Bulletin of Electrical Engineering and Informatics
Vol. 11, No. 2, April 2022, pp. 1062~1068
ISSN: 2302-9285, DOI: 10.11591/eei.v11i2.3594

Classification improvement of gene expression for bipolar disorder using weighted sparse logistic regression

Abdulnasir Y. Ahmed, Mohammed Abdulrazaq Kahya, Suhaib Abduljabbar Altamir
Department of Computer Sciences, College of Education for Pure Sciences, University of Mosul, Mosul, Iraq

Article Info

Article history:

Received Jan 3, 2022
Revised Mar 6, 2022
Accepted Mar 15, 2022

Keywords:

Bipolar disorder data
Classification
Gene expression data
Gene selection
Logistic regression

ABSTRACT

The computer-aided diagnosis system plays an important role in the classification of diseases and genes such as psychological or other diseases. Bipolar disorder (BD) is a commond psychological disease nowdys. Genes that describe this type of disease may include irrelative values to bipolar disorder disease. These values may adversely impact the classification performance. Logistic regression (LR) and recently sparse logistic regression (SLR) were used as a common technique to solve such binary classification problems. Gene selection has been applied to be a successful technique to get better classification output by excluding the irrelative values of genes. In this work we go further in improving the classification accuracy by restoring to incorporating the weight of these genes utilizing integrating the standardization of T-test with the sparse logistic regression, aiming to accomplish high classification accuracy. A bipolar dataset of gene expressions measured for 22283 genes using Affymetrix technology was used. Two performance indicators; classification accuracy, and geometric-mean of specificity and sensitivity are considered in evaluating the proposed method. Experimental results show an improvement over the two competitor methods; SLR-smoothly clipped absolute deviation (SCAD) and SLR-lasso in three indicators: classification accuracy, geo-means, and area under the curve. Therefore, our technique is beneficial to predict and classify BD psychopaths.

Corresponding Author:

Mohammed Abdulrazaq Kahya
Department of Computer Sciences, College of Education for Pure Sciences, University of Mosul
Mosul, Iraq
Email: mohammedkahya@uomosul.edu.iq

1. INTRODUCTION

Bipolar disorder (BD) is a debilitating mental disease from which approximately 4 to 7% of the US population suffered [1] and affects up to 3% of the worldwide population [2]. BD is characterized by frequent mood alteration between two different situations (mania and depression). Mostly, inheritance plays a major role in BD affection. The environmental factors have their contribution to pathogenesis [1]. BD has a nuclear connection to other illnesses such as schizophrenia or hypomania. There is a difficulty in differentiating BD from other psychosis, especially unipolar disorder depression (UDD), because of the similar syndromes that appear on other psychological patients [2]. With early diagnosis and a good treatment plan, BD patients can be successfully managed. Over the last years, great efforts have been done to identify disease-related genes or biomarkers which lead to early detection and treatment [1]. Correct early diagnosis is very important, that is BD patients who are misdiagnosed within the first years of treatment reach 80%. This lateness has the consequence that some of the patients go to suicide. In addition, incorrect therapy may cost a large amount of money. So, preventing suicide behavior is the great gain of early detection and treatment of mood disorder [2].

DNA microarray technology made a great revolution in the field of biology and genetic research. This technology deals with the expression value of a huge number of genes concurrently. The gene expression data that represent the genes that were elicited from a specific tissue provides useful information for understanding the biological effect and value of that tissue [3], [4]. These gene expression datasets are utilized in various fields of application, such as breast, lung, and prostate cancer classification and early detection of tumors [5], [6].

2. METHOD

2.1. Gene expression

No gene set has been known as the effective gene on BD. In the last years, gene expression using the microarray technique was commonly used for the inspection of complex disorders [2]. Gene expression data is always represented by a matrix; the samples are the matrix rows while the genes are the matrix columns. In these matrices, the number of columns (genes) is always more than the number of rows (samples) [7]–[10].

As the classification performance is highly affected by the many irrelevant and redundant genes contained in these datasets, the reduction of the dimensionality of these datasets emerges as a necessity and it has received increasing attention over the last thirty years [11], [12]. Gene selection methods were applied to reduce the dimensionality of the gene expression datasets [7], [13]. In general, these methods can be divided into filter, wrapper, and embedded methods [14]–[16]. The first is the most common of the three. These methods depend on a special tactic to get the gene information for every gene alone. This tactic is used separately before the classification process and is not subject to the classification method. As for the second group, the optimization of the classification performance is obtained by a process based on the interpretation of the classification algorithm. In the last group, gene selection and data classification merge in one technique concurrently [17]–[20]. Compared to these methods, wrapper methods are considered as having a lower computational efficiency [13], [21]–[23]. Improving classification performance is considered an important goal of gene selection. Gene selection downscales the high dimension microarray by eliminating the unrelated genes which lead to speeding up the classifying process and decreasing the risk of overfitting and increasing computational time.

2.2. Sparse logistic regression

Logistic regression (LR) is one of the most common techniques applied in binary classification problems. The regression function has a nonlinear structure of the variables or genes. The logistic regression response value is either one or zero, always one is used for the positive and zero for the negative cases or healthy [19], [24]. LR is considered unfeasible in gene expression data classification as the gene expression matrix is singular [25].

Recently, much attention is given to sparse logistic regression (SLR) as a classification method where it merges classical LR with a penalty to classify data with the fewest number of features or genes [14], [24], [25]. Different penalties could be imposed to produce different SLR models. The most popular and widely panelized method is the least absolute shrinkage and selection operator, lasso (L1-Norm) [25]. Despite its popularity, SLR-lasso has its drawbacks, first, SLR-lasso selects only one gene and ignores the rest from a group of genes even if there are high correlations among them, second, the same magnitude of shrinkage to each gene coefficient is used, which leads to inconsistently gene selection.

Let n represents the independent observations $\{y_i, t_i; i = 1,2,\dots,n\}$ where $y_i \in \{1,0\}$ are response values (class labels), and $t_i = \left(t_{i1},\dots,t_{ip}\right)^T$ is a vector of predictor genes & p represents the number of genes.

Consequently, the LR formula is explained as:

$$Pr\ o\ b(y_i = 1 : t_i) = \psi(t_i), \quad (1)$$

$$Pr\ o\ b(y_i = 0 : t_i) = 1 - \psi(t_i), \quad (2)$$

This likelihood can be simplified as:

$$\psi(x_i) = \frac{exp(\alpha + t_i^T \beta)}{(1 + exp(\alpha + t_i^T \beta))}, \quad (3)$$

where $\alpha + t_i^T \beta = \alpha + t_{i1}\beta_1 + t_{i2}\beta_2 + \dots + t_{ip}\beta_p$, and $ln\left[\frac{\psi(t_i)}{(1-\psi(t_i))}\right] = \alpha + t_i^T \beta$.

The function of log-likelihood concerning response variables y_i can be clarified as:

$$\Gamma(\beta) = \sum_{i=1}^{n}[y_i(\alpha + t_i^T\beta) - ln\{1 + exp(\alpha + t_i^T\beta)\}], \quad (4)$$

The SLR-lasso imposes a nonnegative penalty term to (4) to enable controlling the number of the genes in high dimension and is defined as:

$$\Gamma(\beta) = \sum_{i=1}^{n}\left[y_i(\alpha + t_i^T\beta) - ln\{1 + exp(\alpha + t_i^T\beta)\} - \lambda\sum_{j=1}^{p}|\beta_j|\right], \quad (5)$$

$$\hat{\beta}WSLR = \underset{\beta}{\text{argMax}}\{\sum_{i=1}^{n}[y_i(\alpha + t_i^T\beta)] - ln\{1 + exp(\alpha + t_i^T\beta)\} - \lambda(\beta)\}, \quad (6)$$

As shown in (5) [5] if we supposed the genes are standardized, then the values of the parameters $\alpha\&\beta$ are got by maximizing the SLR-lasso as shown in (6). This equation $\lambda(\beta)$ represents the penalty term that improves the estimation. This term controls the degree of shrinkage. The $\lambda = 0$ value leads to the maximum probability method solution while a large value λ increases the impact of the penalty term on the coefficient evaluations [19], [21], [26].

2.3. Weighted sparse logistic regression

The information obtained from the measurements of genes has a great role in improving gene selection. This paper suggests a weighting scheme that integrates the standardization of T-test (S-Ttest) into WSLR to get better performance of the identification of the related genes. The measurements related to genes that have incorporeal correlation are based on finding the weights so that they can make a better representation for the differential expressions among genes in the data. The main purpose of combining weights is to underline the differential expression of genes and it permits to make good detecting of these genes via the process of gene selection utilizing sparse logistic regression. The S-Ttest weights are calculated as in (7):

$$W_j = \left|\frac{(|T_j| - \mu)}{\sigma}\right|, \quad (7)$$

where $|T_j|$ represents the jth gene value provided by the two-sample T-test. In this paper, the weight indicates the expression difference of genes in samples. Therefore, the most differentially expressed genes own the largest weight and will achieve the highest degree of accuracy in the classification. The classification of these weights is as shown in Figure 1.

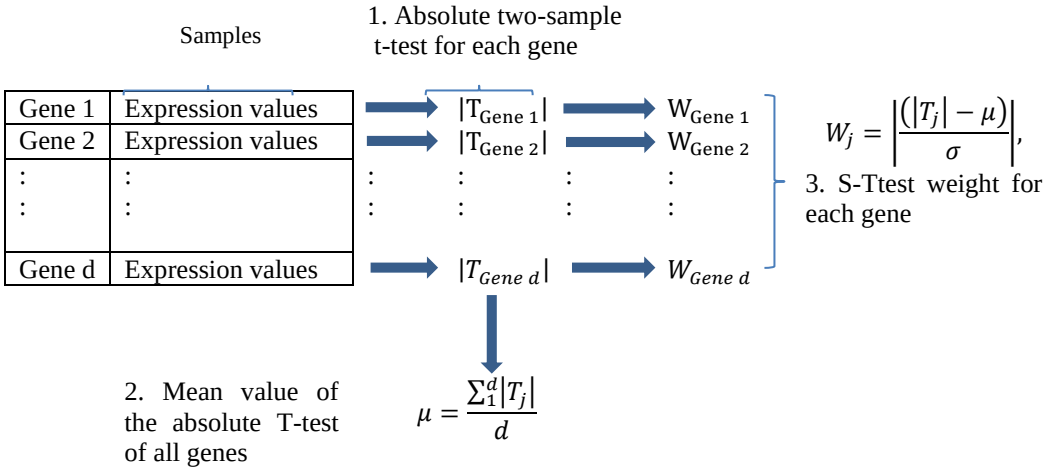

Figure 1. Example of S-Ttest weight calculations of genes

As illustrated in the figure, the weight for each gene is computed by subtracting the mean value from the T-test value for the gene, then dividing the result by the standard deviation. Then a filtration is done to exclude genes that do not have a differential expression between the two groups according to the p-values

greater than (0.05) generated during the two-sample test. Then each of the remaining genes is multiplied by the calculated weights. The result is the weighted gene expression for the microarray as shown in Figure 2.

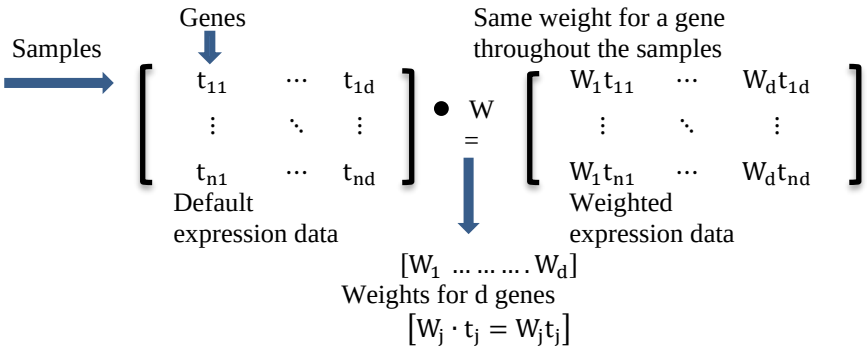

Figure 2. Processing the gene expression for bipolar disorder data with S-Ttest weight

2.4. Experimental setting

The bipolar dataset [27] consists of 30 bipolar disorder and 31 controls. Gene expressions are measured for 22283 genes using Affymetrix technology. To show the improvement of the proposed method over other methods, experiments were done and compared with SLR-lasso and the SLR-SCAD. Initially, the dataset was arbitrarily split to train and test data, where 70% of the samples were chosen for the training group and 30% for the testing group.

For alleviating the effect of the partitioning and for a fair comparison, the classification accuracy was used as an evaluation indicator for all the competitor methods by finding the average up to 20 partitioned times using 10 folds cross-validation for the training group in each time. The parameter value λ was adjusted to basing on the training group while, as for the SCAD penalty the constant value was adjusted to 3.7 as Fan and Li (2001) proposed [22].

2.5. Performance evaluation criteria

Two performance indicators; classification accuracy (CA), and geometric-mean of specificity and sensitivity (geo-mean) are considered in evaluating the proposed method. The CA represents the exact percentage of the classified psychopath and healthy persons; this measure is used to evaluate the classifier power. This indicator can be calculated as:

$$CA = \frac{(TN+TP)}{(TP+TN+FP+FN)} \times 100\% \quad (8)$$

where: TP is no. of true-positive, TN is no. of true-negative, FP is no. of false-positive, FN is no. of false negative. Maximizing the accuracy in both categories of humans is considered as a goal of the common classification methods. The second metric Geo-means has been proposed to show the united performance of specificity and sensitivity, and can be defined as:

$$Geo - mean = Sqrt(spesificity \times sensitivity), \quad (9)$$

where specificity (TN rate) is the proportion of properly classified healthy humans, and sensitivity (TP rate) is the proportion of successfully classified psychopath humans [25].

2.6. Bipolar disorder gene expression data classification

Gene expression data preprocessed (weighted gene matrix) will be used with the SLR to get more accuracy in classification than traditional methods. The algorithm of the weighted SLR (WSLR) calculation is as:

```
Algorithm 1:
Applying a two-sample T-test to extract significant genes.
Find wj (Equation (7)), (Figure1.).
Define t̃i = Wj.ti (Figure 2.)
Applying β̂WSLR (Equation (6))
```

3. RESULTS AND DISCUSSION

3.1. Classification indicator

The improvements in the classification accuracy of our proposed technique and the results of SLR-SCAD, and the SLR-lasso are summarized in Table 1. The table also shows the number of selected genes and the Geo-mean for the training data set. Each value in the table is attached to the corresponding standard deviation. As for classification accuracy and depending on the training group, the WSLR yields 91.936 defeating SLR-SCAD by 6.08% and SLR-lasso by 12.935%.

Table 1. Classification indicators of the proposed technique and its rivals

Techniques	#Genes	Train CA	Geo-mean	Test CA
WSLR	22	91.936 (1.878)	0.895 (0.034)	86.817 (3.416)
SLR-SCAD	29	86.345 (2.756)	0.846 (0.009)	80.115 (1.459)
SLR-lasso	35	80.044 (4.818)	0.797 (0.003)	76.887 (3.309)

The geo-mean of the WSLR achieved 0.895, which shows that WSLR has a clear distinction between healthy individuals and individuals with BP. In addition, it is shown that SLR-SCAD yields a value of 86.345, which is better than that for SLR-lasso. This result is expectant because the SLR-SCAD has high consistency selector efficiency. Moreover, WSLR yields better results than the other techniques regarding classification accuracy as it yields 86.817 which is 7.7197% and 11.438% better than SLR-SCAD and SLR-lasso respectively. The superiority above is related to the number of genes. Our approach has made progress in gene selection, where it selects 22 genes versus 29, and 35 for SLR-SCAD and SLR-lasso respectively. Overall, the classification indicator of the proposed technique has obtained the best classification performance compared to SLR-SCAD and SLR-lasso. That indicates that the proposed technique is useful to give us more information about genes' influence on bipolar disorder.

3.2. Classification indicator

To emphasize how classification performance is affected by WSLR in selecting the most important relevant genes, a pairwise comparison was done between WSLR and other competitor techniques using T-test results considering the area under the curve (AUC) of the train data. Figure 3 presents the AUC boxplot of our proposed technique and its rivals. It was clear that the AUC of our proposed technique is better than the results of the competing techniques. Table 2 reports the test result at significance level α=0.05. It's clear our WSLR technique statistically records better significance than those of SLR-SCAD and SLR-lasso

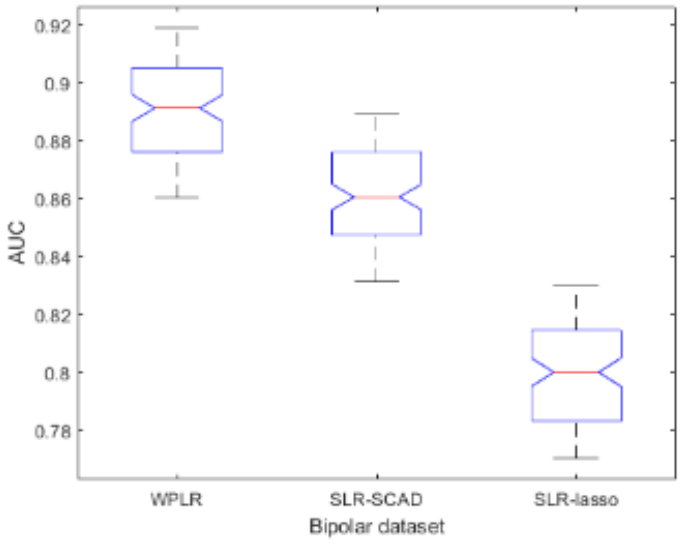

Figure 3. AUC boxplot for the bipolar-disorder dataset achieved using (weighted/non-weighted) sparse logistic regression

Table 2. P-values for the t-test of the proposed technique and its rivals

Dataset	WSLR vs SLR-SCAD	WSLR vs SLR-lasso
Bipolar-disorder	0.0029 (*)	0.0009 (*)

(*) significant differences

4. CONCLUSION

In this work, sparse logistic regression has been used after data weight. The main objective of data weight is to identify the relevant genes in bipolar disorder data and detect genes' influence on bipolar disorder disease. The proposed technique is WSLR, it has beaten clearly in the results of classification and genes selection to boost the used technique. Three comparative aspects: high classification accuracy, Geo-mean, and AUC were considered to show the classification performance efficiency of WSLR. Taking on these three aspects concurrently puts on WSLR as a favorable gene selection method. Overall, WSLR presents its utility and applicability in other fields of huge data classification related to the psychological diseases domain and other fields such as features selection in pattern recognition or variable selection in the statistical domain.

REFERENCES

[1] E. Kyriacou, S. Christofides, and C. S. Pattichis, "Gene Expression Data Analysis for Classification of Bipolar Disorders," *XIV mediterranean conference on medical and biological engineering and computing 2016*, vol 57, pp. 500–506, 2016, doi: 10.1007/978-3-319-32703-7_97.

[2] J. R. De Jesus, B. K. De Campos, R. M. Galazzi, J. L. C. Martinez, and M. A. Z. Arruda, "Bipolar disorder: Recent advances and future trends in bioanalytical developments for biomarker discovery," *Anal. Bioanal. Chem.*, vol. 407, no. 3, pp. 661–667, 2015, doi: 10.1007/s00216-014-8341-3.

[3] J. Apolloni, G. Leguizamón, and E. Alba, "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments," *Appl. Soft Comput.*, vol. 38, no. c, pp. 922–932, 2016, doi: 10.1016/j.asoc.2015.10.037.

[4] Z. Y. Algamal and M. H. Lee, "Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification," *Expert Syst. Appl.*, vol. 42, no. 23, pp. 9326–9332, 2015, doi: 10.1016/j.eswa.2015.08.016.

[5] M. A. Kahya, "Classification enhancement of breast cancer histopathological image using penalized logistic regression," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 1, pp. 405–410, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp405-410.

[6] R. Maulana, A. A. Baharin, and H. Fitriyah, "Classification of lung condition for early diagnosis of pneumonia and tuberculosis based on embedded system," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 1262–1270, 2021, doi: 10.11591/eei.v10i3.3033.

[7] Z. Y. Algamal and M. H. Lee, "Penalized Poisson Regression Model using adaptive modified Elastic Net Penalty," *Electron. J. Appl. Stat. Anal.*, vol. 8, no. 2, pp. 236–245, 2015, doi: 10.1285/i20705948v8n2p236.

[8] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Comput. Biol. Med.*, vol. 66, pp. 1–10, 2015, doi: 10.1016/j.compbiomed.2015.08.010.

[9] J. Kalina, "Classification methods for high-dimensional genetic data," *Biocybern. Biomed. Eng.*, vol. 34, no. 1, pp. 10–18, 2014, doi: 10.1016/j.bbe.2013.09.007.

[10] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics," *Brief. Bioinform.*, vol. 9, no. 5, pp. 392–403, 2008, doi: 10.1093/bib/bbn027.

[11] M. Y. Park and T. Hastie, "Penalized logistic regression for detecting gene interactions," *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2007, doi: 10.1093/biostatistics/kxm010.

[12] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification," *Appl. Intell.*, vol. 49, no. 4, pp. 1245–1259, 2019, doi: 10.1007/s10489-018-1320-1.

[13] S. Mishra and D. Mishra, "SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm," *Karbala Int. J. Mod. Sci.*, vol. 1, no. 2, pp. 86–96, 2015, doi: 10.1016/j.kijoms.2015.10.002

[14] O. S. Qasim and Z. Y. Algamal, "Feature selection using particle swarm optimization-based logistic regression model," *Chemom. Intell. Lab. Syst.*, vol. 182, pp. 41–46, 2018, doi: 10.1016/j.chemolab.2018.08.016.

[15] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "An ensemble of filters and classifiers for microarray data classification," Pattern Recognit., vol. 45, no. 1, pp. 531–539, 2012, doi: 10.1016/j.patcog.2011.06.006.

[16] A. J. Ferreira and M. A. T. Figueiredo, "Efficient feature selection filters for high-dimensional data," *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1794–1804, 2012, doi: 10.1016/j.patrec.2012.05.019.

[17] Z. Algamal, "Variable Selection in Count Data Regression Model based on Firefly Algorithm," *Stat. Optim. Inf. Comput.*, vol. 7, no. 2, pp. 520–529, 2019, doi: 10.19139/soic.v7i2.566.

[18] Q. Mai and H. Zou, "The Kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, vol. 100, no. 1, pp. 229–234, 2012, doi: 10.1093/biomet/ass062.

[19] Z. Y. Algamal, "Classification of gene expression autism data based on adaptive penalized logistic regression," *Electron. J. Appl. Stat. Anal.*, vol. 10, no. 2, pp. 561–571, 2017, doi: 10.1285/i20705948v10n2p561.

[20] S. M. Ayyad, A. I. Saleh, and L. M. Labib, "Gene expression cancer classification using modified K-Nearest Neighbors technique," *Biosystems*, vol. 176, pp. 41–51, 2019, doi: 10.1016/j.biosystems.2018.12.009.

[21] Z. Y. Algamal and M. H. Lee, "A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression," J. Chemom., vol. 31, no. 10, pp. 1–11, 2017, doi: 10.1002/cem.2915.

[22] Z. Y. Algamal and H. T. Mohammad Ali, "An efficient gene selection method for high-dimensional microarray data based on sparse logistic regression," *Electron. J. Appl. Stat. Anal.*, vol. 10, no. 1, pp. 242–256, 2017, doi: 10.1285/i20705948v10n1p242.

[23] S. P. Potharaju and M. Sreedevi, "Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance," *Clin. Epidemiol. Glob. Heal.*, vol. 7, no. 2, pp. 171–176, 2019, doi: 10.1016/j.cegh.2018.04.001.

[24] S. J. Shappot and H. M. Gorgees, "Use of Logistic Regression Approach to Determine the Effective Factors Causing Renal Failure Disease," Ibn AL-Haitham J. Pure Appl. Sci., vol. 31, no. 3, pp. 143–150, 2018, doi: 10.30526/31.3.2007.

[25] Z. Y. Algamal and M. H. Lee, "A two-stage sparse logistic regression for optimal gene selection in high-dimensional microarray data classification," *Adv. Data Anal. Classif.*, vol. 13, no. 3, pp. 753-771, 2018, doi: 10.1007/s11634-018-0334-1.
[26] Z. Y. Algamal, M. K. Qasim, and H. T. M. Ali, "A QSAR classification model for neuraminidase inhibitors of influenza A viruses (H1N1) based on weighted penalized support vector machine," SAR QSAR Environ. Res., vol. 28, no. 5, pp. 415–426, 2017, doi: 10.1080/1062936X.2017.1326402.
[27] M. M. Ryan, H. E. Lockstone, S. J. Huffaker, M. T. Wayland, M. J. Webster, and S. Bahn, "Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes," *Mol. Psychiatry*, vol. 11, no. 10, pp. 965–978, 2006, doi: 10.1038/sj.mp.4001875.

BIOGRAPHIES OF AUTHORS

Abdulnasir Y. Ahmed Bachelor of Computer Science from Computer Science and Mathematics University of Mosul, Iraq, Master of Computer Science, College of Computer Science and Mathematics, University of Mosul, Iraq. And Ph.D. in Operating System and Distributed Computing, College of Computer Science and Mathematics, Iraq. He was a lecturer in a management technical institute. Nineveh, A lecturer in Al-Hadba'a University College in Mosul, Iraq, and in the Computer Department, College of Education, Salahaddin University. Currently, he is a lecturer in the Department of Computer Science, College of education for Sciences, University of Mosul. His research interests are parallel programming, distributed computing, cloud computing, machine learning, and image processing. He can be contacted at email: abdulnasir.younus@uomosul.edu.iq.

Mohammed Abdulrazaq Kahya received the Bachelor's degree in Mathematics Education from the University of Mosul in 2004, Master of Mathematics in Numerical Analysis, College of the Education for Pure Sciences University of Mosul in 2006, Iraq, and Ph.D. in Computational Mathematics, College of Computer Sciences and Mathematics University of Mosul in 2017, Iraq. Currently, he is an Assistant Professor in the Computer Sciences Department at the University of Mosul. His research interests are computational mathematics, mathematical programing, bioinformatics, and numerical analysis. He can be contacted at email: mohammedkahya@uomosul.edu.iq.

Suhaib Abduljabbar Altamir received Bachelor of Mathematics- Computer in University of Mosul College of Education for Pure Sciences Department of mathematics Science Mosul, Nineveh, Iraq in 1996, He received the Master degree of Numerical Analysis in University of Mosul College of Education for Pure Sciences Department of mathematics Science Mosul, Nineveh, Iraq in 2003, He received the Ph.D. of Mathematics/Computational-Computer technologies in University of Mosul College of mathematics and Computer Science Department of mathematics Science Mosul, Nineveh, Iraq in 2018, Currently He is Lecturer in Department of Computer Sciences College of Education for Pure Sciences, University of Mosul. He can be contacted at email: suhaib.altamir@uomosul.edu.iq.