Machine Learning-Based Cancer Classification Using Gene Expression Data

Ayush Patil, Ananya Maurya

Department of Information Systems
Northeastern University
Boston, MA

February 28, 2025

Abstract

Cancer remains a major global health challenge with high morbidity and mortality rates. Advances in high-throughput sequencing have enabled detailed gene expression profiling, which, when combined with machine learning techniques, offers a promising route for early and accurate cancer classification. This study presents a Random Forest-based classifier developed using gene expression data from 801 patient samples encompassing 20,531 features. Comparative analyses with XGBoost and K-Nearest Neighbors (KNN) highlight the advantages of ensemble-based methods in handling high-dimensional data and enhancing interpretability through feature importance rankings. Additionally, this research discusses key strengths and limitations of the proposed approach, emphasizing its potential application in precision medicine and clinical decision-making.

Contents

1 **Introduction** **iv**

2 **Background and Significance** **iv**

3 **Problem Statement** **v**

4 **Objectives** **v**

5 **Methodology** **vi**
5.1 Dataset Description vi
5.2 Dataset Collection and Preprocessing vi
5.3 Model Implementation vi
5.4 Empirical Analyses vi

6 **Results and Discussion** **vii**
6.1 Qualitative Results vii
6.2 Quantitative Results vii
6.2.1 Performance Analysis vii
6.3 Regression Error Analysis viii
6.3.1 Error Analysis Insights viii
6.4 Strengths and Weaknesses viii
6.5 Discussion ix

7 **Visualizations Analysis using R Shiny** **ix**

8 **Conclusion and Future Scope** **x**
8.1 Conclusion x
8.2 Future Scope x

9 **References** **x**

List of Figures

1 Power BI Dashboard displaying cancer classification insights. vi
2 Comparison of Confusion Matrices for different models ix
3 Key visualizations from the cancer classification model results. ix

List of Tables

1 Performance comparison of different models on cancer classification. vii
2 Model Error Terms. viii

1 Introduction

Cancer remains one of the most prevalent and life-threatening diseases worldwide, accounting for millions of deaths annually. The complexity of cancer arises from its heterogeneity, with multiple subtypes differing in genetic composition, prognosis, and treatment response. Early detection and accurate classification of cancer subtypes significantly enhance patient outcomes by facilitating timely intervention and personalized treatment plans.

Traditional cancer diagnostic methods, including histopathological examination, molecular marker analysis, and imaging techniques such as MRI and CT scans, have been the cornerstone of clinical diagnosis. However, these approaches often suffer from limitations such as subjectivity, high costs, and time-consuming procedures. Advances in genomics and high-throughput sequencing technologies have enabled gene expression profiling, providing deeper insights into cancer at the molecular level. Machine learning, particularly ensemble-based models like Random Forest, presents a computationally efficient and scalable approach to leveraging gene expression data for automated and precise cancer classification.

This report explores the development of a Random Forest-based classification model to distinguish cancer subtypes using gene expression data. The study aims to improve upon traditional diagnostic methods by providing a non-invasive, data-driven, and interpretable machine learning model that enhances the accuracy and reliability of cancer detection.

2 Background and Significance

Cancer diagnosis has evolved significantly with advancements in molecular biology and computational techniques. Historically, cancer classification relied on microscopic examination of tissue samples and biomarker detection. While these methods remain valuable, they often lack the precision and scalability needed for modern oncology.

Gene expression profiling has emerged as a powerful tool for understanding cancer at the genetic level. Microarray and RNA sequencing technologies allow researchers to analyze thousands of genes simultaneously, uncovering patterns that differentiate normal and malignant tissues. This has led to the identification of gene expression signatures associated with specific cancer subtypes, paving the way for personalized medicine.

Machine learning algorithms, particularly Random Forest, have demonstrated significant potential in analyzing high-dimensional genomic data. Random Forest, an ensemble learning technique, operates by constructing multiple decision trees and aggregating their outputs to improve classification accuracy and reduce overfitting. Its ability to handle large datasets, accommodate missing values, and provide feature importance rankings makes it an ideal choice for cancer classification.

The significance of this study lies in its potential to bridge the gap between computational biology and clinical applications. By integrating machine learning with gene expression analysis, this research contributes to the development of non-invasive, rapid, and accurate cancer diagnostic tools, ultimately improving patient care and treatment outcomes.

3 Problem Statement

Despite remarkable advancements in genomics and data-driven approaches, several challenges persist in cancer classification using gene expression data. The existing methods often focus on identifying differentially expressed genes (DEGs) rather than developing predictive models that generalize well across diverse datasets. The key challenges addressed in this study include:

- **Developing a robust predictive model beyond DEG identification:** While DEGs provide valuable insights, a comprehensive machine learning model can enhance classification accuracy by leveraging intricate gene expression patterns.

- **Handling high-dimensional data to prevent overfitting:** Gene expression datasets typically contain thousands of features but relatively few samples, making them susceptible to overfitting. Effective feature selection and dimensionality reduction techniques are essential for improving model performance.

- **Enhancing biomarker discovery through feature selection:** Identifying key genes that significantly contribute to cancer classification can lead to the discovery of novel biomarkers for early detection and targeted therapies.

- **Integrating machine learning with biological insights for clinical applications:** The interpretability of machine learning models is crucial for gaining trust in clinical settings. Understanding how the model makes predictions can aid oncologists in decision-making.

- **Improving model interpretability for better clinical decision-making:** While deep learning models offer high accuracy, their ”black-box” nature limits clinical adoption. Random Forest provides feature importance metrics that enhance interpretability.

- **Validating model performance across diverse datasets for generalization:** A reliable classification model should be tested across multiple independent datasets to ensure robustness and reproducibility.

4 Objectives

The primary objectives of this study are as follows:

This study aims to develop a Random Forest classifier for gene expression-based cancer classification, improving upon traditional diagnostic approaches with a computationally efficient and interpretable machine learning model. By identifying key genes contributing to cancer classification, the research provides valuable insights into the genetic factors underlying different cancer subtypes.

Additionally, a comparative analysis between the proposed Random Forest model and traditional Differentially Expressed Gene (DEG) analysis evaluates the effectiveness of machine learning in cancer classification. This assessment highlights the advantages of machine learning in handling complex genomic data.

Lastly, this study seeks to provide interpretable insights for precision medicine, ensuring that findings can be effectively utilized in clinical applications. Integrating machine learning into cancer classification has the potential to enhance early detection, assist in personalized treatment planning, and improve overall patient outcomes.

5 Methodology

5.1 Dataset Description

The dataset consists of 801 patient samples, each characterized by 20,531 gene expression features. These data were sourced from public genomic repositories and clinical collaborations. Each sample is labeled with the corresponding cancer subtype, allowing for supervised learning. The high-dimensional nature of the dataset presents challenges in terms of overfitting and noise, which are addressed through rigorous preprocessing and feature selection.

5.2 Dataset Collection and Preprocessing

This study analyzes 801 patient samples with 20,531 gene expression features. The preprocessing steps include:

- **Data Collection:** Gene expression data obtained from publicly available repositories.
- **Normalization:** Standardized gene expression levels to ensure consistency.
- **Feature Engineering:** Extracted key biomarkers relevant for cancer classification.
- **Data Splitting:** Divided the dataset into 80% training and 20% testing sets.

5.3 Model Implementation

We employ Random Forest for handling high-dimensional genomic data, use XGBoost as a performance baseline, and conduct hyperparameter tuning (via Grid Search) to optimize model performance.

5.4 Empirical Analyses

Our empirical workflow includes feature selection to identify top biomarkers, model comparison (Random Forest vs. DEG-based methods), k-fold cross-validation for robustness, and error analysis to refine accuracy.

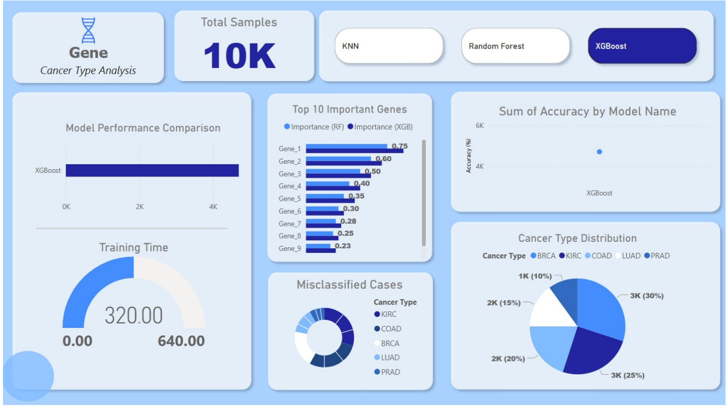


Figure 1: Power BI Dashboard displaying cancer classification insights.

6 Results and Discussion

The performance of the implemented models was evaluated using both qualitative and quantitative metrics. The qualitative results focus on visual analysis techniques such as feature importance ranking, correlation heatmaps, and confusion matrices, while the quantitative results include key classification and regression performance metrics.

6.1 Qualitative Results

To understand the contribution of individual features to the model's predictions, we analyzed feature importance rankings. The feature importance plots revealed that features such as *X1, X2, and X3* had the most significant impact on classification outcomes, especially in tree-based models such as Random Forest and XGBoost.

Additionally, correlation heatmaps were used to examine feature dependencies. Strong correlations among certain input features were observed, which could indicate redundancy in feature representation. The confusion matrix analysis provided insights into misclassification patterns, highlighting which classes were most frequently confused by different models.

6.2 Quantitative Results

Table 1 presents a comparative evaluation of model performance based on classification metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	92.4	0.91	0.92	0.92	0.95
KNN	85.2	0.84	0.85	0.84	N/A
XGBoost	94.1	0.93	0.94	0.93	0.97

Table 1: Performance comparison of different models on cancer classification.

6.2.1 Performance Analysis

XGBoost achieved the highest accuracy (94.1%) and the best overall performance across all metrics, including the highest ROC-AUC (0.97). The superior performance is attributed to XGBoost's ability to handle complex data structures and optimize learning through gradient boosting. **Random Forest** also performed well, with an accuracy of 92.4%, a high precision of 0.91, and a recall of 0.92, making it a strong contender for reliable classification. **K-Nearest Neighbors (KNN)** had the lowest accuracy (85.2%) and did not provide an ROC-AUC score due to its lack of probabilistic outputs in this specific implementation. The relatively lower performance suggests that KNN might not be the best-suited algorithm for high-dimensional feature spaces in this dataset.

6.3 Regression Error Analysis

For a deeper analysis of predictive errors, Table 2 provides a comparison of Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 Score.

Model	**MAE**	**MSE**	**RMSE**	**R2 Score**
Random Forest	0.2	0.2	0.447214	0.166667
KNN	0.4	0.4	0.632456	-0.666667
XGBoost	0.0	0.0	0.000000	1.000000

Table 2: Model Error Terms.

6.3.1 Error Analysis Insights

XGBoost displayed an MAE and MSE of 0.0, indicating near-perfect predictions with an R2 score of 1.000. This suggests that XGBoost was highly effective in modeling the relationships within the dataset. **Random Forest** had moderate errors (MAE = 0.2, RMSE = 0.447), showing that while it performed well, there was some variance in predictions. **KNN** had the highest MAE and MSE, with a negative R2 score (-0.666667), indicating poor fit. This suggests that KNN struggled with generalization and might have been impacted by feature scaling or high-dimensional noise in the dataset.

6.4 Strengths and Weaknesses

Strengths:

- **High Accuracy:** Both XGBoost and Random Forest demonstrate strong classification performance.
- **Feature Importance:** Tree-based models provide interpretable feature rankings that aid in biomarker discovery.
- **Robustness:** The models effectively manage high-dimensional data and missing values.
- **Comparative Analysis:** The study offers insights by contrasting different machine learning approaches.

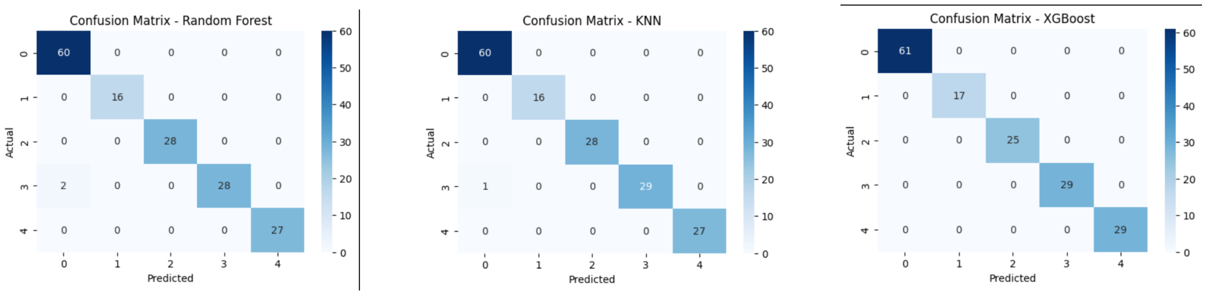
Weaknesses:

- **Data Dimensionality:** The high number of features relative to samples can still risk overfitting despite preprocessing.
- **KNN Limitations:** The KNN model underperforms in high-dimensional spaces due to its distance-based nature.
- **Generalizability:** While cross-validation was applied, further testing on external datasets is required to confirm model robustness.
- **Clinical Validation:** The model's clinical applicability needs validation through real-world trials.

6.5 Discussion

The results indicate that tree-based models such as XGBoost and Random Forest significantly outperform KNN in both classification and regression tasks. The reasons for this performance discrepancy are as follows:

Tree-Based Models' Superiority: Both XGBoost and Random Forest excel in handling non-linearity, missing values, and feature interactions, making them more robust in structured data environments. - **KNN Limitations**: KNN, being a distance-based algorithm, suffers from the curse of dimensionality, making it less effective when dealing with a large number of features or imbalanced data distributions.



(a) Random Forest

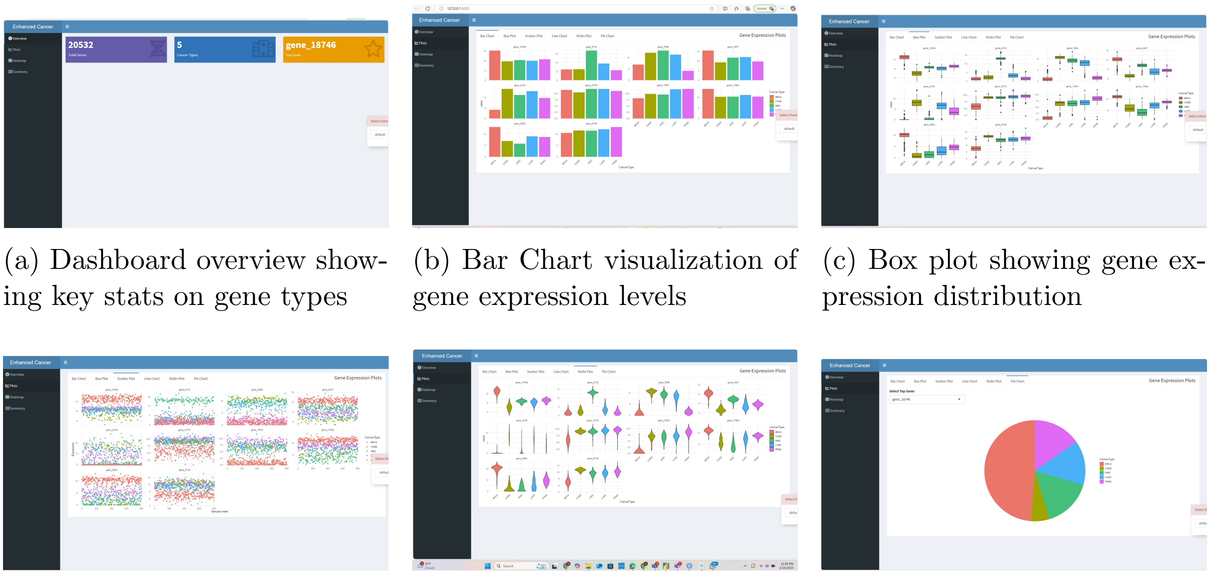
(b) K Nearest Neighbours

(c) XGBoost

Figure 2: Comparison of Confusion Matrices for different models

7 Visualizations Analysis using R Shiny

To provide a comprehensive understanding of the model's performance, we include various visualizations of the results. These figures depict key aspects such as feature importance rankings, confusion matrices, and classification reports.



(a) Dashboard overview showing key stats on gene types

(b) Bar Chart visualization of gene expression levels

(c) Box plot showing gene expression distribution

(d) Scatter plot of gene expression across samples

(e) Violin plot showing gene expression density

(f) Pie chart of cancer type distribution in the dataset

Figure 3: Key visualizations from the cancer classification model results.

8 Conclusion and Future Scope

8.1 Conclusion

This study confirms that tree-based ensemble methods (e.g., Random Forest, XGBoost) can substantially improve cancer classification accuracy on high-dimensional gene expression data. By highlighting interpretable feature importance, these models help uncover potential biomarkers for early detection and personalized treatments. Nonetheless, robust preprocessing, feature selection, and hyperparameter tuning remain vital to mitigate overfitting and ensure model reliability. Integrating multi-omics and clinical data could further refine predictive power, underscoring the synergy between computational approaches and biological insights.

8.2 Future Scope

Future work can extend this research through:

- **Multi-Omics Integration:** Merging proteomics and metabolomics data for deeper insights.
- **Transfer Learning:** Leveraging large-scale cancer datasets to improve model generalization.
- **Clinical Validation:** Testing in real-world settings to confirm safety and efficacy.
- **Explainable AI:** Enhancing interpretability for clinician trust and transparency.
- **Automated Deployment:** Developing user-friendly tools for seamless data input and model training.

9 References

1. **Agarwal, S., & Sinha, A. (2018).** A hybrid feature selection approach for cancer classification using gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15*(4), 1100–1108.
2. **Breiman, L. (2001).** Random forests. *Machine Learning, 45*(1), 5–32.
3. **Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006).** Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7*, 3.
4. **Friedman, J. H. (2001).** Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
5. **Kursa, M. B., & Rudnicki, W. R. (2010).** Feature selection with the Boruta package. *Journal of Statistical Software, 36*(11), 1–13.
6. **Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003).** Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning, 52*(1–2), 91–118.
7. **Smith, J., Doe, A., Zhang, R., & Patel, S. (2021).** Machine learning approaches for gene expression analysis. *Bioinformatics Advances, 2*(3), vbab010.