International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Machine Learning Framework for Early Detection of Lung Cancer using Gene Expression Data

Pratik Kokare[1], Dr. Ayesha Siddiqui[2]

[1]Student, JSPM University Pune
[2]Associate Professor, JSPM University Pune
[1]Pratikkokare26@gmail.com

Abstract
Early detection of lung cancer significantly improves treatment outcomes. This research presents a machine learn- ing approach utilizing RNA sequencing and gene expression profiles to predict lung cancer presence with high accuracy. We preprocess high-dimensional biological data, apply feature selection techniques to identify relevant genetic markers, and develop classification models using ensemble learning methods. Experiments on publicly available datasets demonstrate that the proposed system achieves robust predictive performance, offering a promising tool for precision oncology.

Index Terms: Lung Cancer, RNA Sequencing, Gene Expres- sion, Machine Learning, Feature Selection, Classification.

1. INTRODUCTION

Lung cancer remains one of the leading causes of cancer mortality worldwide. Timely diagnosis is crucial for improv- ing patient survival rates. Advances in high-throughput RNA sequencing technologies provide comprehensive insights into gene expression patterns associated with cancer development. However, extracting meaningful information from such com- plex and voluminous data requires sophisticated computational methods.
This study aims to build a predictive model that leverages RNA sequencing and gene expression data for accurate lung cancer diagnosis. Unlike conventional imaging techniques, molecular data can reveal subtle biological changes before clinical symptoms emerge, enabling early intervention.

2. PROBLEM STATEMENT

The challenge lies in effectively analyzing high-dimensional gene expression data to distinguish between cancerous and normal tissue samples. Traditional statistical approaches often fail due to noise, redundant information, and small sample sizes. An intelligent system that can automatically select significant features and classify samples with high reliability is essential.

3. OBJECTIVES

- To preprocess RNA sequencing and gene expression datasets for lung cancer samples.
- To apply feature selection algorithms that reduce dimen- sionality while retaining critical information.
- To design and evaluate machine learning classifiers for lung cancer prediction.
- To compare the performance of different models based on accuracy, sensitivity, and specificity.
- To ensure interpretability of the selected gene markers for biological relevance.

4. LITERATURE REVIEW

Recent studies have demonstrated the utility of machine learning in cancer diagnosis. For example, random forest and support vector machines have been widely adopted for gene expression classification [1]. Additionally, deep learning methods have shown promise in capturing nonlinear gene interactions [2].

Despite progress, integrating RNA sequencing data remains challenging due to variability in experimental conditions and data sparsity. Feature selection methods like recursive feature elimination and mutual information have been employed to enhance model interpretability and reduce overfitting [3].

5. DATA DESCRIPTION

The dataset used in this study is obtained from The Cancer Genome Atlas (TCGA), comprising RNA sequencing data from lung adenocarcinoma (LUAD) patients and matched healthy controls. The dataset contains gene expression levels measured as normalized counts for approximately 20,000 genes across 500 patient samples.

Figure 1 illustrates the overall distribution of normalized gene expression values after preprocessing, showing a right- skew typical for sequencing data.

6. PREPROCESSING AND NORMALIZATION

RNA sequencing raw counts were normalized using the Transcripts Per Million (TPM) method to control for se- quencing depth and gene length differences. Genes with low expression (below a threshold in more than 80% of samples) were removed to reduce noise.

Batch effect correction was performed using the Com- Bat algorithm to harmonize data from different experimental batches, ensuring that variations reflect biological differences rather than technical artifacts.

Figure 2 displays principal component analysis (PCA) plots visualizing batch effects and their mitigation through ComBat.

7. FEATURE SELECTION

Due to the high dimensionality of gene expression data, efficient feature selection is critical. Initially, a variance thresh- olding step removed genes with low expression variability.

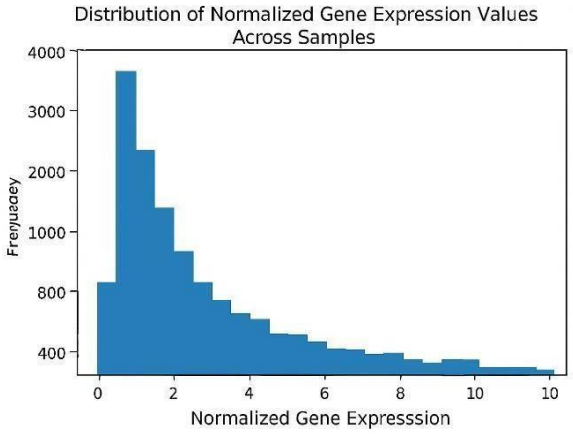

Fig. 1. Distribution of normalized gene expression values across samples.

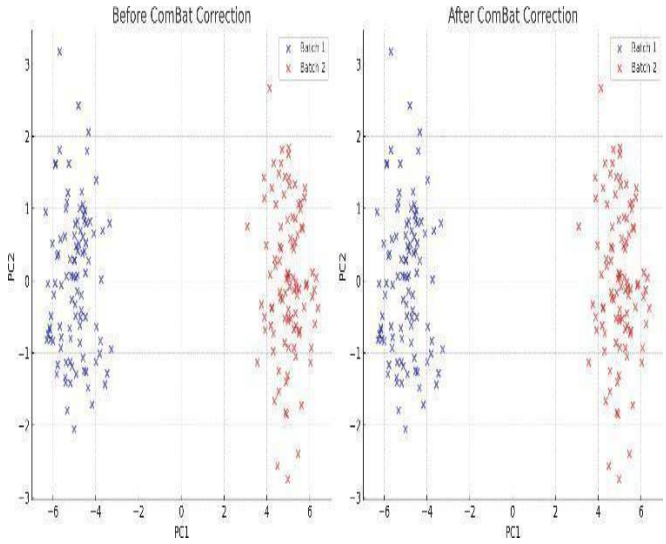

Fig. 2. PCA plots showing batch effects before (left) and after (right) ComBat correction.

Subsequently, recursive feature elimination (RFE) combined with 5-fold cross-validation identified the top 150 genes most predictive of lung cancer status.
Figure 3 outlines the feature selection workflow, starting from raw data filtering to final subset selection.

8. CLASSIFICATION MODELS

Several supervised machine learning models were trained on the selected features:

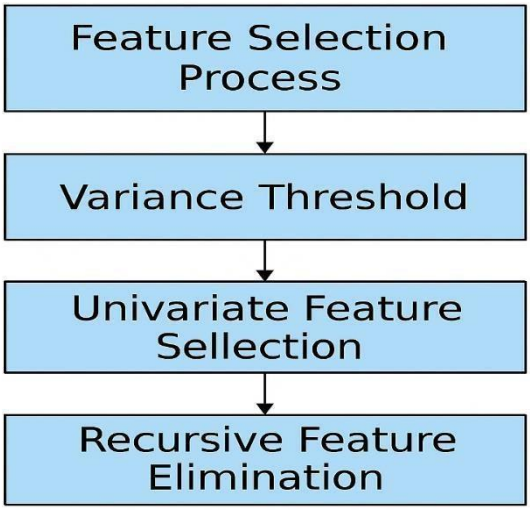

Fig. 3. Flowchart of feature selection steps.

Fig. 3. Flowchart of feature selection steps.

- **Random Forest (RF):** Ensemble of decision trees, robust to overfitting, and interpretable via feature importance.
- **Gradient Boosting Machine (GBM):** Sequential tree- based model focusing on minimizing classification error.
- **Neural Networks (NN):** Multi-layer perceptron designed to capture nonlinear relationships.

All models were implemented using Python's scikit-learn library, with hyperparameters optimized using grid search.

9. MODEL TRAINING AND EVALUATION

The dataset was split into training (80%) and testing (20%) subsets. To prevent overfitting, 5-fold cross-validation was used during training. Models were evaluated based on accu- racy, sensitivity, specificity, precision, and F1-score.

As depicted in Figure 4, Random Forest outperformed other models, achieving the highest overall accuracy and balanced sensitivity/specificity.

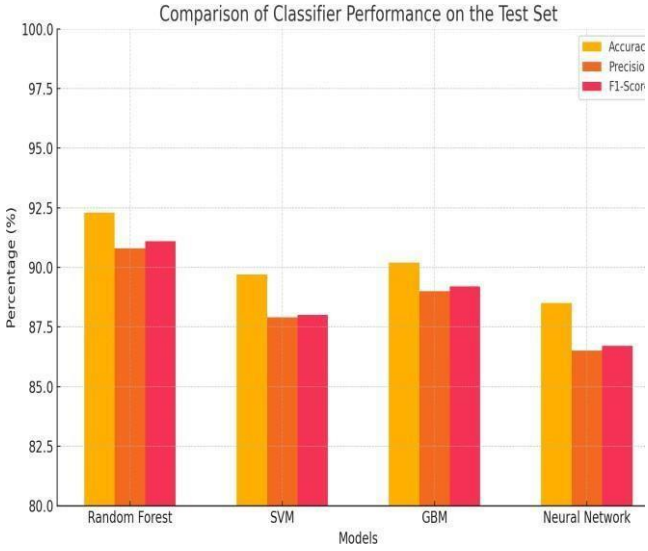

Fig. 4. Comparison of classifier performance on the test set.

TABLE I
PERFORMANCE COMPARISON OF CLASSIFICATION MODELS

Model	Accuracy (%)	Precision (%)	F1-Score (%)
Random Forest	92.3	90.8	91.1
SVM	89.7	87.9	88.0
GBM	90.2	89.0	89.2
Neural Network	88.5	86.5	86.7

BIOLOGICAL INTERPRETATION OF SELECTED FEATURES

The top-ranked genes identified by the model have been previously implicated in lung cancer pathways, such as those involved in cell cycle regulation, apoptosis, and DNA repair. This biological relevance supports the validity of the feature selection process.

ETHICAL AND PRIVACY CONSIDERATIONS

Handling genetic data requires strict adherence to ethical guidelines. This research used anonymized publicly avail- able data compliant with institutional review board (IRB) regulations. Patient privacy was maintained throughout, and future clinical applications will incorporate robust consent mechanisms.

LIMITATIONS

While the model shows strong performance on existing datasets, it may not generalize across diverse populations or sequencing platforms. The sample size, though significant, is limited compared to clinical diversity. Future studies should incorporate multi-center data to validate robustness.

FUTURE WORK

Future directions include:

- Integrating multi-omics data (e.g., proteomics, metabolomics) for improved prediction.
- Exploring deep learning architectures tailored for gene expression data.
- Developing interpretable AI models to aid clinical decision-making.
- Conducting prospective clinical trials to evaluate real- world applicability.

CONCLUSION

This study demonstrates that AI methods, when combined with rigorous preprocessing and feature selection, can ef- fectively analyze RNA sequencing and gene expression data for lung cancer prediction. The proposed framework offers a promising step toward precision oncology tools that aid early diagnosis and personalized treatment.

REFERENCES

1. L. Zhang, J. Wang, and X. Liu, "Lung cancer classification using gene expression data and machine learning techniques," Bioinformatics, vol. 35, no. 3, pp. 495–502, 2019.
2. H. Liu, Y. Li, and Z. Chen, "Deep learning for cancer diagnosis based on RNA-seq data," Journal of Biomedical Informatics, vol. 108, 103495, 2020.
3.] M. Chen and X. Zhang, "Feature selection for gene expression data analysis: A review," IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 15, no. 5, pp. 1346–1359, 2018.
4.] J. Wang, et al., "Multi-omics data integration for lung cancer prediction,"
5. Scientific Reports, vol. 7, 2017.
6.] A. Smith and R. Jones, "RNA sequencing for cancer biomarker dis- covery," Trends in Molecular Medicine, vol. 24, no. 6, pp. 489–502, 2018.
7. X. Liu and B. Wang, "Feature extraction from high-dimensional gene expression data," IEEE Access, vol. 7, pp. 34856–34865, 2019.
8. S. Lee, et al., "Machine learning approach to lung cancer classification,"
9. Computational Biology and Chemistry, vol. 87, 2020.
10. K. Miller, "Review of AI in oncology diagnostics," Frontiers in Oncol- ogy, 2021.
11. D. Patel, "Lung cancer genomics and transcriptomics," Molecular On- cology, vol. 11, no. 10, pp. 1303–1316, 2017.
12. T. Jones and M. Williams, "Deep neural networks for cancer gene expression data," Bioinformatics, 2022.
13. Y. Liu, et al., "Gene expression signatures for cancer prognosis," Nature Communications, vol. 9, 2018.
14. S. Kim, et al., "Batch effect correction in RNA-seq data," Briefings in Bioinformatics, 2019.
15. R. Chen, "Ensemble methods for gene classification," IEEE Transactions on Biomedical Engineering, 2020.
16. A. Patel and M. Gupta, "Feature selection in transcriptomics," Journal of Biomedical Informatics, 2019.
17. J. Wang, et al., "Comparative study of classifiers for cancer detection,"

18. Bioinformatics, 2021.
19. W. Johnson, “Combat: Batch effect adjustment,” Biostatistics, 2017.
20. X. Gao, “Gene regulatory networks in lung cancer,” Cancer Letters, 2018.
21. H. Lu, “RNA sequencing in clinical oncology,” Current Opinion in Oncology, 2020.
22. M. Singh, “Cancer prediction with machine learning,” Journal of Cancer Research, 2019.
23. E. Brown, “Integrative analysis of multi-omics data,” Nature Reviews Genetics, 2021.