



Enhancing lung cancer screening accuracy through a dual-pipeline deep learning and machine learning framework for integrated risk assessment

Patthanan Uthaitipitak*

Ruamrudee International School, Bangkok 10510, THAILAND

*Corresponding author: patthanan.utha@gmail.com

ABSTRACT

Lung cancer is a leading cause of death worldwide, with delayed or misdiagnosis further hindering professionals' ability to lower lung cancer-related mortality rates. Misdiagnosis could look like false positive results, where benign lung diseases are mistaken for malignant lung carcinomas, or more life-threatening false negative results, where harmful eosinophilic pneumonia is misinterpreted as a benign lung hamartoma. The objective of this study consists of the following goals: (1) to develop CNN-based classification models for lung nodule analysis, (2) to compare the performance of multiple deep learning models under consistent controlled environments, (3) to evaluate and present the interpretability of the models through a clinically grounded probability equation, and (4) to propose an integrated risk-scoring framework that combines the imaging and clinical pipelines into a single interpretable measure of lung cancer risk. However this framework is only theoretical as imaging and clinical models were developed using separate datasets; future studies should note that a practical development will require validation on a unified multimodal patient cohort. In regard to the creation of the models, the deep learning convolutional neural network, which utilized a YOLOv8n backbone trained on the IQ-OTH/NCCD dataset of 649 annotated CT images, achieved a mean average precision (mAP50) of 0.687 when evaluated against expert-verified annotations. The machine learning branch utilizes a structured clinical dataset of 309 patient records with symptom- and risk-based features, including age, smoking status, chronic disease, fatigue, wheezing, and shortness of breath, all evaluated under standardized conditions. These machine learning models, showed similarly promising results, with the Logistic Regression (Log) achieving the highest mean accuracy of 0.929, while CatBoost (CB) demonstrated the strongest overall balance with a Mean F1 score of 0.953 at optimal hyperparameter settings of 1,000 iterations and a tree depth of 8. Interpretability is further enhanced through a clinically grounded cancer probability equation. This framework that showcases both deep and machine learning models will offer insight into the effectiveness of incorporating artificial intelligence in the screen accuracy of lung cancer classification.

Keywords: Lung cancer detection, Deep learning, Machine learning, CT scan, Pulmonary nodule classification

INTRODUCTION

Cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 [1, 2], with lung cancer, together with prostate and colorectal cancers, accounting for nearly half of all incident cancer cases [3]. Delays in diagnosis and treatment may lead to an uptick in advanced-stage case and mortality specifically [4]. CT imaging is often used in clinical practice to detect and classify pulmonary nodules at early stages [5], potentially allowing for the early classification of malignant tumors and therefore improving patient survivability. With this technology in mind, differentiating between the subtle features of malignant and benign tumors remains a challenge, especially when considering

how radiologist interpretation may be influenced by intrapersonal factors such as workload; where some studies found radiologists made 21% higher than their normal productivity [6]. With the combined concerns of both alarming cancer mortality rates and the possible human-related risks linked to tumor classification, the application of artificial intelligence (AI) in the practice looks increasingly promising. Developments in deep learning technology, especially convolutional neural networks (CNNs) have demonstrated many advantages. Most crucially, CNNs are capable of automatically learning hierarchical features from large volumes of imaging data [7], which is unique to traditional computerized detection systems, as they can process data without external feature engineering. These two factors allow for efficient

recognition and classification of benign and malignant tumors. On top of efficient deep learning models, AI-based diagnostics have proven high accuracy in identifying tumors [8]. Ardila et al. developed a deep learning system that outperformed radiologists in detecting lung cancers in CT scans [9]. Nam et al. similarly reported that their deep learning algorithm outperformed physicians in “radiograph classification and nodule detection performance for malignant pulmonary nodules on chest radiographs” [10]. The combination of classification efficiency, elimination of human error, and accurate pattern recognition indicates effective serviceability of AI in supporting radiologic decision-making.

Despite AI's promising to support clinical decision-making many considerations must first be accounted for as well. In particular, more research must be conducted into the generalizability of models in using datasets from various institutions [11]. Aside from the practical aspects of AI integration into clinical environments, professionals must also answer questions related to gaps in AI-assisted lung tumor classification. Firstly, many studies rely on only a single dataset, with one study finding that only around 23% of ML-based healthcare papers used multiple datasets [12]. Secondly, effective models should be transparent with their decisions, yet many act as black boxes and do not provide justifications that clinicians require for informed decision-making. Thirdly, regarding deep learning models, direct systematic comparisons between multiple deep learning architectures under identical experimental conditions must also be addressed. These gaps in key information will be investigated to gauge the effectiveness of AI algorithms in applying them to lung cancer diagnosis.

To address these gaps, this study proposes a dual-pipeline framework that, to the authors' knowledge, is among the first to integrate a convolutional object detection model operating on CT scan images with a gradient-boosting classifier trained on structured clinical and symptom data, combining their outputs into a single interpretable risk score. This integration is novel in three respects: it leverages complementary information modalities-imaging and clinical data-within a unified evaluation framework; it enables systematic, controlled comparison of multiple deep learning architectures; and it produces a clinically grounded, interpretable risk score rather than an opaque black-box prediction, directly addressing the transparency gap identified above.

This study aims to evaluate the effectiveness of multiple deep learning and machine learning architectures in differentiating between benign and malignant tumors on CT scans. To break it down into clear objectives, this study aims to: (1) develop CNN-based classification models for lung nodule analysis, (2) compare the performance of multiple deep learning models under similar controlled environments, (3)

evaluate and present the interpretability of the models through a clinically grounded probability equation, and (4) propose an integrated risk-scoring framework that combines the imaging and clinical pipelines into a single interpretable measure of lung cancer risk. The findings of this study contribute to the evidence base on AI-assisted lung cancer diagnostics and provide practical insights for future integration into clinical practice.

MATERIALS AND METHODS

Overall research design

This study follows a two-branch parallel pipeline for developing AI-based models to differentiate between benign and malignant tumors in CT scans, as illustrated in Figure 1. The overall workflow begins with planning for the AI model and then divides into two parallel tracks: (1) an Object Detection model branch and (2) a Machine Learning model branch. Both branches converge at a shared model evaluation stage, after which a decision is made on whether the model meets performance criteria. If the criteria are not met, the model is retrained with adjusted parameters. Once the criteria are satisfied, the model is built and deployed for clinical applications.

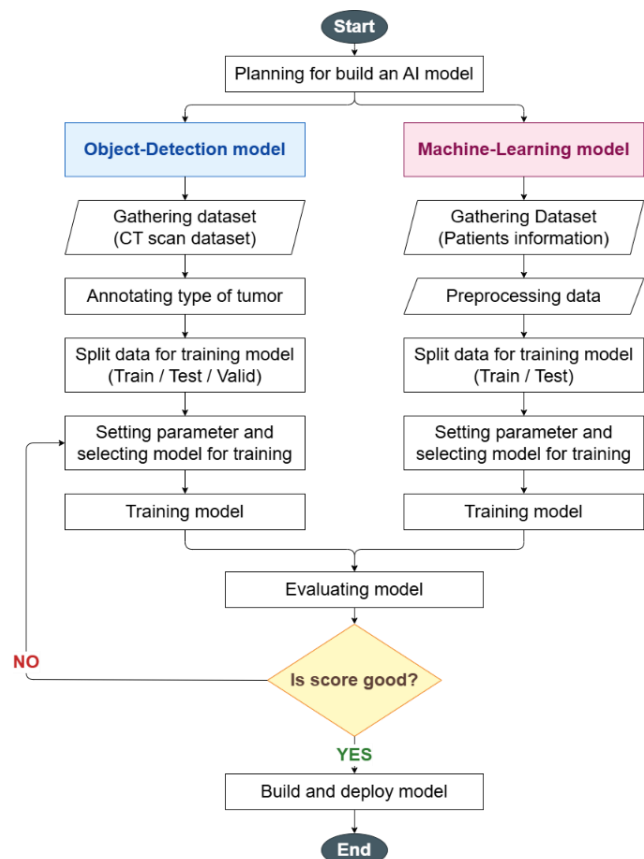


Figure 1 Overall research design flowchart for developing the AI-based tumor classification model.

Tumor and Nodule Identification

Pulmonary nodules are defined as small, well-defined, and typically spherical areas of abnormal tissue growth which could be benign or malignant, whereas tumors represent a broader classification encompassing abnormal growths of various sizes and morphologies. Accurately classifying these anomalies as either benign or malignant is fundamental to the development and training of a high-performance Object Detection Model. The differentiation between benignity and malignancy involves several physiological and radiological factors. Key indicators include calcification patterns, nodule size, growth rate, overall shape, and margin characteristics (such as spiculation). Additionally, location within the lungs is also crucial for the classification of the aforementioned anomalies, hence why the inclusion of multiple datasets in model training is vital to recording variation between the locations of the CT scan images. This information was primarily used to annotate the type of tumor under the Object Detection pathway. See below a comprehensive summary of these variables in determining the probability of whether a nodule is benign or malignant, presented in Table 1.

Table 1 summary of qualities used differentiating benign and malignant nodules on CT scans [13 – 18].

Qualities	Benign	Malignant
Calcification	Diffuse, central, laminated, or popcorn calcifications	Expect other calcification patterns
Size	Typically less than 3cm in size	Typically more than 3cm in size
Growth	Relative comparisons with prior imaging studies	
Shape	Three-dimensional ratio > 1.78	
Margin	Smooth margins (unless metastatic in origin)	Corona radiata signs
Location	Hamartoma: Peripheral areas	Adenocarcinoma: Peripheral areas
	Chondroma: Location varies	Squamous Cell Carcinoma: Central areas
	Carcinoid Tumors: Often central but may appear in other areas	Large Cell Carcinoma: Peripheral areas
		Small Cell Lung Cancer: Central areas

Probability of lung cancer

In addition to the symptom-based clinical data used to train machine learning models in this study, the established Mayo Clinic malignancy prediction model [34] considers several key radiological features for estimating the probability of malignancy. (1), and one of the most vital characteristics of tumors is the Nodule Diameter. Measured in millimeters at the widest point across all views, Nodule Diameter serves as a baseline indicator of growth, with tumors larger than 3 cm in diameter considered as malignant. (2) Spiculation, star-like patterns, are clinically recognized as a sign of malignancy. (3) The location of the tumor, specifically a position in the upper lobe, may indicate that the tumor is malignant. (4) Reported patient data, such as Age, Smoking History, and Family Cancer history, provides invaluable context for malignancy.

$$x = -6.8272 + (0.0391 * Age) \quad (1)$$

$$+ (0.7917 * SH)$$

$$+ (1.3388 * CH)$$

$$+ (0.1274 * DN)$$

$$+ (1.0407 * SPICULATION)$$

$$+ (0.7838 * UPPER LOBE)$$

$$PC = (e^x) / (1 + e^x) \quad (2)$$

- Age: In years.
- Smoke history (SH): 1 if current/former smoker, 0 if never.
- Cancer history (CH): 1 if history of cancer 5+ years ago, 0 if not.
- Diameter of nodules (DN): Largest diameter of the nodule in mm.
- Spiculation: 1 if spiculation (ragged edge) is present, 0 if not.
- Upper Lobe: 1 if the nodule is in the upper lobe, 0 if not.
- Probability of cancer (PC)

Multi-view morphological and positional analysis

While multi-view analysis provides a more complete perspective into tumors, this study emphasizes a focus on Axial View CT scans for Object Detection CT scan analysis. Although this may cause a small number of tumors to be missed due to the difficulty of precisely identifying tumor shapes during the labeling phase, this study is still willing to move forward with this blueprint. Seeing as a key focus of this study is on evaluating the performance of numerous AI models in screening for tumors, by focusing on one perspective, this study better enhances computational efficiency and reduces complexity, something which is essential for real-time or high-throughput processing in clinical environments. This streamlined approach allows the models to maintain high sensitivity in identifying benign and malignant nodules without compromising processing speed or accuracy. For

informative purposes, see below for the types of views considered for the model to built on:

- **Axial View:** A perspective that slices the body into top and bottom sections. Axial views can be used for initial nodule detection and horizontal localization. The images are presented in Figure 2 (a)
- **Coronal View:** This view images the body in slices from front to back; coronal views are used to determine the vertical position of the nodule relative to landmark organs like the heart and diaphragm. The images are presented in Figure 2 (b)
- **Sagittal View:** This view is critical for identifying the specific lung lobe, allowing radiologists to determine whether the nodule is in the upper or lower lobe and figure out the possibility of malignancy. The images are presented in Figure 2 (c).

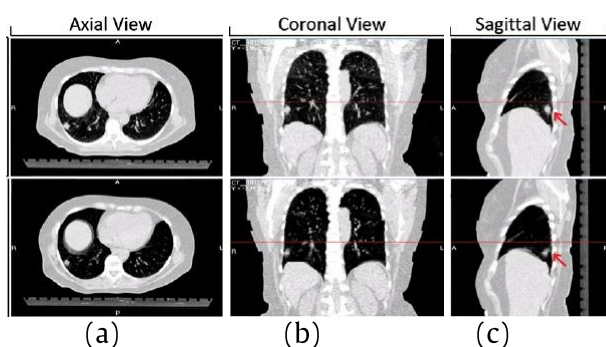


Figure 2 Morphology (Source: Li et al. 2012).

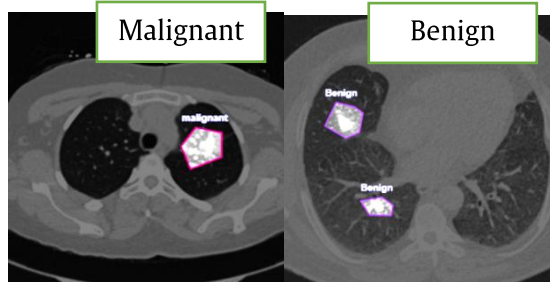


Figure 3 Annotating malignant and benign.

Dataset preparation

Two separate datasets are used for each model branch in this study. The Object Detection branch is built on CT scan images from the publicly available IQ-OTH/NCCD dataset, which contains annotated low-dose CT scans with labelled pulmonary nodules [19]. Each identified nodule is annotated as either benign or malignant based on radiologist consensus following established annotation protocols for pulmonary nodule classification. Samples of labelled CT scan images are presented in Figure 3. The Machine Learning branch, on the other hand, incorporates patients' demographic and symptom-based clinical information, such as age, smoking status, and respiratory symptoms, to compile a structured tabular dataset.

Both datasets undergo preprocessing steps, including normalization, removal of incomplete records, and class balancing using the Synthetic Minority Oversampling Technique (SMOTE) before model training [20].

Object detection model branch

The Object Detection branch is designed to identify and localize tumors directly within CT scan images. After gathering the CT scan dataset, each nodule is manually annotated with a bounding box and labeled as benign or malignant. The dataset is then split at the image level into training (note that these per-image patient identifiers were not exposed in the public release), validation, and test sets at a ratio of 70:15:15 [21]. The detection dataset comprised 649 axial CT images derived from the IQ-OTH/NCCD lung cancer dataset and annotated using Roboflow, partitioned into 456 training, 96 validation, and 97 test images and containing 277 benign and 1,173 malignant annotated nodules. Although this study acknowledges potential leakage risk within this internal split during training. However, final model performance was independently evaluated on a separate, Memorial Sloan Kettering Cancer Center (MSKCC) verified dataset [22] with no patient overlap with the training data, providing an external generalization check unaffected by the internal split's leakage risk.

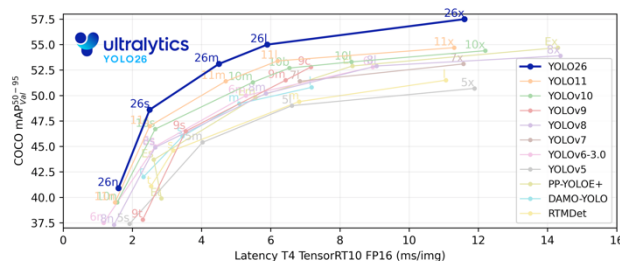


Figure 4 YOLO models (Source: Ultralytics, 2024).



Figure 5 Data augmentation techniques.

In the first stage, the bounding boxes were annotated by the researchers themselves, who identified nodule regions based on radiological characteristics based on radiological characteristics, including nodule size and location within the lung, using class labels provided by the IQ-OTH/NCCD dataset. The models were then trained and tested on these annotations. In the second stage, the same model predictions were compared against a separate dataset whose annotations had been confirmed by the MSKCC verified dataset [22], which provided a more reliable ground truth for

evaluating performance. Model parameters such as the number of training epochs, batch size, and learning rate are configured prior to training. See Figure 4 for the deep Malignant Benign learning frameworks based on YOLO models, which have been fine-tuned on the annotated CT scan images. Random flipping, rotation, and brightness adjustments are shown in Figure 5 in order to improve model generalizability in training.

1	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY
2	M	69	1	2	2
3	M	74	2	1	1
4	F	59	1	1	1
5	M	63	2	2	2
6	F	63	1	2	1
7	F	75	1	2	1
8	M	52	2	1	1
9	F	51	2	2	2
10	F	68	2	1	2
PEER_PRESSURE	CHRONIC_DISEASE	FATIGUE	ALLERGY	WHEEZING	
1		1	2	1	2
1		2	2	2	1
2		1	2	1	2
1		1	1	1	1
1		1	1	1	2
ALCOHOL_CONSUMING	COUGHING	SHORTNESS_OF_BREATH			
2		2			2
1		1			2
1		2			2
2		1			1
1		2			2
1		2			2
2		2			2
1		1			2
SWALLOWING_DIFFICULTY	CHEST_PAIN	LUNG_CANCER			
2		2 YES			
2		2 YES			
1		2 NO			
2		2 NO			
1		1 NO			
1		1 YES			

Figure 6 Example of patient information.

Machine learning model branch

The Machine Learning branch uses structured patient data rather than raw images. Patient information including clinical features is shown in Figure 6 such as patient age, gender, smoking history, anxiety, peer pressure, chronic disease, fatigue, allergy, wheezing, alcohol consumption, coughing, shortness of breath, swallowing difficulty, and chest pain are gathered and preprocessed. Feature engineering steps include encoding categorical variables, handling missing values, and applying standardization. The dataset is evaluated using 10-fold stratified cross-validation, with SMOTE applied exclusively within each training fold to prevent data leakage; reported metrics represent the mean across all folds. Although the 10-fold stratified cross-validation is somewhat practical in preventing data leakage, it must be stated that the lack of a nested cross-validation may act as a limitation when it comes to ensuring that the final test set remains unseen. The clinical dataset comprised 309 patient records,

with 270 positive and 39 negative cases, reflecting a class imbalance of approximately 6.9:1 that motivated the use of SMOTE. Multiple machine learning classifiers are trained and compared, including Logistic Regression (Log), XGBoost (XGB), LightGBM (LGBM), CatBoost (CB), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB) [23]. A summary of the performance comparison across all models is presented in Table 4. Hyperparameter tuning is performed using grid search with cross-validation to identify the optimal model configuration for each algorithm [24, 25].

Model evaluation and deployment

Both model branches feed into a unified evaluation stage. Model performance is assessed using a comprehensive set of classification metrics derived from the confusion matrix. These metrics include accuracy (Eq.3), precision (Eq.4), recall or sensitivity (Eq.5), and F1 score (Eq.6) [26]. In these equations, TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Since k-fold cross-validation is applied, the mean value of each metric across all k folds is used as the final reported performance indicator, where k denotes the number of cross-validation folds. The mean accuracy (MA), mean precision (MP), mean recall (MR), and mean F1 score (MF1) are defined in Equations 7 through 10 respectively.

$$MA = \frac{1}{k} \sum_{i=1}^k Accuracy_i \quad (7)$$

$$MP = \frac{1}{k} \sum_{i=1}^k Precision_i \quad (8)$$

$$MR = \frac{1}{k} \sum_{i=1}^k Recall_i \quad (9)$$

$$MF1 = \frac{1}{k} \sum_{i=1}^k F1\ Score_i \quad (10)$$



Figure 7 Representative CT scan images with model-predicted malignancy labels and confidence scores.

A model is considered to pass evaluation if it achieved an AUC-ROC of at least 0.85 and a mean sensitivity of at least 0.80 on the held-out test set, thresholds selected based on clinically accepted standards for diagnostic screening tools. If a model

fails to meet these criteria, the training parameters are adjusted and the model is retrained. Once the performance criteria are met by either or both branches, the final model is built and packaged for deployment as a decision-support tool that can be integrated into a clinical CT image review workflow, as shown in Figure 7.

Integrated Risk Scoring

To translate the complementary outputs of the two pipelines into a single, clinically interpretable measure, this study proposes an integrated risk score. Although the object detection and machine learning pipelines are trained independently on domain-appropriate datasets, both can be applied to the same patient at inference time: the object detection model contributes the predicted probability of malignancy, $P(\text{malignant})$, from the detected nodule, while the machine learning model contributes the predicted cancer probability, $P(\text{cancer})$, from the patient's clinical and symptom data. The two probabilities are weighted equally and combined into a composite score on a 0-100 scale, as defined in Equation 11:

$$\text{Risk Score} = (P(\text{malignant}) \times 50) + (P(\text{cancer}) \times 50) \quad (11)$$

Table 2 Representative examples of the proposed integrated risk score. $P(\text{mal})$: imaging-derived probability of malignancy; $P(\text{cancer})$: clinical probability of cancer; Score: composite value on a 0-100 scale.

Case	$P(\text{mal})$	$P(\text{cancer})$	Score	Risk
A	0.00	0.70	35.0	Moderate
B	0.85	0.60	72.5	High
C	0.20	0.15	17.5	Low

Here $P(\text{malignant})$ and $P(\text{cancer})$ each range from 0 to 1, yielding a final score between 0 and 100, where higher values indicate greater estimated risk. The equal 50:50 weighting reflects the absence of prior evidence favouring either modality and can be recalibrated once validation data become available. This formulation allows the framework to remain informative even when one modality is uncertain; for example, when no nodule is detected ($P(\text{malignant}) = 0$), the clinical branch still contributes a risk estimate, mirroring evidence that clinical predictors retain value in the absence of visible nodules. The use of complementary imaging and clinical predictors for risk stratification is consistent with established nodule-malignancy models [27, 28] and standardized risk-reporting systems such as Lung-RADS [29] and recent multimodal fusion approaches [30]. It should be emphasised that, because the two branches were trained on separate datasets, the integrated risk score is presented here as a proposed framework; its empirical validation requires a multimodal cohort

containing CT scans, clinical records, and confirmed diagnoses for the same patients, as noted in the limitations. Table 2 illustrates representative cases showing how the score behaves across differing imaging and clinical inputs.

RESULTS AND DISCUSSION

Machine learning model comparison

Table 3 and Figure 8 summarize the machine learning model performance across each evaluated metric, including accuracy, precision, recall, and F1-score. Among the evaluated models, logistic regression emerged as the top performer in terms of overall predictive accuracy, achieving the highest mean accuracy of 0.929. The CB model followed closely with an accuracy of 0.906 and demonstrated superior balance, yielding a strong F1-score of 0.945. Upon closer analysis of the models, it has become apparent that different models stand out in different areas of strength, in particular between efficiency and accuracy. Comparing the top machine learning architectures, the logistic regression model excels in its simplicity and efficiency while the CB model offers robust predictive power through its advanced gradient-boosting framework.

Table 3 Performance comparison of various machine learning models (default parameter).

Model	MA	MP	MR	MF1
Log	0.929	0.960	0.959	0.959
XGB	0.900	0.938	0.948	0.942
LGBM	0.900	0.938	0.948	0.942
CB	0.906	0.948	0.944	0.945
RF	0.903	0.941	0.948	0.944
DT	0.887	0.945	0.926	0.934
SVM	0.479	0.925	0.441	0.588
KNN	0.871	0.947	0.904	0.924
NB	0.900	0.920	0.970	0.944

Table 4 AUC comparison of various machine learning models (Default parameter).

Model	Mean AUC	Std AUC
Log	0.937	0.0535
XGB	0.9006	0.0847
LGBM	0.9043	0.0904
CB	0.9145	0.0817
RF	0.9224	0.0669
DT	0.7949	0.1423
SVM	0.6781	0.1346
KNN	0.8559	0.1439
NB	0.8986	0.0888

Following up on concerns of false negative evaluations, high recall models are vital to ensure that all potentially malignant nodules are accounted for, even if it results in a less harmful false positive outcome. The NB model fits this precautionary criterion, achieving the highest recall (MR) of 0.970. The NB model favors high sensitivity; however, it achieves lower precision than the ensemble methods. These results suggest that while NB is highly effective at minimizing false negatives, ensemble-based models like CB and RF offer a greater balance of metrics; making them highly suitable for robust clinical support systems.

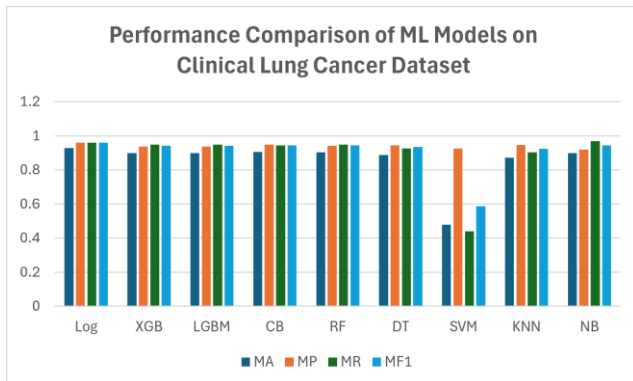


Figure 8 Performance Comparison of ML Models on Clinical Lung Cancer Dataset.

Table 5 Evaluation of CB Model Efficiency with Varying Iteration Values.

Iteration	MA	MP	MR	MF1
100	0.913	0.951	0.948	0.949
200	0.910	0.948	0.948	0.947
500	0.913	0.949	0.952	0.95
1000	0.919	0.956	0.952	0.953
2000	0.919	0.956	0.952	0.953

Table 6 Evaluation of CB Model Efficiency with Varying Eta Values.

Eta	MA	MP	MR	MF1
0.01	0.9100	0.948	0.948	0.947
0.05	0.919	0.956	0.952	0.953
0.15	0.916	0.955	0.948	0.951
0.20	0.919	0.959	0.948	0.953
0.30	0.916	0.955	0.948	0.951

Table 7 Evaluation of CB Model Efficiency with Varying Depth Values.

Depth	MA	MP	MR	MF1
2	0.900	0.938	0.948	0.942
4	0.919	0.953	0.956	0.953
6	0.913	0.949	0.952	0.950
8	0.919	0.956	0.952	0.953
10	0.916	0.956	0.948	0.951

Deep learning analysis

The experimental results summarized in Tables 8 and 9 provide a comparative evaluation of the

CB hyperparameter analysis

Further investigation of the CB model in Tables 5-7 shows that adjusting certain hyperparameters can directly affect diagnostic efficiency [31]. The number of iterations played a crucial role in model convergence. Increasing the iterations to 1,000 gave the optimal mean F1-score (MF1) of 0.953. Past this point, however, the model's performance slightly decreased and plateaued, indicating that beyond the 1,000 iteration threshold, no significant additional predictive gains are achieved. Another critical factor in the CB model is the depth of the decision trees in the dataset, with tree depths of 4 and 8 both yielding the highest recorded accuracy of 0.919 and an F1-score of 0.953; depth 8 was selected for its greater representational capacity.

Within the training process, Learning Rate (eta) was identified as key to ensuring model stability. Decrease yielded comparable results; however, 0.05 was selected for its stability in preventing overfitting. Furthermore, higher learning rates lead to a slight decrease in overall accuracy. Collectively, these findings highlight how systematic hyperparameter tuning is fundamental in maximizing model effectiveness in tumor classification.

performance of various YOLO architectures, emphasizing the influence of ground truth quality on model evaluation. Out of the YOLO architectures, YOLOv8n

is shown to be the most effective architecture for the dataset (see Table 7). Despite newer architectures such as YOLOv9 to YOLOv11, YOLOv8n outperformed these models and achieve the highest mAP50 of 0.3857 [32].

To further support YOLOv8n's effectiveness over additional models, the YOLOv6n performance must be addressed. While YOLOv6n exhibited the highest precision at 0.694, its significantly lower recall of 0.456 and an mAP50 of 0.232 may prove ineffective in real clinical settings. The aforementioned metrics in particular indicate that YOLOv6n operates with an overly conservative detection threshold, potentially leading to a higher rate of false negatives and therefore a likelihood that the model will miss positive instances. Conversely, YOLOv8n remains balanced on factors of precision and recall, allowing for high sensitivity in diagnostics. Additionally, the longitudinal performance of the YOLO models demonstrated a consistent upward trend in the number of epochs directly correlated with improved model convergence and detection accuracy (see in Figure 9), highlighting the importance of training duration in the optimization of deep learning performance in medical imaging.

Table 8 Performance comparison of various deep learning models.

Model	mAP50	Precision	Recall
YOLOv6n	0.232	0.694	0.456
YOLOv8n	0.385	0.647	0.435
YOLOv9c	0.346	0.657	0.429
YOLOv10n	0.311	0.602	0.432
YOLOv11n	0.372	0.664	0.446

Table 9 Performance evaluation of YOLOv8n with the 'MSKCC verified dataset' [22].

Model	mAP50 (Original Val)	mAP50 (MSKCC Verified Dataset)
YOLOv8n	0.385	0.687

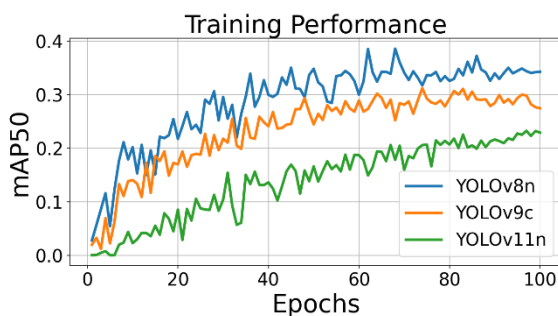


Figure 9 Training Performance (mAP50/Epochs).

A key observation across the YOLO architectures is the relatively low recall, with YOLOv8n achieving a recall of 0.435 against the researcher-annotated

test set. In a cancer screening context, recall is a critical metric, since missed malignant nodules (false negatives) carry the greatest clinical risk. Several factors likely contributed to this limitation. First, the relatively small size of the annotated dataset (approximately 1,000 images) constrains the ability of the model to generalize across the morphological variability of pulmonary nodules. Second, the first-stage bounding boxes were annotated by the researchers rather than qualified radiologists, introducing label noise that disproportionately reduces recall. The increase in mAP50 from 0.385 to 0.687 under the MSKCC verified dataset (Table 9) supports this interpretation, indicating that a substantial portion of the apparent error originated from ground-truth quality rather than model capacity. Future work should therefore prioritise expanding the dataset, obtaining expert annotations, and applying nodule-focused augmentation to raise sensitivity to a clinically acceptable level.

Comparison with previous studies and limitations

When situated against prior work, the present findings are consistent with the broader literature in which deep learning systems for lung nodule analysis report strong performance under expert-curated ground truth while remaining sensitive to annotation quality and dataset scale [10, 11]; the improvement observed under MSKCC verified dataset [22] aligns with this pattern. Nevertheless, several limitations should be acknowledged. First, the deep learning and machine learning pipelines were trained and evaluated on separate datasets, and the proposed integrated risk score is therefore presented as a conceptual framework rather than an empirically validated outcome. Second, the object detection dataset is limited in size and was partially annotated by non-experts, which constrains generalisability. Third, the detection data were split at the image level rather than the patient level, and no external multi-institutional validation was performed, limiting evidence for cross-population robustness. Finally, the clinical machine learning dataset is symptom-based and does not include the radiological measurements used in established models such as the Mayo Clinic equation [27]; integrating these complementary feature sets within a single multimodal cohort represents an important direction for future work. These limitations temper any claim of immediate clinical deployment and instead position the study as evidence supporting the feasibility of AI-assisted lung cancer screening.

CONCLUSIONS

This research evaluated the effectiveness of deep learning and machine learning techniques in differentiating between benign and malignant lung nodules in CT scans. Through the use of a two-branch

parallel pipeline consisting of an object detection branch for image analysis and a Machine Learning branch for clinical data, the study provides a comprehensive approach to automated lung cancer screening.

Performance of machine learning models

This study has demonstrated the effectiveness of machine learning algorithms for classification tasks by incorporating contextual patient data and radiological features. The log models showed excellent accuracy with a mean accuracy of 0.929, making them most optimal for screening accuracy. Additionally, hyperparameter optimization of the CB model yielded a mean F1 score of 0.953 with 1,000 iterations and a tree depth of 8. These findings demonstrate the above adjustments will best improve the performance of the CB model.

Performance of object detection models

The object detection framework was able to classify benign and malignant tumors based on CT scan images. The YOLO architectures were compared, YOLOv8n was found to be the best model in terms of mAP50, precision and recall. Notably, the performance of YOLOv8n was heightened to mAP50 of 0.6875 when tested against datasets verified by medical experts. These results highlight the potential of deep learning architectures for tumor classification and the crucial need for high-quality expert annotated data to build trustworthy AI tools for the clinic. The low recall of 0.435 is likely influenced by the relatively small size of the IQ-OTH/NCCD dataset (110 cases, 1,190 images) and the use of single-annotator labeling. While bounding boxes were drawn based on established radiological criteria, the absence of multi-expert consensus annotation may have introduced variability in ground truth quality, contributing to missed detections. Future work should prioritize expanding the dataset with direct physician-verified annotations, which, as demonstrated in this study through the MSKCC verified dataset [22], improved mAP50 from 0.385 to 0.687."

Clinical significance and future work

The findings of this study underscore the potential of AI-driven systems to support clinical decision-making tools in clinical workflows. Through the incorporation of key identifying features such as nodule diameter, spiculation, and anatomical location into a predictive probability equation, along with deep learning object detection models designed to differentiate between benign and malignant tumors in CT scan images; dual-pipeline frameworks such as the one in this study show great promise towards early detection and improving survival outcomes for lung cancer patients. Future research should focus on enhancing model generalizability across diverse patient populations and further improving the

transparency of black box algorithms for to facilitate their potential clinical application, future studies may also develop a more practical integrated risk scoring system using a unified multimodal cohort.

DECLARATION OF AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The author utilized ChatGPT to improve the legibility and clarity of the text during the preparation of this work.

ACKNOWLEDGEMENT

The author would like to thank Mr. Pichayapazz Naewthong and Mr. Dechanin Tengcharoenkul for their support and facilities provided throughout this research.

REFERENCES

1. World Health Organization. Cancer Son Erişim Tarihi. 2021;24. Available from: <https://www.who.int/news-room/fact-sheets/detail/cancer>.
2. Neacşu F, Vârban AŞ, Simion G, Şurghie R, Pătraşcu OM, Sajin M, et al. Lung cancer mimickers—a case series of seven patients and review of the literature. *Rom J Morphol Embryol*. 2021;62(3):697-704.
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA: a cancer journal for clinicians*. 2023;73(1):17-48.
4. Ellis PM, Vandermeer R. Delays in the Diagnosis of Lung Cancer. *J Thorac Dis*. 2011;3(3):183-8.
5. Ahmed S, Subhan F, Mohd Su'ud M, Alam MM, Waheed A. A Survey of Lung Nodules Detection and Classification from CT Scan Images. *Comput Syst Sci Eng*. 2024;48(6):1483–511
6. Hanna TN, Lamoureux C, Krupinski EA, Weber S, Johnson JO. Effect of Shift, Schedule, and Volume on Interpretive Accuracy: A Retrospective Analysis of 2.9 Million Radiologic Examinations. *Radiology*. 2018;287(1):205-12.
7. Mienye ID, Swart TG, Obaido G, Jordan M, Ilono P. Deep Convolutional Neural Networks in Medical Image Analysis: A Review. *Information*. 2025;16(3):195.
8. Wang M, Chang W, Zhang Y. Artificial Intelligence for the Diagnosis and Management of Cancers: Potentials and Challenges. *MedComm*. 2025;6(11):e70460.
9. Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*. 2019;25(6):954-61.

10. Nam JG, Park S, Hwang EJ, Lee JH, Jin KN, Lim KY, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology*. 2019;290(1):218-28.
11. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*. 2018;15(11):e1002686.
12. Yang J, Soltan AA S, Clifton DA. Machine Learning Generalizability across Healthcare Settings: Insights from Multi-Site COVID-19 Screening. *Npj Digital Medicine*. 2022;5(1). doi: 10.1038/s41746-022-00614-9.
13. Zhen F, Schraeder R. The Changing Pathology of Lung Cancer. *Surg Oncol Clin N Am*. 2011;20(4): 637-53.
14. Constantin AA, Cotea AA, Mihăițan FD. The Multiple Faces of Lung Adenocarcinoma: Challenges in Diagnosis. *Internal Medicine*. 2023;20(3):69-80.
15. Rudin CM, Brambilla E, Faivre-Finn C, Sage J. Small-cell lung cancer. *Nature Reviews Disease Primers*. 2021;7(1):3.
16. Tian D, Wen H, Zhou Y, Fu M. Pulmonary chondroma: A clinicopathological study of 29 cases and a review of the literature. *Mol Clin Oncol*. 2016;5(3): 211-5.
17. Ann L, Smithuis R. The Radiology Assistant: Benign versus Malignant. Available from: <https://radiologyassistant.nl/chest/solitary-pulmonary-nodule/benign-versus-malignant>.
18. Onno M, Smithuis R. The Radiology Assistant: Fleischner 2017 Guideline. Available from: <https://radiologyassistant.nl/chest/plummonary-nodules/fleischner-2017-guideline>.
19. Al-Yasriy HF, AL-Husieny MS, Mohsen FY, Khalil EA, Hassan ZS. Diagnosis of Lung Cancer Based on CT Scans Using CNN. *IOP Conference Series: Materials Science and Engineering*. 202;928 (2):022035.
20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-57.
21. Joseph VR. Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 2022;15(4):531-8.
22. The Coffee-break lung CT collection with scan images reconstructed at multiple imaging parameters [Internet]. The Cancer Imaging Archive (TCIA). 2025. Available from: <https://doi.org/10.7937/k9/tcia.2015.u1x8a5nr>.
23. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):1-48.
24. Breiman L. Random forests. *Machine learning*. 2001;45(1):5-32.
25. Bergstra J, Bengio Y. Random search for hyperparameter optimization. *J Mach Learn Res*. 2012 Feb 1;13(2):281 - 305.
26. Fawcett T. An introduction to ROC analysis. *Pattern recognition letters*. 2006;27(8):861-74.
27. Swensen SJ, Silverstein MD, Ilstrup DM, Schleck CD, Edell ES. The probability of malignancy in solitary pulmonary nodules. Application to small radiologically indeterminate nodules. *Archives of Internal Medicine*. 1997;157(8):849-55.
28. McWilliams A, Tammemagi MC, Mayo JR, Roberts H, Liu G, Soghrati K, et al. Probability of cancer in pulmonary nodules detected on first screening CT. *New England Journal of Medicine*. 2013;369(10):910-9.
29. Lung CT. Screening Reporting and Data System (Lung-RADS) version 1.1. American College of Radiology Web site [Internet]. 2019.
30. Huang H, Zheng D, Chen H, Wang Y, Chen C, Xu L, et al. Fusion of CT images and clinical variables based on deep learning for predicting invasiveness risk of stage I lung adenocarcinoma. *Medical Physics*. 2022;49(10):6384-94.
31. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*. 2018;31. Available from: <https://doi.org/10.48550/arXiv.1706.09516>
32. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. 2023. Available from: <https://github.com/ultralytics/ultralytics>.