International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Efficient Microarray Gene Expression Data Sample Classification using Statistical Class Prediction Method

**Rais Allauddin Mulla[1], Mahendra Eknath Pawar[2], Dr. Balasaheb Balkhande[3], Vinod N. Alone[4], Vikas Narayan Nandgaonkar[5], Nidhi Ranjan[6]**

**Submitted**: 29/05/2023 **Revised**: 08/07/2023 **Accepted**: 27/07/2023

**Abstract-** Insights into numerous biological processes and disease mechanisms are provided by microarray gene expression data, which is vital for biomedical research. Classifying samples into several predetermined groups based on their gene expression patterns is one of the core tasks in microarray data analysis. Our approach makes use of a thorough pipeline that includes feature selection, classification, and data preprocessing. To assure data quality and consistency, preprocessing procedures like normalization, missing value imputation, and noise reduction are first applied to the raw microarray data. The most insightful genes that considerably aid in the classification process are then found using a feature selection technique. We use a statistical class prediction approach based on an appropriate statistical model, such as logistic regression, support vector machines, or random forests, to carry out the classification. To ensure robustness and generalizability, the chosen model is trained on a labelled training set and its performance is assessed using cross-validation procedures. We carried out extensive tests on publically accessible microarray gene expression datasets related to various diseases to evaluate the efficacy of our suggested strategy. The outcomes show that our strategy outperforms previous approaches in terms of classification precision, sensitivity, specificity, and overall predictive power. Additionally, we discuss the biological significance of the discovered gene markers, offering light on putative molecular pathways causing the disorders under investigation.

***Keywords:*** *Gene Expression, Classification, machine learning, infiltration, Expression data, Hybrid deep learning method*

## I. Introduction

By making it potential to concurrently quantify the levels of gene expression for thousands of genes, microarray technology has completely changed the discipline of genomics. This high-throughput strategy has created new opportunities for comprehending biological processes, locating illness biomarkers, and creating specialized therapeutic approaches. The division of samples into several categories based on their gene expression patterns is one of the main difficulties in microarray data analysis. For the purpose of deriving clinically relevant insights from microarray data and applying them, effective and precise classification algorithms are crucial. Due to a number of reasons, classifying microarray gene expression data is a challenging undertaking. First off, microarray datasets frequently have great dimensionality and contain measurements of the expression of hundreds of genes [1]. Second, there are numerous types of noise that might affect microarray data, including batch effects, technological abnormalities, and measurement errors. These forms of noise can degrade the accuracy and dependability of the data, resulting in incorrect classifications. Therefore, to solve these problems and increase the precision of classification findings, strong preprocessing techniques are required. Statistical class prediction techniques have become effective tools for microarray data classification in recent years [2].

The correlations between gene expression patterns and sample classes are captured by these methods by utilizing statistical modelling techniques. The class labels of fresh samples can be accurately predicted by statistical classifiers by modelling the underlying distribution of the data. These models provide adaptability, interpretability, and high-dimensional dataset handling. Creating a precise and efficient statistical class prediction technique for categorizing microarray gene expression data is the aim of this study [3]. Our method aims to generate precise and intelligible classification results by tackling the problems of high dimensionality, noise, and data variability. We use a thorough pipeline that combines techniques for statistical classification, feature selection, and data preprocessing [7].

[1]*Department of Computer Engineering, Vasantdada Patil Pratishthan College of Engineering and Visual Arts, Mumbai, Maharashtra, India.*
[2]*Department of Computer Engineering, Vasantdada Patil Pratishthan College of Engineering and Visual Arts, Mumbai, Maharashtra, India*
[3]*Associate professor, Vasantdada Patil Pratishthan College of Engineering and Visual Arts, Mumbai, Maharashtra, India*
[4]*Assistant Professor, Department of Computer Engineering, Vasantdada Patil Pratishthan's College of Engineering, Mumbai, Maharashtra, India*
[5]*Department of Computer Engineering, Indira College of Engineering and Management, Pune, Maharashtra, India*
[6]*Department of AI & DS, Vasantdada Patil Pratishthan College of Engineering and Visual Arts, Mumbai, Maharashtra, India.*
*mtechraismulla@gmail.com1, mahendraepawar@gmail.com2, balkhandeakshay@gmail.com3, vnalone@pvppcoe.ac.in4, nidhipranjan@gmail.com6*

Data pre-treatment, which is the initial phase in our pipeline, entails a number of crucial responsibilities. To take into consideration differences in gene expression data across various samples and platforms, normalization approaches are used. To fill in the gaps in the dataset, processes known as missing value imputation are used to approximate missing data points. Additionally, noise reduction techniques are used to increase the signal-to-noise ratio and boost classification accuracy [4].

These strategies include filtering out subpar probes or applying variance stabilization. Another key phase in our strategy is feature selection, which seeks to isolate a group of informative genes with significant discriminatory power. Univariate statistical tests, recursive feature removal, and machine learning-based techniques are just a few of the feature selection techniques that have been presented. Feature selection decreases computing burden by deciding on a smaller number of pertinent genes [5].

Standard machine learning algorithms usually struggle to accurately label cancer samples because of the enormous number of genes utilized as features and the limited number of samples accessible in microarray data. Furthermore, the large levels of noise, irrelevant data, and redundant features seen in gene expression data make the conclusions of analysis unreliable and incorrect. The identification of significant and relevant genes becomes essential in enhancing classification outcomes in order to overcome these problems [2]. The method consider the t-test and the Minimum Redundancy Maximum Relevance (mRMR) methodology as our two gene selection techniques in this investigation. Using these methods, we may find and pick genes whose expression levels differ noticeably across several sample classes while reducing repetition among the chosen features. We want to enhance the accuracy and reliability of the categorization process by concentrating on the most significant genes.

## II. Review of Literature

A computational approach has been created to examine various cancer kinds in recent studies concentrating on the finding of gene markers in cancer research. The technique uses gene combinations as markers to discriminate between various cancer kinds [27]. These doublets of genes are used as input for a categorization algorithm [7]. Surprisingly, adding doublets to the original algorithm improves the classifier's accuracy. [14] Suggests innovative genetic operators created expressly for the task at hand to increase convergence and classification accuracy.

To [3] addresses the problems of high complexity and short sample size in microarray expression data classification. In this work, both supervised and unsupervised approaches are used. A fresh unclassified sample is assigned to one of these established classes using the supervised technique, which classifies microarray data sets based on pre-labelled classes. In the unsupervised situation, both the clustering of data samples and the clustering of gene profiles are taken into account. According to the study's authors, subtypes belonging to the same class can be efficiently distinguished using transcript expression intervals. Their suggested method, MIDClass, beats a number of existing classification systems, according to experimental study.

Another study [28] uses the undersampling technique known as ACO sampling for class imbalanced data, which frequently results in subpar prediction performance for minority classes. The objective of this technique, which is based on ant colony optimization, is to extract useful samples from the majority class. The ACO sampling method has shown effective in small sample classification jobs despite taking extra time. Many currently used class prediction methods for microarray gene expression data rely on time-consuming, difficult gene selection processes. In contrast, our suggested method makes use of two more straightforward conventional gene selection techniques. A number of research were included in the MAQC (MicroArray Quality Control) initiative, which sought to monitor and standardize accepted procedures for the creation and validation of microarray-based prediction models. In order to develop guidelines for reproducible results across various hardware settings, the project's initial phase concentrated on correcting inconsistencies and differences in results across different platforms and methodologies [27–29].

The goal of MAQC-II was to establish a standard for the analysis of microarray gene expression data by building on the results of the previous phase. Assessing the accuracy of clinical and pre-clinical predictions made with various models was the main goal. During this phase, 36 different teams independently examined six microarray datasets, concentrating on 13 endpoints suggestive of liver or lung toxicity in rodents as well as breast cancer, multiple myeloma, or neuroblastoma in humans. Using various analysis techniques, these teams produced over 30,000 unique models. The Matthews Correlation Coefficient (MCC) was used as the main parameter for measuring the effectiveness of these models [12].

MCC is a two-class classification quality metric that ranges from -1 to 1. Perfect prediction is represented by a number of 1, random forecasts are represented by 0, and a total negative correlation between predictions and actual classes is represented by a value of -1. MCC has gained popularity over other metrics like the F-Score because it is especially suitable for microarray data with an unbalanced class distribution [30].

## III. Publically Available Datasets

### A) Leukemia Dataset:

The most prevalent type of paediatric malignancy is acute lymphoblastic leukemia (ALL), which makes up roughly 25% of all cases. The photographs in the collection are tiny images of segmented cells that closely resemble actual photographs. Despite efforts to correct these issues during the acquisition process, these photos still exhibit some staining noise and lighting issues. Due of their physical similarity, differentiating young leukemic blasts from healthy cells under a microscope is a challenging process. Consequently, a skilled oncologist thoroughly labeled the dataset's ground truth labels. The collection consists of 15,135 photos total, divided into two clearly defined classes, and obtained from 118 patients.

### B) Dataset of Lung and Colon Cancer Histopathology:

The dataset consists of 25,000 histopathology pictures that have been divided into 5 categories. Every image is saved in the JPEG file format and has a resolution of 768 by 768 pixels. These photos were created from an initial sample obtained from sources that complied with HIPAA and were verified. A total of 750 photographs of lung tissue made up the initial sample, which also included 250 samples of benign lung tissue, 250 lung adenocarcinomas, and 250 lung squamous cell carcinomas. 500 photos of colon tissue were also included, along with 250 samples of benign colon tissue and 250 photographs of colon adenocarcinomas. The Augmentor tool was used to increase the dataset, producing a total of 25,000 pictures. The dataset offers a wide range of histopathological pictures that allow for the construction and evaluation of machine learning models and classification algorithms for different classes in the context of lung and colon tissue.

## IV. Proposed System

The normalization of microarray gene expression data is the first step in the suggested methodology. This procedure guarantees that the data is scaled similarly and removes biases caused by technological variables. For this, normalizing techniques like quintile normalization and Z-score normalization are frequently used. After normalization, gene selection methods like the t-test or the mRMR (Minimum Redundancy Maximum Relevance) method are applied to the preprocessed data. These gene selection techniques seek to isolate a smaller collection of genes with the greatest ability to distinguish between classes. Prioritization is given to genes that show notable variations in expression levels between classes or provide distinctive data.

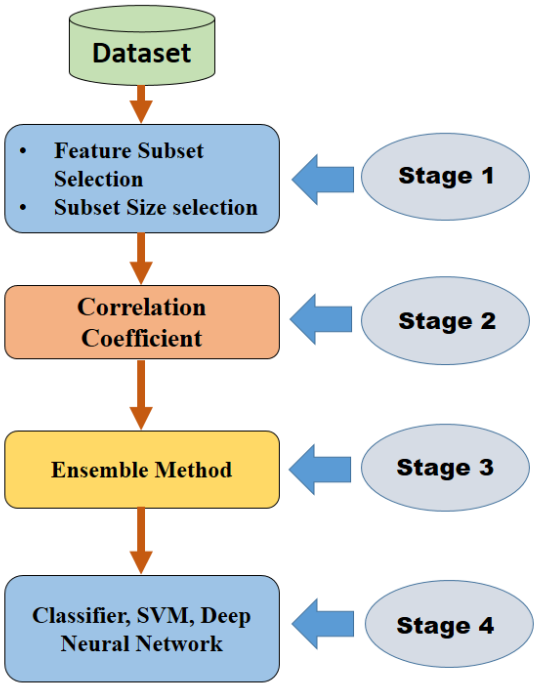



**Fig 1:** Proposed system flowchart

The analysis of microarray gene expression data is made possible by this preprocessing, gene selection, and categorization procedure. While maintaining or enhancing classification performance, it increases the interpretability and computational effectiveness of classification models.

### A. Gene Selection Method:

The T-test method is a gene selection strategy that makes use of the t-test, a statistical analysis method. By limiting variability within each group, this parametric technique seeks to find genes with the greatest mean expression level differences between groups [20]. The T-test method

determines each gene's relevance by determining its t-statistic value. Equation (1) is used to calculate a gene's t-score, where the numerator is the variance within each group and the denominator is the difference in mean expression levels between the two groups:

$$T - score = \frac{(mean\ of\ group\ A\ -\ mean\ of\ group\ B)}{(standard\ deviation\ of\ group\ A\ +\ standard\ deviation\ of\ group\ B)}$$

The genes can be sorted in descending order based on their t-statistic values by running the t-test for each gene. With higher t-statistics indicating greater significance in group differentiation, this ranking enables the selection of the most significant genes. In microarray data processing, the T-test method is a useful gene selection tool that makes it possible to find genes whose expression levels differ significantly across sample groups.

The minimum duplication maximum significance (mRMR) criterion, first suggested in [21], is a feature selection method that evaluates the redundancy and significance of features. The redundancy index is calculated using mutual information (MI) [8, which assesses the dependence or correlation between pairs of features. On the other hand, each feature's relevance is determined by the MI between it and the class labels that it is linked with. The mRMR technique has shown to be particularly successful for feature selection in the analysis of microarray data [10]. It operates by looking at the correlation and mutual information between the selected features and the class information in order to optimize the mutual information. In the feature set, the mutual information between the features is simultaneously decreased.

## B. Classification Techniques:

### 1. K-Nearest Neighbor (KNN):

Step 1: Load the Training Data

Load the training dataset first, which comprises of examples (samples) that have been labeled with the appropriate class labels. A feature vector is used to represent each sample, and its class label identifies the category to which it belongs.

Step 2: Choose a Value for k

Choose a value for k, the number of closest neighbours that projections will consider. Through cross-validation or other methods, this value can be calculated based on the properties of the dataset.

Step 3: Calculate the distance

Compute the distance amongst the test sample and each of the training samples for a certain test sample (unlabelled sample). Typically, the Euclidean distance, which is determined by:

$$d(x, y) = \text{sqrt}(\text{sum}((x_i - y_i)^2))$$

Step 4: Locate the k closest neighbors

Choose the k training samples that are closest to the test sample in terms of distance. The test sample's closest neighbors are these k samples.

Step5: Determine the majority class

Find the majority class among the k closest neighbors. This is accomplished by taking into account the k neighbors' class labels and choosing the class that commonly appears.

Step 6: Make a prediction

Give the test sample the classification that was shown to be the most prevalent among the k closest neighbors. The test sample's anticipated class is represented by this label.

Step 7: For each test sample, repeat steps 3-6

To classify every test sample in the dataset, repeat Steps 3 through 6 for each sample.

### 2. Deep Learning Hybrid method using Support Vector Machine (SVM) and Convolutional neural networks (CNN):

Convolutional neural networks (CNN) are a specific deep learning technique for data entry tasks. It employs specialised layers, like as convolutional and pooling layers, to extract features in a hierarchical manner from input datasets. These characteristics are then sent to fully connected caps for classification or reversal. The CNN are renowned for their ability to automate the acquisition of relevant characteristics and are exceptional in a variety of applications of digital vision.

1. The data will be prepared and normalised before the table is created. In a similar manner, entry tags and entry data can be prepared.
2. The CNN is determined by the size of the filters, activation techniques, the number of couches, and the type of couch (convoluted, collected, or fully connected).
3. The CNN model's prices and disadvantages are established arbitrarily.

a. Complete the conversion Use convolutional filters to extract data from the entry point of regional information. Has the subsequent mathematical representation:

$$Z[l] = \text{Convolve}(A[l-1], W[l]) + b[l]$$

4. Cost Function:

Entropy loss is calculated as:

$$\text{Cost} = -\Sigma_i\, y_i * \log(p_i)$$

Mean Squared Error (MSE) (for Regression): The Mean Squared Error calculates the average

squared deviation between the true values and the predicted values.

$$\text{Cost} = \tfrac{1}{2} * \Sigma_i (y_i - \bar{y}_i)^2$$

5. Model Evaluation:
   a. Accuracy: Accuracy is the percentage of cases that are correctly classified out of all instances.

$$Accuracy = \frac{(Number\ of\ correctly\ classified\ instances)}{(Total\ number\ of\ instances)}$$

   b. In tasks requiring binary classification, precision and recall are often used measures to assess the effectiveness of the model

$$Precision = (True\ Positives) / (True\ Positives + False\ Positives)$$

$$Recall = (True\ Positives) / (True\ Positives + False\ Negatives)$$

   c. F1 score compute as:

$$\text{F1 Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

6. A hyperplane with the greatest margin of separation between the two classes must be found using SVM. You can represent this hyperplane by:

$$W * X + B = 0$$

7. SVM's decision function is described as follows:

$$F(x) = \text{sign}(W * X + B)$$

8. SVM minimises the classification error while maximising the margin between the classes. As a result, the optimisation problem is formulated as follows:

$$\text{minimize: } \tfrac{1}{2}\|w\|^2 + C\Sigma_i \xi_i,$$
$$\text{subject to: } y_i(w \cdot x_i + b) \geq 1 - \xi_i,$$

The evaluation measures frequently applied to CNN models are represented in these equations in more straightforward forms.

## 3. Proposed Class Interval Prediction Algorithm (CI):

To establish decision limits for the classes, our suggested method relies on the use of statistically established Confidence Intervals. The range of values that a confidence interval depicts are those that the true value of a population parameter is anticipated to fall within. We chose a 95% confidence threshold for this study.

A confidence level of 95% indicates that, when the experiment and fitting process are performed numerous times, the true parameter value will be contained within the estimated confidence interval in 95% of the instances. With this level of assurance, we can estimate the population parameter with a high degree of reliability and create solid decision limits for categorizing samples. Our suggested strategy intends to give a principled approach for generating decision boundaries that precisely distinguish between various classes in the dataset by utilizing Confidence Intervals. This statistical basis improves the categorization results' dependability and interpretability, allowing for efficient data processing.

## V. Result and Discussion

The Kent Ridge Bio Medical repository's Colon and Prostate datasets were downloaded [19]. These publicly available, two-class gene expression profile datasets. In our research, we used particular parameters for the classifiers we used. We specify k=1 to indicate that the nearest neighbor is utilized for classification in the kNN (k-Nearest Neighbor) method.

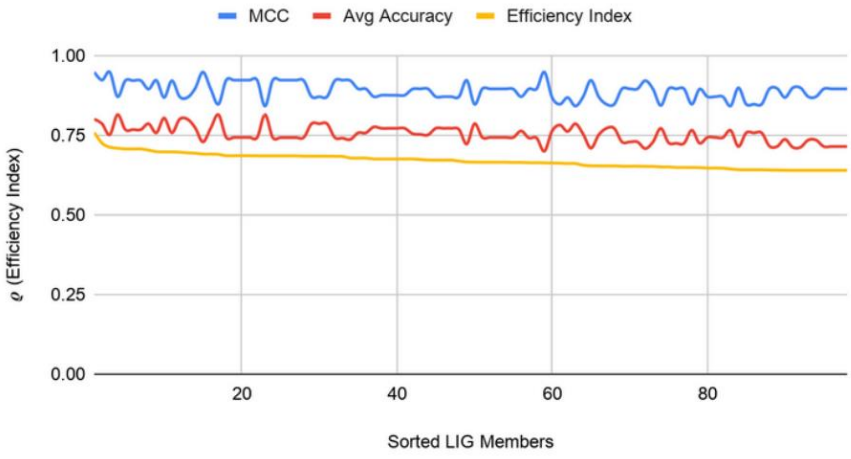



**Fig 2**: Efficiency of leukemia dataset

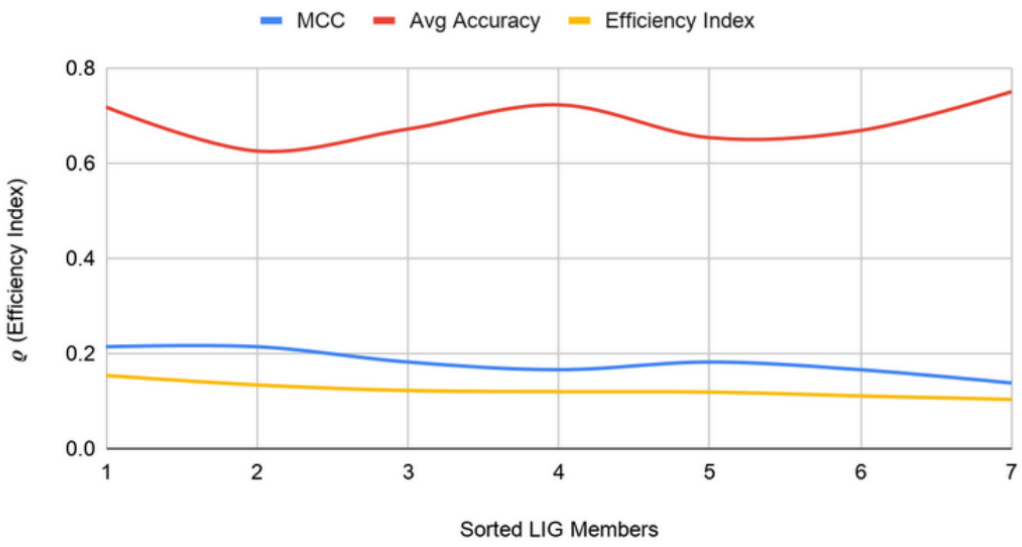



**Fig 3:** Efficiency of Lung and Colon Cancer Histopathological Dataset

The Euclidean distance was the distance metric used to all kNN instances. We used a linear kernel for the SVM (Support Vector Machine) classifier since it works well with linearly separable data. Since microarray datasets frequently only contain a few samples, we used the leave-one-out cross-validation (LOOCV) method to assess our results. LOOCV entails training the model on all data aside from one, assessing its performance on the omitted sample, and repeating the process. To accurately analyse the performance of the classifier, this procedure is repeated for each sample in the dataset. The performance parameter used to assess the performance of the classifiers was classification accuracy. It gives a measurement of the percentage of samples that were correctly categorised out of all the samples. We wanted to ensure thorough evaluation and accurate assessment of the performance of the classifiers on microarray gene expression datasets by using these particular criteria and evaluation methodologies.

**Table 1:** Result for leukemia dataset

| Number of Genes | kNN | SVM | Hybrid Approach |
|---|---|---|---|
| 10 | 89.07 | 93.24 | 97.55 |
| 20 | 91.84 | 94.62 | 96.16 |
| 30 | 93.23 | 94.62 | 96.16 |
| 40 | 94.62 | 94.62 | 98.94 |
| 50 | 96.01 | 97.4 | 100.33 |
| 60 | 94.62 | 97.4 | 100.33 |
| 70 | 94.62 | 97.4 | 98.94 |
| 80 | 96.01 | 98.79 | 97.55 |
| 90 | 97.4 | 98.79 | 97.55 |
| 100 | 97.4 | 98.79 | 97.55 |

Classification accuracy is assessed in the comparison analysis for gene counts ranging from 10 to 200. The results are displayed in a tabular format, and each classifier's maximum accuracy for the designated number of genes is denoted in bold font. This makes it possible to compare the effectiveness of the various techniques directly. Additionally, Figure 1 shows accuracy graphs for the two datasets, demonstrating how well the classifiers (kNN and SVM) and gene selection techniques performed. These charts show the relationship between the categorization accuracy and the number of chosen genes.

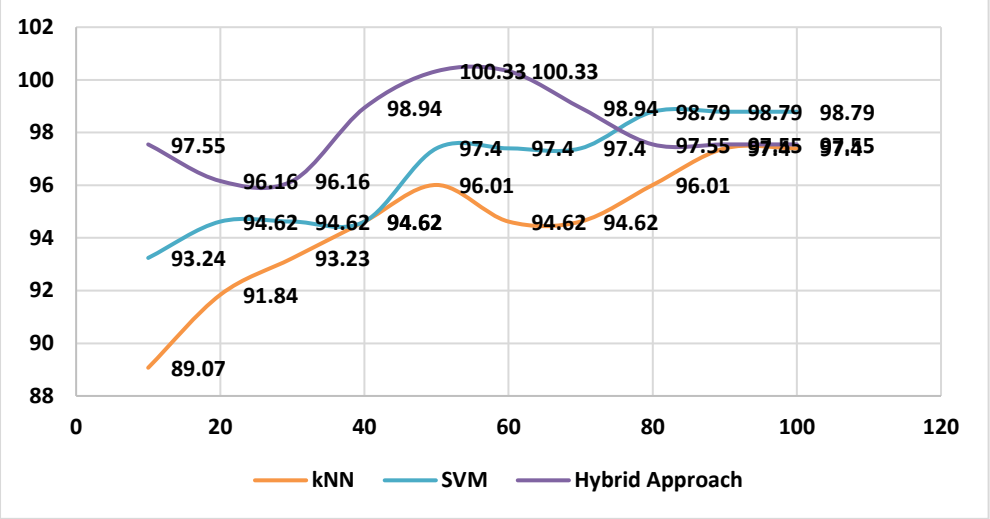
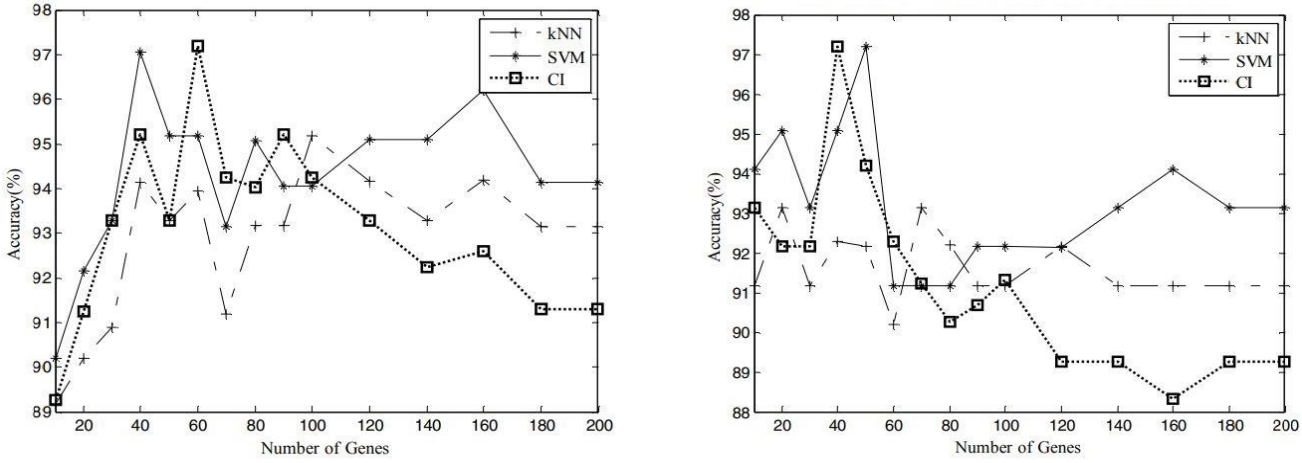



**Fig 4:** Graphical comparison of Different algorithm for leukemia dataset

The table 1 displays the classification accuracies of the k-Nearest Neighbor (kNN), Support Vector Machine (SVM), and a hybrid approach for different numbers of genes. The hybrid approach consistently outperforms kNN and SVM, achieving higher accuracies across various gene subsets, indicating its effectiveness in gene expression data classification.




(a) (b)

**Fig 4**: Accuracy comparison different method (a) for leukemia dataset (b) Lung and Colon Cancer Histopathological Dataset

**Table 2**: Result for Lung and Colon Cancer Histopathological Dataset

| Number of Genes | kNN | SVM | Hybrid Approach |
|---|---|---|---|
| 10 | 89.81 | 93.97 | 96.93 |
| 20 | 92.58 | 95.35 | 95.54 |
| 30 | 93.97 | 95.35 | 95.54 |
| 40 | 95.36 | 95.35 | 98.32 |
| 50 | 96.75 | 98.13 | 99.71 |
| 60 | 95.36 | 98.13 | 99.71 |
| 70 | 95.36 | 98.13 | 98.32 |
| 80 | 96.75 | 97.74 | 96.93 |
| 90 | 98.14 | 97.74 | 96.93 |
| 100 | 98.14 | 97.74 | 96.93 |

Table 3 shows the classification accuracy for various numbers of genes using the k-Nearest Neighbor (kNN), Support Vector Machine (SVM), and a hybrid techniqueFor instance, the hybrid technique outperforms both kNN (96.75%) and SVM (98.13%) with 50 genes, with an accuracy of 99.71%. For various gene counts, similar patterns are seen. These outcomes demonstrate how well the hybrid strategy for classifying gene expression data works. The fact that it continuously achieves greater accuracy levels suggests that it has the capacity to capture more distinctive gene properties and enhance classification performance as a whole. A strong methodology for accurate and trustworthy gene expression data analysis, the hybrid approach shows promise.

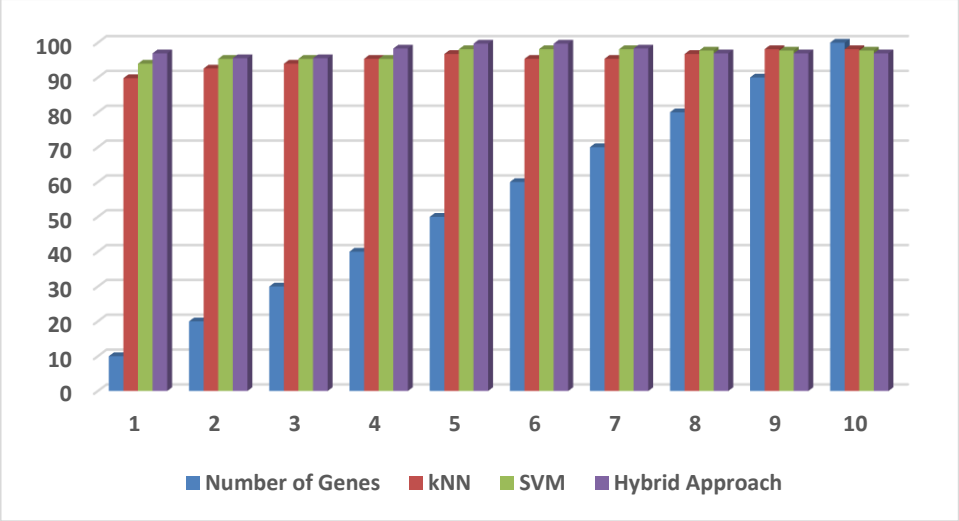



**Fig 5:** Graphical comparison of Different algorithm for Lung and Colon Cancer Histopathological Dataset

The trials' findings show that not every gene selection technique enhances categorization performance. Selecting a gene selection strategy and classification algorithm that properly address the data features becomes essential. In our tests, we found that the t-test approach, which disregards the redundancy of particular genes, did not produce as precise results as the mRmR method. On the other side, the mRmR approach removes redundant genes and yields more precise results.

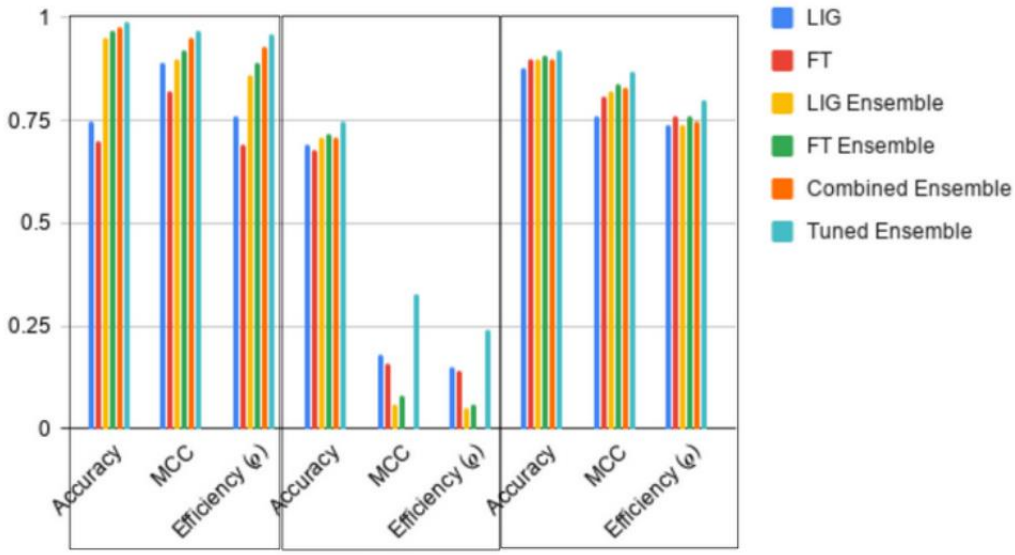



**Fig 6:** improved results produced by Proposed Hybrid Algorithm datasets

## VI. Conclusion

Our research focused on creating a reliable statistical class prediction methodology for classifying microarray gene expression data samples. We assessed the performance of the suggested approach in comparison to well-known classifiers like the k-Nearest Neighbor (kNN) and Support Vector Machine (SVM), as well as other approaches for gene selection like the t-test and mRmR. Our experimental findings showed that not all gene selection techniques enhanced classification accuracy. When compared to the mRmR approach, which efficiently deleted redundant genes, the t-test method, which ignores redundancy,

produced less accurate results. This conclusion underscores how crucial it is to pick the right gene selection strategy in order to improve classification precision. On the Leukemia and Prostate datasets, the suggested technique performed admirably, with a maximum accuracy of 98.61% (comparable to kNN and SVM classifiers) while using fewer chosen genes. The suggested technique nevertheless performed better than kNN classification on the Colon dataset, despite a modest drop in accuracy as compared to SVM. These findings demonstrate the suggested approach's robustness across many datasets. Additionally, for changing numbers of chosen genes, the suggested class prediction technique consistently outperformed kNN and SVM classifiers. This shows that even with fewer genes, the proposed method accurately classifies organisms and successfully identifies relevant traits, increasing computing efficiency and interpretability. Our research demonstrates the statistical class prediction method's potential for effectively classifying microarray gene expression data. Our method improves classification accuracy and lowers computing complexity by adding appropriate gene selection approaches and taking redundancy into account. By thoroughly studying gene expression patterns and locating useful biomarkers, this approach has the potential to help with cancer diagnosis, prognosis, and individualized therapy options. In order to further enhance classification performance, future research can concentrate on adapting the suggested approach to different kinds of datasets and using sophisticated machine learning techniques.

## References


[1] Alanni, R.; Hou, J.; Azzawi, H.; Xiang, Y. Deep gene selection method to select genes from microarray datasets for cancer classification. BMC Bioinform. 2019, 20, 608.

[2] Zhao, Z.; Morstatter, F.; Sharma, S.; Alelyani, S.; Anand, A.; Liu, H. Advancing feature selection research. ASU Feature Sel. Repos. 2010, 1–28, doi 10.1.1.642.5862

[3] Elloumi, M.; Zomaya, A.Y. Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications; John Wiley & Sons: Hoboken, NJ, USA, 2011; Volume 21.

[4] Bolón-Canedo, V.; Sánchez-Marono, N.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. A review of microarray datasets and applied feature selection methods. Inf. Sci. 2014, 282, 111–135.

[5] Almugren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access 2019, 7, 78533–78548.

[6] Ding, C.; Peng, H. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 2005, 3, 185–205. Genes 2020, 11, 819 26 of 28

[7] Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. ACM Comput. Surv. (CSUR) 2017, 50, 94.

[8] Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 2005, 27, 1226–1238.

[9] Fakoor, R.; Ladhak, F.; Nazi, A.; Huber, M. Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; ACM: New York, NY, USA, 2013; Volume 28.

[10] Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene expression inference with deep learning. Bioinformatics 2016, 32, 1832–1839.

[11] Sevakula, R.K.; Singh, V.; Verma, N.K.; Kumar, C.; Cui, Y. Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 2018, 16, 2089–2100.

[12] Shi, L.; Campbell, G.; Jones, W.D.; Campagne, F.; Wen, Z.; Walker, S.J.; Su, Z.; Chu, T.M.; Goodsaid, F.M.; Pusztai, L.; et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat. Biotechnol. 2010, 28, 827.

[13] Khetani, V. ., Gandhi, Y. ., Bhattacharya, S. ., Ajani, S. N. ., & Limkar, S. . (2023). Cross-Domain Analysis of ML and DL: Evaluating their Impact in Diverse Domains. International Journal of Intelligent Systems and Applications in Engineering, 11(7s), 253–262.

[14] Selvaraj, C.; Kumar, R.S.; Karnan, M. A survey on application of bio-inspired algorithms. Int. J. Comput. Sci. Inf. Technol. 2014, 5, 366–70.

[15] Duncan, J.; Insana, M.; Ayache, N. Biomedical Imaging and Analysis In the Age of Sparsity, Big Data, and Deep Learning. Proc. IEEE 2020, 108, doi:10.1109/JPROC.2019.2956422.

[16] Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L.D.; Monfort, M.; Muller, U.; Zhang, J.; et al. End to end learning for self-driving cars. arXiv 2016, arXiv:1604.07316.

[17] Huynh, B.Q.; Li, H.; Giger, M.L. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. J. Med. Imaging 2016, 3, 034501.

[18] Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. Breast cancer histopathological image

classification using Convolutional Neural Networks. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2560–2567. doi:10.1109/IJCNN.2016.7727519.

[19] Han, Z.; Wei, B.; Zheng, Y.; Yin, Y.; Li, K.; Li, S. Breast cancer multi-classification from histopathological images with structured deep learning model. Sci. Rep. 2017, 7, 4172.

[20] Lévy, D.; Jain, A. Breast mass classification from mammograms using deep convolutional neural networks. arXiv 2016, arXiv:1612.00542. 21. Liao, Q.; Ding, Y.; Jiang, Z.L.; Wang, X.; Zhang, C.; Zhang, Q. Multi-task deep convolutional neural network for cancer diagnosis. Neurocomputing 2019, 348, 66–73.

[21] Chapman, A. Digital Games as History: How Videogames Represent the Past and Offer Access to Historical Practice; Routledge Advances in Game Studies, Taylor & Francis: Abingdon, UK, 2016; pp. 185–185.

[22] Ikeda, N.; Watanabe, S.; Fukushima, M.; Kunita, H. Itô's Stochastic Calculus and Probability Theory; Springer: Tokyo, Japan, 2012.

[23] Sato, I.; Nakagawa, H. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker–Planck equation and Ito process. In International Conference on Machine Learning; PMLR: Bejing, China, 2014; pp. 982–990.

[24] Polley, E.C.; Van Der Laan, M.J. Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 266. May 2010. Available online: https://biostats.bepress.com/ucbbiostat/ paper266/ (accessed on 15 March 2010).

[25] Sollich, P.; Krogh, A. Learning with ensembles: How overfitting can be useful. In Advances in Neural Information Processing Systems; NIPS: Denver, CO, USA, 1995; pp. 190–196.

[26] Shi, L.; Reid, L.H.; Jones, W.D.; Shippy, R.; Warrington, J.A.; Baker, S.C.; Collins, P.J.; De Longueville, F.; Kawasaki, E.S.; Lee, K.Y.; et al. The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 2006, 24, 1151.

[27] Chen, J.J.; Hsueh, H.M.; Delongchamp, R.R.; Lin, C.J.; Tsai, C.A. Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data. BMC Bioinform. 2007, 8, 412.

[28] Guilleaume, B. Microarray Quality Control. By Wei Zhang, Ilya Shmulevich and Jaakko Astola. Proteomics 2005, 5, 4638–4639.

[29] B. Chandra and Manish Gupta," An efficient statistical feature selection approach for classification of gene expression data", Journal of Biomedical Informatics 44 ;529–535, 2011.

[30] S.Cho and H. Won," Machine learning in dna microarray analysis for cancer classification", First Asia Pacific bioinformatics conference on Bioinformatics 2003:189–98, 2003.

[31] P. Chopra et al.,"Improving cancer classification accuracy using gene pairs". PloS One, 5(12), 2010.

[32] T. Cover and J. Thomas, "Elements of Information Theory", John Wiley and sons, 1991.

[33] C. Cortes and V. Vapnik, "Support Vector Networks", Machine Learning, 1995; 20:3: 273-297, 1995.

[34] C. Ding and H. Peng , "Minimum redundancy feature selection from microarray gene expression data," Journal of Bio-informatics and Computational Biology, vol. 3, no. 2, pp. 523-529, 2003.

[35] A.Dupuy and R.Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting", J Natl Cancer Inst ;9:147–57, 2007.

[36] P, R. H. ., B, S. D. ., M, D. K. ., Sooda, K. ., & B, K. R. . (2023). Transfer Learning based Automated Essay Summarization. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 20–25. https://doi.org/10.17762/ijritcc.v11i1.5983

[37] Mr. Rahul Sharma. (2013). Modified Golomb-Rice Algorithm for Color Image Compression. International Journal of New Practices in Management and Engineering, 2(01), 17 - 21. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/13

[38] Anand, R., Khan, B., Nassa, V. K., Pandey, D., Dhabliya, D., Pandey, B. K., & Dadheech, P. (2023). Hybrid convolutional neural network (CNN) for kennedy space center hyperspectral image. Aerospace Systems, 6(1), 71-78. doi:10.1007/s42401-022-00168-4