A Systematic Review of Machine Learning and Deep Learning Techniques for Lung Cancer Prediction

Dr.K.Sathya[1], Ms.B.Sowmya[2]
[1,2]*Assistant Professor, Department of Computer Science, VET Institute of Arts and Science (Co-Education) College, Erode*

Abstract: Lung cancer is the foremost cause of cancer-related mortality worldwide, approximately 1.8 million deaths annually. Timely and accurate detection is pivotal to improving patient outcomes. Over the Long period, the field of AI-assisted lung cancer prediction witnessed a paradigm shift from conventional machine learning approaches toward deep learning architectures—including convolutional neural networks (CNNs), vision transformers (ViTs), and, most recently, foundation models. This review systematically examines the methodological evolution, benchmark datasets, reported performance metrics, and outstanding challenges in ML- and DL-based lung cancer prediction. We cover nodule detection, malignancy classification, tumor segmentation, and survival prognosis prediction across CT, chest X-ray, and multi-modal data modalities. Key findings include: (i) 3D CNNs achieve accuracies of 94–98% on the LIDC-IDRI benchmark; (ii) transformer-based models exhibit superior performance on large datasets but demand substantial computational resources; (iii) federated learning and differential privacy are emerging solutions to multi-site data governance; and (iv) explainability (XAI) via Grad-CAM and SHAP is increasingly mandated for clinical translation. Critical gaps remain in external validation, benchmark standardisation, and regulatory approval pathways.

Keywords: ***lung cancer prediction; convolutional neural network; deep learning; CT imaging; nodule detection; vision transformer; federated learning; explainable AI; LIDC-IDRI; radiomics***

I. INTRODUCTION

Lung cancer represents one of the most lethal malignancies globally, with a five-year survival rate of people who had toxic distortion that remains below 20% when diagnosed at advanced stages, compared to over 60% when detected at stage I or II [1]. The global burden is immense: the World Health Organization (WHO) estimates that lung cancer caused approximately 1.8 millions of death in 2020 alone, with incidence rising in low- and middle-income countries [2]. The principal histological subtypes—non-small cell lung cancer (NSCLC, ~85%) and small cell lung cancer (SCLC, ~15%)—exhibit distinct molecular profiles and therapeutic sensitivities, underscoring the need for precise subtype prediction. Conventional diagnostic pathways rely on low-dose computed tomography (LDCT) screening, followed by expert radiologist interpretation, bronchoscopic biopsy, and histopathological confirmation. These workflows are resource-intensive, subject to inter-observer variability, and poorly scalable in regions with limited radiological expertise. The advent of artificial intelligence specifically machine learning (ML) and deep learning (DL offers a compelling opportunity to augment and partially automate these pipelines. The period 2020 to 2026 marks a distinctive era in this domain. Several converging developments catalysed rapid methodological progress: (1) the maturation of large annotated CT datasets such as LIDC-IDRI and NLST; (2) the democratisation of GPU computing and cloud infrastructure; (3) the proliferation of transformer architectures from natural language processing into medical imaging; and (4) growing regulatory and clinical interest in AI-assisted diagnostics. This review provides a structured synthesis of the literature across this period, covering ML approaches, deep learning architectures, benchmark datasets, performance metrics, explainability methods, and translational barriers. The remains of this paper is organised as follows. Section 2 describes the review methodology. Section 3 surveys classical ML approaches. Section 4 covers deep learning architectures. Section 5 examines transformer and self-supervised methods. Section 6 addresses datasets and benchmarks. Section 7 presents a comparative performance analysis. Section 8 discusses explainability and clinical translation. Section 9 identifies open

challenges, and Section 10 concludes with future directions.

II. REVIEW METHODOLOGY

A systematic literature search was conducted across PubMed, IEEE Xplore, Scopus, and Google Scholar for articles published between January 2020 and April 2026. Search strings combined MeSH terms and free-text keywords: ("lung cancer" OR "pulmonary nodule" OR "NSCLC") AND ("machine learning" OR "deep learning" OR "neural network" OR "convolutional" OR "transformer") AND ("detection" OR "classification" OR "segmentation" OR "prediction" OR "prognosis"). Articles were included if they (i) presented a primary ML or DL model for lung cancer-related prediction tasks; (ii) reported quantitative performance metrics; and (iii) were published in peer-reviewed journals or major conference proceedings (MICCAI, CVPR, NeurIPS, ISBI). Reviews, editorials, and non-English-language papers were excluded. Following deduplication and screening, 214 articles were retained for full-text analysis.

III. CLASSICAL MACHINE LEARNING APPROACHES

3.1 Support Vector Machines

Support vector machines (SVMs) were among the earliest ML algorithms applied to lung nodule classification, leveraging hand-crafted radiomic features—including nodule diameter, volume, spiculation index, and texture descriptors (GLCM, LBP)—as inputs to a kernel-based classifier. Studies from 2020–2021 reported SVM accuracies ranging from 83% to 90% on the LIDC-IDRI dataset when combined with PCA-based dimensionality reduction [4,5]. The primary advantages of SVMs in this context are interpretability of the feature space and robustness in low-sample-size regimes. However, their dependence on manual feature engineering limits scalability and generalisability across scanner protocols.

3.2 Ensemble Methods: Random Forest and Gradient Boosting

Random forests and gradient boosting algorithms (XGBoost, LightGBM) demonstrated competitive performance in both nodule malignancy prediction and patient-level risk stratification. A notable 2022 study by Chen et al. [6] combined 107 radiomic features from CT with structured clinical variables (pack-years, age, COPD status) in an XGBoost classifier, achieving an AUC of 0.91 on an external validation cohort of 450 patients. The built-in feature importance scores of tree-based ensembles facilitated partial explainability, a property increasingly valued for clinical deployment. Ensemble methods continue to function as strong baselines against which DL models are benchmarked.

3.3 Logistic Regression and Naïve Bayes

Logistic regression remains a clinical staple due to its coefficient interpretability and compatibility with regulatory frameworks. Several papers in the review period combined logistic regression with radiomic feature selection (LASSO, elastic net) for malignancy probability estimation, achieving AUCs of 0.82–0.87 on NLST subsets [7]. Naïve Bayes classifiers, though less frequently employed, appeared in studies targeting low-resource environments due to their minimal computational requirements and acceptable sensitivity on standard nodule classification tasks.

IV. DEEP LEARNING ARCHITECTURES

4.1 Convolutional Neural Networks for 2D Classification

Transfer learning from ImageNet-pretrained CNNs (VGG-16, ResNet-50, InceptionV3, EfficientNet-B4) became the dominant paradigm for 2D slice-based lung nodule classification from 2020 onward. By fine-tuning on CT axial slices or chest radiograph patches, studies consistently achieved accuracies of 89–95% on LIDC-IDRI and ChestX-ray14 [8,9]. Key design choices included focal loss to address nodule class imbalance, multi-scale feature fusion, and test-time augmentation. A systematic limitation of 2D approaches is the discard of volumetric spatial context, which is clinically significant for nodule shape analysis.

4.2 3D CNNs and Volumetric Segmentation

The most impactful architectural advance of the 2020–2022 period was the widespread adoption of 3D CNNs—principally 3D U-Net and its variants—for volumetric CT analysis. Operating on 3D patches centred on candidate nodules, these models jointly address detection, segmentation, and malignancy scoring within a unified framework. Li et al. [10] reported a 3D CNN achieving a sensitivity of 96.2% at a false-positive rate of 1.0 per scan on LUNA16. Attention-gated U-Net variants further improved segmentation accuracy on small nodules (<6 mm),

which are most diagnostically challenging. Computational demands—typically 24–48 GB GPU memory for full-volume inference remain a deployment barrier in resource-limited settings.

4.3 Recurrent Architectures for Longitudinal Data

Longitudinal CT screening generates temporal sequences of nodule measurements. LSTM and GRU networks have been applied to model nodule growth trajectories and predict malignant transformation over time, with reported AUCs of 0.88–0.93 on NLST follow-up cohorts [11]. These models require alignment of multi-timepoint scans to a common coordinate frame, a non-trivial registration challenge that constrains their broader adoption.

4.4 Generative Adversarial Networks for Data Augmentation

Class imbalance—where malignant nodules comprise as few as 15–20% of samples in screening cohorts—critically constrains DL model performance. GAN-based synthetic augmentation, particularly conditional DCGAN and CycleGAN approaches, was explored extensively from 2021 to 2023. Synthetic CT nodule patches generated by GANs consistently improved minority-class recall by 4–8 percentage points when incorporated into training pipelines [12]. Diffusion model-based synthesis, emerging in 2024–2025, produced higher-fidelity nodule textures and is expected to supersede GAN-based augmentation in the near term.

V. TRANSFORMER-BASED AND SELF-SUPERVISED METHODS

5.1 Vision Transformers for Nodule Classification

The adaptation of the Vision Transformer (ViT) and its hierarchical variant, the Swin Transformer, to medical imaging tasks commenced in earnest in 2022. By partitioning CT patches into non-overlapping tokens and applying self-attention across the full sequence, ViTs capture long-range spatial dependencies that local convolutional filters miss. On the LIDC-IDRI benchmark, Swin Transformer-based classifiers achieved accuracies of 93–96%, competitive with 3D CNNs while offering superior scalability with dataset size [13]. A practical challenge is their data hunger: ViTs require large training sets to outperform CNNs, necessitating extensive pre-training or fine-tuning from ImageNet-pretrained weights.

5.2 Self-Supervised Pre-Training

Self-supervised learning (SSL) methods—notably masked autoencoders (MAE), DINO, and SimCLR—have demonstrated that high-quality representations can be learned from large volumes of unlabelled chest CT scans, sidestepping the labelling bottleneck. A 2024 study pre-trained a ViT on 120,000 unlabelled LDCT scans and fine-tuned on 2,000 labeled nodule scans, achieving 94.1% AUC—within 1.2% of a fully supervised model trained on the complete labelled set [14]. SSL is particularly promising for low-resource clinical settings where labelled data acquisition is costly.

5.3 Foundation Models and Zero-Shot Segmentation

The emergence of large-scale foundation models—trained on diverse medical imaging datasets—marks the frontier of the field in 2025–2026. Models such as SAM-Med3D and MedSAM extend Meta AI's Segment Anything Model (SAM) to volumetric medical images, enabling zero-shot and few-shot nodule segmentation without task-specific fine-tuning. Preliminary results on LUNA16 show Dice coefficients of 0.81–0.87 for nodules >8 mm in zero-shot settings, approaching the performance of task-specific 3D U-Nets [15]. Integration of large language models (LLMs) for structured extraction from radiology reports—enabling the mining of surveillance histories, comparison studies, and clinical context—represents a complementary frontier.

VI. BENCHMARK DATASETS

Table 1 summarises the principal widely available datasets used across the reviewed literature.

Dataset	Modality	Size	Primary Task	Notes
LIDC-IDRI	CT	1,018 scans	Nodule detection & malignancy	4-radiologist consensus; most widely used
LUNA16	CT	888 scans	Nodule detection	Subset of LIDC-DRI; challenge benchmark
NLST	LDCT	26,000+ subjects	Screening; survival prediction	Randomised controlled trial data
ChestX-ray14	CXR	112,120 images	Multi-label classification	14 pathologies; NLP-derived labels

Dataset	Modality	Size	Primary Task	Notes
TCIA (NSCLC-R1)	CT+PET	422 subjects	Staging; radiomics	Multi-institutional; DICOM
RIDER Lung CT	CT	32 patients	Longitudinal; reproducibility	Test-retest; small but precise
In-house (various)	CT/CXR	Varies	All tasks	Single-institution; reproducibility concerns

Table 1. Summary of principal benchmark datasets used in lung cancer ML/DL research, 2020–2026.

VII. COMPARATIVE PERFORMANCE ANALYSIS

Table 2 presents aggregated performance ranges by model category across CT-based classification tasks on the LIDC-IDRI and LUNA16 benchmarks.

Method Category	Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)
SVM + Radiomics	83–90	0.85–0.91	80–88	85–92
Random Forest / XGBoost	85–90	0.88–0.92	82–90	87–93
2D CNN (ResNet/VGG)	89–95	0.91–0.96	87–94	90–96
3D CNN / U-Net	94–98	0.95–0.99	93–97	94–98
Swin Transformer / ViT	91–96	0.93–0.97	90–95	92–97
Ensemble / Multi-modal	93–97	0.95–0.98	91–96	93–97
Self-supervised (SSL)	92–96	0.93–0.97	89–95	92–97

Table 2. Aggregated performance ranges by model category. Values drawn from reviewed literature; vary by dataset split, evaluation protocol, and task definition.

Numerous methodological factors account for performance variation beyond architectural differences. (1) Evaluation protocol: many high-accuracy reports use cross-validation on LIDC-IDRI without an external validation set, inflating apparent performance. (2) Task definition: malignancy classification (binary) is inherently easier than fine-grained staging (multi-class). (3) Nodule size distribution: performance degrades substantially for sub-centimetre nodules, which carry the highest clinical uncertainty. (4) Scanner heterogeneity: models trained on single-vendor data frequently exhibit 5–12% accuracy drops when applied to scans from different manufacturers or reconstruction kernels.

VIII. EXPLAINABILITY AND CLINICAL TRANSLATION

8.1 Explainability Methods

The opacity of deep neural networks remains a principal barrier to clinical adoption. Class activation mapping (CAM) and its gradient-weighted variant Grad-CAM produce saliency maps that spatially localise the regions driving model predictions, enabling radiologists to verify that the network attends to clinically relevant features rather than confounders [16]. SHAP (SHapley Additive exPlanations) has been widely applied to classical ML models operating on radiomic features, attributing prediction scores to individual features with game-theoretic guarantees. Integrated Gradients and LIME provide complementary perspectives. A 2023 survey found that 61% of reviewed DL papers from 2021–2023 included some form of XAI analysis, compared to 18% in the preceding period [17]—a marked shift driven by regulatory expectations and journal policies.

8.2 Federated Learning for Privacy-Preserving Collaboration

Patient data privacy regulations (HIPAA in the USA; GDPR in the EU) create substantial barriers to aggregating multi-site CT datasets for model training. Federated learning (FL) circumvents this by training local models at each participating institution and aggregating only model weights—not raw data—on a central server. Landmark studies in 2023–2024 demonstrated that FL-trained nodule classifiers achieved within 2–3% accuracy of centrally trained models while preserving patient privacy [18]. Differential privacy and secure aggregation protocols further harden these frameworks against gradient inversion attacks. FL is now considered a standard recommendation in multi-institutional AI study designs.

8.3 Regulatory and Clinical Deployment Landscape

As of 2026, fewer than 20 AI-based lung nodule management tools have received FDA 510(k) clearance or CE marking for clinical use. The gap between published accuracy and real-world deployment reflects multiple friction points: (i) studies rarely report calibration metrics (Brier score, reliability diagrams), making uncertainty quantification incomplete; (ii) prospective clinical trial evidence of outcome improvement remains sparse; (iii) workflow integration—embedding AI recommendations into PACS and electronic health record systems—requires substantial engineering; and (iv) reimbursement pathways for AI-assisted radiology remain undefined in many jurisdictions.

IX. OPEN CHALLENGES

Data scarcity and class imbalance: Malignant nodules are substantially outnumbered in screening cohorts. GAN and diffusion-based augmentation partially mitigate this, but synthetic data quality and diversity remain limitations.

Domain shift: Model performance degrades when CT acquisition parameters (slice thickness, reconstruction kernel, kVp) differ from training conditions. Domain adaptation and test-time batch normalisation are partial solutions requiring further validation.

Benchmark inconsistency: Heterogeneous reporting of metrics (accuracy, AUC, FROC sensitivity, Dice coefficient) across studies impedes fair comparison. Community adoption of standardised evaluation frameworks—analogous to MIMIC-IV benchmarks in ICU prediction—is urgently needed.

Multi-omics integration: Fusing imaging features with genomic, proteomic, and liquid biopsy data offers the prospect of subtype-specific prediction and treatment response modelling, but requires aligned multi-modal cohorts that are rarely available at scale.

Prospective validation: The vast majority of reviewed studies are retrospective. Prospective trials evaluating whether AI-assisted screening changes diagnostic yield, time-to-diagnosis, or survival outcomes are essential for clinical adoption and regulatory approval.

X. CONCLUSIONS AND FUTURE DIRECTIONS

The review of this paper has charted the rapid and substantive evolution of ML- and DL-based lung cancer prediction between 2020 and 2026. The field has transitioned from hand-crafted radiomic features has been combined with SVM classifiers toward end-to-end deep learning pipelines, with 3D CNNs and vision transformers now defining the state-of-the-art on standard benchmarks. Foundation models and self-supervised learning are poised to reduce dependence on large annotated datasets—a persistent bottleneck—while federated learning addresses the data governance barriers that have historically confined AI development to single-institution silos. Looking ahead, several priorities merit emphasis. First, the community must invest in prospective validation trials that measure patient-level outcomes rather than image-level classification metrics in isolation. Second, standardised benchmarking frameworks and reporting guidelines analogous to STARD for diagnostic test studies should be adopted universally. Third, multi-omics fusion, integrating CT radiomics with genomic and liquid biopsy biomarkers, represents the most promising pathway toward personalised lung cancer risk stratification and treatment response prediction. Fourth, regulatory science must develop adaptive frameworks capable of evaluating continuously learning AI systems deployed in dynamic clinical environments. The promise of AI in lung cancer care is well-evidenced by the literature reviewed here. Realising that promise in practice will require sustained collaboration among computer scientists, radiologists, oncologists, patient advocates, and regulators working toward systems that are not only accurate, but transparent, equitable, and clinically trustworthy.

REFERENCES

[1] Siegel, R.L., et al. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1), 17–48.

[2] World Health Organization. (2024). Global Cancer Observatory: Lung Cancer Fact Sheet. Lyon: IARC.

[3] Ardila, D., et al. (2019). End-to-end lung cancer malignancy prediction from CT imaging. Nature Medicine, 25(6), 954–961.

[4] Huang, X., & Liu, Y. (2020). SVM-based pulmonary nodule malignancy classification using GLCM and LBP features. Medical Image Analysis, 61, 101652.

[5] Kumar, A., et al. (2021). Radiomic feature selection for lung nodule classification: A comparative

study. IEEE Journal of Biomedical and Health Informatics, 25(4), 1210–1219.

[6] Chen, S., et al. (2022). XGBoost-based multi-modal lung cancer risk prediction integrating CT radiomics and clinical variables. European Radiology, 32(8), 5421–5431.

[7] Park, J., et al. (2021). LASSO-regularised logistic regression for lung cancer risk stratification in NLST. Thorax, 76(3), 278–286.

[8] Wang, L., et al. (2020). Transfer learning with EfficientNet for pulmonary nodule classification. Computers in Biology and Medicine, 125, 103986.

[9] Zhang, Y., et al. (2021). ResNet-50 fine-tuning for chest radiograph lung mass detection on ChestX-ray14. Radiology: Artificial Intelligence, 3(2), e200195.

[10] Li, R., et al. (2022). 3D attention-gated U-Net for volumetric lung nodule detection and segmentation. Medical Image Analysis, 75, 102278.

[11] Zhao, H., et al. (2022). LSTM-based longitudinal nodule growth modelling for malignancy prediction in NLST. IEEE Transactions on Medical Imaging, 41(6), 1488–1498.

[12] Sun, W., et al. (2021). GAN-based CT nodule synthesis for training data augmentation. Journal of Medical Imaging, 8(3), 034502.

[13] He, J., et al. (2023). Swin Transformer for lung nodule classification: Benchmarking against 3D CNNs on LIDC-IDRI. Diagnostics, 13(4), 712.

[14] Liu, M., et al. (2024). Self-supervised masked autoencoder pre-training on unlabeled chest CT for low-shot nodule classification. Nature Biomedical Engineering, 8(2), 198–211.

[15] Gao, X., et al. (2025). SAM-Med3D: Adapting the Segment Anything Model for zero-shot volumetric medical image segmentation. Medical Image Analysis, 98, 103345.

[16] Selvaraju, R.R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localisation. ICCV, 618–626.

[17] Ibrahim, A., et al. (2023). Explainability in AI-based lung cancer detection: A systematic review. npj Digital Medicine, 6(1), 77.

[18] Roth, H., et al. (2024). Federated learning for lung nodule classification across 20 institutions. The Lancet Digital Health, 6(4), e231–e242.