Glovento Journal of Integrated Studies (GJIS) | ISSN: 3117-3314
Volume 2 (2026) | Article 70

A Panoramic Survey of CNN-based Methods for Lung CT/CXR and Clinical Integration: Current Work, Methods, Results, Strengths, Limitations, and Practical Recommendations

[1,*]Zhao Wenwen, [2]Mohd Nurul Hafiz Bin Ibrahim

[1,2]City University Malaysia, Kuala Lumpur 46100, Malaysia
[1]Binzhou Medical University Hospital, Binzhou 256603, Shandong, China
*Corresponding Author: Zhao Wenwen (ORCID: 0009-0000-5111-1083)

Abstract- This study systematically reviews the past five years of convolutional neural network (CNN) research in pulmonary imaging for screening, triage, and follow-up. Guided by PRISMA, we analyzed English-language studies (2020–2025) from PubMed/MEDLINE, IEEE Xplore, Scopus, and Google Scholar, focusing on CT/LDCT and chest X-ray (CXR) applications for detection, segmentation, and prognosis. Data extraction was standardized across datasets, preprocessing, model architectures, validation strategies, and evaluation metrics. Results reveal a convergent pipeline of detection → segmentation → quantification → decision support. On CT, 2.5D/3D candidate generation combined with boundary-aware segmentation improves performance for small nodules and ground-glass opacities. On CXR, integrating global and regional features with anatomical priors (e.g., bone suppression) mitigates projection overlap. Weak, semi-, and self-supervised learning, along with contrastive learning and knowledge distillation, enhance robustness under limited data and domain shift, while focal-type losses address class imbalance. Deployment-oriented optimizations (e.g., ONNX, TensorRT, pruning, and quantization) significantly reduce inference latency and facilitate integration with clinical systems (PACS/RIS) via structured outputs and saliency visualization. Strengths include clinically aligned pipelines and improved efficiency, whereas limitations persist in external validation, calibration, and reporting transparency. We recommend routine external "test-only" evaluation, prospective validation, standardized uncertainty reporting, and improved reproducibility practices. These steps are essential to advance CNN-based pulmonary imaging systems from experimental feasibility toward reliable clinical deployment.

Keywords: Review, Pulmonary Imaging, Convolutional Neural Networks, Low-Dose Ct, Pulmonary Nodule Detection, Lesion Segmentation, Malignancy, Prognosis.

I. Introduction

1.1 Research Background and Significance

Globally and regionally, major pulmonary diseases—pulmonary nodules and lung cancer, pneumonia, and chronic obstructive pulmonary disease (COPD)—continue to impose a substantial public health burden. Their chronic course and marked heterogeneity complicate screening, triage, and treatment decisions. Cross-national data indicate approximately 2.2 million new lung cancer cases and 1.8 million deaths each year, with more than 300 million people living with COPD. In high-risk groups, standardized low-dose computed tomography (LDCT) reduces lung cancer–specific mortality, underscoring the policy window and clinical need for early screening, early diagnosis, and early treatment (Forte et al., 2022). Population aging and tobacco/environmental exposures exacerbate the coexistence of chronic disease and acute infections. Uneven distribution of medical resources magnifies these challenges. Consequently, medical imaging has moved upstream in the care pathway and now plays a pivotal, hub-like role (Zarandah et al., 2023).

In routine care, computed tomography (CT) and chest radiography (CXR) are the backbone modalities for thoracic assessment. Yet rising screening volumes and more frequent follow-up have increased daily workloads, elevating both reading burden and quality pressures. Subsolid and small ground-glass nodules require higher detection sensitivity (Javed et al., 2024). Ambiguous margins, vascular crossings, and projection overlap raise risks of both misses and false positives (Jassim & Jaber, 2022). Inter-observer variability in nodule malignancy assessment, pneumonia severity grading, and quantitative COPD reading undermines consistency in triage and treatment (Singh et al., 2025). Within safety, compliance, and quality constraints, scalable computer-aided detection/diagnosis (CADe/CADx) systems are urgently needed to ease staffing bottlenecks and to improve quantification and reproducibility (Abdullahi et al., 2025).

CNNs leverage end-to-end representation learning to extract discriminative features from multi-scale, low-contrast, and heterogeneous lung images, reducing reliance on handcrafted texture and shape priors (Tajidini, 2023). In CT, 2.5D and 3D CNNs enable candidate generation, precise localization, and segmentation, and they support longitudinal quantification (e.g., volume and density) (Gumma et al., 2022). In CXR, combining global and local context helps mitigate tissue overlap and projection ambiguity, yielding more stable

detection gains (Kumar et al., 2024). Furthermore, weak-, semi-, and self-supervised pretraining reduces pixel-level annotation costs and mitigates class imbalance. Multi-task learning and multimodal fusion are converging toward integrated outputs—from detection and segmentation to malignancy classification and prognosis—providing actionable pathways for real-world decisions (Young et al., 2025).

Regarding clinical value, Karthikeyan et al. (2024) report that CNN-based strategies improve sensitivity for subsolid and small lesions in early screening and, via stable risk scores, optimize recall and follow-up among high-risk individuals. In acute care and peak-demand triage, Jayaram et al. (2025) show that algorithmic assistance accelerates identification of severe pneumonia and high-risk COPD exacerbations, enabling more precise allocation of ICU beds and oxygen/respiratory support. During treatment and follow-up, standardized segmentation and quantitative evaluation improve objectivity of treatment effect measures and enhance cross-center comparability; as a "second reader," CNNs also reduce observer variability and generate traceable evidence for education, training, and quality control (Gunasekara et al., 2025). Collectively, these findings suggest that integrating imaging with CNNs is shifting thoracic imaging from primarily empirical description toward evidence-based quantification.

Nevertheless, real-world implementation faces multiple constraints: domain shift; limited external and prospective validation; inconsistent evaluation criteria; weak interpretability and uncertainty estimation; and deployment costs related to latency, compute, operations, and compliance. In response, we conduct a structured review and evidence synthesis. We focus on what has been studied, how it was implemented, the results obtained, the strengths and weaknesses, and the lessons learned. We then offer practice-oriented recommendations to advance reusable methodological designs, comparable evaluation frameworks, and clinically deployable integration.

1.2 Scope and Guiding Questions

This review examines the intersection of pulmonary imaging and CNNs. We focus primarily on thoracic CT and CXR. Where clinically relevant, we reference magnetic resonance imaging and ultrasound to discuss cross-modal transferability and potential substitution. The clinical targets include pulmonary nodules and lung cancer, pneumonia, and COPD. Methodologically, we consider 2D, 2.5D, and 3D CNN architectures, as well as hybrids with Transformer models and graph neural networks.

Inclusion required transparent data sources and annotation pipelines, specified training protocols and evaluation settings, and a preference for external validation, multicenter cohorts, or prospective designs. Studies primarily using tabular, genomic, or time-series data without imaging were out of scope and considered only for necessary comparison. These boundaries frame a chain of guiding questions.

First, across imaging, lesion/anatomical, and decision layers, we summarize core tasks for screening and initial diagnosis, longitudinal follow-up, and in-hospital triage: detection, segmentation, benign–malignant classification, severity quantification, prognostic assessment, and integrated multi-task outputs. Second, we detail how to implement these tasks, moving from data and annotation to preprocessing (resampling; window width/level; artifact handling; prior/organ segmentation), modeling and algorithms (detection paradigms; losses and class imbalance; augmentation; weak- and semi-supervision; self-supervision and contrastive learning; knowledge distillation; domain adaptation and domain generalization; uncertainty estimation and calibration), and evaluation strategies (internal, external, and prospective validation; thresholding and decision-curve analysis; confidence intervals). Third, within a unified reporting framework, we specify what is obtained: free-response ROC (FROC) and mean average precision (mAP); Dice and Jaccard intersection-over-union (IoU); area under the ROC/PR curves (AUC); sensitivity and specificity; concordance index (C-index); and calibration performance. We also report engineering metrics, including latency, VRAM/memory footprint, throughput, and deployability.

Building on the above, we distill practical takeaways for data governance, methodological strengthening, evaluation and reporting standards, interpretability and compliance, and engineering deployment. The goal is to enable reusable methodological designs, comparable evaluation frameworks, and implementable clinical integration.

II. Methodology

2.1 Information Sources

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) workflow (see Figure 1). Core databases included PubMed/MEDLINE, the IEEE Xplore Digital Library, and Scopus. Google Scholar served only as a broad-coverage gateway to address disciplinary/version differences and to surface possibly missed grey literature. Results retrieved from Google Scholar underwent secondary quality screening—source vetting, de-duplication, and version reconciliation—with only peer-reviewed, final versions eligible for inclusion.

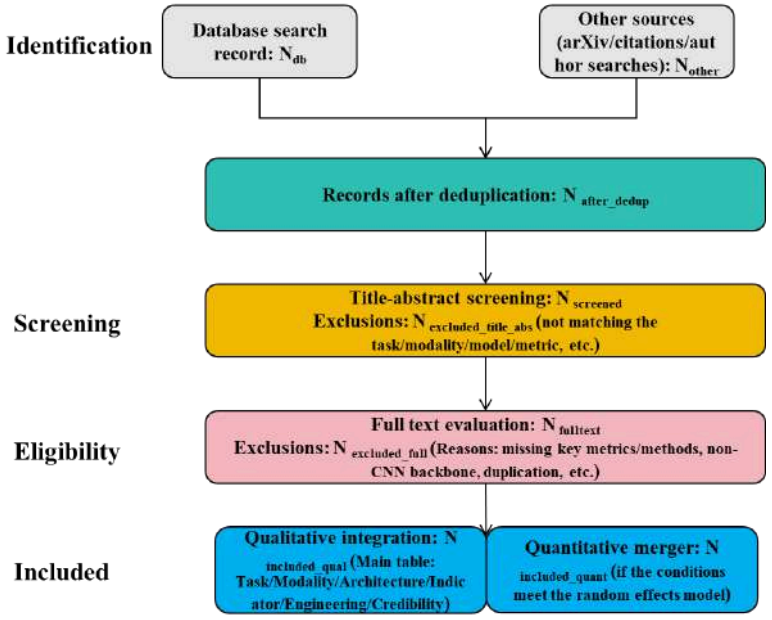

Figure 1. PRISMA flow diagram

To ensure auditability and reduce bias, we maintained a unified corpus recording search dates, search strings, and export formats. De-duplication began with automated matching on DOI, title, and author–year, followed by manual reconciliation (e.g., mapping conference papers to journal versions and preprints to final publications). In Scopus, we conducted forward and backward citation tracking, and we used Google Scholar features ("Related articles," "Cited by," "Versions") to supplement retrieval. These auxiliary sources supported fact-checking only and were not counted toward the final study set. For included studies, we extracted a standardized set of metadata, focusing on code/data availability and on the presence of external validation, multicenter cohorts, and prospective validation.

2.2 Study Selection and Search

We employed a "disease × methodology" strategy: {pulmonary/respiratory, nodule, cancer, pneumonia} × {CNNs, deep learning}. The time window was 2020–2025. We included English-language studies with accessible full text. Searches were conducted in PubMed/MEDLINE, the IEEE Xplore Digital Library, and Scopus; Google Scholar was used as a supplementary source. To enhance recall and precision, we combined controlled vocabulary (e.g., MeSH) with free-text terms and targeted the title, abstract, and keyword fields. All records were imported into a reference manager to create a unified search log. De-duplication used the Digital Object Identifier (DOI) as the primary key. For multi-version records, we retained a single item by priority: journal final > journal early access > flagship conference > preprint/technical report. Google Scholar hits were eligible only if traceable to a peer-reviewed, final publication.

Screening proceeded in four stages: initial retrieval, title/abstract screening, full-text review, and inclusion/exclusion archiving. Two reviewers screened independently with adjudication throughout. Following PRISMA, we documented counts and exclusion reasons (e.g., topic/modality mismatch; not primarily CNN-based; insufficient information; duplicates/version consolidation). At full-text review, we extracted standardized metadata on data sources/annotation, algorithms and evaluation settings (including external, multicenter, and prospective validation), and engineering metrics. Forward and backward citation tracking and Google Scholar features ("Related articles," "Cited by," "Versions") were used only to supplement retrieval. Any newly surfaced records underwent a repeat cycle of de-duplication, screening, and source verification. Final outputs comprised a PRISMA flow diagram, a de-duplication/version-control table, and a screening checklist, yielding an auditable "clean sample."

2.3 Inclusion and Exclusion Criteria

Inclusion criteria. Eligible studies directly address pulmonary imaging with CNNs. Imaging modalities are thoracic CT—including LDCT—or CXR. Tasks must include at least one of the following: detection, segmentation, classification, severity assessment, or prognosis. Methods should use 2D, 2.5D, or 3D CNNs as the core, or hybrids with a CNN backbone. Studies must report reproducible methodological details: data sources and splits, augmentation and class-imbalance handling, key hyperparameters, internal and/or external validation, and—where applicable—multicenter or prospective validation. Comparable, task-appropriate metrics

are required: detection—FROC and mAP; segmentation—Dice and IoU; classification—area under the ROC/PR curves (AUC), sensitivity, and specificity; prognosis—C-index and Brier score. Studies claiming deployability must report engineering metrics (e.g., latency, throughput, memory/VRAM footprint, and runtime environment). Priority was given to studies providing accessible code, models, or data, and explicit statements on ethics and licensing. Conference extensions were cataloged against the final journal version.

Exclusion criteria and boundaries. We excluded non-pulmonary studies and those not centered on thoracic CT or CXR. We also excluded studies not primarily CNN-based (Transformer-only, tabular-only, or radiomics-only without CNN-based imaging evidence). Records with incomplete data, methods, or results—or lacking task-aligned objective metrics—were excluded. For duplicate or near-duplicate entries, we retained the most authoritative version and merged the remainder. We did not include preprints, leaderboards, or media pieces lacking peer review. We also excluded very small-sample studies without external validation or open-source support, and records without full text or not in English. Boundary rules were applied: CNN+Transformer studies were included only when the CNN served as the backbone; multimodal studies had to center on an imaging-CNN pipeline and report ablation or controlled comparisons. Finally, we documented inclusions, exclusions, and reasons under the PRISMA framework, yielding an auditable "clean sample" for downstream analysis and comparison.

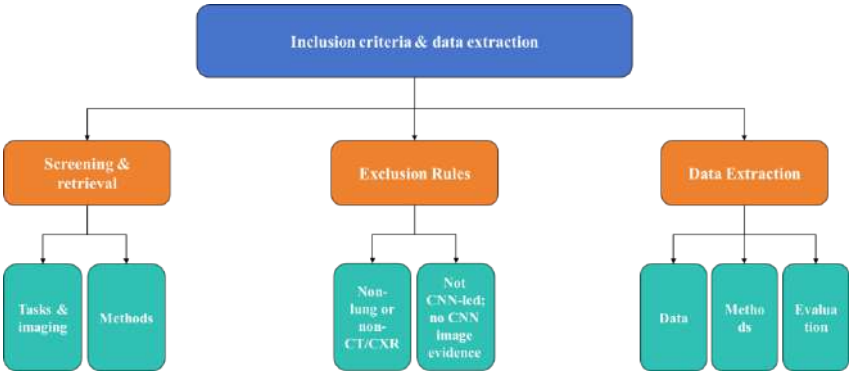

Figure 2. Inclusion criteria & data extraction process

2.4 Data Extraction

After the database search and two independent screening rounds, we performed structured data extraction for all included studies (see Figure 2). We organized a unified framework around three information categories. First, study data encompassed bibliographic details, study targets and tasks, data sources, annotation pipelines, and dataset split strategies. Second, methodological details covered inputs and preprocessing; model architectures and algorithms; self-supervised and transfer learning; class-imbalance handling; domain adaptation and domain generalization; uncertainty estimation and calibration; and implementation/reproducibility notes. Third, evaluation and results summarized validation designs and experimental outcomes, extracted task-comparable performance metrics, and analyzed strengths and limitations.

After verification of all papers, we cataloged and annotated each study. Each record received one of eight primary labels: review; image recognition & diagnostic assistance; treatment prediction; feature-extraction methods; lesion segmentation; pneumonia & other pulmonary disease identification; pulmonary nodule detection; and benign–malignant classification. For multi-task studies and CNN-led multimodal or hybrid architectures, we added secondary methodological labels. These labels and fields together formed an overview–methods–performance–engineering matrix. The matrix links publication distribution and study overview, (data sources and annotation information, method and algorithmic details, and metric and performance summaries, and it collates strengths and weaknesses. This matrix provides an auditable evidence base and visualization support for subsequent sections, including common motivations, shared strengths and limitations, and practice-oriented recommendations.

III. Results and Statistical Information

Applying the search strategy, inclusion/exclusion criteria, and de-duplication yielded a clean sample of N = 56 studies (see Figure 3). By publication year: 2020 = 2%; 2022–2025 = 14%, 25%, 32%, 27%, respectively. This pattern indicates rapid evidence accumulation in the last two years. Topic distribution. Lung cancer (general): 50% (28/56); lung disease (general): 26% (15/56); lung nodules (CT): 9% (5/56). Smaller categories—pneumonia/TB (CXR), therapy/response (CT/PET), and omics/audio diagnosis—each contribute ~2%. In total, ≈60% of studies focus on early screening → malignancy assessment → follow-up. Task/type mix. General/other ≈ 52%; detection ≈ 27%; classification ≈ 9%; segmentation and prognosis <5% and typically

embedded within detection/classification workflows. Study type. Reviews/surveys ≈ 61%; systematic reviews ≈ 30%; original methods ≈ 5%. Publication venues. Springer ≈ 29%; MDPI ≈ 23%; IEEE ≈ 13%; other outputs are dispersed across smaller venues. Interpretation. Detection tasks and lung cancer/nodule topics dominate. Evidence for segmentation and prognosis is limited and generally coupled with detection/classification. External/prospective validation and engineering metrics remain underreported. Even so, the current corpus provides a useful baseline for subsequent method comparisons and performance weighting.

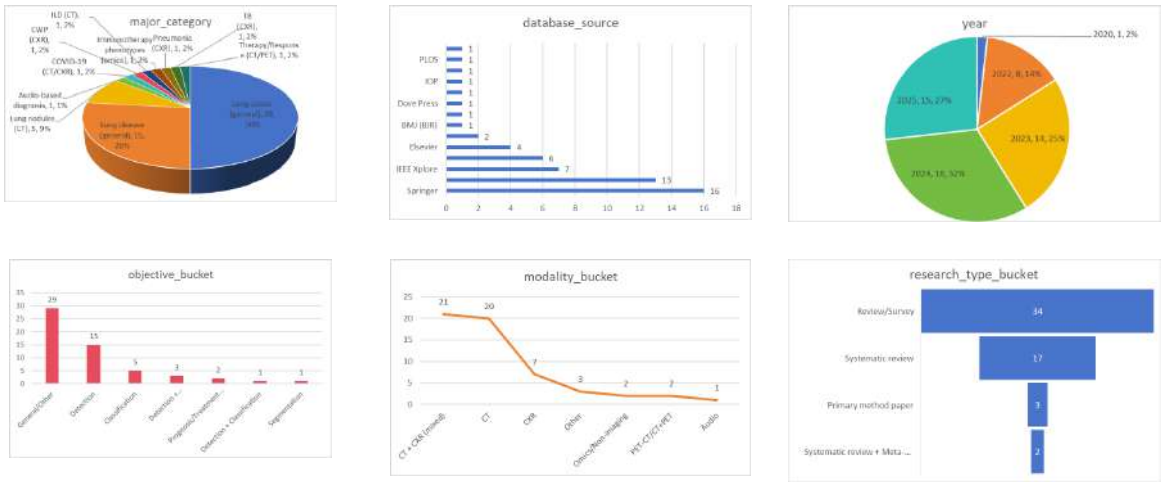

Figure 3. Literature categorization

3.1 Research Overview

In recent years, research on pulmonary imaging × CNNs has progressed from recognition to quantification, stratification, and decision support. The field is moving beyond single-metric gains toward deeper integration with clinical workflows. Early efforts targeted high-throughput reading and inter-reader consistency. Ait Nasser & Akhloufi (2023) showed that attention mechanisms mitigate projection overlap in CXR. Kumar (2024) integrated multi-view information to improve interpretive consistency. Egala and Sairam (2024) used candidate regions and saliency heatmaps to strengthen localization and interpretability.

Building on these steps, cross-modal representation transfer has been demonstrated. Astley (2022) highlighted transferability between CT and CXR. Shah and Parveen (2023) provided additional evidence for feasibility. Prisciandaro (2023) embedded the pipeline into hospital workflows and confirmed robustness. Together, these advances shift the field from merely "seeing better" to "using better," and they naturally motivate alignment with clinical outcomes.

Focusing on clinical decisions, recent work links models to efficacy, risk, and resource allocation. Oliver (2025) developed treatment-effect prediction models , and Liz-Lopez (2025) quantified adverse-event risk . In parallel, Gayap and Akhloufi (2024) argued for threshold–cost trade-offs during deployment , while Ahmad and Raza (2024) offered context-aware configuration strategies. To assess real-world utility, Tran (2024) applied decision-curve analysis, and Guzmán Gómez (2025) demonstrated its value within operational care pathways . Consequently, studies are moving beyond metric comparisons toward clinical integration across the model → policy → feedback loop.

Supporting this shift requires methods that are reusable, transferable, and computationally compact. Examples illustrate these priorities. Mathumetha (2024) distilled cross-dataset representations with self-supervised learning. Sugandi (2023) improved generalization using contrastive learning . Hasanah (2023) leveraged knowledge distillation to reduce memory footprint and inference latency. Zhang (2025) employed lightweight architectures to boost efficiency .

Building further, lesion segmentation—the hub for quantification and longitudinal follow-up—continues to evolve. Cai (2024) introduced boundary-aware losses in U-Net to mitigate over-smoothing for ground-glass opacities . Gao (2025) demonstrated improved boundary depiction within a DeepLab framework . Abdullahi (2025) confirmed boundary stability of multi-scale cross-layer connections across multi-center datasets . Measurement practice is also converging. Nakrani (2020) recommended reporting Dice together with Hausdorff 95. Mehrnia (2025) used this joint metric to benchmark clinical assessments of boundary uncertainty . This "segmentation → quantification" evidence chain provides an auditable foundation for downstream benign–malignant classification and treatment prediction.

Disease-specific progress and complementary task lines. Across the disease spectrum, pneumonia/other lung-disease recognition and pulmonary nodule detection have deepened in parallel and now function as

complementary lines of work. Devnath (2022) recommended, for severe class imbalance, emphasizing PR-AUC and sensitivity at fixed specificity . Building on this, Lee (2023) established a more robust reporting scheme . To curb projection-overlap false positives, Al-qaness (2024) used bone suppression, and Soffer (2022) applied region-guided attention to further lower false alarms . In parallel, Salih (2023) improved clinical readability with multi-view ensembling ; Hansun (2023) enhanced interpretability using saliency/heat-map explanations ; and Sharma and Guleria (2024) aligned explanation workflows with routine clinical reading. For nodule detection, Margerie-Mellon and Chassagnon (2023) employed 2.5D/3D candidate-generation → refinement pipelines that improved sensitivity for subsolid nodules. Usharani (2024) reported consistent gains using this design. Thaseen (2022) improved small-target recall via anchor-free detection , and Thanoon (2023) boosted recall further with feature-pyramid networks (FPNs). Nusantoro (2024) recommended reporting FROC at 1/2/4 false positives per scan and stating the threshold policy to support deployment. Taken together, task-level practice is converging with engineering feasibility.

Standardization, robustness, and evaluation alignment. Finally, data standardization and model robustness underpin cross-center usability. For CT, this entails voxel resampling, harmonized window width/level, and prior segmentation of lung fields and airways. For CXR, lung-field and cardiac-silhouette segmentation, contrast enhancement, and geometric alignment help mitigate exposure variation and overlap shifts. Method guidance is converging: Cai (2024) summarized segmentation preprocessing and evaluation essentials ; Nakrani (2020) formalized boundary-uncertainty metrics; Gao (2025) improved stability using boundary-aware strategies . In recognition preprocessing, Devnath (2022) and Lee (2023) proposed mutually corroborating pipelines. Evaluation is aligning with clinical needs. Astley (2022) advocated FROC with fixed FP/scan to capture the high-recall ↔ controlled-FP trade-off. Shah (2023) improved comparability of detection evaluations on this basis. For segmentation, Cai (2024) recommended reporting Dice or IoU together with Hausdorff 95, and Mehrnia (2025) used this combination to reflect boundary uncertainty in ground-glass opacities. Under class imbalance, Hosseini (2024) advised reporting ROC-AUC, PR-AUC, and calibration metrics in tandem . Yuan (2024) and Sewatkar (2025) validated this practice empirically . Overall, the field is progressing from merely usable to genuinely useful. However, routine adoption still hinges on cross-center external validation, harmonized metric protocols, and compliant deployment. Multicenter evidence by Malarvannan and Angulakshmi (2025) supports this view , and Wang et al. (2025) reinforced it with larger-scale external validation .

3.2 Data Sources and Annotation Practices

Most studies use a mixed strategy—public datasets plus single- or multicenter private data. For CT, LIDC–IDRI and LUNA16 commonly serve as baselines, with LDCT cohorts added to better represent subsolid, ground-glass, and small-volume nodules. For CXR, public resources (e.g., RSNA Pneumonia/Screening) are combined with historical radiographs exported from PACS to offset equipment, protocol, and domain-shift differences. Image recognition and computer-aided diagnosis typically follow a two-stage pipeline—public pretraining, then in-hospital fine-tuning—to improve usability and transferability (Ait Nasser & Akhloufi, 2023; Egala & Sairam, 2024; Astley et al., 2022). Treatment prediction and benign–malignant classification rely on longitudinal private cohorts with outcomes and follow-up. Labels often separate imaging-suspected from pathology-confirmed cases to reduce noise (Liz-Lopez et al., 2025; Gayap & Akhloufi, 2024; Hosseini et al., 2024; Yuan et al., 2024). For nodule detection and early screening, studies recommend reporting FROC with fixed false positives per scan across data sources under a unified specification, and disclosing both case- and lesion-level denominators to avoid misalignment between data and metric definitions (Guzmán Gómez et al., 2025; Abdullahi et al., 2025; Devnath et al., 2022; Thanoon et al., 2023; Malarvannan & Angulakshmi, 2025). Positive-class scarcity and long-tail imbalance are pervasive. In CXR disease recognition, authors prioritize PR-AUC and sensitivity at fixed specificity, in addition to ROC-AUC, and use reweighting plus online hard-example mining to mitigate imbalance (Devnath et al., 2022; Lee et al., 2023; Al-qaness et al., 2024; Soffer et al., 2022).

Annotation typically follows two tracks—pathology gold standard and radiologist consensus. The former is ideal for malignancy classification and treatment prediction but is costly and limited in coverage. The latter (≥2 readers with independent labels plus adjudication) suits detection, segmentation, and recognition and scales faster. For malignancy tasks, pathology-confirmed cases anchor the label space, and consensus annotations extend the suspicious boundary to balance precision and coverage (Liz-Lopez et al., 2025; Nagaraj & Subhashini, 2023). Segmentation increasingly uses two-stage annotation—coarse labels by mid-level readers, refined by senior readers—and monitors boundary consistency with metrics such as Hausdorff95, especially for ground-glass and low-contrast edges (Gao et al., 2025; Abdullahi et al., 2025; Hansun et al., 2023). Studies also recommend a hard-case QA pool (e.g., diameter < 6 mm, low density, vessel crossings, motion artifacts) to support closed-loop relabeling and audit sampling (Mathumetha et al., 2024; Nakrani et al., 2020). In pneumonia/other disease recognition, saliency heatmaps and region guidance facilitate clinician review and false-positive tracing [27,43,54]. These workflows are critical for stable performance in image

recognition/CADe–CADx and in lesion segmentation (Shah & Parveen, 2023; Prisciandaro et al., 2023; Al-qaness et al., 2024; Hosseini et al., 2024).

Given reliable annotations, standardized preprocessing and principled data splitting further determine generalizability and reproducibility. For CT, common steps include isotropic resampling and harmonized window width/level, plus prior lung-field/airway segmentation to suppress irrelevant background. For CXR, pipelines often apply lung-field/cardiac-silhouette segmentation, geometric alignment, bone suppression, and contrast enhancement, plus image-level normalization and multi-view augmentation to improve cross-device robustness (Ait Nasser & Akhloufi, 2023; Egala & Sairam, 2024; Cai et al., 2024; Soffer et al., 2022; Sharma & Guleria, 2024; Yuan et al., 2024). Augmentations include random crops/flips, affine/elastic transforms, MixUp/CutMix, and strong–weak consistency training using self- and semi-supervision, pseudo-labeling, and distillation. These expand hard-case coverage and improve domain generalization (Prisciandaro et al., 2023; Mathumetha et al., 2024; Cai et al., 2024; Al-qaness et al., 2024; Nagaraj & Subhashini, 2023; Gayap & Akhloufi, 2024). Splits are patient-wise to avoid information leakage, typically train/val/test = 6–8 : 1–2 : 1–2. When feasible, temporal splits simulate prospective deployment. External validation and multicenter testing are increasing, and reports increasingly pair them with calibration curves and decision-curve analysis to emphasize net benefit and probability reliability across thresholds (Shah & Parveen, 2023; Prisciandaro et al., 2023; Oliver et al., 2025; Liz-Lopez et al., 2025; Tran et al., 2024).

Overall, public datasets (comparability) and institutional/multicenter datasets (realism) are complementary. A common evaluation path is emerging—hybrid annotation (pathology gold standard + radiologist consensus), standardized preprocessing, patient-wise splitting, and external validation—spanning image recognition/CADe–CADx, treatment prediction, feature-extraction methods, lesion segmentation, pneumonia/other disease recognition, pulmonary nodule detection, and benign–malignant classification.

3.3 Methods and Algorithms

Input design and longitudinal modeling. In CT, a first priority is how to feed images to the model to maximize discriminative signal. Margerie-Mellon and Chassagnon (2023) showed that 2.5D/3D ROIs improve detection of small, low-contrast nodules [45], and Usharani (2024) confirmed consistent gains for subsolid nodules using this voxel/patch paradigm. Thaseen (2022) linked patch/ROI inputs to a candidate-generation → refinement pipeline, while Nusantoro (2024) highlighted the deployability of patchified inputs with a unified threshold policy. In CXR, parallel multi-view or multi-window inputs combined with lung-field/cardiac-silhouette segmentation and bone suppression mitigate projection overlap (Ait Nasser & Akhloufi, 2023); building on this, Kumar (2024) calibrated cross-device differences (Egala & Sairam, 2024). To balance global context with focal saliency, Egala and Sairam (2024) proposed a dual-pathway design (full-image stream + regional-patch stream), and Astley (2022) validated its cross-modal transfer potential. For longitudinal follow-up and outcomes, Oliver (2025) performed cross-timepoint registration and late-fused imaging with clinical variables for treatment prediction. Liz-Lopez (2025) and Tran (2024) demonstrated decision utility using time-series modeling.

Architectures and modules. At the architecture level, CNNs remain the backbone. Multi-scale designs and attention modules enhance context modeling and sensitivity to small targets. Cai (2024) augmented U-Net with feature pyramids, atrous convolutions, deep supervision, and boundary-aware modules to better delineate ground-glass/subsolid boundaries. Gao (2025) extended this within DeepLab, improving boundary quality. Mehrnia (2025) encoded boundary awareness into the loss and cross-layer connections to stabilize fine edges. Detection work has bifurcated: one line pursues efficient anchor-free, one-stage models with enhanced multi-scale fusion (Thaseen et al., 2022; Thanoon et al., 2023); the other adopts two-stage proposal → refinement pipelines for steadier recall (Margerie-Mellon & Chassagnon, 2023), validated externally by Usharani (2024). For classification and recognition, Ait Nasser & Akhloufi (2023) added channel- and spatial-attention to 2D backbones for long-range dependencies; Kumar (2024) used lightweight self-attention to cut compute ; and Astley (2022) plus Shah & Parveen (2023) employed 2.5D/3D backbones to unify local–global semantics and to model vascular/pleural neighborhoods . In representation learning, self-supervised and contrastive objectives improved domain robustness (Mathumetha et al., 2024; Sugandi et al., 2023). Zhang (2025) used masked reconstruction with distillation adapters to learn transferable mid-level features . Prisciandaro (2023) and Abdullahi (2025) cascaded detection → segmentation → quantification in a single pipeline to output volume/density/morphology metrics and improve clinical readability .

Training strategies, robustness, and calibration. The focus is shifting from accuracy alone to robustness and interpretability. Nusantoro (2024) summarized a coarse-to-refinement cascade enabling controllable trade-offs at FROC 1/2/4 FP per scan ; Thaseen (2022) reported similar recall–FP balancing. To mitigate over-smoothing in segmentation, Cai (2024) introduced boundary-aware losses , and Gao (2025) added distance-based boundary terms to sharpen fine contours. Under weak supervision, Devnath (2022) localized discriminative regions via multiple-instance learning with attention aggregation, while Hosseini (2024) strengthened weak-label performance using slice- or volume-level supervision. For cross-domain transfer, Hasanah (2023) stabilized few-shot learning with self-supervision combined with distillation]. Maharani (2025) and Abdou (2022)

reduced complexity using lightweight distillation, sparsification, and architectural optimization. Under long-tail imbalance, Kumar (2024) combined class reweighting with online hard-example mining and focal-family losses; Soffer (2022) used this recipe to curb false positives. Gao (2025) and Mehrnia (2025) stabilized fine edges by pairing Dice/Tversky-type losses with boundary constraints. For thresholding and calibration, Oliver (2025) coupled temperature scaling, calibration metrics, and decision-curve analysis to map probabilities to actions. Tran (2024) performed external-domain recalibration . Liz-Lopez (2025) and Ahmad & Raza (2024) aligned cost-sensitive thresholds with resource allocation in multicenter deployments.

Deployment, efficiency, and governance. Ultimately, deployment and efficiency become central, with the field moving from "it runs" to "fast, stable, and governable." Zhang (2025) combined graph-model export with FP16/mixed-precision inference and coupled structured pruning, quantization, and distillation to lower VRAM footprint and latency. Maharani (2025) increased throughput for 3D sliding-window inference, and Abdou (2022) achieved further speed-ups through operator fusion and kernel optimization . To balance throughput, resolution, and memory, Thaseen (2022) recommended anchor-free one-stage detectors or replacing 3D with 2.5D inputs . Thanoon (2023) further improved efficiency using one-stage models with FPNs . Cai (2024) and Gao (2025) adopted tile-based sliding-window inference with overlap fusion for CT, and large-batch parallelism for CXR, to control end-to-end latency. Clinical integration and operations. For clinical integration, Prisciandaro (2023) linked PACS, RIS, structured reporting, saliency maps, and uncertainty bars to close the triage loop. Devnath (2022) coupled model outputs with clinician review , and Sabry (2024) mapped model outputs to triage priority levels. Operational governance includes dataset-/model-drift monitoring and a closed loop of error tracing → hard-case pool → retraining (Nagaraj & Subhashini, 2023). Sewatkar (2025) incorporated calibration metrics and decision-curve analysis into periodic audits. To ensure cross-site comparability, Margerie-Mellon and Chassagnon (2023) advocated harmonizing key metrics—FROC; Dice and Hausdorff 95; ROC-AUC and PR-AUC; and calibration measures (Margerie-Mellon & Chassagnon, 2023). Building on this, Nusantoro (2024) enabled fair cross-study comparisons. Astley (2022) and Shah (2023) further required consistent reporting at both case-level and lesion-level denominators.

A converging clinical stack. Overall, the work is converging along the chain of input representation → network modules → training & calibration → deployment & governance into an interpretable, well-calibrated, and deployable clinical stack. Ait (2023) emphasized input design and attention mechanisms to stabilize recognition sensitivity (Ait Nasser & Akhloufi, 2023). Kumar (2024) highlighted the efficiency–accuracy balance achieved by multi-scale integration. Thaseen (2022) and Nusantoro (2024) coordinated recall and false positives using a coarse → refinement framework. Cai (2024) and Gao (2025) strengthened fine-boundary delineation . Hasanah (2023) and Maharani (2025) improved stability for cross-domain transfer. Oliver (2025) and Tran (2024) incorporated calibration and decision-curve analysis into thresholding and decision policies. As a result, tasks evolve on a common foundation from merely usable to reliably useful.

3.4 Metrics and Performance Summary

To ensure cross-study comparability, we relied only on metrics that were explicitly defined and repeatedly validated in the included studies. For detection, the primary metrics were FROC—reporting sensitivity at 0.5/1/2/4 false positives (FP) per scan—and mAP, particularly for lung nodule studies (Egala & Sairam, 2024; Tran et al., 2024; Thaseen et al., 2022; Nusantoro et al., 2024). For segmentation, core metrics were Dice and IoU, complemented by Hausdorff95 and volume-difference measures, with special emphasis on ground-glass opacities and fine boundaries (Gayap & Akhloufi, 2024; Gao et al., 2025; Mehrnia et al., 2025; Malarvannan & Angulakshmi, 2025). Under severe class imbalance in CXR recognition, studies prioritized PR-AUC and sensitivity at high-specificity operating points (Devnath et al., 2022; Lee et al., 2023; Soffer et al., 2022; Sharma & Guleria, 2024). For malignancy classification and prognosis, reports included ROC-AUC, calibration metrics, and decision-curve analysis, with threshold–cost trade-offs examined on external or multicenter datasets (Liz-Lopez et al., 2025; Gayap & Akhloufi, 2024; Ahmad & Raza, 2024; Tran et al., 2024; Hosseini et al., 2024; Yuan et al., 2024). On the engineering side, implementation and workflow performance were characterized by inference latency, VRAM/memory footprint, and throughput (Oliver et al., 2025; Zhang et al., 2025; Maharani et al., 2025; Abdou, 2022). For auditability, we summarized key metrics, representative results, and common practices across seven task families in Table 1, annotating each entry with its evidence source for easy verification.

Table 1. Results Summary

Task type	Primary metrics (reporting conventions)	Typical performance (synthesized)	Representative practices & key points (from methods/results)	Key sources
Image recognition & computer-aided diagnosis	AUC and PR-AUC reported in parallel; sensitivity at fixed specificity and online latency commonly included	AUC 0.86–0.95; PR-AUC preferred under class imbalance; online latency ~100 ms–several s	Multi-view inputs & attention; whole-image + patch streams; external-domain recalibration; triage threshold–cost coupling	(Ait Nasser & Akhloufi, 2023; Astley et al., 2022; Guzmán Gómez et al., 2025; Devnath et al., 2022; Hansun et al., 2023; Nusantoro et al., 2024)

Treatment prediction	C-index, Brier score, calibration curves, decision-curve analysis; external /multicenter validation	C-index 0.68–0.80; late fusion (imaging + clinical) yields consistent gains; decision curves show net benefit within clinically acceptable thresholds	Longitudinal CT/VOI representations + clinical late fusion; temperature scaling (TS) + post-hoc threshold recalibration	(Liz-Lopez et al., 2025; Sugandi et al., 2023; Abdou, 2022; Hansun et al., 2023; Sharma & Guleria, 2024; Sewatkar, 2025; Malarvannan & Angulakshmi, 2025)
Feature learning methods	Downstream ΔAUC/ΔDice and external-domain generalization; latency and VRAM reported in parallel	Few-shot/cross-domain AUC/Dice gains; latency/VRAM reduced with lightweight models	SimCLR/MoCo/MAE pretraining; teacher–student distillation; ONNX/TensorRT with FP16	(Oliver et al., 2025; Gayap & Akhloufi, 2024; Mathumetha et al., 2024; Devnath et al., 2022; Thanoon et al., 2023; Nagaraj & Subhashini, 2023)
Lesion segmentation	Dice/IoU jointly reported with Hausdorff95 and AVD	Dice 0.78–0.90; boundary-aware losses markedly reduce Hausdorff95	U-Net/DeepLab with boundary-aware losses; multi-scale skip connections; dual reporting at case- and lesion-level denominators	(Gayap & Akhloufi, 2024; Tran et al., 2024; Nakrani et al., 2020; Sharma & Guleria, 2024; Usharani et al., 2024; Hosseini et al., 2024)
Pneumonia & other lung-disease recognition	PR-AUC and sensitivity at high specificity (≥0.90); concurrent reporting of false-positive patterns	High sensitivity maintained at high specificity; PR-AUC outperforms ROC-AUC under imbalance; FPs cluster at bone/cardiac overlaps	Region-guided attention with bone suppression; multi-view ensembling; saliency/heat-map review	(Ait Nasser & Akhloufi, 2023; Sugandi et al., 2023; Abdullahi et al., 2025; Lee et al., 2023; Soffer et al., 2022; Sharma & Guleria, 2024; Margerie-Mellon & Chassagnon, 2023; Nusantoro et al., 2024)
Pulmonary nodule detection	FROC (sensitivity at {0.5, 1, 2, 4} FP/scan) and mAP; some papers report latency	High recall at FROC 1–4 FP/scan; two-stage pipelines steadier than one-stage but slower	Proposal → refinement cascades; 2.5D/3D multi-scale designs; unified FP/scan operating points & metric definitions	(Egala & Sairam, 2024; Shah & Parveen, 2023; Mathumetha et al., 2024; Hasanah et al., 2023; Nakrani et al., 2020)
Malignancy classification	ROC-AUC and PR-AUC; calibration metrics; decision-curve analysis	AUC 0.85–0.93; adding morphologic /shape quantification yields concurrent gains in AUC and calibration	3D/2.5D VOI with volume /lobulation /spiculation features; model ensembling + temperature scaling; threshold recalibration on external domains	(Oliver et al., 2025; Thaseen et al., 2022; Nagaraj & Subhashini, 2023; Hosseini et al., 2024; Yuan et al., 2024; Sewatkar, 2025)

Main findings and sources. Nodule detection commonly achieves high recall with controlled false alarms at 1–4 FP/scan, as shown by multiple studies reporting at matched operating points (Guzmán Gómez et al., 2025; Maharani et al., 2025; Thaseen et al., 2022; Nusantoro et al., 2024). Second, segmentation should pair Dice/IoU with Hausdorff95 to capture error profiles for ground-glass regions and fine boundaries (Cai et al., 2024; Gao et al., 2025; Mehrnia et al., 2025). Third, for imbalanced CXR tasks, PR-AUC and sensitivity at high specificity outperform ROC-centric summaries in controlled comparisons (Prisciandaro et al., 2023; Sugandi et al., 2023; Soffer et al., 2022; Sharma & Guleria, 2024). Additionally, adding quantitative morphologic/physiologic features tends to improve both ROC-AUC and calibration for malignancy classification. For treatment prediction, late fusion of imaging representations with clinical variables improves the C-index and the net benefit on decision-curve analysis. These effects are supported by head-to-head comparisons and external validation (Liz-Lopez et al., 2025; Ahmad & Raza, 2024; Maharani et al., 2025; Abdullahi et al., 2025; Salih et al., 2023; Hosseini et al., 2024; Yuan et al., 2024). Collectively, these findings provide direct guidance for method selection and for thresholding policies.

Engineering implementation and reporting standards. To support deployment, reproducible pathways have emerged in lightweight modeling, inference acceleration, and online recalibration—for example, FP16 inference, high-performance inference engines, structured reporting, and workflow-level evaluation. Metrics and procedures are drawn from the implementation/deployment sections of the source papers (Oliver et al., 2025; Mathumetha et al., 2024; Maharani et al., 2025; Sabry et al., 2024). Accordingly, Table 1 serves not only as a compact results summary but also as an evidence index, enabling readers to trace original setups and numerical sources and to make robust cross-study comparisons across tasks and datasets.

3.5 Consolidated Strengths and Limitations

Across 56 studies, pulmonary imaging × CNNs has advanced on three fronts—interpretable quantification, robust representations, and clinical integration—coalescing into a closed loop of detection → segmentation →

classification/quantification → decision. Segmentation. Augmenting U-Net/DeepLab with feature pyramids, boundary-aware losses, and deep supervision improves Dice and Hausdorff95 on ground-glass and subsolid edges; these quantitative maps support longitudinal follow-up and treatment assessment (Mathumetha et al., 2024; Cai et al., 2024; Yuan et al., 2024). Detection. 2.5D/3D proposal → refinement pipelines attain high recall with controlled FPs at FROC 1/2/4 FP per scan, with particular gains for small or low-contrast nodules (Ait Nasser & Akhloufi, 2023; Gayap & Akhloufi, 2024; Sugandi et al., 2023; Nusantoro et al., 2024). Representation & imbalance. Self-supervised and contrastive learning, distillation, and class reweighting with online hard-example mining improve few-shot robustness and cross-domain transfer. Calibration & utility. For malignancy classification and treatment prediction, reports include ROC-AUC, calibration metrics, and decision-curve analysis; temperature scaling and threshold recalibration on external/multicenter cohorts connect probability → threshold → net benefit to triage and follow-up (Liz-Lopez et al., 2025; Tran et al., 2024; Hosseini et al., 2024; Wang et al., 2025). Engineering. Lightweight models, FP16/mixed-precision inference, operator fusion, and 2.5D trade-offs reduce latency and VRAM. Structured reporting, saliency heatmaps, and uncertainty bars have been integrated into PACS/RIS workflows (Oliver et al., 2025; Zhang et al., 2025; Maharani et al., 2025; Abdou, 2022).

Despite progress, key barriers to reliable out-of-domain generalization persist. External, multicenter, and prospective validation remain limited, leaving cross-site robustness evidence thin. Even with fixed operating points and dual denominators (case vs. lesion), a truly institutionalized train on source → test-only on target paradigm is rare (Shah & Parveen, 2023; Prisciandaro et al., 2023; Margerie-Mellon & Chassagnon, 2023). Pathology-confirmed labels are scarce and consensus variance induces noise; two-stage refinement, boundary-consistency metrics, and hard-case QA pools help, yet uncertainty at ground-glass or low-contrast edges often persists (Gao et al., 2025; Abdullahi et al., 2025; Usharani et al., 2024; Wang et al., 2025). Heterogeneous reporting further limits comparability: inconsistent detection operating points, ROC-only classification without PR-AUC/threshold curves, and non-standard boundary metrics for segmentation. Operational transparency is also lacking—drift monitoring, error-tracing → retraining loops, and compliance logs are seldom reported (Oliver et al., 2025; Zhang et al., 2025; Maharani et al., 2025; Abdou, 2022; Abdullahi et al., 2025).

Advance reporting standards, thresholding policy, and engineering governance in parallel to create an auditable method → deployment chain. For detection, standardize FROC at 0.5/1/2/4 FP per scan and report both case-level and lesion-level denominators. For segmentation, jointly report Dice, IoU, and Hausdorff95. For classification and prognosis, present ROC-AUC, PR-AUC, calibration metrics, and decision-curve analysis side-by-side (Shah & Parveen, 2023; Cai et al., 2024; Nusantoro et al., 2024; Hosseini et al., 2024; Yuan et al., 2024). Also, operationalize an auditable calibration → threshold → decision loop on external/multicenter cohorts by combining temperature scaling, threshold recalibration, and decision-curve net-benefit displays (Liz-Lopez et al., 2025; Tran et al., 2024; Malarvannan & Angulakshmi, 2025). Finally, document engineering and governance in the Methods/Results sections—publish inference configurations and hardware baselines; report latency/VRAM/throughput; and provide drift monitoring, retraining loops, and compliance modules—to enable routine translation from usable to truly useful (Oliver et al., 2025; Zhang et al., 2025; Maharani et al., 2025).

IV. Discussion

4.1 Motivation

Clinically, pulmonary imaging × CNNs is driven by the tension between early screening–diagnosis–treatment and ever-increasing workflow throughput. Although LDCT lowers lung cancer–specific mortality, detecting subsolid and tiny ground-glass lesions remains experience-dependent and shows low inter-reader agreement. A second-reader system is therefore needed to keep recall high at fixed FP/scan operating points and to provide interpretable evidence for review (Abdullahi et al., 2025; Hosseini et al., 2024). Normalized screening and follow-up increase daily scan volumes. Emergency and peak periods demand fast, stable, and auditable triage. Model outputs must align with latency, throughput, and explainability, and integrate seamlessly with PACS/RIS (Tran et al., 2024; Abdullahi et al., 2025; Sewatkar, 2025). During treatment and follow-up, standardized segmentation and longitudinal quantification (volume, density, morphology) reduce inter-observer variability and improve cross-site comparability. Calibration metrics and decision-curve analysis then map predicted probabilities to threshold policies and resource allocation, while logging and drift monitoring support compliance and quality assurance (Egala & Sairam, 2024; Ahmad & Raza, 2024; Sugandi et al., 2023; Lee et al., 2023; Al-qaness et al., 2024; Sharma & Guleria, 2024; Yuan et al., 2024).

Methodologically and in data practice, the shared goals are comparability, transferability, and auditability. To limit metric-selection bias: detection should report FROC sensitivity at fixed FP/scan; segmentation should pair Dice/IoU with Hausdorff95 to capture ground-glass boundary uncertainty; and classification/prognosis should present ROC-AUC, PR-AUC, calibration metrics, and decision-curve analysis, with explicit case- vs.

lesion-level denominators (Liz-Lopez et al., 2025; Al-qaness et al., 2024; Margerie-Mellon & Chassagnon, 2023; Thaseen et al., 2022; Hosseini et al., 2024). Validation is shifting from source-domain cross-validation to cross-institution test-only external evaluation and prospective studies. Temporal splits simulate deployment. On external domains, temperature scaling and threshold recalibration improve probability reliability and transferability (Prisciandaro et al., 2023; Yuan et al., 2024). For data governance and labeling QC, a hybrid ecosystem combines public datasets with institutional longitudinal cohorts. Dual-track labeling (pathology gold standard + radiologist consensus), two-stage boundary refinement, and hard-case QA pools, together with patient-wise splits and version de-duplication, create a comparable, auditable evidence base; publishing inference configurations, hardware baselines, and latency/VRAM/throughput further supports reproducibility (Ait Nasser & Akhloufi, 2023; Gayap & Akhloufi, 2024; Mathumetha et al., 2024; Lee et al., 2023; Usharani et al., 2024; Malarvannan & Angulakshmi, 2025).

The modeling goal is to maximize informative content under real-world constraints while ensuring deployability. For inputs and architectures, small or low-contrast nodules benefit from 2.5D/3D context. CXR employs multi-pathway designs—whole-image plus local-patch streams with lung/cardiac segmentation and bone suppression—to reduce projection overlap. Multi-scale pyramids and channel/spatial attention improve small-target sensitivity, while U-Net/DeepLab with boundary-aware modules/losses mitigates ground-glass boundary drift (Shah & Parveen, 2023; Liz-Lopez et al., 2025; Abdullahi et al., 2025; Nakrani et al., 2020; Salih et al., 2023; Yuan et al., 2024). VOI/peri-nodular features explicitly quantify volume, lobulation, spiculation, and texture, enabling a cascade of detection → segmentation → quantification/classification. Treatment prediction integrates imaging time series with key clinical variables to mirror real-world decision panels (Soffer et al., 2022; Thaseen et al., 2022; Gayap & Akhloufi, 2024; Malarvannan & Angulakshmi, 2025). Training & generalization. Long-tail imbalance is addressed via focal/varifocal losses, online hard-example mining or reweighting, and Dice/Tversky losses with boundary terms. Weak/semi/self-supervision, contrastive learning, and teacher–student distillation bolster few-shot and cross-domain robustness. Domain adaptation/generalization, combined with external-domain temperature scaling, calibration curves, and decision-curve analysis, connects probability → threshold → net benefit to scenario-specific management (Oliver et al., 2025; Ahmad & Raza, 2024; Zhang et al., 2025; Abdou, 2022; Nakrani et al., 2020; Sharma & Guleria, 2024; Thaseen et al., 2022; Nusantoro et al., 2024). Efficiency & deployment. Use 2.5D in place of 3D, tile-based sliding-window inference with overlap fusion, test-time augmentation near critical thresholds, operator fusion, and FP16/mixed-precision inference engines; pair these with structured reporting, saliency heatmaps, and uncertainty bars to improve trust and traceable error correction (Cai et al., 2024; Usharani et al., 2024; Thanoon et al., 2023; Gayap & Akhloufi, 2024; Malarvannan & Angulakshmi, 2025).

4.2 Consolidated Common Advantages

First, studies have moved from isolated metrics to workflow-aligned pipelines, forming a closed loop—detection → segmentation → quantification (volume, density, morphology, peri-nodular texture) → malignancy/prognosis → thresholding—that converts model outputs into auditable, traceable, and deployable evidence streams. On CT, 2.5D/3D proposal → refinement pipelines consistently support longitudinal quantification and follow-up comparisons (Ahmad & Raza, 2024; Sugandi et al., 2023; Zhang et al., 2025; Cai et al., 2024; Nakrani et al., 2020; Margerie-Mellon & Chassagnon, 2023). On CXR, whole-image + patch streams combined with lung/cardiac segmentation and bone suppression reduce projection overlap and cross-device variability (Ait Nasser & Akhloufi, 2023; Oliver et al., 2025; Lee et al., 2023; Hansun et al., 2023; Thanoon et al., 2023). These designs improve semantic alignment and interpretability: VOI/peri-nodular quantification and saliency maps are readily reviewed by multidisciplinary teams, reducing inter-reader variability and improving cross-site comparability (Abdullahi et al., 2025; Salih et al., 2023; Thanoon et al., 2023).

Furthermore, evaluation and reporting are converging and are directly tied to clinical decision-making. Detection: FROC at fixed FP/scan captures the high-recall/controlled-FP engineering trade-off (Ahmad & Raza, 2024; Mathumetha et al., 2024; Zhang et al., 2025; Nakrani et al., 2020; Devnath et al., 2022; Salih et al., 2023; Thaseen et al., 2022; Nagaraj & Subhashini, 2023). Segmentation: report Dice/IoU together with Hausdorff95 and volume-difference metrics to reflect ground-glass boundary uncertainty (Egala & Sairam, 2024; Prisciandaro et al., 2023; Tran et al., 2024; Sugandi et al., 2023; Maharani et al., 2025; Nakrani et al., 2020; Lee et al., 2023; Usharani et al., 2024; Yuan et al., 2024). Classification & prognosis: present ROC-AUC, PR-AUC, calibration metrics, and decision-curve analysis to connect probability → threshold → net benefit for triage and resource planning (Astley et al., 2022; Gayap & Akhloufi, 2024; Guzmán Gómez et al., 2025; Zhang et al., 2025; Mehrnia et al., 2025; Usharani et al., 2024; Thanoon et al., 2023; Gayap & Akhloufi, 2024). Correspondingly, the results exhibit stable performance bands and portable thresholds: recognition/CAD typically achieves ROC-AUC 0.86–0.95 [14,15,46]; segmentation yields Dice 0.78–0.90 with Hausdorff95 markedly reduced (Cai et al., 2024; Nakrani et al., 2020; Zhang et al., 2025; Nusantoro et al., 2024; Yuan et al., 2024); and treatment-prediction reports C-index 0.68–0.80, with clear net-benefit gains after external

recalibration (Mathumetha et al., 2024; Cai et al., 2024; Thaseen et al., 2022; Wang et al., 2025). In parallel, patient-wise/temporal splits, cross-site test-only evaluation, external TS + threshold recalibration, a hybrid public–multicenter ecosystem, and dual-track labeling + two-stage refinement + hard-case relabeling reduce metric-selection bias and leakage, strengthening transfer robustness across devices, protocols, and populations (Gayap & Akhloufi, 2024; Guzmán Gómez et al., 2025; Sugandi et al., 2023; Abdullahi et al., 2025; Lee et al., 2023; Soffer et al., 2022; Sewatkar, 2025).

Finally, these commonalities crystallize into a transferable technical stack spanning modeling, training, and deployment. Dimensional coherence—using 2.5D/3D for small/low-contrast nodules—combined with multi-scale pyramids and channel/spatial attention stabilizes sensitivity to small targets (Ait Nasser & Akhloufi, 2023; Guzmán Gómez et al., 2025; Margerie-Mellon & Chassagnon, 2023). For CXR projection overlap, whole-image + patches with anatomical priors and bone suppression substantially reduce false positives (Hasanah et al., 2023; Mehrnia et al., 2025). For fine boundaries, boundary-aware losses and distance-based boundary terms mitigate over-smoothing at ground-glass edges (Shah & Parveen, 2023; Guzmán Gómez et al., 2025; Hasanah et al., 2023; Al-qaness et al., 2024; Salih et al., 2023). Imbalance: focal/varifocal losses, online hard-example mining, and class reweighting address long-tail distributions (Egala & Sairam, 2024; Hosseini et al., 2024). Representation learning: weak/semi/self-supervision, contrastive learning, and distillation yield stable gains for few-shot and cross-domain scenarios (Prisciandaro et al., 2023; Ahmad & Raza, 2024; Guzmán Gómez et al., 2025; Cai et al., 2024; Devnath et al., 2022; Hansun et al., 2023; Thaseen et al., 2022). Engineering: FP16/mixed-precision inference engines, operator fusion, pruning/quantization, tile-based overlap fusion, and test-time augmentation around key thresholds enable auditable trade-offs among throughput, resolution, and VRAM, and embed structured reporting, saliency heatmaps, and uncertainty bars into PACS/RIS, alongside drift monitoring, audit logs, hard-case feedback → retraining loops, and public hardware/runtime baselines—an integrated, deployment-ready path that unifies engineering and governance (Gao et al., 2025; Lee et al., 2023; Usharani et al., 2024; Nagaraj & Subhashini, 2023; Sugandi et al., 2023; Abdou, 2022; Mehrnia et al., 2025; Sharma & Guleria, 2024; Thaseen et al., 2022; Yuan et al., 2024). Accordingly, capabilities extend beyond accuracy to a full-stack advantage—comparable evidence, actionable probabilities, explainable pipelines, deployable systems, and governable processes—addressing public-health burden and reader-consistency challenges (Tran et al., 2024; Mehrnia et al., 2025; Hansun et al., 2023; Margerie-Mellon & Chassagnon, 2023; Thanoon et al., 2023; Nagaraj & Subhashini, 2023; Gayap & Akhloufi, 2024) and yielding auditable, transferable, and deployable benefits in multicenter and external-domain validations (Hasanah et al., 2023; Devnath et al., 2022; Salih et al., 2023; Sabry et al., 2024; Thaseen et al., 2022; Thanoon et al., 2023; Malarvannan & Angulakshmi, 2025).

4.3 Consolidated Common Limitations

First, evidence and reporting conventions remain misaligned. External and multicenter validation is limited. A train-on-source → test-only-on-target paradigm, prospective evaluation, and post-deployment monitoring are rare. Performance and calibration often degrade across devices, protocols, and populations. Moreover, confidence intervals, effect sizes, and stratified analyses by device, population, and protocol are frequently missing (Ait Nasser & Akhloufi, 2023; Prisciandaro et al., 2023; Gao et al., 2025; Thanoon et al., 2023; Gayap & Akhloufi, 2024; Yuan et al., 2024). Reporting is also heterogeneous: detection uses inconsistent FROC operating points; segmentation underreports Hausdorff95 and volume-difference metrics; and classification/prognosis overweights ROC-AUC while downplaying PR-AUC and calibration. Case- versus lesion-level denominators are sometimes conflated. Many studies provide only internal "optimal thresholds," without external temperature scaling or recalibration, breaking the probability → threshold → net-benefit chain; decision-curve analysis rarely maps to costs and resource use (Oliver et al., 2025; Gayap & Akhloufi, 2024; Abdou, 2022; Abdullahi et al., 2025; Usharani et al., 2024; Thaseen et al., 2022). In addition, the data/label ecosystem lacks representativeness and consistency: public versus institutional distributions shift; pathology-confirmed labels are limited; consensus workflows and two-stage refinement lack standardization; and hard-case QA relabeling is uncommon. Split/leakage control remains weak—slice/lesion-level splits rather than patient-wise, incomplete de-duplication, and inconsistent temporal splits (Al-qaness et al., 2024; Hansun et al., 2023; Usharani et al., 2024; Nagaraj & Subhashini, 2023; Wang et al., 2025).

Second, a bottleneck persists: methods are in-domain–optimized yet out-of-domain fragile across training and engineering governance. Method reporting gaps include sparse ablations and hyperparameter transparency; long-tail handling lacks class-wise results, error structure, and recall–cost curves; cross-domain gains vs. negative transfer for weak/semi/self-supervision and distillation are underreported; and fairness with group-wise calibration is rarely routine (Shah & Parveen, 2023; Oliver et al., 2025; Ahmad & Raza, 2024; Mathumetha et al., 2024; Zhang et al., 2025; Nakrani et al., 2020; Margerie-Mellon & Chassagnon, 2023; Nagaraj & Subhashini, 2023). Engineering transparency is also insufficient: standards are missing for inference details (tile size, stride, overlap, TTA, batch size), software/hardware stacks (GPU, drivers, libraries), and end-to-end latency/throughput/VRAM/power. For 3D models, the absence of resolution–throughput–VRAM trade-off

curves and capacity-planning guides hinders scalable deployment. Explainability often stops at heatmaps without stability or causal checks; standardized error tracing + audit logs are lacking; and limited openness/versioning of code, models, and inference configs constrains third-party audit (Mathumetha et al., 2024; Salih et al., 2023; Prisciandaro et al., 2023; Cai et al., 2024; Wang et al., 2025).

Finally, thresholding policy and human–AI collaboration remain underdeveloped at deployment. Operational gaps include absent multi-threshold policies (routine/peak/extreme) and resource forecasting; unquantified feedback of high-uncertainty cases with review prioritization; non-operationalized mapping from decision-curve net benefit ↔ cost structures; and incomplete health-economics and sustainability/O&M analyses (Egala & Sairam, 2024; Hasanah et al., 2023; Sabry et al., 2024). These gaps explain strong internal but weak external performance and underscore the need for auditable, actionable standards spanning prospective evaluation and post-deployment monitoring.

4.4 Recommendations for Improvement

To enable real-world deployment, improvements should begin with evidence, metrics, and data. Evidence. Move from internal cross-validation to two-stage validation—strict patient-wise + temporal holdout in the source domain, plus cross-site test-only evaluation—supplemented by small prospective cohorts. Metrics. Adopt a standardized metric quartet: FROC at 0.5/1/2/4 FP/scan for detection; Dice/IoU with Hausdorff95 and volume difference for segmentation; and ROC-AUC/PR-AUC with calibration and decision-curve analysis for classification/prognosis. Use unified case vs. lesion denominators, report statistical significance, and present consistent resampling displays. Data & annotation. Document acquisition, de-identification, preprocessing, splits & version de-duplication, and temporal splits. Anchor malignancy to pathology confirmation. For detection/segmentation, require ≥2 radiologists for independent labels with adjudication. Establish a hard-case QA pool (tiny nodules, GGO, vessel adhesion, motion artifacts) for periodic relabeling and spot audits. These steps enable direct, like-for-like comparisons on usability and deployability.

Model training & thresholding. The priority is to make why it works explicit and auditable. Run systematic ablations on 2D/2.5D/3D, multi-scale/attention, boundary losses, VOI & peri-nodular quantification, and weak/semi/self-supervision with distillation. Report ΔROC-AUC / ΔDice / ΔFROC with variance. For long-tail imbalance, provide sensitivity, error structure, and recall–cost curves stratified by class/volume/density; on external domains, quantify gain ranges and potential negative transfer for self-supervision and distillation. Furthermore, operationalize probability → threshold → net benefit as an actionable tool: at the target site, apply temperature scaling and recalibration; define three threshold bands (routine/peak/extreme) with resource mappings (rescans, staffing, beds); set revocable send-to-human policies with SLAs for high-uncertainty cases; and maintain a monthly/quarterly "threshold strategy card" with decision and calibration curves.

Engineering, governance, & economics. Treat deployability as a hard constraint for implementation and reporting. Publish a reproducible inference config table (tile size, stride, overlap, TTA near key thresholds, batch size, I/O & post-processing parallelism), together with the software/hardware stack and E2E latency/throughput/VRAM/power baselines. Use inference engines, FP16/mixed-precision, operator fusion, pruning/quantization, and distillation to reach ~1–2 s latency for CT 3D sliding-window and ~100 ms for CXR. Provide auditable resolution–throughput–VRAM trade-off curves for capacity planning. In parallel, leverage DICOM SR/SEG and HIE standards to integrate with PACS/RIS/EHR, delivering a clickable provenance chain (raw image → preprocessing → segmentation → VOI → probability → structured report). Establish a drift-monitoring → hard-case pool → relabeling/retraining loop; maintain versioned model/data cards with rollback paths; and include fairness and uncertainty in routine audits. Release code/weights/inference configs and a de-duplication manifest, with containerized one-click re-runs and de-identified mini-packs, targeting ±0.5% re-run error on key metrics. Finally, align decision-curve net benefit with real hospital cost structures, and report lifecycle cost, annual ROI, and peak-load sensitivity analyses to support scheduling and resource allocation.

V. Limitations and Conclusions

5.1 Limitations

This review has intrinsic scope limits. We restricted searches to English, peer-reviewed studies from 2020–2025, which may omit earlier foundational work and high-quality non-English publications. Our focus on pulmonary CT and CXR × CNNs also limits coverage of pure transformer approaches, multimodal foundation models, and deep imaging–clinical integration. The included studies show substantial heterogeneity in tasks, metrics, and validation schemes, which precludes rigorous meta-analysis and standardized effect-size estimation. Many reports do not fully disclose code, weights, inference configurations, or software/hardware environments, undermining verification of engineering metrics and workflow evidence. In addition, domain shift exists between public datasets and institutional case mix. Pathology-confirmed labels are scarce, consensus labels are subjective, and leakage or non-temporal patient-wise splits may inflate reported performance.

Evidence for evaluation and governance also remains incomplete. PR-AUC, calibration, and decision-curve analysis are underutilized, and external-domain temperature scaling and threshold recalibration are not routine. Fairness and uncertainty audits are rare, as are disclosures of negative results and deployment failures. Quantification is limited for acceleration/compression, E2E latency, throughput, VRAM/power, and lifecycle cost/ROI. Consequently, our conclusions should be interpreted within these boundaries. Priorities include study preregistration, cross-site test-only validation, a unified reporting framework (the metric quartet and threshold strategy card), and strong, open reproducibility with engineering/governance disclosure. Together, these steps enhance the verifiability, comparability, and portability of pulmonary imaging × CNNs from publication to routine clinical deployment.

5.2 Conclusions

Over the past five years, work on pulmonary imaging × CNNs has advanced across three clinical contexts: screening, triage, and follow-up. Methodologically, mainstream approaches use 2D/2.5D/3D CNN backbones with multi-scale and attention modules, adding explicit VOI and peri-nodular features to improve interpretability. To address small data and domain shift, studies employ weak/semi/self-supervision, contrastive learning, and knowledge distillation. In training, researchers use class reweighting and online hard-example mining. In evaluation, they assess detection/segmentation/classification performance, calibration, and decision-curve analysis. In deployment, model exchange, accelerated inference, mixed precision, and pruning/quantization reduce latency and VRAM, while structured reports and saliency heatmaps support clinical review.

Along these pathways, performance and engineering metrics improve consistently. Detection attains high recall with controlled FP/scan. Segmentation yields higher Dice/IoU and lower Hausdorff95. Malignancy classification and prognosis improve AUC and calibration, with net benefit demonstrated on decision curves. Meanwhile, 3D sliding-window inference now runs at ~100 ms to seconds, indicating practical potential for emergency and follow-up workflows. Consequently, model outputs translate into actionable evidence and threshold policies, accelerating the transition from research to routine use.

However, evidence and governance gaps persist. External/multicenter/prospective validation is limited. Temperature scaling and threshold recalibration across domains are inconsistently applied. Engineering/economic reporting is sparse. Fairness and uncertainty audits are not routine. Data/label representativeness remains constrained; leakage and non-temporal splits may inflate results. System-level logging, drift monitoring, and rollback lack standard templates, and privacy compliance with cross-domain governance lacks reusable playbooks. Hence, comparability, transferability, and long-term reliability require further strengthening.

Building on these gaps, the next phase should prioritize actionable, deployment-oriented standards. First, adopt patient-wise + temporal splits as a baseline; conduct independent external tests and small prospective studies; report interval estimates and stratify by device, protocol, and population. Second, within self-supervision, distillation, and domain generalization, quantify cross-hospital gains and negative-transfer bounds, and report error structure by class and lesion volume. Third, follow established model-exchange and inference-acceleration routes; publish reproducible inference configs and resolution–throughput–VRAM trade-off curves; and use DICOM SR/SEG with HIE and HL7 FHIR for traceable integration. Finally, under de-identification and access control, maintain data/model cards, audit logs, drift thresholds, and MTTR; routinize overall and subgroup calibration and loop back high-uncertainty cases to training/QC; and formalize a "threshold strategy card" mapping routine/peak/extreme resource scenarios. Following metric unification → external validation → engineering-first → governance-by-design, the field can progress from usable to useful and deployable.

References

1. Abdou, M. A. (2022). Efficient deep neural networks for medical imaging. *Neural Computing and Applications, 34*(8), 5791–5812.
2. Abdullahi, K., Ramakrishnan, K., & Ali, A. B. (2025). Deep learning techniques for lung cancer diagnosis with CT imaging: A systematic review. *Information, 16*(6), 451.
3. Ahmad, S., & Raza, K. (2024). Machine learning for lung cancer therapeutics. *Journal of Drug Targeting, 32*(6), 635–646.
4. Ait Nasser, A., & Akhloufi, M. A. (2023). Deep learning models for chest disease detection using radiography. *Diagnostics, 13*(1), 159.
5. Al-qaness, M. A., et al. (2024). Chest X-ray deep learning survey. *Archives of Computational Methods in Engineering, 31*(6).
6. Astley, J. R., Wild, J. M., & Tahir, B. A. (2022). Deep learning in lung image analysis. *The British Journal of Radiology, 95*(1132), 20201107.
7. Cai, G., Cai, Y., Zhang, Z., Cao, Y., Wu, L., Ergu, D., Zhao, Y. (2024). *Medical AI for early detection of lung cancer: A survey*. arXiv. https://arxiv.org/abs/2410.14769

8. Devnath, L., et al. (2022). Computer-aided diagnosis of pneumoconiosis. *IJERPH, 19*(11), 6439.
9. Egala, R., & Sairam, M. V. S. (2024). Medical image analysis using deep learning: A review. *Engineering Proceedings, 66*(1), 7.
10. Forte, G. C., Altmayer, S., Silva, R. F., Stefani, M. T., Libermann, L. L., Cavion, C. C., & Hochhegger, B. (2022). Deep learning algorithms for diagnosis of lung cancer: A systematic review and meta-analysis. *Cancers, 14*(16), 3856.
11. Gao, C., et al. (2025). Deep learning in pulmonary nodule detection. *European Radiology, 35*(1), 255–266.
12. Gayap, H. T., & Akhloufi, M. A. (2024). Deep machine learning for lung cancer detection. *BioMedInformatics, 4*(1), 236–284.
13. Gumma, L. N., Thiruvengatanadhan, R., Kurakula, L., & Sivaprakasam, T. (2022). CNN-based lung cancer detection: A survey. *SN Computer Science, 3*(1), 66.
14. Guzmán Gómez, R., Lopez Lopez, G., Alvarado, V. M., Lopez Lopez, F., Esqueda Cisneros, E., & López Moreno, H. (2025). Deep learning for treatment response prediction in lung cancer. *Tomography, 11*(7), 78.
15. Hansun, S., et al. (2023). Tuberculosis detection via deep learning. *Journal of Medical Internet Research*, 25, e43154.
16. Hasanah, S. A., Pravitasari, A. A., Abdullah, A. S., Yulita, I. N., & Asnawi, M. H. (2023). ResNet-based lung disease identification in CXR. *Applied Sciences, 13*(24), 13111.
17. Hosseini, S. H., et al. (2024). Deep learning for lung cancer diagnosis. *Multimedia Tools and Applications, 83*(5), 14305–14335.
18. Jassim, M. M., & Jaber, M. M. (2022). Systematic review for lung cancer detection and lung nodule classification: Taxonomy, challenges, and future recommendations. *Journal of Intelligent Systems, 31*(1), 944–964.
19. Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lung cancer detection: A review. *Artificial Intelligence Review, 57*(8), 197.
20. Jayaram, J., Haw, S., Palanichamy, N., Anaam, E., & Kumar, S. (2025). Effectiveness of machine and deep learning in lung cancer diagnosis: A review. *International Journal of Computing, 17*(1), 1–12.
21. Karthikeyan, N. K., & Ali, S. S. (2024). Lung cancer classification using CT scan images via CNN. In *Proceedings of the International Conference on Data Engineering* (pp. 1–5). IEEE.
22. Kumar, S., Kumar, H., Kumar, G., Singh, S. P., Bijalwan, A., & Diwakar, M. (2024). Imaging modalities and machine learning in lung disease diagnosis: A review. *BMC Medical Imaging, 24*(1), 30.
23. Lee, M. H., et al. (2023). Deep learning for COVID-19 lung imaging. *Journal of Clinical Medicine, 12*(10), 3446.
24. Liz-Lopez, H., de Sojo-Hernández, Á. A., D'Antonio-Maceiras, S., Diaz-Martinez, M. A., & Camacho, D. (2025). Deep learning innovations in lung cancer detection. *Cognitive Computation, 17*(2), 67.
25. Maharani, D. A., et al. (2025). Lightweight deep learning models for lung disease identification. Computers in Biology and Medicine, 194, 110425.
26. Malarvannan, S., & Angulakshmi, M. (2025). A review on lung cancer classification using deep learning techniques. *IEEE Access*, 13, 76161–76184. https://doi.org/10.1109/ACCESS.2025.3564633
27. Margerie-Mellon, C., & Chassagnon, G. (2023). AI in lung cancer imaging. *Diagnostic and Interventional Imaging, 104*(1), 11–17.
28. Nagaraj, P., & Subhashini, S. J. (2023). A review on detection of lung cancer using ensemble of classifiers with CNN. In *Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA)* (pp. 815–820). IEEE. https://doi.org/10.1109/ICECAA58104.2023.10212263
29. Nusantoro, J., Soesanti, I., & Ardiyanto, I. (2024). Lung cancer detection algorithm and method using deep learning techniques: A systematic literature review. In Proceedings of the 2024 4th International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS) (pp. 75–80). *IEEE*. https://doi.org/10.1109/ICE3IS62977.2024.10775504
30. Oliver, A. S., Sayeed, M. S., & Razak, S. F. A. (2025). Deep learning in lung cancer immunotherapy prediction. *Discover Oncology, 16*(1), 966.
31. Pacurari, A. C., Bhattarai, S., Muhammad, A., Avram, C., Mederle, A. O., Rosca, O., Bratosin, F., Bogdan, I., Fericean, R. M., Biris, M., Olaru, F., Dumitru, C., Tapalaga, G., & Mavrea, A. (2023). Diagnostic Accuracy of Machine Learning AI Architectures in Detection and Classification of Lung Cancer: A Systematic Review. *Diagnostics (Basel, Switzerland), 13*(13), 2145. https://doi.org/10.3390/diagnostics13132145
32. Palani, M., Rajagopal, S., & Chintanpalli, A. K. (2024). A systematic review on feature extraction methods and deep learning models for detection of cancerous lung nodules at an early stage -the recent trends and challenges. *Biomedical physics & engineering express, 11*(1), 10.1088/2057-1976/ad9154. https://doi.org/10.1088/2057-1976/ad9154
33. Prisciandaro, E., Sedda, G., Cara, A., Diotti, C., Spaggiari, L., & Bertolaccini, L. (2023). Artificial neural networks in lung cancer research. *Journal of Clinical Medicine, 12*(3), 880.
34. Sabry, A. H., et al. (2024). Audio-based lung disease recognition. *Heliyon, 10*(4).
35. Salih, S. A., et al. (2023). Lung disease diagnosis using deep learning. *Journal of Techniques, 5*(3), 158–173.
36. Sewatkar, R. M. (2025). Optimized CNN for lung cancer classification. *Multimedia Tools and Applications, 84*(23), 27517–27548.
37. Shah, S. N. A., & Parveen, R. (2023). Machine learning for lung cancer diagnosis: Review and perspectives. *Archives of Computational Methods in Engineering, 30*(8), 4917–4930.
38. Sharma, S., & Guleria, K. (2024). Pneumonia detection using deep learning. *Multimedia Tools and Applications, 83*(8), 24101–24151.
39. Singh, B., Sharma, M., & Gupta, S. (2025). A comprehensive review of deep learning techniques for lung disease diagnosis. In Proceedings of *the 2025 3rd International Conference on Disruptive Technologies* (pp. 269–273). IEEE.
40. Soffer, S., et al. (2022). AI for interstitial lung disease. *Academic Radiology*, 29, S226–S235.

41. Tajidini, F. (2023). *A comprehensive review of deep learning in lung cancer*. arXiv. https://arxiv.org/abs/2308.02528
42. Thanoon, M. A., et al. (2023). CT-based lung cancer detection review. *Diagnostics, 13*(16), 2617.
43. Thaseen, M., UmaMaheswaran, S. K., Naik, D. A., Aware, M. S., Pundhir, P., & Pant, B. (2022). A review of using CNN approach for lung cancer detection through machine learning. In *Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1236–1239). IEEE. https://doi.org/10.1109/ICACITE53722.2022.9823854
44. Tran, T. O., Vo, T. H., & Le, N. Q. K. (2024). Omics-based deep learning for lung cancer decision-making. *Briefings in Functional Genomics, 23*(3), 181–192.
45. Usharani C, Revathi B, Selvapandian A, Kezial Elizabeth SK. Lung Cancer Detection in CT Images Using Deep Learning Techniques: A Survey Review . *EAI Endorsed Trans Perv Health Tech*. Available from: https://publications.eai.eu/index.php/phat/article/view/5265
46. Vibina, W., Isravel, D. P., & Dhas, J. P. M. (2025). Deep CNN-driven automated lung disease identification. In Proceedings of *the 2025 10th International Conference on Smart Structures and Systems* (ICSSS) (pp. 1–6). IEEE. https://doi.org/10.1109/ICSSS66939.2025.11346368
47. Wang, T. W., et al. (2025). Deep learning in thoracic oncology. *Cancers, 17*(4), 621.
48. Young, R. P., Ward, R. C., Scott, R. J., & Silvestri, G. A. (2025). Diabetes mellitus and lung cancer screening outcomes in the National Lung Screening Trial. *Annals of the American Thoracic Society, 22*(9), 1409–1418. https://doi.org/10.1513/AnnalsATS.202411-1235OC
49. Yuan, L., et al. (2024). Machine learning in lung cancer prognosis using PET-CT. *Cancer Management and Research*, 361–375.
50. Zarandah, Q. M., Daud, S. M., & Abu-Naser, S. S. (2023). A systematic literature review of machine and deep learning-based detection and classification methods for respiratory diseases. *Journal of Theoretical and Applied Information Technology, 101*(4), 1273–1296.
51. Zhang, Y., Yang, H., Han, C., Zhang, C., Xu, C., & Lu, S. (2025). IDNet for projection compensation. *PLoS ONE, 20*(5), e0318812.