Early Lung Cancer Prediction Using Machine Learning

[1] Pavan Kumar, [1] Pranay Sai, [1] Niteesh Kumar, [*] Sibendu Samanta

[1] Department of Computer Science and Engineering, SRM University - Andhra Pradesh

[*] Department of Electronics and Communications Engineering, SRM University - Andhra Pradesh

Corresponding Author - Pavankumar_doppalapudi@srmap.edu.in

Abstract

Lung cancer remains a leading cause of mortality worldwide, underscoring the importance of early detection to improve patient outcomes. This study proposes a framework for predicting lung cancer risk using machine learning algorithms combined with a robust feature selection strategy. A dataset of 309 samples with 15 features spanning lifestyle factors and health indicators is processed using techniques like label encoding, Synthetic Minority Oversampling Technique (SMOTE) for class balance, and outlier management. Principal Component Analysis (PCA) and low-variance filtering are applied to refine the most impactful features, reducing data dimensionality. Six machine learning models K-Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), XGBoost, Naive Bayes (NB), and Logistic Regression (LR) are used to evaluate lung cancer risk, with accuracy, precision, recall, and F1-score as performance metrics. Results show that Random Forest and Logistic Regression achieved the highest accuracy, with RF reaching 96.91% after PCA-based feature selection. This study demonstrates that combining machine learning and feature selection offers a valuable approach to early lung cancer prediction. Future work could incorporate neural networks and ensemble methods to further enhance predictive reliability.

Keywords - Lung Cancer, Early Prediction, Machine Learning, SMOTE, KNN, Decision Tree, Random Forest, Performance Evaluation

1. Existing System

Current methods for detecting lung cancer rely primarily on advanced imaging technologies, including Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), and X-rays, which are the mainstay in assessing tumour location, size, and disease progression. These methods are highly effective at identifying and staging lung cancer, yet they require significant infrastructure, are costly, and involve skilled medical professionals for accurate analysis. Due to the resource-intensive nature of these technologies, they remain out of reach for many, particularly in remote or low-income areas. Routine screening, which is essential for early detection, is often not feasible in these regions, leading to late diagnoses when the disease has already progressed. This reliance on imaging has created disparities in healthcare access, as these systems are limited by both high operational costs and the demand for specialized personnel, leading to avoidable delays in diagnosis and care.

In recent years, machine learning (ML) has emerged as a promising tool in medical diagnosis, offering potential solutions to these accessibility issues. However, early Machine Learning based lung cancer prediction models often lack sophisticated data preprocessing and feature selection techniques, which limits their effectiveness. Many existing ML systems depend on basic dataset features without addressing common data issues like noise, class imbalance, and redundancy. These limitations reduce the reliability and accuracy of predictions, as the models may either overfit or underrepresent crucial data features. Additionally, simpler ML models do not account for nuanced patient-specific factors, such as detailed smoking history, environmental exposure, and genetic predispositions, which are crucial in accurately assessing lung cancer risk. Without a targeted feature selection approach, these systems may suffer from low prediction accuracy, further limiting their applicability in clinical scenarios.

Most early ML models lack advanced techniques like data balancing and outlier handling, which are essential for making accurate and unbiased predictions. For example, imbalances in data classes can cause models to favour the more common class (e.g., non-cancerous cases), resulting in missed detections of cancer cases. Noise and redundant data also complicate the prediction process, often overwhelming the model with irrelevant information. In many cases, the lack of refined feature selection prevents the model from isolating the most predictive variables, which are crucial for improving model performance and reducing computational load. As a result, these models may fall short in providing reliable early detection solutions for lung cancer, particularly when compared to traditional imaging, which remains the preferred method despite its limitations. Consequently, there is a clear need for an enhanced approach that integrates ML with comprehensive preprocessing and feature selection to support more accurate, accessible, and timely lung cancer detection.

2. Proposed System

The proposed system introduces an optimised framework for lung cancer prediction that utilizes a combination of machine learning algorithms and hybrid feature selection techniques. The dataset for this model comprises 309 samples with 15 attributes relevant to lung cancer risk, including lifestyle factors, smoking habits, and the presence of chronic conditions. To ensure the dataset is of high quality and can support accurate predictions, preprocessing steps such as label encoding are applied to convert categorical data into numerical values, making it suitable for ML algorithms. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance class distributions, addressing the challenge of class imbalance by adding synthetic examples to the minority class. This step ensures that the model does not favour one class over another, which is crucial for maintaining prediction reliability.

The system further incorporates a hybrid feature selection process that combines Principal Component Analysis (PCA) with low-variance filtering to reduce the dataset's dimensionality while preserving the most informative features. PCA helps by transforming original features into principal components, which retain the highest variance and thus the most valuable predictive information. Low-variance filtering complements this by removing features with minimal variation, which are less likely to influence the model's accuracy. Together, these techniques allow the model to focus on the most impactful features, streamlining the dataset and improving model performance. This approach reduces noise and irrelevant data, which can otherwise obscure meaningful patterns in the data and reduce the model's effectiveness in making accurate predictions.

For this study, six machine learning algorithms those are K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), XGBoost, Naive Bayes (NB), and Logistic Regression (LR) were implemented to evaluate lung cancer risk. The models' performance was assessed using metrics such as accuracy, precision, recall, and F1-score, ensuring a thorough evaluation. Experimental results indicated that both Random Forest and Logistic Regression achieved the highest accuracy, with Random Forest attaining a 96.91% accuracy rate after applying PCA-based feature selection. By concentrating on relevant features and reducing dimensionality, the proposed system outperformed traditional models by effectively minimizing noise and irrelevant information. This framework offers a scalable, cost-effective alternative to traditional imaging, making it an accessible option for early lung cancer detection, particularly in resource-limited settings. Future research could further refine this approach by integrating more complex models like neural networks and testing against both balanced and unbalanced datasets to increase robustness and reliability.

3. Theory Behind the Project

This project combines insights from medical science, statistical data processing, and machine learning to build a robust predictive model for lung cancer detection. By leveraging feature selection techniques, advanced data preprocessing, and multiple machine learning algorithms, this framework aims to deliver an accessible and accurate solution for early lung cancer diagnosis. Below are the key theoretical principles guiding the project:

3. 1. Medical Basis of Lung Cancer and Early Detection

Lung cancer is characterised by the abnormal growth of cells within lung tissues, leading to the formation of tumours that can disrupt normal lung function and spread to other parts of the body. It primarily manifests in two main types: Non-Small Cell Lung Cancer (NSCLC) and Small Cell Lung Cancer (SCLC). NSCLC is the more common type, accounting for approximately 85% of all lung cancer cases. It typically grows more slowly and can often be treated more effectively in its early stages. On the other hand, SCLC is less common but more aggressive, rapidly spreading to other organs and tissues, making it harder to treat and more likely to result in poor outcomes. Understanding the differences between these two types is crucial for determining appropriate treatment strategies, as each type has distinct clinical features and responds differently to therapies.

Several major risk factors contribute to the development of lung cancer, with smoking being the most significant. Tobacco smoke contains carcinogenic compounds that damage lung cells, leading to mutations that cause uncontrolled cell division and tumour formation. Smoking alone is responsible for around 90% of lung cancer cases, but other factors also contribute to the risk, including exposure to secondhand smoke, air pollutants, asbestos, and genetic predisposition. Individuals with a family history of lung cancer or certain chronic health conditions, such as chronic obstructive pulmonary disease (COPD), are also at a higher risk. Given these multiple risk factors, identifying individuals who are at risk for lung cancer is essential for early intervention and prevention.

Early detection of lung cancer is critical, as it significantly improves treatment outcomes and survival rates. However, the disease often presents with subtle symptoms in its early stages, such as persistent cough or shortness of breath, which can easily be mistaken for less serious conditions like asthma or a respiratory infection. Traditional diagnostic methods, such as CT scans and MRIs, are effective but expensive and may not be widely accessible, especially in low-resource settings. In contrast, a machine learning-based predictive model could offer a more cost-effective and accessible alternative. By analysing data from various risk factors such as age, smoking history, environmental exposures, and family history, machine learning algorithms can identify individuals at higher risk for lung cancer, allowing for earlier detection and intervention. This approach holds great potential in enhancing early diagnosis and ultimately improving patient outcomes.

3. 2. Machine Learning Algorithms for Classification

This project evaluates six distinct machine learning models those are K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), XGBoost, Naive Bayes (NB), and Logistic Regression (LR) each of which brings unique strengths in classification and pattern recognition. Here’s a breakdown of each model’s theoretical basis:

- **K-Nearest Neighbor (KNN)**: This algorithm classifies a data point by analysing its ‘k’ nearest neighbours and assigning it to the class most common among them. It’s particularly effective for smaller datasets, where it can capture local patterns, though it may become computationally intensive with larger datasets.

- **Decision Tree (DT)**: Decision Tree models classify data by creating a tree-like structure that asks sequential questions based on feature values, branching off as needed. This interpretable model is effective at isolating key factors that influence classification outcomes.

- **Random Forest (RF)**: As an ensemble method, Random Forest builds multiple decision trees and averages their results, enhancing accuracy and robustness by minimizing overfitting and handling complex, noisy datasets effectively.

- **XGBoost**: This model uses gradient boosting, where each subsequent tree corrects the errors of the previous trees, leading to improved accuracy. It is powerful for high-dimensional data and can handle interactions among features.

- **Naive Bayes (NB)**: Based on Bayes' theorem, this probabilistic classifier assumes feature independence and calculates the probability of each class given certain features. It is highly efficient, especially with small datasets, but may oversimplify relationships among features.

- **Logistic Regression (LR)**: Logistic Regression is a statistical model that predicts the probability of a binary outcome by mapping input features to a logistic function. It is interpretable, straightforward, and often serves as a benchmark for classification tasks.

Together, these algorithms allow for a comparative analysis to determine which model performs best for lung cancer prediction, enabling a data-driven selection process.

3. 3. Feature Selection and Dimensionality Reduction Techniques

Feature selection and dimensionality reduction are essential components of this project, as they help in enhancing model performance and computational efficiency. This study uses a hybrid approach combining **Principal Component Analysis (PCA)** and **low-variance filtering**:

- **Principal Component Analysis (PCA)**: PCA is a dimensionality reduction method that identifies patterns in high-dimensional data by transforming original features into a set of principal components. These components capture the maximum variance, preserving only the most informative features. By reducing the dataset's dimensionality, PCA minimises redundancy and helps the model focus on the variables that contribute most to predictions.

- **Low-Variance Filtering**: This technique removes features with little variability across the dataset, as they are unlikely to be predictive. Features with low variance are often redundant or irrelevant, so excluding them improves the model's focus on impactful data, reducing computational complexity and preventing overfitting.

These feature selection techniques work together to streamline the dataset, filtering out noise and irrelevant features, which in turn enhances the accuracy and interpretability of the models.

3. 4. Data Preprocessing for Model Reliability

Data preprocessing ensures the quality and consistency of the data, creating a reliable foundation for machine learning models. The steps include:

- **Label Encoding**: Categorical data, such as gender and smoking status, is converted into numerical form using label encoding, enabling machine learning algorithms to process this information effectively.

- **Synthetic Minority Oversampling Technique (SMOTE)**: Since the dataset may have an imbalance between lung cancer cases and non-cancer cases, SMOTE is applied to balance the class distribution by generating synthetic instances of the minority class. This step is essential to avoid biased predictions that favor the majority class.

- **Outlier Management**: The process uses a method called Winsorization, which limits extreme values within a reasonable range to reduce their effect on the model without distorting the overall dataset.

- **Normalization**: Z-score normalization standardises the dataset by scaling numerical features, ensuring that variations in scale do not skew the model's learning process.

Together, these preprocessing steps increase data quality, allowing the model to identify patterns more effectively and produce consistent predictions.

4. Results and Discussion

In this study, machine learning models were employed to predict lung cancer outcomes based on a comprehensive dataset. The dataset underwent extensive preprocessing, including label encoding, class balancing with SMOTE, outlier handling, and normalization, to enhance the model's reliability and performance. Feature selection techniques, such as low-variance filtering and Principal Component Analysis (PCA), were used to reduce dimensionality and focus on the most relevant features. The models were evaluated using accuracy, precision, recall, and F1-score, with Random Forest and Logistic Regression emerging as the top performers. The impact of feature selection on model performance was significant, highlighting its importance in improving prediction accuracy, especially in medical applications.

4. 1. Data Preprocessing and Class Balancing

In this study, the dataset underwent rigorous preprocessing to ensure its quality and reliability for machine learning algorithms. The preprocessing steps included label encoding, Synthetic Minority Oversampling Technique (SMOTE), outlier handling, and normalization. Label encoding was used to convert categorical variables, such as gender and smoking status, into numerical values that could be processed by machine learning algorithms. SMOTE was applied to balance the dataset by generating synthetic samples for the minority class (positive lung cancer cases), thereby preventing the model from being biased toward the majority class (non-cancerous cases). This balancing step was crucial in ensuring that the model could effectively identify rare but critical instances of lung cancer, which is often underrepresented in medical datasets.

Outlier handling was done using Winsorization, where extreme values above the 95th percentile were capped, and those below the 5th percentile were adjusted to ensure the data distribution remained within a reasonable range. This process minimizes the impact of extreme values, which could otherwise skew the model’s predictions. Normalization, specifically Z-score normalization, was then applied to standardize numerical features, ensuring all data was on a uniform scale. These preprocessing steps were vital in enhancing the quality of the dataset, allowing the machine learning models to make accurate and unbiased predictions based on the most relevant features.

Additionally, the choice of preprocessing techniques is critical in ensuring the model's generalisability and performance across different data scenarios. These steps are particularly important in medical datasets, where imbalances, noise, and inconsistencies are common. By carefully addressing these issues, the model was better equipped to handle real-world clinical data and provide more reliable predictions for lung cancer diagnosis. Proper preprocessing is thus essential for building a robust model that can deliver actionable insights in medical practice.

4. 2. Feature Selection and Dimensionality Reduction

Feature selection and dimensionality reduction play a crucial role in improving the efficiency and accuracy of machine learning models. In this study, a hybrid feature selection approach was employed, combining low-variance filtering and Principal Component Analysis (PCA). Low-variance filtering removed features with minimal variation across the dataset, as such features are less likely to contribute to meaningful predictions. By eliminating these uninformative features, the model was better able to focus on the most relevant data, reducing the complexity of the model and improving its overall performance. This step is particularly useful when dealing with high-dimensional datasets, where many features may be redundant or irrelevant.

PCA was then applied to further reduce the dimensionality of the dataset. PCA transforms the original features into principal components that capture the most variance within the data. By retaining only the most significant components, PCA reduces the noise and computational complexity of the model. This dimensionality reduction helps to improve both the accuracy and generalizability of the machine learning models, as fewer features lead to better performance without overfitting. The combination of low-variance filtering and PCA allowed for the extraction of the most impactful features, which directly contributed to the success of the lung cancer prediction models.

4. 3. Model Performance Evaluation

The performance of the machine learning models was evaluated using various metrics, including accuracy, precision, recall, and F1-score. Accuracy was the primary metric used to measure the overall performance of each model, while precision and recall provided more detailed insights into the model’s ability to correctly classify positive and negative cases of lung cancer. Precision, which measures the proportion of true positives among all positive predictions, was essential for evaluating how well the models avoided false positives. Recall, on the other hand, indicated the model’s ability to correctly identify true positive cases, which is crucial in medical predictions where missing a positive case (false negative) could have serious consequences.

The F1-score, which is the harmonic mean of precision and recall, provided a balanced view of the models’ performance, especially in cases where there was an imbalance between the positive and negative classes. The results showed that the Random Forest (RF) and Logistic Regression (LR) models performed the best, with RF achieving an accuracy of 96.91%. This high performance can be attributed to RF's ensemble nature, which aggregates the results of multiple decision trees, reducing overfitting and improving generalisability. The F1-scores for both RF and LR were also high, indicating that they balanced precision and recall effectively, making them the most reliable models for lung cancer prediction in this study.

4. 4. Impact of Feature Selection on Model Performance

The application of feature selection techniques, particularly PCA, had a significant impact on the performance of the models. Before feature selection, models such as K-Nearest Neighbours (KNN) and Naive Bayes (NB) exhibited moderate performance, with accuracy values around 90%. However, after applying PCA and reducing the dimensionality of the dataset, these models showed notable improvements in accuracy. PCA helped in focusing the model's attention on the most critical features, minimizing the noise created by irrelevant or redundant data. As a result, the models were able to better differentiate between cancerous and non-cancerous cases, leading to more accurate predictions.

The impact of feature selection was particularly evident in the Random Forest model, which performed better after the feature reduction process. RF, as an ensemble method, benefits from reduced feature complexity because it can focus on the most informative features, thus improving its ability to generalize and make accurate predictions. Similarly, Logistic Regression, a simpler model, showed an improvement in both accuracy and F1-score with the reduced feature set. This reinforces the importance of feature selection techniques in enhancing model performance, particularly when dealing with high-dimensional datasets in medical applications where precision is critical.

4. 5. Model Comparison and Selection

After evaluating the performance of all six machine learning models—KNN, Decision Tree (DT), Random Forest (RF), XGBoost, Naive Bayes (NB), and Logistic Regression (LR)—the Random Forest and Logistic Regression models emerged as the most effective for lung cancer prediction. Random Forest achieved the highest accuracy of 96.91%, followed closely by Logistic Regression at 96.3%. These two models outperformed the others in terms of both accuracy and F1-score, suggesting that they were the best at distinguishing between positive and negative cases of lung cancer in the dataset.

The Decision Tree model, while effective, struggled with overfitting, especially when applied to high-dimensional data, leading to lower performance. XGBoost, although a powerful model in many scenarios, did not significantly outperform Random Forest or Logistic Regression in this particular study. Naive Bayes, despite being a simple and computationally efficient model, performed poorly compared to the other algorithms. This comparison highlights the importance of model selection in machine learning applications, particularly in fields like medical diagnostics, where high accuracy and low false negatives are critical.

4. 6. Future Directions and Model Enhancement

While the Random Forest and Logistic Regression models performed well in this study, there are several ways in which the model can be further improved in future work. One potential avenue for enhancement is the integration of deep learning models, such as neural networks, which have shown great promise in fields like image and sequential data analysis. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated their ability to detect intricate patterns in large, high-dimensional datasets, which may be overlooked by traditional machine learning models. These models excel in handling complex data, especially when dealing with unstructured data like images or time-series information. Incorporating these advanced models into the predictive framework could improve accuracy and uncover hidden correlations in the data, thereby providing a more robust and reliable model. Furthermore, the increasing availability of computational power makes it feasible to implement deep learning models, which could substantially enhance the model's performance in making predictions.

Another area for improvement lies in the inclusion of more diverse datasets, which can contribute to a more comprehensive and accurate model. The current model may be limited by the data it has been trained on, and including additional clinical features such as genetic information, medical history, environmental factors, and even lifestyle choices could enhance its predictive power. By incorporating a broader range of data, the model would be able to account for a wider array of variables that can influence patient outcomes, providing a more holistic and nuanced prediction. Additionally, integrating multimodal data sources, such as genomic data, patient records, and imaging data, could offer a more complete picture of the patient’s health. This would make the model more adaptable and scalable, ensuring it is suitable for various medical settings with diverse patient profiles and an extensive set of clinical features. With such improvements, the model could provide more accurate, personalized predictions for patient care.

Another promising direction for future work is the exploration of ensemble methods, where multiple models are combined to improve prediction accuracy. Ensemble techniques like stacking, boosting, and bagging can leverage the strengths of various models to reduce errors, improve stability, and enhance performance. Stacking, for example, combines the predictions of several models to generate a more accurate final result, while boosting improves weak learners by focusing on previously misclassified data points. Bagging, such as Random Forest, helps reduce variance by training multiple models on different data subsets. These methods could make the model more robust, enabling it to better handle noisy or imbalanced data, which is often present in real-world clinical datasets.

5. Flowchart, Tables and Outputs

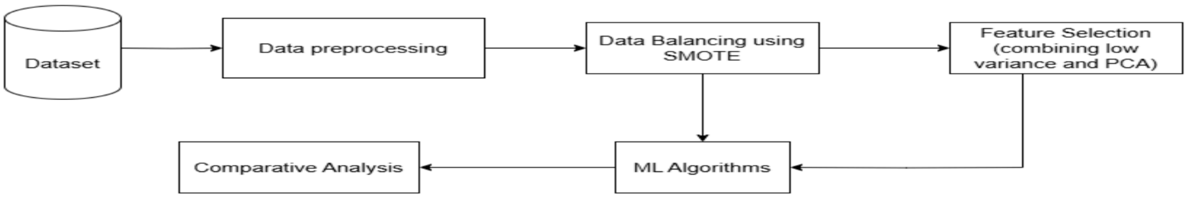
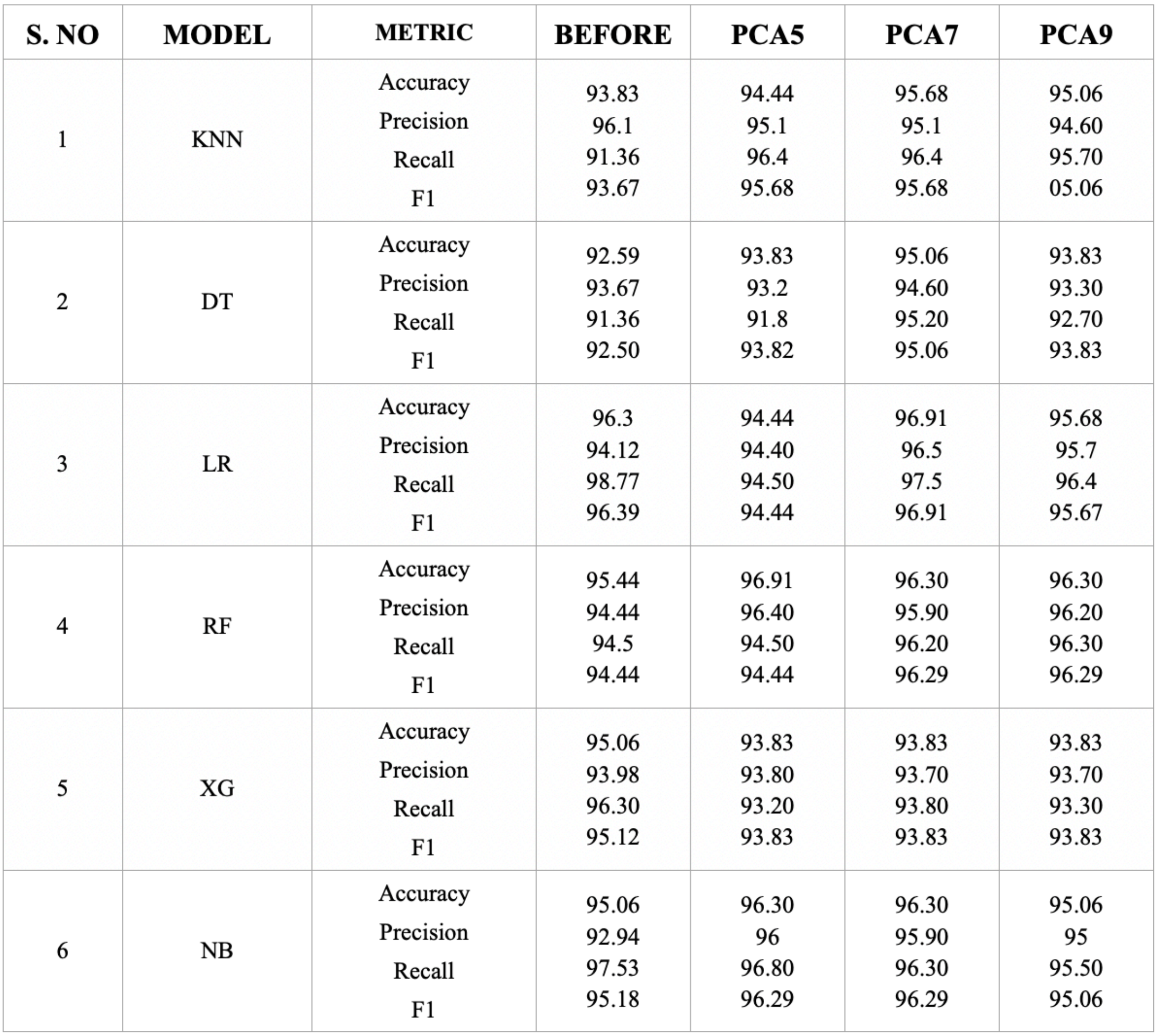

Dataset
Data preprocessing
Data Balancing using SMOTE
Feature Selection (combining low variance and PCA)
ML Algorithms
Comparative Analysis

Fig. 1: Workflow of our method



S. NO	MODEL	METRIC	BEFORE	PCA5	PCA7	PCA9
1	KNN	Accuracy Precision Recall F1	93.83 96.1 91.36 93.67	94.44 95.1 96.4 95.68	95.68 95.1 96.4 95.68	95.06 94.60 95.70 05.06
2	DT	Accuracy Precision Recall F1	92.59 93.67 91.36 92.50	93.83 93.2 91.8 93.82	95.06 94.60 95.20 95.06	93.83 93.30 92.70 93.83
3	LR	Accuracy Precision Recall F1	96.3 94.12 98.77 96.39	94.44 94.40 94.50 94.44	96.91 96.5 97.5 96.91	95.68 95.7 96.4 95.67
4	RF	Accuracy Precision Recall F1	95.44 94.44 94.5 94.44	96.91 96.40 94.50 94.44	96.30 95.90 96.20 96.29	96.30 96.20 96.30 96.29
5	XG	Accuracy Precision Recall F1	95.06 93.98 96.30 95.12	93.83 93.80 93.20 93.83	93.83 93.70 93.80 93.83	93.83 93.70 93.30 93.83
6	NB	Accuracy Precision Recall F1	95.06 92.94 97.53 95.18	96.30 96 96.80 96.29	96.30 95.90 96.30 96.29	95.06 95 95.50 95.06

Table - 1 : Performance metrics of various models before and after PCA with different components

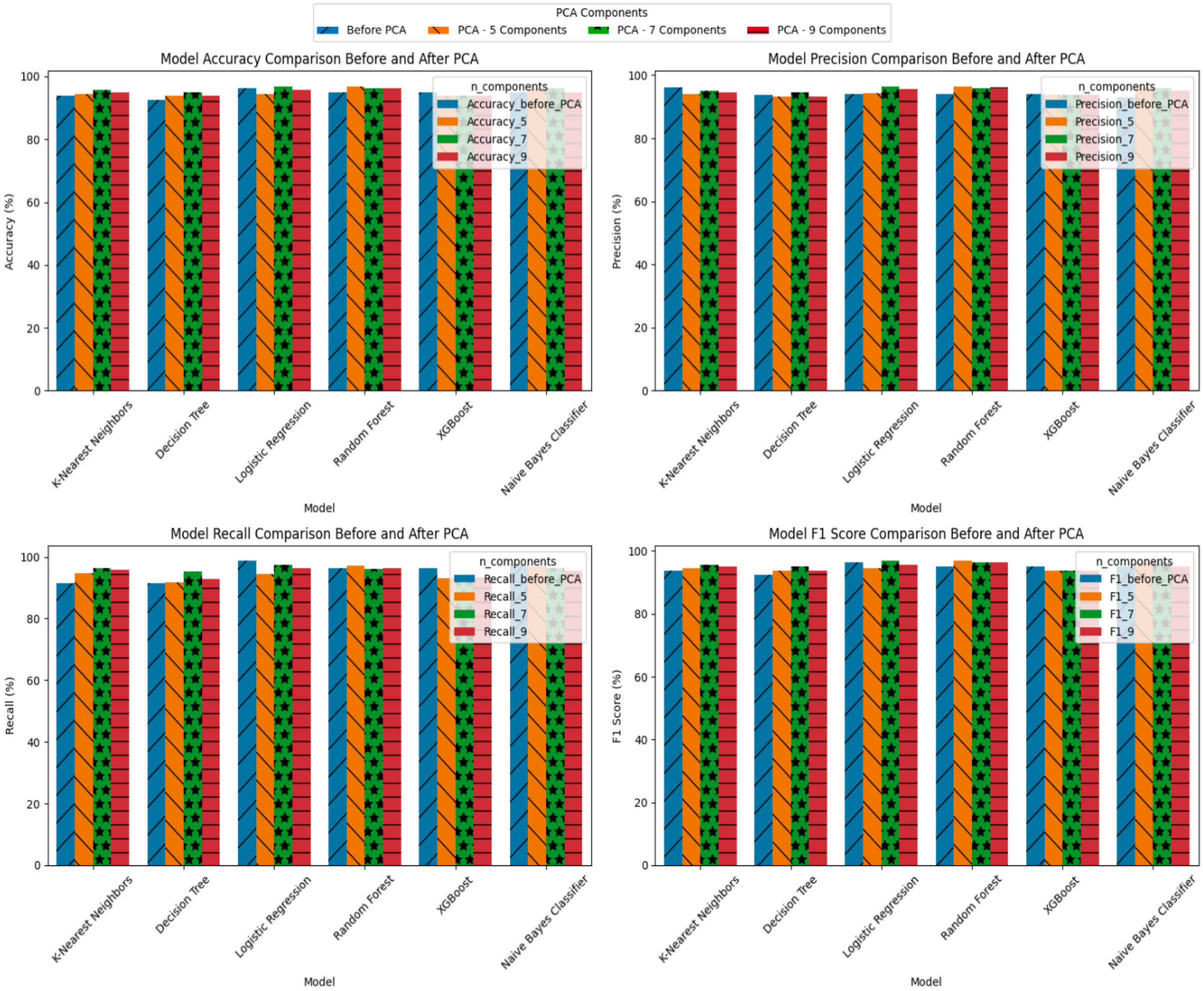
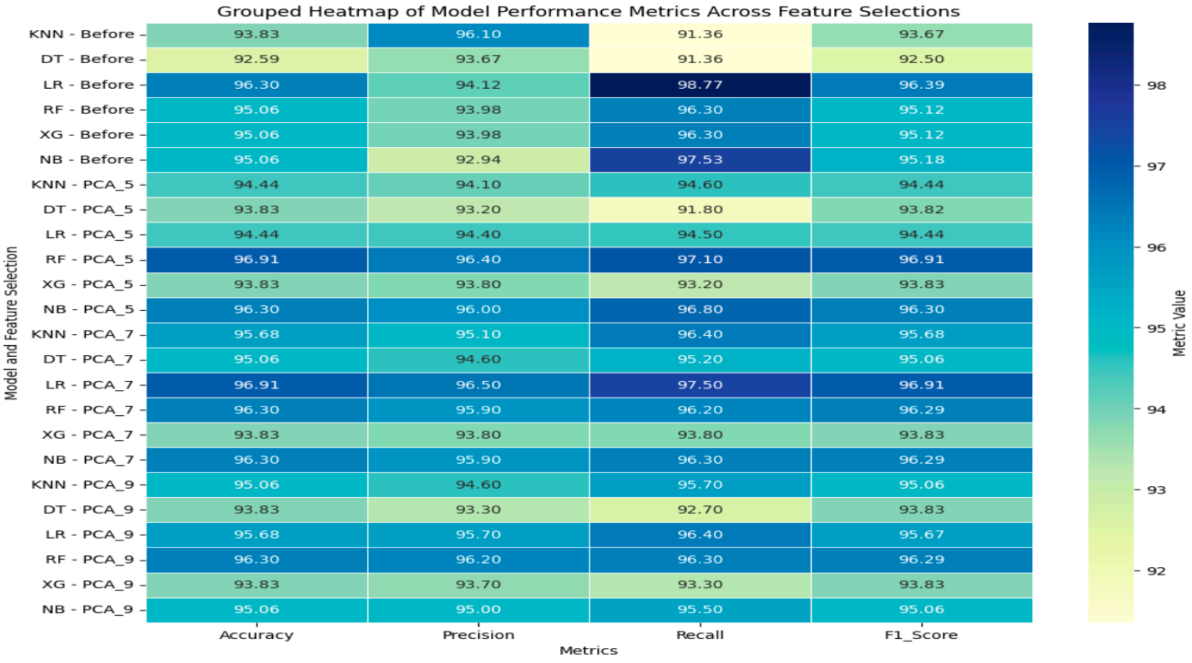


Fig. 2: Performance of With and Without PCA



	Accuracy	Precision	Recall	F1_Score
KNN - Before	93.83	96.10	91.36	93.67
DT - Before	92.59	93.67	91.36	92.50
LR - Before	96.30	94.12	98.77	96.39
RF - Before	95.06	93.98	96.30	95.12
XG - Before	95.06	93.98	96.30	95.12
NB - Before	95.06	92.94	97.53	95.18
KNN - PCA_5	94.44	94.10	94.60	94.44
DT - PCA_5	93.83	93.20	91.80	93.82
LR - PCA_5	94.44	94.40	94.50	94.44
RF - PCA_5	96.91	96.40	97.10	96.91
XG - PCA_5	93.83	93.80	93.20	93.83
NB - PCA_5	96.30	96.00	96.80	96.30
KNN - PCA_7	95.68	95.10	96.40	95.68
DT - PCA_7	95.06	94.60	95.20	95.06
LR - PCA_7	96.91	96.50	97.50	96.91
RF - PCA_7	96.30	95.90	96.20	96.29
XG - PCA_7	93.83	93.80	93.80	93.83
NB - PCA_7	96.30	95.90	96.30	96.29
KNN - PCA_9	95.06	94.60	95.70	95.06
DT - PCA_9	93.83	93.30	92.70	93.83
LR - PCA_9	95.68	95.70	96.40	95.67
RF - PCA_9	96.30	96.20	96.30	96.29
XG - PCA_9	93.83	93.70	93.30	93.83
NB - PCA_9	95.06	95.00	95.50	95.06

Fig. 3: Grouped Heatmap of model performance metrics across Feature selections

6. Conclusion

This study effectively demonstrated the potential of machine learning models in predicting lung cancer outcomes with high accuracy. By applying a series of preprocessing techniques, including label encoding, class balancing using SMOTE, outlier handling, and normalization, the dataset was optimised to ensure reliable and unbiased predictions. Feature selection methods, such as low-variance filtering and Principal Component Analysis (PCA), played a critical role in reducing dimensionality and focusing on the most relevant features. These steps helped eliminate noise and irrelevant data, improving the overall performance and efficiency of the models. Among the models evaluated, Random Forest and Logistic Regression stood out as the most reliable, achieving high accuracy and F1-scores, making them suitable candidates for predicting lung cancer outcomes.

The success of this study highlights the importance of thoughtful data preprocessing and feature selection in machine learning applications, particularly in medical fields where precision is crucial. The Random Forest model's ensemble nature and Logistic Regression's simplicity allowed them to perform well despite the challenges posed by high-dimensional medical datasets. Additionally, the application of class balancing techniques like SMOTE ensured that the model did not become biased toward the majority class. While these models performed admirably, future work could explore the integration of deep learning models, which have shown promise in uncovering more complex patterns, and the inclusion of diverse datasets, such as genetic and medical history data, to further enhance predictive accuracy. By advancing these techniques, the predictive capabilities of machine learning models in clinical settings can be significantly improved, offering more reliable early detection systems for lung cancer and other medical conditions.

References -

1. K. Sivanagireddy, S. Yerram, S. S. N. Kowsalya, S. Sivasankari, J. Surendiran, and R. Vidhya, “Early lung cancer prediction using corre- lation and regression,” in 2022 International Conference on Computer, Power and Communications (ICCPC), pp. 24–28, IEEE, 2022.

2. C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, “Early stage lung cancer prediction using various machine learning techniques,” in 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1285–1292, IEEE, 2020.

3. N. Banerjee and S. Das, “Prediction lung cancer–in machine learning perspective,” in 2020 International conference on computer science, engineering and applications (ICCSEA), pp. 1–5, IEEE, 2020.

4. A. Chauhan et al., “Detection of lung cancer using machine learning techniques based on routine blood indices,” in 2020 IEEE international conference for innovation in technology (INOCON), pp. 1–6, IEEE, 2020.

5. I. Mohamed, M. M. Fouda, and K. M. Hosny, “Machine learning algorithms for copd patients readmission prediction: A data analytics approach,” IEEE Access, vol. 10, pp. 15279–15287, 2022.

6. M. S. Kumar and K. V. Rao, “Prediction of lung cancer using machine learning technique: A survey,” in 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–5, IEEE, 2021.

7. A. A.-C. Omar and A. B. Nassif, “Lung cancer prediction using machine learning based feature selection: A comparative study,” in 2023 Advances in Science and Engineering Technology International Conferences (ASET), pp. 1–6, IEEE, 2023.

8. S. Shanthi and N. Rajkumar, “Lung cancer prediction using stochastic diffusion search (sds) based feature selection and machine learning methods,” Neural Processing Letters, vol. 53, no. 4, pp. 2617–2630, 2021.

9. S. S. Raoof, M. A. Jabbar, and S. A. Fathima, “Lung cancer prediction using machine learning: A comprehensive approach,” in 2020 2nd Inter- national conference on innovative mechanisms for industry applications (ICIMIA), pp. 108–115, IEEE, 2020.

10. P. Chaturvedi, A. Jhamb, M. Vanani, and V. Nemade, “Prediction and classification of lung cancer using machine learning techniques,” in IOP conference series: materials science and engineering, vol. 1099, p. 012059, IOP Publishing, 2021.

11. J.Li,Y.Tao,andT.Cai,“Predictinglungcancersusingepidemiological data: A generative-discriminative framework,” IEEE/CAA Journal of Automatica Sinica, vol. 8, no. 5, pp. 1067–1078, 2021.

12. S. Alagarsamy, R. R. Subramanian, T. Shree, M. Balasubramanian, V. Govindaraj, et al., “Prediction of lung cancer using meta-heuristic based optimization technique: Crow search technique,” in 2021 In- ternational Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 186–191, IEEE, 2021.

13. G. Gupta, V. Kumar, and P. Karuppanan, “Study & analysis of lung cancer risk prediction techniques using ml and dl algorithms,” in 2024 IEEE Students Conference on Engineering and Systems (SCES), pp. 1–6, IEEE, 2024.

14. J. A. Bartholomai and H. B. Frieboes, “Lung cancer survival prediction via machine learning regression, classification, and statistical tech- niques,” in 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 632–637, IEEE, 2018.

15. C. Modak, M. A. Shahriyar, M. S. Taluckder, M. S. Haque, and M. A. Sayed, “A study of lung cancer prediction using machine learning algorithms,” in 2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS), pp. 213–217, IEEE, 2023.

16. R. Patra, “Prediction of lung cancer using machine learning classifier,” in Computing Science, Communication and Security: First International Conference, COMS2 2020, Gujarat, India, March 26–27, 2020, Revised Selected Papers 1, pp. 132–142, Springer, 2020.

17. P. K. Singh, V. Chaudhary, S. Jain, and G. A. Kumar, “Compare and contrast of machine learning classification algorithms to predict accuracy and performance of lung cancer disease,” in 2021 International Conference on Technological Advancements and Innovations (ICTAI), pp. 96–101, IEEE, 2021.

18. K.Ingle, U.Chaskar and S.Rathod,“Lung cancer types prediction using machine learning approach,” in 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 01–06, IEEE, 2021.

19. S. M. Nimmagadda, K. Likhitha, G. Srilatha, and S. M. Sree, “Lung cancer prediction and classification using machine learning algorithms,” in 2024 International Conference on Expert Clouds and Applications (ICOECA), pp. 1012–1015, IEEE, 2024.