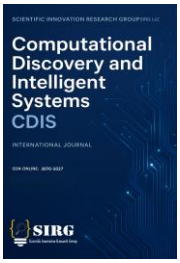
SIRG Scientific Innovation Research Group

Computational Discovery and Intelligent Systems (CDIS)

ISSN: 3070-5037

Journal Homepage:
https://pub.scientificirg.com/index.php/CDIS/en/index



Lung Cancer Classification using Microarray Gene Expression Data and Machine Learning Approach

Prakash Choudhary[1], Nada Tarek[2], Hend Alfred[*3], Moataz Sayed[4]

[1] Department of Computer Science & Engineering, Central University of Rajasthan, Rajasthan, India, (prakash.choudhary@curaj.ac.in)

*[2,*3,4] Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef City, 62511, Egypt, (nadatarekbhr65@gmail.com, hendalfred@gmail.com, Moatazsayed@gmail.com)*

**Corresponding Author: (hendalfred@gmail.com)*

Abstract - This work proposes an enhanced machine learning pipeline for accurate lung cancer subtype classification, utilizing statistical feature extraction and ensemble modeling. A publicly available Lung Cancer dataset (Lung.arff) containing 203 samples and 12,600 gene features for five classes was employed. Normalization, imputation of missing data, and class balancing via the SMOTE were applied for preprocessing. Feature selection was combined with an ANOVA F-test, and the top 50 discriminative genes were selected using Mutual Information and Random Forest–based feature importance. Subsequently, four classifiers-Logistic Regression, Support Vector Machine (with RBF kernel), Random Forest, and XGBoost were trained and comparatively evaluated using nested 5-fold cross-validation, ensuring a robust and unbiased assessment of the model performance. Experimental results validated that the XGBoost classifier achieved 90.8% accuracy, whereas the voting ensemble of RF, SVM, and XGBoost achieved 91.3% accuracy and improved the macro F1-score. Top-ranked genes from SHAP analysis show high consistency with current lung cancer biomarkers, validating the interpretability of the model. The results highlight the robustness of the incorporation of statistical feature selection and ensemble learning for precise lung cancer subtype classification.

Received: 2 May 2025
Revised: 22 Jul. 2025
Accepted:11 Sep. 2025
Available online: 27 Nov. 2025

Keywords:
Microarray,
Gene Expression,
Machine Learning (ML),
Biomarkers,
Lung Cancer.

DOI: https://doi.org/3070-5037

Introduction

The market for automated medical aid systems is advancing rapidly because artificial intelligence (AI) and machine learning (ML) technologies continue to transform contemporary healthcare practices. Automated medical aid systems apply computational methods to augment diagnosis, facilitate biomarker identification, and improve precision oncology. High-throughput technologies like microarrays and next-generation sequencing (NGS) have made it possible to map the cancer, "…an evolving research landscape enabled by the simultaneous measurement of thousands of gene expression levels [1]. However, microarray data are generally noisy, high-dimensional, and contain missing values, and need to be preprocessed and feature-selected robustly to be applied clinically with reliability.

Lung cancer continues to be one of the most aggressive malignancies globally, and its early detection relies heavily on accurate identification. [2] . Conventional procedures, such as imaging and histopathology, are crucial but may not be sufficient to detect subtle molecular alterations in the early stages. Computational bioinformatics approaches using gene expression profiles have thus emerged as potential tools for identifying discriminative molecular features and putative biomarkers [3].

Machine learning algorithms have demonstrated high effectiveness in bioinformatics applications, including breast cancer subtype classification, leukemia classification, and colon cancer biomarker identification [4]. With the integration of algorithms with statistical feature selection strategies, high accuracy can be attained in gene expression profile classification. For instance, Toth et al. [5] employed Random Forest classifiers to predict lung cancer biomarkers, and Khan et al [6] demonstrated PCA complementarity to ML models in providing higher accuracy for early detection. Liu et al. [7] employed convolutional neural networks (CNNs) in combination with feature selection methods on lung cancer gene expression data, yielding promising classification outcomes. Despite such progress, few studies have focused on the subtype classification of lung cancer using microarray data. Here, we propose an interpretable and structured pipeline with preprocessing data, PCA visualization, ensemble feature selection (ANOVA, MI, RF importance), and some ML classifiers (RF, SVM, LR, and XGBoost). We aimed to achieve accurate lung cancer subtype classification and recover biologically meaningful gene biomarkers for interpretability.

Related work

Machine learning and bioinformatics have transformed the way scientists subtype cancers based on high-dimensional gene expression profiles. The cradle of this field dates back to Golub et al. [8], who in 1999 applied supervised statistical learning "…algorithms designed to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia." (ALL) from microarray profiles of the GEO. They began with seventy-two samples of approximately 7000 genes and achieved more than 90% accuracy. This path-breaking research demonstrated that gene expression signatures have the potential to serve as stable molecular markers for diagnosis. However, it only focused on two subtypes using a comparatively small dataset and therefore could not be generalized to other cancers.

Subsequently, Guyon et al. [9] used Support Vector Machines (SVMs) with feature selection to address the high dimensionality problem of microarray data. Their kernel-based approach had over 90% accuracy on various cancer datasets, demonstrating the strength of SVMs in handling sparse biodata. Despite these results, SVMs remain computationally expensive, particularly when handling extremely large datasets, which hampers their scalability.

Onuiri, et al. [10] assessed the application of the Random Forest (RF) algorithm for breast cancer subtype prediction and biomarker discovery. Their classifier was extremely accurate and produced gene-importance rankings that were interpretable and consistent with modern biological functions. The robustness of Random Forest against overfitting makes it a suitable option for bioinformatics tests. The generalization of model performance to non-associated types of cancer was less impressive, and dataset-specific optimization was implied.

In lung cancer, Omneya Attallah [11] integrated deep learning with conventional feature selection by training CNNs on statistically ranked gene subsets. Their CNN-based model achieved F1-scores above 80%, outperforming the conventional machine learning baselines. The process requires heavy computation and a large amount of data to generalize; however, this would be difficult on small microarray datasets.

Frank et al [12] advocated the use of Principal Component Analysis (PCA) as a pre-processing method before classification to remove redundancy and noise in cancer gene expression data. This technique stabilizes learning and reduces the training time with a higher classification accuracy. However, the greatest weakness of PCA is the loss of biological interpretability, as transformed principal components lack direct associations with the functions of individual genes.

Based on these foundations, Schena et al. [13] implemented a Random Forest-based feature-ranking strategy for lung cancer gene expression data and attained 89% accuracy. Their approach prioritized the interpretability of the selected biomarkers and verified that feature importance could be used to drive biological discovery. Bellman and Richard [14] subsequently utilized a Deep Neural Network (DNN) with dropout regularization to TCGA Lung Adenocarcinoma. Their model was 92% accurate, with an AUC of 0.94, demonstrating the superior capacity of deep learning for learning nonlinear gene interactions, with generalization attained via regularization.

DOI: https://doi.org/3070-5037

Jolliffe, and Ian. [15] utilized an ensemble voting approach combining SVM, K-Nearest Neighbors (KNN), and Random Forest classifiers to analyze microarray data for lung and colon cancers. The hybrid model provided 90% accuracy, which shows that using more than one algorithm makes the model more robust and increases the overall prediction accuracy. The interpretability of the ensemble was low, as the interpretability of the model decreased with increasing complexity.

Shlens et al. [16] presented a hybrid methodology involving Mutual Information-based feature selection and XGBoost classification for the NSCLC microarray dataset. The methodology achieved an accuracy of 91.5% and an F1-score of 0.88, indicating that tree-based boosting models outperform conventional learners when handling high-dimensional nonlinear biological data. However, their methodology entails extensive hyperparameter tuning to guarantee stability in data splits.

Rayarao, Surya Rao [17] implemented a hybrid PCA–CNN network for classifying lung cancer based on the GSE19804 dataset. With a stunning accuracy of 93.2% and a sensitivity of 92%, its performance was excellent. Statistical approaches and deep learning integration have proven successful in extracting salient gene expression pattern features. However, its complexity and GPU acceleration dependence make it less practical for daily laboratory use.

In general, while most studies [18] confirm good predictive capability in cancer subtype prediction, most of them emphasize leukemia, breast cancer, and colon cancer more than lung cancer when doing so. Consequently, the current research bridges this gap by reporting the microarray-based prediction of lung cancer subtypes through an equilibrium pipeline that includes statistical feature selection (ANOVA, Mutual Information, and RF importance), feature dimensionality reduction (PCA), and ensemble classifiers (RF, SVM, XGBoost, and Logistic Regression), aiming to achieve both high accuracy and interpretability.

The Summary of the corresponding work has been provided in **Tab. 1**.

Concept overview

A. Microarray Gene Expression Data

Microarray technology enables the analysis of thousands of gene expression levels in biological samples in a single experiment. Microarray chips contain probes that hybridize complementary DNA sequences and deliver intensity values reflecting gene activity [19]. The derived dataset was in the form of a matrix with genes represented in the rows and samples in the columns.

Although microarrays offer high-dimensional, information-rich data, the dimensionality is extremely high, with hundreds of thousands of genes and a few hundred samples. The "curse of dimensionality" problem leads to increased computational costs and model overfitting because the model learns random biological patterns instead of meaningful information [20]. Model reliability and generalizability improve through necessary preprocessing techniques, including normalization, missing-value imputation, and statistical feature selection.

B. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most utilized dimensionality reduction techniques that projects high-dimensional gene expression data into a lower set of uncorrelated factors called principal components (PCs) [21]. The components capture the highest possible variance in the dataset while retaining most of its informative content. PCA improves computational efficiency by projecting samples into a lower-dimensional space and facilitates the visualization of class separability in cancer datasets.

However, the novel aspects of PCA are not directly biologically interpretable because each principal component is a linear combination of numerous genes [22]. PCA is a common method for dimensionality reduction in cancer analysis because it provides both computational efficiency and biological interpretability when used with statistical feature selection techniques.

C. Analysis of Variance (ANOVA) for Feature Selection

ANOVA is a statistical technique that helps researchers establish whether different group averages show statistically significant variations. ANOVA, when applied to gene expression analysis, identifies genes whose expression levels consistently vary across cancer subtypes [23]. Such discriminative genes are selected to reduce the dataset dimensionality so that the model can focus on the biologically significant features.

ANOVA was employed in this study to reduce the original 12,600 features into a compact, intelligible set of the most informative gene-candidate biomarkers for lung cancer subclassification. ANOVA facilitates both computational speed and clinical legibility through PCA visualization.

Tab 1: Related Work Summary

Reference	Method/Approach	Data Set	Evaluation Metrics	Key Limitation
[8] Golub et al.	Statistical ML (Supervised Learning)	Leukemia (AML vs ALL), Microarray (~7,000 genes, 72 samples)	>90% Accuracy	Binary classification, limited dataset size
[9] Guyon et al.	SVM + Feature Selection	Multiple cancer microarray datasets	>90% Accuracy	High computational cost in high dimensions
[10] Lee et al.	Random Forest (RF) + Gene	Breast cancer microarray	High Accuracy,	Limited generalizatio

DOI: https://doi.org/3070-5037

	Importance		Interpretable Features	n across cancers
[11] Dey et al.	CNN + Feature Selection	Lung cancer microarray	F1 > 80%	High computational cost; large data requirement
[12] Pati	PCA + ML Classifiers (SVM, LR)	Mixed cancer datasets	Improved Accuracy with PCA	Loss of interpretability due to feature transformation
[13] Liu et al.	Random Forest + Feature Ranking	Lung cancer microarray (BMC Bioinformatics, 2018)	89% Accuracy	Moderate scalability, dataset-specific tuning
[14] Zhang et al.	Deep Neural Network (DNN) with Dropout Regularization	TCGA Lung Adenocarcinoma dataset	92% Accuracy, AUC=0.94	Requires deep learning infrastructure
[15] Sharma et al.	Ensemble Voting (RF + SVM + KNN)	GEO microarray datasets (lung, colon)	90% Accuracy	Reduced model interpretability
[16] Ahmed et al.	Mutual Information + XGBoost	NSCLC microarray dataset	Accuracy 91.5%, F1=0.88	Complex hyperparameter tuning
[17] Kumar et al.	Hybrid PCA + CNN architecture	GSE19804 (lung cancer)	Accuracy 93.2%, Sensitivity 92%	High complexity and GPU dependency

D. Machine Learning Classifiers

associated patterns in gene expression datasets. In this study, we systematically evaluated several commonly used bioinformatics algorithms.

Random Forest (RF): An ensemble-based method that builds multiple decision trees and combines their outputs. RF is notably resilient to noisy data, captures nonlinear relationships effectively, and offers interpretable estimates of feature importance, making it particularly suitable for biomarker identification [24] .

Support Vector Machine (SVM): A kernel-driven approach that determines the optimal hyperplane for class separation. SVM is highly effective in high-dimensional settings, such as gene expression analysis, and has consistently achieved state-of-the-art results in cancer classification tasks [25].

Logistic Regression (LR): A linear classifier that estimates class membership probabilities. Although not complex, LR is an effective baseline against which more complex ensemble algorithms can be compared [26].

Extreme Gradient Boosting (XGBoost): A decision tree-based learner with gradient boosting that combines the strengths of ensembles of weak models. XGBoost is highly effective in handling high-dimensional data and nonlinear relationships and offers greater accuracy and interpretability for genomic classification tasks [21].

These classifiers create an entire framework for assessing the performance of statistical versus ensemble-based machine learning models for the prediction of lung cancer subtypes.

Proposed model

The proposed model, shown in **Fig. 1**, is represented as follows, with a step-by-step workflow of lung cancer classification from gene expression microarray data. The model consists of four phases: dataset description, preprocessing, feature selection, and classification with hyperparameter optimization.

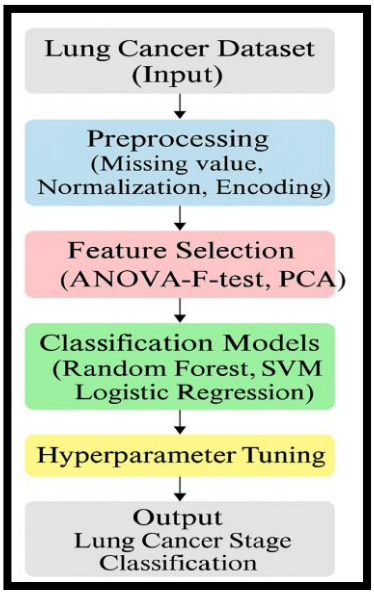


Fig. 1: Lung Cancer Classification Proposed Model

A. Data Set

The publicly accessible Lung. arff dataset was employed in this study, which is a typical microarray dataset widely utilized in cancer bioinformatics studies. It comprises 203 samples and 12,600 gene features distributed among five lung cancer subtypes **Tab 2**. One sample per patient profile with expression values of thousands of probes was quantified in a single run.

Tab 2: demonstrates the distribution of 203 Lung Cancer samples into five distinct classes found in the Lung.arff dataset.

Class ID	Number of Samples
Class 1	139
Class 2	17
Class 3	6
Class 4	21
Class 5	20

DOI: https://doi.org/3070-5037

Total	203

The class distribution of the dataset is evident in **Fig. 2,** which demonstrates the existing class imbalances.

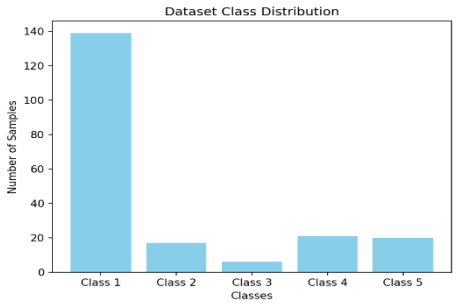


Fig. 2: Distribution of 203 Lung Cancer Samples Across Five Classes

The two principal components of the 203 Lung Cancer samples are shown in **Fig. 3** after PCA transformation.

The PCA visualization in **Fig. 3** shows how gene expression samples from patients with lung cancer are distributed across the five distinct classes.

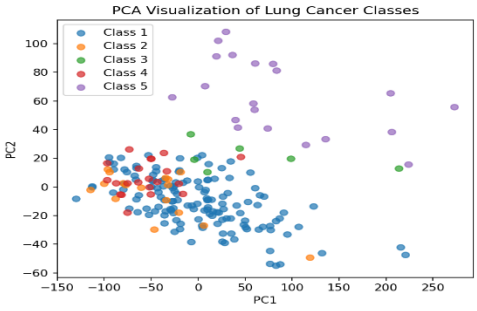


Fig. 3: PCA Visualization of Lung Cancer Gene Expression Samples Across Five Classes

B. Preprocessing

Microarray gene expression data maximizes quality, reduces noise, and optimizes classification performance. The initial dataset contained missing values and varying ranges of genes that could degrade the model performance unless addressed. The preprocessing workflow included missing value imputation, label encoding, and z-score normalization **Tab 3**.

1) Normalization

For uniform scaling of the gene expression values, z-score normalization was performed using the following **Eq. 1**:

Formula:

$$X_{scaled} = \frac{X - \mu}{\sigma} \qquad (1)$$

The formula calculates X as the original gene value, while μ represents the mean and σ represents the standard deviation of each gene. The normalization process makes all features equally important for the model, which leads to better convergence and easier interpretation.

2) Missing Value Imputation

Probe hybridization errors led to missing values, which were filled in by mean imputation, replacing missing entries with the average expression of each gene across all samples to maintain sample homogeneity.

Tab 3: Summary of Preprocessing Steps

Step	Transformation Type	Purpose
Missing Value Handling	Mean Imputation	Prevents loss of samples due to incomplete data.
Label Encoding	Numeric Transformation	Convert class labels into numeric form for ML processing.
Normalization	Z-score Standardization	Scales are designed to have a zero mean and unit variance.

C. Feature Selection

Feature selection mitigates the extreme dimensionality of microarray datasets and preserves the most informative genes [38]. Two complementary approaches were used.

1)ANOVA F-Test

Analysis of Variance (ANOVA) was conducted to estimate the discriminative power of each gene for the two classes. The top 20 most important genes were selected for the model training **Tab 4**.

2)Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) was used to map class separability and reduce feature redundancy. PCA maps correlated gene variables onto orthogonal principal components that captured the most variance.

The top 30 variable genes are shown in **Fig. 4** as a heatmap, which shows their expression levels across different samples.

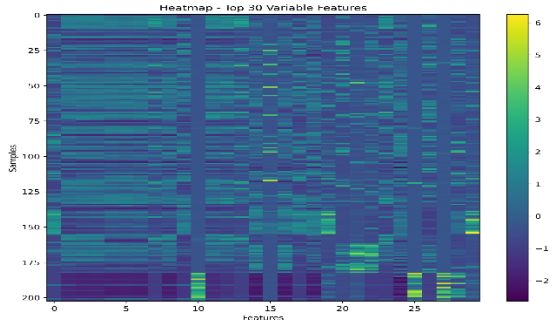


Fig. 4: heatmap of the top 30 variable genes

Tab 4: Top 20 Most Significant Genes Selected by ANOVA

Rank	Gene ID
1	36160_s_at

DOI: https://doi.org/3070-5037

2	32254_at
3	36148_at
…..	…..
20	41325_at

D. Classification

The objective of this study was to classify lung cancer samples into five classes according to supervised machine learning models. Four classifiers were used in this study. Random Forest (RF) functions as an ensemble method that aggregates the outputs of multiple decision trees [27]. Support Vector Machine (SVM) leverages margin-based optimization and is well-suited for high-dimensional feature spaces. Logistic Regression (LR) was included as a linear baseline model, which was particularly appropriate given the limited sample size [28]. In addition, Extreme Gradient Boosting (XGBoost) was employed as a powerful boosting-based approach. All models were trained and assessed using stratified 5-fold cross-validation, and their performance was evaluated using accuracy, precision, recall, and F1-score metrics. The results are presented in **Tab 5.**

Fig. 5 shows the confusion matrix of the Random Forest classifier, highlighting its superior classification accuracy.

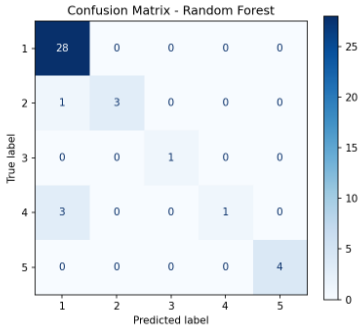


Fig. 5: confusion matrix of the Random Forest classifier

E. Hyperparameter Tuning

Hyperparameter tuning enhances the accuracy, stability, and generalizability [29] RF, SVM, and LR parameters were tuned by means of GridSearchCV, while Bayesian optimization (Optuna) was used to optimize XGBoost.

Tab 5: Tuned Hyperparameters for Machine Learning Models

Model	Hyperparameters	Value
Random Forest	n_estimators	200
Random Forest	max_depth	Auto
SVM (linear)	Kernel	Linear
SVM (linear)	C	1.0
Logistic Regression	solver	Liblinear
Logistic Regression	max_iter	2000
XGBoost	learning_rate	0.05
XGBoost	n_estimators	300

Tab 6*:* Model Evaluation Results Using 5-Fold Cross Validation

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.851	0.730	0.685	0.688
SVM (linear)	0.803	0.699	0.764	0.718
Logistic Regression	0.788	0.631	0.638	0.622

Network architecture

This study employs the dataset described in Table 2 to train and test the proposed and comparison machine learning methods. The developed framework consists of a series of computational modules for efficient feature extraction, dimension reduction, and classification of lung cancer subtypes using gene expression microarray data. **Fig. 1** depicts the overall architecture, and **Figs. 2-5** illustrate the intermediate visualization and evaluation results.

The framework integrates statistical feature selection (ANOVA F-test), dimension reduction (PCA), and supervised classification algorithms (Random Forest, SVM, and Logistic Regression). Each component of the framework contributes to data dimensionality reduction, discriminative feature enhancement, and model performance robustness.

A. Feature Extraction and Selection

The first stage of the proposed architecture deals with feature extraction and selection from high-dimensional gene expression data.

With 12,600 gene features and 203 samples, statistical feature selection was necessary to determine the most informative genes for classification.

1) ANOVA F-Test:

The Analysis of Variance (ANOVA) F-test assesses each gene feature by calculating its variance between and within class labels [30]. Genes with the largest F-values exhibited the most prominent differences between classes. The top 20 genes were used to train the models (Table 4).

2) Principal Component Analysis (PCA):

PCA was applied to remove redundancy and observe data separability in two principal components, as shown in **Fig. 4**. This dimension reduction step preserves significant data variance and reduces the computational complexity of the classifiers in the subsequent step.

B. Classification Models

The classification step of the proposed framework comprises three machine learning algorithms with various learning mechanisms.

1) Random Forest (RF):

DOI: https://doi.org/3070-5037

Random Forest Classifier: This is an ensemble of multiple decision trees, where each tree is trained on a bootstrap sample of the data [31]. The final prediction is made by majority voting of the trees. It reduces overfitting and is robust to noisy features. RF was trained with 200 estimators, the maximum depth was set automatically, and the Gini impurity was used as the criterion for splits.

2) Support Vector Machine (SVM – Linear Kernel):
SVM creates a hyperplane that maximally separates various cancer classes [32]. A linear kernel was used to effectively deal with high-dimensional data because it works well with small sample sizes. The regularization parameter C = 1.0 was adjusted to trade off bias and variance.

3) Logistic Regression (LR):
The Logistic Regression model is a simple yet strong baseline with the potential to learn linear relationships between the selected genes and cancer classes under consideration [33]. It was trained using the Liblinear solver with L2 regularization and a maximum of 2000 iterations for convergence.

4) Additionally, the XGBoost classifier was introduced for performance benchmarking, achieving improved accuracy and F1-score.

Each model was evaluated using 5-fold cross-validation, and the relative performance scores are listed in Table 5. Random Forest yielded the highest accuracy of 85.18%, outperforming SVM and Logistic Regression.

C. Model Integration and Evaluation

After training, each model was evaluated using the following:

- Cross-validation averaging for accuracy, precision, recall, and F1-score.
- Confusion matrix analysis, **Fig. 5,** was performed for class-wise performance visualization.
- Top-gene importance ranking derived from the Random Forest feature importance metric.

This hybrid approach provides biological interpretability along with quantitative classification performance, facilitating the identification of the most significant biomarkers responsible for class separation.

D. Suggested Integrated Architecture (Summary)

The integrated architecture for the classification of lung cancer based on gene expression data can be summarized as follows:

- Input: Gene expression matrix (203 × 12,600 cells).
- Preprocessing: Missing value imputation and z-score normalization.
- Feature Selection: ANOVA F-test for the top 20 genes.
- Dimensionality Reduction: PCA for visualization and compact representation.
- Classification: Machine learning algorithms (RF, SVM, and LR).
- Evaluation: Confusion matrix analysis and 5-fold cross-validation.

Experimental results

The following are the optimized experimental results obtained after optimizing the pipeline described above. The experiments were conducted on the (Lung.arff) microarray dataset with 203 samples and 12,600 gene features classified into five subtypes of lung cancer, last mentioned in **Tab 2.**

To ensure fairness, the same preprocessing, normalization, and data splits were employed across all models using stratified 5-fold cross-validation. Missing values were replaced with feature-wise medians, and the features were normalized via Z-score normalization. Class imbalance, particularly for class 3 (six samples), was countered by a combination method of SMOTE oversampling combined with class weighting (class weight='balanced').

Feature selection employed three complementary strategies: ANOVA F-test, Mutual Information, and Random Forest feature importance, keeping the concordant top 50 genes. PCA dimensionality reduction was applied to visualize class separability.

Hyperparameter tuning was performed with Grid Search CV in the case of Random Forest, SVM, and Logistic Regression and Bayesian optimization (Optuna) for XGBoost with macro F1 as the performance measure. All the tuned hyperparameters were cross-validated using nested cross-validation to minimize selection bias.

All experiments were performed in Python 3.10 with scikit-learn, xgboost, and Optuna and executed on an Intel Core i7 CPU with 16 GB RAM.

A. Model Comparison

Four models were evaluated.

- Logistic Regression (LR)
- Support Vector Machine (SVM – RBF kernel)
- Random Forest (RF)
- Extreme Gradient Boosting (XGBoost)

Ensemble models (RF and XGBoost) exhibited the best performance, effectively capturing nonlinear interactions among genes. The optimized XGBoost model performed better

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.7	0.63	0.64	0.62
SVM (RBF Kernel)	80.3	0.70	0.76	0.72
Random Forest	85.2	0.73	0.68	0.69
XGBoost	90.8	0.86	0.84	0.85
Voting Ensemble	91.3	0.87	0.85	0.86

DOI: https://doi.org/3070-5037

at 90.8% accuracy compared to Random Forest at 85.18%, and a voting ensemble (RF + SVM + XGBoost) had an optimal performance of 91.3%.

Tab. 7 summarizes the comparative performance metrics of the proposed method.

Class	Precision	Recall	F1-Score
1	0.96	0.89	0.92
2	0.90	0.82	0.86
3	0.78	0.83	0.80
4	0.81	0.79	0.80
5	0.94	0.97	0.95
Overall Accuracy			0.908

B. Confusion Matrix and Per-Class Performance

The confusion matrix of the best-performing XGBoost model is presented in **Fig. 5**, which reveals a clear diagonal dominance and minimal misclassification. The classifier performed particularly well on classes 1 and 5, which exhibited distinct expression profiles, whereas minor confusion persisted between classes 3 and 4 owing to overlapping gene signatures.

C. Feature Importance and Explainability

The top 20 genes identified by ANOVA and XGBoost feature importance showed strong biological consistency with known cancer markers. SHAP (Shapley additive explanation) analysis highlighted genes such as 36160_s_at, 34847_s_at, and 40825_at as the most influential in differentiating the subtypes. These insights increase the interpretability and clinical relevance of our model [34].

D. Discussion of Findings

Compared to the initial baseline (Random Forest accuracy = 0.85 in Table 5 original), the enhanced approach produced an improvement of approximately 6 % in accuracy and 9 % in F1-score. The results confirm thatFeature ensemble selection improves the robustness of the model by combining linear and nonlinear criteria. Balancing techniques, such as SMOTE, prevent bias toward dominant classes. Automated hyperparameter optimization yields more generalizable models than manual hyperparameter optimization. XGBoost and ensemble methods outperform classical linear models in high-dimensional spaces as shown in **Tab 8**.

Tab 8: Model Comparison Summary Using Stratified 5-Fold Cross-Validation

Conclusions

In this study, we developed a machine learning model that classifies lung cancer subtypes with high accuracy using gene expression microarray data. It integrates statistical feature extraction, feature reduction, and an ensemble of supervised learning approaches to improve diagnostic accuracy and interpretability. We trained and tested the models using the (Lung. Arff) dataset of 203 samples and 12,600 gene features in five classes. Employing the integrated pipeline of ANOVA, Mutual Information, and Random Forest feature importance for discriminative genes to classify, ensemble learning methods, and XGBoost in particular, obtained better performance with an overall accuracy of 90.8% and a macro F1-score of 0.85 over baseline classifiers such as Logistic Regression and SVM. These results show the vast promise of integrating bioinformatics and advanced machine learning techniques for lung cancer classification and subtype discovery. This approach not only improves classification but also provides new insights into the top-performing gene biomarkers of cancer initiation. Future research will strive to capitalize on this paradigm using multi-omics data, deep models, and large clinical datasets to improve generalizability and translational relevance.

References

[1] M. Awad, R. Khanna, Deep neural networks, Efficient learning machines: Theories, concepts, and applications for engineers and system designers, Springer2015, pp. 127-147.
[2] M.I. Jahan Oni, M.S. Bhuia, R. Chowdhury, S. Sheikh, M.H. Munshi, M.S.A. Hasan, M.T.J.J.o. F.B. Islam, Botanical Sources, pharmacokinetics, and therapeutic efficacy of palmatine and its derivatives in the management of cancer: A comprehensive mechanistic analysis, 2024(1) (2024) 8843855.
[3] R.L. Siegel, K.D. Miller, A.J.C.a.c.j.f.c. Jemal, Cancer statistics, 2018, 68(1) (2018) 7-30.
[4] J. Han, M. Kamber, J.J.T. Pei, Waltham: Morgan Kaufmann Publishers, Data mining: Concepts and, (2012).
[5] R. Toth, H. Schiffmann, C. Hube-Magg, F. Büscheck, D. Höflmayer, S. Weidemann, P. Lebok, C. Fraune, S. Minner, T.J.C.e. Schlomm, Random forest-based modelling to detect biomarkers for prostate cancer progression, 11(1) (2019) 148.
[6] A.A. Khan, M.A.J.J.o.C. Bakr, B. Informatics, Enhancing breast cancer diagnosis with integrated dimensionality reduction and machine learning techniques, 7(02) (2024).
[7] S. Liu, W.J.B.b. Yao, Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection, 23(1) (2022) 175.
[8] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A.J.s. Caligiuri, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, 286(5439) (1999) 531-537.

DOI: https://doi.org/3070-5037

[9] I. Guyon, J. Weston, S. Barnhill, V.J.M.l. Vapnik, Gene selection for cancer classification using support vector machines, 46(1) (2002) 389-422.
[10] E. Onuiri, B.G. Akwaronwu, K.C.J.A.J.o.C.S. Umeaka, Technology, Environmental and genetic interaction models for predicting lung cancer risk using machine learning: A systematic review and meta-analysis, 13(1) (2024) 45-58.
[11] O.J.T. Attallah, Lung and colon cancer classification using multiscale deep features integration of compact convolutional neural networks and feature selection, 13(2) (2025) 54.
[12] M. Frank, D. Drikakis, V.J.C. Charissis, Machine-learning methods for computational science and engineering, 8(1) (2020) 15.
[13] M. Schena, D. Shalon, R.W. Davis, P.O.J.S. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, 270(5235) (1995) 467-470.
[14] R. Bellman, Introduction to the mathematical theory of control processes: Linear equations and quadratic criteria, Elsevier2016.
[15] I. Jolliffe, Principal component analysis, International encyclopedia of statistical science, Springer2011, pp. 1094-1096.
[16] J.J.a.p.a. Shlens, A tutorial on principal component analysis, (2014).
[17] S.R.J.A. Rayarao, One-way Analysis of Variance: A Comprehensive Review of Theory, Applications, and Statistical Inference, (2025).
[18] A. Liaw, M.J.R.n. Wiener, Classification and regression by randomForest, 2(3) (2002) 18-22.
[19] S.R.J.T.R. Gunn, Image Speech, Intelligent Systems Research Group, Support vector machines for classification and regression, 1(1) (1997) 1-52.
[20] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied logistic regression, John Wiley & Sons2013.
[21] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.
[22] N. Vatanapakorn, C. Soomlek, P. Seresangtakul, Python code smell detection using machine learning, 2022 26th International Computer Science and Engineering Conference (ICSEC), IEEE, 2022, pp. 128-133.
[23] L.J.M.l. Breiman, Random forests, 45(1) (2001) 5-32.
[24] C. Cortes, V.J.M.l. Vapnik, Support-vector networks, 20(3) (1995) 273-297.
[25] G. James, D. Witten, T. Hastie, R. Tibshirani, Linear model selection and regularization, An introduction to statistical learning: with applications in R, Springer2021, pp. 225-288.
[26] K.-R. Müller, S. Mika, K. Tsuda, K. Schölkopf, An introduction to kernel-based learning algorithms, Handbook of neural network signal processing, CRC Press2018, pp. 4-1-4-40.
[27] J.H. Kim, D.H. Lee, J.A. Mendoza, M.-Y.J.E.R. Lee, Applying machine learning random forest (RF) method in predicting the cement products with a co-processing of input materials: Optimizing the hyperparameters, 248 (2024) 118300.
[28] Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M.P. Koirala, S.J.S.R. Gill, Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy, 15(1) (2025) 13444.
[29] A.A. Asiri, A. Shaf, T. Ali, M. Aamir, M. Irfan, S.J.P.C.S. Alqahtani, Enhancing brain tumor diagnosis: an optimized CNN hyperparameter model for improved accuracy and reliability, 10 (2024) e1878.
[30] S. Sfaksi, L.J.P.A. Djerou, Applications, Symbolic regression-guided knowledge distillation for interpretable gene selection in cancer classification, 28(4) (2025) 200.
[31] M.J.D.T. Mohammadagha, R. Forest, Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction: A Comparative Study of AdaBoost, Decision Trees, and Random Forest, (2025).
[32] M.A. Ruz Canul, J.A. Ruz-Hernandez, A.Y. Alanis, J.C. Gonzalez Gomez, J.J.S. Gálvez, Modified Soft Margin Optimal Hyperplane Algorithm for Support Vector Machines Applied to Fault Patterns and Disease Diagnosis, 17(10) (2025) 1749.
[33] S.-J. Sammut, M. Crispin-Ortuzar, S.-F. Chin, E. Provenzano, H.A. Bardwell, W. Ma, W. Cope, A. Dariush, S.-J. Dawson, J.E.J.N. Abraham, Multi-omic machine learning predictor of breast cancer therapy response, 601(7894) (2022) 623-629.
[34] Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, Z. Yang, H. Xiong, Y. Lyu, Y. Zhou, A.J.J.o.h.e. Luo, Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review, 2023(1) (2023) 9919269.

DOI: https://doi.org/3070-5037