful tool for detecting prostate cancer in medical diagnosis. Even though traditional machine learning methods have been successfully applied for detecting prostate cancer, the large number of attributes with a
small sample size of microarray data is still a challenge that limits their ability for effective medical diagnosis. Selecting a subset of relevant features from all features and choosing an appropriate machine learning method can exploit the information of microarray data to improve the accuracy rate of detection. In this paper, we propose to use a correlation feature selection (CFS) method with random committee (RC) ensemble learning to detect prostate cancer from microarray data of gene expression. A set of experiments are conducted on a public benchmark dataset using 10-fold cross-validation technique to evaluate the proposed approach. The experimental results revealed that the proposed approach attains 95.098% accuracy, which is higher than related work methods on the same dataset.

**Keywords**
prostate cancer, microarray data, machine learning, random committee, ensemble learning, feature selection, 10-fold cross-validation

**Introduction**

Prostate cancer disease is the second leading source of mortality in men. It comes after lung cancer in terms of exceeding and threatening in the world. The etiology of this cancer type is still not fully recognized yet. However, some factors such as heredity, diet, and environmental influences that affect male hormones have been mentioned in epidemiological studies.1–6

Microarray technology is a leading-edge technology in molecular biology when it comes to the contribution of information in quantitating hundreds or thousands of genes that are used in the diagnosis of different diseases and to predict the possible outcomes of an ailment. The genes which are regulated due to disease condition can be analyzed through the expression extracted from the microarray data samples. These measurements also assist in the investigation of cancer for clinical medicine at the biology and molecular level.7

Cancer can alter the gene expression profile of the body cells. This fatal genetic disease transpired from the mutations or epigenetic changes. Therefore, the microarray data is utilized in clinical diagnosis to recognize down or up the regulated gene expression, which is the reason for activating some oncogenic pathways, generating new biomarkers, and leading to cancer disease.8 Nevertheless, this approach comes with a considerable cost and time. Moreover, it is not clinically applicable to all patients. The algorithms used in the data analysis are not helping the researchers either due to their restrictions, which is a massive setback for microarray technology. Microarray array data analysis has been used as a source for profiling gene expression for many decades.9 However, it suffers from noise and the difficulty of range detection as it involves transcriptome and genome references.10 Mainly, it utilizes the sequence-specific probe hybridization accompanied by fluorescence detection to estimate gene expression levels.11

Progression of cancer could be closely monitored through building a set of genes markers with data analysis techniques. Quantity of genes utilized in the analysis is highly crucial for microarray data. A large number of genes might lead to redundancy and correlation among features affecting the accuracy results of a cancer diagnosis. Besides, a small number of genes could result in unreliable accuracy results.12 Therefore, an optimal number of genes is needed to predict the class labels of prostate cancer efficiently with high accuracy results. A significant drawback of microarray data is the massive number of genes captured from microarray data with a small number of samples. This drawback will not only significantly cause to overfitting problem and decrease the accuracy of microarray data classification, but also increase the computational cost.13 A small sample size of microarray data with large attributes is still ongoing research direction to date. Many studies seek
to solve this problem through many techniques; for example, Li et al. proposed a novel method called the semi-supervised maximum discriminative local margin (semiMM). The authors used mutual information theories and spectral graph for selecting gene from expression data that has a small sample size with high dimensionality. In the other recent works proposed in Nirmalakumari et al., Raj and Mohanasundaram, Bentkowska, and Santhakumar and Logeswari, the authors have presented some solutions for this research issue in gene expression of microarray data.

In this case, selecting the significant and most essential genes is needed for reducing the high dimensionality of genes feature space. The relevance of genes is grouped into three groups: Strongly related, weakly related, and unrelated genes. The group of strongly related is for those genes that are shown in the cancer cell formation and required in the optimum set. The weakly related and unrelated groups are ignored from the optimum set.

Genes are selected for the investigation of disease for the following reasons: (a) by rendering only the significant genes make the classification process easy, (b) accuracy of classification is improved, and (c) the dimensionality of the data set is reduced. There are several methods like neighborhood-based analysis, Bayesian variable-based selection, principle component analysis (PCA)-based reduction, and genetic-based evolution of sequence expressions are used for choosing the optimal subset of genes for classification.

The effectiveness of genes selection is assessed by the accuracy of classification methods, which is highly crucial. There is also a variety of machine learning-based classification methods that can be used with genes feature selection for improving classification accuracy results. In recent years, machine learning algorithms are used in several applications and tasks, including brain tumor classification, fall detection in connected home healthcare, lymphoma prediction, diabetes disease classification, breast segmentation using k-means algorithm, human activity recognition, and medical decision support.

In the same context, some other machine learning algorithms such as k-nearest neighbor (kNN), support vector machine (SVM), artificial neural network (ANN), deep learning model, random forest (RF), convolutional neural network, and maximum margin linear programming (MMLP) are applied for analyzing the microarray data.

There are some other works that have been employed the microarray data for cancer detection and diagnosis. The recent work introduced a method that aims to increase accuracy by using microarray data for cancer detection. The earliest works proposed to use microarray datasets with different machine learning methods for diagnosing prostate cancer have been introduced in Penney et al., Cuzick et al., Erho et al., Mo et al., Tyekucheva et al., and Sharifi-Noghabi et al. The studies of these works are designed to predict if the tumor is metastasized or not. Even though the experimental results obtained from the microarray data of these studies are relatively acceptable, the gene expression features gained from the microarray data should be reduced to improve classification accuracy.

In Takeuchi et al., the authors proposed a machine learning approach to diagnose cancer from clinical prostate data using an ANN model. They trained the ANN on clinical data containing 22 features. Although they reported that the model performed well, some improvements are needed on that model before being applicable for clinical diagnosis. However, there is still a need for more robust, accurate, and easily interpretable classification methods.

In the literature studies, some rule-based evolutionary machine learning models, such as BioHEL and GAssist are used for prostate tumor classification in Glaab et al. through evaluating them on a large-scale public microarray prostate cancer dataset, which consists of expression measurements for 12,600 genes, acquired from 50 healthy normal tissues and 52 prostate cancer tissues. Other works such as Huerta et al., Chen et al., Dashtban and Balafar, Dashtban et al.,
and Shen and Tan\textsuperscript{52} focused on improving the accuracy of cancer classification and prediction through gene feature selection or gene feature reduction. In Wessels et al.,\textsuperscript{53} the authors introduced a protocol to build and evaluate machine learning predictors of a disease state that applied to microarray data.

Recently, in Bouazza et al.,\textsuperscript{54} a comparative study has conducted for prostate cancer diagnosis using a set of feature selection methods and machine learning algorithms on microarray gene expression data. The authors in this study reported that feature selection using signal to noise ratio (SNR), correlation coefficient (CC), and SVM-recursive feature elimination (SVM-RFE) with classification using linear discriminant analysis (LDA) achieved a high accuracy result up to 95%.

However, there is still room for improving the accuracy of prostate cancer classification from microarray data. In other words, the previous works need an effective feature selection method to reduce the high dimension of gene features to a limited subset of relevant features and then use an appropriate machine learning algorithm for cancer classification.

In this study, we propose a practical approach for classifying prostate cancer from gene expression of microarray data by using a correlation feature selection (CFS) method with a random committee (RC) ensemble learning algorithm. The reason behind using a CFS method is its ability to take the correlation between features for selection, and the reason behind using the RC algorithm is its capability to solve the overfitting problems. Moreover, we evaluate the proposed approach using a 10-fold cross-validation technique.

The rest of the paper is structured as follows: Section 2 presents the proposed approach in more detail. Section 3 introduces the experiment and results. Finally, Section 4 summarizes the conclusion of the proposed work.

**Proposed approach**

The proposed approach takes the gene expression of microarray data samples as input. Then, it selects the relevant features using a correlation feature selection (CFS) method. Finally, the input samples with selected features are classified using a random committee (RC) model to output the classification results as normal or tumor cases. Figure 1 shows the flowchart of the proposed approach steps.

In the feature selection process, the less significant features are removed using the CFS method. The CFS calculates the feature-to-class and feature-to-feature correlations and then searches the correlations space of the feature subsets for selecting the best feature subset. Correlation is a statistical similarity measure used to evaluate the relationship between two variables. If the two variables are uncorrelated, the value of their correlation coefficient is 0. On the other hand, if the features are correlated, their correlation coefficient will always be between −1 and +1. Two coefficients can be commonly used to measure the correlation between two arbitrary variables or features. One is using a linear correlation coefficient, and the other is using an information theory coefficient. The linear correlation coefficient is the most familiar measure that can be computed between a pair of variables \((x, y)\) in the training dataset of \(n\) instances as follows:

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

(1)
where $\bar{x}$ represents the mean of $x_i$ values and $\bar{y}$ is the mean of $y_i$ values.

The CFS method removes all redundant and irrelevant features to improve classification accuracy results and speed up the execution time.

In the classification phase, the RC model is a supervised machine learning classifier, used to classify the microarray data samples with the selected features. It is an ensemble of random trees (RTs) base learners. Each base learner is built using a training dataset at a different random number of seeds. The RTs are a collection of individual decision trees (DTs) in which each tree is produced from different subsets and samples of the training dataset. The main idea behind building these DTs is that for each instance to be classified, a number of decisions are generated in rank order according to their importance. When the features of an instance are built, it looks like a branch, and when the features of the entire dataset are constructed, the branches will form a tree. These trees are called RTs because they are actually trained on the dataset a number of times using a random subset of training instances, hence resulting in many DTs. Through this process, the overfitting problem will be mitigated. The final classification model is an average of the classifications generated by the individual RTs base classifiers. At the beginning of this phase, a number of RTs base learners of the RC classifier is initialized. The RC classifier is then trained and tested using a 10-fold cross-validation technique. The 10-fold cross-validation technique divides the dataset into 10 subsets through 10 iterations. For each iteration, a different subset is applied for
testing, and the remaining nine subsets are used for training. The final testing result of the RC model is the average of all testing results obtained by those individual base learners in the 10 iterations.

**Experiment and discussion**

The experiment of this study is conducted on a public dataset of microarray prostate cancer gene expression, consisting of 102 tissue samples (52 prostate tumor and 50 normal tissues) with 2135 genes. For more explanation about microarray data, a microarray is a laboratory tool that can be utilized to record thousands of genes expressed at the same time (see Figure 2). Similarly, DNA microarrays are microscope slides printed with thousands of tiny spots in defined positions, and each spot has a known gene or DNA sequence. Figure 3 presents the distribution of samples in the dataset.

The experimental evaluation metrics and results with comparisons are given in the following subsections.

**Evaluation metrics**

To assess the experimental results, we use a set of evaluation metrics. These evaluation metrics can be explained as follows:

- **Confusion Matrix**
  - It can be defined as a table that visualizes and describes the performance of the classification task on a test dataset in which the true positive and true negative samples are correctly classified. Figure 4 shows the confusion matrix of binary classification.
Recall (True Positive Rate)
It is the number of samples that are correctly recognized as positive out of total true positives samples, computed as:

$$\text{Recall} = \frac{TP}{TP + FN}$$  \hspace{1cm} (2)

Specificity (True Negative Rate)
It is the number of samples that are correctly recognized as negatives out of total negatives samples, computed as:

$$\text{Specificity} = \frac{TN}{TN + FP}$$  \hspace{1cm} (3)

Precision
It is the number of samples that are correctly recognized as positives out of total samples identified as positives samples, computed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$  \hspace{1cm} (4)

Figure 3. The distribution of samples in the dataset of prostate cancer.

Figure 4. Confusion matrix of binary classification.
• **F1-Score**
  It can be defined as a harmonic-mean of recall and precision, given by:

\[
F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  

(5)

• **Accuracy**
  It is the ratio of total samples that are classified correctly, computed as:

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}
\]  

(6)

**Results and comparisons**

After reading the microarray dataset of prostate cancer, the feature selection process selects only 38 genes out of 2135 genes as a significant set of features. These selected features are the most correlated features of the class labels. Then, the RC model is initialized with 30 base learners, trained, and tested on the dataset with these selected features a 10-fold cross-validation technique. Figure 5 shows the confusion matrix of 10-fold cross-validation results based on the selected features to classify normal tissue and prostate tumor.

From the confusion matrix, we can see that 49 samples out of 50 for normal tissues are correctly classified as normal tissues, and 48 samples out of 52 for the prostate tumors are correctly classified as prostate tumors. Therefore, the results of accuracy and weighted average F1-Score for the approach are 95.098% and 95.1%, respectively. Moreover, in Table 1 and Figure 6, we present and visualize the results of other evaluation metrics.

As shown in Table 1 and Figure 6, we can notice that the proposed approach achieves high results of recall metric up to 0.98 for normal tissue and 0.923 for prostate tumors. In total, the

![Figure 5. Confusion matrix of 10-fold cross-validation results based on the selected features.](image)

|                  | Normal tissue | Prostate tumor |
|------------------|---------------|---------------|
| Normal tissue    | **49**        | **1**         |
| Prostate tumor   | **4**         | **48**        |

**Table 1. Evaluation metrics results of 10-fold cross-validation on microarray prostate cancer dataset.**

|                  | FP rate | Precision | Recall | F1-score |
|------------------|---------|-----------|--------|----------|
| Normal tissue    | 0.077   | 0.925     | 0.980  | 95.1%    |
| Prostate tumor   | 0.020   | 0.980     | 0.923  | 95.0%    |
| Weighted avg.    | 0.048   | 0.953     | 0.951  | 95.1%    |
Figure 6. Results of evaluation metrics obtained from the 10-fold cross-validation technique.

To verify the effectiveness of the proposed approach, we conducted another experiment using all features and compared the accuracy result against the accuracy result of selected features. Figure 7 shows the accuracy results of 10-fold cross-validation to classify normal tissue and prostate tumor for the RC model on the dataset with all features and with selected features.

As shown in Figure 7, we can see a higher accuracy result of using the selected features compared to using all the features. It achieves a remarkable improvement.

In Table 2 and Figure 8, we compare the accuracy result of the proposed approach with the accuracy results of related works based on cross-validation (CV) and holdout testing techniques. We notice that the proposed approach outperforms the related work methods and techniques. The highest accuracy result is highlighted with a boldface font in Table 2.

| Score | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Normal tissue | 0.951 | 0.920 | 0.890 | 0.860 | 0.830 | 0.800 | 0.770 | 0.740 | 0.710 | 0.680 |
| Prostate tumor | 0.920 | 0.890 | 0.860 | 0.830 | 0.800 | 0.770 | 0.740 | 0.710 | 0.680 | 0.650 |
| Weighted Avg. | 0.900 | 0.870 | 0.840 | 0.810 | 0.780 | 0.750 | 0.720 | 0.690 | 0.660 | 0.630 |

Figure 7. Accuracy results of 10-fold cross-validation for the RC model on the dataset with all features and selected features.

weighted average result of the recall (TP rate) metric is 0.951 for the two classes. Furthermore, the approach attains a low output of the FP rate up to 0.077 for normal tissue, 0.020 for prostate tumor, and 0.048 for the weighted average result of the two classes.

Figure 6. Results of evaluation metrics obtained from the 10-fold cross-validation technique.
Table 2. Comparison of accuracy results for the proposed approach and other related work approaches on the same microarray prostate cancer dataset.

| Author (ref.)          | Approach                                           | Accuracy   |
|------------------------|----------------------------------------------------|------------|
| Shen and Tan\textsuperscript{52} | PLR, Monte-Carlo CV (30 iterations)               | 94.6%      |
| Wessels et al.\textsuperscript{53} | RFLD (0), Monte-Carlo CV                           | 93.4%      |
| Glaab et al.\textsuperscript{47}   | BioHEL, 10-fold CV                                 | 94%        |
| Dashtban and Balafar\textsuperscript{50} | SVM, Holdout test (34 samples)                    | 91.2%      |
| Gunavathi and Premalatha\textsuperscript{58} | Genetic algorithm with kNN and SVM, 5-fold CV     | 85.71%     |
| Our approach           | CSF-RC, 10-fold CV                                 | 95.098%    |

From Table 2 and Figure 8, we can see that the proposed approach outperforms the state-of-the-art works in terms of accuracy on the same dataset and using a cross-validation technique.

Conclusion and future work

In this paper, we have proposed an effective approach for prostate cancer classification using microarray gene expression data. It consists of two phases: the feature selection phase and the classification phase. In the process, we used the CFS method for gene selection and the RC model for classification. The experiments are conducted on a public microarray dataset using a 10-fold cross-validation technique.

Experimental results are reported using a set of evaluation metrics, showing the effectiveness of efficiency proposed approach for prostate cancer classification and diagnosis. In addition, the comparison results confirmed the importance of selected features for improving the accuracy result.
against using all features and demonstrated the superiority of the proposed approach against the related works. In future work, we will collect more datasets on microarray gene expression for further improvement of machine learning-based prostate cancer diagnosis. Furthermore, we will conduct a comprehensive comparative study of using the machine and deep learning methods for prostate cancer detection.

Acknowledgements
The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs.

Author contributions
AG provided the idea and conducted the experiment, AG, RS, and MA-R wrote the manuscript. All authors analyzed the results and reviewed the final manuscript.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The authors are grateful to the Deanship of Scientific Research, King Saud University for funding through Vice Deanship of Scientific Research Chairs.

ORCID iD
Abdu Gumaei https://orcid.org/0000-0001-8512-9687

References
1. Li M, Chen W, Jiang D, et al. Immuno-PET imaging of the VEGFR-2 expression in prostate cancer using 89Zr-labeled ramucirumab. *J Nucl Med* 2019; 60(Suppl 1): 1006–1006.
2. Guo L, Lin M, Cheng Z, et al. Identification of key genes and multiple molecular pathways of metastatic process in prostate cancer. *PeerJ* 2019; 7: e7899.
3. He Q, Wan Z, Xu X, et al. Poultry consumption and prostate cancer risk: a meta-analysis. *PeerJ* 2016; 4: e1646.
4. Abel SD, Dadhwal S, Gamble AB, et al. Honey reduces the metastatic characteristics of prostate cancer cell lines by promoting a loss of adhesion. *PeerJ* 2018; 6: e5115.
5. Cuypers M, Lamers RE, Kil PJ, et al. A global, incremental development method for a web-based prostate cancer treatment decision aid and usability testing in a Dutch clinical setting. *Health Informatics J* 2019; 25(3): 701–714.
6. Vidrighin C and Potolea R. ProICET: a cost-sensitive system for prostate cancer data. *Health Informatics J* 2008; 14(4): 297–307.
7. Trevino V, Falciani F and Barrera-Saldaña HA. DNA microarrays: a powerful genomic tool for biomedical and clinical research. *Mol Med* 2007; 13(9): 527–541.
8. Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* 2002; 32(4): 502–508.
9. Chen Z, Gerke T, Bird V, et al. Trends in gene expression profiling for prostate cancer risk assessment: a systematic review. *Biomed Hub* 2017; 2(2): 1–15.
10. Wang Z, Gerstein M and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; 10(1): 57–63.
11. Wolf JB. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. *Mol Ecol Resour* 2013; 13(4): 559–572.
12. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978; 6(2): 461–464.
13. Dagliyan O, Uney-Yuksektepe F, Kavakli IH, et al. Optimization based tumor classification from microarray gene expression data. *PLoS One* 2011; 6(2): e14579.
14. Li Z, Liao B, Cai L, et al. Semi-supervised maximum discriminative local margin for gene selection. *Sci Rep* 2018; 8(1): 1–11.
15. Nirmalakumari K, Rajaguru H and Rajkumar P. Performance analysis of classifiers for colon cancer detection from dimensionality reduced microarray gene data. *Int J Imaging Syst Technol*. Epub ahead of print April 2020. DOI: 10.1002/ima.22431.
16. Raj DD and Mohanasundaram R. An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data. *Arab J Sci Eng* 2020; 45: 2619–2630.
17. Bentkowska U. Optimization problem of k-NN classifier in DNA microarray methods. In: Carter J, Chiclana F, Khuman AS, and Chen T (eds) *Interval-valued methods in classifications and decisions* (pp.107–120). Cham, Switzerland: Springer, 2020.
18. Santhakumar D and Logeswari S. Efficient attribute selection technique for leukaemia prediction using microarray gene data. *Soft Comput* 2020; 24: 14265–14274.
19. Kohavi R and John GH. Wrappers for feature subset selection. *Artif Intell* 1997; 97(1–2): 273–324.
20. Wang Y, Tetko IV, Hall MA, et al. Gene selection from microarray data for cancer classification—a machine learning approach. *Comput Biol Chem* 2005; 29(1): 37–46.
21. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; 286(5439): 531–537.
22. Sha N, Vannucci M, Tadesse MG, et al. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* 2004; 60(3): 812–819.
23. Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001; 7(6): 673–679.
24. Deutsch J. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics* 2003; 19(1): 45–52.
25. Gumaei A, Hassan MM, Hassan MR, et al. A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access* 2019; 7: 36266–36273.
26. Hassan MM, Gumaei A, Aloj G, et al. A smartphone-enabled fall detection framework for elderly people in connected home healthcare. *IEEE Netw* 2019; 33(6): 58–63.
27. Parodi S, Manneschi C, Verda D, et al. Logic learning machine and standard supervised methods for Hodgkin’s lymphoma prognosis using gene expression data and clinical variables. *Health Informatics J* 2018; 24(1): 54–65.
28. Nilashi M, Ibrahim OB, Mardani A, et al. A soft computing approach for diabetes disease classification. *Health Informatics J* 2018; 24(3): 379–393.
29. Gumaei A, El-Zaat A, Hussien M, et al. Breast segmentation using k-means algorithm with a mixture of gamma distributions. In: *2012 symposium on broadband networks and fast internet (RELABIRA)*, Baabda, Lebanon, 2012, pp. 97–102. New York: IEEE.
30. Gumaei A, Hassan MM, Alelaiwi A, et al. A hybrid deep learning model for human activity recognition using multimodal body sensing data. *IEEE Access* 2019; 7: 99152–99160.
31. Budnik M and Krawczyk B. On optimal settings of classification tree ensembles for medical decision support. *Health Informatics J* 2013; 19(1): 3–15.
32. Dudoit S, Fridlyand J and Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* 2002; 97(457): 77–87.
33. Statnikov A, Aliferis CF, Tsamardinos I, et al. A comprehensive evaluation of multiclassification classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 2005; 21(5): 631–643.
34. Furey TS, Cristianini N, Duffy N, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 2000; 16(10): 906–914.
35. Eminaga O, Al-Hamad O, Boegemann M, et al. Combination possibility and deep learning model as clinical decision-aided approach for prostate cancer. *Health Informatics J*. Epub ahead of print June 2019. DOI: 10.1177/1460458219855884.
36. Diaz-Uriarte R and De Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 2006; 7(1): 3.

37. Zeebaree DQ, Haron H and Abdulazeez AM. Gene selection and classification of microarray data using convolutional neural network. In: *2018 international conference on advanced science and engineering (ICOSAE)*, Duhok, Iraq, 9–11 October 2018, pp.145–150. New York IEEE.

38. Aksu Y, Miller DJ, Kesidis G, et al. Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions. *IEEE Trans Neural Netw* 2010; 21(5): 701–717.

39. Mitra S, Saha S and Acharya S. Fusion of stability and multi-objective optimization for solving cancer tissue classification problem. *Expert Syst Appl* 2018; 113: 377–396.

40. Penney KL, Sinnott JA, Fall K, et al. mRNA expression signature of Gleason grade predicts lethal prostate cancer. *J Clin Oncol* 2011; 29(17): 2391.

41. Cuzick J, Swanson GP, Fisher G, et al. Prognostic value of an RNA expression signature derived from cell cycle proliferation genes in patients with prostate cancer: a retrospective study. *Lancet Oncol* 2011; 12(3): 245–255.

42. Erho N, Crisan A, Vergara IA, et al. Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS One* 2013; 8(6): e66855.

43. Mo F, Lin D, Takhar M, et al. Stromal gene expression is predictive for metastatic primary prostate cancer. *Eur Urol* 2018; 73(4): 524–532.

44. Tyekucheva S, Bowden M, Bangoo C, et al. Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nat Commun* 2017; 8(1): 1–10.

45. Sharifi-Noghabi H, Liu Y, Erho N, et al. Deep genomic signature for early metastasis prediction in prostate cancer. *BioRxiv* 2019: 276055.

46. Takeuchi T, Hattori-Kato M, Okuno Y, et al. Prediction of prostate cancer by deep learning with multilayer artificial neural network. *Can Urol Assoc J* 2019; 13(5): E145.

47. Glaab E, Bacardit J, Garibaldi JM, et al. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS One* 2012; 7(7): e39932.

48. Huerta EB, Duval B and Hao JK. A hybrid LDA and genetic algorithm for gene selection and classification of microarray data. *Neurocomputing* 2010; 73(13–15): 2375–2383.

49. Chen H, Zhang Y and Gutman I. A kernel-based clustering method for gene selection with gene expression data. *J Biomed Inform* 2016; 62: 12–20.

50. Bouazza SH, Auhmani K and Zeroual A. Prostate cancer diagnosis based on microarray gene expression profiles. *J Eng Technol* 2018; 6(2): 282–291.

51. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002; 1(2): 203–209.

52. MediaWiki. What is microarray technology?, http://teachercenter.insidecancer.org/icwiki/index.php/What_is_Microarray_Technology%3F (2008, accessed 25 September 2019).

53. Gunavathi C and Premalatha K. Performance analysis of genetic algorithm with kNN and SVM for feature selection in tumor classification. *Int J Comput Electr Autom Control Inform Eng* 2014; 8(8): 1490–1497.