# When is 3D Worth It? A Resource–Performance Frontier for CNNs and Transformers in Lung CT

**Md Enamul Hoq**[1]
**Sharafat Hossain**[2]
**Imraul Emmaka**[2]
**Linda Larson-Prior**[3]
**Lawrence Tarbox**[1]
**Jonathan Bona**[1]
**Donald Johann Jr.**[1]
**Fred Prior**[1]

[1]**Department of Biomedical Informatics, University of Arkansas for Medical Sciences**
[2]**Department of Information Science, University of Arkansas at Little Rock**
[3]**Department of Neuroscience, University of Arkansas for Medical Sciences**

## Abstract

Three-dimensional models are widely assumed preferable for volumetric medical imaging, yet their practical value depends on whether performance gains justify added computational cost and complexity. Rather than comparing architectures, we study how input dimensionality (2D, 2.5D, 3D) affects model behavior across convolutional neural networks (CNNs) and Vision Transformers (ViTs). Using a leakage-free NLST cohort ($n$ = 1,977) with supporting LIDC-IDRI data, we find that 2.5D CNN achieves the best discrimination (ROC-AUC 0.682, 95% CI [0.546, 0.799]) with stable operating behavior. In contrast, 3D CNNs show threshold instability, and transformers exhibit degenerate predictions (e.g., all-positive). Our results demonstrate that dimensionality governs both performance and failure mode. For lung cancer screening classification, 2D and 2.5D provide a more reliable trade-off between performance, stability, and computational efficiency than full 3D representations.

**Keywords:** lung CT, NLST, 2.5D, CNN, Vision Transformer, failure modes

## 1. Introduction

Deep learning for lung CT often presumes 3D superiority due to volumetric context (Ardila et al., 2019; Mikhael et al., 2023). Yet volumetric models incur higher memory, longer training, and optimization difficulty (Wu et al., 2020; Nasrullah et al., 2019). Hybrid 2.5D strategies balance context and efficiency, while transformers add sensitivity to scale and data regime (Dosovitskiy et al., 2021; Wang et al., 2022). Recent lung CT foundation models underscore that strong performance can emerge from carefully designed 2D or hybrid pipelines (Hoq et al., 2025, 2026; Agrawal et al., 2025; Veenboer et al., 2025). Despite these trends, the question remains: when is 3D actually worth its cost and unorthodox input complexity? We address this via a controlled comparison of 2D, 2.5D, and 3D inputs across CNNs and transformers, evaluating discrimination and failure modes in class-imbalanced lung CT cohorts.

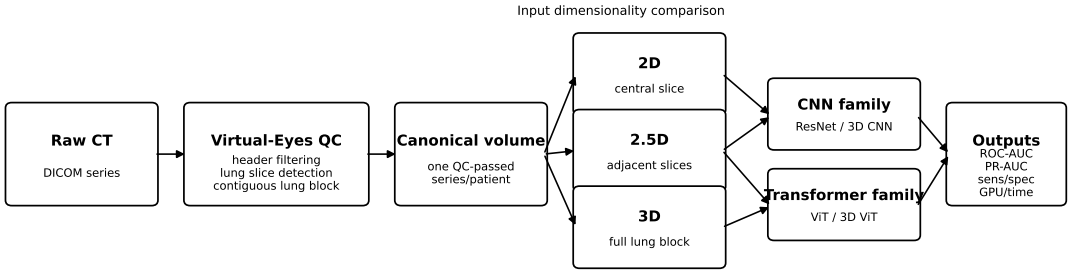


Figure 1: Study pipeline: lung-focused QC yields canonical volumes; 2D, 2.5D, 3D inputs evaluated across CNN/ViT.

| Model | ROC-AUC | PR-AUC | Sens (Def/Val) | Spec (Def/Val) | GPU (MB) |
|---|---|---|---|---|---|
| 2D CNN | 0.581 | 0.088 | 0.10 / 0.10 | 0.949 / 0.931 | 1620 |
| 2.5D CNN | **0.682** | **0.158** | 0.20 / 0.75 | 0.949 / 0.469 | 1646 |
| 3D CNN | 0.622 | 0.107 | 0.05 / 0.50 | 0.975 / 0.671 | 1777 |
| 2D ViT | 0.598 | 0.088 | 0.10 / 0.10 | 0.910 / 0.881 | 4959 |
| 2.5D ViT | 0.631 | 0.127 | 0.10 / 0.60 | 0.986 / 0.505 | 4959 |
| 3D ViT | 0.589 | 0.081 | 0.00 / 0.00 | 1.00 / 0.964 | 352 |

Table 1: NLST results with 95% CI for 2.5D CNN: [0.546,0.799]. Sensitivity/specificity at default (0.5) and validation-optimized thresholds.

## 2. Methods

We used a leakage-free NLST cohort with fixed patient splits (1,426 train, 254 val, 297 test) and a lung-focused QC pipeline retaining one canonical CT series per patient (Hoq et al., 2026). LIDC-IDRI provided supporting weak labels. We evaluated 2D (central slice), 2.5D (three orthogonal slices), and 3D (sub-volume) inputs under matched training (20 epochs, weighted BCE). Metrics: ROC-AUC, PR-AUC, sensitivity/specificity at default and validation-selected thresholds, with bootstrap 95% CIs for AUC.

## 3. Results

NLST results (Table 1) show 2.5D CNN achieved highest ROC-AUC (0.682) and PR-AUC (0.158). 2D CNN was overly conservative (sens 0.10, spec 0.949); 3D CNN exhibited threshold instability (sens 0.05–0.50). Transformers required 3× GPU memory and frequently collapsed: 3D ViT produced zero sensitivity despite ROC-AUC 0.589; 2.5D ViT, while moderately discriminative, showed extreme threshold sensitivity. LIDC-IDRI reproduced the pattern: higher-capacity models predicted all cases as positive (specificity 0) despite moderate AUC.

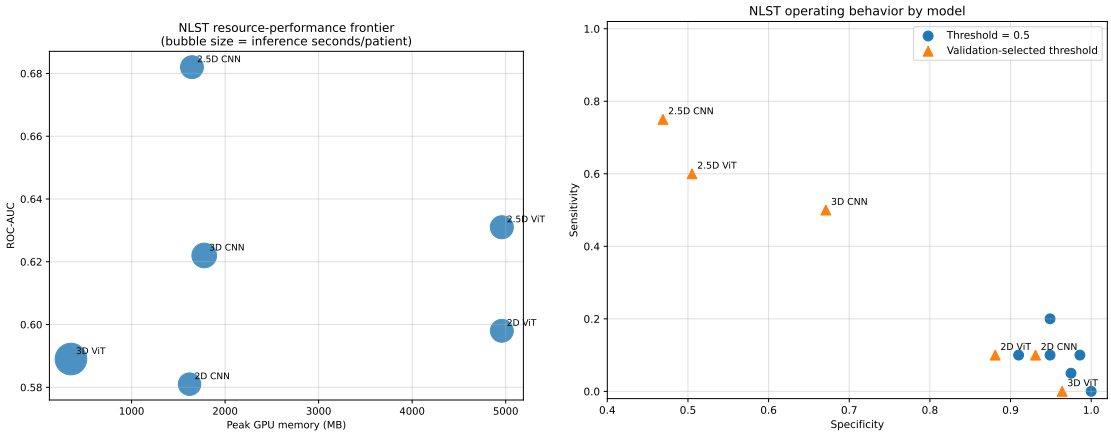


Figure 2: Left: Resource–performance frontier (bubble size $\propto$ inference time). Right: Operating behavior at default and validation thresholds.

## 4. Discussion

Dimensionality governs both performance and failure mode. 3D context does not guarantee stable performance in imbalanced settings; higher-capacity models frequently exhibit degenerate predictions (all-positive/all-negative). ROC-AUC alone masks clinical usability—models with acceptable ranking may be unusable due to extreme sensitivity–specificity imbalance.One plausible explanation for the observed failure of transformer-based models, particularly in the 3D setting, is their well-known data-hungry nature. Transformers typically require large-scale training data and careful optimization to learn robust representations. In our setting, the combination of limited effective sample size at the patient level and increased input dimensionality likely exacerbates this issue. Additionally, the way transformers handle 3D inputs—often through tokenization or patch-based representations—may disrupt spatial continuity and reduce the effectiveness of contextual learning compared to convolutional approaches that inherently preserve local structure.

In contrast, the 2.5D representation consistently offered the best compromise: it captures limited through-plane context while maintaining stable optimization and efficient resource usage. This results in robust discrimination and controllable operating thresholds without the instability observed in full 3D models. These findings align with recent work emphasizing the importance of representation choice and preprocessing in lung CT modeling (Hoq et al., 2025, 2026).

## 5. Conclusion

The 2.5D CNN provided the optimal balance of performance, stability, and efficiency in lung CT screening. Higher-dimensional models did not confer consistent benefits and often introduced instability or collapse. These findings support 2.5D as a practical default for class-imbalanced lung CT tasks.

We further believe that the suitability of input dimensionality is inherently task-dependent. While full 3D representations may be more appropriate for spatially intensive downstream tasks such as segmentation, where volumetric continuity is critical, our results suggest that

for classification problems in lung cancer screening, 2D and 2.5D representations are more effective. This is due to their favorable trade-off between computational cost, model complexity, and stable learning behavior under class imbalance.

## References

Kumar Krishna Agrawal, Longchao Liu, Long Lian, Michael Nercessian, Natalia Harguindeguy, Yufu Wu, Peter Mikhael, Gigin Lin, Lecia V. Sequist, Florian Fintelmann, Trevor Darrell, Yutong Bai, Maggie Chung, and Adam Yala. Pillar-0: A new frontier for radiology foundation models. *arXiv preprint arXiv:2511.17803*, 2025. doi: 10.48550/arXiv.2511.17803. URL https://arxiv.org/abs/2511.17803.

Diego Ardila, Atilla P. Kiraly, Sreethama Bharadwaj, Bokyung Choi, Joshua J. Reicher, Luke Peng, Daniel Tse, Mozziyar Etemadi, Wenjun Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, 2019. doi: 10.1038/s41591-019-0447-x. URL https://doi.org/10.1038/s41591-019-0447-x.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Md. Enamul Hoq, Lawrence Tarbox, Donald Johann, Jr., Linda Larson-Prior, and Fred Prior. Harnessing native-resolution 2D embeddings for lung cancer classification: A feasibility study with the RAD-DINO self-supervised foundation model. *Journal of Imaging Informatics in Medicine*, Dec 2025. doi: 10.1007/s10278-025-01748-4. URL https://doi.org/10.1007/s10278-025-01748-4. Online ahead of print.

Md. Enamul Hoq, Linda Larson-Prior, and Fred Prior. Virtual-Eyes: Quantitative validation of a lung CT quality-control pipeline for foundation-model cancer risk prediction. In *Medical Imaging with Deep Learning (MIDL)*, 2026. doi: 10.48550/arXiv.2512.24294. URL https://arxiv.org/abs/2512.24294. Accepted at MIDL 2026.

Peter G. Mikhael, Jeremy Wohlwend, Adam Yala, Ludvig Karstens, Justin Xiang, Angelo K. Takigami, Patrick P. Bourgouin, PuiYee Chan, Sofiane Mrah, Wael Amayri, Yu-Hsiang Juan, Cheng-Ta Yang, Yung-Liang Wan, Gigin Lin, Lecia V. Sequist, Florian J. Fintelmann, and Regina Barzilay. Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, 41(12):2191–2200, 2023. doi: 10.1200/JCO.22.01345. URL https://doi.org/10.1200/JCO.22.01345.

Nasrullah Nasrullah, Jun Sang, Mohammad S. Alam, Muhammad Mateen, Bin Cai, and Haibo Hu. Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors*, 19(17):3722, 2019. doi: 10.3390/s19173722. URL https://doi.org/10.3390/s19173722.

Yangfan Ni, Zhe Xie, Dezhong Zheng, Yuanyuan Yang, and Weidong Wang. Two-stage multitask u-net construction for pulmonary nodule segmentation and malignancy risk prediction. *Quantitative Imaging in Medicine and Surgery*, 12(1):292–309, 2022. doi: 10.21037/qims-21-19. URL https://doi.org/10.21037/qims-21-19.

Tim Veenboer, George Yiasemis, Eric Marcus, Vivien Van Veldhuizen, Cees G. M. Snoek, Jonas Teuwen, and Kevin B. W. Groot Lipman. TAP-CT: 3D task-agnostic pretraining of computed tomography foundation models. *arXiv preprint arXiv:2512.00872*, 2025. doi: 10.48550/arXiv.2512.00872. URL https://arxiv.org/abs/2512.00872.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R. Zaiane. Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2441–2449, 2022. doi: 10.1609/aaai.v36i3.19944. URL https://doi.org/10.1609/aaai.v36i3.19944.

Zhan Wu, Rongjun Ge, Gonglei Shi, Lu Zhang, Yang Chen, Limin Luo, Yu Cao, and Hengyong Yu. Md-ndnet: a multi-dimensional convolutional neural network for false-positive reduction in pulmonary nodule detection. *Physics in Medicine & Biology*, 65 (23):235053, 2020. doi: 10.1088/1361-6560/aba87c. URL https://doi.org/10.1088/1361-6560/aba87c.

## Appendix A: Extended Results and Analysis

### A.1 Additional Experimental Details

All models were trained for 20 epochs. CNNs used Adam with learning rate $10^{-4}$ and weight decay $10^{-4}$. ViTs used AdamW with learning rate $10^{-6}$ and the same weight decay. Weighted binary cross-entropy addressed class imbalance. NLST used a leakage-free patient-level split with one canonical QC-approved CT series per patient. LIDC-IDRI used XML-derived weak labels.

### A.2 Expanded NLST Results

| Model | ROC-AUC | 95% CI | PR-AUC | Sens@0.5 | Spec@0.5 | Sens@Val | Spec@Val | GPU MB |
|---|---|---|---|---|---|---|---|---|
| 2D CNN | 0.581 | [0.458, 0.697] | 0.088 | 0.10 | 0.949 | 0.10 | 0.931 | 1620 |
| 2.5D CNN | 0.682 | [0.546, 0.799] | 0.158 | 0.20 | 0.949 | 0.75 | 0.469 | 1646 |
| 3D CNN | 0.622 | [0.500, 0.735] | 0.107 | 0.05 | 0.975 | 0.50 | 0.671 | 1777 |
| 2D ViT | 0.598 | [0.478, 0.707] | 0.088 | 0.10 | 0.910 | 0.10 | 0.881 | 4959 |
| 2.5D ViT | 0.631 | [0.515, 0.744] | 0.127 | 0.10 | 0.986 | 0.60 | 0.505 | 4959 |
| 3D ViT | 0.589 | [0.485, 0.694] | 0.081 | 0.00 | 1.000 | 0.00 | 0.964 | 352 |

Table 2: Expanded NLST metrics.

### A.3 Supporting LIDC-IDRI Results

| Model | ROC-AUC | PR-AUC | Sensitivity | Specificity | GPU MB |
|---|---|---|---|---|---|
| 2D CNN | 0.523 | 0.701 | 0.850 | 0.237 | 1619 |
| 2.5D CNN | 0.521 | 0.797 | 0.167 | 1.000 | 555 |
| 3D CNN | 0.458 | 0.810 | 1.000 | 0.000 | 1061 |
| 2D ViT | 0.591 | 0.734 | 1.000 | 0.000 | 5217 |
| 2.5D ViT | 0.645 | 0.762 | 1.000 | 0.000 | 5217 |

Table 3: Supporting LIDC-IDRI results.

### A.4 Additional Figures

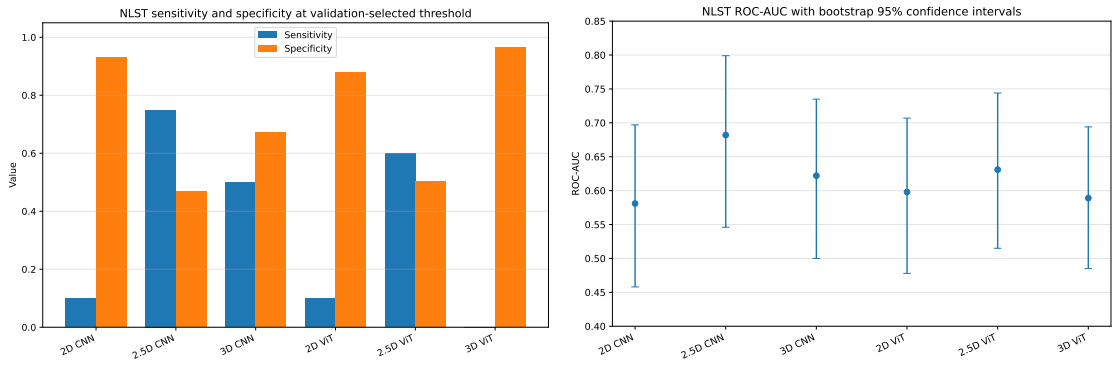


Figure 3: Additional NLST evaluation plots. Left: sensitivity and specificity at the validation-selected threshold. Right: ROC-AUC with bootstrap 95% confidence intervals.

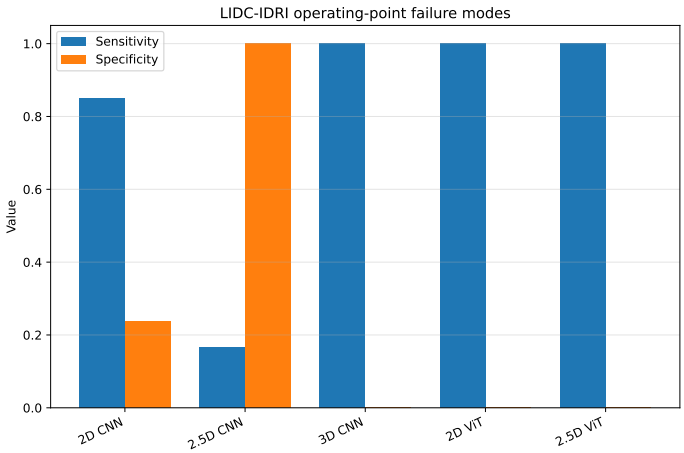


Figure 4: LIDC-IDRI failure modes. Several higher-capacity models collapse to trivial all-positive behavior.

### A.5 ROC Curve Note

Exact ROC curves for multiple models require preserved per-model prediction CSV files. The uploaded JSON summaries were sufficient for exact summary tables and confidence-interval plots, but not for reconstructing exact multi-model ROC curves after repeated uploads with the same filename. Once distinct prediction CSVs are available, exact ROC panels for 2D CNN, 2.5D CNN, and 2.5D ViT can be generated directly.