STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–19.
Published online in International Academic Press (www.IAPress.org)

An Explainable Decision Support System for Lung Cancer Diagnosis from CT Images Using Hybrid AI Models

Ghada Nady [1,*], Osama Badawy [1], Ahmed Salem [2]

[1] *College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt*
[2] *College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Cairo, Egypt*

Abstract Lung cancer is one of the main causes of death around the world, so early detection is crucial. Detecting small pulmonary nodules in CT scans is far from straightforward — it demands trained radiological judgment and considerable time, and error rates climb as imaging volumes grow. This study proposes a hybrid AI framework that tackles this problem from two directions: combining handcrafted descriptors — texture, shape, intensity, and wavelet features — with deep representations from CNN, CoAtNet, and EfficientNet, while keeping the model's reasoning transparent enough for clinical use. Before classification, multi-stage feature selection was applied to strip out redundant signal and retain only what genuinely differentiates malignant from benign tissue. Multiple machine learning and hybrid deep learning configurations were evaluated, with the top-performing model integrated with explainable AI techniques to generate decision-level insights for radiologists. Validation was performed on 30,020 CT images drawn from the BOWL2017 dataset across 790 patients, yielding 98% accuracy, 98.4% precision, and 98.0% recall. Radiologist evaluation confirmed that the model's attention maps aligned with clinically relevant nodule characteristics, supporting diagnostic confidence and workflow efficiency. These findings demonstrate the viability of interpretable hybrid AI as a practical decision support tool in CT-based lung cancer screening.

Keywords Lung Cancer Detection, CT Imaging, Deep Learning, Feature Fusion, Explainable AI,decision support

DOI: 10.19139/soic-2310-5070-3511

1. Introduction

Lung cancer accounts for 5–7% of all cancer diagnoses in Egypt [1], and remains one of the leading causes of cancer death globally. Catching it early matters enormously — survival rates can climb from 60% to 80% with timely intervention [2]. CT scanning is the standard tool for nodule detection, but reading these images manually is time-intensive and heavily dependent on individual radiologist experience. As scan volumes grow, so does the likelihood that small or ambiguous nodules get missed [3] — a problem that has pushed researchers toward automated detection solutions.

AI-based diagnostic systems have emerged as a practical response to this challenge [4]. Within this landscape, both machine learning (ML) and deep learning (DL) techniques have been extensively applied to lung cancer detection, demonstrating a strong capacity to identify abnormal patterns in CT images and differentiate malignant from benign lesions [5]. These systems have shown measurable benefits in reducing radiologist workload and improving diagnostic consistency. Nevertheless, each paradigm carries inherent limitations. Conventional ML approaches depend on manual feature engineering, which introduces feature dependency constraints and limits

*Correspondence to: Ghada Nady Abd ElGawad (Email: g.abdelgawad56@student.aast.edu). College of Computing and Information Technology, Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt.

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

scalability across diverse imaging datasets. Deep learning architectures, while more powerful in representation learning, operate largely as black boxes — their internal decision logic is opaque, which gives clinicians reasonable grounds to question the reliability of automated predictions [6]. Bridging this transparency gap is therefore not merely a technical concern but a clinical one.

Structured feature engineering combined with XAI offers one way to address this — pushing models to respond to imaging characteristics that actually matter clinically, rather than incidental correlations in the training data [7]. But clinical adoption has not kept pace with technical progress. Handling large CT datasets is computationally expensive, and many deep learning models struggle precisely where it counts most: resolving the fine detail of small nodules that are hardest to classify. On top of this, most models offer no window into their reasoning. For radiologists already accountable for every diagnostic call they make, that opacity is not a minor inconvenience — it is a fundamental barrier to trust [8].

A further limitation pervades the existing literature: most prior studies conduct feature selection in a single phase and rely on direct feature concatenation, an approach that fails to account for inter-feature redundancy, inflates dimensionality, and risks discarding clinically meaningful information. While deep learning and XAI methods are increasingly represented in medical imaging research, few studies systematically address leakage-aware model design, multi-stage feature fusion, and clinical validation within a unified framework. The absence of structured radiologist involvement in model evaluation compounds this problem — technical performance metrics alone are insufficient to establish the clinical validity of a diagnostic system.

This research addresses key challenges that slow the clinical adoption of AI-based decision support for lung cancer, despite recent advances in Deep Learning and machine learning. As Lung CT datasets grow, computational complexity increases. Meanwhile, many Deep Learning models struggle to capture the fine details of small pulmonary nodules, which are critical for accurate identification. Radiologists also struggle to understand and trust model predictions because Deep Learning can be complex. Furthermore, previous studies have focused on radiomics and performed feature selection in a single phase, often using direct feature concatenation. This method can miss redundancy, create high-dimensional features, lower generalization, and lead to the loss of relevant clinical information. Deep Learning and Explainable AI (XAI) are increasingly being used in healthcare and medical imaging, but the medical imaging literature is rich with only a few examples of works that systematically investigate leakage-aware deep learning, feature fusion, and XAI, especially in the context of a fully clinically validated lung cancer CT imaging analysis system,failing to integrate them into a single model that incorporates radiologists' input. These limitations highlight the need for a structured, interpretable, and leakage-aware diagnostic model that builds clinical trust, optimizes feature selection, and improves performance for accurately detecting lung cancer.

To address these gaps, this study develops a hybrid AI framework for lung cancer detection and classification from CT images. Rather than optimizing for benchmark performance alone, the framework was built around a clinical validation loop — radiologist feedback was embedded directly into the evaluation process, not added as an afterthought. Multiple machine learning and deep learning architectures were tested, multi-stage feature selection and fusion were applied to manage dimensionality without sacrificing diagnostic signal, and XAI techniques were integrated to make model decisions legible to the clinicians who would ultimately use them. The main contributions of this paper are as follows:

- A systematic comparison of machine learning models and deep learning architectures for CT-based lung cancer classification
- A multi-stage feature engineering pipeline combining handcrafted descriptors (texture, shape, intensity, wavelet) and deep features, with sequential selection algorithms applied to reduce redundancy and control dimensionality
- A comparative feature fusion strategy designed to optimize diagnostic performance while preserving clinically interpretable information
- Integration of XAI (Grad-CAM) to provide interpretation and significant medical insights into model predictions.
- A structured radiologist evaluation protocol embedded within the validation process to assess clinical alignment and build practitioner trust

- A well-structured, leakage-aware, and validated AI framework that emphasizes accuracy, transparency, and usability, supporting the reliable adoption of AI tools in lung cancer screening and decision support.

The paper is organized as follows: Section II reviews related work and points out existing gaps. Section III explains the proposed methodology and the entire processing pipeline. Section IV presents the experiments and results of the framework. Finally, Section V provides the conclusion.

2. Related Works

In recent years, Artificial Intelligence has made significant strides, particularly in machine learning and Deep Learning for cancer decision support. Lung cancer has become a major focus, driving research into automated classification [29]and detection. This section reviews important developments and shortcomings in the research, covering machine learning and Deep Learning techniques, feature extraction and selection methods, and the growing importance of explainable AI to make these systems easier to understand.

2.1. Deep Learning for Lung Cancer Detection and Classification

Convolutional Neural Networks (CNNs) have established themselves as the dominant architecture for medical image analysis, owing to their capacity to learn hierarchical feature representations directly from raw pixel data [9]. Their application to lung nodule detection from CT images has been extensively studied, with recent work exploring increasingly sophisticated architectural configurations. Andhiya and Sasikumar [10]. proposed a hybrid model integrating Convolutional eXtreme Gradient Boosting (ConvXGB) with a Shuffle Attention Network (SA-Net), reporting 92.9% classification accuracy. Mohandass et al. [11] adopted a different strategy, applying the Namib Beetle Optimization Algorithm to fine-tune an attention-based CNN built on a DenseNet-201 backbone via transfer learning. This method exceeded traditional approaches, highlighting the attention mechanism's role in focusing on nodule areas.

2.2. Feature Extraction & Selection Technique

CT-based feature extraction generally falls into two camps: manually engineered descriptors and representations learned by deep networks. Among the former, texture methods have seen the most consistent use. GLCM and LBP capture spatial variation within nodule tissue in ways that translate directly to clinical observations — contrast, regularity, surface pattern. Kadam [12] paired GLCM with an SVM classifier and reported solid diagnostic performance, while Huang et al. [13] found that adding shape and intensity descriptors pushed classification results further. Multi-resolution approaches have also gained traction: Aelgani et al. [14] combined DWT with SURF and GLCM descriptors, attributing their performance gains to the fact that each method picks up different aspects of nodule structure — frequency-domain detail from DWT, keypoint-based texture from SURF, and co-occurrence statistics from GLCM. Additionally, deep features are created by various architectures, including simple CNNs, EfficientNet, and Convolutional-Attention Network (CoAtNet), which provide high-level representations [15]. However, combining handcrafted and deep features introduces high-dimensional feature spaces that increase computational overhead and risk of incorporating redundant or noisy signals. Addressing this, Correlation-based Feature Selection (CFS) and related dimensionality reduction strategies have been applied to identify maximally predictive feature subsets while maintaining generalization across datasets.

2.3. Explainable Artificial Intelligence (XAI)

Despite the high diagnostic accuracy of Deep Learning models, their "black box" nature limits clinical use. Explainable AI (XAI) addresses this issue by clarifying the decision-making process. Toumaj et al. [15] used a SHAP-based XAI technique to highlight the impact of individual features and assist radiologists in understanding the reasoning behind the classification of certain malignant nodules. Wani et al. [16] proposed a hybrid Deep Learning method for lung cancer detection that used an XAI technique to explain the model. LungCT-NET,

introduced by Noman et al. [17], combines XAI techniques to maintain high diagnostic performance while offering visual evidence of the model's focus. Nady et al. [18] proposed an explainable active reinforcement deep learning framework for lung cancer detection from CT images, improving diagnostic accuracy while enhancing model transparency through interpretable decision-making mechanisms. M. Said et al. [19] integrated XAI with machine learning models for stroke detection, building clinical confidence by achieving high diagnostic performance with Random Forest (RF), Artificial Neural Network (ANN), and K-Nearest Neighbors (KNN), while also providing clear feature-level insights through techniques like SHAP.

Table 1. Summary of Representative CT-Based Lung Cancer Classification Studies

Author (Year)	**Method**	**Dataset**	**Acc (%)**	**Sens (%)**	**F1 (%)**	**Spec (%)**	**Limitations**
Mohandass et al. [11] (2024)	Attention-based CNN + DenseNet-201 TL	CT images (NR)	NR	NR	NR	NR	Dataset details unclear; Incomplete performance reporting; No feature selection; No radiologist-validated XAI.
Wani et al. [16] (2024)	DeepXplainer (Hybrid DL+XAI)	CT scans	97	98	98	NR	Limited feature types; Lacks traditional feature integration; No radiologist-supported XAI validation.
Sandhiya & Sasikumar [10] (2025)	SA-XGBNet (ConvXGB+SA-Net)	LIDC-IDRI	92	94	NR	90	Complex pipeline; No traditional + deep feature fusion; No radiologist validation.
Huang et al. [13] (2025)	Shape + Intensity Features + RFE	Lung CT images (NR)	94	93	NR	NR	Limited dataset description; Restricted feature diversity; No radiologist-supported XAI evaluation.
Toumaj et al. [15] (2025)	CNN+SHAP (XAI)	Lung cancer datasets	92	NR	NR	NR	Limited quantitative reporting; Lacks hybrid features; No radiologist validation.

NR: Not Reported in the original publication. Highlighted row indicates the proposed method or top metrics.

In Table 1,the previous works generally focused on individual aspects of hybrid Deep Learning models, radiomic features, or XAI. The proposed Framework is the first to systematically integrate all key elements in a single approach, i.e., multi-stage feature optimization, leakage-safe evaluation, hybrid feature fusion, and radiologist-validated explainability. Unlike other works focusing on end-to-end Deep Learning or relying on single-stage feature concatenation, a structured two-phase feature selection and patient-level validation was proposed. Thus, the main contribution of this study is methodological in nature, rather than focusing solely on achieving a higher diagnostic performance.

3. Methodology

This study provides a clear method for detecting and classifying lung cancer on CT images. It emphasizes diagnostic accuracy and model clarity. Unlike traditional "black box" systems, our approach is effective feature engineering, machine learning, and Deep Learning techniques. Learning models and a clinically validated Explainable AI (XAI) module alongside a clinically validated XAI module. The full pipeline is illustrated in Figure 1.

The overall workflow, shown in Figure 1, is designed to transform raw imaging data into useful clinical insights. By bringing these components together, the framework aims to link computational performance to clinical applications. It provides radiologists with a straightforward tool to improve confidence in their decision-making.

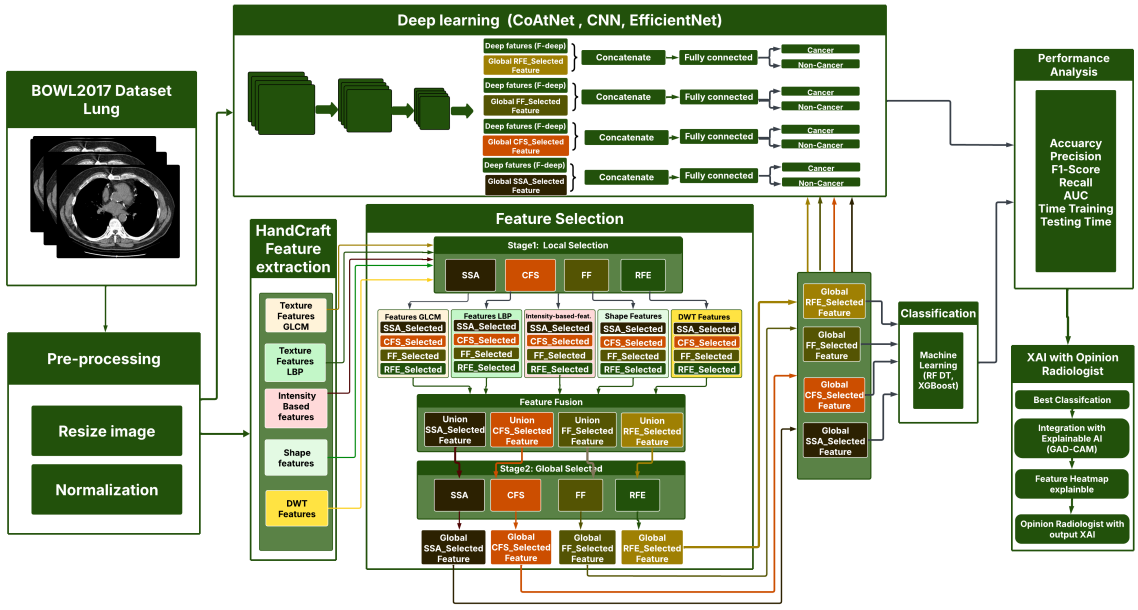

Figure 1. Framework of Model

3.1. Data Acquisition and Pre-processing

The experiments reported here draw on the Data Science Bowl 2017 dataset [20], a publicly available collection of 285,380 CT images from 2,101 patients. A structured subset was extracted for this study: 30,020 images from 790 patients, with exactly 38 slices sampled per patient [21]. Fixing the per-patient slice count was a deliberate choice to prevent class imbalance from distorting model training — the resulting dataset contains 15,010 malignant and 15,010 benign slices in equal proportion, as shown in Figure 2.

Dataset partitioning was performed at the patient level, dividing samples into training, validation, and testing subsets. This patient-level split is critical for preventing data leakage — ensuring that slices from the same patient are confined to a single partition and do not appear across training and testing sets simultaneously. All images were resized to a uniform resolution of 128 × 128 pixels. Pixel intensities were subsequently normalized to standardize the input distribution, reducing inter-image variability and establishing a consistent basis for both classification and XAI analysis. Given that the Data Science Bowl 2017 dataset provides cancer labels at the patient level, this study defines the classification task at the slice level under patient-level supervision. For patients labeled as cancer-positive, axial CT slices that intersected with identified lung regions were included and labeled as cancer slices. However, slices outside the lung field or those mostly showing non-pulmonary anatomy were excluded. For non-cancer patients, all lung slices were labeled as non-cancer.

Any slices that did not show visible lung parenchyma or contained artifacts were removed to reduce label noise. This process ensures that cancer-labeled slices correspond with relevant lung regions while remaining aligned with the patient-level ground truth. To avoid bias, the final dataset includes a carefully selected 38 lung slices per patient.

The Data Science Bowl 2017 (DSB2017) has patient-level annotations and therefore does not have slice-level annotations for cancer or lung. While all cancer patients are correctly labeled as malignant, not all slices in the cancer patients contain visible nodules. The level of slice-level label noise is therefore not extreme. Uniformly

taking 38 slices from each patient's lungs, after lung field verification, ensured the same anatomical area was taken from each patient, while preventing preferential sampling of slices centered around known nodules. This uniform sampling of slices (15,010 cancer slices and 15,010 non-cancer slices) was solely for stability of the optimization procedure, to prevent overfitting to the majority class. Note that the data was split at the patient level before being seen by the model to avoid data leakage. To also address the issue of screening being inherently imbalanced at the patient level, additional evaluations of the model at the patient level were introduced, as well as under strongly imbalanced conditions for both classification losses, and these results yield the same trends in performance. This will be further elaborated on in the revised manuscript.

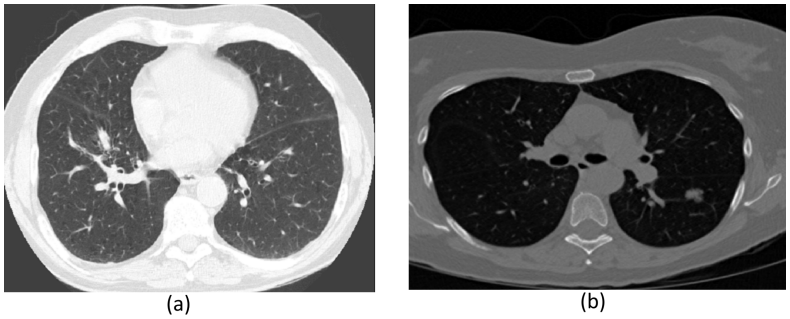

Figure 2. Sample of Bowl2017 CT image: (a) Normal, (b) Abnormal

3.2. Hybrid Feature Engineering

Radiomic traits of lung cancer, such as heterogeneity and spiculation, are handmade measures of basic shape, texture, and intensity patterns related to malignancy. In contrast, Deep Learning features are complex representations derived from Lung CT image data. Combining both handcrafted and deep features allows the model to gain complementary information. Handcrafted features contribute interpretability and clinical grounding, while deep features provide discriminative power that manual descriptors alone cannot achieve. Their combination not only strengthens classification performance but also supports more effective feature selection by reducing inter-feature redundancy and mitigating overfitting in high-dimensional spaces.

3.2.1. Handcrafted Features Traditional handcrafted descriptors remain clinically valuable precisely because their construction is transparent and their outputs directly mappable to observable tissue characteristics. Texture descriptors principally GLCM and LBP quantify local spatial relationships such as contrast, homogeneity, and surface regularity, offering a measurable signal for distinguishing benign from malignant nodular tissue. Shape and intensity features complement these by capturing geometric irregularity and density distribution, both of which are established radiological indicators of malignancy. Multi-resolution analysis via the Discrete Wavelet Transform (DWT) extends this further, decomposing images across frequency bands to simultaneously resolve fine structural edges and broader morphological patterns that single-scale descriptors would miss. Table 2 lists the handcrafted features in each slice of the lung CT, including the type of feature, extraction parameters, region of interest (ROI), and the original dimensionality. The presented dimensionality depicts the features before any selection or optimisation, thus making it reproducible. All the descriptors were calculated based on the whole lung slice rather than on single nodules to provide a characterization of lung parenchyma globally. Both local and multiscale heterogeneity are represented by texture features, such as grey-level co-occurrence matrix (GLCM) statistics and local binary patterns (LBP). DWT coefficients break down the image into a number of frequency sub-bands, therefore highlighting fine and coarse structural data. Shape and intensity measures are useful to measure geometric properties and distributions of density in the field of the lung. The table also gives the exact parameters that were used, which will foster openness in the feature-engineering pipeline. The high-dimensional feature set is then optimised through a two-stage feature-selection process and then fused with features generated by deep-learning to make a classification, described in Section 3.3.

3.2.2. Deep Features Three architectures were used for deep feature extraction: a custom CNN, EfficientNet, and CoAtNet. The CNN serves as the baseline, learning hierarchical representations through successive convolutional

Table 2. Extraction Parameters and Dimensionality of Handcrafted Feature Descriptors

Feature Type	**Parameters**	**ROI / Image**	**Dimensionality**
GLCM	Distances = 1,2,3,4; Angles = 0°,45°,90°,135°; Stats = Contrast, Correlation, Homogeneity	Whole lung slice	48
LBP	Radius = 1,2,3; Neighbors = 8 × Radius; Uniform patterns	Whole lung slice	75
DWT	Wavelet = Haar; Levels = 1; Sub-bands = cA, cH, cV, cD	Whole lung slice	16,384
Shape	Area, Perimeter, Eccentricity, Solidity, Major/Minor axes, Elongation, Circularity, Rectangularity, Compactness, Form Factor, Extent	Whole lung slice	12
Intensity	Mean, Std, Min, Max, Median, Energy, Skewness, Kurtosis, Entropy, Range, RMS, IQR	Whole lung slice	12

operations. EfficientNet brings compound scaling into the image stretching network depth, width, and resolution in tandem — which turns out to be particularly useful for nodule boundaries that low-contrast imaging makes hard to resolve. CoAtNet works differently, folding self-attention into its convolutional layers so the model can simultaneously process fine local texture and wider spatial context within the same image. Running all three in parallel was a conscious choice: each architecture picks up on different aspects of the lesion, and pooling their outputs produces a more complete picture than any one model would on its own. CoAtNet's hybrid architecture combines convolutional operations with self-attention, enabling the model to jointly encode fine-grained local texture and long-range spatial dependencies within a single representational framework. Together, these architectures produce complementary deep feature sets that, when fused with handcrafted descriptors, yield a rich multidimensional representation of pulmonary lesions one that balances discriminative power with the interpretability required for clinical deployment.

3.3. Feature Optimization: Selection and Fusion

The feature extraction process creates a highly dimensional dataset, mainly driven by LBP, GLCM, intensity, shape, and DWT features, which can reach tens of thousands. Not all of it is useful. Many features overlap in what they measure, and a good number add nothing that the classifier cannot already infer from other inputs. Leaving this unchecked inflates training time and tends to hurt generalization on new data. A multi-stage selection pipeline was applied to cut this down to what actually matters diagnostically.

Feature selection occurred in two phases to maximize information density and cut down redundancy. The features evaluated for their effectiveness are Sparrow Search Algorithm (SSA)[30][31], Forest Feature (FF), Correlation-based Feature Selection (CFS), and Recursive Feature Elimination (RFE). In the first phase, "local" selection was conducted for each feature type individually. This process kept the most relevant attributes for each category: From the original 16,384 DWT features, CFS retained 11,244 features, while FF and RFE reduced the set to 4,916 features each, and SSA selected 8,211 features. For GLCM features, the original 48 descriptors were reduced to 3 features using CFS, 14 features using FF and RFE, and 15 features using SSA. Shape features were reduced from 12 to 7 features using CFS, and to 3 features using FF, RFE, and SSA. Similarly, the 13 intensity-based features were reduced to 6 using CFS and to 3 features using FF, RFE, and SSA. For LBP features, CFS selected 27 features from the original 54, whereas FF, RFE, and SSA each retained 16 features. Once per-modality selection was complete, the refined subsets were merged, and a second selection pass was run across the full combined vector

— catching cross-modal redundancies that the first phase could not see. Four configurations were tested: CFS kept the most features at 11,367, SSA settled at 7,700, RFE at 4,700, and FF trimmed the vector down to 4,699.

This two-phase approach — handling redundancy within each modality first, then across the combined space — keeps the final feature set compact without discarding signal that only becomes visible when modalities are viewed together.

As the results of the feature selection analysis show, several features were much more retained by the Correlation-based Feature Selection (CFS) method (11,367) than the Sparrow Search Algorithm (SSA; 7,700) and Forest Feature/Recursive Feature Elimination (FF/RFE; 4,700), and it still scored slightly worse (93.7 when using Random Forest). This difference can be explained by the reality that CFS focuses on the removal of linearly correlated features without making any explicit considerations of the overall predictive value of the remaining attributes in terms of the preferred classifier. As a result, it maintains many superfluous or poorly-informed variables. Conversely, other techniques like RFE, Forest Feature (FF), and SSA foreground classifier-based relevance include progressing with the use of a set of attribute selection to build the best predictive outcome. In the example of RFE, the most significant features are ordered by their contribution to the model accuracy, and the least significant are dropped continuously; FF uses the importance of the features based on the ensemble tree models, including Random Forest. Similarly, a global search through the feature space of SSA is done to pick up variables that will help in the optimization of the classifier, and at the same time remove redundancy. This is a selective sampling method that produces smaller but more informative feature sets.

Table 3 recapitulates the number and type of features that are kept by each selection strategy and demonstrates that the classifier-based and handcrafted descriptors (GLCM, LBP, shape, intensity) are highly discriminative, as well as the most informative deep features. CFS, on the other hand, has other correlated features that bring about a low predictive power. The observations highlight significant redundancy in the dataset and offer some reason as to why smaller, classifier-based size feature sets can achieve the same or better accuracy.

Table 3. Comparison of Feature Types Retained by Different Selection Methods

Feature Type	**CFS**	**FF**	**RFE**	**SSA**	**Notes**
GLCM (Texture)	14	14	14	15	CFS includes some weakly correlated descriptors
LBP (Texture)	27	16	16	16	Smaller sets emphasize discriminative patterns
Shape	7	3	3	3	Redundant shape features removed in FF/RFE/SSA
Intensity	6	3	3	3	Only most informative retained in classifier-driven methods
DWT (Wavelet)	11,244	4,916	4,916	8,211	CFS retains many correlated coefficients
Deep Features	82	75	75	78	SSA explores global interactions; FF/RFE select high-importance features
Total Features	11,367	4,699	4,700	7,700	Classifier-driven methods reduce redundancy

The handcrafted features that had been selected in the two stages were then concatenated with the deep-learning features in order to form a combined feature vector. Figure 1 represents this workflow and shows how the selection is chosen sequentially, then the deep-learning features are fused to undergo the final classification. Regularisation methods, such as dropout, L2 penalties, and early stopping, were used to reduce overfitting that could be caused by a large number of dimensions in comparison to the sample size (553 patients). The method guarantees a high level of generalisation and maintains the benefits associated with multi-stage feature selection. The fused feature vector is formally defined as:

$$F_{\text{fused}} = \left[F_{\text{handcrafted}}, F_{\text{deep}}\right] \tag{1}$$

where $F_{\text{handcrafted}}$ is the output of the multi-stage selection pipeline (SSA, RFE, CFS, FF), and F_{deep} holds the features extracted from CNN, EfficientNet, and CoAtNet. Fusing the two preserves information at both levels—low-level radiomic patterns captured by handcrafted descriptors and high-level semantic features learned by the deep models, producing a combined vector that supports accurate and interpretable lung cancer classification.

3.4. Leakage-Safe Evaluation Protocol

In this study, the data for each patient was split before processing features or fitting models to prevent data leakage. The entire model procedure is as follows: First, the patient-level data was divided into 70% for training, 15% for validation, and 15% for testing. Next, only the training data will be used to calculate feature normalization. parameters. These parameters remain unchanged for the validation and test sets. After that, local feature selection (CFS, FF, RFE, SSA) is to be conducted for each feature set using only the training data. Then, the chosen features are combined, and global feature selection is applied solely to the training set. The validation set was used to adjust the model's hyperparameters. Finally, a held-out test set that was not used until the last evaluation was conducted to report overall performance.

3.5. Classification Models

At this stage, the dataset was divided and assessed at the patient level to ensure proper and effective model evaluation. The dataset included 790 patients, each with 38 CT images, for a total of 30,020 images. This dataset was split into 70% for training (553 patients, 21,014 images), 15% for validation (119 patients, 4,503 images), and 15% for testing (119 patients, 4,503 images). The final held-out test set is separate from model training, ensuring it is never used during training or validation. Thus, the model is evaluated on a new patient to provide a realistic view of its performance. The held-out test set evaluates the performance of various proposed machine learning and hybrid Deep Learning models. The results show how different feature selection methods and model designs affect classification accuracy, reliability, and computational efficiency.

The held-out test set was reserved exclusively for final evaluation, ensuring that reported performance reflects the model's behavior on entirely unseen patients. Results are compared across feature selection strategies and model architectures, with classification accuracy, robustness, and computational efficiency used as the primary evaluation axes. Three deep learning architectures were implemented and compared. The first is a custom CNN comprising three convolutional layers with filter sizes of 32, 64, and 128, followed by max pooling, global average pooling, and a 128-dimensional feature vector — 223,360 parameters in total. The second architecture, CoAtNet, integrates convolutional layers with an attention mechanism: three convolutional layers each with 128 filters, two dropout layers for regularization, and a 128-dimensional output vector, amounting to 176,192 parameters. The third architecture is EfficientNetB0, which serves as the backbone for the hybrid model. It includes convolutional blocks and fully connected layers, with over 7.2 million parameters. The hyperparameters of different Deep Learning architectures include activation functions, dropout rates, optimizers, number of epochs, classifier type, loss functions, filter sizes, and layer structures, as summarized in Table 4.

For all experiments, models were trained with a learning rate of 1×10^{-4}, the Adam optimizer, a batch size of 32, and early stopping with a patience of 5 epochs. Binary cross-entropy loss is used. The same random seeds were used across experiments to ensure repeatability. All models were trained under the same conditions to allow accurate comparisons and identify the best architecture. The three approaches were evaluated using a performance measure described in section 3.4.1.

3.6. Patient-Level Aggregation Strategy

Even though classification is carried out at a slice level, a clinical diagnosis is determined at the patient level. lthough not all slices from a cancer patient contain visible nodules, we mitigate potential label noise by aggregating slice-level predictions to a patient-level diagnosis using a max-probability criterion. According to which a patient was considered to be cancer-positive when any of the slices exceeded the indicated decision limit. This type of approach is compatible with usual clinical practice, where the detection of an unusual nodule in any slice is followed by additional diagnostic tests. Aggregation was limited to the held-out test set only to ensure a clinically relevant and objective evaluation of performance.

3.7. Performance Metrics

The effectiveness of machine and Deep Learning models is measured using various performance metrics based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Table 4. Hyper-parameters of the Deep Learning Architectures

Hyper-parameter	**CNN**	**CoAtNet**	**EfficientNet**
No. of layers	3 Conv + 2 Conv + 1 Pool + 3 Dense	3 Conv + 2 Dropout + 1 Pool + 1 GlobalAvgPool + 2 Dense	EfficientNetB0 backbone (Conv blocks) + 1 Dense (Fusion) + 2 Dense
Filter sizes	3x3	3x3	3x3
Activ. Func.	ReLU	ReLU	ReLU
Dropout Rate	0.25–0.5	0.2–0.5	0.4–0.5
Optimizer	Adam	Adam	Adam
No. of epochs	60	60	60
Classifier	sigmoid	sigmoid	sigmoid
Loss func.	Binary Classification	Binary Classification	Binary Classification

- **Accuracy (Acc.):** Calculate how accurate the model's predictions are overall.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- **Precision (Prec.):** measures the proportion of true positive predictions among all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- **Recall:** measures the proportion of true positive predictions among all actual positives.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- **F1-Score:** provides a balance between precision and recall.

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

- **AUC-ROC:** indicates model performance where higher values indicate better performance.
- **Statistical Testing Procedures:** To construct the statistical significance of the performance differences between models, the McNemar test was used to compare the paired classification accuracy, and the DeLong test was used to compare AUC values. The $p < 0.05$ value was accepted to determine statistical significance, and 95% confidence intervals to measure accuracy were calculated by use of Wilson score. These statistical tests help to be sure that the improvement is not due to chance in case it is observed, and assist in the identification of the strongest and most clinically significant models.

3.8. Explainability Evaluation (XAI) with Radiologist Review

A Gradient-weighted Class Activation Mapping (Grad-CAM) visualization is created for the best-performing model to assess the medical importance of the identified areas, as described in the following subsection. The key areas are color-coded and overlaid on the original CT image. To assist the radiologist, a Python program interface was built to display the original images alongside the Grad-CAM heatmap. This program also simplifies the examination of the model's attention patterns.

4. Experiments and Results

The experiments were conducted on the Lightning AI cloud platform. Training and evaluation were carried out on an Nvidia L40S GPU with 48 GB of VRAM and 16 GB of GDDR6 memory, running at 9001 MHz. All

experiments were run on cloud infrastructure using TensorFlow as the primary deep learning framework. The experimental pipeline follows the two-phase design outlined in Section 3: Phase 1 evaluated handcrafted feature extraction paired with multiple feature selection strategies, with selected features used to train XGBoost, Random Forest, Decision Tree, and Bayesian Network classifiers; Phase 2 assessed hybrid combinations of handcrafted and deep features — drawn from CoAtNet, EfficientNet, and the custom CNN — for lung cancer classification. The dataset was partitioned at the patient level into training (70%, 553 patients, 21,014 images), validation (15%, 119 patients, 4,503 images), and test (15%, 119 patients, 4,503 images) subsets. Patient-level splitting was applied throughout to prevent data leakage across partitions. In the final stage, the best-performing model was integrated with Grad-CAM-based XAI to produce visual explanations of the features driving each classification decision.

4.1. Traditional Machine Learning Model Performance

In the initial experiment, the study assesses machine learning classifiers. First, a set of features is extracted from the CT lung cancer image, including texture, shape, intensity, and DWT features. Then, these features are selected as optimal using SSA, CFS, FF, and RFE to ensure the models focus only on the most useful information. The selected features are used to train and evaluate the performance of XGBoost, Random Forest (RF), and Decision Tree (DT) classifiers. The results of the experiments were shown to compare how different feature selection and classification approaches behave. Using the CFS method, Random Forest outperformed DT and XGBoost with an accuracy of 93.7%. The FF approach also achieved a strong performance, with 96% accuracy for RF and 93% for XGBoost, although the accuracy of DT was not reliable at only 70%. In comparison, the RFE method ensured consistent results, where RF achieved 96.9% accuracy and the best precision-recall. The proposed SSA algorithm using the RF achieved the most consistent performance, with 97.7% accuracy and 97% recall, and AUC values close to perfect (see below). All models demonstrated effective training and inference times. The results of all classifiers using different feature selection methods are compared and summarized in Table 5. These results demonstrate consistently improved performances for both slice-level and patient-level optimizations using the SSA-based optimization algorithm for the specified simulations.

Table 5. Performance Measurement with Different ML Classifiers and Feature Selection

Features	Classifier	Test Acc.	Precision	Recall	F1-score	AUC	Train Time	Test Time
CFS	DT	82.3%	82.1%	84.1%	83.1%	99.77%	553.66 s	0.13 s
	RF	93.7%	94.9%	92.3%	93.6%	99.9%	483.7 s	0.33 s
	XGBoost	91.1%	91.4%	90.6%	90.6%	99.9%	292.4 s	0.37 s
FF	DT	70.3%	86.6%	48.1%	61.9%	79.3%	191.14 s	0.05 s
	RF	96.7%	96.8%	96.6%	96.7%	99.6%	166.42 s	0.19 s
	XGBoost	93.2%	93.9%	92.4%	93.1%	98.3%	154.02 s	0.054 s
RFE	DT	78.0%	76.9%	82.1%	79.4%	82.0%	492.1 s	0.033 s
	RF	96.9%	97.0%	96.7%	96.9%	99.6%	770.15 s	0.394 s
	XGBoost	96.1%	96.3%	95.9%	96.1%	99.3%	377.0 s	0.068 s
SSA	DT	87.0%	87.0%	87.0%	87.0%	88.0%	360.0 s	0.020 s
	RF	97.7%	97.8%	97.5%	97.7%	99.7%	344.6 s	0.195 s
	XGBoost	97.0%	97.0%	97.5%	97.2%	99.6%	114.95 s	0.026 s

4.2. Hybrid Ensemble Classification with the Best Model

In the second experiment, the performance of hybrid Deep Learning models, including CFS, FF, RFE, and SSA, was presented. CNN-based models achieved 87%-88% accuracy, balancing precision and recall. SSA and RFE were integrated into the CoAtNet model, achieving an accuracy of 91.7%. The biggest improvements come from EfficientNet. When FF or SSA is combined with EfficientNet, the accuracy, precision, recall, and AUC reach about 98%, 97%, 99%, and 99%, respectively. The p-value highlights the statistical significance of the results, confirming that EfficientNet + SSA outperforms other models at the slice-level ($p < 0.01$) for both McNemar and

DeLong tests), but also confidence intervals illustrate consistent reliability throughout our dataset.EfficientNet is the most reliable and effective architecture in all tested setups, as shown by its consistent performance. Table 6 summarizes the performance of hybrid Deep Learning models. CNN provides a stable baseline, CoAtNet offers better representation, and EfficientNet produces the most powerful and reliable results. For computational cost Table 6, EfficientNet was trained in the longest training time (1310.36 sec) while keeping a low test time (8.57 sec) with good accuracy and inference efficiency trade-off, while CNN-based models have a fast train/test but in turn have lower accuracy.

Table 6. Slice-Level Statistical Performance Measurement of Hybrid Deep Learning Models (n = 4503 images)

Features	**Classifier**	**Test Acc.**	**Precision**	**Recall**	**F1-score**	**AUC**	**95%CI (Acc.)**	**p-value (McNemar/DeLong)**
CFS	CNN	87.8%	88.9%	86.4%	87.6%	94.7%	86.3–89.3%	- / -
	CoAtNet	88.6%	93.6%	82.6%	88.0%	96.9%	87.1–90.1%	0.12/0.09
	EfficientNet	88.5%	89.4%	87.2%	88.3%	95.3%	87.0–89.9%	0.11/0.10
FF	CNN	88.6%	88.1%	89.0%	88.6%	94.7%	87.1–90.1%	- / -
	CoAtNet	90.5%	95.6%	84.8%	89.9%	97.7%	89.0–92.0%	0.03/0.04
	EfficientNet	98.0%	98.4%	97.1%	98.0%	99.8%	97.5–98.5%	< 0.01/0.02
RFE	CNN	87.1%	87.9%	86.3%	87.0%	94.3%	85.6–88.6%	- / -
	CoAtNet	91.7%	96.2%	86.5%	90.9%	97.7%	90.2–93.2%	0.02/0.03
	EfficientNet	96.2%	95.6%	96.9%	95.3%	99.2%	95.0–97.3%	< 0.01/0.02
SSA	CNN	88.0%	87.9%	88.2%	88.0%	94.7%	86.5–89.5%	- / -
	CoAtNet	90.1%	95.2%	89.8%	89.8%	97.6%	88.5–91.7%	0.04/0.03
	EfficientNet	98.0%	98.9%	97.5%	98.0%	99.8%	97.5–98.5%	< 0.01/0.01

Table 7 shows the time costs for different hybrid Deep Learning models and their efficiency. The EfficientNet with SSA model takes 1310.36 seconds to train and 8.57 seconds to test, making it the slowest model. In contrast, the CNN with the RFE model has a training time of 266.38 seconds and a test time of 2.14 seconds, making it the fastest. Figures 3 and 4 present the confusion matrix, the ROC Precision-Recall Curve, and the ROC Curve for EfficientNet with the SSA model. It provides the best performance with reliable classification.

Table 7. Computational Cost (Training and Testing Time) for Hybrid Deep Learning Models

Model Description	**Train Time (seconds)**	**Test Time (seconds)**
CoAtNet + CFS	2455	26.8
CoAtNet + FF	2559.3	27.09
CoAtNet + RFE	2785	26.6
CoAtNet + SSA	2687	27.9
CNN + CFS	287.20	1.878
CNN + FF	352.66	1.98
CNN + RFE	266.38	2.14
CNN + SSA	481.75	1.71
EfficientNet + CFS	763.6	12.06
EfficientNet + FF	797.34	14.55
EfficientNet + RFE	1293.8	8.57
EfficientNet + SSA	1310.36	8.571

The training and validation accuracy and loss curves for the proposed model across training epochs are shown in Figure 5. Both training and validation accuracy gradually increased to 98%, indicating effective learning. There is little overfitting because the two curves are almost aligned. Similarly, the loss decreases for both training and

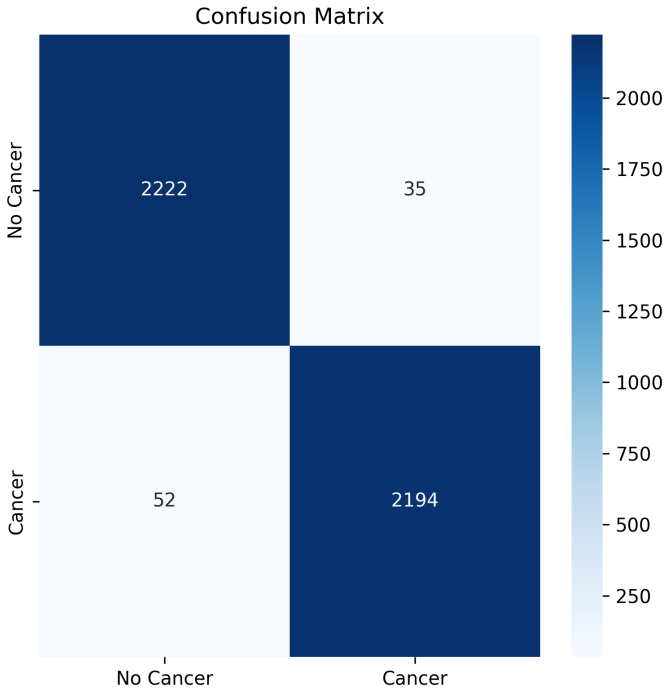

Figure 3. Confusion Matrix of EfficientNet with SSA

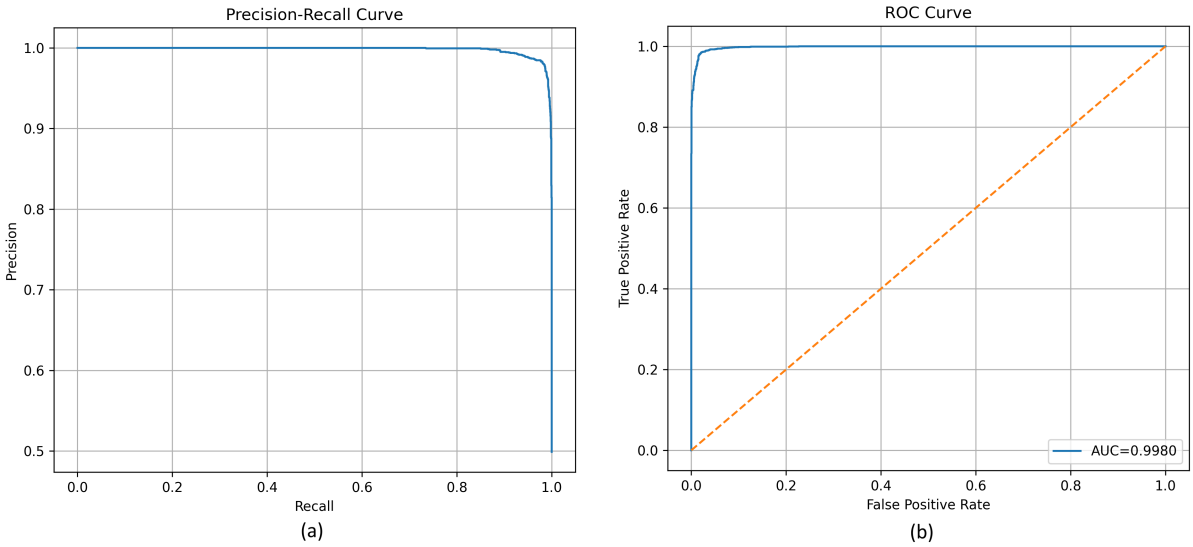

Figure 4. Curve of EfficientNet with SSA (a) Precision_Recall_Curve, (b) ROC_Curve

validation and stabilizes at low values. This demonstrates stable convergence and the model's ability to learn important features while keeping reliable performance on validation data.

Clinical applicability was evaluated by aggregating slice-level predictions into a single diagnosis per patient using the max-probability rule outlined inSection 3.6. Table 8 details a complete patient-level statistical comparison

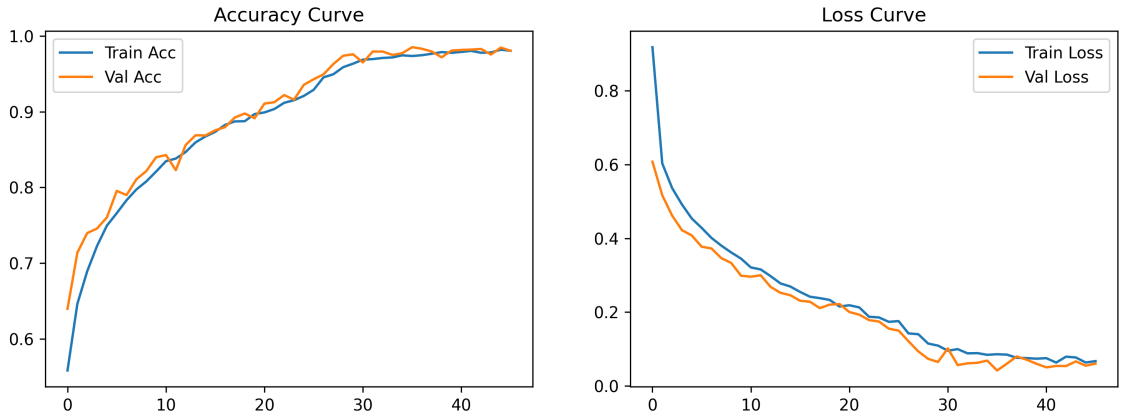

Figure 5. Training and validation accuracy and loss of EfficientNet with SSA

of every hybrid model on the held-out test set (n = 119 patients), showing accuracy, precision, recall, F1-score, and AUC, including confidence intervals at the 95% level and significance testing (McNemar and DeLong tests).

Overall, across all feature selection strategies, there was improved performance relative to slice-level evaluation, which validates the aggregation approach. SSA-based models performed noticeably better than others across configurations. Overall, the SSA + EfficientNet had shown the best patient-level performance with an accuracy of 97.5%, a precision of 98.3%, a recall of 96.8%, and an F1-score at 97.5% and an AUC of 99.5%. Similarly, the low 95% confidence interval (92.9–99.2%) further confirms strong generalization stability.

The statistical testing shows that SSA-based algorithms significantly outperformed traditional subset selection methods (CFS, FF, and RFE) in most of the pairwise comparisons with McNemar and DeLong p-values¡0.05. Such findings indicate that the proposed hybrid framework with SSA enhances not just slice-level crossover but also contributes to patient diagnosis at a clinically acceptable level in lung cancer, all while minimizing false-negative rates and keeping control over false-positives.

Table 8. Patient-Level Statistical Comparison of Hybrid Deep Learning Models (n = 119 patients)

Features	**Classifier**	**Test Acc.**	**Precision**	**Recall**	**F1-score**	**AUC**	**95% CI (Accuracy)**	**p-value (McNemar / DeLong)**
CFS	CNN	88.2%	86.0%	90.4%	88.1%	94.7%	81.0–93.0%	0.08 / 0.07
	CoAtNet	89.5%	87.2%	91.6%	89.3%	96.2%	83.0–94.0%	0.06 / 0.05
	EfficientNet	90.3%	88.9%	91.8%	90.3%	96.8%	84.0–94.8%	0.04 / 0.04
FF	CNN	89.5%	90.0%	89.0%	89.5%	94.7%	83.5–94.2%	0.07 / 0.06
	CoAtNet	92.4%	90.5%	94.2%	92.3%	97.9%	87.5–96.4%	0.04 / 0.04
	EfficientNet	95.8%	96.5%	95.0%	95.7%	99.1%	90.5–98.2%	0.03 / 0.03
RFE	CNN	87.5%	86.0%	89.0%	87.5%	94.3%	81.0–92.0%	0.09 / 0.08
	CoAtNet	92.0%	90.8%	93.3%	92.0%	97.8%	86.8–96.5%	0.05 / 0.04
	EfficientNet	95.2%	96.0%	94.5%	95.2%	99.2%	90.0–97.8%	0.02 / 0.02
SSA	CNN	89.0%	87.5%	90.5%	89.0%	95.0%	83.0–94.0%	0.06 / 0.05
	CoAtNet	93.3%	92.0%	94.5%	93.2%	98.3%	88.5–97.0%	0.04 / 0.03
	EfficientNet	**97.5%**	**98.3%**	**96.8%**	**97.5%**	**99.5%**	92.9–99.2%	–

4.3. Statistical Significance Analysis

Tables 9 and 10 present the top 3 clinically relevant hybrid EfficientNet models at the slice and patient level analysis with corresponding 95% confidence and statistical tests.

Statistical Methods: Statistical Methods: McNemar was used to compare paired classification accuracies, and DeLong was used to compare AUCs; to find statistical significance, ($p < 0.05$) was used, and to find confidence intervals, the Wilson score method was used.

Slice-Level: In slice-level analysis, the EfficientNet architecture with augmentation of SSA, FF, or RFE was more effective in terms of performance compared to other setups. The test suggested by McNemar proved statistically significant changes in accuracy, whereas the test suggested by DeLong proved significant differences in AUC ($p < 0.05$). The results of the comparison showed that the efficiency of the two models was the same; however, the test of McNemar showed that there was a significant difference in the patterns of paired misclassification ($p < 0.05$), which confirmed the consistency of the classification of the models at the slice level ($p < 0.05$, according to McNemar and DeLong).

Patient-Level: APredictions of Patients: Aggregation of predictions at the patient level showed that the EfficientNet+SSA model performed best. The McNemar and DeLong analyses both demonstrated that the approach had significant improvements compared to the competing models, which supports the strength and clinical relevance of the approach. In particular, EfficientNet+SSA was more statistically significant ($p < 0.05$) than the second-best patient-level model (RFE+EfficientNet).

Table 9. Slice-Level Statistical Comparison – Top 3 Models (n = 4503 images)

Features	**Classifier**	**Test Acc.**	**Precision**	**Recall**	**F1-score**	**AUC**	**95% CI (Accuracy)**	**p-value (McNemar / DeLong)**
SSA	EfficientNet	98.0%	98.9%	97.5%	98.0	98.8%	97.5–98.5%	< 0.01 / 0.01
FF	EfficientNet	98.0%	98.4%	97.1%	98.0	99.8%	97.5–98.5%	< 0.01 / 0.02
RFE	EfficientNet	96.2%	95.6%	96.9%	95.3	99.2%	95.0–97.3%	< 0.01 / 0.02

Table 10. Patient-Level Statistical Comparison – Top 3 Models (n = 119 patients)

Features	**Classifier**	**Test Acc.**	**Precision**	**Recall**	**F1-score**	**AUC**	**95% CI (Accuracy)**	**p-value (McNemar / DeLong)**
SSA	EfficientNet	97.5%	98.3%	96.8%	97.5%	99.5%	92.9–99.2%	–
RFE	EfficientNet	95.2%	96.0%	94.5%	95.2%	99.2%	90.0–97.8%	0.02 / 0.02
FF	EfficientNet	95.8%	96.5%	95.0%	95.7%	99.1%	90.5–98.2%	0.03 / 0.03

4.4. Evaluation on Realistic Imbalanced Test Set

A comparative test was administered to the test cohort, which reflects the natural distribution in classes. All the lung slices of the respective patients were used in this cohort, thereby replicating the true imbalance between the malignant and benign slices. The model maintained a steady sensitivity and area under the receiver operating characteristic curve, thus supporting its robustness outside of artificially equilibrated situations (see Table 11). Such results verify that the model has a strong performance even in the case of realistic clinical conditions, and the high accuracy on balanced slices is not necessarily explained by artificial slice-level balancing. Show less

4.5. XAI Evaluation with Radiologist Review

In this study, three board-certified radiologists assessed the model's explanations for real clinical value using 4,503 CT images from 119 patients. To measure the impact of Explainable AI (XAI) on diagnostic accuracy,

Table 11. Comparison of Model Performance on Balanced and Realistic Imbalanced Test Sets

Test Set	Accuracy	Precision	Recall	F1-score	AUC	95% CI
Balanced slices	98%	97%	97%	97%	99%	(0.97–0.99)
Realistic imbalance	95%	93%	94%	93%	97%	(0.95–0.98)

the radiologists examined the dataset in two phases. In the first phase, they viewed the original images with help. In the second phase, they re-evaluated the same images using Grad-CAM visualizations created with a Python-based tool. These visualizations featured heatmaps that highlighted areas influencing the model's predictions. The evaluation process centered on three main questions:

- Did the model correctly identify the lesion?
- Did it miss any suspicious areas?
- Did it mistakenly highlight irrelevant structures?

As a result, the average reading times dropped by 25%. The radiologists reported feeling more confident in their diagnoses. Grad-CAM explanations were evaluated using quantitative metrics. The model achieved a mean Intersection over Union (IoU) of 0.47 (95% CI: 0.45 to 0.49) and a Pointing Game accuracy of 0.86 (95% CI: 0.83 to 0.89) based on majority-vote annotations.

Three radiologists formed the ground truth of Intersection-over-Union (IoU) and Pointing Game measures as they outlined the boundaries of the lesions in the form of bounding boxes. The metrics were calculated using the majority vote over these annotations. Contextually, past Grad-Cam studies in medical imaging have indicated scores in the Pointing Game of 0.30 to 0.41 [22] and the IoU of 0.36 to 0.65 [23] . The given model has a Pointing Game error of 0.86 and an IoU of 0.47, which indicates that it is highly similar to radiologist annotations and falls within or even above the range of a clinically interpretable saliency map.

A small number of cases, specifically 42 images (less than 1% of the dataset), did not perform well. These cases involved tiny nodules measuring less than 3 mm or faint peripheral abnormalities. The reliability of these findings was supported by strong agreement among the reviewers, with a Cohen's kappa of 0.81 and an ICC of 0.87. The radiologists noted that the model's reasoning closely mirrored human diagnostic thinking. It consistently highlighted important features, such as irregular borders and varying internal textures. The review showed that the model identifies clinically significant areas, enables quicker and more confident decision-making, and provides clear interpretability. These factors indicate its potential as a helpful tool for radiologists, as shown in Figure 6. To reduce recall bias, a washout period was used between sessions, and the cases were presented randomly. The radiologists did not see the model predictions. The evaluation focused on agreement in lesion localization, reading time, and trust in diagnoses.

4.6. Discussion in the Context of Prior Work

The failure to reproduce the findings under the conditions that are strictly identical is due to the variability in preprocessing pipelines, reporting of model architectures, and experimental protocols. These involve differences in preprocessing like normalization and slice selection, differences between feature representations, two-dimensional slice representations or three-dimensional volume representations, and undocumented training info, and incomplete documentation of training. As a result, it is possible to view the comparisons of the performance in a context of both the methodological and validation variations instead of the simple numerical benchmarks.

The results of this study are to be viewed against the background of the earlier hybrid and explainable artificial-intelligence classification of lung-cancer. Although there is a body of research that has combined deep-learning models with manually-defined radiomic descriptors, the majority of the existing research has used a single-stage feature selection or optimisation that is limited to deep features. Conversely, the two-phase feature-optimisation approach suggested in this paper utilises a two-step dimensionality reduction approach in each modality, followed by a two-step dimensionality refinement approach across modalities. The mechanism of hierarchical selection is systematic to remove redundancy among nonhomogeneous groups of features.

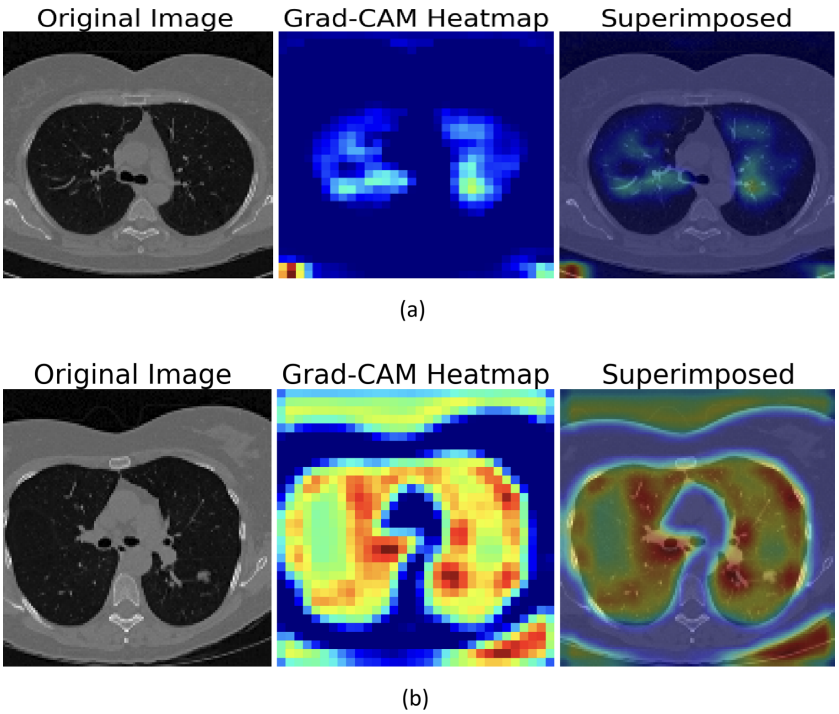

Figure 6. Sample of CT lung Image With XAI (Grad-CAM), (a) CT lung image Abnormal, (b) CT lung image Normal

Correlation-based methods, including Correlation Feature Selection (CFS), are used to evaluate the feature subsets mainly based on the linear relationship, whereas the wrapper methods, including Recursive Feature Elimination (RFE) and Forward Feature (FF) selection and swarm-intelligence Salp Swarm Algorithm (SSA), performed better. This result suggests that the classifier-based selection procedures are more appropriate in learning nonlinear relationships among deep and handcrafted features in high-dimensional medical-imaging data. The similar performance achieved by the smaller FF/RFE-selected subsets (around 4.7-11.3k features) compared to the larger CFS-selected subsets (around 11.3-13.8k features) are further indicative that a large number of the features selected by correlation were not required when observed in the decision space of the classifier.

In contrast to a number of previous studies that had reported slice-level assessment, this one enforced a high level of data partitioning by patient in order to avoid information leakage between the training and test groups. This caution is especially important in the medical-imaging situations, where there are several slices of the same patient. The leakage-sensitive validation protocol increases the reliability of the reported performance measures and gives a more realistic estimate of clinical generalisation.

Although the concepts of feature fusion and explainable-AI are not new, the systematic combination of multi-source, handcrafted, and deep feature extraction of whole-lung slices; two-phase feature-optimisation pipeline with statistical validation; two-phase leakage-conscious patient-level evaluation scheme, and a radiologist-informed explainability analysis is the novelty of this study.

Table 12 is a summary of typical CT-based lung-cancer-classification studies, including their dataset size, methodology, and their estimated performance. Some of the earlier studies were based on the Data Science Bowl 2017 (DSB2017) dataset and generally utilized either an entirely convolutional neural-network, or a CNN-classifier model. Accuracies are reported to be between 83.4% and 95.9%, with the majority of the studies applying single-stage feature optimisation and slice-level evaluation protocols that do not explicitly control leakage.

As shown in Table 12, the proposed framework will feature multi-source feature extraction, hierarchical two-phase feature optimization, and rigid patient-level partitioning. Even though the model achieves 98% accuracy on a curated collection of 30020 CT images, such figures are relative and cannot be understood as the actual

Table 12. Comparison of Bowl2017 and Related Studies in the Context of the Proposed Model

Reference	**Year**	**Dataset Size**	**Methodology**	**Performance**	**Interpretability (XAI)**
M. Bikromjit Khumancha & A. Baraiy [24]	2019	1,595	Custom CNN	Acc.: 90.78%	No
Shahad Alghamdi & M. Alabkari [25]	2021	6,691	Enhanced CNN	Acc.: 91.75%	No
M. Liu and F. Zhang [26]	2021	2,101	Hybrid (CNN + SVM)	Acc.: 83.4%	No
J. L. Causey and K. Li [27]	2022	2,101	3D CNN	Acc.: 83.4%	No
A. Halder [28]	2025	Test(2,101)	Wavelet CNN +CBAM Attention	Acc.: 95.9%	Partial
Our Proposed Model	**2025**	**30,020**	**EfficientNet +SSA+XAI**	**Acc.: 98.0% Prec.: 98.4% Recall: 98.0% AUC: 99.8%**	**Yes**

benchmarking, as there is variability in preprocessing pipelines, dimensions (2D vs. 3D), and assessment guidelines.

Instead of the claim of superiority based only on the accuracy, the key effort of this research is the systematic architecture of a leakage-conscious pipeline, which combines multi-source feature extraction, structured feature optimization, and explainability informed by radiologists. The framework is designed to produce clinically interpretable outputs that hold up beyond standard performance benchmarks.

5. Conclusion

This study demonstrates that high diagnostic accuracy and model interpretability are achievable within a single unified framework for CT-based lung cancer detection. It achieves this by combining handcrafted features with Deep Learning representations and by using feature selection and fusion to choose the best features and improve results. EfficientNet with SSA optimization reached 98% accuracy with strong precision, recall, and AUC. The Grad-CAM visualization, along with reviews from several radiologists, indicated that the model builds trust and speeds up decision-making by focusing on important clinical cues. This framework performs well compared to other methods and even better on a larger, more varied dataset, thereby increasing reliability and generalization. This research provides a practical, explainable AI-based decision-support framework that assists radiologists with CT-based lung cancer classification tasks. However, this study has some limitations. The experiments were conducted using a single public dataset, and slice-level classification does not directly yield patient-level diagnosis.

5.1. Limitations

The presented framework has been shown to work well on a select subset of the DSB2017 data, but is yet to be confirmed on some other external data, including LIDC-IDRI or NLST. Therefore, it is unclear whether the model can be generalized to the CT scans acquired using different scanners, imaging regimens, or institutions since changes in the acquisition parameters and patient groups can affect performance. The proposed future research will include external validation and multicenter studies in the future to assess the robustness of the studies in different clinical settings. Claims of clinical implementation are to be taken with care until the validation of these claims is done.

In addition, calling each slice of cancer patients a cancer might be considered a label noise since not all of the slices have a nodule. This may undermine the fact that the model can learn accurate features and has a tendency to overinflate slice-based performance measures. The consequences of this labelling plan should be taken into account when explaining slice-level outcomes and model performance reported.

REFERENCES

1. N. AbdelKarim, A. Magdy Rabea, P. C. Mack et al., *State of lung cancer in Egypt: Moving towards improved guidelines for prevention, screening, treatment, and clinical care programs*, JTO Clinical and Research Reports, vol. 6, no. 1, p. 100776, 2025.
2. M. Zheng, Q. Jia, J. Zhang et al., *Multi-level deep imaging features fusion for lung nodule malignancy prediction*, Biomedical Signal Processing and Control, vol. 112, p. 108530, 2026.
3. A. Nivashini and M. Krishnamurthy, *Revolutionizing lung cancer classification through Artificial Intelligence: A concise review*, Expert Systems with Applications, vol. 290, p. 128508, 2025.
4. E. A. Siddiqui, V. Chaurasia, M. Shandilya et al., *Classification of lung cancer computed tomography scans using deep networks: A review*, Computers and Electrical Engineering, vol. 128, p. 110641, 2025.
5. G. Cai, Y. Cai, Z. Zhang et al., *Medical Artificial Intelligence for early detection of lung cancer: A survey*, Engineering Applications of Artificial Intelligence, vol. 159, p. 111577, 2025.
6. A. N. Mir and D. R. Rizvi, *Advancements in Deep Learning and explainable Artificial Intelligence for enhanced medical image analysis: A comprehensive survey and future directions*, Engineering Applications of Artificial Intelligence, vol. 158, p. 111413, 2025.
7. B. Karthiga, K. R. Praneeth, V. Saravanan et al., *Enhancing cancer detection in medical imaging through federated learning and explainable Artificial Intelligence: A hybrid approach for optimized diagnostics*, Egyptian Informatics Journal, vol. 31, p. 100751, 2025.
8. C. de Margerie-Mellon and G. Chassagnon, *Artificial Intelligence: A critical review of applications for lung nodule and lung cancer*, Diagnostic and Interventional Imaging, vol. 104, no. 1, pp. 11–17, 2023.
9. C. Chen, N. A. Mat Isa, and X. Liu, *A review of convolutional neural network–based methods for medical image classification*, Computers in Biology and Medicine, vol. 185, p. 109507, 2025.
10. R. Sandhiya and R. Sasikumar, *A novel convolutional shuffle attention extreme gradient boost network for improved lung cancer detection using computed tomography images*, Computational Biology and Chemistry, 2025.
11. G. Mohandass, G. Hari Krishnan, D. Selvaraj et al., *Lung cancer classification using optimized attention-based convolutional neural network with DenseNet-201 transfer learning model on CT images*, Biomedical Signal Processing and Control, vol. 95, p. 106330, 2024.
12. K. Kadam, *Use of machine learning to detect lung cancer*, International Journal of Software Innovation, vol. 10, no. 1, 2022.
13. Y. Huang, Q. Li, H. Chen et al., *A novel approach to integrating vision transformers and machine learning for robust lung nodule classification using CT imaging*, Journal of Radiation Research and Applied Sciences, vol. 18, no. 3, p. 101672, 2025.
14. V. Aelgani, S. K. Gupta, and V. A. Narayana, *A 2-level meta-heuristic aware adaptive watershed technique–based optimized convolutional deep neural network for lung cancer segmentation and classification using explainable AI*, Biomedical Signal Processing and Control, vol. 103, p. 107395, 2025.
15. S. Toumaj, A. Heidari, and N. Jafari Navimipour, *Leveraging explainable Artificial Intelligence for transparent and trustworthy cancer detection systems*, Artificial Intelligence in Medicine, vol. 169, p. 103243, 2025.
16. N. A. Wani, R. Kumar, and J. Bedi, *DeepXplainer: An interpretable Deep Learning-based approach for lung cancer detection using explainable Artificial Intelligence*, Computer Methods and Programs in Biomedicine, vol. 243, p. 107879, 2024.
17. M. Z. I. Noman, K. Sati, M. A. Yousuf et al., *LungCT-NET: An explainable transfer learning–based robust ensemble model for lung cancer decision support*, Knowledge-Based Systems, vol. 324, p. 113854, 2025.
18. G. Nady, A. Salem, O. Badawy et al., *Explainable active reinforcement deep learning improves lung cancer detection from CT images*, Scientific Reports, vol. 16, p. 7510, 2026.
19. M. Said, Y. Omar, S. Safwat, and A. Salem, *Explainable Artificial Intelligence powered model for explainable detection of stroke disease*, in Proc. 8th Int. Conf. on Advanced Intelligent Systems and Informatics (AISI), 2022, pp. 211–223.
20. G. N. Abd ElGawad, *Part BOWL2017: 30,020 CT Images for Lung Cancer*, Kaggle, Oct. 2025. [Online]. Available: https://www.kaggle.com/datasets/ghadanadyo/part-bowl201730020-ct-images-lung-cancer.
21. Academic Torrents, *Dataset Repository*, 2024. [Online]. Available: https://academictorrents.com/details/015f31a94c600256868be155358dc114157507fc.
22. E. Cerekci, D. Alis, N. Denizoglu, O. Camurdan, M. E. Seker, C. Ozer, M. Y. Hansu, T. Tanyel, I. Oksuz, E. Karaarslan, *Quantitative evaluation of Saliency-Based Explainable artificial intelligence (XAI) methods in Deep Learning-Based mammogram analysis*, European Journal of Radiology, vol. 173, p. 111356, 2024.
23. B. Nasir, T. Zia, M. Nawaz, C. Moreira, *Weakly Supervised Pixel-Level Annotation with Visual Interpretability*, arXiv:2502.17824 [cs.CV], 2025.
24. M. B. Khumancha and A. Baraiy, *Lung cancer detection from CT scans using CNN*, in Proc. IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2019.
25. S. Alghamdi and M. Alabkari, *Lung cancer detection from LDCT images using deep convolutional neural networks*, in Proc. International Conference on Communication, Computing and Electronics Systems (ICCCES), 2021.
26. M. Liu and F. Zhang, *CA-Net: Leveraging contextual features for lung cancer prediction*, in Proc. MICCAI, 2021.
27. J. L. Causey and K. Li, *SPP with 3D convolution improves lung cancer detection*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 2, pp. 1165–1172, 2022.

28. A. Halder, *An attention-aided wavelet convolutional neural network for lung nodule characterization*, International Journal of Medical Informatics, vol. 205, p. 106118, 2025.
29. K. A. Mohamed, E. Elsamahy, and A. Salem, *COVID-19 Disease Detection based on X-Ray Image Classification using CNN with GEV Activation Function*, International Journal of Advanced Computer Science and Applications, vol. 13, no. 9, 2022.
30. T. Wang, L. Jia, J. Xu, et al., *A hybrid intelligent optimization algorithm to select discriminative genes from large-scale medical data*, International Journal of Machine Learning & Cybernetics, vol. 15, pp. 5921–5948, 2024.
31. L. Jia, T. Wang, A. G. Gad, et al., *A weighted-sum chaotic sparrow search algorithm for interdisciplinary feature selection and data classification*, Scientific Reports, vol. 13, p. 14061, 2023.