INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A Refined Approach Involving Feature Correlation Analysis Followed by Classification using Python for Gene Assortment

S S Machchhar[1]*, U L Solanki[1], H S Sanghavi[1]

1 Assistant Professor, Computer Engineering Department, Government Engineering College, Bhavnagar, Gujarat, India

OPEN ACCESS

Received: 05-04-2024

Accepted: 29-06-2024

Published: 19-07-2024

Citation: Machchhar SS, Solanki UL, Sanghavi HS (2024) A Refined Approach Involving Feature Correlation Analysis Followed by Classification using Python for Gene Assortment. Indian Journal of Science and Technology 17(27): 2873-2879. https://doi.org/10.17485/IJST/v17i27.1110

* **Corresponding author**.

sahista.mtech@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To make a study on hybrid scheme of selecting a most representative subset of genes to improve microarray data classification accuracy using machine learning algorithm. **Method:** ML-based SGSA approach proposed here has two parts of execution on gene expression datasets. First it utilizes an entropy based IGKullback–Leibler divergence to select the most informative genes. Subsequently, an attribute evaluation performed using correlation-based feature selection. After that a random forest based classifier is employed with 10-fold cross-validation. The proposed method involves data pre processing, testing, training, fitting an algorithm and then finds the best accuracy with comparable CPU cost. **Findings:** The rationale behind this study is that only most informative genes are submitted to classifier for classification task. Proven accuracy by this approach is 98.48 for Lymphoma, 89.69 for Breast, 86.67 for CNS, 97.22 for Leukemia, 98.4 for Lung cancer, 96.6 for MLL, 97.21 for Ovarian cancer and 98.5 for SRBCT over traditional machine learning algorithms like naïve bias, J48 and SMO. These results demonstrate the effectiveness of the suggested approach in accurately classifying tumors. The numerical illustration also showed that the new estimator is more efficient in terms of CPU cost. **Novelty:** The major problem with traditional machine learning algorithms is that all features are treated equally important during the classification process which makes them susceptible to the influence of outliers and difficult to find a meaningful class in the dataset. In this study, classification accuracy is improved by processing the most informative features in the classifiers. The primary contribution of the research is a hybrid ML model which uses IG based feature selection followed by correlation analysis. The result is then fed to RF based classifier which significantly enhance accuracy as well as CPU cost.

Keywords: Machine Learning; Classification; Correlation; Feature Selection; Gene Expression

1 Introduction

The challenge of making an accurate disease prediction remains a critical issue for biomedical practioners and researchers. For that reason, clear recognition, classification and timely prediction of disease are essential aspects of study. Microarray gene expression datasets are high dimensional which has many features, and they are tricky for efficient computational study. These high dimensional features may contain insignificant genes may critically degrade the performance of machine learning algorithms. To ensure optimal performance, machine learning classifiers depends heavily on well prepared training data. For that reason, it is expected to provide refined training data to machine learning classifiers. Ongoing research is being conducted in this area to improve outcomes; however, there are additional flaws that need to be addressed. Lugagne's work introduced a deep model predictive control for driving single-cell gene expression which allows the model to identify the most informative features [(1)].

Feature selection is employed to address the issue of dimensionality. It involves identifying and removing irrelevant features to distinguish between relevant and irrelevant ones. Do and Tran-Nguyen's 2023 paper presents a novel method for classifying gene expression through the utilization of a multi-class support vector machine (SVM) integrated with feature selection [(2)]. The investigation offered by ghaleb encompasses basically two models namely regression and decision trees for mining gene expression data. These models are used in gene silencing approach, ultimately halting or impeding the advancement of cancer [(3)]. The study introduced by Senbagamalar have used a genetic cluster algorithm (GCA) for feature selection along with a divergent forest (DF) classifier for multiple classifications of data samples and have shown exceptional performance in terms of accuracy, specificity and sensitivity [(4)]. An important part of studying gene expression mining is to identify the group of genes that are correlated and are highly expressed in different types of tumor cells. A model was developed to integrate clustering and classification of gene expression RNA sequence data across various cancer types [(5)]. In studied background literature, the main objective is structured to encompass three main goals: enhancing classification accuracy, optimizing the reduction rate of specific attributes, and minimizing the complexity of the resulting models [(6)]. Achieving high precision in cancer prognosis is crucial for enhancing the treatment approach and the survival rate of patients. Utilizing machine learning methods can significantly aid in the prediction and early detection of breast cancer, making it a focal point of research and a powerful tool in the field. Aim of research by Naji is to find a best machine learning algorithm that is precise, reliable and finds the higher accuracy, incorporating new parameters, on larger datasets with a wider range of disease classes is most challenging while dealing with gene expression datasets [(7)]. Extraction of significant features on crop dataset performed gave a unique knowledge involving a strategy for planting the areca nut crop, limitation of these work is very small test and train samples were processed to forecast crop yield and tree lifetime [(8)]. Borodulin has constructed the mathematical model to determine dependencies exhibits a high level of accuracy around 93.06, effectively capturing the interrelationships among attributes [(9)]. Scope of work in this study is to improve accuracy level along with CPU cost too. Two new Bayesian network scoring functions were proposed by ajmal in [(10)]. They have used standard search method which limits the performance of said model on gene expression data.

The major challenges faced for constructing the model needed are explained in this section. Lugagne's work is limited by its inability to process multi-cell gene expression datasets. The research proposed in paper [(2)] has excluded 99% of the data and solely utilized the remaining data for their investigation. This approach may occasionally result in the exclusion of highly informative genes as well. Senbagamalar have used random forest classifier which has ensured greater accuracy compared to state of art methods. Their limitation or future work is to concentrate on computational complexity [(4)].

The authors in literature did not focus on both computational complexity and accuracy, which introduced a new objective to the study discussed in this paper. In this research article, a novel approach is introduced called Significant Gene Selection Algorithm (SGSA) in pursuit of enhancing the performance of cancer detection along with CPU cost. Previous works which are nearly related to this paper are [(4,9)]. The objectives are achieved by SGSA with the data preparation step followed by feature selection which is then fed to random forest (RF) based classifier.

Rest of this article is organized as follows: revision and review of a few primary design methodologies for the related work using gene expression data was described in this section. An algorithmic design methodology using a hybrid framework based on a combination of multiple filters on classifiers and discusses the concept of dependency on accuracy obtained by classifiers in section 2. The framework is applied to microarray gene expression datasets in for further biomedical study. Then the performance-based comparisons is presented while keeping a goal to provide a broad review of the recent advances on hybrid framework model-based genomic data processing for disease detection and prediction. The discussion on simulations and results are presented in section 3 and Conclusions are given in section 4.

1.1 The aim of this research

This study is intended to develop an automated method to the predicting and identifying disease classes by preparing the processing of gene expression data. The proposed procedure comprises of a streamlined model which consists of preprocessing and searching followed by a correlation analysis which is then fed to the classification process, this addresses certain weaknesses by conventional analysis of microarray gene expression data. This approach may give a successful, productive and reasonable alternative to help identifying the most informative gene among wide available gene expression data and it provides comparable accuracy with existing strategies. The aim of the predictive project is to discover recurring patterns by analyzing specific genes known or believed to play a role in the progression of the disease.

2 Methodology

2.1 Dataset descriptions

In this study, the datasets utilized are carefully considered along with their inherent data characteristics, and the effectiveness of conventional feature selection algorithms employed for microarray gene expression datasets. The gene expression datasets were obtained from multiple sources like Kaggle platform and ncbi repository. The Eight benchmark datasets are considered for empirical experimentation namely Leukemia, Lung cancer, Breast, CNS, Lymphoma, MLL, Ovarian cancer and SRBCT. Each data set has two classes, the experimental group and the control group, denoted by Class1 and Class2 respectively. The detail descriptions of the data sets are shown in Table 1.

Table 1. Dataset Description

Sr no	Dataset	Size in MB	No of features	No of tuples
1	Breast	15377	24482	97
2	Central Nervous System	1900	7130	60
3	Leukemia	2240	7130	72
4	Lymphoma	16145	4027	66
5	Lung cancer	1486	12601	203
6	MLL	4233	12583	72
7	Ovarian cancer	33401	15155	253
8	SRBCT	3239	2309	83

2.2 Proposed method

Algorithmic framework of proposed method Gene Selection and Classification framework is modeled in below mentioned Figure 1. The innovative nature of the proposed method makes it significantly more efficient when compared to traditional methods. The primary objective is to reduce the dimensionality of microarray data and enhance the convergence speed of classifiers, thereby enhancing the accuracy of detection. In the available literature, the methods utilizes machine learning to meager extent, however, the proposed approach combines the Random Forest algorithm with an optimized feature extraction process. After obtaining the feature vectors from the gene expression data IGKullback–Leibler divergence based entropy is calculated. This is then followed by a correlation-based Feature Selection (CFS) processing which selects most informative genes among high dimensional attributes present in datasets. Further four state-of-art classification techniques is applied and found that RF works better in major cases when accuracy and cost is concerned.

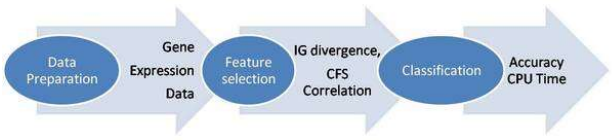

Fig 1. Flow of a Gene Selection and Classification framework

Pseudocode Significant Gene Selection Algorithm (SGSA)

Step 1: Pre-processing

Establish attribute rich file format from gene expression microarray datasets.

Step 2.1: Gene Selection

A selected subset of genes using the statistical methods such as Information Gain (IG), the Pearson's correlation coefficient, Gain Ratio, Gini Index and Principal Component Analysis (PCA). However, the results obtained are not recorded for every case instead which is comparable with Senbagamalar [4] and Borodulin [9]. By using gene screening strategy mentioned below, it is observed that the better results can be obtained.

Details of step 2.1 execution:

The significance of an attribute is evaluated by quantifying the entropy concerning the class. To obtain entropy of the attribute, calculation of an entropy based IGKullback–Leibler divergence is done using below-mentioned formulae:

$$\mathrm{IG}\left(\mathrm{A_i}, \mathrm{Y}\right) = -\sum_{u=1}^{n_i}\left(p_u \sum_{v=1}^{c} p_{uv} \log_2 p_{uv}\right)$$

Where p_{uv} is the probability that any arbitrary tuple belongs to particular class. Here logarithmic function with base 2 is used because the information is having bit level encoding. Entire gene expression is divided in partition based on available classes of attribute A_i. Specifically, genes were ranked based on their IG ratio values, measured as the mean value of IG ratios in all eight. Genes with a good IG percentage value were chosen as the mainly revealing ones in terms of variability. Then ranker model ranks and select top 20% attributes by their individual evaluations done by entropy based IG values.

Step 2.2: Correation Analysis of selected genes

Group of feature selection methods were studied from literature and empirically evaluated which leads to the conclusion that attributes evaluation using correlation-based Feature Selection (CFS) based processing along with best fit model was given comparable effect on the accuracy of classifiers. Below mentioned formulae are used to finalize correlation between two genes x and y:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

CFS calculates the merit of a subset of attributes by considering the individual predictive ability of each characteristic all along with the degree of redundancy between them. Only subsets of features that are remarkably correlated with the class while having low inter-correlation are preferred.

Step 3: Gene Classification followed by visualization

The main four state-of-art classification technique including Naïve Bayes, Support vector machine (SMO) and other two tree-based methods J48 and Random Forest were applied on resultant files leading us towards better accuracy.

The flow diagram of the framework of proposed study is as shown in Figure 2.

2.3 Experimental set-up

Proposed idea is implemented in Python. A personal computer installed with Jupiter notebook is used for all simulations on the datasets, with the following configuration: Processor: Intel(R) Core (TM) i3-4170 CPU @ 3.70 GHz, RAM: 8 GB, System type: 64-bit operating system with $\times$64 based processor. The classification accuracy and CPU time are used as the performance measures. The numbers recorded represent the highest values in these experiments.

3 Results and Discussion

The work of this part shows the performance of SGSA compared and contrasts with different state of art models which is analyzed in terms of accuracy, CPU cost. This study offers an ML-based method which uses Entropy as a statistical approach for selecting the most informative genes, CFS based analysis is the performed to find out significance of features and then after the classifier is used with 10 fold cross-validation. In the following experimentation, the performance of proposed model is compared with those of several state-of-the-art classifier approaches studied in literature survey. Based on the data analysis, four machine learning classification algorithms are used to classify separately the accuracy and CPU cost using the 10-fold cross-validation method. The performance of the proposed scheme is assessed in a series of simulations have also been performed using python.

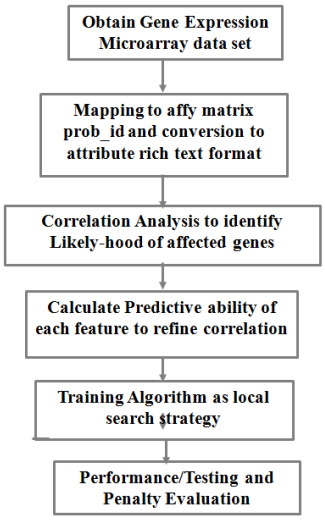

Fig 2. Flow diagram of the framework of Proposed Approach

Table 2. Classification accuracy with CPU time for state of art methods

Knowledge Discovery Algorithm / Type of Disease	RF [4]	SMO [7]	J48 [9]	Naïve Bias [11]
Lymphoma	100	92.42	96.97	92.42
	0.06	0.08	0.19	0.08
Ovarian Cancer	100	95.65	94.07	92.49
	0.45	0.62	1.17	0.34
SRBCT	100	84.34	98.79	98.79
	0.34	0.08	0.19	0.03
Breast Cancer	68.04	62.88	65.98	54.64
	0.42	1.29	1.75	0.34
CentralNervousSystem	68.33	58.33	66.67	61.67
	0.05	0.14	0.33	0.06
Lukemia	95.83	83.33	88.89	98.61
	0.05	0.11	0.3	0.08
Lung cancer	95.56	93.10	88.18	80.78
	0.5	1.19	1.06	0.27
MLL	97.22	84.72	94.44	95.83
	0.13	0.19	0.9	0.08

As the benchmark J48, Random Forest, the SMO method, and the Naive Bayes methods are used for training which were found in the literature study. The datasets used in this work are publicly available on the internet. Dataset of Leukemia has 72 instances, features count as 3572 and classes count as 2, lymphoma dataset consists of 77 instances, 2647 features, and 2 classes. Prostate dataset has 102 cases, 2135 features and 2 classes. Results of all four selected state-of-the-art classification algorithms for all eight datasets are recorded. Accuracy has been achieved and verified for three groups of frames, which are then analyzed in the subsequent section. The classification results of these classifiers in terms of average accuracy (AA), and CPU Cost are presented in Table 2 for the Naïve bias, SMO, J48 and Random forest (RF) respectively.

After this execution, gene selection methods have been applied by choosing a better small set of discriminative genes which has increased performance that is recorded and shown Figures 3 and 4. And, as shown by the results, proposed method outperforms others; the results are shown in Table 2. Shows achieved the highest classification accuracy, up to 98.4% in a few cases using RF which is so then employed for classification. The classification score of proposed approach compared against the

Table 3. Summary of comparative study of state of art methods with proposed method

	Year	# of features	Method	Key findings	Accuracy
Lugagne [1]	2024	8	multi-layer perceptron	Predict gene expression in single cells with high accuracy	RMSE time 190
Senbagamalar [4]	2024	21	Divergent forest	Genetic cluster algorithm used for feature selection	95.21
Naji [7]	2021	11	SVM based algorithm	Five different models were used for comparison	97.2
Borodulin [9]	2024	18	Deductor Academic version 5.3	Decision-tree based classifier gives better performance	93.06
SGSA (proposed)	NA	2309	BN scoring functions	Dynamic Bayesian Network	98.4

existing approaches shown in Table 3. Two resulting parameters are shown graphically in below-mentioned graphs, wherein Figure 1 represented an accuracy of proposed SGSA Algorithm is compared with all four state-of-art methods. From Figure 1, it can be seen that the CPU time of proposed SGSA method on all eight benchmark datasets outperforms all four state-of-art methods. Therefore, the effectiveness of this method is verified.

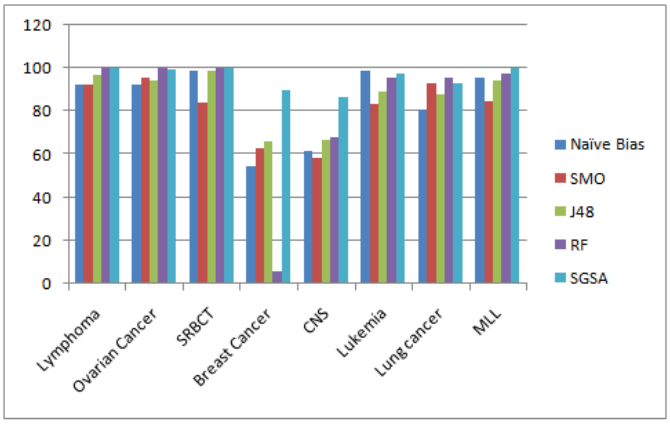

Fig 3. Comparison of the accuracy of the proposed method with state of art methods

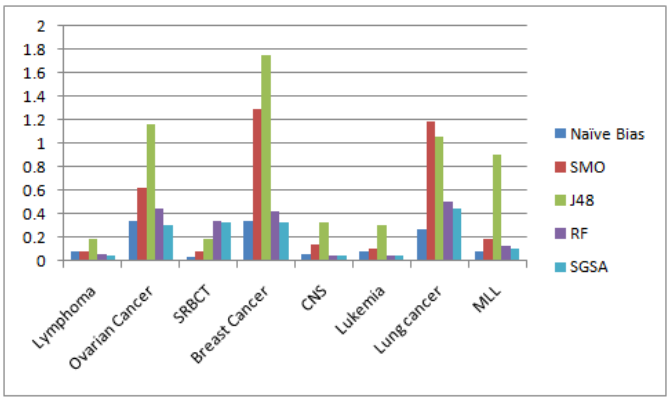

Fig 4. Comparison of the CPU cost of the proposed method with state of art methods

4 Conclusion

Extensive research has been conducted on classification methods utilizing gene expression and a variety of techniques were applied for classifying gene expression microarray data. This study offers a competent hybrid scheme for gene selection where a set of feature selection method have undertaken on multi-variant data. 70% are used for training, and 30% are used for testing and validation purposes. Furthermore, the Python programming language has proven to be a proficient instrument for generating numerical results and simulating them across various α-fold values.

Parameter values are utilized from established literature to carry out numerical simulations for the model. It is found that the anticipated way amplifies accuracy level of classifiers significantly and hence better than the existing process and the current approach [(1–10)]. Better accuracy and improved CPU cost has been achieved due to finding and involving most representative genes using correlation analysis in a classification process. As seen in Table 3 , the performance of the proposed plan is better than and sometimes comparable to existing techniques. It is also noted from Table 3 that, with feature selection either by using correlation and attribute evaluation, proposed method shows better performance on all datasets. The application of this model to an extended set of real data could reveal more insights about the characteristics of some specific gene which is invaluable for a more appropriate selection of genes during the diagnosis process. This observation further is a witness the usefulness and importance of the gene selection for cancer classification with gene expression datasets.

While this research has laid a strong foundation for the use of feature selection and classification, there are several promising directions for future work:

- Develop intuitive visualization tools to enhance the understanding and utilization of key concepts by biomedical practitioner and decision-makers.
- Perform additional comprehensive real-world case studies and validations across various fields to evaluate the feasibility and efficiency of the suggested method in different situations.

Acknowledgement

The authors thank the Government Engineering College, Bhavnagar, Gujarat, India for providing resources and other facilities to carry out this research work. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

1) Lugagne JB, Blassick CM, Dunlop MJ. Deep model predictive control of gene expression in thousands of single cells. *Nat Commun*. 2024;15:2148–2148. Available from: https://doi.org/10.1038/s41467-024-46361-1.
2) Do TN, Tran-Nguyen MT. Improved gene expression classification through multi-class support vector machines feature selection. Intelligent Systems and Data Science (ISDS). *Communications in Computer and Information Science*. 2023;1950:119–149. Available from: https://doi.org/10.1007/978-981-99-7666-9_10.
3) Ghaleb MS, Ebied HM, Tolba MF. Lung cancer stages classification based on differential gene expression. *The 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023)*. 2023;164:272–281. Available from: https://doi.org/10.1007/978-3-031-27762-7_26.
4) Senbagamalar L, Logeswari S. Genetic clustering algorithm-based feature selection and divergent random forest for multiclass cancer classification using gene expression data. *International Journal of Computational Intelligence Systems*. 2024;17:23–23. Available from: https://doi.org/10.1007/s44196-024-00416-9.
5) Das A, Chatterjee S, Sanyal G, Travieso-González CM, Awasthi S, Pinto CMA, et al. Cancer classification based on an integrated clustering and classification model using gene expression data. In: International Conference on Artificial Intelligence and Sustainable Engineering;vol. 836. Springer. 2022;p. 461–70. Available from: https://doi.org/10.1007/978-981-16-8542-2_37.
6) Abdulqader D, Abdulazeez AM, Zeebaree D. Machine learning supervised algorithms of gene selection: A Review. *Technology Reports of Kansai University*. 2020;62(3):233–277. Available from: https://www.researchgate.net/publication/341119469_Machine_Learning_Supervised_Algorithms_of_Gene_Selection_A_Review.
7) Naji AM, Filali SE, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*. 2021;191:487–492. Available from: https://doi.org/10.1016/j.procs.2021.07.062.
8) Mehla A, Deora SS, Dalal S. Improving crop yield prediction models with optimization-based feature selection and filtering approaches. *Indian Journal of Science and Technology*. 2023;16(47):4512–4524. Available from: https://doi.org/10.17485/IJST/v16i47.1602.
9) Borodulin A, Gladkov A, Gantimurov A, Kukartsev V, Evsyukov D. Using machine learning algorithms to solve data classification problems using multi-attribute dataset. *BIO Web Conf*. 2024;84:2001. Available from: https://doi.org/10.1051/bioconf/20248402001.
10) Ajmal HB, Madden MG. Dynamic Bayesian Network Learning to Infer Sparse Models From Time Series Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2022;19(5):2794–2805. Available from: https://doi.org/10.1109/TCBB.2021.3092879.
11) Michal I, Treepop W, Nicholas A, Kaehler BD. Beating Naive Bayes at Taxonomic Classification of 16S rRNA Gene Sequences. *Frontiers in Microbiology*. 2021;12. Available from: https://doi.org/10.3389/fmicb.2021.644487.