Classification of gene expression dataset for type 1 diabetes using machine learning methods

Noor AlRefaai, Sura Zaki AlRashid
Department of Software, College of Information Technology, University of Babylon, Babylon, Iraq

Article Info

Article history:

Received Jun 28, 2022
Revised Jul 15, 2022
Accepted Nov 27, 2022

Keywords:

Classification
Gene expression
Naive Bayes
Random forest
Support vector machine
Type 1 diabetes

ABSTRACT

Type 1 diabetes (T1D) disease is considered one of the most prevalent chronic diseases in the world, it causes a high level of glucose in the human blood. Despite the seriousness of this disease, T1D may affect people and their condition develops to an advanced stage without feeling it, which makes it difficult to control the disease. Early prediction of the presence of this disease may significantly reduce its risks. There are many attempts to overcome this disease, some of them are heading towards biological solutions and others towards bioinformatic solutions. Several studies have used a single feature selection method with a machine learning (ML) model to predict or classify T1D. In this paper, ML techniques were used for classification, such as Naive Bayes (NB), support vector machine (SVM), and random forest (RF) using a T1D gene expression dataset that has a multiclass to classify the genes associated with this disease. The proposed model can identify the genes related to T1D with high efficiency, which helps a lot in predicting a person carrying the disease before symptoms appear. The highest accuracy of 89.1% was obtained when applying SVM with chi^2 as the feature selection method.

Corresponding Author:

Noor AlRefaai
Department of Software, College of Information Technology, University of Babylon
Babylon, Iraq
Email: noorali17790@gmail.com

1. INTRODUCTION

High blood glucose is one of the most important features of diabetes, which is caused by a defect in insulin secretion or insulin action, or both. One of the complications of high blood glucose is an imbalance in the body’s functions and failure of a number of organs, and the disease becomes long-term [1]. Signs of high blood glucose include constant thirst, excessive urination, and excessive hunger [2]. Diabetes disease included a three types: type 1 diabetes (T1D), type 2 diabetes (T2D), and gestational diabetes [1]. Gene expression offers the possibility of diagnosing different diseases, by using precise deoxyribo nucleic acid (DNA) arrays containing powerful gene expression data acquisition technology. The level is interpreted as a mixture of different messenger molecule ribonukleat acid (RNA) in the cell and using it can identify good treatments and detect diseases early as well as detect mutations [3]. In many different sectors, research has been conducted to improve the quality of care for patients with diabetes and reduce its effects, including artificial intelligence (AI) and machine learning (ML). In several studies, ML models used to predict and classify diabetes have been mentioned [4].

According to Deberneh and Kim [5], the gene expression dataset (GSE55098) for T1D was used to classify the immune cell-related genes that have a role in the occurrence and development of T1D and possibly contribute to finding an immunotherapy for it. Hub genes were investigated using least absolute shrinkage and selection factor (LASSO) and support vector machines (SVM). Gene prognostication in T1D was evaluated using receptor operating characteristic curves and the highest value was receiver operating characteristic area under the curve (ROC) AUC=0.918 for the NCR3 immune-related genes.

According to Roy *et al.* [6], ML models were developed to classify two categories of cancer between early and late stages. The RNA-seq gene expression dataset was used for invasive ductal carcinoma, as it contains different stages of the disease for 610 patients. The models used are random forest (RF), SVM, logistic regression (LR), decision tree (DT), and Naive Bayes (NB), where the highest accuracy was obtained by RF 0.95. The resulting analysis provides insight into the events associated with the progression of this disease.

According to Ma and Zheng [7], three datasets were analyzed to determine the causes of beta-cell dysfunction and T2D deficiency. In order to distinguish between healthy cells and cells with T2D, attention and focus were given to the levels of expression of essential genes. To obtain the influencing factors, the mutual information (MI) and correlation coefficients were calculated. Five ML classifiers were used: bayesian network, SVM, RF, neural network (NN), and LR. The highest accuracy (ACC) obtained for RF classifier is 0.907.

According to Kazerouni *et al*. [8], six lncRNA expression for T2D was used with four classification ML models, including SVM, nearest neighbors (KNN), LR, and artificial neural networks (ANN) to diagnose T2D. These algorithms were compared with each other in terms of diagnostic accuracy. The best AUC had SVM and LR among the classification methods with an average AUC of 95%.

Research by Alshamlan *et al.* [9], two datasets (GSE38642) and (GSE13760) of T2D gene expression data were used from the gene expression omnibus (GEO) database. The feature selection methods used were fisher score and chi-squar. By these methods, a subset of genes was obtained, numbering between (1,800–2,700). LR and SVM classifiers ML models were applied on the subset of genes for predicting genes that cause T2D. The highest accuracy was 90.23% for the LR model when it was used on the Fisher Score method, and the same model on chi-square accuracy was 88.81%, while SVM classification did not produce satisfactory results.

Research by Li *et al*. [10], the RNA-seq dataset (GSE164416), which contains T2D samples and healthy samples, was used to discover biomarkers that directly affect the diagnosis of T2D and early prediction of its risk. The validity of the biomarkers was validated using SVM ML model. The identification (ID) of patients with T2D was evaluated with a sensitivity and specificity of 100%.

Research by Lee and Lee [11], blood gene expression datasets (GSE63060) and (GSE63061) were used for prediction purpose. Five method of ML models were used to predict the Alzheimer’s disease: LR, L1-regularized LR (L1-LR), SVM, RF, and deep neural network (DNN). Variational autoencoder (VAE) method was used for feature selection. For the two dataset, the best average values of the AUC were 0.874 and 0.804, respectively. In this paper, classifying for gene expression dataset of T1D using ML methods has been proposed. It is totally presented as follows: section 1 is the introduction, section 2 is the proposed method, section 3 is the method, section 4 is the results and discussion, and section 5 is the conclusion.

2. THE PROPOSED METHOD

The proposed system implemented for classifying the gene expression dataset of T1D. Classification using machine learning models may avoid major problems for patients with T1D. The system was executed using four main stages. The stages are explained in detail next subsection:

2.1. Preprocessing stage

In machine learning models, there is a necessity for the suitability of the dataset to the business requirements as the initial stage. The pre-processing stage includes many steps, such as handling missing values and data cleaning. Some of which will be discussed below [12].

2.1.1. Missing value

Gene expression datasets that are highly dimensional contain a huge number of features. It may include missing values for some of the features [12]. There are several ways to solve the missing values problem such as guessing the missing values, ignoring the missing values, and removing the data objects [13].

2.1.2. Normalization

Normalization is one of the initial processing steps for processing the dataset that is applied before the data is used, such as increasing or decreasing the range of values. Normalization is convenient and useful in dataset problems that depend on classification, the conversion of feature values for a specific and small range such as 0 to 1. There are many ways to normalize, such as z-normalization and min-max normalization. Min-max normalization is a linear transformation technique used in processors in which preserving the relationship between the original dataset is important. In addition, it is considered one of the simple techniques that are suitable for the dataset within predefined limits [14], [15]. Normalization is done according to (1):

$$x\ new = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

where $x\ new$ represent normalized $x.$

2.2. Ranking stage

Student’s t-test was used for ranking genes, it is a statistical test of the parametric type. This test is used in the case of comparing the mean of two sets of data, the formula of the test as in (2):

$$t = \frac{\tilde{x} - \mu}{s/\sqrt{n}} \tag{2}$$

Where $\tilde{x}$ is the mean of sample, μ is the mean of expected population, and s is the variance of estimated population [16], [17].

2.3. Feature selection stage

Feature selection stage is performed before the classification stage to effectively reduce the data. Feature selection methods are used as a pre-processing step for selecting the related features (genes). This stage helps to reduce the execution time and increase the classification accuracy [18].

2.3.1. Chi-square

Chi^2 is a non-parametric statistical method of analysis data. Using (3), the value of the chi-square is calculated and the best set of features is selected.

$$x^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei} \tag{3}$$

Where Oi is observed value, Ei is expected value and k is a number of classes [4].

2.3.2. Analysis of variance (ANOVA)

ANOVA is also known as the F statistic, it is used for a dataset that contains multiple categories in case the mean values contain a large difference between them [19]. ANOVA is a technique used to reduce the dimensions of datasets that contain huge numbers of features. In the result, the dataset can be expressed with the fewest number of variables [20].

2.3.3. Mutual information

For each original random variable that contains information about another random variable, the exchanged information measures the amount of this information and is also considered to reduce the uncertainty between the original variable in relation to the other variable. The use of MI to select features is done by selecting a subset of features n from the dataset X that includes all the features N. The value of the MI for this subset is greater with the category [21].

2.3.4. Principal of component analysis (PCA)

PCA is one of the linear transformation techniques that used for reducing high-dimensional datasets. It creates basic components for the input features that have been converted from correlated features to unconnected featuers [22]. As the resulting, the reduced data contains fewer unrelated features [20].

2.3.5. K-means clustering

One of the aggregation algorithms, where a single object is assigned to one set of aggregates. Using an objective function value, the quality of each group is measured. Where k is considered as centers of groups and are initially empty, then each object is assigned to a group according to the closest distance between them in the end, a number of classes is obtained as the number of k [23].

2.4. Machine learning stage

The machine learning models have the ability to solve problems within many domains and facilitate dealing with data. Some of problems such as prediction or detection, are solved using the classification and the regression models. The learning of the ML models are supervised or unsupervised, this indicated accordeing to the type of the problem [23].

2.4.1. Random forest

RF is a set of decision trees, that assembled after averaging the prediction for each tree in the forest [24]. RF is considered one of the useful methods used recently in biological studies, due to the simplicity and flexibility with the presence of variables in large numbers. As well as it determines for each variable used its role in responding to the prediction. It is noteworthy that it provides high accuracy and interpretability [25].

2.4.2. Support vector machine

A supervised ML algorithm is used for solving classification and regression problems, which is a supervised classifier. Classification is commonly used in many applications by allocating the points of the dataset by the hyperlevel as a limit to the classification decision. Since there is a maximum margin between the hyperplane and the classes, the data is sorted using the hyperplane [2].

2.4.3. Naïve Bayes

A supervised ML method, it is most commonly used with datasets that include classification problems, due to its accuracy in classification results. NB is a classifier that depends on probability in its classification, where for each class in the dataset guesses its probability in the prediction. A classifier that based on learning using training data and then predicting a class for the test record that has a high subsequent probability [26], [27].

3. METHOD

Gene expression dataset of T1D was used in this work. The proposed model shows in Figure 1, includes five stages: the pre-processing as a first stage, which consists of missing values and normalization, second stage is ranking for the features using the student's t-test. Then a subset of features were selected according to the feature selection methods. Then, ML models were used to classify dataset. Finally, the proposed system was evaluated using accuracy metric.

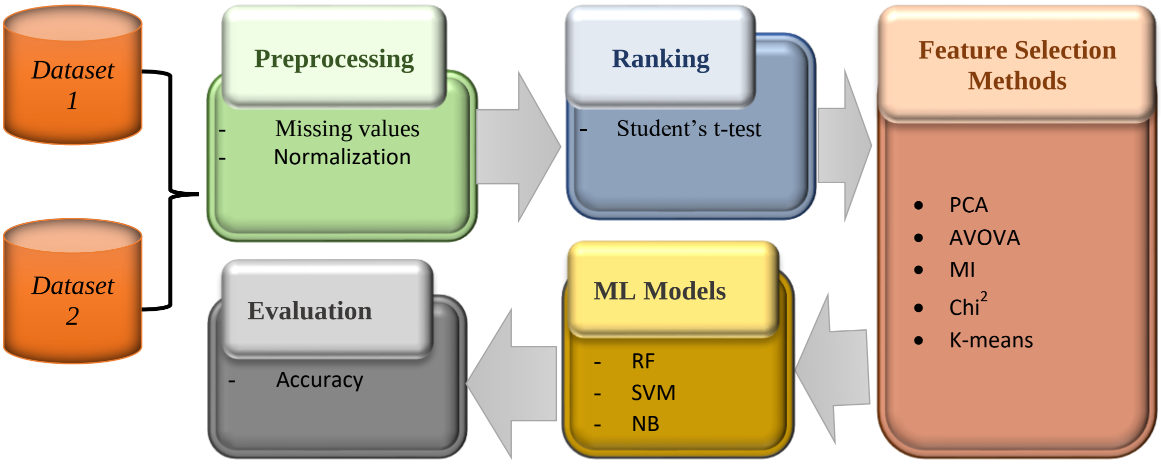


Figure 1. Block diagram of main stages for proposed model

4. RESULTS AND DISCUSSION

4.1. Dataset

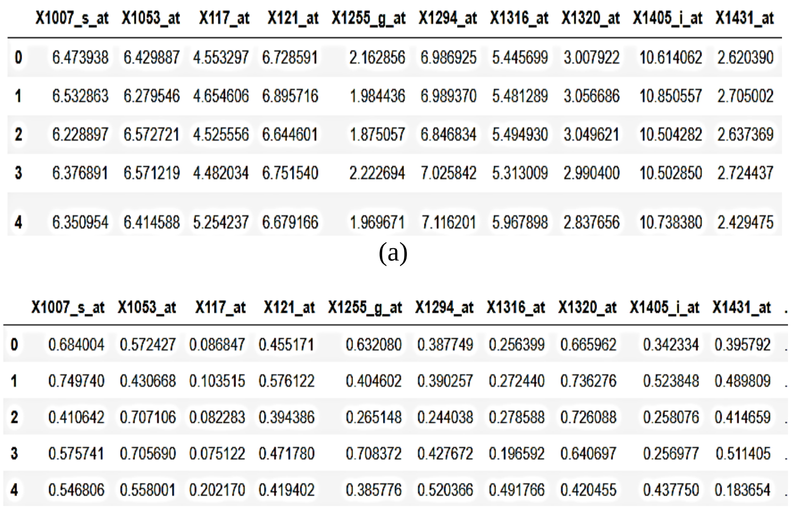
Gene expression dataset was used can be accessed from national center for biotechnology information (NCBI) GEO database [28]. Microarray technology the coefficients in genes at the same time. Gene expression data for T1D were used in the proposed system [29]. The values obtained by mapping the RMA algorithm using Limma package in R language. In this work, the raw data was collected according to what was used in the research [30] the clarification as follows: only the samples on which the experiment (longitudinal) on auto-antibody-negative (AA-) high HLA risk siblings (60) and on low HLA risk siblings (31) has been taken from GSE52724 and concatenated it with unrelated healthy control plasma (44) and recent onset T1D plasma (46) from GSE35725 in new dataset of 181 sample and 54675 genes. Two datasets were used the first dataset with code GSE35725 is downloaded in raw file. It has 114 sample, the samples ID between GSM874033 to GSM874146. The second dataset with code GSE52724. It has 286 samples; the samples ID are between GSM1274585 to GSM1274870.

4.2. Pre-processing

According to the data that was used in this work, there was a need to make two steps as pre-processing to facilitate the process of implementing the model. The first step was handling missing values, the raw data contained a set of genes that did not contain values or NaN. Therefore, (4) calculates the mean is used to be the estimated value of the missing values of the corresponding column that contains values for the same gene for all remaining samples.

$$Mean = \frac{\sum_{i=1}^{N} Xi}{N} \tag{4}$$

Where X represent value of data and N represent number of data values in column [13]. The other step was Min-Max normalization; it was applied according to (1) and the Figure 2(a) shows the dataset before normalization, and Figure 2(b) shows the dataset after normalization.



	X1007_s_at	X1053_at	X117_at	X121_at	X1255_g_at	X1294_at	X1316_at	X1320_at	X1405_i_at	X1431_at
0	6.473938	6.429887	4.553297	6.728591	2.162856	6.986925	5.445699	3.007922	10.614062	2.620390
1	6.532863	6.279546	4.654606	6.895716	1.984436	6.989370	5.481289	3.056686	10.850557	2.705002
2	6.228897	6.572721	4.525556	6.644601	1.875057	6.846834	5.494930	3.049621	10.504282	2.637369
3	6.376891	6.571219	4.482034	6.751540	2.222694	7.025842	5.313009	2.990400	10.502850	2.724437
4	6.350954	6.414588	5.254237	6.679166	1.969671	7.116201	5.967898	2.837656	10.738380	2.429475

(a)

	X1007_s_at	X1053_at	X117_at	X121_at	X1255_g_at	X1294_at	X1316_at	X1320_at	X1405_i_at	X1431_at	.
0	0.684004	0.572427	0.086847	0.455171	0.632080	0.387749	0.256399	0.665962	0.342334	0.395792	.
1	0.749740	0.430668	0.103515	0.576122	0.404602	0.390257	0.272440	0.736276	0.523848	0.489809	.
2	0.410642	0.707106	0.082283	0.394386	0.265148	0.244038	0.278588	0.726088	0.258076	0.414659	.
3	0.575741	0.705690	0.075122	0.471780	0.708372	0.427672	0.196592	0.640697	0.256977	0.511405	.
4	0.546806	0.558001	0.202170	0.419402	0.385776	0.520366	0.491766	0.420455	0.437750	0.183654	.

(b)

Figure 2. Data min-max normalization (a) before normalization and (b) after normalization

After the two processes above, the result was a dataset has the same original numbers of genes and samples. The difference was in the data values only, where the missing values had compensated, and the data had transformed to the required range.

4.3. Ranking

Student's t-test was used for ranking genes, the number of genes required for all samples was pre-determined. The number of genes has been reduced from 54,675 to 10,000. The resulting dataset after applying the ranking contained 10,000 genes and 181 samples.

4.4. Feature selection

The gene expression data are characterized by high dimensions; due to a large number of genes. The feature selection methods provided the ability to select the most important genes and those most closely related to the disease and neglect the redundant genes. Therefore, these methods had used to select a subset of the dataset that includes the most necessary genes associated with T1D disease. The methods that had used in this paper explained with their results in the following paragraphs:

a. MI: genes with a value of MI greater than 0.05 were selected and 7,542 genes out of 10,000 genes were selected as informational and important features.

b. Chi-square: the Chi^2 test was applied to all dataset with the number of genes of 10,000 genes and 8415 genes were selected according to the threshold of 0.5.

c. ANOVA: the number of features (genes) was reduced from 10,000 to 8583 after using ANOVA and the p-value was 0.05.

d. Principle component analysis (PCA): the dataset was entered in the form of an array with a size of 10,000 genes with 181 samples. After applying PCA to it, the dimensions were reduced to 181 genes and with 181 samples.

e. K-means clustering: the k-means algorithm requires defining a predetermined number of clusters before applying it to the data. Therefore, in this work the elbow method was used to estimate the appropriate number of clusters, this method estimated that four clusters are the appropriate number of clusters, the number of genes reduced from 10000 to 8005. The result is either entered to another feature selection method or to a classification model. Table 1 illustrate the result after applying feature selection methods on 10,000 genes.

4.5. Machine learning models

ML models were developed to process gene expression data to classify many different diseases. RF, SVM and NB models were used as classification models for T1D. When implementing RF, the dataset was divided into two sets, a training set and a test set. The dividing percent of 80% for training set and 20% for testing set. The number of trees used was determined by 100 trees of the classification model. The result of RF model showed the data obtained from chi-square produced the accuracy of 86.4%. The other ML model was linear kernel SVM, when classify the subset of data selected by the chi-square produce accuracy of 89.1%. Finally, gaussian NB model was used in the observation and the dataset was divided into 80% training set and 20% test set. All the results are summarized in Table 2. The work was compared with some related work in which different datasets for T1D and T2D were used for classification and prediction using ML methods for the most influential and most relevant genes for this disease. Table 3 shows the evaluation of ML models were used in the related works.

Table 1. Selected genes number of many feature selection methods

Feature selection method	Threshold	Number of genes selected
MI	0.05	7,542
ANOVA	0.05	8,583
Chi^2	0.5	8415
PCA	___	181
K-means clustering	___	8,005

Table 2. Accuracy of ML models for result of feature selection method

Feature selection method	Number of genes	Accuracy %		
		RF	SVM	NB
MI	7,542	83.7	86.4	83.7
ANOVA	8,583	83.7	81	83.7
Chi^2	8,415	86.4	89.1	83.7
PCA	181	78.3	83.7	51.3
K-means clustering	8,005	83.7	81	83.7

Table 3. Evaluation of ML models for different diabetes dataset

Reference	Dataset	ML model	Evaluation
[5]	GSE55098 for T1D	LASSO-SVM	AUC=0.918
[7]	Single-cell RNA-sequencing for T2D	Bayesian network, SVM, RF, LR and NN	ACC=0.907
[8]	lncRNA expression for T2D	KNN, SVM, LR and ANN	AUC=0.95
[9]	GSE38642 and GSE13760 for T2D	LR and SVM	ACC=90.23%
[10]	GSE164416 for T2D	SVM	Sensitivity=100%

5. CONCLUSION

To classify the genes affecting T1D diabetes, the gene expression dataset GSE52724 and GSE35713 were used. After applying the pre-processing methods, we concluded that this data is not suitable for working directly with ML techniques, as it needs to apply normalization to all data values, while it does not contain missing values. Then implementing the feature selection methods, we concluded that many of the features are not related to T1D diabetes, as they are not useful for classification. Therefore, during the implemention of feature selection stage, a very large number of these features were canceled, and the classification was based on the features related to T1D diabetes only. The highest accuracy of 89.1% is obtained from SVM model, the accuracy can be improved by using other methods of selecting features and applying another ML classification model.

REFERENCES

[1] A. Czmil, S. Czmil, and D. Mazur, “applied sciences A Method to Detect Type 1 Diabetes Based on Physical Activity Measurements Using a Mobile Device,” *Applied Sciences*, vol. 9, no. 12, pp. 1–16, 2019, doi: 10.3390/app9122555.

[2] N. P. Tigga and S. Garg, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods,” in *Procedia Computer Science*, 2020, vol. 167, pp. 706–716. doi: 10.1016/j.procs.2020.03.336.

[3] G. Selection, A. El-gawady, and M. A. Makhlouf, “Machine Learning Framework for the Prediction of Alzheimer ’ s Disease Using Gene Expression Data Based on Efficient Gene Selection,” *Symmetry*, vol. 14, no.3, 2022. doi: 10.3390/sym14030491.

[4] H. M. Deberneh and I. Kim, “Prediction of type 2 diabetes based on machine learning algorithm,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 6, Mar. 2021, doi: 10.3390/ijerph18063317.

[5] J. Lin, Y. Lu, and B. Wang, “Analysis of immune cell components and immune- related gene expression profiles in peripheral blood of patients with type 1 diabetes mellitus,” 2021.

[6] R. Shikha, R. Ku, V. Mi, and D. Gu, “Classification models for Invasive Ductal Carcinoma Progression , based on gene expression data-trained supervised machine learning,” pp. 1–15, 2020, doi: 10.1038/s41598-020-60740-w.

[7] L. M. and J. Zheng, “Single-cell gene expression analysis reveals β-cell dysfunction and deficit mechanisms in type 2 diabetes,” vol.

19(Suppl 19), no. 515, pp. 38-48, 2018, doi: 10.1186/s12859-018-2519-1.

[8] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, "Type2 diabetes mellitus prediction using data mining algorithms based on the long- noncoding RNAs expression : a comparison of four data mining approaches," *BMC Bioinformatics*, vol. 21pp. 1–13, 2020.

[9] H. Alshamlan and H. Bin Taleb, "A Gene Prediction Function for Type 2 Diabetes Mellitus using Logistic Regression," 2020 11th *International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2020, pp. 1-4, doi: 10.1109/ICICS49469.2020.239549.

[10] J. Li, J. Ding, D. U. Zhi, K. Gu, and H. Wang, "Identification of Type 2 Diabetes Based on a Ten-Gene Biomarker Prediction Model Constructed Using a Support Vector Machine Algorithm," *Biomed Res. Int.*, vol. 2022, 2022, doi: 10.1155/2022/1230761.

[11] T. Lee and H. Lee, "Prediction of Alzheimer ' s disease using blood gene expression data," *Sci. Rep.*, pp. 1–13, 2020, doi: 10.1038/s41598-020-60595-1.

[12] C. Sen Seah *et al.*, "An effective pre-processing phase for gene expression classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 3, pp. 1223-1227, 2018, doi: 10.11591/ijeecs.v11.i3.pp1223-1227.

[13] A. K. Al-Mashanji and S. Z. AL-Rashid, "Predicting with the quantify intensities of transcription factor-target genes binding using random forest technique," *Int. J. Nonlinear Anal. Appl.*, vol. 12, no. 2, pp. 145–161, 2021, doi: 10.22075/ijnaa.2021.5026.

[14] S. G. K. Patro and K. Kumar, "Normalization : A Preprocessing Stage," International Advanced Research Journal in Science, Engineering and Technology, vol. 2, no 3, pp. 20-22, March 2015, doi: 10.17148/IARJSET.2015.2305.

[15] L. A. Shalabi , Z. Shaaban and B. Kasasbe, "Data Mining : A Preprocessing Engineh," *Journal of Computer Science, Applied Science University,* Amman , Jordan," vol. 2, no. 9, pp. 735–739, 2006.

[16] T. K. Kim, "T test as a parametric statistic," *Korean Journal of Anesthesiology*, vol. 68, no. 6, p. 540, 2015, doi: 10.4097/kjae.2015.68.6.540.

[17] J. Englund, "Another Student ' s T -test Proposal and evaluation of a modified T-test," 2014. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:752341/FULLTEXT01.pdf

[18] A. Jovic, K. Brkic, and N. Bogunovic, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* May 2015, doi: 10.1109/mipro.2015.7160458.

[19] M. Kumar, N. K. Rath, A. Swain, and S. K. Rath, "Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor," *Procedia - Procedia Comput. Sci.*, vol. 54, pp. 301–310, 2015, doi: 10.1016/j.procs.2015.06.035.

[20] E. Odhiambo, G. Onyango, and M. Waema, "Feature Selection for Classification using Principal Component Analysis and Information Gain," *Expert Syst. Appl.*, vol. 174, p. 114765, 2021, doi: 10.1016/j.eswa.2021.114765.

[21] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Applied Intelligence*, vol. 52, no. 5, pp. 5457–5474, Aug. 2021, doi: 10.1007/s10489-021-02524-x.

[22] K. Sekaran and M. Sudha, "A survey on machine learning approaches in gene expression classification in modelling computational diagnostic system for complex diseases," *International Journal of Engineering and Advanced Technology*, vol. 8, pp. 182–191, Jan. 2018.

[23] M. Bramer, "Principles of Data Mining," in *Principles of Data Mining*, London: Springer London, 2016, pp. 1–343, doi: 10.1007/978-1-4471-7307-6_2.

[24] O. Okun and H. Priisalu, "Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues," in *Pattern Recognition and Image Analysis*, Springer Berlin Heidelberg, pp. 483–490, doi: 10.1007/978-3-540-72849-8_61.

[25] M. Ram, A. Najafi, and M. T. Shakeri, "Classification and Biomarker Genes Selection for Cancer Gene Expression Data Using Random Forest," *Iranian Journal of Pathology*, vol. 12, no. 4, pp. 339–347, Dec. 2017, doi: 10.30699/ijp.2017.27990.

[26] B. Chandra and M. Gupta, "Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1293–1298, Mar. 2011, doi: 10.1016/j.eswa.2010.06.076.

[27] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles," in *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2006, pp. 106–115, doi: 10.1007/11691730_11.

[28] "Gene Expression Omnibus" NCBI, [Online]: Available: http://www.ncbi.nlm.nih.gov/geo /, (accessed on 22 January 2023)

[29] A. K. Al-Mashanji and S. Z. Al-Rashi, "Computational Methods for Preprocessing and Classifying Gene Expression Data- Survey," *4th Sci. Int. Conf. Najaf, SICN 2019*, pp. 121–126, 2019, doi: 10.1109/SICN47020.2019.9019349.

[30] S. Gao, N. Wolanyk, Y. Chen, S. Jia, M. J. Hessner, and X. Wang, "Investigation of coordination and order in transcription regulation of innate and adaptive immunity genes in type 1 diabetes," *BMC Med. Genomics*, vol. 10, no. 1, pp. 1–14, Jan. 2017, doi: 10.1186/s12920-017-0243-8.

BIOGRAPHIES OF AUTHORS

Noor AlRefaai received the bachelor degree in computer science from University of Babylon in 2012. She is master student in information technology, University of Babylon 2021-2022. She can be contacted at email: noorali17790@gmail.com.

Sura Zaki AlRashid is an Assist Professor in computer Science, she received her Ph.D/Computer Science degree in Science Faculty, the University of Babylon, Iraq, 2015. Her area of interest includes artificial intelligent, image processing, data mining, bioinformatics, data analysis and text mining. She can be contacted at email: sura_os@itnet.uobabylon.edu.iq.