Journal of Education for Pure Science
Vol.15, No.2 (2025)
DOI: https://doi.org/10.32792/jeps.v15i2.453
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

# Improving DNA microarray classification with graph-based gene selection and maximum Clique (MC)

**Hadeel Najm[1] Firas Sabar Miften [2]**

[1]Computer Science Department, Thi-Qar University, Iraq

[*] Corresponding email: hadeelnajm.comp@utq.edu.iq

**Received 9/6/2024, Accepted 17/9/2024, Published 1/6 /2025**



**Abstract:**

Nowadays, one of the most important uses of molecular biology for disease detection is microarray data analysis. Gene selection, which searches for a subset of genes with the lowest internal similarity and highest relatedness to the target class, is an important task in microarray data processing. Data dimensionality can be reduced by removing duplicate, annoying, or unnecessary information. This study uses graph theory to support the gene selection strategy for disease diagnosis. To improve diagnostic skills, this study proposes a gene selection technique to evaluate DNA microarray data that takes advantage of graph theory and social network analysis. Gene selection effectively addresses this problem because it reduces computational complexity while increasing the classification accuracy of microarray data. This research proposes a unique gene selection method based on social network analysis. The two primary goals of the proposed approach are to reduce redundancy and increase its relevance for the selected genes. This algorithm iteratively determines the maximum population size in each cycle. Then, using node clustering, relevant genes are selected from this community's list of genes currently present. According to the published results, the new gene selection method will reduce time complexity while improving microarray data's classification accuracy. It is a bipolar disorder (Bip) dataset. Thirty of the 61 observations in the sample had bipolar disorder, while the remaining thirty were observational observations. The LS-SVM classifier was used to test the suggested approach on several sets of data, with the main focus of the study being classification accuracy. With an average classification accuracy using LS-SVM, the findings demonstrated notable increases in classification accuracy. The study demonstrated that the suggested methodology can efficiently identify genes and greatly increase classification accuracy by utilizing a variety of metrics to assess them. These findings demonstrate how well the suggested strategy analyzes microarray data and increases

classification accuracy, which represents a significant advancement in the field of gene expression-based illness categorization.

**Keywords**: Microarray data classification ,Gene selection, graph, LS-SVM, maximum Clique

## 1. Introduction

The amount of DNA microarray data has expanded due to advances in microarray technology. This development will enable medical professionals and biologists to investigate several aspects of gene expression simultaneously in an experiment to identify or classify different forms of cancer and other diseases. Data processing and analysis techniques face great challenges due to the rapidly increasing amounts of data accumulated in medical fields [1]. Deoxyribonucleic acid (DNA) is the molecule that carries the genetic instructions necessary for the development and functioning of living organisms and most viruses. DNA is made up of basic units called nucleotides, and each nucleotide consists of three main components: A phosphate group: This gives DNA its acidic character. Deoxyribose sugar: It is the pentose sugar that distinguishes DNA from RNA. Nitrogenous base: There are four main nitrogenous bases in DNA: adenine (A), thymine (T), Guanine (G), cytosine (C). DNA consists of two complementary strands that wrap around each other to form a double helix. Nitrogenous bases are linked to each other through hydrogen bonds, where adenine is bonded to thymine, and guanine is bonded to cytosine [2].

Based on gene expression patterns, researchers may use microarray technology to differentiate between malignant and normal tissue. Gene expression databases have been the subject of much research recently to identify different types of cancer. They also predict clinical findings in order to provide prognosis for cancer patients. Several characteristics of microarray gene expression datasets hinder the advancement of these approaches. The large dimensions of data sets are one of these characteristics [3]. Microarray technology is a modern molecular technique used to study the gene expression of many genes simultaneously. This technology relies on microarrays containing thousands of probes arranged in a regular grid. These probes are used to identify and quantify gene expression (mRNA) metabolites or genes present in a given samp[4]. .

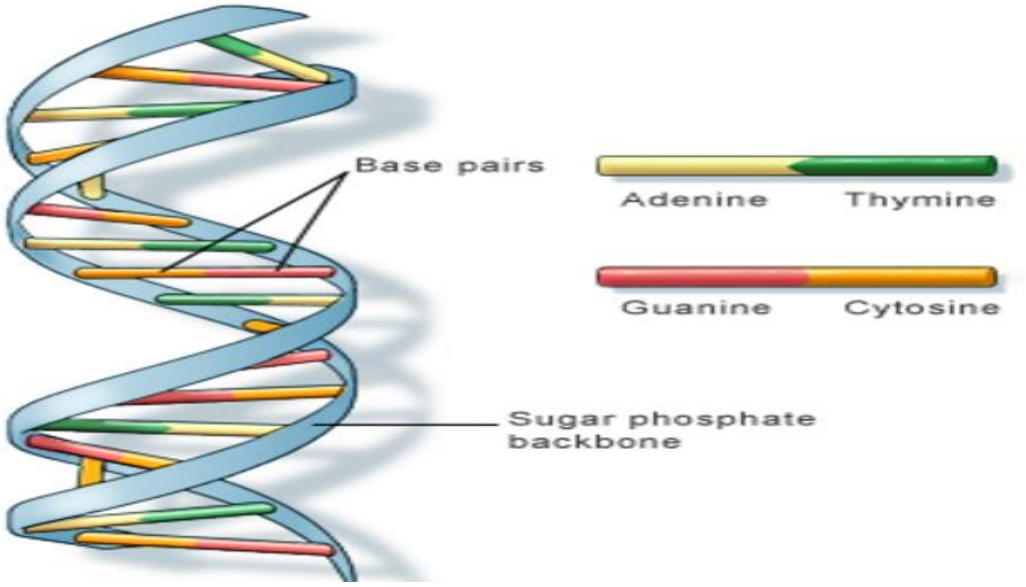


Fig. 1. DNA

Early treatment of diseases increases the possibility of cure. Reduces the mortality rate and the possibility of its recurrence. For this purpose, microarray technology is essential for widespread use. Gene expression data represents the exact state of a cell's DNA at the molecular level [5].

Technological advances address the exponential growth in data. For scientific purposes, these technologies make it easier to arrange huge amounts of data into logical groups. Big data may contain duplicate or redundant features in gene expressions. Therefore, scientists select a small set of salient traits from the available information in order to discover essential genes. Gene selection accelerates the learning process and improves model performance [6]. Machine learning (ML) is an intelligent, autonomous learning technique that allows machines to learn without explicit programming. Machine learning methods are frequently used to address a variety of challenging real-world issues and have proven to be effective[3]. Compared with previous methods, this technology has also been effectively used to solve a variety of problems and has achieved better results, especially in the medical sector. In addition, microarray has proven its ability to identify people with certain diseases.

As a result, diseases such as cancer are detected using this technique. Large dimensions and complex interrelationships between features are the main weaknesses of microarray dataset. The complexity of these problems should be reduced, and irrelevant genes should be removed [7].

In terms of gene expression, finding relevant genes usually involves feature selection methods and microarray data analysis. Differentially expressed genes were found by feature selection. Feature selection is also known as biomarker discovery or gene prioritization. Microarray data analysis is difficult

because there are so few samples with so many characteristics. The procedure of performing the tests is what causes the microarray data to be sparse. There are missing values in many microarray datasets, which affects post-processing. The main goal of this work is to present a unique multi-objective graph-based genetic technique for selecting the initial gene set that meets the above requirements [8].

The established approach consists of three main stages. In the first stage, key genes are depicted as a diagram. Each gene in this network is represented by a node, and similarities between genes are indicated by edge weights. In the second stage, an iterative procedure is employed to identify gene communities, aiming to group related genes together. The final stage involves introducing a scoring system to assign a significance score to each gene. Ultimately, the top-scoring genes from each community are selected.

To generate the complete set of genes, the developed gene selection technique aims to enhance the significance of these genes by identifying them within medical data, while ensuring the selection of genes with minimal redundancy. In this research method, gene frequencies are estimated through genetic similarities and community detection, and gene importance is determined accordingly. This gene selection approach offers several advantages.

## 2. Related works

**Mohamed Loey et al (2020)[10]:** This study introduces an automated system for diagnosing breast and colon cancer early, leveraging gene expression patterns and data mining. It emphasizes the significance of timely diagnosis for survival rates. Using information gain and the grey wolf algorithm, the system selects pertinent genes, reducing data noise. Employing intelligent decision support enhances classification accuracy, evaluated on Breast and Colon datasets. Results indicate the system can accurately identify illness-related genes with up to 95.935% accuracy, improving classification stability.

**Muhammed Abd-Elnabya et al(2021) [11]** : The study examines prospects and problems in the detection and treatment of cancer using microarray gene expression data. It presents a table of several approaches and talks about the most recent feature selection and classification strategies. The most popular classification technique, according to the results, is Support Vector Machine (SVM), which achieves a high accuracy of around 94.87% when combined with hybrid feature selection (IG-GWO). The authors point out that the scarcity of available samples and the variety of characteristics affect the accuracy of cancer categorization.

**Malek Alzaqebah et al(2021) [12]:** This work investigates the feature selection process of the Cuckoo search algorithm for cancer outcome prediction using gene expression datasets. The selection of

features is essential for precise categorization because of the intricacy of these datasets. The suggested approach makes use of a memory-based technique to hold onto the most illuminating characteristics found in the best solutions. Testing the algorithm on twelve microarray datasets produced results that show how effective and accurate the categorization is in comparison to other methods.

**Aiedh Mrisi Alharthia et al (2021)[13]:** Based on an updated version of the Elastic Net methodology, the paper presents a new approach to gene selection and cancer classification using microarray data. The strategy improves the efficiency and flexibility of gene selection and classification by including an L1-norm penalty. Evaluation measures demonstrate better performance when compared to alternative approaches, suggesting that this approach may increase the accuracy of cancer detection. Impressive accuracy 93% are attained using the suggested PLRAEN approach.

**Logeshwaran et al(2022) [14]:** This article provides insights into state-of-the-art medical systems and methods for evaluating tumor sequencing data by presenting a deep DNA machine learning model for tumor genome categorization. It goes over the benefits of using this kind of model, contrasts it with current systems, and lists some difficulties in putting it into practice in clinical settings. The following are the classification accuracies for the various models: DDML (95.47%), MLDNA (69.76%), DERCA (83.06%), DTL (65.80%), and CDNA (79.08%).

**Saeid Azadifara , et al (2022)[15]:** In comparison to previous approaches that employ iterative procedures, the research offers a unique gene selection method for cancer detection using a graph-based approach. This method is quicker and more efficient. It offers a thorough analysis of relevant works in machine learning, biomedical informatics, statistical pattern recognition, and

**Mehrdad Rostami ,et al (2022) [16]:** A novel multi-objective approach based on social network theory for gene selection in microarray data. The CDNC technique outperformed earlier approaches in terms of gene selection and classification accuracy, scoring 0.67, SVM 0.22, NB 0.75, 0.22, and Ada Boost (AB) 20.38, according to testing this strategy on many real-world data sets**.**

**Abdul Wahid, M. Tariq Banday (2023)[17]:** CNN is utilized for classification, while the ABC optimization technique is employed for feature selection. The accuracy of classification is increased by employing the ABC algorithm to choose the best characteristics.The proposed methodology combines feature selection and deep learning techniques to achieve excellent accuracy in predicting cancer subtypes and disease biomarkers.The accuracy of the ABC algorithm makes it useful for feature selection in additional healthcare data types, including clinical trial findings, medical pictures, ECG, EEG, and EMG records.

## 3. The proposed gene selection method

This section introduces a novel approach to gene selection that is founded on the idea of community detection. The concept of community detection in genomic data serves as the foundation for the suggested gene selection approach. An algorithm is used to classify genes related to one another into communities, and the genes within these communities are ranked using a method based on weight estimates for each community. Subsequently, each population's genes are chosen such that only highly significant genes remain in the final population and no duplicate genes are selected. Until a final group of genes is found that accurately reflects the intended categorization of genetic data, this procedure is repeated. There are four primary phases in the process:

1) A graph describes the gene selection issue, with each node representing a gene and edges denoting gene similarity.
2) Maximum clique identification, approach is utilized to ascertain gene subsets that exhibit the highest degree of mutual similarity.
3) After that, use a weighted community identification strategy to categorize the genes inside each community. Genes ranking.
4) Genes with high scores are chosen first to prevent repetition in the selection process. The method is continued until no more populations are detected in the genetic graph after removing the greatest population. The final gene collection is then expanded to include all remaining genes.

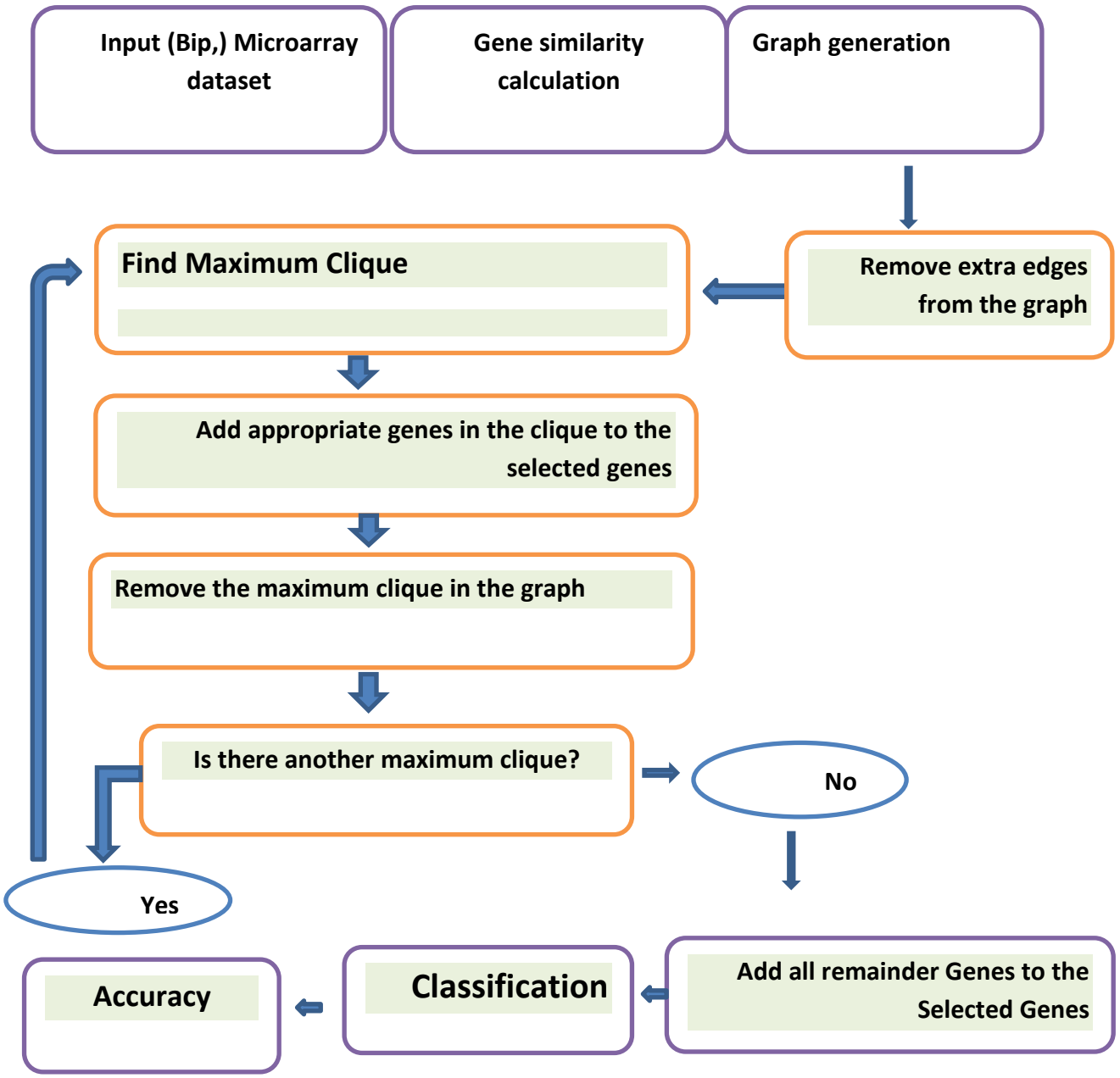


Fig. 2. Flowchart of the proposed model

The first steps and the suggested technique make up the two primary phases of the created gene selection method. Initially, a weighted graph is used to describe the gene selection issue. Genes are iteratively picked from populations in the suggested technique phase, and the final gene set is derived from the genes selected from various populations.

### 3.1 Graph construction

**F**irst, an unweighted, undirected graph is used to represent the solution space in the graph-theoretic gene selection approach. Every node in the graph represents a gene found in the microarray dataset; these nodes are grouped together and designated as G(F,E). Based on the contrast between genes, the edges in the graph (E) are created and given weights. A metric called mutual information (MI) is used to evaluate how similar two genes are to one another.

First, the mutual information (MI) between gene pairs is computed. Next, each set of data's regression (entropy) is calculated. As shown by Eq (1), genes with high MI values generally obtain high values . In order to overcome these constraints, Symmetric Uncertainty (SU) is used, which limits values to the interval [0,1] and discourages the assignment of genes with higher values. The computation of the SU of genes based on their common values is described in Eq (2).

Values below the mean, or THREECH SU VALUES, are then found. Row-wise averaging of the joint values (SU) matrix yields averages that may be compared. Zero is assigned to values that fall below the average. Common values are utilized to build the network after they have been found. Understanding interactions and linkages is aided by the network's representation of data and the interconnections between them.

The procedure is summed up by Algorithm 1, which aims to evaluate data and create a network that shows linkages, making it easier to understand data interactions and connections.

#### 3.1.1 Mutual Information (MI)

One of the well-known measures to assess the degree of similarity between two genes is Mutual Information (MI). Given two genes, Gi and Gj, what is the Mutual Information (MI) between them

$$\mathbf{MI}(\mathbf{Gi},\mathbf{Gj}) = \sum_{\mathbf{Values(Gi)}} \sum_{\mathbf{Values(Gj)}} \mathbf{p}(\mathbf{Gi},\mathbf{Gj}) \log \frac{\mathbf{p}(\mathbf{Gi},\mathbf{Gj})}{\mathbf{p(Gi).p}(\mathbf{Gj})} \quad (1)$$

where Gi and Gj, respectively, have potential values denoted by Values(Gi) and Values(Gj). Moreover, p(Gi), p(Gj), and p(Gi,Gj) are the distributions of Gi, Gj, and their combined distribution, respectively.

Genes with high values are probably naturally allocated high Mutual Information (MI) values. Symmetrical Uncertainty (SU), which discourages the assignment of genes with greater values and confines the values of genes in the range [0,1], is used in this work to alleviate this constraint. Gi and Gj's genes

### 3.1.2 Symmetric Uncertainty (SU)

is a straightforward technique to evaluate the quality of the characteristics. In its normalized version, this is just Mutual Information (MI). A normalized number between 0 and 1 is the basis for SU, which is based on information advantages. Thus, SU (X, Y) and SU (Y, X) are the same. SU treats the characteristic symmetrically, which is essentially what the word "symmetrical" means. Fewer comparisons will ultimately result from this.

$$\mathbf{SU}\ (\mathbf{Gi}, \mathbf{Gj}\ ) = \frac{\mathbf{2 \times MI\ (Gi\ ,Gj)}}{\mathbf{H(Gi) + H(\ Gj)}} \qquad (2)$$

An indicator of the level of uncertainty in a variable's distribution, entropy is defined as follows:

$$\mathbf{H(X)} = -\sum \mathbf{P(X_i)}\ \ \mathbf{log_2(P(X_i)} \qquad (\mathbf{3})$$

where X and X symbolize the discrete variable P(X) and the probability mass function of X, respectively, and H(X) indicates the variable's entropyIn order to decrease similarity, maintain node connectedness, and minimize the number of nodes with significant internal correlation, the threshold for nodes is set by averaging the SU for nodes.and after that In contrast If the SU value in the visible array is less than the average, set it to zero. Representing a graph using the "SU" matrix Consequently, there are less superfluous nodes.

### 3.2 Identification of maximal cliques

This subsection of the developed method outlines the first step of the second stage, which involves identifying the largest community in the graph. The goal of this step is to achieve initial partitioning. Nodes are divided into several groups where the members of each group exhibit maximum similarity with each other. Previous gene combination methods have faced several drawbacks, as noted in [13]. These methods often require the specification of group numbers before implementation, which can be limiting.

In most previously proposed methods for gene clustering, the distribution of genes within a cluster a crucial measure in gene clustering—is not adequately considered. Additionally, these methods tend to treat all genes as equal throughout the clustering process. However, more influential genes should have a greater impact on the clustering process. This developed method addresses these issues by allowing for more dynamic and accurate clustering based on genetic similarities and community detection.

This new representation model can efficiently reflect the characteristics of societies. In this algorithm, first, the optimal number of communities with very high representativeness is determined from which models are selected according to Similarity between nodes and models, and other nodes are mapped To Prototypes In the prototyping method, the selection of models is a vital
problem. For this purpose, there are two indicators of degree and similarity. Used to measure node representation. In this algorithm, The k nodes with high degree and low similarity with other models are Specified. In this algorithm(1)

| Algorithm 1: Find Max Weight Clique Algorithm |
|---|
| **Input:**<br>• Data set<br>**Output:**<br>• Cliques |
| **Begin:**<br>**Step 1: Calculate Symmetrical Uncertainty**<br>1. Initialize an N×NN \times NN×N matrix with dimensions proportional to the number of columns in the given data matrix.<br>2. Determine the symmetric uncertainty values between all possible pairs of columns in the data matrix.<br>3. Store these values in the NNN matrix.<br>**Step 2: Build Graph**<br>1. Use the NNN matrix to build a graph.<br>2. Each node in the graph represents a column in the data matrix.<br>3. The connections (edges) between nodes reflect the corresponding uncertainty values between the columns.<br>**Step 3: Find Maximum Weight Clique**<br>1. Use the graph built in the previous step to search for the largest nodes in the graph.<br>2. Find the maximum weight clique in graph GGG.<br>**End Algorithm** |

The symmetric uncertainty values between each pair of columns in the dataset are essentially the first thing the algorithm calculates. After that, it creates a graph in which the uncertainty values are represented by edges and columns, respectively. Ultimately, it looks for the graph's largest weight cliques, which are basically collections of strongly connected or mutually significantly informative columns.

## 4. Results and discussion

Gene selection techniques were employed in this study. This section outlines the proposed approach and evaluates its effectiveness. The tests were conducted on a system running Windows 10, equipped with an NVIDIA graphics card, 8GB of RAM, and 16GB of storage. Various metrics, including the number of selected genes and classification accuracy, were used to assess the results.

### 4-1: Description of the dataset

A well-known gene expression dataset with different gene numbers and sample sizes was used to evaluate the efficiency of the proposed technique and highlight its advantages over competing approaches. These freely accessible materials have been used by many academics in the past. [11]. They were used in this study to evaluate the performance of our approach. It is a bipolar disorder (Bip) dataset. Thirty of the 61 observations in the sample had bipolar disorder, while the remaining thirty were observational observations. (16,383 ) gene expressions of human genes were recorded using Affymetrix technology [9,17]

Table 1: Characteristics of the Dataset Used

| Datasets | # Samples(n) | # Genes(p) | # Classes |
|---|---|---|---|
| Bip | 61 | 16,383 | 31 control /30 bipolar disorder |

### 4.2 The applied classifier

By assessing the efficacy of the different gene selection strategies on two popular classifiers—the Least Square Support Vector Machine (LS-SVM) and K-Nearest Neighbor (KNN), a least squares variant of the SVM classifier by expressing the classification problem—the designed experiments assessed the adaptability of the developed method on different classifiers. Machine learning methods for handling

regression, classification, and pattern recognition include Least Square Support Vector Machine (LS-SVM). developed Support Vector Machines (SVM), a common machine learning-based technique for classification problems that has gained prominence recently . The classifier's objective is to maximize the margin between the available patterns.

K-Nearest Neighbor is a well-liked machine learning method (KNN). K-nearest Neighbor (KNN), for issues with regression and classification, is the basis for a lot of algorithms for lazy learning. It is one of the simplest categorization techniques and has a reasonable level of accuracy.

### 4.3 Experimental results

This section gives experimental findings in terms of feature reduction, parameter effect, execution time, and accuracy of classification term to illustrate the efficacy of the suggested strategy. First, a number of classifiers are used in the tests to assess the effectiveness of the suggested strategy. It gives a summary of the tables. precise categorization LS SVM is used to pick genes for the dataset across many groups.

The suggested approach worked well in the beginning trials. They are assessed using several classifiers. The average classification accuracy on L S SVM and kNN is summarized in Tables 2 through 4. The suggested strategy outperforms alternative gene selection techniques in every scenario. Additionally included in this table is the average accuracy.

The maximum accuracy achieved using the described technique on all microarray data was 99%. Moreover, the list of clusters that received accurate scores for every cluster that contained the most significant genes overall using KNN is displayed in Table 3.

A comparison of the LSSM classifier and KNN is presented in Table No. 4, with the findings indicating that the LSSM classifier outperforms KNN in terms of accuracy.

Table 2: Results of LS-SVM Classifier

| Classifier | Specificity | Sensitivity | Precision | Accuracy |
|---|---|---|---|---|
| LS-SVM | 1.00 | 1.00 | 1.00 | 99% |

Table 3: Results of KNN Classifier

| Classifier | Specificity | Sensitivity | Precision | Accuracy |
|---|---|---|---|---|
| KNN | 1.00 | 1.00 | 0.89 | 88% |

Table 4: Comparison of LS-SVM and KNN

| Metric | LS-SVM | KNN |
|---|---|---|
| Accuracy (%) | 99 | 88 |

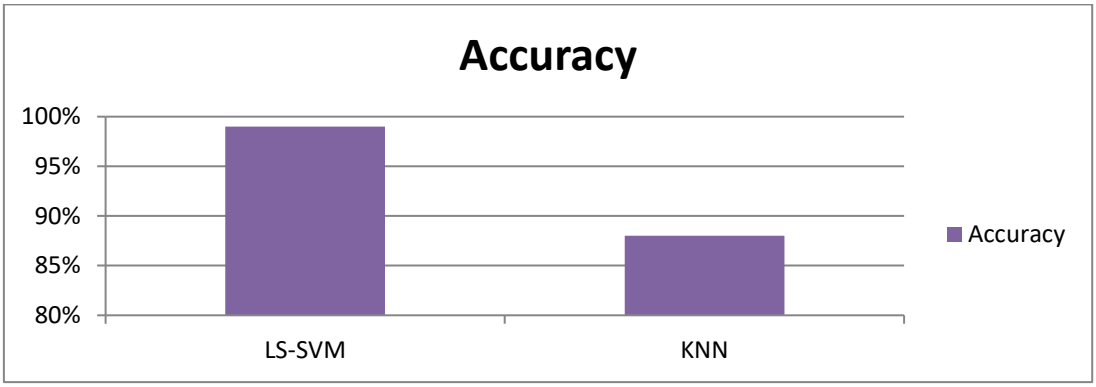


Fig. 3. Comparison of LS-SVM and KNN

The results indicate that the LS-SVM classifier outperforms KNN in terms of accuracy. The highest accuracy achieved using the proposed method on all microarray data was 99%, showcasing its potential effectiveness for gene selection in the context of bipolar disorder. The analysis revealed that the LS-SVM classifier provides more precise categorization and better overall performance compared to KNN.

## 5-Conclusion

In this study using the Max Weight Clique Algorithm, the study successfully identified the most significant and distinctive genes in the gene expression data related to bipolar disorder. These selected genes showed high classification performance, contributing to a better understanding of genetic relationships and providing effective tools for biological data analysis.Demonstrated excellent performance with specificity, sensitivity, and precision all at 1.00, and an overall accuracy of 99%. This indicates that LS-SVM was able to classify the samples with high accuracy using the selected genes.KNN Classifier Showed good performance but was less accurate than LS-SVM, with specificity and sensitivity at 1.00, precision at 0.89, and overall accuracy at 88%. This indicates that KNN was able to classify the samples well, but not as accurately as LS-SVM.This indicates the effectiveness of the gene selection algorithm in identifying important gene.

**References**

[ 1] Rostami, M., Forouzandeh, S., Berahmand, K., Soltani, M., Shahsavari, M., & Oussalah, M**.** (2022). Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artificial Intelligence in Medicine, 123, 102228. https://doi.org/10.1016/j.artmed.2021.102228.

[ 2] Stoughton, R. B**.** (2005). Applications of DNA microarrays in biology. Annu. Rev. Biochem., 74(1), 53-82. https://doi.org/10.1146/annurev.biochem.74.082803.133212.

[ 3] Alharthi, A. M., Lee, M. H., & Algamal, Z. Y. (2021). Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty. Informatics in Medicine Unlocked, 24, 100622. https://doi.org/10.1016/j.imu.2021.100622.

[ 4] Kerr, M. K., & Churchill, G. A. (2001). Experimental design for gene expression microarrays. Biostatistics, 2(2), 183-201. https://doi.org/10.1093/biostatistics/2.2.183

[ 5] Cilia, N. D., De Stefano, C., Fontanella, F., Raimondo, S., & Scotto di Freca, A. (2019). An experimental comparison of feature-selection and classification methods for microarray datasets. Information, 10(3), 109. https://doi.org/10.3390/info10030109

[ 6] Alzaqebah, M., Briki, K., Alrefai, N., Brini, S., Jawarneh, S., Alsmadi, M. K., ... & Alqahtani, A**.** (2021). Memory based cuckoo search algorithm for feature selection of gene expression dataset. Informatics in Medicine Unlocked, 24, 100572. https://doi.org/10.1016/j.imu.2021.100572.

[7 ]

Almugren, N., & Alshamlan, H**.** (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE access, 7, 78533-78548 10.1109/ACCESS.2019.2922987

[ 8] AbdElNabi, M. L. R., Wajeeh Jasim, M., El-Bakry, H. M., Hamed N. Taha, M., & Khalifa, N. E**. M.** (2020). Breast and colon cancer classification from gene expression, https://doi.org/10.3390/sym12030408.

[ 9] Abd-Elnaby, M., Alfonse, M., & Roushdy, M. (2021). Classification of breast cancer using microarray gene expression data: A survey. Journal of biomedical informatics, 117, 103764 https://doi.org/10.1016/j.jbi.2021.103764.

[ 10] Alzaqebah, M., Briki, K., Alrefai, N., Brini, S., Jawarneh, S., Alsmadi, M. K., ... & Alqahtani, A. (2021). Memory based cuckoo search algorithm for feature selection of gene expression dataset. Informatics in Medicine Unlocked, 24, 100572. https://doi.org/10.1016/j.imu.2021.100572.

[ 11] Alharthi, A. M., Lee, M. H., & Algamal, Z. Y. (2021). Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty. Informatics in Medicine Unlocked, 24, 100622. https://doi.org/10.1016/j.imu.2021.100622.

[ 12] Logeshwaran, J., Adhikari, N., Joshi, S. S., Saxena, P., & Sharma, A. (2022). The deep DNA machine learning model to classify the tumor genome of patients with tumor sequencing. International Journal of Health Sciences, 6(S5), 9364-9375.

[ 13] Azadifar, S., Rostami, M., Berahmand, K., Moradi, P., & Oussalah, M. (2022). Graph-based relevancy-redundancy gene selection method for cancer diagnosis. Computers in Biology and Medicine, 147, 105766 https://doi.org/10.1016/j.compbiomed.2022.105766.

[ 14] Yang, D., & Zhu, X. (2021)**.** Gene correlation guided gene selection for microarray data classification. BioMed Research International, 2021. https://doi.org/10.1155/2021/6490118

[ 15] Wahid, A., & Banday, M. T. (2023). Classification of DNA microarray gene expression Leukaemia data through ABC and CNN method. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 119-131

[ 16] Moradi P, Rostami M**.** Integration of graph clustering with ant colony optimization for feature selection. Knowl-Based Syst 2015;84:144–61. 2015/08/ 01/. https://doi.org/10.1016/j.knosys.2015.04.007.

[ 17] Rostami, M., Forouzandeh, S., Berahmand, K., Soltani, M., Shahsavari, M., & Oussalah, M. (2022). Gene selection for microarray data classification via multi-objective graph theoretic-based method. Artificial Intelligence in Medicine, 123, 102228. https://doi.org/10.1016/j.artmed.2021.102228