International Journal of Multidisciplinary Research Excellence (IJMRE)

Volume-II (Issue 3) – July- Sept 2026 ISSN: 3107-3697

Lung Cancer Detection Using Convolutional Neural Networks (CNN): A Deep Learning-Based Approach for Medical Image Analysis

[1.] **Dr. Ajay Saini**
Associate Professor
Arya College of Engineering
Jaipur (Rajasthan)

[2.] **Shashank Gujrati**
Research Scholar
Arya College of Engineering
Jaipur(Rajasthan)

ABSTRACT

Lung cancer remains a major cause of cancer-related mortality, and timely identification of suspicious pulmonary nodules is an important component of risk assessment. Computed tomography (CT) provides detailed volumetric information, yet nodule detection and malignancy estimation are affected by small lesion size, variable morphology, reader variability, and differences among scanners and institutions. This paper reviews convolutional neural network (CNN)-based approaches for pulmonary nodule detection and malignancy classification, emphasizing work available from 2022 through the preparation date in 2026. Public resources considered include LIDC-IDRI, LUNA16, and the Lung Nodule Database (LNDb); these datasets support different tasks and must not be treated as interchangeable sources of histopathologically confirmed cancer labels. A reproducible pipeline is proposed covering DICOM handling, Hounsfield-unit conversion, lung-region processing, candidate extraction, patient-level data partitioning, augmentation, transfer learning, class-imbalance management, model selection, calibration, and explainability. Candidate architectures include compact custom CNNs, ResNet, DenseNet, EfficientNet, three-dimensional CNNs, and CNN–Transformer hybrids, selected according to task and available data. Evaluation is specified using sensitivity, specificity, precision, F1-score, ROC-AUC, false positives per scan, calibration, and external validation. Recent systematic reviews report promising performance for deep-learning systems but also identify methodological heterogeneity and limited external validation. No original model was trained for this paper; therefore, it reports no new accuracy, confusion matrix, or inference-time result. The contribution is a source-grounded synthesis and a practical evaluation protocol that prioritizes leakage prevention, transparent reporting, uncertainty assessment, and clinician oversight over unverified headline metrics.

KEYWORDS

Lung cancer detection; convolutional neural networks; computed tomography; pulmonary nodules; deep learning; transfer learning; explainable artificial intelligence; medical image analysis.

I. INTRODUCTION

Lung cancer is a substantial global health burden. The World Health Organization reports approximately 2.5 million new cases and 1.8 million deaths worldwide in 2022, and notes that diagnosis frequently occurs at advanced stages [1]. The

International Agency for Research on Cancer likewise identifies lung cancer as a leading cause of cancer incidence and mortality [2]. Screening selected high-risk populations with low-dose CT can support earlier detection, but screening decisions and follow-up remain clinical processes rather than purely image-classification tasks.

CT scans represent anatomy as a stack of slices, enabling assessment of nodule size, shape, density, margins, and growth over time. Human interpretation is demanding because nodules may be tiny, adjacent to vessels or pleura, or visually similar to benign structures. Inter-reader variation and high scan volumes motivate computer-aided detection (CADe) and computer-aided diagnosis (CADx) tools. CADe proposes candidate locations; CADx estimates the likelihood or risk associated with an identified nodule. A patient-level screening or cancer prediction model is a distinct task requiring clinical outcomes and longitudinal context.

CNNs learn spatial filters and hierarchical image representations directly from pixel or voxel data. Residual connections, dense feature reuse, and efficient convolutional blocks provide useful design choices, while three-dimensional models can incorporate through-plane context. Nevertheless, performance on a public benchmark can be inflated by data leakage, label shortcuts, class imbalance, and differences between development and clinical populations. The central objective of this paper is therefore not to claim a new state-of-the-art result, but to synthesize the evidence and define a reproducible, clinically cautious framework for future evaluation. Foundational CNN work established the effectiveness of learned convolutional features [17], and very deep convolutional architectures further explored depth as a representational advantage [18]. Attention-based models [19] and vision transformers [20] motivate hybrid comparisons, but do not remove the need for task-specific validation.

The paper makes four practical contributions: (1) separates detection, malignancy classification, and patient-level prediction; (2) summarizes the strengths and limitations of widely used datasets and CNN families; (3) provides an implementation-oriented methodology and metric definitions; and (4) specifies safeguards for explainability, calibration, external testing, and transparent reporting.

II. LITERATURE REVIEW

Recent literature covers candidate detection, segmentation, nodule malignancy estimation, and end-to-end decision support. Gao et al. reviewed deep learning for pulmonary nodule detection and segmentation and highlighted the importance of standardized processing, external test sets, data/code sharing, and balancing complexity with efficiency [3]. A separate systematic review and meta-analysis of CT-detected nodules included 17 studies, 8,553 participants, and 9,884 nodules; it reported pooled diagnostic performance for externally validated deep-learning systems and compared them with physician judgement and clinical risk models [4]. These findings are encouraging, but the review's eligibility criteria, population mix, reference standards, and thresholds limit direct transfer of pooled results to any one proposed model.

A 2025 review of CT-based lung cancer diagnosis organizes research around detection, segmentation, and classification and summarizes widely varying results across datasets and evaluation designs [5]. Such tables are useful for mapping the field, but numbers from different test populations, label definitions, split strategies, and class balances cannot be ranked as though they came from one controlled experiment. A study using a curated nodule patch classification set does not establish the sensitivity of a full-volume screening pipeline, and neither task alone establishes patient-level cancer prediction.

A 2024 study proposed a deep-learning-based pyramid network with a honey-badger optimization component and reported high classification accuracy on its stated LIDC-IDRI/NIH chest image setup [6]. Its reported numbers should be interpreted within the paper's own data preparation and task definition; a mixed CT/CXR setup and unclear comparability of image-level classes do not provide a direct benchmark against prospective CT screening. This illustrates why reported accuracy must be read alongside cohort definition, independent split, ground truth, and availability of code/data.

Across the recent literature, common approaches include 2-D CNNs trained on selected axial slices or cropped nodules, 2.5-D inputs that stack neighboring slices as channels, and 3-D CNNs that process volumetric patches. Transfer learning can stabilize training when labelled samples are limited, but natural-image pretraining may not match CT intensity statistics.

Study / year	Evidence type / task	Dataset or scope	Method / reported evidence	Key limitation
Gao et al., 2025 [3]	Systematic review: detection & segmentation	Studies across pulmonary CT datasets	CNN/deep learning dominates; emphasizes standardization and external tests	Heterogeneous protocols and metrics
Deep-learning malignancy meta-analysis, 2024 [4]	Systematic review + meta-analysis; malignancy risk	17 studies; 8,553 participants; 9,884 nodules	Pooled comparisons with physician judgement and clinical risk models	Population and reference-standard heterogeneity
Deep-learning CT diagnosis review, 2025 [5]	Systematic review; detection, segmentation, classification	Multiple public/private datasets	Synthesizes CNN and related methods; reports varied metrics	Reported scores not directly comparable across tasks
DL-LCD pyramid network, 2024 [6]	Original method; classification	LIDC-IDRI and NIH CXR as described by authors	Reports high dataset-specific accuracy; CLAHE and proposed segmentation/classifier pipeline	Mixed modalities/task setup limits comparability
LNDb v4, 2024 [7]	Dataset resource	294 CT scans, Portugal	Radiologist annotations and structured data for development/external evaluation	Annotation disagreement; limited cohort size

Table 1. Summary of selected recent literature

The literature supports CNNs as a useful research family, not a universal solution. In particular, malignancy classification requires an appropriate outcome label such as pathology or sufficiently long follow-up; radiologist suspicion scores and nodule appearance ratings are not equivalent to confirmed malignancy. Studies should disclose whether the target is nodule presence, malignancy risk, histology, or patient-level future diagnosis.

III. DATASET AND PROBLEM DEFINITION

A. Public datasets and their intended use

LIDC-IDRI is a multi-institutional thoracic CT collection with 1,018 cases and annotations from four experienced thoracic radiologists [8], [9]. It is valuable for nodule localization and characterization, but it is not a simple binary cancer-versus-normal dataset: radiologist marks and subjective nodule characteristics should not be silently converted into pathology-confirmed labels. LUNA16 is a benchmark derived from LIDC-IDRI; it includes 888 CT scans after excluding scans with slice thickness greater than 2.5 mm and defines a nodule reference standard based on agreement among radiologists [10].

Dataset	Modality	Cases / scans	Annotations	Primary task	Important limitations
LIDC-IDRI [8], [9]	Thoracic CT	1,018 cases	Four-reader lesion marks; nodule characteristics	Nodule detection, segmentation, characterization	Not uniformly pathology-labelled; reader disagreement
LUNA16 [10], [11]	Thoracic CT	888 CT scans	Consensus-based nodule reference standard	Nodule candidate detection	Derived subset; not a cancer outcome dataset
LNDb [7], [12]	Thoracic CT	294 CT scans	Radiologist nodule coordinates and attributes; held-out test subset	Detection/segmentation and cross-dataset evaluation	Small cohort; reader variability; access rules for test labels
Clinical institutional cohort	CT; exact protocol must be reported	Not applicable until acquired	Pathology and/or longitudinal follow-up, with adjudication	Malignancy risk or patient-level prediction	Governance, selection bias, privacy, and label quality

Table 2. Dataset comparison (counts refer to cases/scans, not slices or nodules)

B. Task definitions

Let a CT volume be$X \in R^(H \times W \times D)$, where H and W are in-plane dimensions and D is the number of slices. A detection model returns candidate locations B = {b1, …, bk} and confidence scores. A nodule classifier receives a crop x_i around a

candidate and predicts a class or risk score. A patient-level model receives a complete scan and potentially clinical covariates c, then estimates $P(y_patient = 1|X, c)$. These outputs have different labels, evaluation units, and clinical meanings.

For binary nodule classification, the training set is $D = (x_i, y_i)_(i = 1)^N$ with $y_i \in 0,1$. A parameterized modelf_θ $maps$ x_i to a $logitz_i$ and probability $p_i = \sigma(z_i)$. The objective is to learn θ that minimizes empirical loss while selecting hyperparameters only on validation data. If labels are ordinal reader ratings rather than confirmed malignancy, the target and conclusions must explicitly reflect that distinction.

IV. PROPOSED METHODOLOGY

The following pipeline is a proposed protocol, not a completed experiment. It should be implemented with a versioned dataset manifest, patient identifiers, immutable split files, code/configuration versions, and a record of all preprocessing and label decisions.

A. Image acquisition and preprocessing

For DICOM CT, read pixel data and rescale using the modality-specific RescaleSlope and RescaleIntercept to obtain Hounsfield units (HU). Confirm orientation, spacing, reconstruction kernel, contrast status, slice thickness, and missing/corrupt slices. If NIfTI or MetaImage is used, preserve affine/orientation metadata and verify that voxel coordinates map correctly to physical space. Do not apply arbitrary histogram equalization as a substitute for consistent HU handling.

A common lung window is approximately −1000 to 400 HU, but windowing parameters must be fixed and reported. Resample volumes to a chosen voxel spacing only when appropriate for the task; interpolate images and masks with methods suitable for each data type. Lung segmentation can reduce irrelevant background, but it must not remove pleural nodules or lesions near the mediastinum. Candidate generation may use a detector, radiologist annotations for research crops, or a benchmark-provided candidate list; these alternatives change the task and must be documented.

B. Patient-level splitting and augmentation

Create training, validation, and test partitions by patient, never by individual slice or crop. All scans, repeated studies, and nodules belonging to one patient must remain in a single partition. Where feasible, reserve an institution or time period for external validation. Augmentation is applied to training samples only and may include small rotations, flips when anatomically valid, modest scaling, translation, noise, and intensity perturbations. Transformations must preserve clinical plausibility and coordinate labels.

Class imbalance can be addressed through class-weighted loss, balanced mini-batches, focal loss, or carefully controlled resampling. Report the natural class prevalence in validation and test data; oversampling the test set invalidates prevalence-dependent metrics and calibration. Any synthetic or augmented sample must remain derived only from training patients.

C. CNN feature extraction and transfer learning

A baseline should be deliberately simple: a small CNN with convolution, normalization, ReLU, pooling, and a global pooling classifier. Stronger comparators may include ResNet [13], DenseNet [14], and EfficientNet [15]. For volumetric

analysis, a 3-D CNN can preserve spatial context but increases memory and data requirements. A 2.5-D model offers a compromise by stacking adjacent slices. CNN–Transformer hybrids may capture longer-range relationships, but should be justified by a controlled comparison rather than assumed superior.

For pretrained 2-D networks, adapt the first layer to CT input channels and fine-tune using a low learning rate. Compare frozen-backbone and full fine-tuning variants. Pretraining source, checkpoint version, layer-freezing schedule, input size, and initialization seed should be reported. If a model is trained from scratch, justify sample size and regularization. Patient-level clinical covariates can be fused late with image features, but this requires an explicitly defined multimodal baseline and missing-data plan. Transfer learning for medical image analysis should be evaluated empirically rather than assumed to outperform full training [23].

D. Mathematical formulation

For input feature map X and kernel K, a convolutional feature map at location (i,j) can be written as:

$$Z(i,j) = \Sigma_m \Sigma_n X(i+m, j+n) K(m,n) + b$$

Here m,n index the kernel, b is a bias term, and channel summation is implicit. A nonlinear rectified linear unit is:

$$\mathrm{ReLU}(z) = \max(0, z)$$

Pooling reduces spatial resolution and can provide local translation tolerance. For a pooling window Ω:

$$P(i,j) = \max_(u,v \in \Omega)\ Z(i+u, j+v)$$

For a binary target, the sigmoid output is$p = 1/(1 + e^{(} - z))$; a multiclass output uses softmax:

$$p_k = exp(z_k)/\Sigma_j exp(z_j)$$

Binary cross-entropy for N examples is:

$$L_B CE = -(1/N)\Sigma_i [y_i log(p_i) + (1 - y_i) log(1 - p_i)]$$

For class-weighted training, multiply each example's loss by a weight based on the training-set class distribution. Weights must be determined without using the test set. Optimization may use AdamW or stochastic gradient descent; the optimizer, learning-rate schedule, weight decay, batch size, stopping rule, and random seed must be logged.

E. Training, selection, and reproducibility

A reasonable starting configuration is provided in Table IV as a proposed setting only. Tune a small, predeclared search space using validation data. Use early stopping on a preselected metric and save the best checkpoint. Report mean and dispersion across multiple seeds or patient-level folds where computationally feasible. Lock the pipeline before final test evaluation; do not repeatedly inspect the test set to make design decisions. Adam is a commonly used adaptive optimizer [21], while focal loss was introduced to address foreground/background imbalance in dense detection [22]; these are candidate tools rather than evidence that one setting is optimal for this dataset.

F. Explainability and risk output

Grad-CAM uses gradients flowing into a convolutional layer to produce a coarse localization map [16]. In this application, the map can be overlaid on a CT slice to inspect whether the classifier attends to the candidate nodule or to confounding

structures. A heatmap is not a proof of causal reasoning, and a visually plausible explanation does not establish clinical validity. Use model randomization checks, perturbation tests, expert review, and failure-case analysis.

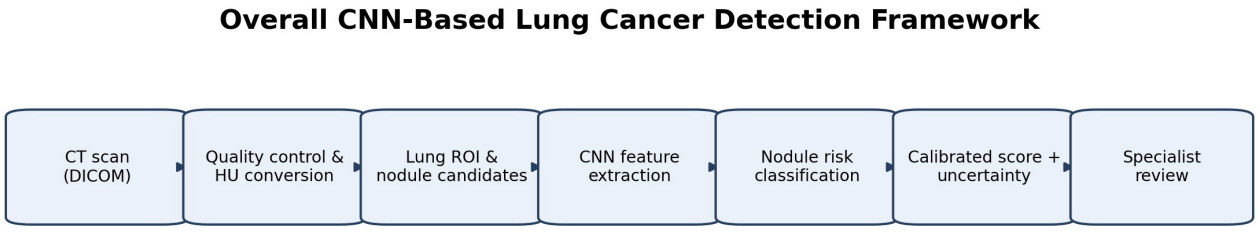


Fig. 1. Overall CNN-based lung cancer detection framework. The risk output is intended for decision support and requires specialist review.

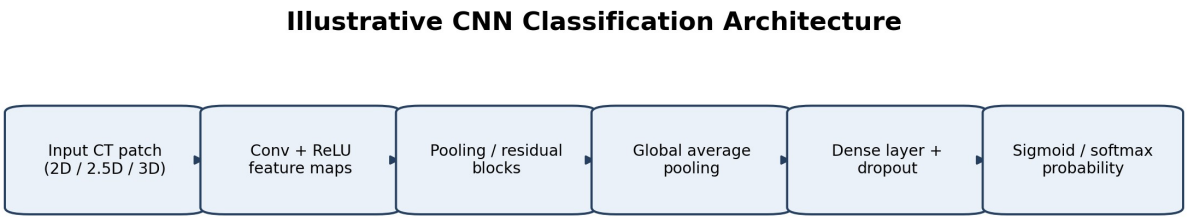


Fig. 2. Illustrative CNN architecture for nodule classification. Layer counts and dimensions are to be selected and reported during implementation.

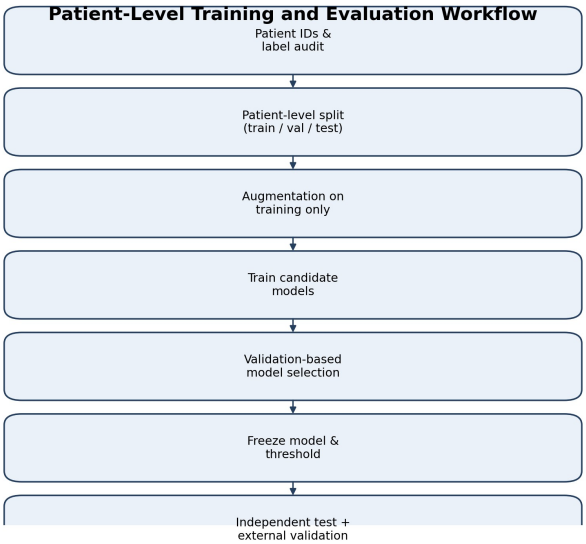


Fig. 3. Proposed training and evaluation workflow. Patient-level separation precedes augmentation and model fitting.

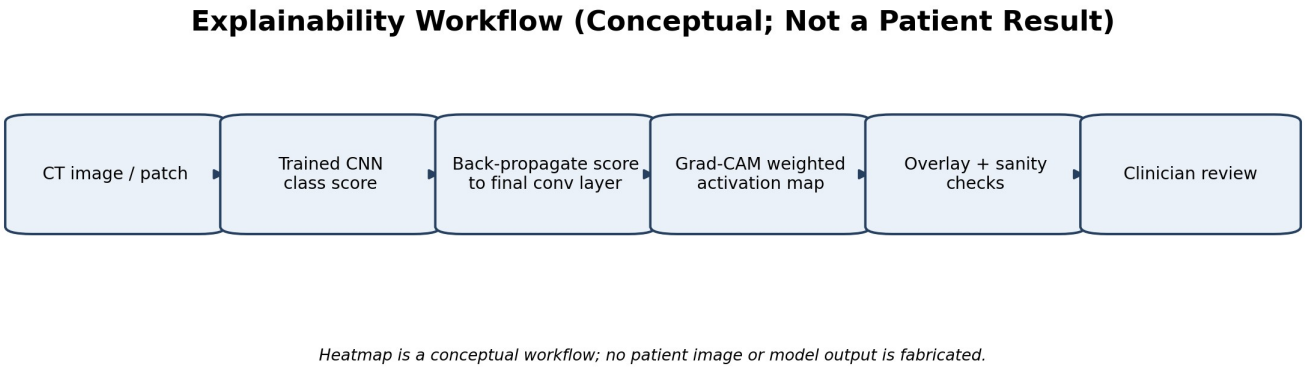


Fig. 4. Conceptual Grad-CAM workflow. This figure is a method diagram, not a heatmap from a real patient or trained model.

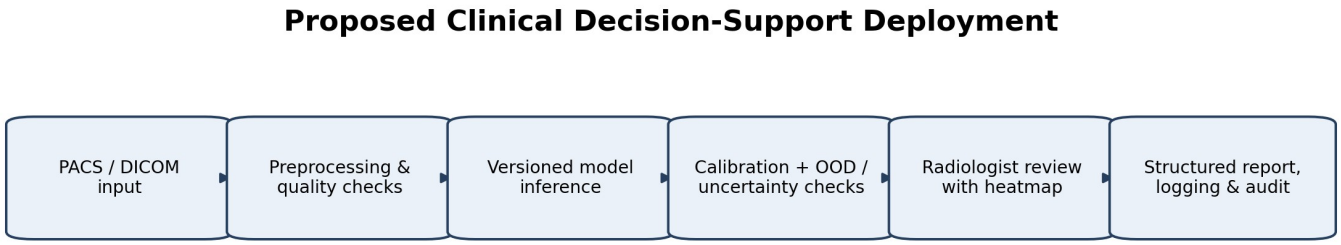


Fig. 5. Proposed deployment architecture with quality checks, uncertainty handling, clinician review, and audit logging.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION CONSIDERATIONS

The architecture is modular: (1) an ingestion layer validates DICOM metadata and access permissions; (2) preprocessing converts intensities and standardizes geometry; (3) candidate detection identifies regions for analysis; (4) a CNN estimates candidate-level features and probabilities; (5) calibration and out-of-distribution checks determine whether the output is reliable enough to display; and (6) a clinician-facing interface presents candidate coordinates, probability, heatmap, and supporting metadata. The design should retain traceability to source images and the exact model version.

In production, inference should be separated from the picture archiving and communication system (PACS) through controlled interfaces, authentication, encryption, audit logs, and monitoring. The model must not silently overwrite a radiologist's interpretation. A report should distinguish an algorithmic suggestion from a confirmed diagnosis and record whether the clinician accepted, modified, or rejected the suggestion. Deployment is not justified by a benchmark score alone. Clinical reporting conventions such as Lung-RADS provide a management-oriented context but are not substitutes for validating the model's intended use [25].

VI. EXPERIMENTAL SETUP

Architecture	Strengths	Trade-offs	Suggested role
Custom 2-D CNN	Simple, low compute, easy baseline	Limited context; sensitive to design choices	Baseline for cropped axial images
ResNet [13]	Residual connections support deeper networks	More parameters than a small custom CNN	Transfer-learning comparator
DenseNet [14]	Feature reuse and dense connectivity	Memory use can rise with depth	Small-data comparator where memory permits
EfficientNet [15]	Compound scaling; efficient variants	Pretraining/input adaptation still matters	Resource-constrained baseline
3-D CNN	Uses volumetric context	Higher memory and data requirements	Nodule patches/volume-level analysis
CNN–Transformer hybrid	Combines local features and longer-range context	Higher complexity; needs careful ablation	Exploratory model after strong CNN baselines

Table 3. CNN architecture comparison (qualitative)

The proposed experiment uses LIDC-IDRI for model development and LUNA16 for a detection-specific benchmark where appropriate; LNDb can be used as an external dataset only after confirming compatible annotations, access rules, and task definition. These resources should not be combined into one label pool without harmonizing the target and reference standard. If the study aims to predict malignancy, a dataset with pathology or longitudinal outcome labels is required; nodule ratings alone are insufficient.

In production, inference should be separated from the picture archiving and communication system (PACS) through controlled interfaces, authentication, encryption, audit logs, and monitoring. The model must not silently overwrite a radiologist's interpretation. A report should distinguish an algorithmic suggestion from a confirmed diagnosis and record whether the clinician accepted, modified, or rejected the suggestion. Deployment is not justified by a benchmark score alone. Clinical reporting conventions such as Lung-RADS provide a management-oriented context but are not substitutes for validating the model's intended use [25].

Patient-level leakage occurs when images from one patient appear in more than one split. Other leakage routes include duplicate scans, derived crops crossing partitions, preprocessing fitted on all data, and selection based on test outcomes. Use a patient-level manifest, hash duplicate files, and document how repeated examinations are handled. Radiologist annotations can disagree; preserve individual annotations where possible or predefine a consensus rule rather than implying that all nodules have a single unquestionable label.

Setting	Proposed value	Rationale / reporting requirement
Development data	LIDC-IDRI; task-specific labels	Record exact release, case IDs, annotation policy
External evaluation	LNDb where task-compatible	Keep external cohort untouched during tuning
Partition	Patient-level 70/15/15 or fixed benchmark folds	Freeze split manifest; do not split slices
Input	CT crop; compare 2-D, 2.5-D and/or 3-D	Report physical spacing and tensor dimensions
HU window	Predeclared lung window; validate sensitivity	Document clipping, normalization, interpolation
Optimizer	AdamW, initial LR 1e−4 (starting point)	Tune only on validation set
Batch size	8–32 depending on input dimensionality and memory	Report actual value and hardware
Epochs	Up to 100 with early stopping	Report stopping metric and selected epoch
Loss	Binary cross-entropy; class weighting if needed	Compute weights on training partition only
Random seed	42 as a declared initial seed; repeat seeds if feasible	Seed does not guarantee bitwise determinism
Software/hardware	To be recorded at execution	No hardware or runtime is claimed in this paper

Table 4. Proposed experimental configuration (not executed)

VII. RESULTS AND DISCUSSION

No new experiment was run for this paper. Consequently, this section provides a literature-based interpretation and a plan for reporting results rather than invented performance values. A valid empirical study should report the number of patients, scans, nodules, positive cases, excluded cases, and missing labels separately. A flow diagram should document all exclusions and the final partition counts.

At a fixed decision threshold, define TP, TN, FP, and FN as true positives, true negatives, false positives, and false

negatives. Accuracy alone can conceal clinically important errors when malignancy prevalence is low or class imbalance is substantial. Sensitivity measures the fraction of positive cases detected; specificity measures the fraction of negative cases correctly rejected. Precision reflects the proportion of positive predictions that are true positives. ROC-AUC summarizes ranking over thresholds but does not specify a usable clinical threshold or guarantee calibration.

Metric	Formula	Interpretation / caution
Accuracy	(TP+TN)/(TP+TN+FP+FN)	Can be misleading under class imbalance
Sensitivity / Recall	TP/(TP+FN)	Fraction of positive cases detected
Specificity	TN/(TN+FP)	Fraction of negative cases correctly rejected
Precision / PPV	TP/(TP+FP)	Depends on prevalence and threshold
F1-score	2·Precision·Recall/(Precision+Recall)	Balances precision and recall; ignores TN
ROC-AUC	Area under TPR vs. FPR curve	Threshold-independent ranking; not calibration
False positives per scan	Number of FP candidates / number of scans	Important for detection workflow burden
Calibration	Compare predicted risk with observed frequency	Use reliability plot, Brier score or calibration slope

Table 5. Evaluation metrics and formulas

$$\text{Sensitivity} = TP / (TP + FN) \quad (6)$$

$$\text{Specificity} = TN / (TN + FP) \quad (7)$$

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (9)$$

$$FPR = FP / (FP + TN) \quad (10)$$

Thresholds should be chosen according to the intended use and validation evidence, not optimized on the test set. In screening support, missed malignancy and excessive false-positive workups have different consequences; report sensitivity and false positives per scan together. Confidence intervals should be computed at the patient level, for example using patient-level bootstrap resampling. When comparing models, use the same cohort, split, labels, and operating point; a difference between published values from unrelated datasets is not a controlled model comparison.

Study / year	Dataset / cohort	Model / task	Reported metric	Evaluation notes / limitations
Pyramid network + honey-badger method, 2024 [6]	LIDC-IDRI and NIH chest X-ray dataset as described	Proposed deep-learning classification pipeline	Authors report 98.5% accuracy for normal and 97.44% for abnormal lungs	Do not compare directly with CT-only nodule detection; mixed modality and task-specific setup
External malignancy meta-analysis, 2024 [4]	17 studies; 8,553 participants; 9,884 nodules	Deep-learning CADx for malignancy risk	Pooled comparisons reported; see original review for estimates and confidence intervals	Heterogeneous populations, reference standards, thresholds
Gao et al., 2025 [3]	Studies of pulmonary nodule detection/segmentation	Systematic review	No single pooled score used here	Metrics and protocols vary across studies
Current proposed framework	No experimental dataset evaluated in this paper	CNN baseline + stronger comparators proposed	Not reported — no new model training performed	Requires implementation, locked test set, and external validation

Table 6. Performance comparison of published methods (evidence-level summary)

A confusion matrix should be populated only after a locked evaluation and should include the threshold and unit of analysis (candidate, nodule, or patient). ROC curves and precision–recall curves should be generated from actual held-out predictions. Report inference latency and parameter count only after measuring them on stated hardware and software. A robust study should also compare the model with radiologist performance or an established clinical risk model under an appropriate blinded protocol.

VIII. EXPLAINABILITY, ROBUSTNESS, AND CLINICAL CONSIDERATIONS

A. Leakage and label quality

Patient-level leakage occurs when images from one patient appear in more than one split. Other leakage routes include duplicate scans, derived crops crossing partitions, preprocessing fitted on all data, and selection based on test outcomes. Use a patient-level manifest, hash duplicate files, and document how repeated examinations are handled. Radiologist annotations can disagree; preserve individual annotations where possible or predefine a consensus rule rather than implying that all nodules have a single unquestionable label.

B. Domain shift, calibration, and uncertainty

Scanner vendors, reconstruction kernels, slice thickness, dose, contrast protocols, patient demographics, smoking history, and referral patterns can change the data distribution. Evaluate performance by site and relevant subgroups where sample size permits. Calibration matters when a probability is presented as risk: a model that ranks cases well can still systematically overestimate or underestimate malignancy. Assess calibration on validation data and confirm it externally. An abstention mechanism should refer low-quality or out-of-distribution scans for ordinary clinical review rather than producing false certainty.

C. Clinical impact, privacy, and governance

False negatives may delay follow-up; false positives may cause anxiety, additional imaging, invasive procedures, and cost. A decision-support tool therefore requires a defined intended use, user interface evaluation, prospective workflow assessment, and monitoring after deployment. Data access must comply with institutional approvals, consent or waiver requirements, de-identification standards, and applicable law. .

IX. RESEARCH GAPS AND FUTURE DIRECTIONS

Gap	Why it matters	Proposed experiment / mitigation
Limited external validation	Internal scores may reflect site-specific characteristics	Train on one cohort; lock and test on independent institution(s)
Task and label mismatch	Nodule presence is not cancer diagnosis	Define task and use pathology/follow-up labels for malignancy claims
2-D context loss	A single slice can omit 3-D morphology	Compare 2-D, 2.5-D and 3-D models on the same patients
Class imbalance and calibration	Rare positives distort accuracy and probability estimates	Report PR-AUC, calibration, PPV/NPV at target prevalence
Weak explainability evidence	Heatmaps can be plausible but unfaithful	Use perturbation/randomization tests and expert assessment
Reproducibility gaps	Different splits and preprocessing block comparison	Release split IDs, configs, code, model hashes, and metric scripts
Clinical workflow uncertainty	Model benefit may not translate to improved care	Prospective silent-mode study, then controlled reader/workflow study
Compute and deployment burden	3-D or hybrid models may be difficult to deploy	Measure latency, memory, robustness, and cost on target hardware

Table 7. Research gaps and proposed solutions

Future experiments should start with a preregistered task and evaluation plan, establish a transparent baseline, and add complexity through controlled ablations. The most useful next step is likely not simply to add a larger network, but to determine whether a model improves sensitivity at a tolerable false-positive rate on an independent cohort and whether its output is calibrated and interpretable to clinicians. Prospective studies should assess workflow impact, not only image-level discrimination.

X. CONCLUSION

CNNs provide a flexible foundation for pulmonary nodule detection and CT image analysis, with 2-D, 2.5-D, and 3-D designs offering different trade-offs in context, compute, and data requirements. Public resources such as LIDC-IDRI, LUNA16, and LNDb enable research, but their annotations and intended tasks differ; in particular, nodule detection labels must not be presented as histopathologically confirmed cancer labels. Recent reviews find promising deep-learning performance while emphasizing heterogeneous methods and the need for external validation. This paper has presented a reproducible framework spanning image preprocessing, patient-level splitting, model selection, loss functions, metrics, explainability, calibration, and clinical governance. No new model was trained, and no original performance result is claimed. Future work should prioritize independent multicentre evaluation, transparent reporting, calibrated risk estimates, and prospective evidence that decision support improves clinical practice. CNN systems should complement, not replace, specialist judgement unless and until robust clinical evidence and applicable regulatory requirements support a defined use.

REFERENCES

[1] World Health Organization, “Lung cancer,” Fact sheet, Apr. 16, 2026. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/lung-cancer

[2] International Agency for Research on Cancer, “Lung cancer,” IARC, 2025. [Online]. Available: https://www.iarc.who.int/cancer-type/lung-cancer/

[3] C. Gao, L. Wu, W. Wu, Y. Huang, X. Wang, Z. Sun, M. Xu, and C. Gao, “Deep learning in pulmonary nodule detection and segmentation: a systematic review,” European Radiology, vol. 35, no. 1, pp. 255–266, 2025, doi: 10.1007/s00330-024-10907-0.

[4] “Deep Learning Models for Predicting Malignancy Risk in CT-Detected Pulmonary Nodules: A Systematic Review and Meta-analysis,” Lung, vol. 202, no. 5, pp. 625–636, 2024, doi: 10.1007/s00408-024-00706-1.

[5] “Deep Learning Techniques for Lung Cancer Diagnosis with Computed Tomography Imaging: A Systematic Review for Detection, Segmentation, and Classification,” Information, vol. 16, no. 6, Art. no. 451, 2025. [Online]. Available: https://www.mdpi.com/2078-2489/16/6/451

[6] “Lung cancer detection via deep learning-based pyramid network with honey badger algorithm,” Measurement: Sensors, vol. 31, Art. no. 100993, Feb. 2024, doi: 10.1016/j.measen.2023.100993.

[7] J. Pedrosa et al., “LNDb v4: pulmonary nodule annotation from medical reports,” Scientific Data, 2024, doi: 10.1038/s41597-024-03345-6.

[8] S. G. Armato III et al., “The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI):

A completed reference database of lung nodules on CT scans," Medical Physics, vol. 38, no. 2, pp. 915–931, 2011, doi: 10.1118/1.3528204.

[9] The Cancer Imaging Archive, "LIDC-IDRI collection," National Cancer Institute. [Online]. Available: https://www.cancerimagingarchive.net/collection/lidc-idri/

[10] A. A. A. Setio et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," Medical Image Analysis, vol. 42, pp. 1–13, 2017, doi: 10.1016/j.media.2017.06.015.

[11] LUNA16 Challenge, "Data." [Online]. Available: https://luna16.grand-challenge.org/Data/

[12] LNDb Challenge, "Data description." [Online]. Available: https://lndb.grand-challenge.org/Data/

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[14] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.

[15] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. Machine Learning (ICML), PMLR, vol. 97, pp. 6105–6114, 2019.

[16] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Advances in Neural Information Processing Systems, vol. 25, 2012.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[19] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[20] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015.

[22] T.-Y. Lin et al., "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.324.

[23] N. Tajbakhsh et al., "Convolutional neural networks for medical image analysis: Full training or fine tuning?" IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1299–1312, 2016, doi: 10.1109/TMI.2016.2535302.

[25] American College of Radiology, "Lung-RADS: Assessment categories and management recommendations," ACR, current edition. [Online]. Available: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads