ASEC 2024 Conference

The 5th International Electronic Conference on Applied Sciences

04–06 December 2024 | Online

MDPI

Feature Engineering for Lung Cancer Classification Using Next-Generation Sequencing Data

Syed Naseer Ahmad Shah, Rafat Parveen

Department of Computer Science, Jamia Millia Islamia, New Delhi, India-110025

INTRODUCTION & AIM

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 lung cancer, facilitating more targeted approaches in diagnosis, treatment, and research. While NGS data are high dimensional and complex, it poses significant challenge for data analysis and classification tasks. In this paper, we investigated the feature engineering to improve classification accuracy of lung cancer using NGS data. dimensionality reduction method Principal Component Analysis (PCA) is used to optimise feature selection along with the transformation techniques like normalization and scaling to optimize the data for better model performance. The efficacy of this technique is evaluated using the machine learning classifier Support Vector Machines (SVM). The results demonstrate the efficiency of feature engineering enhances the classification accuracy and robustness of lung cancer prediction models, providing valuable insights for the development of precision medicine approaches in oncology.

METHOD

The study used the GSE32863 dataset containing gene expression profiles. Data preprocessing involved normalization, scaling, and handling missing values. PCA was employed for dimensionality reduction, while machine learning models like SVM were evaluated using metrics such as accuracy, precision, recall, and F1-score as shown in **Figure 1.**

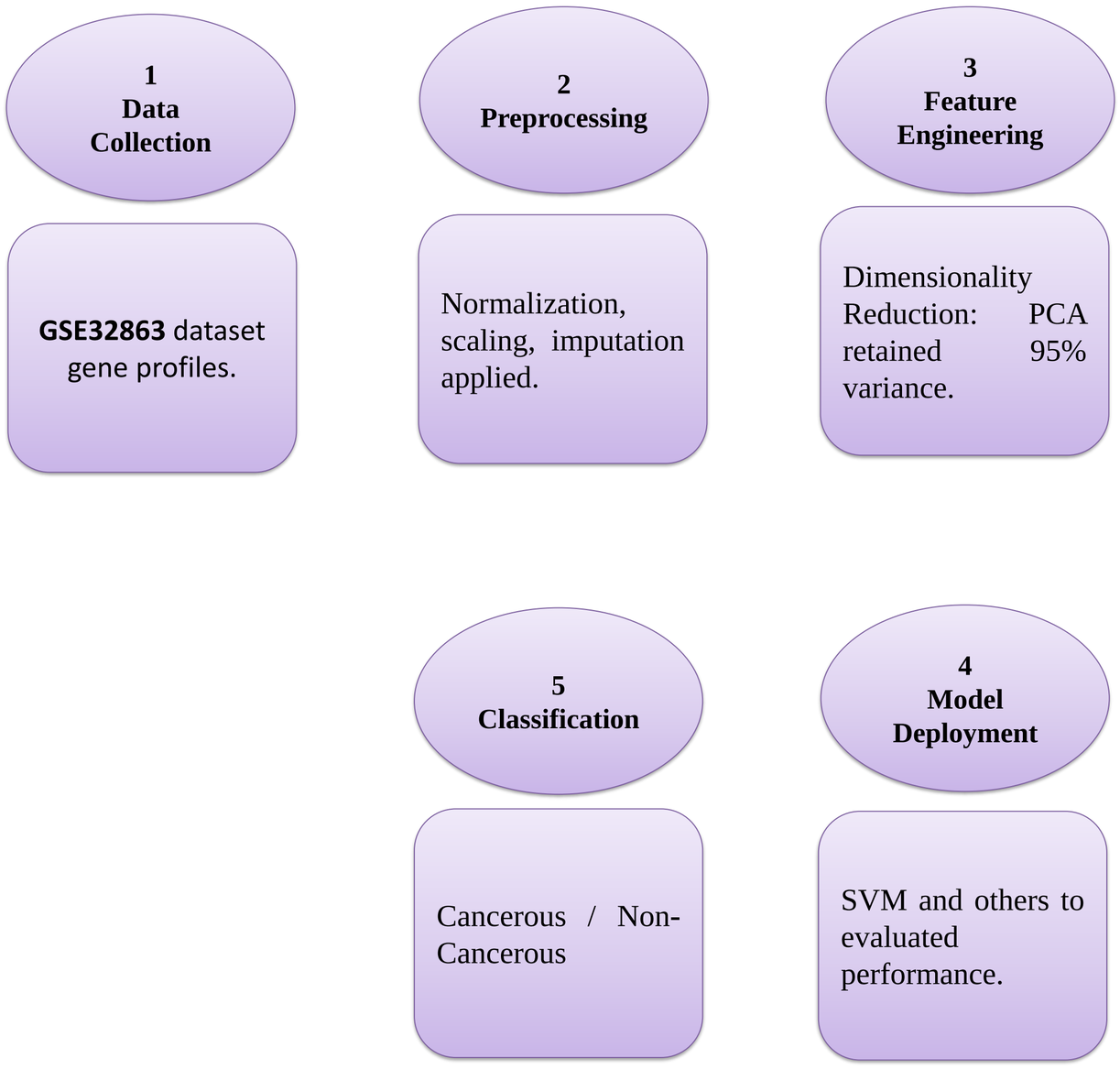

Figure 1. *Step-wise Methodology of the Study*.

RESULTS & DISCUSSION

Feature Reduction with PCA:

PCA reduced the dataset from 20,000 features to 200 principal components, retaining 95% of the variance while eliminating redundant information. This improved computational efficiency, reduced the risk of overfitting, and preserved essential data patterns for accurate classification. SVM and other machine learning algorithms were applied to the engineered dataset, and a comparative analysis shown in **Table 1.** highlighted SVM's superior performance in terms of accuracy, precision, and robustness over other classifiers. These results demonstrate the effectiveness of PCA in optimizing feature engineering for lung cancer prediction.

Table 1. *Shows the Comparative Analysis of the different Machine learning Models.*

Classifier	Accuracy	Precision	Recall	F1-Score
Support Vector Machine (SVM)	96.30%	91.80%	92.60%	92.20%
Random Forest	89.70%	88.50%	90.10%	89.30%
k-Nearest Neighbors (k-NN)	85.20%	84.70%	85.50%	85.10%
Logistic Regression	81.40%	80.90%	81.20%	81.00%

Key Observation

- SVM outperformed other classifiers in terms of accuracy and F1-score, emphasizing its robustness and stability with engineered features.
- The normalization and scaling of features significantly improved model convergence and reduced the risk of overfitting.
- PCA-driven dimensionality reduction streamlined the machine learning pipeline, resulting in faster processing, reduced complexity, and higher accuracy.

CONCLUSION

Efficient feature engineering, including dimensionality reduction through PCA and data transformation techniques, enhances the performance of machine learning classifiers. These findings contribute to the development of precision oncology, aiding in better diagnostic and therapeutic decisions for lung cancer.

FUTURE WORK / REFERENCES

Multi-omics data, deep learning models, and PCA with algorithms to improve results.

1. *Dataset: GSE32863. Accessible via NCBI GEO.*
2. *Smith et al. "Feature Engineering Techniques for High-Dimensional Data." Bioinformatics Journal, 2023.*
3. *Johnson et al. "PCA Applications in Genomics." Journal of Genomic Data Science, 2022.*