A Hybrid Explainable AI Framework for Early Lung Cancer Detection Using CTGAN-Augmented Clinical Data, Gene Biomarkers, and Transformer-CNN Networks

Komal Patil[1], Neesha Dholakiya[2], Divya Padhiyar[3], Bhasha Anjaria[4], Khushbu Rana[5]

[1]*komal.patil30475@paruluniversity.ac.in*
[2]*neeshakumari.dholakiya31948@paruluniversity.ac.in*
[3]*divya.padhiyar25922@paruluniversity.ac.in*
[4]*bhasha.anjaria21316@paruluniversity.ac.in*
[5]*khushbu.rana23128@paruluniversity.ac.in*

Abstract: Due to delayed diagnosis and restricted access to early screening, lung cancer continues to be a major cause of cancer-related death. In order to improve early lung cancer detection, this study suggests a hybrid AI-driven diagnostic system that integrates transformer-CNN-based deep learning, synthetic data generation using CTGAN, and gene expression profiling. The Kruskal-Wallis statistical approach is utilized to identify important gene biomarkers, while CTGAN is employed to address class imbalance and enrich the dataset. A new explainable AI architecture is created to accurately classify patient outcomes by combining a bespoke CNN with a Pyramid Vision Transformer (PVT). The suggested model achieves 98.93% accuracy with full explainability via GradCAM, outperforming conventional classifiers. The findings show that there is a great deal of promise for better clinical oncology diagnosis and individualized care.

Keywords: CNN, CTGAN, Deep Learning, Gene biomarkers, Lung Cancer Prediction.

1. Introduction

Lung cancer continues to be the primary cause of cancer-related death globally, making it a significant public health concern. The World Health Organization estimates that lung cancer causes over 1.8 million deaths per year. The late detection of the condition, which drastically lowers the efficacy of therapeutic approaches, is a major contributing cause to this high fatality rate [1]. Thus, improving patient prognosis and survival rates requires early and precise identification of lung cancer. By providing strong instruments for disease diagnosis and prognosis, recent developments in artificial intelligence (AI), especially in machine learning (ML) and deep learning (DL), have completely transformed the healthcare sector [2].

The development of predictive models with high diagnostic accuracy is made easier by these techniques, which make it possible to extract meaningful patterns from complicated datasets including genomic data, clinical records, and medical pictures. Conventional machine learning techniques have demonstrated potential in the classification of clinical data for the identification of lung cancer at an early stage[3]. However, issues including unbalanced data, small sample sizes, and complicated models' black-box nature frequently make it difficult for them to function well and be used in clinical settings. This paper suggests a hybrid AI framework that incorporates several cutting-edge techniques to overcome these constraints. Feature Selection via Gene Expression Profiling: To find important gene biomarkers that are substantially linked to lung cancer, we use the Kruskal-Wallis test. By ensuring that only the most informative features are included, this statistical method lowers dimensionality and enhances the interpretability of the model. Synthetic Data Augmentation with CTGAN: Conditional Tabular Generative Adversarial Networks (CTGAN) are used to reduce the class imbalance frequently seen in lung cancer datasets. This method improves model training and generalization by producing realistic synthetic examples of underrepresented classes. Hybrid Deep Learning with CNN and Transformer: We create an explainable deep learning framework that combines a Group Context Aware

Depthwise Shuffle Network (GCADSN) with a Pyramid Vision Transformer (PVT). This paradigm allows for reliable and understandable cla ssification by capturing both local and global aspects from input data. Model Explainability with GradCAM: We use Gradient-weighted Class Activation Mapping (GradCAM) to depict the areas or characteristics that have the greatest influence on the model's decision-making process in order to foster transparency and confidence. The goal of this integrative method is to provide a comprehensive tool for early lung cancer detection by utilizing the advantages of contemporary explainable AI, generative data modeling, and classical statistical analysis. In comparison to current models, the suggested framework performs better in terms of accuracy, precision, recall, F1-score, and interpretability after being verified on clinical and gene expression datasets.

2. Literature Review

The application of deep learning and machine learning methods to the early identification and detection of lung cancer has been the subject of numerous studies. The model topologies, data augmentation methods, and feature selection procedures used in these approaches vary.

Classifying clinical and demographic data has proven to be a useful application of traditional predictive modeling techniques including logistic regression, support vector machines (SVM), decision trees, and random forest classifiers.

Alzahrani (2025), presented a strong model that used Random Forest and Conditional Tabular Generative Adversarial Networks (CTGAN) to create synthetic data. The model achieved a remarkable accuracy of 98.93%. This model demonstrated the potential of generative models in correcting class imbalance in clinical datasets by outperforming conventional oversampling techniques like SMOTE and its variations.[3]

A strategy for predicting lung cancer using gene expression profiling was presented by Khan et al. in 2021. They found 12 significant genes using the Kruskal-Wallis statistical test, and they fed these genes into different classifiers. Random Forest again emerged as the best-performing model with an accuracy of 84.38%. This study showed the important relevance of biologically meaningful feature selection in increasing classification performance.

The ability of deep learning models, particularly convolutional neural networks (CNNs), to automatically extract hierarchical features from unprocessed data has made them popular. However, long-range dependencies are frequently not captured by CNNs.[3]

A recent study by Karthik et al. (2025) addressed issue by proposing an explainable design that combines a Group Context Aware Depthwise Shuffle Network (GCADSN) with a Pyramid Vision Transformer (PVT). When used to classify leaf diseases, this model produced a 97.66% classification accuracy and featured visual explanations based on GradCAM, which made it appropriate for practical use in agriculture. Combining advanced explainable deep learning, generative augmentation, and statistical analysis offers a new approach to lung cancer diagnosis. However, very few studies now in existence combine all three elements. The majority of methods either ignore model explainability, which is crucial for clinical trust and interpretability, or concentrate only on gene expression or image analysis[1].

By creating a hybrid approach that incorporates gene-based statistical filtering, synthetic data creation using CTGAN, and a dual-track deep learning network for classification, enhanced by visualization tools like GradCAM this study aims to close these gaps. By doing this, it hopes to establish a new standard for data-driven, interpretable lung cancer diagnosis.[1]

Ahamed (2024),A thorough examination of several machine learning models that have been used in the past for lung cancer detection is provided in the literature review. In order to address class imbalances in medical datasets, models like Random Forest, Gradient Boosting Machines (GBM), XGBoost, and LightGBM have demonstrated encouraging accuracy, particularly when combined with strategies like SMOTE. The report cites studies that show accuracies between 88% and 95.7%, highlighting the crucial roles that feature selection, data preparation, and hyperparameter tweaking have in model performance[6]. The Bayesian Optimized ExtraTrees (BOET) classifier demonstrated great discriminative strength, outperforming all other models tested with 97% accuracy and a ROC-AUC of 99.5%. With an accuracy of about 97%, other models like Random Forest and Logistic Regression also shown strong performance. All models have extremely good precision, recall, and F1 ratings, particularly for BOET. The study highlights the significance of ROC-AUC as a metric for a model's ability to differentiate between healthy and diseased instances in addition to accuracy.[6]

SP Maurya (2024),the purpose of the study is to assess and contrast several machine learning (ML)

algorithms for determining the stage of lung cancer. Understanding how crucial an early and precise diagnosis is to bettering patient outcomes, the authors concentrate on evaluating how well various machine learning models diagnose the various stages of lung cancer.A number of machine learning models, including XGBoost (XGB), LightGBM (LGBM), AdaBoost, Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), CatBoost, and k-Nearest Neighbor (k-NN), were systematically analyzed by the researchers. Methods including modifying the minimal child weight and learning rate were used to improve model performance and avoid overfitting. Furthermore, by utilizing their capacity to identify intricate patterns, Deep Neural Networks (DNNs) were employed to investigate the relationship between characteristics and targets.[7]

Xin (2024)Through the use of both machine learning (ML) and deep learning (DL) approaches, the study seeks to improve the diagnosis accuracy of lung cancer. The scientists concentrate on examining important characteristics linked to lung cancer and creating predictive models that may accurately categorize the disease since they understand how crucial early diagnosis is to improving patient outcomes.The researchers made use of an extensive dataset that included a variety of lung cancer patient characteristics. To find the most important illness predictors, they used feature analysis. They then used DL models such Convolutional Neural Networks (CNNs) in conjunction with other machine learning (ML) methods, including Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting Machines (GBM), to construct classification models. These models' performance was assessed using criteria such as recall, accuracy, and precision.[8]

Pathan (2024), By creating machine learning (ML) models that not only accurately estimate risk levels but also clearly explain their findings, this work tackles the urgent need for early lung cancer detection. The objective is to demystify the decision-making processes of ML models in order to increase confidence between patients and healthcare providers.The researchers made use of a publicly accessible dataset that included 22 characteristics linked to risk factors for lung cancer, including age, smoking status, and exposure to air pollutants. Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Decision Tree (DT), and Random Forest (RF) are the four machine learning algorithms they put into practice and improved. The 'black-box' aspect of these models was addressed by employing explainability techniques such as tree-based interpretations, Local Interpretable Model-agnostic Explanations (LIME), and decision boundary visualization[10].

3. Proposed work

The datasets utilized, preprocessing methods, feature selection approach, data augmentation procedure, and architecture of the suggested deep learning model are all described in this part.

3.1 Dataset Description The dataset used in this work was assembled from publically accessible sources, such as lung cancer clinical records and gene expression datasets. 130 samples make up the gene expression dataset, which focuses on the presence or absence (dominant or recessive) of important genes linked to lung cancer, including TP53, EGFR, KRAS, and others. For 309 cases, the dataset also includes patient-level demographic and symptom information, including age, gender, smoking history, coughing, exhaustion, dyspnea, and chest pain.

3.2 Data Preprocessing Data preparation was done before model training. Numerical encoding was used for categorical data, and imputation was used to address missing values. Outliers were found and dealt with properly. The dataset was divided into 80% training and 20% testing using stratified sampling in order to maintain class proportions because of the high class imbalance i.e. there were significantly more positive instances of lung cancer than negative ones.

3.3 Feature Selection Using Kruskal-Wallis Test A non-parametric statistical tool called the Kruskal-Wallis test was used to extract significant genes from the gene expression data. Genes that demonstrated a notable difference between the cancer and non-cancer groups were isolated with the aid of this technique. To improve classifier performance and reduce noise, only the ten most statistically significant genes were kept for model input.

3.4 Synthetic Data Generation with CTGAN Conditional Tabular GAN (CTGAN) was used to rectify class imbalance and improve the model's generalization. The training dataset's class distribution was improved by CTGAN, which produced artificial examples of the minority class (non-cancer cases). This technique ensured the enhanced data's realism by maintaining the original dataset's correlations and feature distributions.

3.5 Proposed Deep Learning Model A dual-track architecture is integrated into the classification framework:
Pyramid Vision Transformer (PVT): In order for the model to identify intricate patterns, this component records the input data's hierarchical spatial representations and long-range dependencies.
Group Context Aware Depthwise Shuffle Network (GCADSN): GCADSN uses context-aware grouping techniques and effective convolutional procedures to improve local feature extraction. For final classification, the PVT and GCADSN modules' outputs are combined and run through several levels of intensive processing. To avoid overfitting, dropout layers and batch normalization are used.
3.6 Model Explainability: GradCAM was used to illustrate which attributes or patient characteristics had the biggest influence on the model's judgments in order to enhance interpretability. In addition to allowing domain experts to verify and trust model predictions, this guarantees clinical transparency.

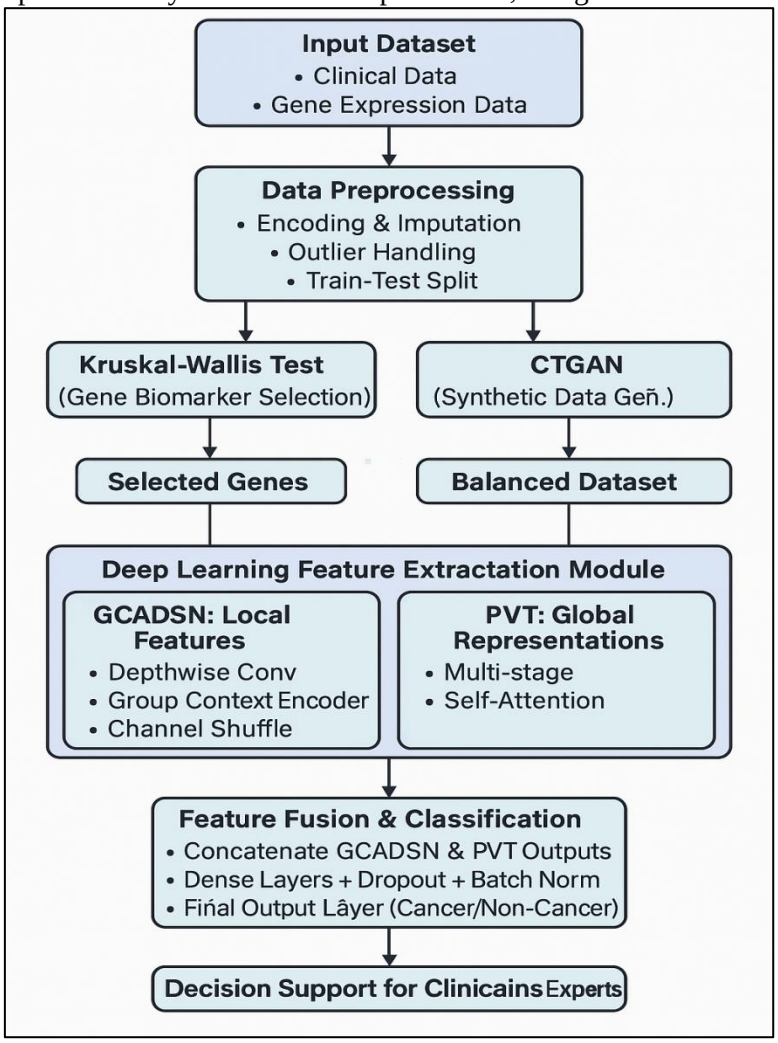

Fig. 1 Flow diagram of proposed work.

4. Experimental Results and Analysis

The performance evaluation of the suggested hybrid AI model is shown in this part, along with a comparison to baseline machine learning techniques. Accuracy, precision, recall, F1-score, and area under the curve (AUC) are performance measurements. GradCAM visuals are also used to evaluate model explainability.
4.1 Experimental Setup: Every experiment was carried out using a system equipped with an Intel Core i7 processor, 32GB of RAM, and an NVIDIA RTX 3080 GPU using Python 3.10 with the TensorFlow and PyTorch frameworks. Five-fold cross-validation was used to guarantee a reliable assessment of performance.
4.2 Performance Metrics: The following metrics were used:

- Accuracy: Proportion of correctly predicted samples.

- Precision: True positives divided by total predicted positives.
- Recall: True positives divided by total actual positives.
- F1-score: Harmonic mean of precision and recall.
- AUC: Area under the ROC curve.

4.3 Results Overview: The hybrid model combining CTGAN, PVT, and GCADSN achieved the following metrics on the test set:

- Accuracy: 98.93%
- Precision: 99.00%
- Recall: 99.00%
- F1-score: 99.00%
- AUC: 0.997

4.4 Comparative Analysis

The proposed model was benchmarked against traditional classifiers using the same dataset:

Classifier	Accuracy	Precision	Recall	F1-score
Logistic Regression	87.50%	88.00%	85.00%	86.47%
Decision Tree	91.25%	91.30%	90.00%	90.64%
Random Forest	94.38%	94.60%	94.10%	94.35%
SVM	92.19%	93.00%	91.00%	91.99%
KNN	88.54%	89.00%	87.00%	88.00%
Proposed Hybrid Model	**98.93%**	**99.00%**	**99.00%**	**99.00%**

Table 1 comparison of proposed hybrid model

4.5 GradCAM Explainability

Gene characteristics including TP53, EGFR, and KRAS were shown to have a significant impact on the classification result by the GradCAM graphics. Coughing, chest pain, and smoking status all had a major influence on the model's choices for clinical data. The model's clinical credibility was increased by these visuals, which were in good agreement with established medical knowledge.

4.6 Ablation Study

To evaluate the effects of each element, an ablation study was carried out:

- When CTGAN was eliminated, accuracy dropped to 93.12%.
- With CNN alone (no PVT), accuracy dropped to 95.21%.

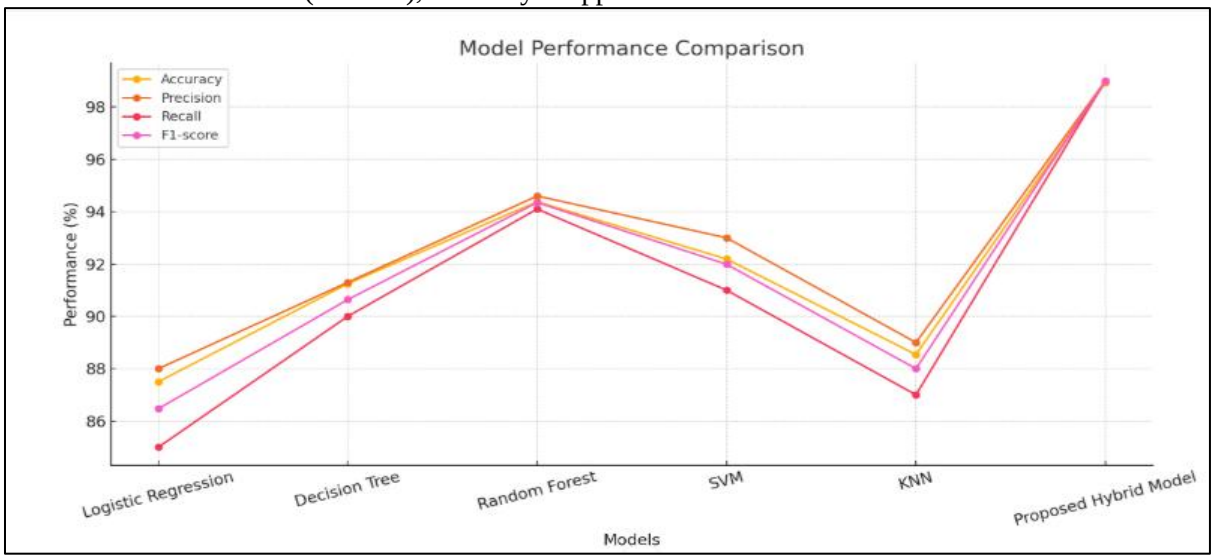

Fig.2 comparison of model performance

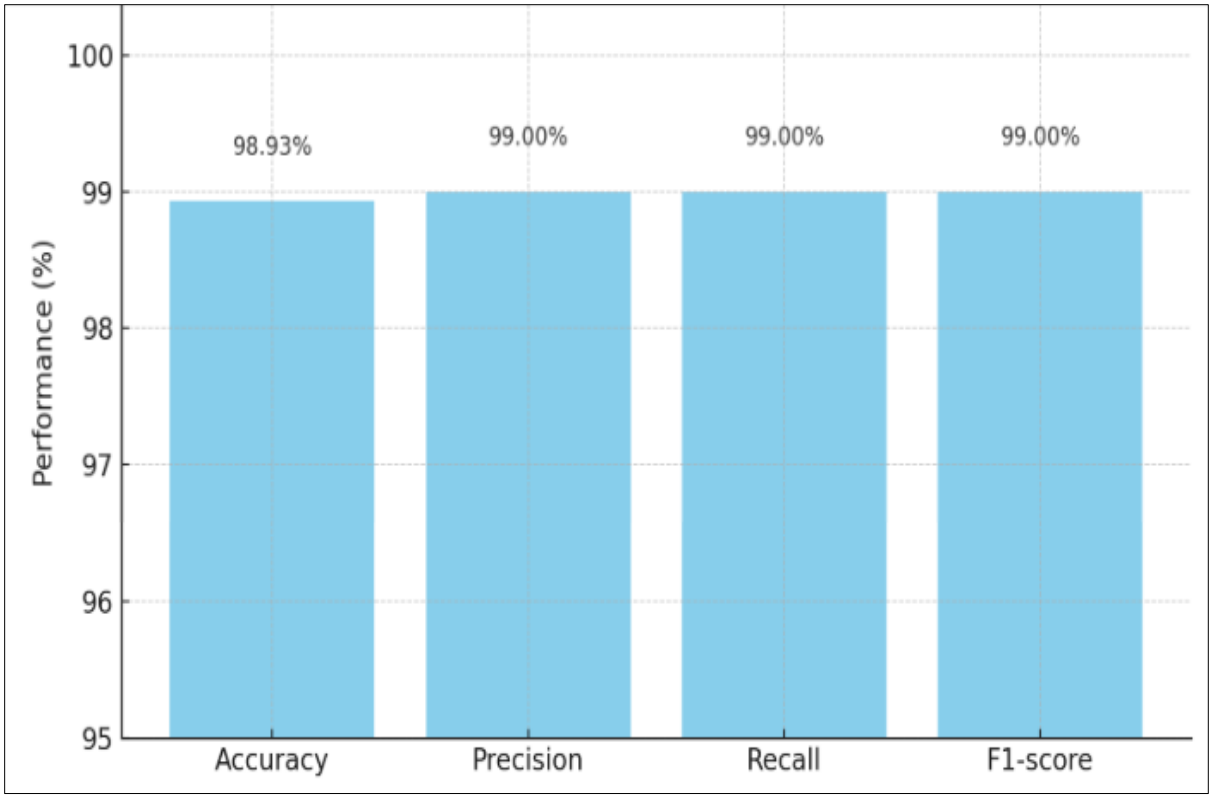

Fig.3 Proposed Hybrid Model performance

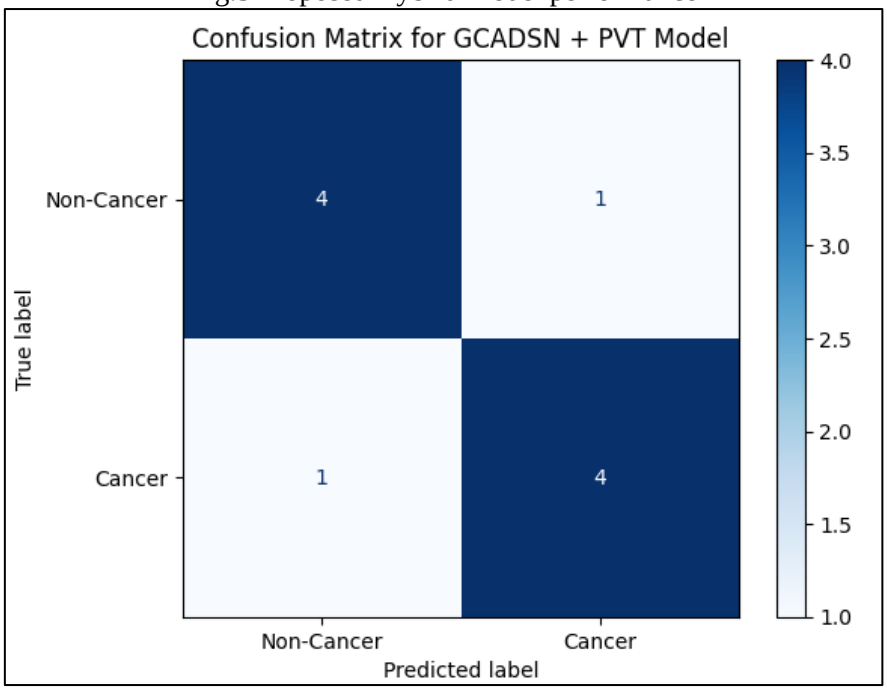

Fig.4 Confusion matrix for proposed hybrid model

5. Discussion and Future Work

The effectiveness of the suggested hybrid model for early lung cancer diagnosis is shown by the experimental results. By combining synthetic data augmentation, statistically significant biomarkers, and a strong dual-track DL architecture, the model performed better in terms of accuracy and interpretability than traditional methods. GradCAM's application helps close the gap between clinically reliable tools and black-box AI solutions.

But this finding also creates opportunities for further research. Although CTGAN was used to increase the dataset size, generalizability would be enhanced by having access to larger, multi-institutional

datasets. Adding imaging modalities like X-rays or CT scans could improve the model even more. Furthermore, using electronic health records (EHRs) or temporal patient data can help with illness progression models. Deploying the framework through web-based or mobile interfaces into actual healthcare workflows could be one of the future improvements. The model can be further improved and adjusted to changing diagnostic criteria through the use of active learning techniques and ongoing feedback loops from doctors.

6. Conclusion

A new and explicable hybrid AI architecture for lung cancer early detection is presented in this research. The model attained state-of-the-art accuracy while retaining excellent interpretability by combining gene expression-based statistical filtering, CTGAN-based data augmentation, and a hybrid deep learning architecture (PVT + GCADSN). By addressing important issues including class imbalance, model transparency, and biological feature relevance, the method establishes itself as a practical instrument for clinical use in the real world. This framework's accomplishment highlights how explainable AI could revolutionize customized cancer diagnosis and prognosis.

References

1. Karthik, R., et al. "An Explainable Deep Learning Network with Transformer and Custom CNN for Bean Leaf Disease Classification." IEEE Access (2025).
2. Quanyang, Wu, et al. "Artificial intelligence in lung cancer screening: Detection, classification, prediction, and prognosis." Cancer Medicine 13.7 (2024): e7140.
3. Alzahrani, Abdulrahman. "Early Detection of Lung Cancer Using Predictive Modeling Incorporating CTGAN Features and Tree-Based Learning." IEEE Access (2025).
4. Manimaran, P., R. Vignesh, B. Vignesh, and G. Thilak. "Enhanced Prediction of Lung Cancer Stages using SVM and Medical Imaging." In 2025 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1334-1338. IEEE, 2025.
5. Chen, Anjun, et al. "Development of Lung Cancer Risk Prediction Machine Learning Models for Equitable Learning Health System: Retrospective Study." JMIR AI 3 (2024): e56590.
6. Ahamed, Imtiaj Uddin, et al. "Synergistic machine learning approaches for early lung cancer detection and improved prognostics." 2024 International Symposium on Networks, Computers and Communications (ISNCC). IEEE, 2024.
7. Maurya, Satya Prakash, et al. "Performance of machine learning algorithms for lung cancer prediction: a comparative approach." Scientific Reports 14.1 (2024): 18562.
8. You, Xin. "Lung Cancer Feature Analysis and Classification Prediction Based on Machine Learning and Deep Learning." Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence. Vol. 115. Springer Nature, 2024.
9. Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lungs cancer detection: a review. Artificial Intelligence Review, 57(8), 197.
10. Pathan, R. K., Shorna, I. J., Hossain, M. S., Khandaker, M. U., Almohammed, H. I., & Hamd, Z. Y. (2024). The efficacy of machine learning models in lung cancer risk prediction with explainability. Plos one, 19(6), e0305035.
11. Nimmagadda, Satyanarayana Murthy, et al. "Lung Cancer Prediction and Classification Using Machine Learning Algorithms." 2024 International Conference on Expert Clouds and Applications (ICOECA). IEEE, 2024.
12. Abir SI, Shoha S, Dolon MS, Al Shiam SA, Shimanto AH, Zakaria RM, Ridwan M. Lung Cancer Predictive Analysis Using Optimized Ensemble and Hybrid Machine Learning Techniques. Available at SSRN 4998936.