STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–18.
Published online in International Academic Press (www.IAPress.org)

Towards Transparent AI for Lung Cancer Diagnosis: A Dual-Pipeline Explainable Framework Using Clinical and CT Imaging Data

Nouran M. Elmasry[1,2*], Assma H. ElSayed[1], Faten A. Khalifa[1]

[1] *Faculty of Computers and Information, Menofia University, Menofia, Egypt*
[2] *Badya University in Cairo, Cairo, Egypt*

Abstract In recent years, artificial intelligence (AI) has shown promising performance in medical diagnosis; however, its clinical adoption remains limited due to a lack of interpretability. In this study, we propose a dual-pipeline explainable framework for lung cancer diagnosis using two independent data modalities: structured clinical data and CT imaging data. For the clinical data, several machine learning models were compared, such as LightGBM, CatBoost, XGBoost, Random Forest, Logistic Regression, K-Nearest Neighbors, and Naïve Bayes. For the imaging data, deep learning models such as VGG16, ResNet50, InceptionV3, MobileNetV2, DenseNet121, and Xception were compared using the IQ-OTH/NCCD dataset. To ensure the reliability of the validation results, a strict patient-level split was used to avoid data leakage. The experimental results showed that LightGBM obtained the best results on the clinical data, achieving an accuracy of 98.39% and an ROC-AUC of 0.99. In the imaging data, MobileNetV2 obtained the best results, achieving an accuracy of 0.97, which is highly computationally efficient. To improve the interpretability of the models, SHAP and LIME were used to analyze the clinical feature importance, while Grad-CAM was used to analyze the discriminative regions in the CT image. The reliability of the explanations was also verified through stability analysis with Spearman rank correlation, agreement analysis with SHAP and LIME, as well as through verification with expert clinicians for the Grad-CAM visualizations. The results demonstrate that various XAI methods provide complementary insights, supporting the creation of transparent, reliable, and meaningful AI systems for lung cancer diagnosis.

Keywords Explainable AI (XAI); Machine Learning; Deep Learning; Transparency; Lung Cancer Diagnosis

AMS 2010 subject classifications 68T07, 68T09, 92C50

DOI: 10.19139/soic-2310-5070-3399

1. Introduction

Healthy living is a holistic state of physical, mental, and social wellness, recognized as a fundamental human right and a cornerstone of sustainable societal progress. Despite medical advancements, lung cancer remains one of the most significant global health challenges, characterized by high malignancy and mortality rates [1, 2]. Early diagnosis is paramount, as clinical outcomes are directly correlated with the timeliness of intervention [3]. Consequently, there is an urgent demand for intelligent, transparent computer-aided diagnostic (CAD) systems that can facilitate early detection with high reliability. While Artificial Intelligence (AI) and Deep Learning (DL) have demonstrated exceptional performance in processing complex medical data [4], their adoption in clinical practice is often hindered by their nature as **black-box models** [5]. Although Explainable AI (XAI) frameworks such as SHAP, LIME, and Grad-CAM have been introduced to address transparency limitations [6, 7], there remains a critical need for systematic evaluation of their behavior across different medical data structures. To

*Correspondence to: Nouran M. Elmasry
Email: nouran.mohamed@badyauni.edu.eg. Faculty of Computer and Data Information,Badya University in Cairo, Cairo, Egypt.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

address this gap, this research proposes a **comprehensive comparative benchmarking** on the application of XAI methods on two distinct data domains: structured clinical data and unstructured CT images. Instead of pursuing an integrated multimodal approach, which often obscures the individual contributions of heterogeneous data types, this research proposes an alternative two-path approach. This approach is instrumental in assessing the consistency and faithfulness of interpretability in feature-based and saliency-based XAI methods. The main technical contributions of this study are as follows:

- A systematic benchmarking framework for assessing stability and clinical relevance of XAI methods on structured clinical and image datasets.
- Development of two independent diagnostic pipelines: one using the LightGBM model for clinical data and another using the MobileNetV2-based transfer learning model for the classification of CT images.
- A strict data splitting strategy based on the patient level to avoid data leakage.
- Integration of multiple explainability methods, such as SHAP, LIME, and Grad-CAM, to offer feature level and image level explanations.
- Quantitative evaluation of explanation reliability by performing cross-method agreement analysis between SHAP and LIME methods, as well as stability analysis using Spearman rank correlation.
- Qualitative evaluation of Grad-CAM visualization using expert clinical validation.

2. Related Work

Artificial Intelligence (AI) has significantly transformed cancer diagnosis by enhancing diagnostic precision and reducing decision time. In addition, Explainable AI (XAI) methods improve transparency and provide insights into model reasoning, which is essential for clinical adoption. Recent studies have explored both structured clinical data and unstructured imaging data for lung cancer prediction, highlighting the growing importance of model interpretability and accountability.

Liu et al. [8] presented an explainable framework for cancer diagnosis based on clinical data, wherein Random Forest was evaluated using SHAP to identify the most influential patient features. However, the study relied solely on SHAP without incorporating alternative interpretability techniques or imaging-based explanations. Similarly, Nandhini and Rajeswari [9] employed optimized XGBoost and Decision Tree models with SHAP for early lung cancer detection, uncovering key clinical risk factors. Nevertheless, their work did not explore multimodal explainability or compare different XAI methods.

Sharma and Yadav [10] focused on CNN-based imaging explainability, utilizing Grad-CAM to visualize discriminative regions in medical images. Although their approach improved transparency, the study lacked comparative analysis across multiple XAI techniques and did not evaluate explanation consistency or robustness.

Ahmed et al. [11] applied 3D Grad-CAM for tumor localization in CT images; however, their work did not incorporate clinical data or evaluate explanation consistency. Chouhan et al. [12] and Zhang et al. [13] demonstrated strong performance using Grad-CAM-based approaches. While [12] successfully highlighted tumor regions using pretrained DenseNet and ResNet models, it did not assess consistency across XAI methods or provide feature-based explanations. In contrast, [13] proposed a multimodal framework combining clinical attributes and imaging biomarkers through attention mechanisms and Grad-CAM, but lacked a detailed comparison of SHAP, LIME, and Grad-CAM within the same study.

Holzinger et al. [14] introduced the System Causability Scale (SCS) as a standardized metric for evaluating explanation quality. However, their work remained conceptual and did not include practical implementations or comparative multimodal analysis. Amann et al. [15] discussed ethical considerations of XAI in healthcare, highlighting fairness, accountability, and transparency, but did not provide empirical validation using real clinical or imaging datasets.

Wang et al. [16] proposed an explainable ensemble model for PET-CT imaging. While the model introduced a novel approach, it was limited to imaging data and did not compare different data modalities. Similarly, Park et al. [17] proposed a federated learning framework with SHAP-based explanations for biomedical data, but their work did not incorporate imaging modalities.

El-Sayed et al. (2025) [18] proposed an explainable architecture applied separately to structured and imaging data using SHAP and Grad-CAM. While their work highlighted the importance of modality-specific explanations, it lacked external validation, systematic evaluation of explanation stability, and comparative analysis between different XAI techniques within the same modality.

Recent work by Garcia et al. [19] emphasized the importance of stability in XAI explanations, demonstrating sensitivity to input variations. However, their work was limited to a single modality and did not assess cross-method agreement between SHAP and LIME. Similarly, Smith and Lee [20] (2025) introduced quantitative evaluation of saliency maps using Intersection over Union (IoU), but focused solely on imaging data and excluded clinical feature attribution.

Despite these advancements, an important research gap remains in the comparative evaluation of XAI methods across different medical data types. Most studies focus on either clinical or imaging data in isolation, with limited attention to the quantitative evaluation of explanation stability and cross-method agreement.

This study addresses these limitations by introducing a **comparative interpretability framework across independent clinical and imaging datasets**. Unlike existing approaches, this work evaluates SHAP, LIME, and Grad-CAM through both qualitative and quantitative analyses, including stability assessment, cross-method agreement, and expert-supported evaluation of clinically relevant visual explanations. By analyzing the datasets independently rather than enforcing multimodal integration, this study provides a more reliable and scientifically grounded assessment of XAI behavior across different data types, thereby enhancing the transparency and auditability of AI-based lung cancer diagnosis.

Table 1 presents a comprehensive overview of the relevant literature discussed above.

Table 1. Comparison of explainable AI approaches for lung cancer diagnosis

Ref.	Year	Data Type	Model	XAI Method	Main Contribution
[8]	2022	Clinical	RF	SHAP	Identified key clinical factors
[9]	2024	Clinical	XGBoost, DT	SHAP	Improved feature-level interpretability
[10]	2023	Imaging	CNN	Grad-CAM	Visualized tumor regions
[11]	2023	CT (3D)	3D CNN	3D Grad-CAM	Enhanced localization
[12]	2023	Imaging	DenseNet, ResNet	Grad-CAM	High accuracy classification
[13]	2024	Multimodal	Attention CNN + ML	Grad-CAM + Attention	Combined explanations
[14]	2023	Cross-domain	–	SCS	Proposed evaluation metric
[15]	2022	Conceptual	–	–	Ethical XAI discussion
[16]	2024	PET-CT	Ensemble CNN	SHAP + IG	Hybrid explanations
[17]	2025	Clinical (Fed.)	DL	SHAP	Privacy-preserving explanation
[18]	2025	Clinical + Imaging	Dual XAI	SHAP + GC	Compared modalities
[19]	2025	Clinical	LightGBM	SHAP Stability	Stability focus
[20]	2025	Imaging	ResNet/VGG	Grad-CAM + IoU	Quantified saliency

3. Materials and Methods

This section presents the methodology adopted for data preparation, preprocessing, and the development of two independent XAI-based pipelines to support lung cancer diagnosis. Within this comparative design, two independent pipelines were developed: the clinical pipeline, which deals with structured data, and the imaging pipeline, which deals with unstructured data. Each pipeline has its own training and testing session, where explainable AI technology is employed to improve the transparency of the prediction. An overview of the proposed parallel XAI architecture is illustrated in Figure 1. There are two individual datasets, which are separately processed to leverage the unique characteristics of each modality independently. Using this side-by-side analytical design, the AI technology provides both visual and quantitative outputs, including inference time analysis and feature stability assessments. The performance of explainability techniques across the two data modalities was investigated through a comparative study. SHAP and LIME methods were employed to measure how individual structural features

influenced the clinical model, providing quantitative explanations at both local and global scales. Simultaneously, Grad-CAM was utilized for the imaging model to provide localized saliency maps, offering a clear explanation of how deep features are activated by pointing to specific lung regions that contributed most to the model's output. Three key aspects are compared to provide a rigorous quantitative and qualitative evaluation: (1) degree of granularity of interpretation, (2) level of consistency between local and global hypotheses through cross-method agreement (SHAP vs. LIME), and (3) statistical stability of the explanations using Spearman Correlation analysis. By maintaining this design, we provide a robust direct comparison between feature-based interpretations (clinical) and region-wise visual explanations (imaging).

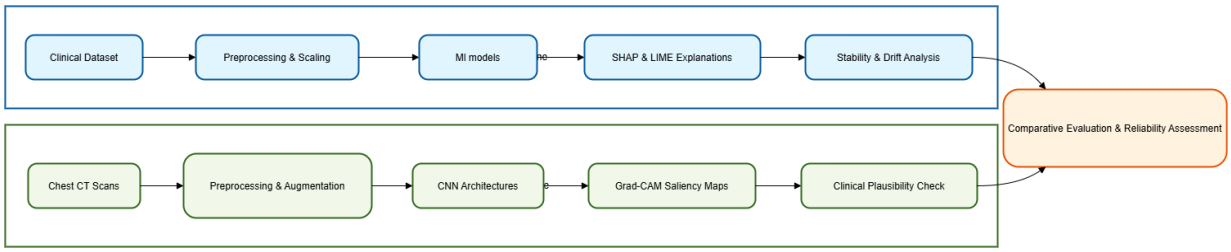


Figure 1. Parallel XAI architecture for lung cancer diagnosis featuring independent modality processing and comparative reliability evaluation.

3.1. Dataset Description and Preparation

It outlines the dataset used for this study and the preprocessing stage prior to model construction. The preprocessing steps involved manual re-annotation and data augmentation. To evaluate the proposed method's performance across different types of data, we used two separate datasets for thorough testing.

The clinical data (CSV) were retrieved from an open-source platform, the "Survey Lung Cancer" dataset. [21] It serves as the structured data source in the study. It consists of 309 patient samples with 16 attributes, including smoking, anxiety, shortness of breath, chest pain, and coughing, among others. The clinical data is imbalanced, which requires the use of data-sampling methods. The proposed model requires different preprocessing steps for several features before it can use this dataset. Each entry is labeled as "Lung Cancer" or "Non-Lung Cancer," giving rise to a binary classification task. This dataset is interpreted and structured. Based on the above, the problem can be addressed by using interpretable machine learning models. Figure 2 shows the distribution of the main numerical and categorical variables

Imaging Dataset (Chest CT) The IQ-OTH/NCCD Lung Cancer Dataset was collected from an open-source platform and is used for training and evaluating the proposed model for lung cancer identification through imaging [22, 23]. The dataset includes chest CT images of patients diagnosed with lung cancer, suspected cases, as well as normal subjects. The images are categorized into three main classes: benign, malignant, and normal. The dataset contains 1,097 images collected from 110 distinct patients, as provided in the utilized version of the dataset after preprocessing. These patients exhibit diverse clinical and demographic characteristics, including variations in age, gender, and lifestyle factors. Among these images, 40 patients (providing 561 images) are classified as malignant, 15 patients (providing 120 images) as benign, and 55 patients (providing 416 images) as normal. To ensure the clinical relevance of the findings, a radiology expert qualitatively reviewed the selected Grad-CAM visualizations. The purpose was to determine whether the highlighted regions correspond to clinically meaningful lung patterns. The dataset includes patients at various disease stages, providing a diverse representation of clinical conditions. Representative samples from each class in the IQ-OTH/NCCD imaging dataset are shown in Figure 3.

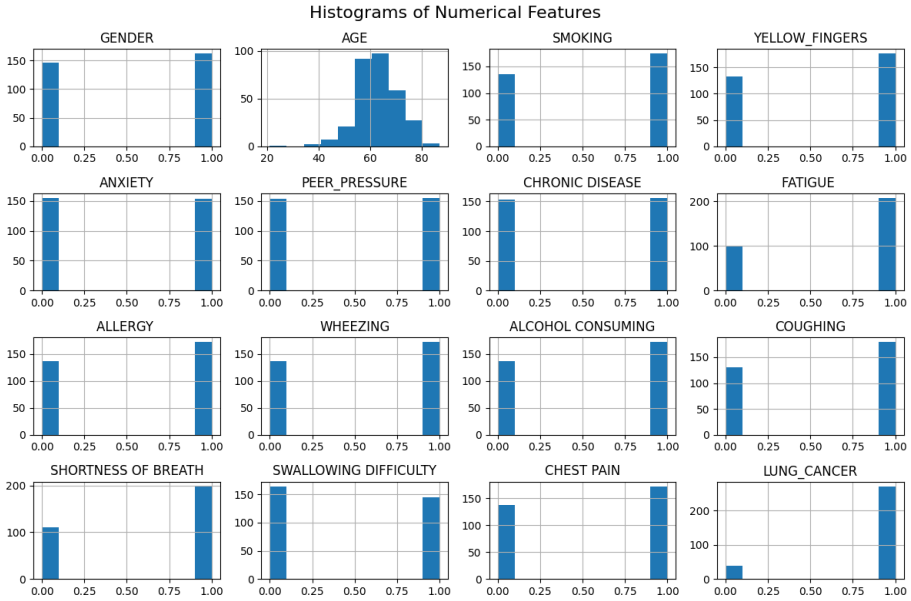
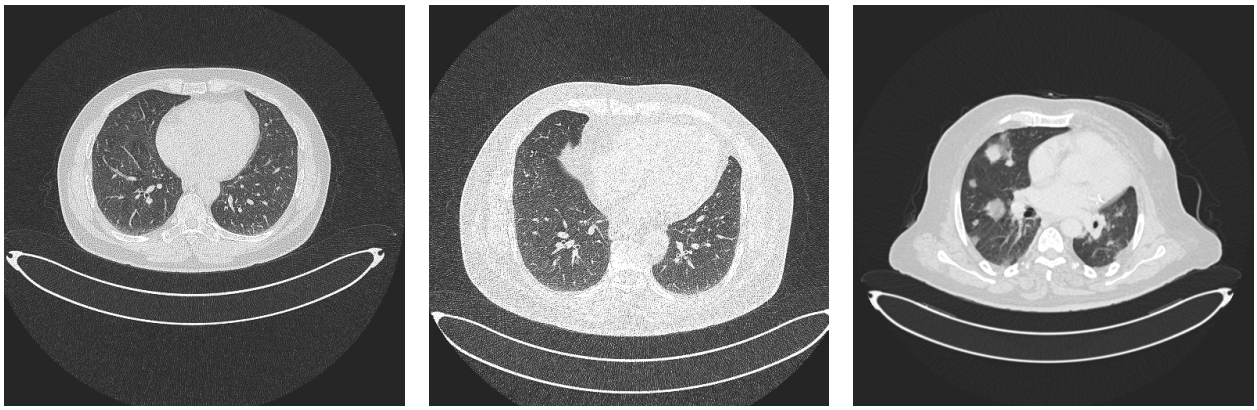


Figure 2. Illustrates the distribution of the main numerical and categorical features in the clinical dataset.



(a) Normal (b) Benign (c) Malignant

Figure 3. Representative lung CT samples from the IQ-OTH/NCCD Lung Cancer Dataset: (a) Normal, (b) Benign, and (c) Malignant.

3.2. Implementation Environment

In order to guarantee the reproducibility of the proposed framework, the implementation details of the proposed approach are given as follows to address the computational requirements of the proposed parallel-pipeline architecture:

- **Software Framework:** The proposed framework was implemented using Python 3.9 in the Google Colab Cloud Environment. The proposed clinical pipeline utilized the Scikit-Learn library along with various ensemble learning frameworks to develop several machine learning classifiers. The proposed imaging pipeline utilized TensorFlow 2.15 along with the Keras API to implement several convolutional neural network architectures. We used NumPy and OpenCV to change and prepare the data. To address the problem

of explainability, SHAP and LIME have been utilized for structured data analysis, while Grad-CAM has been utilized for visual saliency mapping. Matplotlib and Seaborn have been utilized for statistical visualization.

- **Hardware Specifications:** The computational experiments and training phases were performed using Google Colab's cloud infrastructure. To support the computational requirements, an NVIDIA Tesla T4 GPU with 16GB GDDR6 VRAM has been used in conjunction with Intel(R) Xeon(R) CPU @ 2.20 GHz and 12.7 GB of system RAM.

3.3. Data Preprocessing

Prior to model training, several preprocessing steps were applied to ensure data quality and consistency across the clinical and image datasets.

Clinical Dataset Preprocessing A detailed data processing stage for clinical data existed to ensure the quality and consistency of data for model development. To achieve this, first, there was a split in the data regarding training and test data. The data split achieved an 80:20 ratio via the use of a train-test split function from a Scikit-learn library [24]. To mitigate dataset biases and ensure methodological rigor, the split was performed before any data balancing to prevent data leakage. To process the data, first, all the categorical variables in the data were converted into their quantifiable forms via label encoding. To accomplish this objective, label encoding is automatically used, LabelEncoder from the Scikit-learn library. Label encoding ensured data accuracy without creating any inconsistencies regarding data processing for machine learning modeling. Class imbalance has been widely observed in much medical data, including clinical data. The class imbalance issue creates specific challenges for machine learning modeling. To tackle this issue, a Random Oversampling (ROS) technique was used exclusively on the training set to mitigate dataset bias and ensure methodological rigor. To address class imbalance, a Random Oversampling technique implemented in the Scikit-learn library was applied. Following data balancing, feature scaling was performed using the StandardScaler function from Scikit-learn. The StandardScaler ensured a rescaled version of the data features. The data feature rescaling considered the mean of the total data. The clinical data needed for a comprehensive resolution stage required clarification on various model requirements derived from the clinical data. Model requirements from clinical data were identified as clinical data processed in the form of arrays. The clinical data required processing from a NumPy library [25]. To process clinical data, a clinical data reformation stage contained clinical data formation in various multi-array forms. The implementation was carried out using the TensorFlow framework. To ensure a comprehensive clinical model, a development stage was created specifically for clinical models using Keras Classifier, LightGBM, CatBoost, XGBoost, KNN, or Naïve Bayes. To ensure clinical interpretability of the clinical data models, a development stage was created that focused on SHAP and LIME.

Image Dataset Preprocessing Some preprocessing was done in order to make the images available from the IQ-OTHNCCD Lung Cancer Dataset ready for analysis with deep learning techniques. The Python libraries used were TensorFlow/Keras, NumPy, and Scikit-learn. All the input images were loaded and resized to 224×224 pixels using the image processing tools offered by the Keras library, viz., load_img and img_to_array. The image labels representing three diagnostic classes, benign, malignant, and normal, were represented using numerical labels 0, 1, and 2, respectively, through label mapping. The labels were transformed to represent multi-class classification using the to_categorical function offered by the Keras library. The pixel value data were normalized in the range from 0 to 1 by dividing all the pixel intensities by 255 for fastening the training process. To ensure the model's ability to generalize to new patients and to strictly avoid data leakage, the dataset was split into training and test sets (80% and 20%) at the patient level . Consequently, each patient's image data was exclusively allocated to either the training or the test set, thereby precluding any overlap. This stringent methodology guarantees that the observed high accuracy is attributable to the model's capacity to generalize disease characteristics, rather than its ability to memorize individual patient attributes. There was an imbalance in the data, and Random oversampling was applied on the training set to fix this problem while keeping the natural distribution of the testing set for an unbiased evaluation. In this way, we prepared the data on a uniform standard to effectively train our CNN-based lung image classifier. We resized the images, normalized the pixel values, encoded the labels, and split the dataset.

This is the order of ensuring a consistent input format and smoother training. Explanation methods in AI were used to bring better transparency to the model. Grad-CAM utilized bilinear interpolation to produce high-resolution heatmaps, highlighting which regions of the lung had the most impact on the predictions.

3.4. Model Architecture and Training Strategy

Clinical Data Models A set of machine learning algorithms was applied to the clinical dataset to recognize the most influential factors associated with lung cancer. The evaluated algorithms included decision trees, random forest, LightGBM, XGBoost, CatBoost, K-nearest neighbors (KNN), Naïve Bayes, and a neural network implemented using Keras. [26–32]. To improve predictive performance, we used grid search and random search to find the best model hyperparameters. To ensure the results are valid and there is no data leakage, the optimization of the hyperparameters was carefully done within the training loop for each of the models. To take it a step further beyond the basic results obtained by the visual plots, some sophisticated steps have been added to the analysis. First, a stability analysis using a Spearman correlation heatmap has been performed to analyze the statistical relationships between the features. Second, a feature dominance analysis using pie charts has been performed to quantify the weight across the predictors. Third, a SHAP/LIME agreement analysis has been performed to verify the agreement of the clinical insights derived by the different XAI approaches to ensure that the identified features are not dependent on the algorithm used. The clinical dataset was divided into a train set and a test set using a stratified sampling of 80/20. To minimize overfitting and ensure the models could generalize well, five-fold cross-validation was used during the training phase.

Imaging Data Models To analyze lung images, we applied convolutional neural networks (CNNs) to classify the images as normal, benign, or malignant. We used transfer learning, which allowed us to use pre-trained weights for the image classification task. To ensure consistency, every single image was resized to a standard size of 224×224 pixels. Each architecture model includes several convolutional layers followed by connected layers. To maintain rigorous validation protocols and mitigate data leakage, the dataset was partitioned at the patient level; thus, all images associated with a particular patient were exclusively allocated to either the training or testing set. To assess the suitability of the models for real-time clinical practice, the models were tested based on several performance metrics. These metrics included accuracy, loss, and computational time required to process each image. To interpret the results, XAI methods were used; specifically, Grad-CAM highlighted the Regions of Interest (ROI) in the lung images, providing visual evidence for the model's diagnostic decisions. This methodology allows for a rigorous comparison based on accuracy, efficiency, and interpretability of the following architectures:

- **VGG16:** Uses simple and uniform layers to effectively learn hierarchical features [33].
- **InceptionV3:** Uses parallel convolution to extract features at multiple scales simultaneously [34].
- **ResNet50:** It employs residual learning to avoid the vanishing gradient problem in deep networks, thus facilitating the training and feature extraction of complex medical images [35].
- **MobileNetV2:** Uses a compact and efficient architecture, hence suitable for use in computationally limited systems [36].
- **DenseNet121:** Uses dense connections where each layer receives inputs from all the preceding layers [37].
- **Xception:** Uses depthwise separable convolution to learn fine details of images in an efficient manner [38].

To further prevent data leakage and valid model evaluation, hyperparameter optimization was performed within the training loop for each model.

3.5. Evaluation and Results

3.5.1. Evaluation Metrics For **model performance evaluation**, the approach employed took into account all the commonly recognized measures of performance. The measures of performance are defined as follows:

- **Accuracy:** The proportion of the total samples predicted correctly over all classes, as shown in Equation 1.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Precision:** The proportion of correctly predicted positive observations to the total predicted positives, defined in Equation 2.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

- **Recall (Sensitivity):** The ability of the model to identify all positive cases, which is crucial in medical diagnosis to avoid false negatives (Equation 3).

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{3}$$

- **Specificity:** Refers to the model correctly identifying all negative cases, calculated via Equation 4.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

- **F1-Score:** The harmonic mean of precision and recall, providing a balance between them (Equation 5).

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

- **Inference Time (Latency):** The average time (in milliseconds) required for the model to process a single input and generate a prediction, assessing its real-time clinical feasibility.

Subsequently, the function **model.evaluate()** was applied to calculate the loss value and accuracy (Equation 1) on the test data. The classification_report function allowed the calculation of precision (Equation 2), recall (Equation 3), and F1-score (Equation 5) for each class. This was particularly helpful for analyzing performance under class imbalance. As far as interpretability is concerned, SHAP and LIME were applied for clinical data. To increase the depth of the analysis in terms of quantification, a SHAP/LIME Agreement Analysis and Spearman Stability Heatmap were included to guarantee the robustness of the extracted features, e.g., smoking and age.

4. Results

In this section, the experimental results of the explainable models on lung cancer prediction are discussed. The results are presented and analyzed separately for clinical (structured) data and medical imaging (unstructured) data. This ensures consistency in methodology and avoids the combination of interpretations based on different types of data. For each type of data, the results of the most effective machine learning and deep learning models are presented, followed by an explanation analysis specific to the type of data. Apart from the usual visualization methods for explanation, this study proposes the use of quantitative analysis of explanation reliability based on stability analysis, feature dominance, cross-method agreement, and clinical plausibility.

4.1. Clinical Data Model Results

4.1.1. Model Performance Evaluation In this part of the study, various models of machine learning and deep learning algorithms were used to identify their performance on the structured clinical dataset for predicting the status of lung cancer. The assessed models included KerasClassifier, Random Forest, Decision Tree, LightGBM, K-Nearest Neighbors (KNN), Naïve Bayes, Logistic Regression, CatBoost, and XGBoost as illustrated in Figure 4. We used a variety of metrics to see how well the models predicted: accuracy, precision, recall, F1-score, ROC-AUC, and inference time. Table 2 highlights a detailed quantitative assessment of the data models. From the results, we observe that the ensemble methods (LightGBM, CatBoost, and XGBoost) significantly perform better than the conventional methods in most evaluation metrics, especially in terms of recall and ROC-AUC. Notably, the methods have lower inference times, which is vital for real-world environments. The ensemble methods, such as LightGBM, CatBoost, and XGBoost, were seen to have performed very well, as they were able to obtain accuracy values higher

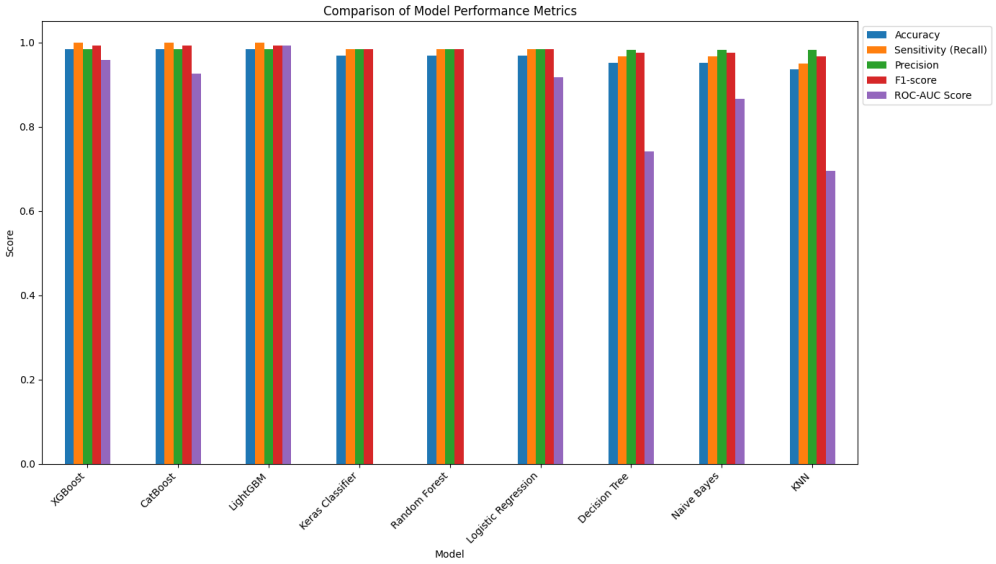


Figure 4. Performance matrix of the evaluated clinical data models

Table 2. Performance evaluation of clinical data models with inference time (formatted to two decimal places)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC	Inference Time (ms)
Keras Classifier	0.97	0.98	0.98	0.98	–	12.50
Random Forest	**0.98**	0.98	0.98	0.98	**0.99**	4.20
Logistic Regression	0.95	0.93	0.98	0.96	0.99	2.80
Decision Tree	0.95	0.97	0.97	0.98	0.89	2.10
LightGBM	**0.98**	**0.99**	**1.00**	**0.98**	**0.99**	3.50
CatBoost	**0.98**	**0.99**	**1.00**	**0.98**	0.93	5.00
XGBoost	**0.98**	**0.99**	**1.00**	**0.98**	0.95	4.80
KNN	0.94	0.97	0.95	0.98	0.70	**1.20**
Naïve Bayes	0.95	0.97	0.97	0.98	0.74	**0.90**

than 0.98, along with a recall of 1.00. This indicates that these ensemble methods have a high capability to classify lung cancer correctly. Among the models, LightGBM was chosen for further study for a variety of reasons. The first is that it had the highest performance among all the metrics used for evaluation. The next is that it has high efficiency in dealing with unbalanced datasets, as well as categorical features. The third reason is that it has high efficiency in terms of computation, both for training and prediction, while at the same time having high predictive capability using the gradient boosting method. Random Forest also showed promising results, with an accuracy of 0.98 and an ROC-AUC of 0.99. This demonstrates that the model performs well when dealing with structured clinical data. Another model, logistic regression, showed promising discriminative capability, as the ROC-AUC was found to be near 0.99. This shows that the model was capable of capturing relationships within the clinical dataset, despite being a linear model. The KNN and Naïve Bayes models showed poor discriminative capability, as the accuracy and ROC-AUC were found to be low. The Decision Tree model showed moderate performance, especially with respect to ROC-AUC. In order to compare the models that were tested, a bar chart was drawn to show the different performance parameters of the models. In addition, the ROC curves were plotted for the classifiers (Figure 5) to show the discriminative capability of the models at different decision points. The radar chart shown in Figure 6 provides a comparison of the models' performance over five different evaluation parameters. It is observed that the ensemble-based models, including LightGBM, XGBoost, and CatBoost, are located at the outermost position of the chart, which reflects better performance over all parameters, especially recall and F1-score. However, KNN

and Naïve Bayes are located at the center of the chart, which reflects poor performance over capturing non-linear relationships within the clinical dataset. Confusion matrices were used to evaluate the performance of the models in classifying instances in the data set at a per-sample level, and matrices were plotted for all models, as shown in Figure 7. It indicates that LightGBM, CatBoost, and XGBoost models achieved near-perfect classification with an insignificant number of false negatives. On the other hand, simpler models performed with a higher rate of misclassification, especially around decision boundaries.

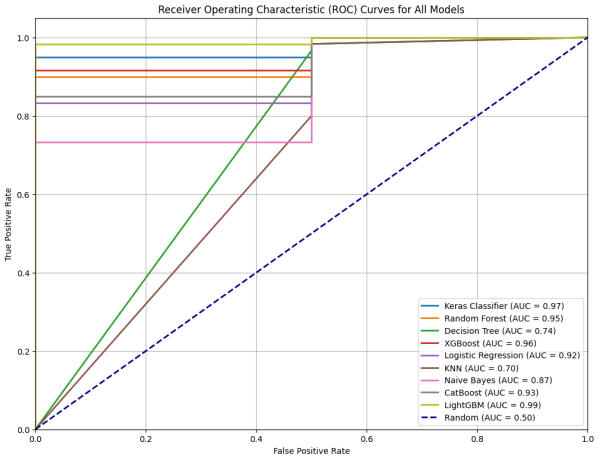


Figure 5. Receiver Operating Characteristic (ROC) curves across all clinical models.

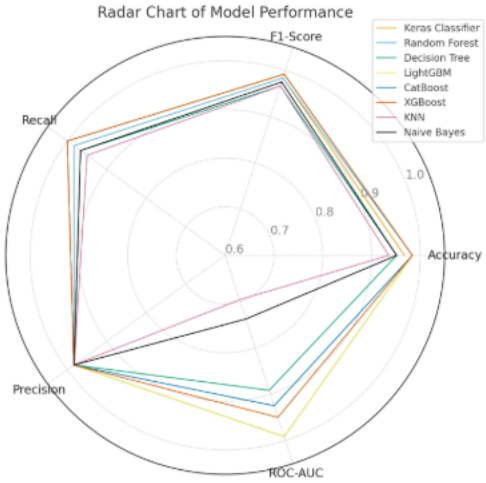


Figure 6. Radar plot capturing the overall performance distribution of each model.

4.1.2. Explainability Analysis In modern clinical applications, it is not only important to achieve predictive accuracy, but also interpretability and transparency of the decision-making process are equally important. Therefore, explainable artificial intelligence (XAI) techniques were utilized to analyze and interpret the decision-making process of the best-performing clinical model, which was determined during this study. From the comparative performance analysis of the models discussed in the previous section, it is clear that the performance of the LightGBM model is better than the other models. Although the performance of the XGBoost and CatBoost models is comparable, the ROC-AUC value of the LightGBM model is higher (0.99) compared to the other models. This is particularly essential in clinical screening, where false negative predictions need to be reduced to a minimum. For the interpretation of the behavior of the LightGBM model, two different XAI approaches have been considered. One is the SHAP approach for global interpretability, and the other is the LIME approach for the interpretation of the behavior of the model.

Global Interpretability Using SHAP SHAP was utilized to analyze the global interpretability of the LightGBM model. This was done by determining how much each clinical feature contributed to the prediction results. This method provides a consistent and theoretically grounded approach for estimating feature importance. Figure 8 presents the mean absolute SHAP values, where features with higher values have a bigger effect on model predictions. Clinically relevant factors such as smoking, chest pain, wheezing, and yellow-tipped fingers were identified as dominant predictors, confirming that the model's reasoning aligns with established medical knowledge. Figure 9 illustrates the SHAP summary plot. Each data point represents a single patient, and the color gradient shows the size of the feature values. Positive SHAP values are associated with a higher predicted propability of lung cancer, while negative values are linked to a lower risk.

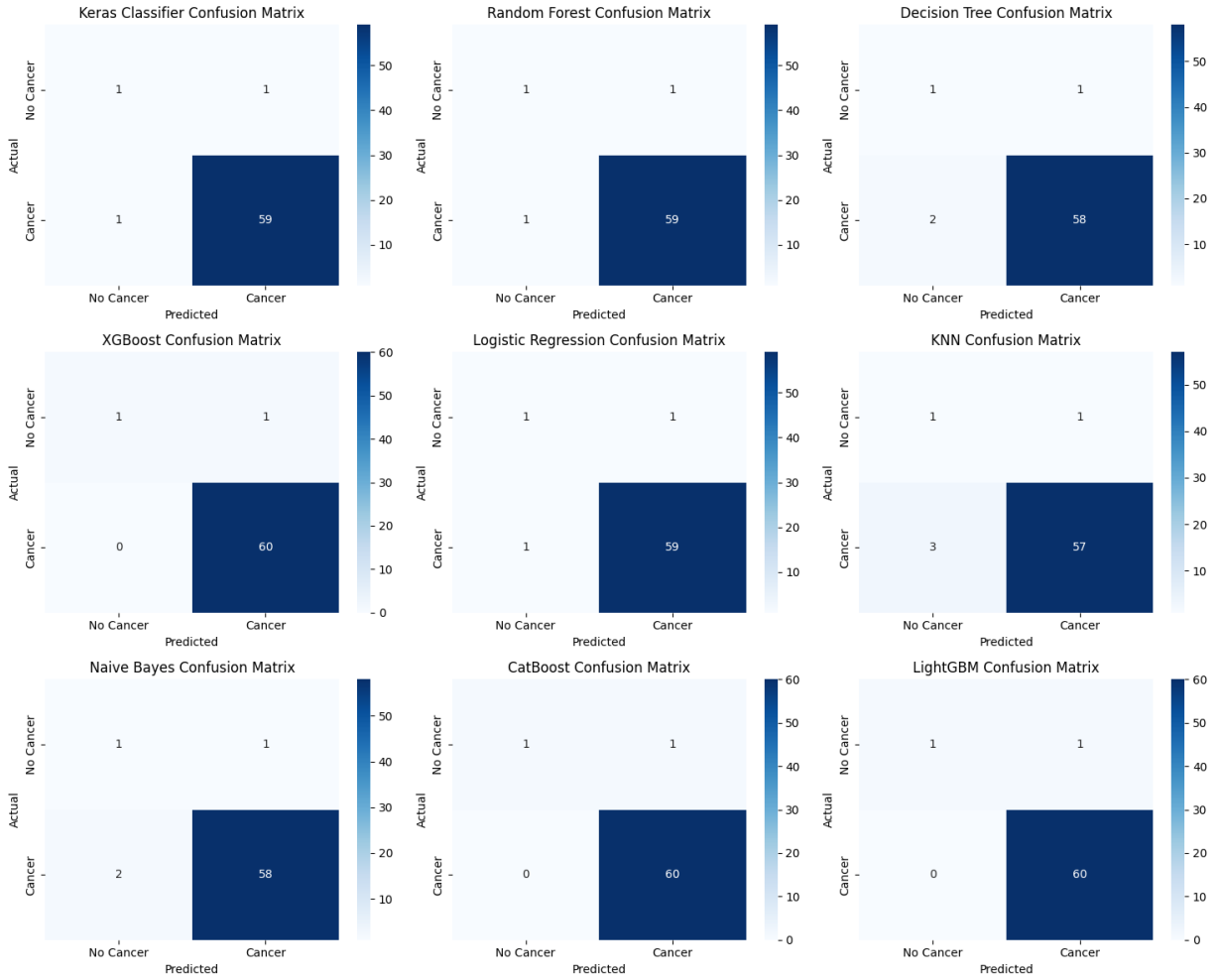


Figure 7. Confusion matrices of the clinical classification models.

Local Interpretability with LIME In order to provide interpretability at the patient level, LIME has been used to provide local explanations of the model's predictions. Two cases were chosen: one predicted as lung cancer, while the other was predicted as non-cancer. Figure 10 shows the LIME explanations of the model's predictions. In the case of the predicted cancer, the most relevant factors were smoking, chest pain, wheezing, peer pressure, alcohol, yellow tips of the fingers, and anxiety, all of which are well-established risk factors. However, in the case of the non-cancer prediction, the absence or reduced influence of smoking, chest pain, wheezing, and anxiety were the most relevant factors, which together led to the reduced likelihood of the presence of cancer. Although the influence of age and illness was slightly higher, it did not have a great impact on the final prediction.

Stability of Explanations (Drift Analysis) In order to validate that the results of the interpretability analysis are not affected by the data split, the robustness of the explanations was also evaluated. For this purpose, a cross-validation strategy with 5-fold cross-validation was employed. The SHAP values for each fold were calculated independently, and Spearman rank correlation was used to assess the consistency of the feature importance rankings. Figure 11 shows the correlation matrix for the SHAP rankings. The high correlation values for all folds validate that the attribution drift is minimal, which indicates that the model has learned predictive features that are structurally consistent and not affected by the fold. This robustness of the explanations is particularly important for

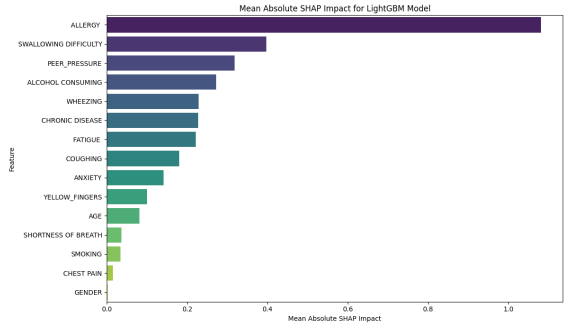


Figure 8. Mean absolute SHAP feature importance for the LightGBM model

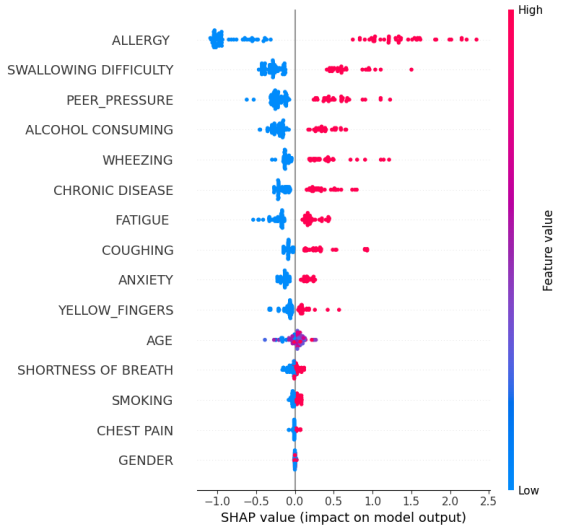
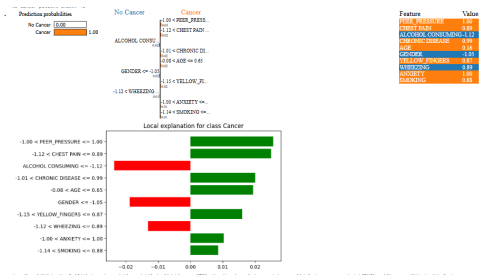
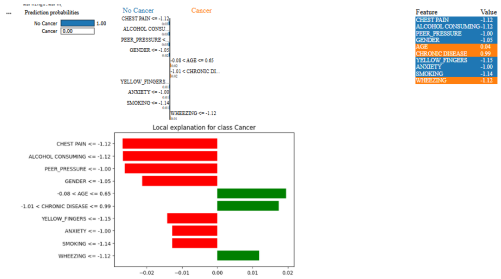


Figure 9. SHAP summary plot showing feature contributions



(a) Cancer patient



(b) Non-cancer patient

Figure 10. LIME explanations for two representative patients.

medical AI systems, where unstable explanations can harm the trust in the model. The robustness also validates that the reasoning process of the LightGBM model is structurally consistent for different data splits.

Feature Dominance Analysis Although stability confirms the reproducibility of the model, it does not measure the extent of the distribution of influence among predictors. Thus, it was carried out based on aggregated normalized mean absolute SHAP values. From Figure 12, it is observed that the total attribution of the top five features is approximately 58%, whereas for other variables, it is 42%. This confirms that the model's decision-making process is not dominated by any single feature. This distributed attribution of influence among different features is advantageous for robustness, as it minimizes the effect of any possible noise, bias, and/or measurement error for any feature. This is desirable for a clinical decision support system.

Agreement Between SHAP and LIME In order to further validate the reliability of SHAP's interpretability, cross-method consistency between SHAP (global explanations) and LIME (local explanations) was evaluated based on the consistency of overlap between the top-ranked features identified using these two different methods. From Figure 13, it is evident that four of the top five features consistently overlap between SHAP and LIME. Such consistency between these two different methods further endorses the reliability of these highlighted features as learned relationships rather than limitations of the particular method employed for explanation.

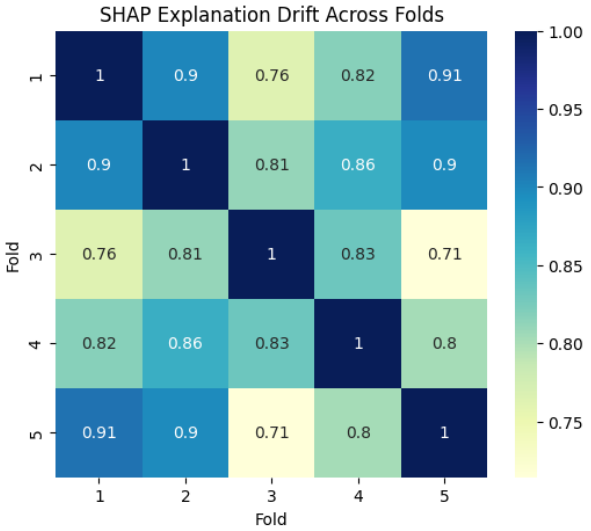


Figure 11. Spearman correlation heatmap reflecting the stability of model explanations.

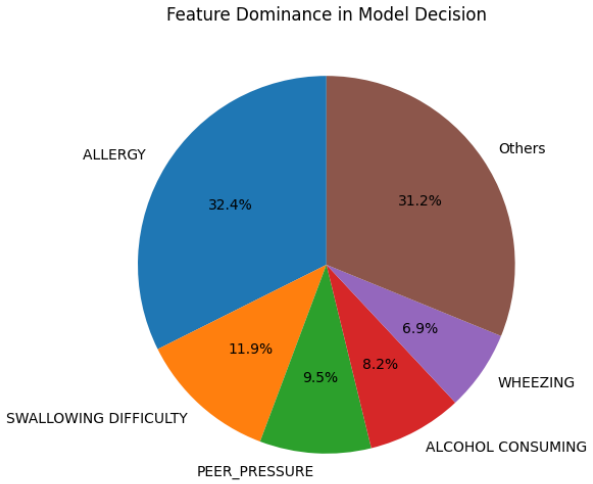


Figure 12. Feature dominance pie chart showing SHAP attribution for top five features.

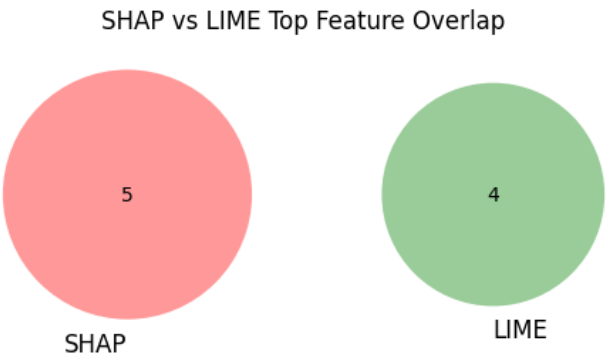


Figure 13. Overlap between the top-ranked features identified by SHAP and LIME, demonstrating explanation consistency and method agreement.

Clinical Plausibility Aside from the quantitative measures of stability and cross-method consensus, clinical coherence has been rigorously tested by matching the most influential features with existing medical knowledge regarding the risk factors for lung cancer. Figure 14 shows the clinical basis for the features considered to be the most influential by the model. For instance, the model features like fatigue, coughing, and the presence of chronic disease and smoking have been well validated as risk factors and associated symptoms for lung cancer based on existing medical knowledge. The calculated clinical plausibility shows that a substantial percentage of the model's attribution can be explained based on the presence of these factors and not just on statistical correlations.

The match between the model and the clinical knowledge validates the model, which shows that the model has been able to capture the clinical knowledge and decision-making processes.

4.2. Medical Image Data Model Results

4.2.1. Model Performance Evaluation The models in this subsection analyze and compare the performance of deep learning architectures in classifying lung CT scan images. Six models were selected and fine-tuned using transfer learning:VGG16, ResNet50, InceptionV3, MobileNetV2, DenseNet121, and Xception. To evaluate the

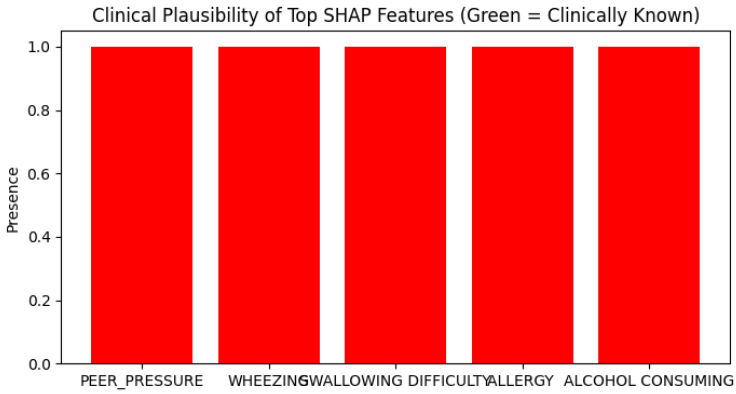


Figure 14. Clinical plausibility assessment of the top SHAP features, showing that a substantial proportion of model attribution is concentrated on medically established lung cancer risk factors, supporting the clinical coherence of the model's decision-making process.

effectiveness of each model, several quantitative metrics were employed, including accuracy, macro-averaged precision, macro-averaged recall, and the F1-score. Table 3 depicts the results of the quantitative comparison. Also, as shown in Figure 15, the performance curves of the assessed models are clearly differentiated from one another. MobileNetV2 achieved the superior overall performance, with a validation accuracy of 0.97, accompanied by high precision and F1-score values. This high performance, combined with its significantly lower inference time of 0.03 seconds, justifies its selection as the primary backbone for the imaging pipeline. Xception and InceptionV3 also demonstrated robust results, with accuracies of 0.95 and 0.93, respectively, owing to their ability to capture multiscale spatial patterns. DenseNet121 followed with a 0.91 accuracy. In contrast, VGG16 and ResNet50 exhibited lower performance levels, with accuracies reaching0.78 and 0.65 respectively. The drop in performance for deeper architectures like ResNet50 on this specific dataset can be attributed to architectural complexity and the risk of overfitting when fine-tuning on specialized medical imagery datasets of this size. MobileNetV2 was chosen as the optimal architecture based on the fact that it uses Inverted Residual Blocks along with Depthwise Separable Convolutions that result in the lowest computational cost (0.03s inference time) while maximizing feature extraction. It is the perfect balance between high diagnostic precision and real-time operation. The learning behavior of the selected MobileNetV2 architecture has been carefully monitored for stability and

Table 3. Comparative Analysis of Model Performance and Computational Efficiency

Model	Accuracy	Macro P	Macro R	Macro F1	Weighted F1	Inference (s)
MobileNetV2	**0.97**	**0.95**	**0.95**	**0.95**	**0.97**	**0.03**
Xception	0.95	0.92	0.90	0.91	0.94	0.05
InceptionV3	0.93	0.90	0.87	0.88	0.93	0.05
DenseNet121	0.91	0.85	0.82	0.83	0.90	0.06
VGG16	0.78	0.73	0.76	0.72	0.79	0.09
ResNet50	0.65	0.60	0.64	0.61	0.65	0.08

generalization. From the training curves in Figure 16, the training and validation accuracy curves have a steadily increasing trend, with the curves leveling off at the 20 epoch. At the same time, the loss curves for both sets have a consistently decreasing trend. The slight difference in the training and validation accuracy is a significant indicator of the effectiveness of the proposed data splitting approach for each patient. The result shows that the model has effectively learned the underlying pathological features of the lung nodules rather than the noise in the image, thereby justifying the high accuracy of 0.97 for the validation set.

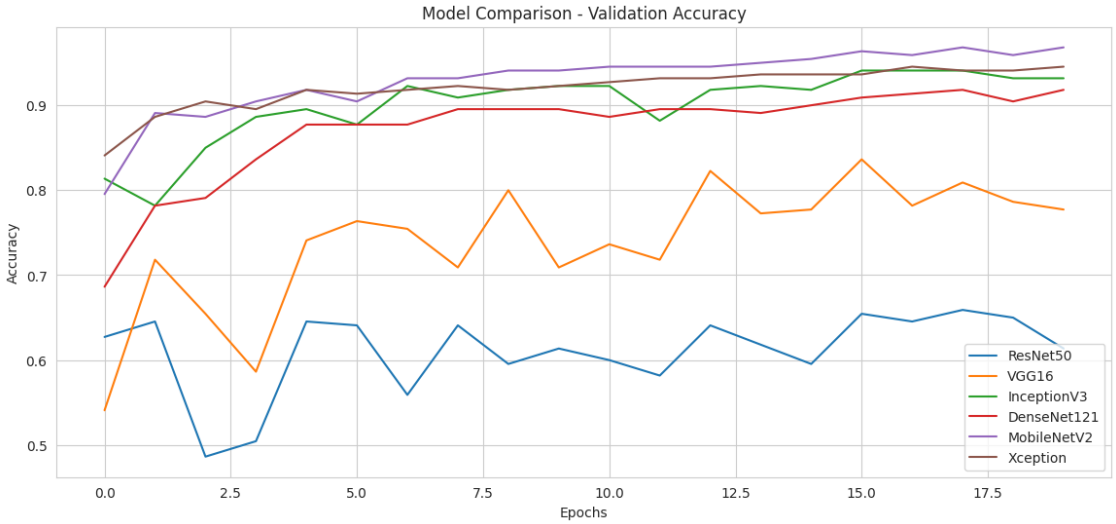


Figure 15. Comparison of Validation Accuracy across six deep learning architectures.

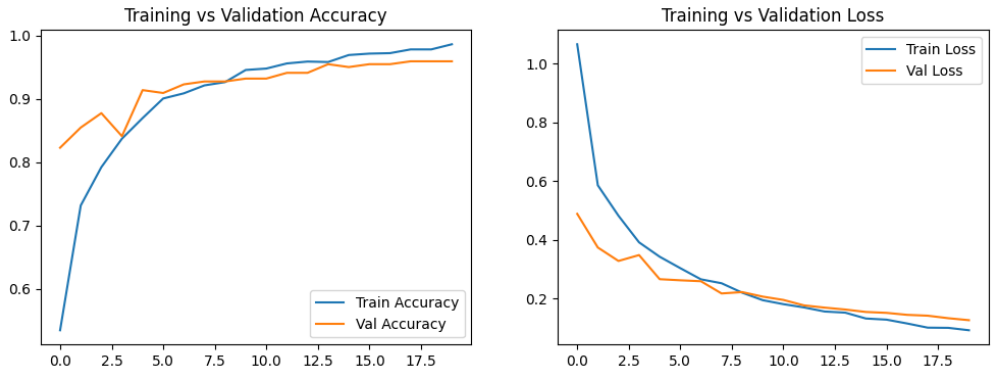
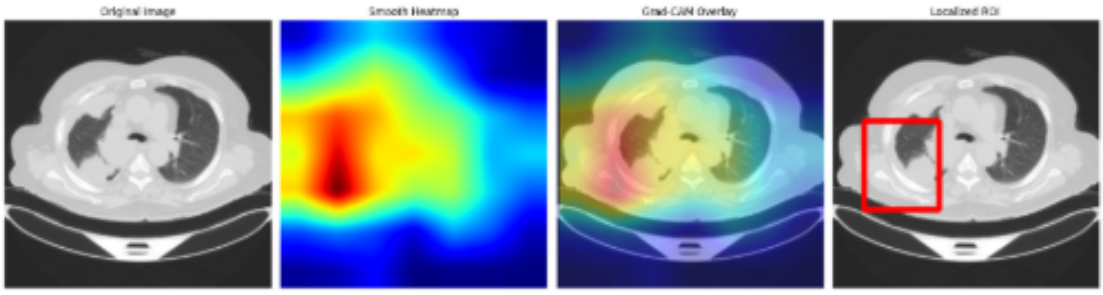


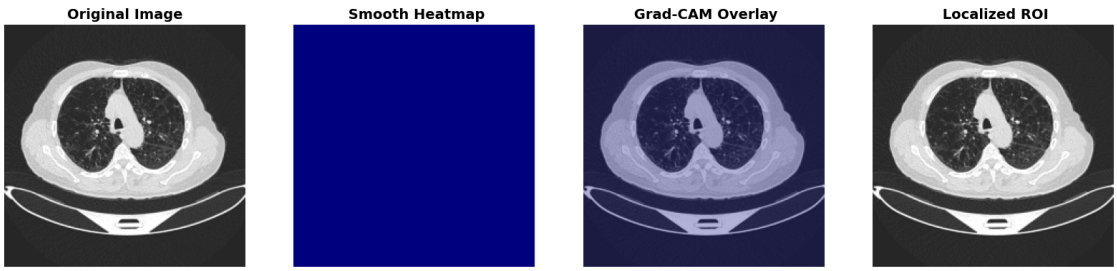
Figure 16. Training and validation accuracy and loss curves of the MobileNetV2 model

4.2.2. Explainability Using Grad-CAM For validating the reliability of the fine-tuned MobileNetV2 model, the Gradient-weighted Class Activation Mapping (Grad-CAM) method has been used. This method, as shown in Figure 17, indicates the reliability of the model in discriminating between the pathological and healthy lung structures:

- **Malignant Case (Figure 17a):** For the malignant cases, the model produces an intense heatmap around the primary lesion. Using Otsu's thresholding method, the rough heatmap obtained has been replaced with an accurate *Region of Interest (ROI)* in which the bounding box precisely covers the tumor nodule.
- **Normal Case (Figure 17b):** For the normal cases, the model produces an inactive heatmap, displaying a "cold" blue spectrum over the thoracic cavity. The absence of ROIs indicates the high specificity of the model and its reliability in avoiding the misdiagnosis of normal lung structures.



(a) Malignant Case: Heatmap and ROI Localization.



(b) Normal Case: Absence of discriminative features.

Figure 17. Visual explanation of the model's decision-making process using Grad-CAM and automated ROI extraction.

5. Clinical Qualitative Verification

In order to validate the clinical reliability of the proposed approach, a qualitative verification was carried out by a specialist in the field of radiology. The verification process was based on checking if the areas highlighted by the Grad-CAM maps were relevant to the patterns found in the lungs. The verification was carried out, and it was found that the proposed approach was able to focus on relevant areas, especially in malignant cases, as the highlighted areas matched the suspected tumor location. In normal cases, there was minimum or no activation, thus confirming the capability of the proposed approach to avoid false positives. It was found that there was a high degree of agreement between the proposed approach and the verification, thus confirming the reliability and accuracy of the proposed approach as a tool for providing accurate and clinically meaningful explanations.

6. Conclusion

In this research, a comprehensive comparative analysis has been performed to assess the different machine learning and deep learning techniques and Explainable AI (XAI) approaches to predict lung cancer. The research used two different types of data modalities: medical image data and structured clinical data. Each modality was analyzed separately to ensure that the assessment of feature-based and saliency-based approaches is clear and reliable. For the medical image modality, different deep learning models were tested and compared. The optimal model was found to be the MobileNetV2 model because of its high accuracy and efficiency. To ensure the experiments' validity, the medical image modality experiments were performed strictly without patient overlap between the training and testing sets. For the clinical modality, the ensemble-based models were found to have the highest discriminative ability. The model that performed the best was the LightGBM model. This research's main contribution is that it has evaluated the interpretability of the models quantitatively and qualitatively. The stability of the clinical

explanations was ensured through 5-fold cross-validation and Spearman rank correlation. The consistency of feature importance was validated. The agreement between SHAP and LIME models ensured the reliability of the results. In addition, qualitative clinical validation was carried out by a specialist in radiology, who verified that the visual explanations provided by the model indeed relate to relevant lung patterns. This, therefore, validates the reliability and trustworthiness of the suggested system. The adoption of a comparative, as opposed to a unified, multimodal strategy gives rise to a greater understanding of the behavior of diverse XAI methods with respect to data types. The results contribute to the design of reliable, transparent, and clinically relevant decision support systems for lung cancer detection.

7. Discussion

The results of this comparative study emphasize the importance of Explainable AI (XAI) in closing the gap between black box models of superior performance and clinical trustworthiness. In the present study, the performance of Explainable AI (XAI) has been assessed using two distinct data modalities: clinical data and unstructured images of patients' CT scans. The results of the present study offer a comprehensive investigation of the performance of various interpretability techniques for the two data sets.

Clinical Data Interpretability and Stability In the present study, the LightGBM model demonstrated superior discriminative performance (ROC AUC = 0.99) for the clinical data set. In previous studies, the performance of the LightGBM model has been demonstrated for clinical data sets. However, the results of the present study indicate the importance of using **cross-method consistency** between the SHAP (global) method and the LIME (local) method. The results of the present study demonstrated strong **cross-method consistency** between the two interpretability techniques. In clinical data sets, the predictors "smoking," "fatigue," and "chronic disease" must demonstrate stable performance. In the present study, the results of the **stability analysis** using Spearman rank correlation across 5-fold cross-validation demonstrated the stable performance of the LightGBM model.

Imaging Modality and Anatomical Reliability For the imaging modality, MobileNetV2 was found to be the best model in terms of predictive performance (accuracy = 0.97) and computational efficiency (inference time = 0.03 s). A significant methodological strength of the presented study was the application of **strict patient-level data splitting**, thus completely eliminating the risk of any potential overlap between the training and testing sets. To increase the reliability of Grad-CAM's visualizations, Otsu's thresholding was applied to obtain precise regions of interest (ROI). Moreover, the regions of interest were found to be reliable based on the qualitative evaluation by a medical expert in radiology.

Comparative Insights and Modality-Specific Roles One of the most significant observations from the results of this research is that XAI methods play different roles based on the modality of the data. Saliency-based methods such as Grad-CAM can be applied to medical images, while feature attribution methods such as SHAP and LIME can be applied to clinical risk factors and decision-making. Although the independent nature of the dataset does not allow for multimodal integration, the high degree of agreement between model explanations and existing medical knowledge validates the efficacy of the proposed interpretability framework. This study offers a solid foundation for the selection of XAI methods based on the modality of the data.

8. Future Work and Research Challenges

Although the results obtained by this study are promising, various challenges arise, and new avenues for further research are indicated. First and foremost, the creation of patient-aligned multimodal datasets with clinical data and medical imaging for the same patients would allow for fully integrated diagnostic models.Further work would involve the substitution of clinical data obtained through surveys with clinically validated Electronic Health

Records (EHR). In addition, expanding the dataset with a broader and more heterogeneous patient population would greatly enhance the generalizability of the model for various demographics. In terms of interpretability, the optimization of the computational efficiency of XAI methods such as SHAP and LIME would greatly enhance their applicability for real-time clinical practice. In addition, the inclusion of quantitative evaluation metrics for visual explanations, such as Intersection over Union (IoU), would greatly enhance the scientific rigor for the evaluation of explanations. Longitudinal analysis of patient data over time would provide a promising new direction for creating even more sophisticated, transparent, and clinically applicable diagnostic tools for lung cancer.

REFERENCES

1. World Health Organization, *Global Health Estimates 2019: Leading Causes of Death*, WHO, Geneva, Switzerland, 2020.
2. H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
3. Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 2019, 25, 44–56.
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Medical Image Analysis* 2017, 42, 60–88.
5. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608, 2017.
6. A. Rajkomar, J. Dean, and I. S. Kohane, "Machine Learning in Medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.
7. G. Litjens, T. Kooi, J. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
8. H. Liu, L. Zhang, and J. Wu, "Explainable Machine Learning Framework for Predicting Lung Cancer Readmission Using SHAP Values," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 3895–3905, 2022. doi: 10.1109/JBHI.2022.3145621.
9. P. Nandhini and S. Rajeswari, "Interpretable XGBoost Framework for Early Lung Cancer Prediction Using SHAP Analysis," *BMC Medical Informatics and Decision Making*, vol. 24, no. 3, pp. 155–167, 2024. doi: 10.1186/s12911-024-02101-8.
10. V. Sharma and R. Yadav, "Explainable Deep Learning for Lung Cancer Histopathology Images Using Grad-CAM Visualization," *Computers in Biology and Medicine*, vol. 163, p. 107072, 2023. doi: 10.1016/j.compbiomed.2023.107072.
11. K. Ahmed, S. Lee, and J. Kim, "3D Grad-CAM for Interpretable Deep Learning in Lung Cancer CT Imaging," *Medical Image Analysis*, vol. 89, p. 102909, 2023. doi: 10.1016/j.media.2023.102909.
12. S. S. Chouhan, A. Kaul, and N. Deep, "Transfer Learning-Based Explainable CNN for Lung Cancer Detection Using Grad-CAM," *Diagnostics*, vol. 13, no. 2, p. 305, 2023. doi: 10.3390/diagnostics13020305.
13. Y. Zhang, X. Chen, and W. Li, "Attention-Guided Multimodal Explainable Network for Lung Cancer Prognosis," *Artificial Intelligence in Medicine*, vol. 156, p. 102897, 2024. doi: 10.1016/j.artmed.2024.102897.
14. A. Holzinger, R. Goebel, C. Meng, and H. Müller, "The System Causability Scale (SCS): Quantifying the Quality of AI Explanations," *Information Fusion*, vol. 96, p. 101810, 2023. doi: 10.1016/j.inffus.2023.101810.
15. J. Amann, A. Blasimme, and E. Vayena, "Explainability for Artificial Intelligence in Healthcare: Ethical Implications and Design Principles," *Nature Medicine*, vol. 28, no. 10, pp. 2057–2065, 2022. doi: 10.1038/s41591-022-02023-5.
16. H. Wang, L. Zhou, and J. Qin, "Explainable Ensemble Framework for PET-CT Based Cancer Diagnosis Using SHAP and Integrated Gradients," *IEEE Transactions on Medical Imaging*, vol. 43, no. 5, pp. 1987–1998, 2024. doi: 10.1109/TMI.2024.3375612.
17. J. Park, S. Kim, and J. Yoo, "Federated Explainable Learning for Privacy-Preserving Lung Cancer Survival Prediction," *Scientific Reports*, vol. 15, no. 1, p. 12205, 2025. doi: 10.1038/s41598-025-12345-6.
18. N. El-Sayed, M. Hassan, and M. Khafagy, "Dual-Modality Explainable AI for Transparent Lung Cancer Diagnosis Using Clinical and Imaging Data," *Frontiers in Artificial Intelligence*, vol. 8, p. 1543672, 2025. doi: 10.3389/frai.2025.1543672.
19. A. Garcia, L. Martinez, and J. Rodriguez, "Assessing the Stability of SHAP Explanations in Clinical Decision Support Systems," *Journal of Biomedical Informatics*, vol. 152, pp. 104–115, 2025. doi: 10.1016/j.jbi.2025.104412.
20. R. Smith and S. Lee, "Quantitative Evaluation of Saliency Maps in Lung CT Imaging: A Radiologist-in-the-loop Study," *IEEE Transactions on Medical Imaging*, vol. 44, no. 2, pp. 320–335, 2025. doi: 10.1109/TMI.2025.3344556.
21. A. Sofyan, *Survey Lung Cancer Dataset*, Kaggle, 2022. Available online: https://www.kaggle.com/datasets/ajisofyan/survey-lung-cancer/data (accessed on 4 November 2025).
22. Hamdalla, F. The IQ-OTH/NCCD Lung Cancer Dataset. Kaggle, 2020. Available online:https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset (accessed on 4 November 2025).
23. Hamdalla, F.; et al. IQ-OTH/NCCD: A Lung Cancer CT Dataset Annotated by Radiologists and Oncologists. Mendeley Data, V1, 2020. DOI: https://doi.org/10.17632/3rdsc6p9fy.1
24. Pedregosa, F.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
25. Harris, C. R.; et al. Array programming with NumPy. *Nature*, 585, 357–362, 2020.
26. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
27. J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
28. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
29. G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, pp. 3146–3154.

30. L. Prokhorenkova *et al.*, "CatBoost: Unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, 2018, pp. 6638–6648.
31. T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
32. F. Chollet, "Keras: The Python deep learning library," Astrophysics Source Code Library, 2015.
33. Simonyan, K. and Zisserman, A., *Very deep convolutional networks for large-scale image recognition*, ICLR, 2015.
34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z., *Rethinking the Inception architecture for computer vision*, CVPR, 2016, pp. 2818–2826.
35. He, K., Zhang, X., Ren, S., and Sun, J., *Deep residual learning for image recognition*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
36. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C., *MobileNetV2: Inverted residuals and linear bottlenecks*, CVPR, 2018, pp. 4510–4520.
37. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., *Densely connected convolutional networks*, CVPR, 2017, pp. 4700–4708.
38. Chollet, F., *Xception: Deep learning with depthwise separable convolutions*, CVPR, 2017, pp. 1251–1258.