Hybrid Filter Feature Selection for Improving Cancer Classification in High-Dimensional Microarray Data

Oluwabukunmi Oyegbile, Faisal Saeed, Samer Bamansoor

DAAI Research Group, College of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, UK
oluwabukunmi.oyegbile@mail.bcu.ac.uk. faisal.saeed@bcu.ac.uk and samer.bamansoor@bcu.ac.uk

Abstract: In this study, we present a novel approach to improve cancer classification using high-dimensional microarray data. The proposed method combines a hybrid filter and a genetic algorithm-based feature selection process, incorporating Chi-square and Recursive Feature Elimination (RFE) techniques to identify critical gene expressions for cancer classification. Experiments using diverse datasets have yielded significant results. In the Lung Cancer Dataset, Logistic Regression Analysis (LR) and Support Vector Machine (SVM) achieved remarkable accuracy rates of 97.56%, with a precision and recall of 98.0%, resulting in an F1-score of 97.0%. This highlights the effectiveness of the feature selection method in enhancing classification accuracy. In the Ovarian Cancer Dataset, Gradient Boosting emerged as the top-performing classifier, achieving an accuracy of 92.85% along with precision, recall, and F1-score values of 94.0%, 93.0%, and 92.0%, respectively. These results demonstrate the versatility of the proposed feature-selection approach. This demonstrates the adaptability of the proposed feature selection technique in improving classifier performance. In summary, the hybrid filter and genetic algorithm-based feature selection method, incorporating Chi-square and RFE, proved to be a valuable tool for enhancing cancer classification in high-dimensional microarray data. The consistently high accuracy, precision, recall, and F1-score across diverse cancer datasets underscore the effectiveness and versatility of the proposed approach, holding promise for the development of more accurate cancer classification models in the future.

Keywords: cancer classification, hybrid Feature Selection, microarray dataset, genetic algorithm,

1 Introduction

Significant advancements in health informatics have been fueled by research focusing on bioinformatics, cheminformatics, cancer prediction, and other related fields. Advances in DNA microarray technology have revolutionized the field of biology by enabling simultaneous monitoring of genes in a single experiment [1]. Despite progress, cancer continues to pose a formidable threat to human life, with its incidence increasing over time. Early detection of this deadly disease remains a critical challenge in health informatics [2].

A powerful method for the classification, diagnosis, and treatment of cancer that has emerged in recent years is the analysis of microarray-based gene expression data. These datasets contain thousands of genes as features and capture crucial molecular information from patient samples [3]. However, Various studies have explored the prediction of gene expression, cancer classification, and discovery of new bioactive molecules using machine learning methodologies. In particular, the analysis of microarray datasets containing information about human genes and their expression has opened up new avenues for efficient research in bioinformatics and cancer prediction [4]. However, this progress is not without its challenges. The sheer number of existing genes, now reaching hundreds of thousands, surpasses the available dataset sizes, leading to what is known as the "curse of dimensionality." As a result, the analysis of microarray datasets for cancer classification has become more complex and computationally demanding. Moreover, the presence of redundant and irrelevant features within datasets further hinders the computational efficiency and accuracy of cancer prediction models [5]. To address these issues and improve the performance of machine-learning techniques, feature-selection methods have been employed to identify the most relevant features in cancerous microarray datasets. Various approaches including filter, wrapper, and hybrid methods have been used for this purpose [6]. Filter methods rank individual features and select a subset without using a learning algorithm, resulting in faster processing. In contrast, wrapper methods employ a learning algorithm to evaluate the feature subset, leading to higher classification accuracies but higher computational costs [7]. The hybrid approach to feature selection capitalizes on the advantages of both the filter and wrapper methods, combining the speed of the filter approach with the accuracy of the wrapper approach. Research studies have demonstrated that using a hybrid approach is more efficient and effective compared to applying the filter and wrapper methods [8].

This study explored the use of filter and wrapper feature detection techniques to improve cancer classification performance when dealing with high-dimensional microarray datasets. The primary objective is to identify a well-suited subset of features that can effectively improve the performance of machine-learning methods in cancer classification. To achieve this, the study utilized a filter method, Chi-squared (CS), to compute scores for each feature within the microarray cancer datasets. These scores play a crucial role in determining the relevance and importance of features for the classification task. Furthermore, by synergizing these filter methods with the wrapper method selection, the research aims to optimize the feature subset, leading to more accurate and efficient cancer classification outcomes.

2 Related Works

This section presents the filter computational techniques explored by various researchers to efficiently select relevant features or genes from a vast amount of data. In [9], the authors proposed a feature selection algorithm based on an Artificial Neural Network

(ANN) for analyzing transcriptomic data from multi-cancer samples. This study aimed to improve the cancer classification accuracy using the filter method as a filter selection method alongside the ANN algorithm as a machine learning classifier. The results revealed that XGBoost outperformed the SVM, achieving an impressive accuracy score of 90.71%. However, the study lacks a comprehensive analysis of the limitations or challenges faced during the implementation of the proposed approach, leaving room for further investigation into potential drawbacks.

Authors in [10] introduced an unsupervised feature selection algorithm for multiclass cancer classification using gene expression RNA-Seq data. They employed filter methods and implemented a voting system with different counting values, resulting in an average accuracy of 98.81%. Although the study demonstrates high accuracy, it does not thoroughly discuss potential limitations or challenges in applying the unsupervised feature selection algorithm, which could be valuable information for future research. In [11], machine learning techniques were utilized in conjunction with a filter-based feature selection method to classify breast cancer types. This study evaluated four widely used machine learning algorithms, namely K-nearest neighbor (kNN), Naïve Bayes (NGB), Decision Trees (DT), and Support Vector Machines (SVM), on the Breast Cancer dataset. The reported accuracy scores for the kNN, NGB, DT, and SVM were 87 %, 85%, 87%, and 90% for SVM. Despite the promising results, the study did not explicitly discuss any limitations or challenges faced during the application of the filter-based feature selection method, leaving potential gaps in understanding the robustness of the approach. In [12], the authors investigated the application of classification of lung adenocarcinoma and lung squamous cancers, along with biomarker identification and gene expression analysis, using filter-based feature selection methods. This study focused on utilizing the Random Forest (RF) algorithm for feature selection on the Lung Cancer dataset, achieving a remarkable accuracy score of 90%. However, the study does not delve into potential limitations or challenges encountered during the implementation of the Random Forest algorithm for feature selection, leaving room for further exploration in this aspect. In [13], the authors studied the prediction of lung cancer using gene expression data and deep learning by employing KL divergence gene selection as a filter method. They employed a Deep Neural Network (DNN) for feature selection on the Lung Cancer dataset, achieving an impressive accuracy score of 99%. While the results show the effectiveness of the proposed approach, the study lacks a discussion of potential limitations or challenges associated with the KL divergence gene selection method, providing an opportunity for future research to address these aspects. The study [14] presented a novel approach for cancer classification in multi-cancer samples using filter methods for feature selection and the k-Nearest Neighbor algorithm. The proposed method achieved an outstanding accuracy score of 100%, surpassing the performance of existing techniques. Despite the remarkable results, the study does not explicitly discuss potential limitations or challenges in applying filter methods for feature selection, presenting an area for future research. In [15], the authors introduced a novel filter method for feature selection and evaluated its impact on classification performance using three different classifiers: Multilayer Perceptron (MLP), Sequential Minimal Optimization (SMO), and Naive Bayes Classifier. The study focuses on two lung cancer datasets and reports accuracy scores of 100%, 96.42%, and

98.59% for MLP, SMO, and Naive Bayes Classifier, respectively. While the study highlights the significant improvement in classification performance, it does not thoroughly discuss the potential limitations or challenges associated with the proposed graph-based feature selection method, providing an opportunity for future research to address these aspects. The study [16] presented a robust approach to pan-cancer classification using wrapper methods for feature selection, incorporating Genetic Algorithm and k-Nearest Neighbor algorithms. The model achieved an impressive accuracy score of over 90% for the test set samples, successfully identifying the majority of tumor types among the 31 categories, with only three exceptions. While the study demonstrates high accuracy, it does not extensively discuss the potential limitations or challenges associated with the wrapper methods employed, leaving room for further exploration in this regard. In [17], the authors proposed a feature-selection strategy using Wrapper Methods and multiple Support Vector Machine (SVM) techniques. This study focused on breast cancer data and evaluated four approaches: SVM-RFE-GS, SVM-RFE-PSO, SVM-RFE-GA, and RFFS-GS. The accuracy scores obtained were 91%, 91.68%, 91.34%, and 92.19%, respectively. The findings of this study demonstrate the effectiveness of the proposed feature selection strategy for accurately classifying breast cancer based on gene expression data. However, the study does not thoroughly discuss the potential limitations or challenges associated with the various SVM techniques employed, providing an opportunity for future research to delve into these aspects. In [18], the authors proposed a novel approach for biomarker identification using the wrapper method. They applied Support Vector Machines (SVM) and SVM-Recursive Feature Elimination (SVM-RFE) for feature selection on the Tumor dataset, achieving an impressive accuracy score of 99.68%. This study showcases the efficacy of their method in accurately identifying biomarkers, emphasizing its potential in the field of bioinformatics and medical research. While the study presents promising results, it does not extensively discuss potential limitations or challenges associated with the proposed approach, providing an opportunity for future research to explore these aspects.

3 Methodology

3.1 Data Collection

A comprehensive compilation of diverse cancer types was meticulously gathered for analysis, drawing on the research carried out in [19]. Encompassing the Lung Cancer Dataset, Ovarian Cancer, Mixed Lineage Leukemia (MLL), and Acute Myeloid Leukemia (AML-3) dataset, this dataset contains microarray data that provide extensive genetic insights. The size and dimensionality of each dataset presented unique challenges and opportunities for the proposed research.

The Lung Cancer Dataset, which contains 203 data points and 12,601 columns, introduces the challenge of handling high-dimensional data. The Ovarian Cancer Dataset, comprising 253 rows and 15,155 columns, emphasizes the need for advanced techniques to sift through vast amounts of information. The MLL dataset with 72 rows and 12,583 columns offers a sizable dataset with a considerable number of attributes. The AML-3 dataset, consisting of 72 rows and 7,130 columns, highlights the challenge of dealing with high-dimensional data even with a relatively smaller number of data entries.

3.2 Data Preparation/Preprocessing

Given the high-dimensional nature of microarray data and the critical importance of accurate feature selection for cancer classification, a meticulous approach was undertaken to enhance the reliability of the subsequent analyses. First, the collected microarray datasets were subjected to comprehensive data cleaning and integration. Measures to eliminate duplicate records and address missing values were meticulously executed, thereby fostering data consistency and completeness. Integration, normalization, and transformation methods were applied to standardize the expression levels of genes across the datasets. This normalization, particularly through Z-score normalization, aimed to mitigate any biases arising from variations in measurement technologies.

Logarithmic transformations were employed to rectify the skewed distribution of gene expression values, aligning them with a more Gaussian-like distribution. Recognizing the challenge of high dimensionality intrinsic to microarray data, dimensionality reduction techniques have been enlisted. Principal Component Analysis (PCA) was harnessed to distill the dataset to its most informative components while upholding crucial insights, thereby mitigating the issues associated with the "curse of dimensionality." The meticulous selection of pertinent features was integral to the study's objective. To this end, a hybrid approach was adopted. In the preliminary stages, filter methods such as correlation analysis and Chi-squared were instrumental in the identification of promising features. The wrapper method iteratively fine-tunes the feature subset and optimizes it for enhanced performance in cancer classification. The key advantage of the wrapper method is that it considers the interplay between features within a specific classification context. To address the potential imbalance in class distribution among different cancer types, strategies such as oversampling and undersampling were considered to ensure equitable representation during model training. Upholding quality assurance, post-preprocessing data validation was executed to affirm the efficacy of the applied techniques and ascertain adherence to statistical assumptions.

3.3 Modeling

In the context of cancer classification using high-dimensional microarray data, the proposed approach combines the strengths of both filter and wrapper methods for feature selection, resulting in a hybrid model. The filter method employed here is Chi-Square, which measures the dependence between categorical variables. It assists in selecting relevant features that exhibit strong associations with the target classes. On the other hand, the wrapper method involves Recursive Feature Selection (RFE), a technique that recursively eliminates less informative features, striving to find an optimal subset that enhances the model's performance. The process of model training and validation was initiated by applying a hybrid feature selection method. Chi-Square is initially utilized to rank the features according to their relevance to the target class labels. This helps in narrowing the feature pool and retaining those most likely to contribute to accurate classification. Subsequently, RFE was employed, which iteratively prunes features with the least significance, producing a subset of features with higher discriminatory power. The diagram in Fig.1 illustrates the Recursive Feature Selection (RFE).

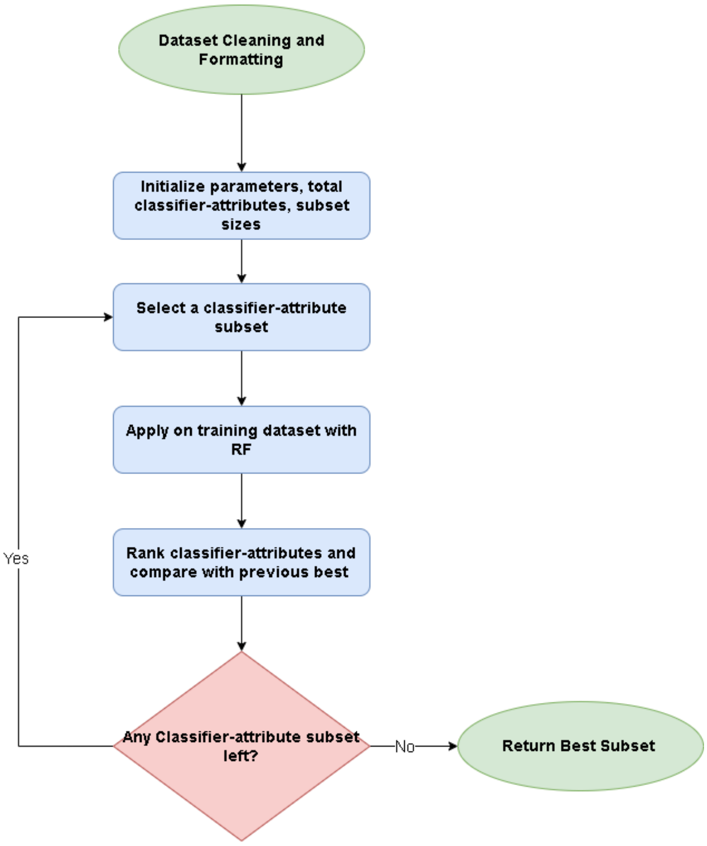
Dataset Cleaning and Formatting
Initialize parameters, total classifier-attributes, subset sizes
Select a classifier-attribute subset
Apply on training dataset with RF
Rank classifier-attributes and compare with previous best
Any Classifier-attribute subset left?
Yes
No
Return Best Subset

Fig. 1. Recursive Feature Selection

With the selected features in hand, four distinct classification algorithms were then employed: K-Nearest Neighbor (kNN), Naïve Bayes (NB), Decision Trees (DT), and Support Vector Machines (SVM). Each algorithm was trained on a refined feature set and optimized using the appropriate hyperparameters. The aim was to assess the performance of each classifier in the context of the hybrid feature selection method. To ensure the generalizability of the model, a robust cross-validation strategy was employed. K-fold cross-validation is commonly employed, where the dataset is divided into k subsets of approximately equal size. In each iteration, one of these subsets was designated as the validation set, whereas the remaining k-1 subsets were used for training. This process was repeated k times, with each subset serving as the validation set. The results from these iterations were averaged to yield a more reliable estimate of model performance. This approach mitigates the risk of overfitting and provides a better understanding of how the model is likely to perform on unseen data.

3.4 Application of Machine Learning

Taking a closer look at the execution of the machine-learning stage, the initial step involves feature selection. The preprocessed microarray data were subjected to Chi-Square analysis to assess the statistical significance of each feature for the target cancer classes. This entails computing the Chi-Square statistic for each feature to measure the dependency between the feature's presence and the different cancer classifications. Features were then ranked based on their Chi-Square scores, with higher scores indicating stronger associations with the target classes. Ex-panding on the Chi-Square-selected features, the subsequent step incorporates Recursive Feature Selection (RFE). This method utilizes a machine-learning algorithm (e.g., LR, K-NN, RF, NB, DT, and SVM) to evaluate the importance of each feature. The algorithm is initiated by considering all the features and iteratively eliminating the least important feature in each iteration. After removing a feature, the algorithm reassessed the performance of the model using the remaining features. This process repeats until a predetermined number of features is reached.

The refined feature subset obtained from the RFE process serves as input for training the selected machine learning algorithms. Each algorithm, such as kNN, NB, DT, and SVM, is applied to the data. However, before training, it is crucial to conduct hyperparameter tuning to optimize the algorithms' performances. This involves systematically exploring a range of hyperparameter values. For instance, in kNN, the optimal number of neighbors may be explored in the range of 1 to 20, and in SVM, the kernel type and associated parameters are tuned within specific ranges, such as linear, polynomial, and radial basis function (RBF) kernel with corresponding parameter ranges. While the paper acknowledges the application of hyperparameter tuning, providing specific details on the ranges explored enhances reproducibility. Following the optimization of the algorithm's hyperparameters, the training phase begins. The algorithms learn from the training data, capturing patterns and relationships between the selected features and cancer classes. Post-training, the models are evaluated for their performance using validation techniques to ensure robust model assessment, with cross-validation being a commonly employed method.

3.5 Evaluation Metrics

To quantitatively measure the models' performance, a set of well-established evaluation metrics is employed. These metrics provide insights into the various aspects of model effectiveness. The confusion matrix provides valuable insights into how well the models perform in differentiating between cancer and non-cancer instances using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Under the ROC curve (AUC-ROC), the model's ability to discriminate between positive and negative instances across various probability thresholds is assessed.

4 Experiments

In this section, we provide comprehensive documentation of all the experiments conducted during this research. These experiments are rooted in the framework outlined in Section 3.2, which offers an overview of the proposed process model. In this model, we delineate each stage of the process, providing a clear structure for the research endeavors.

The conducted experiments were executed using the Python programming language, with the Integrated Development Environment (IDE) of choice being Jupyter Notebook. We initiated the research process by meticulously preparing and transforming the data. To accomplish this, we adopted a hybrid approach combining Chi-squared as the filter method and Recursive Feature Elimination (RFE) as the wrapper method. This preprocessing step was diligently applied to the datasets before subjecting them to various machine-learning algorithms. Now, let us delve into each phase considering the 4 different datasets.

4.1 Lung Cancer Dataset

In the lung cancer dataset, seven machine-learning models were trained and evaluated. Support Vector Machines (SVM) and Logistic Regression emerged as the top-performing models, both achieving an impressive accuracy of 97.56%. As shown in Table 1, the models demonstrated exceptional precision and recall, excelling in different classes. Naïve Bayes, Decision Trees, Random Forest, k-Nearest Neighbors (k-NN), and XGBoost also showcased commendable performance, with accuracies ranging from 90.24% to 95.12%. The choice of the best model depends on specific classification goals and considerations related to interpretability and computational complexity.

The results highlight the robustness of the machine learning models in classifying Lung Cancer cases. The combination of feature selection techniques, including the Chi-Squared test and Recursive Feature Elimination (RFE), helped streamline the feature set and improve the models' performance. The research emphasizes the importance of tailoring the choice of classifier to the unique requirements of a machine learning project, as each model excelled in different aspects, showcasing the versatility of these algorithms in healthcare applications.

Table 1. Classification Results for the lung dataset

Models	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
XGBoost	85.71	88.00	86.00	83.00
Naive Bayes	85.71	81.00	86.00	82.00
KNN	**92.85**	86.00	**93.00**	89.00
Decision Tree	78.57	80.00	79.00	78.00
Gradient Boosting	**92.85**	**94.00**	**93.00**	**92.00**
Random forest	**92.85**	86.00	**93.00**	89.00

4.2 Ovarian Cancer Dataset

In the analysis of the Ovarian Cancer dataset, a thorough exploration of data was conducted, revealing the dataset's substantial dimensionality with 15,155 columns and 253 rows. The presence of missing values in specific columns was addressed, ensuring data integrity and label encoding was applied to the categorical "Class" column to enable compatibility with machine learning algorithms. Feature selection was performed using the chi-squared test and Recursive Feature Elimination (RFE), with the top 50% of important features selected for modeling. The training process followed a consistent and systematic approach, encompassing various machine learning models, including Naïve Bayes, Random Forest, Decision Trees, Gradient Boosting, XGBoost, and k-NN. As shown in Table 2, the evaluation revealed that Random Forest, Gradient Boosting, and k-NN exhibited the highest accuracy, reaching 92.85%, indicating their robustness in making accurate predictions.

While accuracy was a pivotal metric in the model evaluation, the analysis emphasized that the choice of the best classifier should consider factors beyond accuracy alone, such as model complexity, interpretability, and computational efficiency. This research underlines the importance of a structured and standardized approach to model training and evaluation and provides valuable insights into the strengths and weaknesses of each classifier. This comprehensive analysis aids in making informed decisions about selecting the most suitable model for specific classification tasks, ensuring the reliable and effective application of machine learning techniques to healthcare datasets such as Ovarian Cancer.

Table 2. Classification Results for Ovarian Cancer dataset

Models	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
XGBoost	85.71	88.00	86.00	83.00
Naive Bayes	85.71	81.00	86.00	82.00
KNN	**92.85**	86.00	**93.00**	89.00
Decision Tree	78.57	80.00	79.00	78.00
Gradient Boosting	**92.85**	**94.00**	**93.00**	**92.00**
Random forest	**92.85**	86.00	**93.00**	89.00

4.3 Mixed Lineage Leukemia (MLL)

In the analysis of the MLL Cancer dataset, feature selection and model training followed a structured and consistent approach, mirroring the methodology applied to the Lung and Ovarian datasets. The data preparation involved detailed visualization of feature importance using the chi-squared test and Recursive Feature Elimination (RFE). Chi-squared values and p-values helped identify influential features, enabling informed decision-making for modeling. Systematic feature selection, including the top 50% of important features, was performed to strike a balance between dimensionality reduction and information preservation.

As shown in Table 3, the model evaluation results revealed variations in the classifier performance. Gradient Boosting and Random Forest models demonstrated a high accuracy of 93.33%, indicating their effectiveness in making precise predictions. Decision Trees and k-NN also performed well, with accuracies of 80.00% and 93.33%, respectively. Naive Bayes achieved an accuracy on par with the top-performing models, indicating its suitability for the given task. While accuracy was a significant metric, considering other factors like precision, recall, and F1-score can provide a more comprehensive understanding of classifier performance. Overall, the research maintained consistency across datasets, ensuring a reliable and systematic approach to uncover valuable insights while maintaining analytical robustness and integrity.

Table 3 Classification Results for MLL Cancer dataset

Models	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
XGBoost	86.00	87.00	87.00	87.00
Naïve Bayes	93.00	**94.00**	**93.00**	**93.00**
KNN	**93.00**	**94.00**	**93.00**	**93.00**
Decision Trees	80.00	84.00	80.00	79.00
Gradient Boosting	**93.00**	**94.00**	**93.00**	**93.00**
Random Forest	**93.00**	**94.00**	**93.00**	**93.00**

4.4 Acute Myeloid Leukemia (AML)

In the analysis of the Acute Myeloid Leukemia (AML) dataset, a thorough understanding of the data revealed a dataset with 7,130 columns and 72 rows, emphasizing its high dimensionality and potential challenges in data preprocessing. Data preparation included label encoding of the "CLASS" column and careful inspection for missing values, which were not found, ensuring data quality and reliability. Feature selection involved the chi-squared test and Recursive Feature Elimination (RFE), with the top 50% of important features selected to strike a balance between dimensionality reduction and data integrity.

As shown in Table 4, the model evaluation results displayed variations in classifier performance, with Logistic Regression and SVM achieving the highest accuracy of 93.33%. Decision Trees and Random Forest exhibited good performance but could be susceptible to overfitting, with accuracies of 86%. k-NN and XGBoost achieved varying degrees of success, with accuracies of 66.66% and 80%, respectively, highlighting the sensitivity of classifier performance to dataset characteristics. Naive Bayes delivered a moderate accuracy of 73.33%. While accuracy was a significant metric, further evaluation metrics such as precision, recall, and F1-score are crucial for a comprehensive assessment of classifier performance, particularly in the context of imbalanced datasets or varying misclassification costs. This study emphasized a consistent and systematic approach to model training and evaluation, maintaining analytical robustness and integrity.

Table 4. Classification Results for (AML-3) dataset

Models	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-score
XGBoost	80.00	77.00	80.00	76.00
Naïve Bayes	73.33	63.00	73.00	68.00
KNN	66.66	56.00	67.00	60.00

Decision Tree	86.66	92.0	87.00	88.00
Random forest	86.66	92.0	87.00	87.00
Logistic Regression	**93.33**	**95.0**	**93.00**	**93.00**

5 Conclusion and Future Works

In this study, we focused on enhancing cancer classification in high-dimensional microarray data using a hybrid filter feature selection approach. The objective was to improve cancer classification accuracy, with a specific emphasis on datasets such as Lung Cancer, Ovarian Cancer, and AML Cancer. Several models were implemented and evaluated, with the best-performing model varying by dataset. For Lung Cancer, Logistic Regression Analysis with Chi-square filter and Recursive Feature Elimination wrapper methods achieved an accuracy of 97.56%. In the case of Ovarian Cancer, Gradient Boosting with Recursive Feature Elimination displayed potential with an accuracy of 92.85%, and for AML Cancer, both Random Forest and Logistic Regression models achieved competitive results with accuracies of 93.0% and 93.33%, respectively. This study successfully achieved its objectives, including the exploration of feature selection methods, model development, evaluation, and model comparison. However, there are limitations, such as the potential lack of model generalization, dataset age, small sample sizes, and the need for multi-objective optimization. Future work should explore additional datasets, develop tailored feature selection methods for specific cancer types, enhance interpretability and biological relevance of selected features, incorporate advanced machine learning techniques, and continuously update models with new data to improve real-world clinical applications. In summary, this research offers valuable insights into improving cancer classification using high-dimensional microarray data, with logistic regression and feature selection techniques showing promise as robust approach for accurate cancer prediction. Future research should address the identified limitations and explore innovative methods to further advance cancer classification in complex medical datasets.

References

1. Trevino, V., Falciani, F. and Barrera-Saldaña, H.A. (2007). DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. Molecular Medicine, 13(9-10), pp.527–541. doi:https://doi.org/10.2119/2006-00107.trevino.
2. Cosma, G., Brown, D., Archer, M., Khan, M. and Graham Pockley, A. (2017). A survey on computational intelligence approaches for predictive modeling in prostate cancer. Expert Systems with Applications, 70, pp.1–19.
3. Lai, C.-M., Yeh, W.-C. and Chang, C.-Y. (2016). Gene selection using information gain and improved simplified swarm optimization. Neurocomputing, 218, pp.331–338. doi:https://doi.org/10.1016/j.neucom.2016.08.089.
4. Singh, R.K. and Sivabalakrishnan, M. (2023). Feature Selection of Gene Expression Data for Cancer Classification: A Review. Procedia Computer Science, 50, pp.52–57. doi:https://doi.org/10.1016/j.procs.2015.04.060.

5. Veerabhadrappa, Mr. and Rangarajan, L. (2010). Bi-level dimensionality reduction methods using feature selection and feature extraction. International Journal of Computer Applications, 4(2), pp.33–38. doi:https://doi.org/10.5120/800-1137.
6. Bennet, J., Ganaprakasam, C. and Kumar, N. (2015). A Hybrid Approach for Gene Selection and Classification using Support Vector Machine. The International Arab Journal of Information Technology, 12(6A).
7. Jain, A. and Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), pp.153–158. doi:https://doi.org/10.1109/34.574797.
8. Li, B.-Q., Hu, L.-L., Chen, L., Feng, K.-Y., Cai, Y.-D. and Chou, K.-C. (2012). Pre-diction of Protein Domain with mRMR Feature Selection and Analysis. PLoS ONE, 7(6), p.e39308. doi:https://doi.org/10.1371/journal.pone.0039308.
9. Alshamlan, H.M., Badr, G.H. and Alohali, Y.A. (2015). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Compu-tational Biology and Chemistry, 56, pp.49–60. doi:https://doi.org/10.1016/j.compbiolchem.2015.03.001.
10. García-Díaz, P., Sánchez-Berriel, I., Martínez-Rojas, J.A. and Diez-Pascual, A.M. (2020). Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics*, 112(2), pp.1916–1925. doi:https://doi.org/10.1016/j.ygeno.2019.11.004.
11. Wu, J. and Hicks, C. (2021). Breast Cancer Type Classification Using Machine Learning. *Journal of Personalized Medicine*, 11(2), p.61. doi:https://doi.org/10.3390/jpm11020061.
12. Chen, J.W. and Dhahbi, J. (2021b). Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. Scientific Reports, 11(1). doi:https://doi.org/10.1038/s41598-021-92725-8.
13. Liu, S. and Yao, W. (2022). Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection. BMC Bioinformatics, 23(1). doi:https://doi.org/10.1186/s12859-022-04689-9.
14. Mahin, K.F., Robiuddin, Md., Islam, M., Ashraf, S., Yeasmin, F. and Shatabda, S. (2022). PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. Genomics, 114(2), p.110264. doi:https://doi.org/10.1016/j.ygeno.2022.01.001.
15. Gakii, C., Mireji, P.O. and Rimiru, R. (2022). Graph Based Feature Selection for Reduction of Dimensionality in Next-Generation RNA Sequencing Datasets. Algorithms, 15(1), p.21. doi:https://doi.org/10.3390/a15010021.
16. Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M. and Li, L. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics, 18(1). doi:https://doi.org/10.1186/s12864-017-3906-0.
17. Zhang, Y., Deng, Q., Liang, W. and Zou, X. (2018b). An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data. *BioMed Research International*, 2018, pp.1–11. doi:https://doi.org/10.1155/2018/7538204.

18. Al Abir, F., Shovan, S.M., Hasan, Md.A.M., Sayeed, A. and Shin, J. (2022). Biomarker identification by reversing the learning mechanism of an autoencoder and recursive feature elimination. Molecular Omics, 18(7), pp.652–661. doi:https://doi.org/10.1039/d1mo00467k.
19. Zexuan Zhu, Y. S. Ong and M. Dash, "Markov Blanket-Embedded Genetic Algorithm for Gene Selection", *Pattern Recognition*, Vol. 49, No. 11, 3236-3248, 2007.