LatIA. 2025; 3:84
doi: 10.62486/latia202584
ISSN: 3046-403X

ORIGINAL

From Complexity to Clarity: Improving Microarray Classification with Correlation-Based Feature Selection

De la Complejidad a la Claridad: Mejorando la Clasificación de Microarrays con Selección de Características Basada en Correlación

Muhyeeddin Alqaraleh[1] ✉, Mowafaq Salem Alzboon[2] ✉, Mohammad Subhi Al-Batah[2] ✉, Hatim Solayman Migdadi[2] ✉

[1]Zarqa University, Faculty of Information Technology. Zarqa, Jordan.
[2]Jadara University, Faculty of Information Technology. Irbid, Jordan.

Cite as: Alqaraleh M, Salem Alzboon M, Subhi Al-Batah M, Solayman Migdadi H. From Complexity to Clarity: Improving Microarray Classification with Correlation-Based Feature Selection. LatIA. 2025; 3:84. https://doi.org/10.62486/latia202584

Submitted: 24-02-2024 **Revised:** 02-07-2024 **Accepted:** 21-11-2024 **Published:** 01-01-2025

Editor: Dr. Rubén González Vallejo

Corresponding author: Mohammad Subhi Al-Batah ✉

ABSTRACT

Gene microarray classification is yet a difficult task because of the bigness of the data and limited number of samples available. Thus, the need for efficient selection of a subset of genes is necessary to cut down on computation costs and improve classification performance. Consistently, this study employs the Correlation-based Feature Selection (CFS) algorithm to identify a subset of informative genes, thereby decreasing data dimensions and isolating discriminative features. Thereafter, three classifiers, Decision Table, JRip and OneR were used to assess the classification performance. The strategy was implemented on eleven microarray samples such that the reduced samples were compared with the complete gene set results. The observed results lead to a conclusion that CFS efficiently eliminates irrelevant, redundant, and noisy features as well. This method showed great prediction opportunities and relevant gene differentiation for datasets. JRip performed best among the Decision Table and OneR by average accuracy in all mentioned datasets. However, this approach has many advantages and enhances the classification of several classes with large numbers of genes and high time complexity.

Keywords: Feature Selection; Gene Expression Data; Correlation-Based Feature Selection Algorithm; Decision Table; JRip and OneR.

RESUMEN

La clasificación de microarrays de genes sigue siendo una tarea difícil debido al gran tamaño de los datos y al número limitado de muestras disponibles. Por lo tanto, es necesaria una selección eficiente de un subconjunto de genes para reducir los costos computacionales y mejorar el rendimiento de la clasificación. En consecuencia, este estudio emplea el algoritmo de Selección de Características Basada en Correlación (CFS) para identificar un subconjunto de genes informativos, disminuyendo así las dimensiones de los datos y aislando características discriminativas. Posteriormente, se utilizaron tres clasificadores, Tabla de Decisión, JRip y OneR, para evaluar el rendimiento de la clasificación. La estrategia se implementó en once muestras de microarrays, de modo que las muestras reducidas se compararon con los resultados del conjunto completo de genes. Los resultados observados llevan a la conclusión de que CFS elimina eficazmente características irrelevantes, redundantes y ruidosas. Este método mostró grandes oportunidades de predicción y una diferenciación relevante de genes en los conjuntos de datos. JRip tuvo el mejor rendimiento en promedio entre Tabla de Decisión y OneR en todos los conjuntos de datos mencionados. Sin embargo, este enfoque presenta muchas ventajas y mejora la clasificación de varias clases con grandes cantidades de genes y alta complejidad temporal.

Palabras clave: Selección de Características; Datos de Expresión Genética; Algoritmo de Selección de Características Basado en Correlación; Tabla de Decisión; JRip y OneR.

INTRODUCTION

Cancer remains one of the most serious global health threats, underscoring the need for early and accurate diagnosis to enhance treatment outcomes. Cancer diagnosis often involves large datasets containing gene expression levels measured by DNA microarray technology, which captures thousands of genes' activities simultaneously.[1] However, analyzing microarray gene data is particularly challenging due to its high dimensionality and the presence of many noisy, redundant genes. An important problem with these datasets is the greatly over-represented number of genes when compared to the low number of samples available; this presents obstacles to achieving accurate classification.[2] AbstractMicroarray cancer datasets usually have unrelated features and feature selection techniques are significant to microarray cancer datasets as they separate useful features, increase classification accuracy and computational efficiency.[3]

The most important contributions of feature selection are removing redundancy and noise while increasing feature relevance which contributes to improved classifier performance. Correlation Metrics (~ Symmetrical Uncertainty, Mutual Information and Pearson's correlation coefficients) -applied to quantify the relevance and redundancy in gene data.[4] They have given us information about a gene's contribution towards the accuracy of the classification and as to how independent it is from other genes, which is quite important in order to plan out an effective feature selection.[5]

High-dimensional data analysis has become very difficult which has brought about different novel algorithms, and methodologies to enhance feature selection.[6] Two more such strategies have been found to be gaining a lot of interest — both meta-heuristics algorithms known for their power to optimize feature selection especially when the size of the data is large.[7] Principal Component Analysis, Genetic Algorithm, Ant Colony Optimization, Simulated Annealing, and Particle Swarm Optimization are some of the most commonly used meta-heuristic approaches. All of these approaches are aimed at optimal feature selection — leaving the informative genes and taking the noisy genes or irrelevant genes out of the study.[8]

Recently, the Correlation-based Feature Selection (CFS) algorithm has become one of the widely adopted and efficient filter methods for feature selection for high-dimensional datasets. CFS uses a correlation-based heuristic evaluation for ranking feature subsets assuming that the perfect feature subset contains features highly correlated to the target class but not highly correlated to each other. By discarding the overlapping or irrelevant genes, this redundancy filtering guarantees that the features selected contribute independently to classification.[9]

CFS determines the potential of a subset to predict by investigating the evaluation predictiveness of individual features and the redundancy between them. It uses a combination of correlation coefficients for the degree of relevance of the feature subset to the class labels and the amount of inter-feature correlation.[10] Relevance increases as the correlation between features and class increases; and relevance decreases as the inter-correlation between features increases. The CFS selects subsets of variables according to these criteria and reduces the dimensionality of the data, while the selected variables hold the information needed for classification.[11]

CFS has a strong focus when it comes to selecting feature subsets and is often combined with search strategies to accelerate the selection of feature subset. Forward selection, backward elimination, bidirectional search, best-first search, and genetic search are among the commonly used search strategies, which have different pros and cons with respect to exploring the feature space. In this paper, the search method selected is Greedy Stepwise along with CFS for selecting the best gene subset.[12] The approach followed by Greedy Stepwise is to add or remove one feature to or from the current subset based on which feature contributes to the classification accuracy more adding to the efficiency and accuracy of subset selection.[13]

Finally, it can be concluded that using microarray datasets, feature selection plays a vital in reducing the noise which in turn reduces unnecessary computations that can confuse the learner and will facilitate the cancer classification.[14] With Greedy Stepwise search as a wrapper, CFS gives a fore fronting step to select relevant, non-redundant features and thus high accuracy of classification and lower computation. In this study we evaluate the performance of this method on several microarray datasets and we show that it can overcome the huge class-imbalanced nature in large data sets by reducing the search-space to a small group of classification-modifying genes which can serve as a basis for the development of cancer diagnosis.[15]

Related work

Every year, millions lose their life due to cancer, which is a driving factor of death worldwide. Medical practitioners find it exceedingly significant to identify cancer cells perfectly for effectively treating patients. Neural networks are a particular area of research in medical science, especially cardiology, radiology, oncology, and urology — because they can help doctors more accurately make a diagnosis. In (16), the Authors to a paper survey various types of neural network technology used in cancer classification in order to guide the reader on its application in medical discrimination. The overarching aim of this survey is to aid researchers in devising affordable, yet intuitive systems, processes and frameworks that can assist clinicians with accurate and prompt diagnosis, thereby enhancing patient outcomes.(16)

Method: In (17), the aim was to create a model to classify cancers into diagnostic categories based on gene expression signatures, using artificial neural networks (ANNs) as the classification method. The ANNs were trained on small, round blue-cell tumors (SRBCTs), a model that incorporates all four separate diagnostic categories that are often difficult to separate in the clinical arena. The ANNs not only distinguished all samples by classification, but also identified the key genes that were related to the classification process. Several of these genes have been connected with SRBCTs before, but most had not been related with such cancers, indicating novel insights. Blinded samples that had not helped develop the original model were then run with the model and all samples were classified correctly, validating the model as reliable. ANN based approaches for tumor diagnosis are elaborated in (17), paving a path for identification of possible therapeutic targets.(17)

Various classification methods have been developed and applied to differentiate disease classes at the molecular level using microarray data. Recently, hierarchical probabilistic models based on kernel-embedding techniques have emerged as highly effective tools for microarray data analysis. Initially introduced as kernel-embedded Gaussian processes (KIGPs) for binary classification of microarray gene expression data, these models were later extended to multiclass classification within a Bayesian framework. This approach incorporates an adaptive, cascading algorithm that identifies suitable feature kernels, detects potentially significant genes, and provides optimal disease class predictions (e.g., tumor/cancer) with corresponding Bayesian posterior probabilities. Simulation studies and applications on real datasets indicate that KIGPs perform near the Bayesian bound and consistently outperform or match state-of-the-art methods. A key strength of the KIGP approach lies in its capacity to capture both linear and nonlinear relationships between target disease classifications and explanatory gene expression data. This research suggests broader applications of the KIGP model for other types of high-throughput omics data and time-series omics data, particularly where linear methods are insufficient.(18)

Microarray technology enables simultaneous analysis of multiple gene expressions across time points. To classify gene expression data, methods like sparse representation (SR)-based models have been used to cluster gene data. This paper introduces a collaborative representation (CR)-based classification with regularized least squares, encoding test samples as sparse linear combinations of training samples and classifying them based on minimal representation error. This approach is less complex than traditional classifiers but maintains high accuracy. Additionally, compressive sensing reduces the high-dimensional data to a lower-dimensional space, retaining essential information and reducing computational load. Experimental results on disease subtypes, including leukemia and autism, show the CR-based method's superior stability and accuracy over traditional classifiers like support vector machines.(19)

High-throughput microarray technologies have been instrumental in genomic research, providing essential data despite analytical challenges. Their cost-effectiveness made them a popular choice initially, but advances in sequencing technologies—now more affordable and less noisy—have shifted much of the focus toward sequence data. Nonetheless, new and legacy microarray data remain valuable, offering complementary insights into biological systems and disease. The current challenge is to integrate these datasets with sequencing data for enhanced genomic analysis, a theme explored in this Special Issue.(20)

DNA microarray technology has become essential for cancer diagnosis and classification, though analyzing large gene expression datasets remains challenging. This study proposes a two-phase hybrid model combining Correlation-based Feature Selection (CFS) with improved Binary Particle Swarm Optimization (iBPSO) to select a minimal set of prognostic genes for Naive Bayes classification. Tested on 11 cancer datasets, the model demonstrated superior accuracy and efficiency, achieving up to 100 % accuracy on seven datasets with less than 1,5 % of genes selected.(21)

Gene selection and cancer classification are critical for uncovering insights within genomic data. While logistic regression is widely used for classification, it lacks inherent feature selection capabilities. This study introduces a new hybrid L1/2 + L2 regularization (HLR) function, which combines the sparsity of L1/2 with the grouping effect of L2 to identify relevant genes within logistic regression models. Additionally, a novel univariate HLR thresholding approach is proposed to update coefficient estimates, alongside a coordinate descent algorithm tailored for HLR penalized logistic regression. Empirical results and simulations show that this method performs competitively with other leading approaches.[22]

A primary challenge in microarray data analysis is the "curse of dimensionality," which can obscure valuable information and cause computational instability. Consequently, selecting relevant genes is essential. Most existing methods utilize a two-phase process involving feature selection or extraction, followed by classification. This study proposes an ANOVA-based statistical test using the MapReduce framework to select relevant features, followed by a MapReduce-based K-Nearest Neighbor (K-NN) classifier for classification. Both algorithms are implemented on the Hadoop framework, and a comparative analysis is conducted across various datasets, demonstrating the method's effectiveness.[23]

The authors in [24] proposes a three-phase hybrid approach for selecting and classifying high-dimensional microarray data. The method combines Pearson's Correlation Coefficient (PCC) with Binary Particle Swarm Optimization (BPSO) or Genetic Algorithm (GA) and integrates multiple classifiers, creating a PCC-BPSO/GA-multi-classifier framework. In the final phase, five different classifiers are applied. The PCC filter significantly enhances classification accuracy when combined with BPSO or GA, with performance varying across datasets depending on the final classifier used. Comparisons of the hybrid approach in terms of accuracy and gene selection reveal that BPSO not only outperforms GA in speed but also achieves superior accuracy when paired with PCC for feature selection.[24]

In many big-data systems, extensive information is recorded and stored for analytics; however, this data volume can hinder rather than aid optimal decision-making due to high costs of collection, storage, and processing. For example, tumor classification using high-throughput microarray data is challenging, as it often includes numerous noisy features that contribute little to reducing classification errors. The goal in such cases is to identify a minimal set of genes that effectively distinguish between classes. This study focuses on feature selection within support vector machine (SVM) classification, specifically aiming to develop an accurate binary classifier that uses a limited number of features. We propose a novel approach that iteratively adjusts a bound on the l1-norm of the classifier vector, guiding the feature set to converge toward a specified maximum size. The approach is evaluated on two real-world high-dimensional classification tasks: tumor diagnosis based on microarray gene expression data and sentiment classification from Amazon, Yelp, and IMDb reviews. Results demonstrate that the proposed method is computationally efficient, straightforward, and achieves low error rates, which are crucial for developing advanced decision-support systems.[25]

Effective analysis tools for biological data should provide clear, interpretable models. Decision trees, while promising for their transparency, often underfit gene expression data. This study introduces a multi-test decision tree (MTDT) that enhances decision trees' accuracy and stability by using multiple univariate tests at each node and incorporating alternative, lower-ranked features. Tested on multiple gene expression datasets, MTDT showed statistically superior accuracy over traditional decision trees and competitiveness with ensemble methods, outperforming its baseline by an average of 6 % on 14 datasets. MTDT offers a robust, interpretable approach suitable for high-dimensional biological data.[26]

Cancer remains one of the leading causes of mortality globally, emphasizing the need for accurate diagnosis to guide effective treatment. Correct identification of cancer cells is critical for optimal patient care. Neural networks have become a prominent research focus in medical science, particularly within fields such as cardiology, radiology, oncology, and urology, due to their potential to improve diagnostic precision. This paper surveys various neural network technologies applied in cancer classification, aiming to provide researchers with insights for developing cost-effective, user-friendly systems and methodologies that support clinicians in medical diagnostics.[27]

METHOD

We describe a method for gene selection and classification applied to high-dimensional cancer microarray data. First, we applied the Correlation-based Feature Selection (CFS) algorithm to each dataset for filtering irrelevant and redundant genes to reduce dimensionality. The filter method uses Pearson correlation to identify relationships between genes and class labels, subsequently retaining the most predictive genes.[28] CFS uses the assumption that irrelevant features are weakly correlated with the target class and redundant feature have high inter-correlation with other features. The study included the feature selected stepwise approach, a Greedy method, to accelerate the search process by limiting the search space but enhancing the algorithm in finding the optimal subset.[29]

There wore three different algorithms for classification Decision Table, JRip and, OneR. Then each classifier is chosen based on their operational characteristics since there is no model which can work perfectly on every dataset. Decision Table: Owing to its simplicity and interpretability, Decision Table evaluates feature subsets using cross-validation.[30] JRip (an implementation of RIPPER — Repeated Incremental Pruning to Produce Error Reduction) builds up rules-based models that optimize for classification accuracy. Finally, for its simplicity in classification tasks, OneR, which is a simple algorithm that produces one-rule classifiers from a single attribute, was used as a comparative baseline.[31]

We utilized both the original and comprehensive filtered datasets across 11 cancer types, such as breast cancer, CNS cancer, and leukemia for experimental procedures. The classifiers were evaluated for accuracy on each dataset with both full training, as well as 2- to 10-fold cross-validation testing. The number of genes after CFS was greatly reduced, which facilitated computation.[32] We measured the performance by comparing the classification accuracy before and after feature selection. JRip consistently achieved better accuracy on filtered datasets than its non-filtered counterparts, revealing that feature selection not only reduced the feature-space but also positively affected classifier performance.[33]

The gene selection and classification on cancer microarray dataset is explored in this paper through a correlation-based feature selection (CFS) algorithm and several classifiers. Methods Usage One of the ways that methods are applied is through analysis of datasets, where eleven high-dimensional datasets were analyzed, including Breast Cancer, CNS, and Leukaemia as types of cancer.[34,35] Gene features were selected using the CFS algorithm combined with a Greedy Stepwise search and variable were kept as a final gene if its Pearson's correlation coefficients indicated that the relationship of the gene to the target classification variable. This discards features that are both non-informative (in that they donnot add value to classification accuracy) and redundant (meaning, high correlation) features.[36,37]

For classification, three types of algorithms were tried, they are Decision table, JRip and OneR. JRip (RIPPER) is a rule-based classifier with heuristic global optimization, whereas Decision Table is a majority classifier that uses best-first search and cross-validation to assess feature subsets. Another simpler classifier for characteristic selection, OneR, chooses the single best feature based on discretionization parameter.[38,39] The effectiveness of each classifer between two datasets is presented in the figure 1, which shows that CFS opens the door to reduce the number of genes used in the models without losing accuracy, thus enhances the usability of the models through time and space. The strategy showed the capacity of improving the procedure of cancer classification in high-dimensional gene datasets.[40,41]

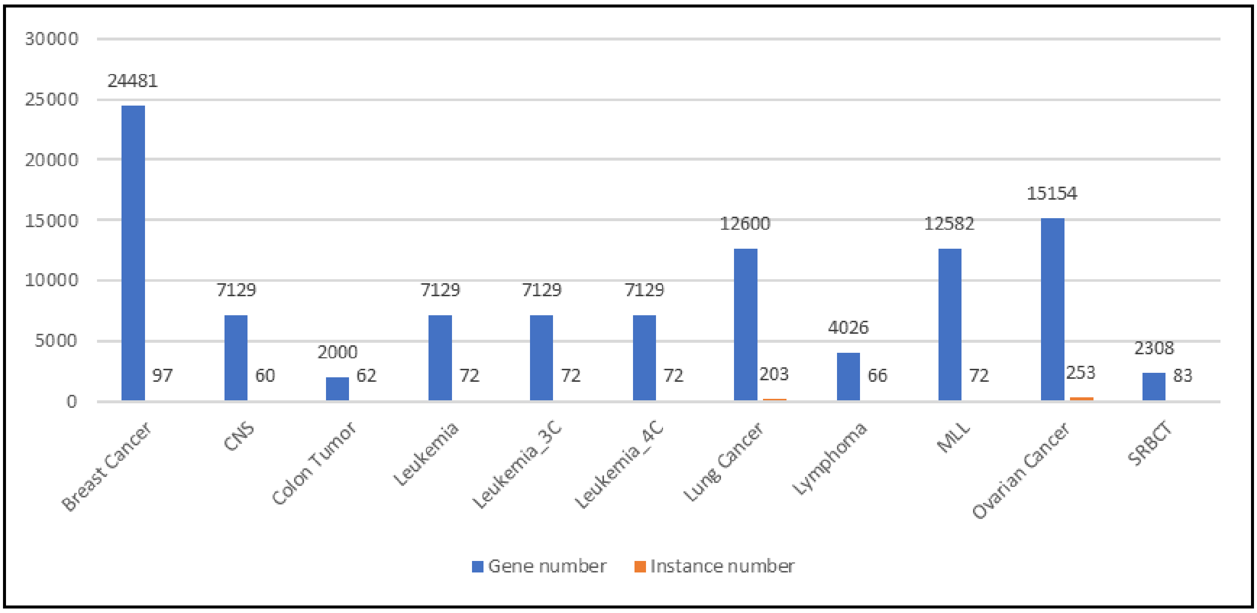

Figure 1. Gene microarray datasets

RESULTS AND DISCUSSION

Initially, the Decision Table, JRip, and OneR classifiers were applied to the original datasets. Subsequently, all eleven datasets were processed using the Correlation-based Feature Selection (CFS) algorithm to filter out irrelevant and redundant features. The filtered datasets were then classified using the same classifiers to compare classification accuracy before and after filtration. Experiments for each dataset utilized both full training and cross-validation methods, ranging from 2-fold to 10-fold validation.[42,43]

The number of selected genes post-CFS filtration is presented in Figure 2 and 3, demonstrating a substantial reduction in gene count across all datasets. Specifically, gene counts were reduced from 24,481 to 138 in Breast Cancer, from 7,129 to 39 in CNS, from 2,000 to 26 in Colon Tumor, from 7,129 to 79 in Leukemia, from 7,129 to 104 in Leukemia_3C, from 7,129 to 119 in Leukemia_4C, from 12,600 to 548 in Lung Cancer, from 4,026 to 175 in Lymphoma, from 12,582 to 142 in MLL, from 15,154 to 35 in Ovarian Cancer, and from 2,308 to 112 in SRBCT. This reduction highlights the CFS algorithm's efficiency in selecting a minimal subset of genes while preserving those most relevant for classification.[44,45]

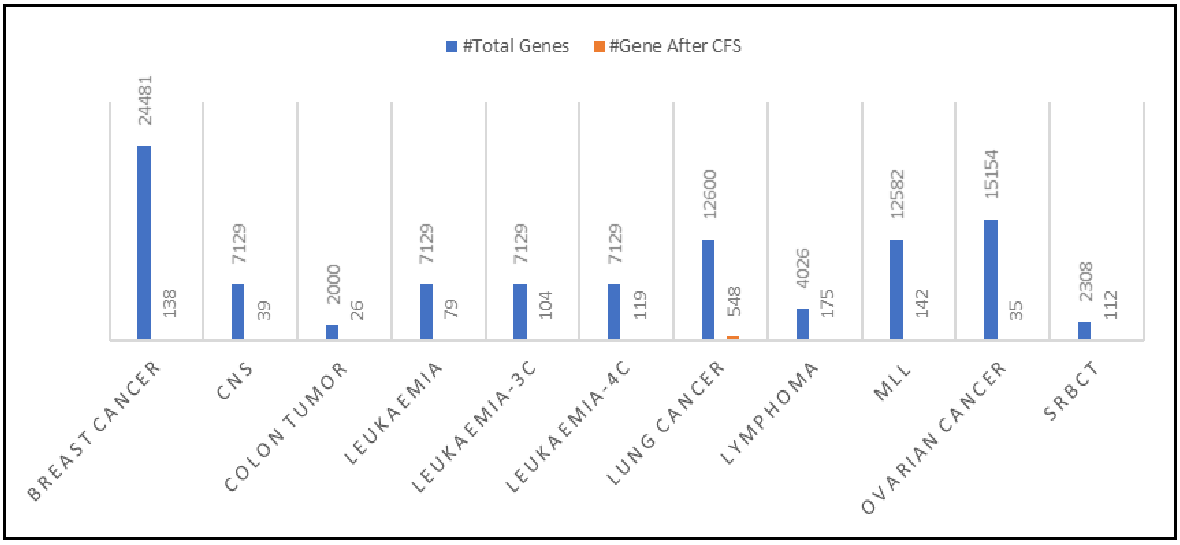

Figure 2. Number of selected genes before/after applying CFS algorithm

The classification accuracy of the classifiers on both the original and filtered datasets was assessed, with results displayed in figure 3 and 4. The highest-performing classifier for each dataset is indicated in bold. Overall, the results indicate that classifiers generally achieved higher accuracy on the filtered datasets compared to the original ones. However, in a few cases, certain classifiers performed equally on both original and filtered datasets.

Furthermore, the Decision Table and JRip classifiers consistently outperformed OneR across the datasets. Notable accuracies include 88,7 % for Breast Cancer with JRip, 90,0 % for CNS with Decision Table, 96,8 % for Colon Tumor with both Decision Table and JRip, 98,6 % for Leukemia with Decision Table, and 100,0 % for Leukemia_3C with Decision Table. Additional results include 98,6 % accuracy for Leukemia_4C with both Decision Table and JRip, 97,5 % for Lung Cancer with JRip, 100,0 % for Lymphoma with JRip, 95,8 % for MLL with JRip, 100,0 % for Ovarian Cancer with Decision Table, and 97,6 % for SRBCT with JRip.

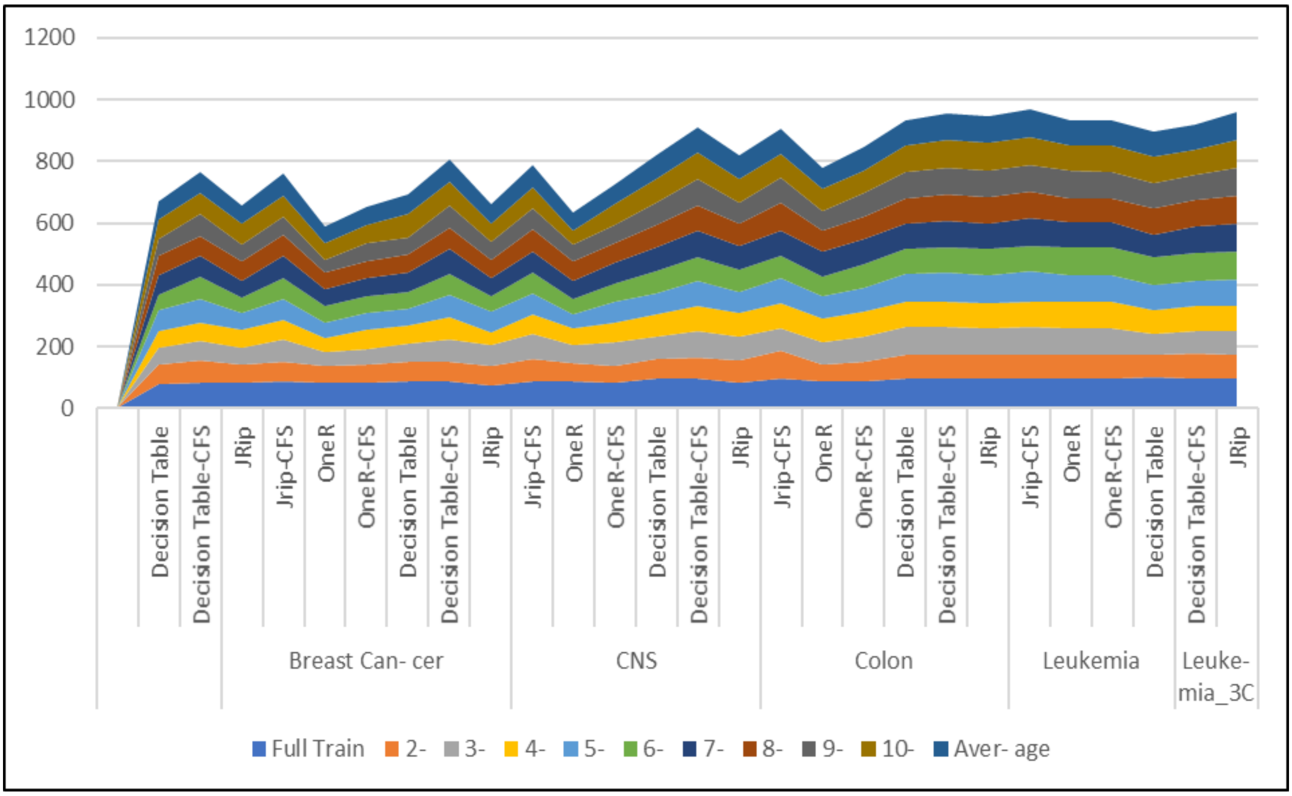

Figure 3. Accuracy for the original and filtered microarray datasets

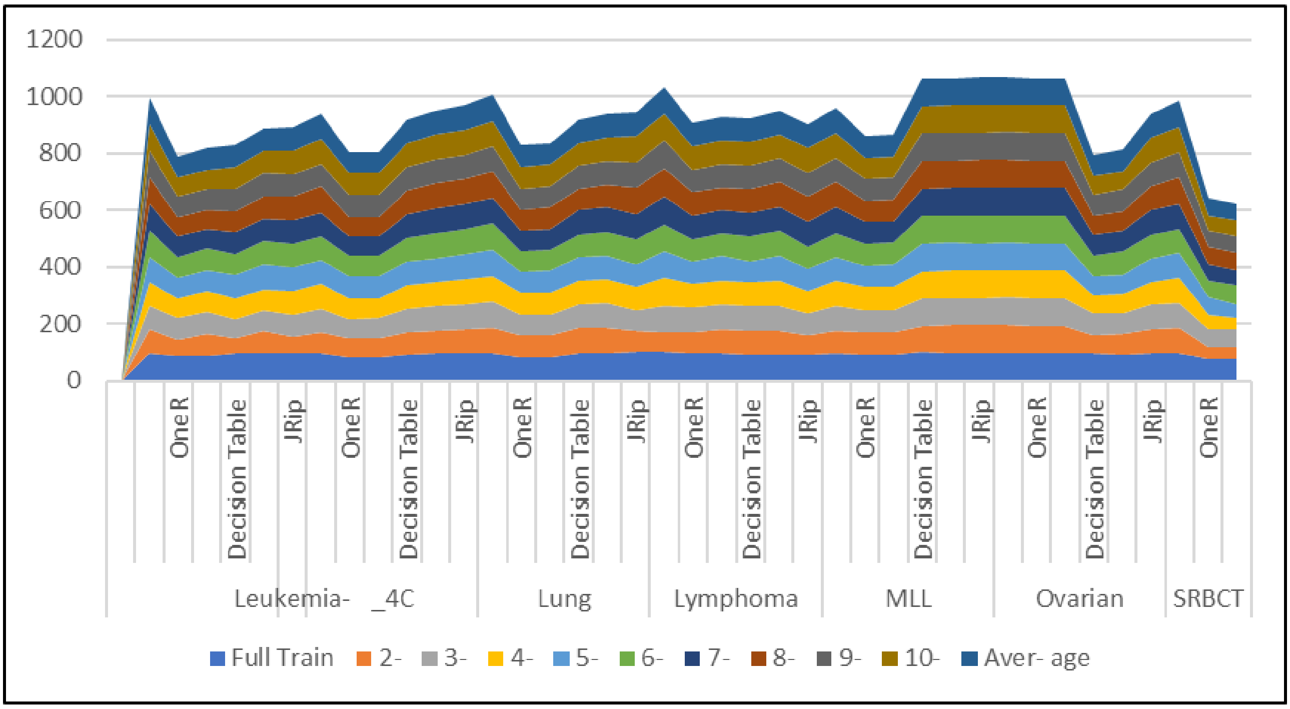

Figure 4. Accuracy for the original and filtered microarray datasets

It is evident that JRip-CFS achieved the highest average accuracy among the classifiers. For instance, on the Lymphoma dataset, JRip-CFS attained an accuracy of 93,8 %, compared to 86,1 % with JRip alone, 85,5 % with Decision Table-CFS, 83,6 % with Decision Table, 84,5 % with OneR-CFS, and 82,6 % with OneR.

Analyzing the datasets using both full training and 2-fold to 10-fold cross-validation methods, JRip consistently outperformed Decision Table and OneR in terms of accuracy. As illustrated in figure 5, the classification accuracy for the SRBCT dataset across ten tests confirmed that JRip combined with CFS yielded superior results, establishing it as the most effective method among the classifiers evaluated.

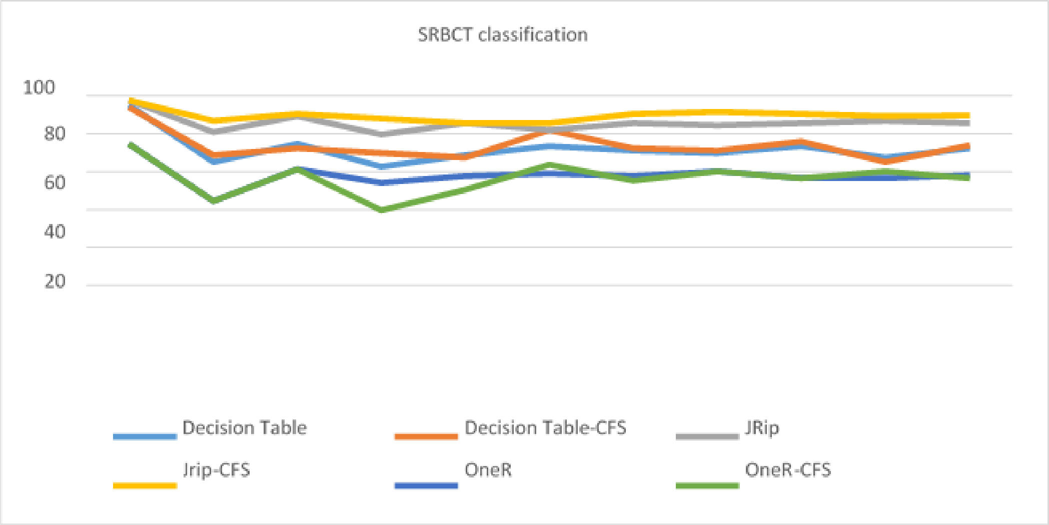

Figure 5. Accuracy of SRBCT using full training and cross validation method

Furthermore, the average accuracy across all 11 datasets, as presented in Figure 6, indicates notable differences among the classifiers. Specifically, the average accuracies were 78,2 % for Decision Table, 82,4 % for Decision Table-CFS, 80,7 % for JRip, 86,0 % for JRip-CFS, 73,1 % for OneR, and 75,3 % for OneR-CFS. These results clearly demonstrate that the feature selection process using CFS enhances classifier accuracy. Additionally, JRip consistently achieved higher average accuracy than both Decision Table and OneR, underscoring its effectiveness in this context.

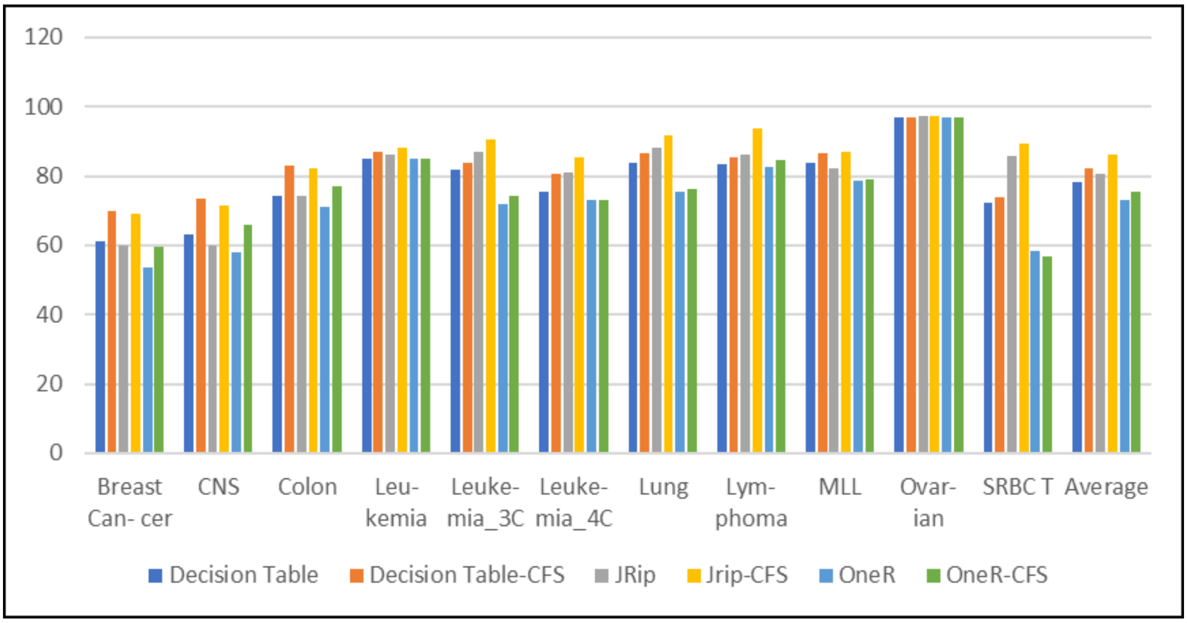

Figure 6. Accuracy of SRBCT using full training and cross validation method

CONCLUSION

To address the challenges of noisiness and high dimensionality in microarray data, this study employs Correlation-based Feature Selection (CFS) to identify relevant features. Additionally, the classifiers Decision Table, JRip, and OneR are applied to classify the microarray data. Comparative analysis shows that the classification accuracy of all classifiers improves when using filtered datasets, indicating that CFS enhances both the efficiency and accuracy of the classification process. Among the classifiers, JRip demonstrated the highest classification accuracy. Future research could extend this work by exploring the application of other feature selection techniques, such as Genetic Algorithm, Principal Component Analysis, Simulated Annealing, Ant Colony Optimization, and Particle Swarm Optimization.

REFERENCES

1. Wahed MA, Alqaraleh M, Alzboon MS, Al-Batah MS. Application of Artificial Intelligence for Diagnosing Tumors in the Female Reproductive System: A Systematic Review. Multidiscip. 2025;3:54.

2. Wahed MA, Alqaraleh M, Alzboon MS, Al-Batah MS. Evaluating AI and Machine Learning Models in Breast Cancer Detection: A Review of Convolutional Neural Networks (CNN) and Global Research Trends. LatIA. 2025;3:117.

3. Muhyeeddin Alqaraleh, Mohammad Al-Batah, Mowafaq Salem Alzboon EA. Automated quantification of vesicoureteral reflux using machine learning with advancing diagnostic precision. Data Metadata. 2025;4:460.

4. Al-shanableh N, Alzyoud M, Al-husban RY, Alshanableh NM, Al-Oun A, Al-Batah MS, et al. Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: A Detailed Examination of Hospital Data. Data Metadata. 2024;3:363.

5. Al-Batah MS, Salem Alzboon M, Solayman Migdadi H, Alkhasawneh M, Alqaraleh M. Advanced Landslide Detection Using Machine Learning and Remote Sensing Data. Data Metadata [Internet]. 2024 Oct 7;3. Available from: https://dm.ageditor.ar/index.php/dm/article/view/419/782

6. Alqaraleh M, Abdel M. Advancing Medical Image Analysis : The Role of Adaptive Optimization Techniques in Enhancing COVID-19 Detection , Lung Infection , and Tumor Segmentation Avances en el análisis de imágenes médicas : el papel de las técnicas de optimización adaptativa para. LatIA. 2024;2(74).

7. Alzboon MS, Alqaraleh M, Al-Batah MS. AI in the Sky: Developing Real-Time UAV Recognition Systems to Enhance Military Security. Data Metadata. 2024;3(417).

8. Mohammad Al-Batah, Mowafaq Salem Alzboon, Muhyeeddin Alqaraleh FA. Comparative Analysis of

Advanced Data Mining Methods for Enhancing Medical Diagnosis and Prognosis. Data Metadata. 2024;3:465.

9. Ahmad A, Alzboon MS, Alqaraleh MK. Comparative Study of Classification Mechanisms of Machine Learning on Multiple Data Mining Tool Kits. Am J Biomed Sci Res 2024 [Internet]. 2024;22(1):577–9. Available from: www.biomedgrid.com

10. Al-Batah MS, Alzboon MS, Alzyoud M, Al-Shanableh N. Enhancing Image Cryptography Performance with Block Left Rotation Operations. Appl Comput Intell Soft Comput. 2024;2024(1):3641927.

11. Alqaraleh M, Alzboon MS, Al-Batah MS, Wahed MA, Abuashour A, Alsmadi FH. Harnessing Machine Learning for Quantifying Vesicoureteral Reflux: A Promising Approach for Objective Assessment. Int J Online \& Biomed Eng. 2024;20(11).

12. Alzboon MS, Al-Batah M, Alqaraleh M, Abuashour A, Bader AF. A Comparative Study of Machine Learning Techniques for Early Prediction of Diabetes. In: 2023 IEEE 10th International Conference on Communications and Networking, ComNet 2023 - Proceedings. 2023. p. 1–12.

13. Alzboon MS, Al-Batah M, Alqaraleh M, Abuashour A, Bader AF. A Comparative Study of Machine Learning Techniques for Early Prediction of Prostate Cancer. In: 2023 IEEE 10th International Conference on Communications and Networking, ComNet 2023 - Proceedings. 2023. p. 1–12.

14. Alzboon MS, Al-Batah MS, Alqaraleh M, Abuashour A, Bader AFH. Early Diagnosis of Diabetes: A Comparison of Machine Learning Methods. Int J online Biomed Eng. 2023;19(15):144–65.

15. Al-Batah MS, Alzboon MS, Alazaidah R. Intelligent Heart Disease Prediction System with Applications in Jordanian Hospitals. Int J Adv Comput Sci Appl. 2023;14(9):508–17.

16. Dash R. An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data. J King Saud Univ - Comput Inf Sci. 2021;33(2):195–207.

17. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673–9.

18. Cheung LWK. Classification approaches for microarray gene expression data analysis. Methods Mol Biol. 2012;802:73–85.

19. Shen L, Jiang H, He M, Liu G. Collaborative representation-based classification of microarray gene expression data. PLoS One. 2017;12(12):e0189533.

20. Ruskin H. Computational Modeling and Analysis of Microarray Data: New Horizons. Microarrays. 2016;5(4):26.

21. Jain I, Jain VK, Jain R. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification. Appl Soft Comput. 2018;62:203–15.

22. Huang HH, Liu XY, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2 +2 regularization. PLoS One [Internet]. 2016;11(5):e0149675. Available from: https://doi.org/10.1371/journal.pone.0149675

23. Kumar M, Rath NK, Swain A, Rath SK. Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor. In: Procedia Computer Science. 2015. p. 301–10.

24. Hameed SS, Muhammad FF, Hassan R, Saeed F. Gene selection and classification in microarray datasets using a hybrid approach of PCC-BPSO/GA with multi classifiers. J Comput Sci. 2018;14(6):868–80.

25. Ghaddar B, Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. Eur J Oper Res. 2018;265(3):993–1004.

26. Czajkowski M, Grześ M, Kretowski M. Multi-test decision tree and its application to microarray data

classification. Artif Intell Med. 2014;61(1):35–44.

27. Agrawal S, Agrawal J. Neural network techniques for cancer prediction: A survey. In: Procedia Computer Science. 2015. p. 769–74.

28. Alzboon MS, Qawasmeh S, Alqaraleh M, Abuashour A, Bader AF, Al-Batah M. Machine Learning Classification Algorithms for Accurate Breast Cancer Diagnosis. In: 2023 3rd International Conference on Emerging Smart Technologies and Applications, eSmarTA 2023. 2023.

29. Alzboon MS, Al-Batah MS. Prostate Cancer Detection and Analysis using Advanced Machine Learning. Int J Adv Comput Sci Appl. 2023;14(8):388–96.

30. Alzboon MS, Qawasmeh S, Alqaraleh M, Abuashour A, Bader AF, Al-Batah M. Pushing the Envelope: Investigating the Potential and Limitations of ChatGPT and Artificial Intelligence in Advancing Computer Science Research. In: 2023 3rd International Conference on Emerging Smart Technologies and Applications, eSmarTA 2023. 2023.

31. Alzboon MS, Bader AF, Abuashour A, Alqaraleh MK, Zaqaibeh B, Al-Batah M. The Two Sides of AI in Cybersecurity: Opportunities and Challenges. In: Proceedings of 2023 2nd International Conference on Intelligent Computing and Next Generation Networks, ICNGN 2023. 2023.

32. Alzboon M. Semantic Text Analysis on Social Networks and Data Processing: Review and Future Directions. Inf Sci Lett. 2022;11(5):1371–84.

33. Alzboon MS. Survey on Patient Health Monitoring System Based on Internet of Things. Inf Sci Lett. 2022;11(4):1183–90.

34. Alzboon MS, Aljarrah E, Alqaraleh M, Alomari SA. Nodexl Tool for Social Network Analysis. Vol. 12, Turkish Journal of Computer and Mathematics Education. 2021.

35. Al-Batah MS, Al-Eiadeh MR. An improved discreet Jaya optimisation algorithm with mutation operator and opposition-based learning to solve the 0-1 knapsack problem. Int J Math Oper Res. 2023;26(2):143-69.

36. Alomari SA, Alqaraleh M, Aljarrah E, Alzboon MS. Toward achieving self-resource discovery in distributed systems based on distributed quadtree. J Theor Appl Inf Technol. 2020;98(20):3088–99.

37. Al-Batah MS, Al-Eiadeh MR. An improved binary crow-JAYA optimisation system with various evolution operators, such as mutation for finding the max clique in the dense graph. Int J Comput Sci Math. 2024;19(4):327-38.

38. Al-Batah M, Zaqaibeh B, Alomari SA, Alzboon MS. Gene Microarray Cancer classification using correlation based feature selection algorithm and rules classifiers. Int J online Biomed Eng. 2019;15(8):62–73.

39. Al-Batah MS. Modified recursive least squares algorithm to train the hybrid multilayered perceptron (HMLP) network. Appl Soft Comput. 2010;10(1):236-44.

40. Al Tal S, Al Salaimeh S, Ali Alomari S, Alqaraleh M. The modern hosting computing systems for small and medium businesses. Acad Entrep J. 2019;25(4):1–7.

41. Al-Batah MS. Testing the probability of heart disease using classification and regression tree model. Annu Res Rev Biol. 2014;4(11):1713-25.

42. Alzboon MS. Internet of things between reality or a wishing-list: a survey. Int J Eng \& Technol. 2018;7(2):956–61.

43. Al-Batah MS. Integrating the principal component analysis with partial decision tree in microarray gene data. IJCSNS Int J Comput Sci Netw Secur. 2019;19(3):24-29.

44. Alzboon M, Alomari SA, Al-Batah MS, Banikhalaf M. The characteristics of the green internet of things

and big data in building safer, smarter, and sustainable cities. Int J Eng \& Technol. 2017;6(3):83-92.

45. Al-Batah MS. Ranked features selection with MSBRG algorithm and rules classifiers for cervical cancer. Int J Online Biomed Eng. 2019;15(12):4.

FINANCING

This work is supported from Jadara University under grant number [Jadara-SR-Full2023], Zarqa University, and The Hashemite University.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

AUTHORSHIP CONTRIBUTION

Conceptualization: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.

Data curation: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Al-Batah.

Formal analysis: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Al-Batah.

Funding acquisition: Mowafaq Salem Alzboon, Mohammad Al-Bata.

Research: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.

Methodology: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon.

Project management: Mowafaq Salem Alzboon, Mohammad Al-Bata.

Software: Muhyeeddin Alqaraleh, Mohammad Al-Batah, Hatim Migdadi.

Supervision: Mowafaq Salem Alzboon, Mohammad Al-Bata.

Validation: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.

Display: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.

Drafting - original draft: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.

Writing: Muhyeeddin Alqaraleh, Mowafaq Salem Alzboon, Mohammad Subhi Al-Batah, Hatim Solayman Migdadi.