International Journal of e-Collaboration
Volume 18 • Issue 2

Collaborative Computing-Based K-Nearest Neighbour Algorithm and Mutual Information to Classify Gene Expressions for Type 2 Diabetes

Sura Zaki Al Rashid, University of Babylon, Iraq*

ABSTRACT

The classification process is used in gene expression data on venous endothelial cells of umbilical cords in humans to reveal the concepts of regulation of insulin using dynamic gene expression data for two classes, namely control and exposed to insulin. The mutual information statistical feature selection method is used on all available datasets to select these significant genes. The data reduction results are divided into training and testing and further supplemented to the KNN classifier for diabetes classification. The results show that the mutual information in KNN reaches the highest ranked 10,000 genes, and the test classification accuracy is 100%. Pathway analysis and gene ontology enrichment are used to evaluate the targeted genes. The results clearly exhibit the importance of finding the most informative genes in the database by using the statistical gene selection technique to achieve a reduction in time and cost and increase the efficiency of the classifier. This method exhibits these significant results that can be applied to other data and diseases.

KEYWORDS

Diabetes Disease, Gene Expression, K-Nearest Neighbour, MMGMOS, Mutual Information

1. INTRODUCTION

Diabetes is a chronic disease that affects humans regardless of their age and its causes, many of which are genetic and related to illness, impact and cause shock symptoms such as thirst, persistent fatigue, mobility issues and sweating. Patients suffering from diabetes die because of nephropathy leading to long-lasting problems such as cardiovascular macroangiopathy because harmful effects of hyperglycemia are prolonged in tissues. In terms of the pathophysiology, the disease is a classic metabolic condition of insulin-resistance in patients of type 2 diabetes. It can lead to compensatory hyperinsulinemia, which brings about a proliferative influence in the cellular vascular wall component, increasing the risk of cardiovascular diseases (Kharroubi, 2015), (di Camillo et al., 2010). There are ways to treat this disease such as by injecting insulin and through pills or herbal aid. The disease can lead to infection and complications of the kidney, eyes, brain, and other organs. In the endothelium, the transcriptional modifications characterisation is a key stage of a well considerate the mechanism of insulin action as well as the relationship between insulin resistance and dysfunction of endothelial cells [2]–(Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). Microarrays are a key tool for profiling the global gene expression patterns of tissues and cells. At present, such findings contain

DOI: 10.4018/IJeC.304044 *Corresponding Author

thousands of genes but few samples (Li, Weinberg, Darden, & Pedersen, 2001). A important challenge in biomedical studies in latest research concerns whether the data from samples can be classified and inferred into specific diseases [6]–(Babu & Sarkar, 2017).

Developing a suitable classifier and using training examples for genetic diagnosis is a problem in this area. herein this study, the challenge is to classify genes into control and exposed to insulin categories (Vanitha, Devaraj, & Venkatesulu, 2015a). Therefore, the k-nearest neighbours (KNN) approach of non-parametric pattern recognition is applied. Since the data set consists of several thousands of genes with few samples, for a specific dataset, many subsets of genes that can be classified under different sample classes may exist. Many subsets were found and the significance of genes was considered in the classification of the samples by examining the membership frequency of the genes in these near-optimal sets [5], [10]–(Bouazza, Hamdi, Zeroual, & Auhmani, 2015). While KNN is simple and clinically attractive, a large number of performance alternatives were found among groups for experienced data analysis (Sheela & Rangarajan, 2018), (Vanitha, Devaraj, & Venkatesulu, 2015b). The dimensionality reduction of the dataset variable space is an important and key pre-processing step for all the classification and clustering methods. Still, it is unknown whether increasing the specific genes' transcription for cellular proliferation is due to insulin itself in the endothelium or not. In this work, the classifier makes decision either control or exposed.

2. METHODS

2.1 Microarray Data Classification

A gene is a segment of DNA. It includes all the essential information to make the proteins in a body. Experiments of microarray expression permit the simultaneous calculation of the levels of expression for tens of thousands of genes. It mainly involves alternately evaluating each gene in a single environment but different tissues, checking each gene that is under altered conditions several times, mainly cancerous tissues (Maher, Mahmoud, & Salem, 2014). The microscopic DNA spots on the chip can be scanned by the machine based on the detection of the fluorescent excitation, which occurs if there is a hybridization. This technology is a powerful process to help with detecting the alterations in gene expression (Sura Zaki Al-Rashid, 2019), which can eventually be applied to distinguish the normal cells from cancerous ones, for instance. Nowadays, DNA microarrays have become an indispensable technology to specifically determine certain subtypes of particular cancers. In the classification of the microarray gene, a set of genes are given specific classes. The aim of this work is to obtain the relationship among the genes in same class, so the corresponding class label is retrieved once the gene is tested. Two-class classification problems are addressed for gene expression datasets. Patterns correspond to patients' samples and the features are coefficients of gene expression (Vanitha et al., 2015a), (Al-Mashanji & Al-Rashi, 2019).

2.2 K-Nearest Neighbours (KNN)

Each sample is classified according to the class memberships of its k-nearest neighbours. This is represented by expression patterns that consist of significant genes (N genes), which are determined by the mutual information distance. If most of the KNNs of a sample belong to the same class, it is classified as that class (Li et al., 2001). For example, in this work, they are classified as 'control'. Otherwise, the sample is classified under another class; for instance, in the case of this study, the sample is classified as 'exposed'.

3. PRELIMINARIES

3.1 Pathway Analysis

The Database for Annotation, Visualization and Integrated Discovery (DAVID) (S. Al-Rashid, Arifur, Al-aaraji, Lawrence, & Heath, 2018; Sura Z Al-Rashid & Al-aaraji, 2015) offers a full set tools of functional annotation for researchers to distinguish the biological meaning of a lot of genes. For

each list of given genes, DAVID tools are capable of the following (S. Al-Rashid et al., 2018; Sura Z Al-Rashid & Al-aaraji, 2015).

- Terms of Cluster redundant annotation.
- Making interacting protein lists, discovering enriched functional-related gene groups, identifying enriched biological topics, particularly GO terms among others, which can be searched on the free website.

3.2 Gene Ontology Enrichment Analysis

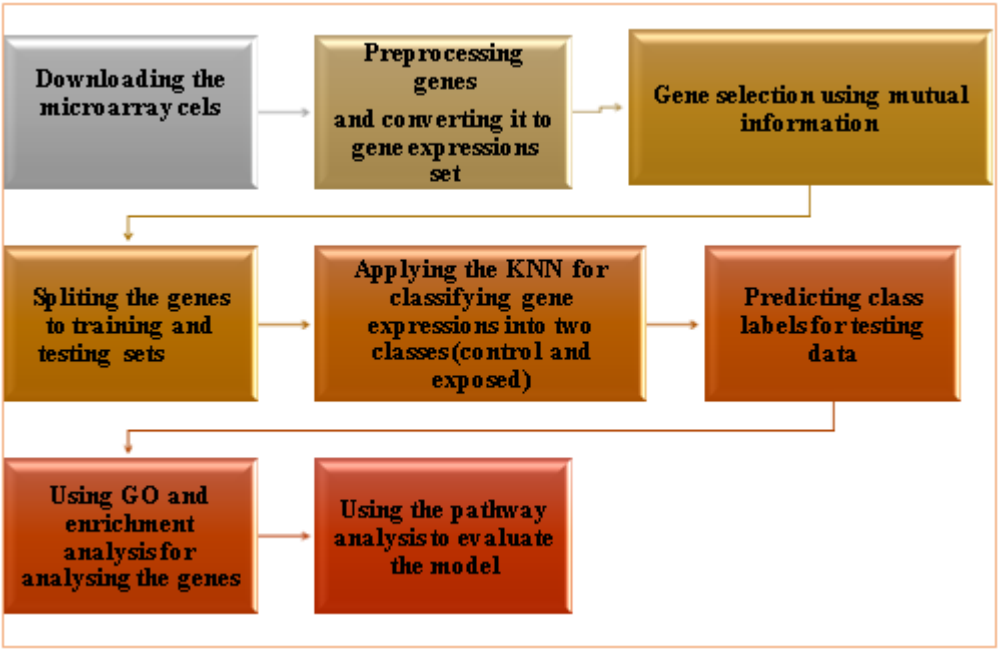
Characterizing the complicated data of biological systems, such as DNA sequencing, without the aid of computational software or tools is a challenge task. Biological systems are very complex and require trust in computers represent the knowledge.

The gene ontology (GO) task is a key bioinformatics initiative to enhance a computational representation of our evolving knowledge of how genes encode biological functions at the tissue system, cellular and molecular levels. The GO resource plays an important role in assistant biomedical research, including analysis of biological knowledge computation and interpretation of large-scale molecular experiments (S. Al-Rashid et al., 2018).

4. THE PROPOSED SYSTEM

Figure 1 shows the proposed work, where first, the data set is downloaded. This is followed by pre-processing steps, and then genes are selected using the mutual information method. However, before applying the KNN to classify the gene expression under the control and exposed labels, it is essential to split the gene for training and testing purposes. Finally, the targeted genes are examined using pathway analysis and GO enrichment analysis and the dataset is downloaded.

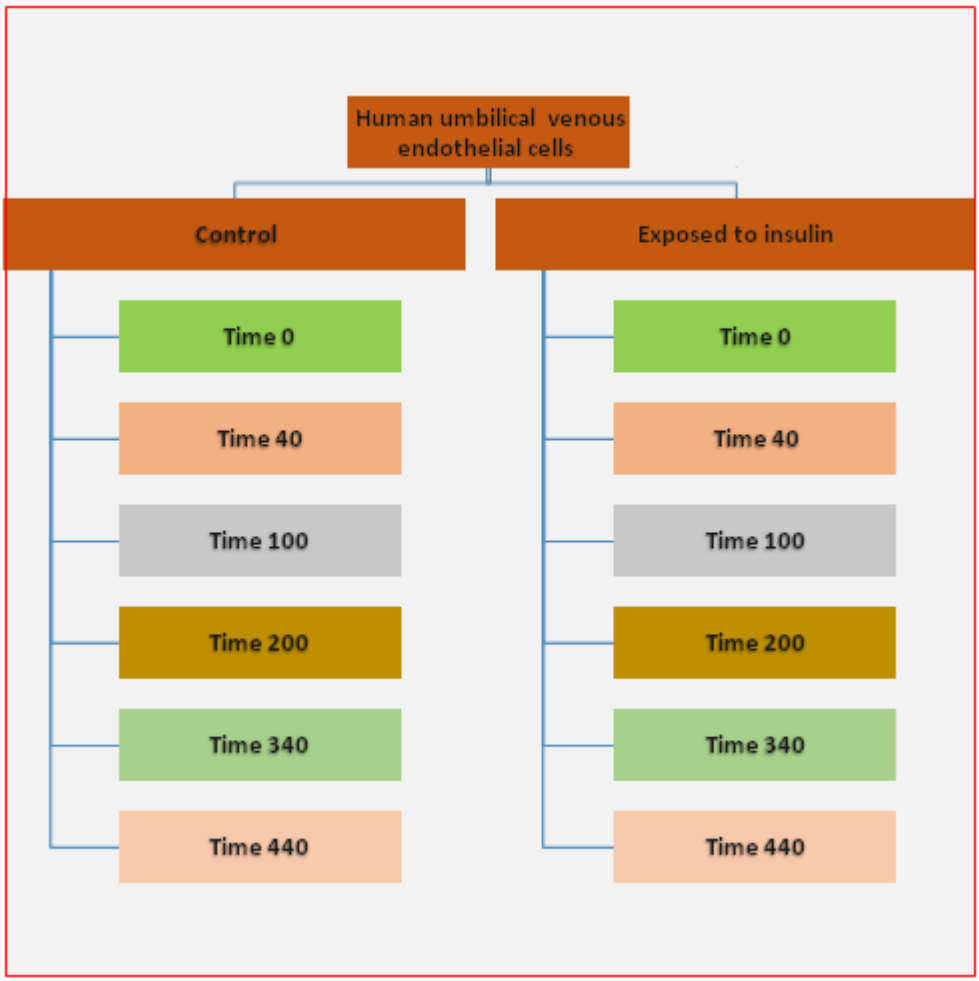
Figure 1. Block diagram of a proposed system



4.1 Downloading the Dataset

Experiments are dependent on the human umbilical venous endothelial cells (HUVEC) to explain the regulatory features of insulin using data of dynamic gene expression. HUVEC are covered through control by insulin and the exposed cells are recorded at 0, 40, 100, 200, 340, and 440 time series, as shown in Figure 2.

Figure 2. The data set is split into two. The first is a controlled gene expression and the other exposed to insulin. Times 0 to 440 are the time series of the experiments which have been done, where the level of gene expression is measured.in each experiment at each specific time



4.2 Computing Gene Expression

The mmgmos method is applied for analysing the process to estimate the gene expression values using the puma bioconductor package in R-language[20].

4.3 Gene Selection Using Mutual Information

The curse of dimensionality is one of the key problems in gene microarray data. It has few samples but thousands of genes. Further, a large number of genes this data exhibits are not informative for classification because they are is redundant or irrelevant (Babu & Sarkar, 2017). Thus, the use of feature selection considerably reduces the burden of computation in the classification and clustering tasks [21]. One of the most important steps in the classification analyses of microarray data is the gene selection step. The efficiency of selecting the gene can considerably ease the computation complexity of the task of subsequent classifications. This can yield a much smaller classification with more compacted gene set without impacting the accuracy. Therefore, a few nominated genes can be more economically and conveniently used for analytical tasks for the purpose of scientific experiments [21].

The gene selection process recognizes the informative genes among thousands of genes. It reduces the dimensionality as well as complexity of the dataset. In this task, the correlation among the genes is calculated applying the MI method. To apply this method, the probability distribution of genes is required, which is unknown in this experiment. Therefore, a histogram of the data has been used (Vanitha et al., 2015a).

4.4 Gene Ranking

Gene ranking simplifies gene expression tests to involve only a few genes from thousands. The genes are ranked using the MI method as shown in algorithm 1. Informative genes are found to greatly reduce the burden of computation and noise arising from irrelevant genes.

Algorithm 1: Mutual information selection method

Inputs: 2D array Dataset(n,m), where n is the number of genes and m is number of conditions. num_genes is number of genes required.

Outputs: list of significant gene indexes that has the highest of MI score.

Step1: Computing the mutual information between gene i and gene j according to following equations

$$H(I) = -\sum_{i \in I} P(I) logp(I) \quad (1)$$

$$H(J) = -\sum_{j \in J} P(J) logp(J) \quad (2)$$

$$H(I,J) = -\sum_{i \in I} \sum_{j \in J} P(I,J) \log(I,J) \quad (3)$$

$$MI(I, J) = H(I) + H(J) - H(I, J) \quad (4)$$

Where i,j=1…n.

Step2: Storing the mutual information values and index (i and j).

Step3:Sorting the mutual information array in descending order based on the mutual information scores.

4.5 Classification Stage

In this stage, KNN is used to classify the gene expressions to two classes, namely, control and exposed. The steps of this algorithm are explained in Algorithm 2. Six factors have been identified, including number of features(genes), feature ranking methods((Mutual Information selection), distance metric(Euclidian distance), vote weighting(higher same class for K neighbour), neighbour numbers(top K entries), and decision threshold(Prediction the class) that is relevant to the KNN method [22], [23], [24], (Maher et al., 2014).

Algorithm 2: k-nearest neighbour algorithm

Inputs: Raw data set and k is the number of neighbours.
Outputs: Test set is aligned to either the control or expose classes
Step1: Splitting the raw data into two sets (training 70% and testing 30%). Step 2: Determining the value of k. Step 3: Using the Euclidian distance to calculate the distance between the samplei and samples in the dataset, where I= 1 to N, N = number of samples in training set. Step 4: Arranging the samples to obtain the neighbours depending on the least distance calculated in the previous step and taking them to be the number of the adjacent k. (Getting the labels of the top K entries. Step 5: Defining the class for the neighbours, where the most common class of neighbours is the expected class for samplei.(Returning a prediction about the test sample using thresholding value)

4.6 The Measures of Validation

Accuracy: The overall accuracy of classification is the final correctness of the model that is computed as the sum of correct classifications divided by the all classifications.

$$Accuracy = \left(True\ \ Classification\right) / \left(Total\ \ no\ \ of\ \ cases\right) \tag{5}$$

Mean-Absolute -Error (MAE): is one of the most common metrics for continuous variables. This measure is used to calculate the error between the predicted value and the actual value. MAE is given by the following equation:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|P_i - A_i\right| \tag{6}$$

where: n: instances number, Pi: predicted value for record i, and Ai: actual value for record i.

Root Mean Squared Error (RMSE): is used to measure the differences between predicted values and observed values in the model. It provides a entire picture about the error distribution. The equation of RMSE can be represented

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(P_i - A_i\right)^2} \tag{7}$$

where: n: number of instances, Pi: predicted value of record i, and Ai:actual value of record i.

5. RESULTS AND DISCUSSION

5.1 Dataset

The Dataset is collected from a public microarray data repository. In this work, the diabetes datasets taken include 54,675 genes and 12 samples (Table 1). The gene expression data set in https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21989 containing about differentially expressed 54,675 genes is chosen as. The transcriptional

Table 1. Description of dataset

Dataset	Samples	Genes	Class
Diabetes Disease	6	54,500	2

response of endothelial cells at the 440 minute-mark after stimulation of insulin by microarrays is examined and compared to a control condition. Clustering genes in two groups is capable of helping the biological effects of insulin. We have applied the mmgmos method for analysing the process to get the gene expression values as shown in Figure 3. We have also applied the RMA function for analysing as shown in Figure 4. The values of gene expressions are showed in figure 5

5.2 Mutual Information

In figure 6, the format of MI matrix, after the ranking the genes depending on algorithm 1, where Table 2 shows the gene ranking sample of the 20 most informative genes. The genes with the highest scores are retained as informative genes.

Figure 3. Boxplot of values of gene expression via mmgmos method from puma package in R language (training set)

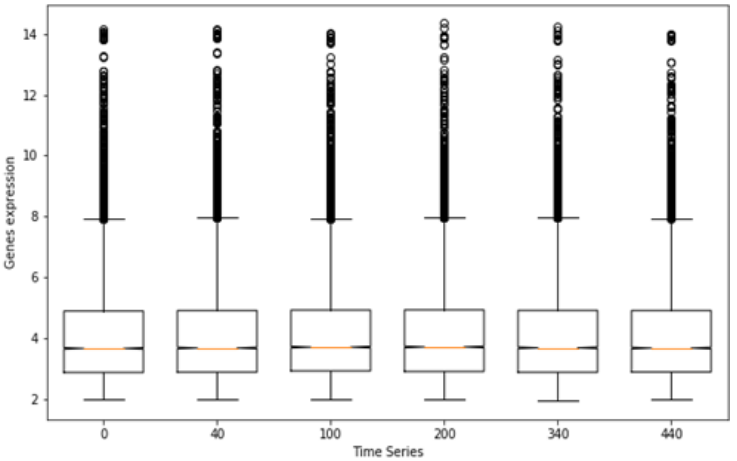


Figure 4. Boxplot for values of gene expression via mmgmos method from puma package in R language(test-set), where, the X_axis represents the experiments which done under set of time series and Y_axis represents the convergence and spacing of gene expression values and the values of outliers

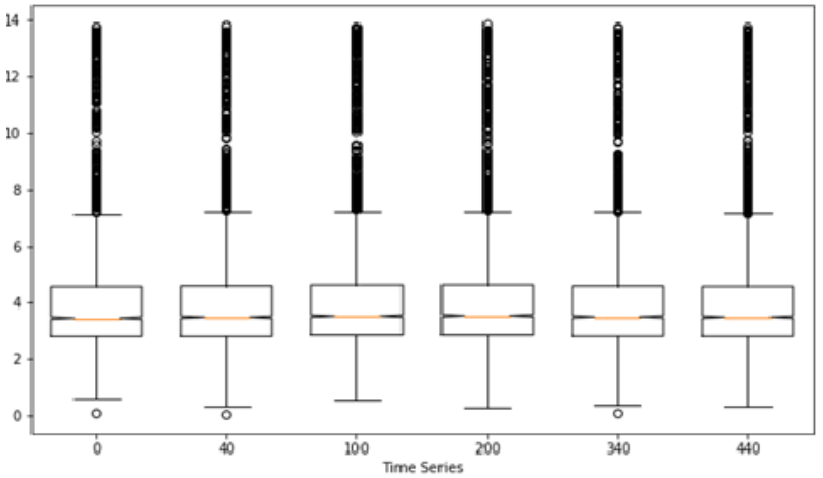


Figure 5. Screenshot for distribution of gene expression

A	B	C	D	E	F	G	H	I	J	K	L	M
	GSM546908.CEL	GSM546910.	GSM546911.	GSM546912.	GSM546913.	GSM546914.	GSM546909.	GSM546915.	GSM546916.	GSM546917.	GSM546918.	GSM546919.(
1007_s_at	6.9528171	6.99507876	6.79998615	6.95580258	7.01873263	6.88270264	7.16988751	7.1201922	7.13386205	7.11828708	6.92341013	7.12133667
1053_at	8.039938494	7.98807407	7.74500589	7.76488905	7.67398373	7.71670935	8.1154709	7.93794657	7.96528983	7.66184735	7.86141272	7.89411357
117_at	5.124377062	5.07760924	4.83257347	5.19920286	5.11903364	4.96996691	5.03319489	5.04673968	5.13841056	5.18407747	5.16469932	5.01674067
121_at	7.501561941	7.37436962	7.36793769	7.38376746	7.23331685	7.37870938	7.51825836	7.43650158	7.59796142	7.5672778	7.51742041	7.54444412
1255_g_at	2.928182133	3.06626086	3.00993108	2.80427812	2.87500354	2.90804678	2.83415331	2.92849097	2.82231474	2.71558087	2.84816778	2.87604036
1294_at	6.979458612	6.9030464	7.04496902	6.84973082	6.73495687	6.91459601	6.92276589	7.00140503	6.99704783	6.98038037	6.92434915	6.86712733
1316_at	5.22212662	5.17826064	4.87111484	5.19704117	5.18995676	5.01373996	5.15946096	5.22223274	5.03403707	5.16573782	5.1644018	4.90923908
1320_at	6.004727689	5.90739207	5.68502638	5.89722256	5.85315864	5.9665097	5.96638605	5.90343544	6.04565027	5.83459579	5.78922004	6.08331385
1405_i_at	2.604216269	2.57756464	2.4049493	2.42902616	2.6023348	2.55071453	2.52699664	2.44797741	2.68863994	2.67009704	2.48637251	2.51298088
1431_at	2.979576601	3.16730437	3.09597966	3.18316003	3.14297206	3.1206343	3.07513449	3.13684167	3.23698241	3.13658463	3.12696961	3.14363548
1438_at	5.35255615	5.13802994	5.05576208	5.19074638	5.17218752	5.16856958	5.3236339	5.18003343	5.29935325	5.11047664	5.16757321	5.20641445
1487_at	7.367652013	7.17435335	7.28230876	7.23079027	7.05596812	7.30338575	7.47609985	7.25564234	7.34228744	7.48745689	7.24364754	7.25326263
1494_f_at	5.700738494	5.34375311	5.55337846	5.4863237	5.36309238	5.26053482	5.4473051	5.45539442	5.45736952	5.40206524	5.20098738	5.37625077
1552256_a_i	8.409295504	8.27402633	8.27954229	8.3658546	8.08185339	8.12207655	8.58520859	8.55271882	8.35919187	8.50280553	8.04100347	8.19203787
1552257_a_i	8.32450323	8.07678121	8.21240753	8.14689989	8.03309974	7.99898123	8.32926699	8.18016715	8.38347589	8.17462645	8.09458459	8.10263985
1552258_at	4.328453851	4.21216642	4.04046876	3.88857491	4.07833614	4.09445566	4.28131872	4.34678921	4.37372624	4.53726901	4.30679118	4.44618253
1552261_at	4.323385691	4.19236279	4.36374408	4.22318552	4.27796445	4.08297293	4.30712132	4.26570216	4.48757897	4.19733726	4.27510277	4.26883228
1552263_at	6.979862535	7.00987854	6.70733907	7.03387897	6.8119336	7.11178509	7.38270105	6.99364754	7.12761068	6.9330513	6.98729942	7.07876399
1552264_a_i	9.487295385	9.26765839	9.40539416	8.94426759	9.09358305	9.35425877	9.39353774	9.1553261	9.2764697	8.92168885	9.10431347	9.30551208
1552266_at	2.993561696	2.9672587	2.76869016	3.19899279	3.0023098	2.77634236	3.02227092	2.95949676	3.30022366	2.90692226	2.9992768	3.19911002
1552269_at	2.42770285	2.56803442	2.56655463	2.44693452	2.63499545	2.52888199	2.70626456	2.3984162	2.48972956	2.44933921	2.39005898	2.73101788
1552271_at	5.590997692	5.59835147	5.62457655	5.64540181	5.53550185	5.77326387	5.51388485	5.8422486	5.79976061	5.69516119	5.64400591	5.77200441
1552272_a_i	4.911697293	4.84554133	4.9137697	4.8236371	4.60905913	5.12564303	4.76534837	4.97790353	4.99246926	5.07410192	5.01646294	4.96260718
1552274_at	6.086496889	6.56804299	6.24924413	6.62778366	6.61435109	6.2868085	6.57642169	6.31977292	6.08138829	6.29099499	6.78958565	6.35322915
1552275_s_i	6.390507891	6.69690523	6.48930477	6.61341438	6.60020758	6.86012223	6.99490848	6.65184966	6.7211021	6.57023484	6.72986463	6.9517695
1552276_a_i	6.063858556	6.21886958	5.9447647	5.56919687	5.78909223	5.67228537	5.97101949	5.96863138	6.18776055	6.08266449	5.9721383	5.96119258
1552277_a_i	9.675557323	9.64433052	9.72426951	9.56815444	9.479294	9.59703399	9.7869209	9.58312678	9.64358919	9.60408654	9.57510777	9.60330615
1552278_a_i	3.882635361	3.569503	3.607098	3.59862892	3.58396089	3.55050283	3.56906389	3.23247116	3.74449772	3.31019441	3.38348288	3.66715409
1552279_a_i	5.825862856	5.54875626	5.66791542	5.58062579	5.31806736	5.74065908	5.53121397	5.36408482	5.52984325	6.06299522	5.53391784	5.71370283

Figure 6. Format of mutual information matrix

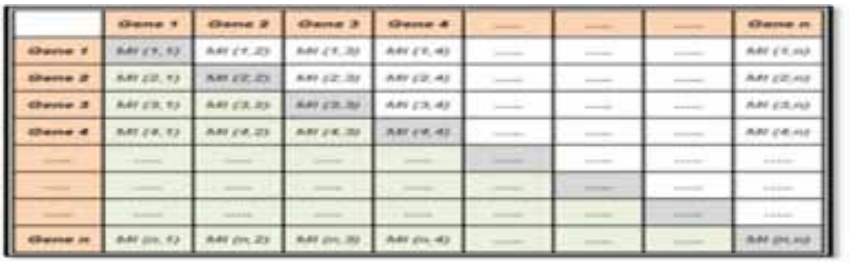


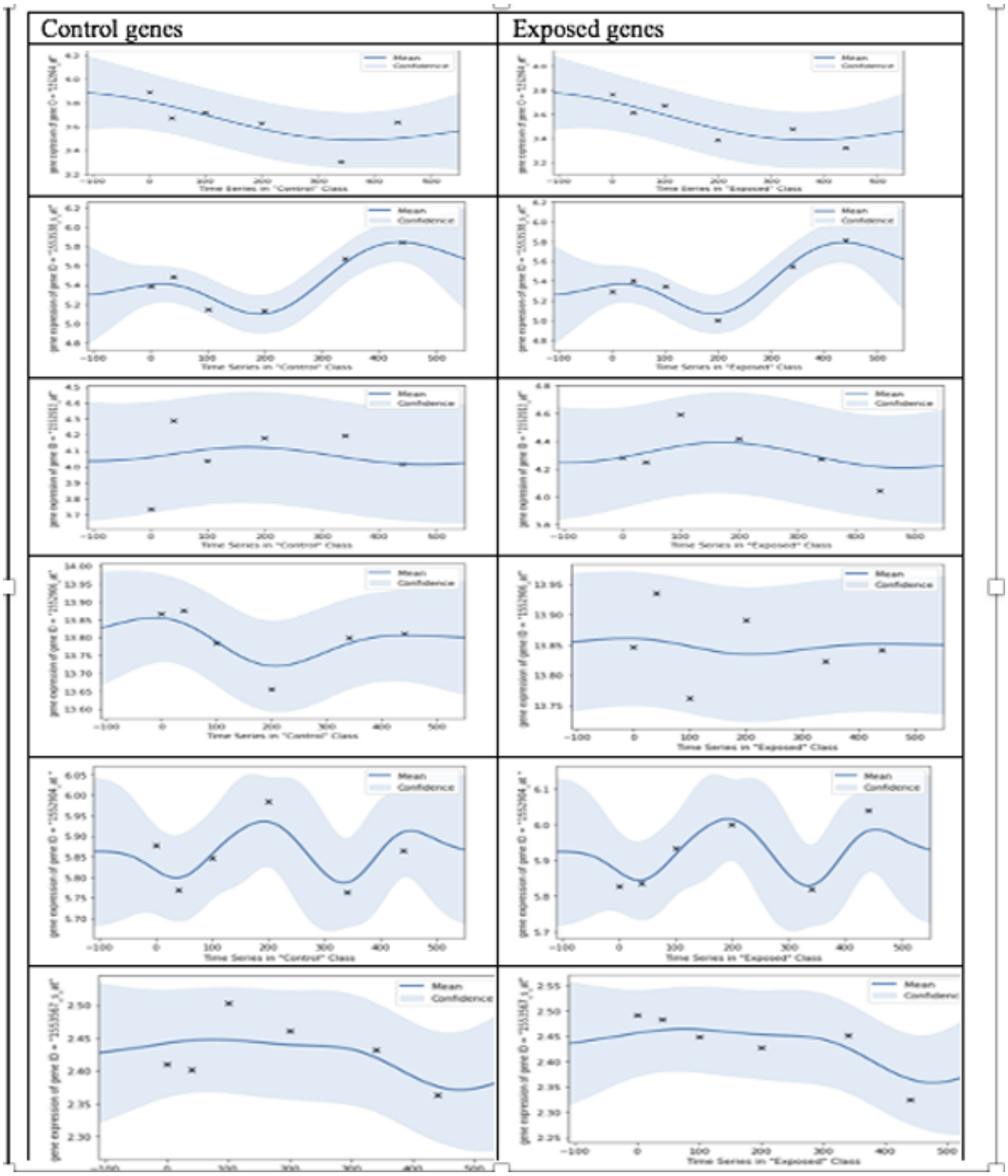
Table 2 Informative Genes based on their MI for Diabetes Dataset

No.	Gene ID	Mutual information value	No.	Gene ID	Mutual information value
1	1552974_at	[2.88919866 511. 935.]	11	1552972_at	[2.88265124 511. 962.]
2	1553538_s_at	[2.88896662 935. 987.]	12	1553214_a_at	[2.88248426 961. 987.]
3	1553604_at	[2.88592311 511. 975.]	13	1553311_at	[2.88241065 962. 987.]
4	1553588_at	[2.88568676 975. 987.]	14	1552913_at	[2.88143223 509. 935.]
5	1553441_at	[2.88413886 857. 935.]	15	1552906_at	[2.88093056 685. 935.]
6	1553567_s_at	[2.88370263 511. 959.]	16	1552904_at	[2.88085376 857. 975.]
7	1553132_a_at	[2.88346338 959. 987.]	17	1553415_at	[2.87998793 756. 935.]
8	1553569_at	[2.88281816 625. 935.]	18	1553362_at	[2.87952436 625. 975.]
9	1552372_at	[2.88272476 511. 961.]	19	1552325_at	[2.87942849 86. 975.]
10	1553570_x_at	[2.8827192 86. 935.]	20	1552480_s_at	[2.87862692 857. 959.]]

5.3 Classification Results

After the gene classification stage is completed, two classes are obtained, namely, control and exposed, for the 10,000 genes that are selected from the gene selection step using mutual information. In the previous step, these informative genes were split into training 70% and test sets 30% and KNN method applied on both sets. Figure 7 shows the behaviour of same gene in both classes, where the first column shows the genes that are controlled by insulin and the second column is exposed.

Figure 7. The behaviour of some genes in both classes, namely, control and exposed is shown in this figure, where the 1st column shows the genes in the control class while the 2nd column shows the same genes but in the exposed class. The X_axis contains the time series of the conditions and Y_axis indicates the behaviour of the genes.



6. CONCLUSION

Classification is considered a significant analysing technique of microarray data and is used for the prediction of classes among the genes of interest. Prediction also plays significant role in the biomedical field for the prediction stage of disease. In the field of biomedical science, the disease type, location, and stage of infection can be predicted based on the computational data obtained by using the microarray technology. In this paper, the classification technique has been applied for predicting the genes using a diabetes gene expression dataset with two classes, namely, control and exposed to insulin. In this work, the KNN gene classification is proposed as the method for classifying gene expression data. Computing MI among genes is applied to select the informative genes in the gene selection step. This step influences the classification performance using KNN for comparison with another algorithm, as shown in Table 3. The experimental analysis illustrates the effectiveness of the mutual information approach with KNN. This work can be further developed by using other data sets for other instances of diseases.

Table 3

	Time of built model	Correlation coefficient	Total number of instances	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
KNN with MI	0.023	0.096	40000	0.4308	0.6700	99.3572	153.2597
KNN without MI	0.03	0.0964	2*54675	0.4518	0.6722	90.3572	134.4297
Multiperceptron	51.87	0.0249	2*54675	0.4949	0.572	98.9885	114.3914
Linear Regression	2.46	0.0361	2*54675	0.4993	0.4997	99.8636	99.9343

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Al-Mashanji, A. K., & Al-Rashi, S. Z. (2019). Computational Methods for Preprocessing and Classifying Gene Expression Data- Survey. *4th Scientific International Conference Najaf, SICN 2019*, 121–126. 10.1109/SICN47020.2019.9019349

Al-Mshanji, A. A., & Al-Rashid, S. Z. (2019). Improving clustering algorithm for gene expression data using hybrid algorithm. *Compusoft*, *8*(9), 3422–3430.

Al-Rashid, S. Z., & Al-Aaraji, N. H. (2015). Bayesian Models with Coregionalization to Model Gene Expression Time Series for Mouse Model for Speed Progression of ALS Disease. *European Journal of Scientific Research*, *132*(1).

Al-Rashid, S. Z. (2019). Studying the effect of Mouse models for Gene Expression using Coregionalization Models in Gaussian process. *4th Scientific International Conference Najaf, SICN 2019*, 210–215. 10.1109/SICN47020.2019.9019355

Al-Rashid, S. (2011). Utilizing a Gath _ Geva Algorithm and Run Length Encode Algorithm for YUV Image Compression. *European Journal of Scientific Research*, *60*(1), 105–118.

Al-Rashid, S. (2013). Performance Evaluation of The Fuzzy C-means Algorithm and Comparison with Gath _ Geva algorithm for Color Images Segmentation Introduction : *Journal of Babylon University/Pure and Applied Sciences*, (1).

Al-Rashid, S. (2015). Inferring Transcription Factors Protein Activities by Combining Binding Information via Gaussian Process Regression. *Journal of Babylon University/Pure and Applied Sciences*, 1–16.

Al-Rashid, S., Arifur, M., Al-aaraji, N. H., Lawrence, N. D., & Heath, P. R. (2018). Increasing Power by Sharing Information from Genetic Background and Treatment in Clustering of Gene Expression Time Series. *Journal of University of Babylon*. *Pure and Applied Sciences*, *26*(4), 253–267.

Babu, M., & Sarkar, K. (2017). A comparative study of gene selection methods for cancer classification using microarray data. *Proceedings - 2016 2nd IEEE International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN 2016*, 204–211. 10.1109/ICRCICN.2016.7813657

Bouazza, S. H., Hamdi, N., Zeroual, A., & Auhmani, K. (2015). Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers. *2015 Intelligent Systems and Computer Vision, ISCV 2015*. 10.1109/ISACV.2015.7106168

di Camillo, B., Sanavia, T., Iori, E., Bronte, V., Roncaglia, E., Maran, A., … Cobelli, C. (2010). The transcriptional response in human umbilical vein endothelial cells exposed to insulin: A dynamic gene expression approach. *PLoS ONE, 5*(12). 10.1371/journal.pone.0014390

Fattah, S. A., & Lafta, H. A., & Alrashid, S. (2020). B-pred: An intelligent and adaptable medical diagnosis system based on bagging machine learning. *International Journal of Scientific and Technology Research*, *9*(3), 1325–1331.

Guyon, Weston, & Barnhill. (2002). *Gene Selection for Cancer Classification using Support Vector Machines*. 10.1108/03321640910919020

Kharroubi, A. T. (2015). Diabetes mellitus: The epidemic of the century. *World Journal of Diabetes*, *6*(6), 850. https://doi.org/10.4239/wjd.v6.i6.850

Kibanov, M., Becker, M., Mueller, J., Atzmueller, M., Hotho, A., & Stumme, G. (2017). *Adaptive kNN using Expected Accuracy for Classification of Geo-Spatial Data*. 10.1145/3167132.3167226

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics (Oxford, England)*, *17*(12), 1131–1142. https://doi.org/10.1093/bioinformatics/17.12.1131

MA. (2017). J. (2017). Prediction of heart disease using k-nearest neighbor and particle swarm. *Biomedical Research*, *28*(9), 4154–4158.

Maher, B. A., Mahmoud, A. M., & Salem, A. M. (2014). Classification of Two Types of Cancer Based on Microarray Data 2. *Related Work*, *38*(2), 56–66.

Sheela, T., & Rangarajan, L. (2018). *An Approach to reduce the large feature space of Microarray Gene Expression Data by gene clustering for efficient sample classification*. Academic Press.

Statnikov, A., Aliferis, C. F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics (Oxford, England)*, *21*(5), 631–643. https://doi.org/10.1093/bioinformatics/bti033

Vanitha, C. D. A., Devaraj, D., & Venkatesulu, M. (2015). Gene Expression Data Classification Using Support Vector Machine and Mutual Information-based Gene Selection. *Procedia Computer Science*, *47*, 13–21. https://doi.org/10.1016/j.procs.2015.03.178

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(5), 1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241

Sura Zaki Alrashid received her Ph.D. in Computer Science from the University of Babylon, Babil, Iraq. She completed part of a Ph.D. at Sheffield University, UK. She is a senior lecturer at the University of Babylon, Information Technology faculty. She has been teaching in the Software Department since 2005. She has been engaged to research works such as conferences and workshops. Her earlier publications are on Data mining and Bioinformatics. The author's current interest is in Bioinformatics, data mining, data analytics, text mining, time series analysis, and statistical modeling.