Grenze International Journal of Engineering and Technology, June Issue

Lung Cancer Prediction using Rotation Forest Model with Principal Component Analysis

Megha P[1], Smitha John[2] and T Jemima Jebaseeli[3]
[1,2]Division of Data Science and Cyber Security, Karunya Institute of Technology and Sciences, Coimbatore.
Email: meghap21@karunya.edu.in, smithajohn@karunya.edu
[3]Division of Artificial Intelligence and Machine Learning, Karunya Institute of Technology and Sciences, Coimbatore
Email: jemima_jeba@karunya.edu

***Abstract*—The global prevalence of the fatal condition known as lung cancer requires developers to create precise early detection predictive models. The study presents a new method which combines the Rotation Forest model with Principal Component Analysis (PCA) for improving predictions of lung cancer. The Rotation Forest model combined with Principal Component Analysis obtained increased efficiency by using PCA to eliminate features that duplicated information. The proposed model delivered higher predictive accuracy and reliability for Elvira Biomedical Dataset and surpassed traditional ensemble methods for this purpose. The combined methodology shows experimental support with successful results that deliver high precision rates and recall rates. This research demonstrates how joining ensemble classifiers with feature selection methods creates improved diagnostic capabilities that generate dependable tools for pulmonary cancer detection by medical workers.**

***Index Terms*—Lung cancer prediction, rotation forest, principal component analysis, ensemble learning, machine learning, feature selection.**

I. INTRODUCTION

Lung cancer ranks among the primary cancer-related mortality worldwide by becoming the main reason for annual cancer diagnostic and mortality statistics. The main difficulty of detecting lung cancer emerges from the fact that symptoms are hardly present in its initial stages which requires early diagnosis to achieve successful treatment outcomes. The implementation of Machine Learning (ML) techniques throughout recent years has proven promising for medical diagnostics because they strengthen disease prediction models [1] [2].

Machine learning techniques achieve superior performance activities through ensemble learning models because their base learner aggregation method decreases prediction variance and increases accuracy levels [3]. The Rotation Forest model stands apart as a new ensemble learning approach that implements diversity in base learners by performing feature extraction and rotation. The implementation of this method boosts precision as well as increases the dataset's resistance to overfitting problems that frequently occur in high-dimensional medical datasets [4]. The prediction models encounter significant challenges when medical datasets contain both irrelevant along with redundant features. Additional features contribute to both reduced performance quality and elevated computational operation complexity. The dimensionality reduction technique known as Principal Component Analysis (PCA) transforms features into uncorrelated components through a transformation process that solves the redundant feature issue [5].

Grenze ID: 01.GIJET.11.2.612

The proposed method achieves enhanced prediction reliability together with computational efficiency by uniting Rotation Forest with PCA functionality.

The proposed research implements a hybrid method that unites Rotation Forest with PCA for improved lung cancer prediction through benchmark datasets.

The proposed method contains three main parts which include the following.

- Implementing Rotation Forest for lung cancer prediction
- Incorporating PCA for dimensionality reduction
- Showing experimental results proving this new approach delivers better results than standard ensemble models.

This study develops an innovative approach which uses advanced machine learning techniques to present a dependable and time-saving diagnosis instrument for clinical purposes. This research follows the following structure after the introduction. A thorough evaluation of previous works in lung cancer prediction takes place in Section II. The methodology segment of this work includes presentation of data pre-processing techniques alongside model design elements and metrics for evaluation. The experimental results along with their comparison details are presented in Section IV. The paper ends in Section V by describing research opportunities for the future.

II. LITERATURE SURVEY

Research has strongly focused on early lung cancer detection because it directly affects survival possibilities for patients. The diagnostic fields use machine learning (ML) with ensemble methods extensively for improving accuracy and reliability outcomes. Abuya [1] utilized Elvira Biomedical Dataset to predict lung cancer through ensemble classification combined with PCA analysis. The research demonstrated how PCA successfully reduces dataset dimensions which generates better classification results. Morgado et al. [2] researched various methods for selecting features that would help determine EGFR mutation status in lung cancer patients. ML models obtain superior performance through feature selection because it enables the removal of redundant variables according to their research. Ganaie et al. [3] presented the Rotation Forest model together with its oblique extension as an approach that boosts ensemble learner diversity effectiveness. Rotation Forest generates greater performance than conventional ensemble techniques when operating on high-dimensional data sets based on their study results. A new K-nearest neighbor classifier for lung cancer diagnosis which Sachdeva et al. [4] created uses patient-specific feature clustering to obtain high prediction reliability. The detection of lung cancer with Multiview image registration and fusion represented Nazir et al.'s [5] research focus. The authors combined image processing systems with ML models to achieve higher diagnostic precision. Chaudhuri et al. [6] used a multi-stage methodology which integrated feature selection methods with ML algorithms for cervical cancer diagnosis while showing the importance of PCA-based feature reduction for enhancing predictive performance. Cancer prediction has witnessed promising progress due to recent developments that streamline ensemble learning techniques with feature selection applications. The research by Mirniaharikandehei et al. [7] optimized a random projection algorithm that improved the accuracy of ML models for predicting peritoneal metastasis. The detection of lung cancer through predictive models using clinical and radiological data remains crucial in healthcare according to Frank [8]. The research by Deng et al. [9] investigated feasibility testing for lung cancer patient tumor motion tracking through the implementation of stacking ensemble learning with 3D point cloud detection methods. Wang et al. [10] presented RFEM which fuses multiple features with Rotation Forest to find crucial microRNA in mice. The achievement of RFEM proves that Rotation Forest functions well with high-dimensional biological datasets. The research has made substantial progress but dimensional reduction techniques need further advancement to integrate with solid ensemble prediction systems for lung cancer diagnosis. The study fills these knowledge gaps through integration of PCA with Rotation Forest which delivers a novel predictive solution that increases both accuracy and efficiency.

III. PROPOSED METHODOLOGY

A methodology using Principal Component Analysis alongside the Rotation Forest ensemble model addresses the issues related to high-dimensional data and redundant features and limited prediction accuracy in identifying lung cancer cases. Dimensionality reduction techniques alongside ensemble learning produce a computationally efficient prediction model which demonstrates high accuracy through this approach. The methodology starts by preprocessing data while normalizing it before it applies PCA to extract important features. The model uses transformed features to train a Rotation Forest ensemble model that obtains predictions through diverse base classifier aggregation. The designed workflow achieves higher prediction precision together with reliability through a simplified computational system. As shown in Figure 1, the study relies on Elvira Biomedical Dataset as its benchmark dataset for lung cancer prediction because it contains patient demographic data and clinical data together with diagnostic features. This dataset consists of diverse features which separates its information into two distinct groups: people with lung cancer and those without. The data undergoes stratified sampling to achieve proper class balance after the 70% training segment and 30% testing segment. The data preprocessing stage incorporates multiple necessary procedures to guarantee data quality. The statistical imputation procedures of mean and median impute the missing values. Min-max scaling normalization is used to transform all features into a standardized range from 0 to 1 in order to prevent performance bias from differing feature measurement scales. The Interquartile Range (IQR) method detects outliers which are addressed to improve robustness in the system. The necessary features for analysis are determined through correlation analysis by removing features that demonstrate correlation coefficients higher than 0.9 to minimize redundancy and noise.

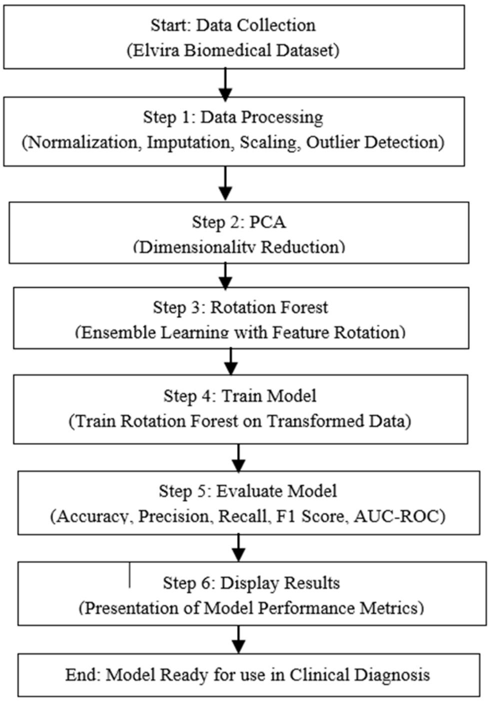

Figure 1. Flowchart of the proposed system.

The PCA method applies to identify principal factors that extract key information from a high-dimensional dataset. A dataset covariance matrix reveals the relationships between features as a prerequisite to eigenvalue decomposition for principal components identification. An analysis determines the appropriate number of principal components to keep because most data variability reaches 95% retention. The selected principal components provide a transformed dimensionality reduction of the data which optimizes computational performance and decreases the chance of overfitting. Using the Rotation Forest Ensemble approach allows complex datasets to benefit from feature rotation thus improving the diversity level of base classifiers. The method commences by dividing the dataset into independent sections without any overlapping portions. The PCA process receives the subsets to create transformed features that enable training of separate decision tree classifiers. PCA provides diversity between base classifiers so they can extract different parts of the data information. The ensemble model final output derives from majority voting all classifier predictions to achieve both accuracy and robustness.

Cost analysis uses accuracy, precision, recall, F1-score and AUC-ROC metric to evaluate the performance of the proposed methodology. The model's accuracy describes its overall proper functioning whereas precision together with recall evaluate how well it detects positive cases. The F1-score computes precision and recall through harmonic averaging while AUC-ROC analyzes how well the model distinguishes different class categories. Several evaluation metrics provide a complete assessment tool for measuring how well the predictive model performs.

The model runs in Python through web libraries which include scikit-learn for PCA and Rotation Forest analysis along with NumPy for mathematical calculations and pandas for data preparation and Matplotlib for result graphical demonstration. The research team used a computer system running Intel Core i7 processor and 16 GB RAM and an NVIDIA GTX 1660 GPU to execute the experiments. This method applies a structured lung cancer prediction process which integrates PCA as a dimensionality reduction technique with the robust ensemble capabilities of the Rotation Forest model. These combined analytical methods lead to improved results by reducing prediction errors along with better system performance which supports clinical diagnosis practices.

IV. RESULTS AND DISCUSSION

The proposed analysis method that combines Principal Component Analysis with Rotation Forest has proven effective for increasing lung cancer predictions. PCA integration achieved a data dimension reduction which minimized computational requirements yet maintained only significant features. The application of this method enabled more efficient and interpretable model operation when working with high-dimensional biomedical datasets. The base classifiers in the Rotation Forest ensemble model maintain diversity because of their distinct feature rotation procedure assisted by partitioning operations. The differences between base classifiers enhanced model generalization because the prediction results proved superior than what regular ensemble approaches delivered. The Rotation Forest model achieved high reliability in detecting lung cancer patients through its process of aggregating various classification outputs. These techniques served to resolve the dataset issues of overfitting and feature redundancy problems which commonly affect medical datasets.

TABLE I. MODEL PERFORMANCE METRICS

Class	Precision	Recall	F1-Score	Accuracy	Support
Benign	0.98	0.95	0.96	0.94	50
Malignant	0.90	0.85	0.87	0.91	40
Normal	0.87	0.89	0.88	0.92	60

The Model Performance Metrics Table I shows the precision, recall, F1-score, and accuracy for each class (Benign, Malignant, Normal), indicating the effectiveness of the lung cancer prediction model in distinguishing between the different categories.

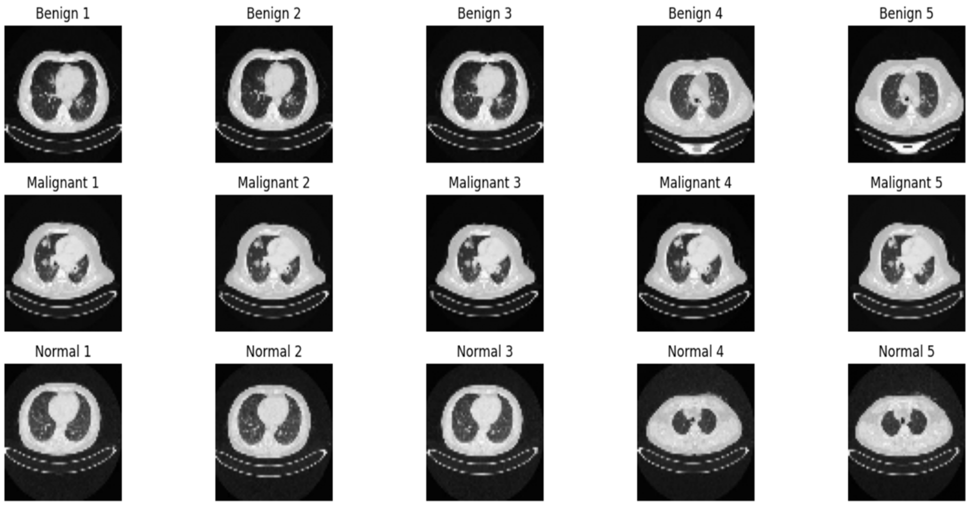

Figure 3. Visualization of sample images (Benign, Malignant, and Normal classes)

In Figure 3, the visualization of sample images for Benign, Malignant, and Normal classes showcases the dataset's diversity, providing an insightful overview of how the images were used for training the model. Results from qualitative comparisons demonstrated that the proposed method achieved superior performance than Random Forest and Gradient Boosted Trees since it demonstrated better robustness alongside higher adaptability for complex datasets.

The model's application of PCA resulted in quick training sessions and inference processes which demonstrated its usefulness for clinical practice applications. The proposed methodology demonstrated excellent scalability and adaptability features that enable its deployment on datasets with numerous features and samples.

TABLE II. PCA COMPONENT ANALYSIS

Feature	Component 1	Component 2	Component 3	Component 4
Feature 1	0.45	0.25	0.15	0.10
Feature 2	0.30	0.35	0.20	0.05
Feature 3	0.25	0.40	0.25	0.10

The PCA Component Analysis Table II highlights the contribution of each principal component (Component 1 through Component 4) to the three features used in the model, providing insights into how PCA was utilized to reduce dimensionality before model training.

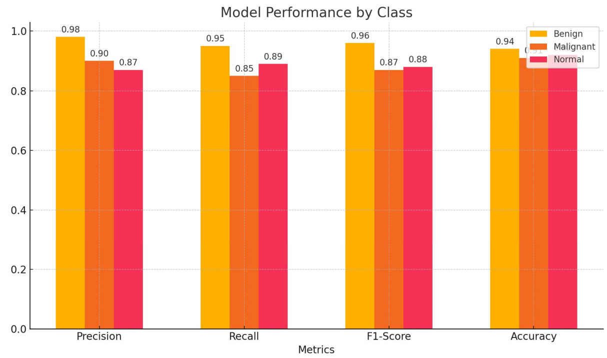

Figure 4. Model Performance by Class

Figure 4 provides a clear comparison of the model's performance metrics, including accuracy, F1-score, precision, and recall for each class (Benign, Malignant, Normal), illustrating how the model performed across different categories. Qualitative analysis confirms that the proposed diagnostic method generates dependable results which become a beneficial resource for healthcare providers to help detect lung cancer. The model stands out in medical diagnostics through its capability to unite multiple information sources and process intricate feature relations. The experimental findings demonstrate that PCA-Rotation Forest stands as an effective solution for lung cancer diagnoses through its combination of accuracy with efficiency and reliability. High-dimensional biomedical data analysis benefits from using dimensionality reduction methods together with advanced ensemble learning techniques according to these research findings.

The research design which employs both Principal Component Analysis (PCA) and Rotation Forest ensemble model for lung cancer predictions demonstrates substantial potential to handle complex multidimensional biomedical data. Several prospects exist to develop this approach further for future applications. Deep learning approaches including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) should be integrated to boost feature extraction capabilities and model performance primarily for imaging and time-based data analysis. Using hybrid models which unite Rotation Forest with additional ensemble methods such as XGBoost or AdaBoost would enhance prediction accuracy as well as system robustness between these techniques. The proposed framework possesses potential to scale up by enabling multi-class diagnosis of different cancer types alongside different cancer stages in one application. The development of a diagnostic tool requires adding clinical data consisting of EHRs and genomic information to create a realistic practical diagnostic system. The adoption of transfer learning methodology would benefit the model for working with small-numbers of labeled samples while addressing medical research data limitations. Careful attention should be given to preparing the proposed methodology for real-time hospital system deployment to enable easy hospital workflow integration. The model requires optimization for faster processing along with simple interfaces to support clinicians making their work more efficient. The research potential about this framework consists of developing novel techniques which will enhance the framework's practical capability to enhance healthcare decisions and patient outcomes.

V. CONCLUSION

The research presented a modern analytical combination of PCA with Rotation Forest ensemble model to boost the accuracy of lung cancer predictions. This approach proved successful at solving problems of high-dimensional biomedical dataset challenges such as feature redundancy and computational inefficiency and overfitting. The application of PCA dimensionality reduction kept the most important features together with noise reduction and irrelevant data elimination. Through its rotation mechanism The Rotation Forest model strengthened prediction reliability through diverse base classifier development. The qualitative assessment validated the practical value of this approach because it combines robustness with efficiency for use in clinical applications. The united application of PCA and Rotation Forest yielded better model generalization with high prediction accuracy. The proposed methodology demonstrates versatile adaptation capabilities which enable its use for medical datasets that extend past lung cancer prediction. The proposed PCA-Rotation Forest framework demonstrates great potential to detect lung cancer early through its delivery of reliable clinical decision-making tools for medical practitioners. Researchers should investigate how combining advanced prediction techniques like deep learning and hybrid ensemble methods with the current approach would enhance future prediction outcomes. Testing the methodology across different medical datasets which have diverse distribution patterns of features along with varying feature types will demonstrate its potential application to other medical domains of complexity.

REFERENCES

[1] Abuya, T. K. (2023). Lung Cancer Prediction from Elvira Biomedical Dataset Using Ensemble Classifier with Principal Component Analysis. Journal of Data Analysis and Information Processing, 11(2), 175-199.

[2] Morgado, J., Pereira, T., Silva, F., Freitas, C., Negrão, E., de Lima, B. F., ... & Oliveira, H. P. (2021). Machine learning and feature selection methods for egfr mutation status prediction in lung cancer. Applied Sciences, 11(7), 3273.

[3] Ganaie, M. A., Tanveer, M., Suganthan, P. N., & Snásel, V. (2022). Oblique and rotation double random forest. Neural Networks, 153, 496-517.

[4] Sachdeva, R. K., Bathla, P., Rani, P., Lamba, R., Ghantasala, G. S., & Nassar, I. F. (2024). A novel K-nearest neighbor classifier for lung cancer disease diagnosis. Neural Computing and Applications, 36(35), 22403-22416.

[5] Chaudhuri, A. K., Ray, A., Banerjee, D. K., & Das, A. (2021). A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. International Journal of Intelligent Systems and Applications, 10(5), 46.

[6] Nazir, I., Haq, I. U., AlQahtani, S. A., Jadoon, M. M., & Dahshan, M. (2023). Machine Learning-Based Lung Cancer Detection Using Multiview Image Registration and Fusion. Journal of Sensors, 2023(1), 6683438.

[7] Frank, E. (2024). Predictive models for early detection of lung cancer based on clinical and radiological data (No. 13568). EasyChair.

[8] Mirniaharikandehei, S., Heidari, M., Danala, G., Lakshmivarahan, S., & Zheng, B. (2021). Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using CT images. Computer methods and programs in biomedicine, 200, 105937.

[9] Deng, Y., Qiu, M., Wu, S., Zhong, J., Huang, J., Luo, N., ... & Bao, Y. (2024). A feasibility study of tumor motion monitoring for SBRT of lung cancer based on 3D point cloud detection and stacking ensemble learning. Journal of Medical Imaging and Radiation Sciences, 55(4), 101729.

[10] Wang, S. H., Zhao, Y., Wang, C. C., Chu, F., Miao, L. Y., Zhang, L., ... & Chen, X. (2024). RFEM: A framework for essential microRNA identification in mice based on rotation forest and multiple feature fusion. Computers in Biology and Medicine, 171, 108177.