FEATURE ENGINEERING FOR LUNG CANCER CLASSIFICATION USING NEXT GENERATION SEQUENCING DATA

Syed Naseer Ahmad Shah[1]*, Rafat Parveen[2]

[1,2]Department of Computer Science, Jamia Millia Islamia, New Delhi-110025, India syeddnaseer@gmail.com, rparveeen@jmi.ac.in*

Corresponding author:

Abstract

Next-generation sequencing (NGS) has profoundly transformed the field of genomics with its ability to detect molecular findings on a large scale, particularly for the somatic genome. Research on complex diseases such as lung cancer has shifted significantly as NGS technology provides an efficient method to unravel the genetic fingerprint of this extensively studied disorder. This advancement has opened new pathways for understanding the molecular underpinnings of lung cancer, facilitating more targeted approaches in diagnosis, treatment, and research. While NGS data are high dimensional and complex, they pose significant challenges to data analysis and classification tasks. In this paper, we investigated feature engineering to improve the classification accuracy of lung cancer using NGS data. The goal of these methods of dimensionality reduction, feature selection, and transformation techniques is to improve machine learning's predictive power. In this work, the dimensionality reduction method, Principal Component Analysis (PCA), is used to optimise feature selection. Advanced transformation techniques like normalisation and scaling are applied to optimise the data for better model performance. The efficacy of these techniques is evaluated through a comprehensive comparison of various machine learning classifiers, with a focus on Support Vector Machine (SVM). The results demonstrate that efficient feature engineering, particularly PCA, enhances the classification accuracy and robustness of lung cancer prediction models, providing valuable insights for the development of precision medicine approaches in oncology.

Keywords: *Next-generation sequencing, lung cancer, feature engineering, machine learning, dimensionality reduction, classification, Support Vector Machine.*

Introduction

Lung cancer remains a leading cause of cancer-related deaths worldwide, posing a significant burden on healthcare systems and societies [1]. Its high mortality rate underscores the critical need for early and accurate detection methods. Advances in genomic technologies, particularly next-generation sequencing (NGS), have revolutionised cancer research, offering unprecedented insights into the genetic alterations associated with various cancers, including lung cancer. By analysing gene expression data, researchers can identify key molecular biomarkers that enable accurate disease classification, prognosis, and therapeutic targeting [2].

However, the high dimensionality and complexity of NGS data present formidable challenges in data analysis. Machine learning (ML) models, often employed for classification tasks, struggle with the curse of dimensionality, which can lead to overfitting and reduced generalizability. Feature engineering, which involves selecting, transforming, and optimising features, plays a pivotal role in addressing these challenges. By extracting the most relevant information from NGS data, feature engineering enhances the performance of ML models[3]. This study focuses on applying Principal Component Analysis (PCA) for dimensionality reduction and implementing advanced normalisation techniques to prepare gene expression data for machine learning classification. A Support Vector Machine (SVM) classifier is employed to classify samples as either cancerous or normal. The results demonstrate the efficacy of these methods in improving classification accuracy and contribute to the development of robust diagnostic tools for lung cancer.

Methodology

A comprehensive methodology was employed to analyse the gene expression dataset for distinguishing cancerous and normal samples. The process included data preprocessing, feature selection, model training, and performance evaluation. A detailed overview of the methodology is illustrated in **Figure 1.** outlining each step from data acquisition to model validation.

Dataset Acquisition: The gene expression dataset used in this study is sourced from the International Cancer Genome Consortium (ICGC). The dataset consists of 488 cancerous and 55 normal samples for lung adenocarcinoma, while other cancer types in the ICGC repository have varying sample sizes. Each sample contains multiple features representing gene expression levels, providing a high-dimensional view of the molecular characteristics associated with cancer and normal conditions [4].

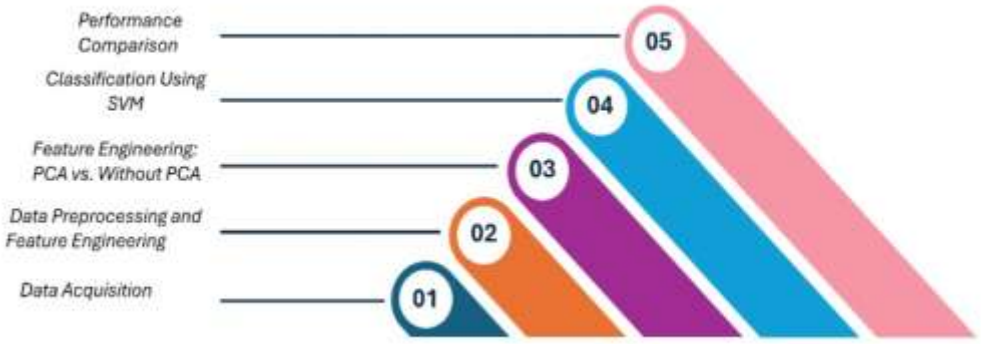

Figure 1. *Step wise flow chart of the methodology*

Data Preprocessing Data Cleaning: Missing values were addressed using mean imputation to ensure data completeness. Normalisation: Z-score normalisation was applied to scale the gene expression levels, ensuring that all features contributed equally to the classification task [5, 6].

Feature Engineering Dimensionality Reduction: PCA was applied to reduce the dimensionality of the dataset, retaining components that accounted for 95% of the variance. This step significantly reduced the feature space while preserving essential information[7].

Machine Learning Model A Support Vector Machine (SVM) classifier was designed and implemented to classify the processed data. The model included: Kernel Selection: A radial basis function (RBF) kernel was chosen to capture complex relationships in the data. Hyperparameter Optimization: The optimal values of C and gamma were determined using grid search cross-validation. Training and Validation: The dataset was split into 80% training and 20% validation subsets [8].

Results and Discussion

To evaluate the impact of dimensionality reduction on classification performance, the Support Vector Machine (SVM) classifier was trained and tested on the gene expression dataset under two conditions: (i) with PCA and (ii) without PCA. The model's performance was assessed using standard metrics, including accuracy, precision, recall, and F1-score. as shown in **Table 1.** presents a comparative analysis of classification performance for both conditions [9].

Table 1.*Performance Metrics of SVM with and without PCA*

Technique	Accuracy	Precision	Recall	F1_Score	ROC AUC
SVM	0.93	98	93	96	97
PCA+SVM	99	1	98	99	99

The results indicate that applying PCA led to a notable improvement in model performance by 85 reducing feature redundancy and enhancing computational efficiency. The SVM model with 86 PCA exhibited a higher accuracy and F1-score, demonstrating its effectiveness in 87 distinguishing between cancerous and normal samples[10].

Confusion Matrix Analysis: To further analyse the classification performance, confusion 89 matrices for both models were generated, as shown in **Figure 2.**

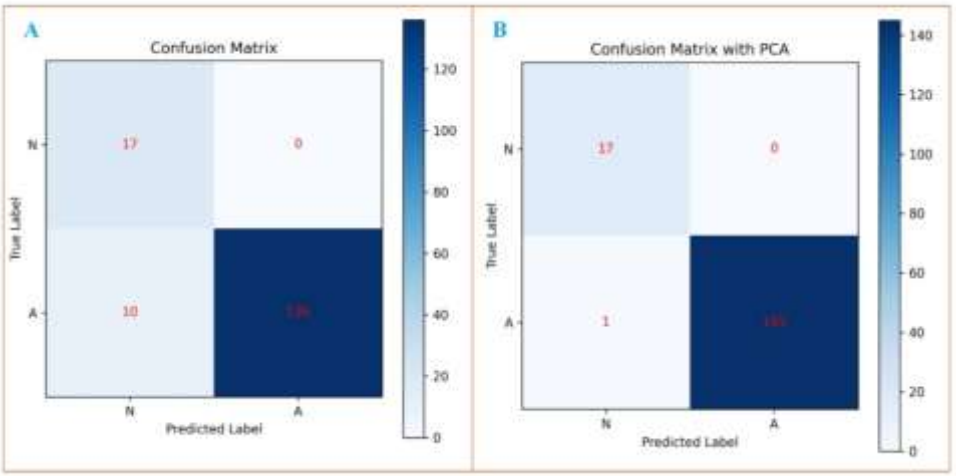

Figure 2.*Confusion Matrices A) SVM without PCA B) SVM with PCA*

The confusion matrices provide insights into the model's ability to correctly classify cancerous 98 and normal samples. The SVM model without PCA misclassified a higher number of normal 99 samples as cancerous, indicating potential overfitting due to high-dimensional data.

Conversely, the SVM with PCA exhibited a lower misclassification rate, suggesting that dimensionality reduction contributed to improved generalization [11]. *Impact of Feature Reduction on Model Performance:* The PCA-transformed dataset significantly reduced 103 computational complexity, leading to faster model training and validation. The PCA-based model demonstrated better performance on the validation set, reducing overfitting and 105 enhancing classification robustness. By retaining the principal components accounting for 95% 106 of variance, the model preserved key gene expression patterns associated with lung 107 adenocarcinoma. The results highlight the advantages of using PCA for dimensionality 108 reduction in high-dimensional gene expression datasets. The SVM classifier with PCA 109 outperformed the model trained on the original feature set, demonstrating higher accuracy, improved classification performance, and reduced computational cost[12, 13]. These findings underscore the significance of feature selection and dimensionality reduction techniques in biomedical data analysis.

Conclusion

This study underscores the indispensable role of feature engineering in optimising machine learning models for lung cancer classification using next-generation sequencing (NGS) data. The high-dimensional nature of NGS datasets necessitates robust techniques for dimensionality reduction and feature selection to ensure that machine learning models can efficiently process and classify the data. By employing Principal Component Analysis (PCA), we effectively reduced the feature space while preserving critical information, leading to improved model efficiency and reduced computational requirements. Advanced normalisation techniques further ensured that the data was appropriately scaled, enabling the model to achieve optimal performance. The implementation of an optimised Support Vector Machine (SVM) classifier demonstrated the tangible benefits of these feature engineering techniques, as evidenced by notable improvements in classification accuracy, precision, recall, and overall robustness. These advancements provide a framework for leveraging machine learning in the realm of precision medicine, enabling early and accurate detection of lung cancer, which is critical for improving patient outcomes.

While this study provides valuable insights, it also opens avenues for future exploration. The integration of multi-omics data, which combines genomics, transcriptomics, and proteomics, could further enhance the predictive capabilities of machine learning models. Additionally, exploring more sophisticated deep learning architectures, such as Convolutional Neural Networks (CNNs) or Graph Neural Networks (GNNs), may provide deeper insights into the complex relationships within genomic data. As technology advances, the methodologies and findings presented in this study can serve as a foundation for developing more effective diagnostic tools in oncology, contributing to the ongoing pursuit of precision medicine.

Author Contributions
Conceptualisation, S.N.A.S.; methodology, S.N.A.S.; implementation and coding, S.N.A.S.; writing—original draft preparation, S.N.A.S.; writing—review, S.N.A.S.; visualisation, S.N.A.S.; supervision, R.P.; editing and improvements, R.P. All authors have read and agreed to the published version of the manuscript.

Funding
This research received no external funding.

Institutional Review Board Statement
This study did not require ethical approval.

Informed Consent Statement Not applicable.

Data Availability Statement
The dataset used in this study is available in the International Cancer Genome Consortium (ICGC) repository, accessible at https://dcc.icgc.org.

Conflicts of Interest
The authors declare no conflicts of interest.

References
1. de Groot, P.M., et al., *The epidemiology of lung cancer.* Translational lung cancer research, 2018. **7**(3): p. 220.
2. Mehta, A., et al., *Biomarker testing for advanced lung cancer by next-generation sequencing; a valid method to achieve a comprehensive glimpse at mutational landscape.* Applied Cancer Research, 2020. **40**: p. 1-12.
3. Singh, D., et al. *Fsnet: Feature selection network on high-dimensional biological data*. in *2023 International Joint Conference on Neural Networks (IJCNN)*. 2023. IEEE.
4. Liu, S. and W. Yao, *Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection.* BMC bioinformatics, 2022. **23**(1): p. 175.
5. Zelaya, C.V.G. *Towards explaining the effects of data preprocessing on machine learning*. in *2019 IEEE 35th international conference on data engineering (ICDE)*. 2019. IEEE.
6. Rahman, A., *Statistics-based data preprocessing methods and machine learning algorithms for big data analysis.* International Journal of Artificial Intelligence, 2019. **17**(2): p. 44-65.
7. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 2016. **374**(2065): p. 20150202.
8. Schuhmann, R.M., A. Rausch, and T. Schanze, *Parameter estimation of support vector machine with radial basis function kernel using grid search with leave-p-out cross validation for classification of motion patterns of subviral particles.* Current Directions in Biomedical Engineering, 2021. **7**(2): p. 121-124.
9. Reddy, G.T., et al., *Analysis of dimensionality reduction techniques on big data.* Ieee Access, 2020. **8**: p. 54776-54788.
10. Sucharita, S., B. Sahu, and T. Swarnkar. *An Empirical Analysis of PCA-SVM Model for Cancer Microarray Data Classification*. in *Advances in Intelligent Computing and Communication: Proceedings of ICAC 2020*. 2021. Springer.
11. Li, G.-Z., et al., *Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis.* BMC genomics, 2008. **9**: p. 1-15.
12. Octaria, E.A., et al. *Kernel PCA and SVM-RFE based feature selection for classification of dengue microarray dataset*. in *AIP Conference Proceedings*. 2020. AIP Publishing.
13. De Souza, J.T., A.C. De Francisco, and D.C. De Macedo, *Dimensionality reduction in gene expression data sets.* IEEE access, 2019. **7**: p. 61136-61144.