
LUNG CANCER DETECTION USING DEEP LEARNING

A PREPRINT

Pranav Gaddam
Texas A&M University

Advay Bhattacharya
Texas A&M University

Kyle Wu
Texas A&M University

April 29, 2026

1 Introduction

Lung cancer is one of the leading causes of cancer-related deaths worldwide, largely because it is often detected at a late stage when treatment options are limited. Early detection can significantly improve survival rates and expand treatment options. Computed tomography (CT) scans are widely used to identify lung abnormalities, but interpreting them is time-consuming and depends very much on radiologist expertise. Small nodules can be missed, and benign and malignant patterns often appear very similar, which makes consistent and accurate diagnosis difficult.

To address this problem, researchers have explored both machine learning and deep learning approaches for automated lung cancer detection. Early methods relied on hand-crafted features such as nodule shape, size, and texture, often using curated datasets like LIDC-IDRI [1]. More recent work has shifted toward deep learning, especially convolutional neural networks (CNNs), which learn features directly from images and have shown strong performance in medical imaging tasks [2]. In lung cancer specifically, large-scale studies have demonstrated that deep learning models can predict malignancy risk from CT scans with performance comparable to expert radiologists [3]. However, despite these advances, important challenges remain. CT data can vary in quality, nodules differ widely in appearance, and class imbalance is common. These issues can lead to overfitting and inconsistent performance across different datasets.

In this project, we build on this prior work by developing and comparing multiple machine learning models for lung cancer detection using CT scan images. Rather than focusing on a single approach, we evaluate how different modeling choices perform under the same dataset and experimental conditions. We also study how preprocessing steps and model selection impact overall performance.

Contributions. This work makes the following contributions: (1) a comparison of four approaches to CT scan classification under identical experimental conditions, ranging from a classical PCA + Logistic Regression baseline to a pretrained Swin Transformer; (2) an analysis of how transfer learning initialization strategy affects performance on a small medical imaging dataset; and (3) a demonstration that multi-phase fine-tuning from an adapted checkpoint outperforms both full fine-tuning and a completely fresh transfer learning process, achieving 91.2% test accuracy on a five-class lung cancer classification task.

2 Methods

2.1 Dataset and Preprocessing

We use the CT Scan Images for Lung Cancer dataset from Kaggle [5], which contains labeled CT scan images split into train, validation, and test directories. We first resolved inconsistent directory names with a canonical mapping function that normalizes all folder names to five class labels regardless of split. We then resolved an issue where the benign class was severely underrepresented relative to other classes. Since benign and normal cases both are defined by the absence of malignancy, we merged them into a single normal class. After preprocessing, the training set contained 1,460 images: adenocarcinoma (195), large cell carcinoma (115), malignant cases (460), normal (535), and squamous cell carcinoma (155). The held-out test set contained 475 images. To address remaining class imbalance, we computed

inverse-frequency class weights and passed them to the loss function during training, so the model has a higher penalty for misclassifying underrepresented classes.

2.2 Image Transformations

All images were resized to 224×224 pixels and converted to grayscale then replicated to three channels, since CT scans are inherently single-channel but the Swin backbone expects RGB input. Training images were augmented with random horizontal flips, rotations up to 10 degrees, small affine translations and scaling, and color jitter on brightness and contrast. Test and validation images received only resizing and grayscale conversion. All images were normalized using ImageNet mean and standard deviation values, which is standard practice when using pretrained backbones.

2.3 Baseline: PCA + Logistic Regression

As a classical lower-bound baseline, each image was flattened to a 150,528-dimensional vector, standardized, and reduced to 100 principal components via PCA. A multinomial logistic regression classifier with balanced class weights was then trained on the reduced features. This approach represents how CT scan images would have been classified before deep learning became dominant.

2.4 Swin Transformer Architecture

Our deep learning models are built on the Swin Transformer [4], a vision transformer architecture that processes images as non-overlapping patches and applies self-attention within shifted local windows. This design captures both local and global spatial structure efficiently, making it well-suited for medical imaging tasks where subtle local patterns are diagnostically relevant. We use the `microsoft/swin-tiny-patch4-window7-224` checkpoint pretrained on ImageNet as our backbone, replacing its original classification head with a single linear layer sized to our five-class problem.

2.5 Model 1: Full Fine-Tuning

In Model 1, all backbone parameters are unfrozen from the start and the entire model is fine-tuned end-to-end. We use the AdamW optimizer with a learning rate of 3×10^{-5} and weight decay, weighted cross-entropy loss using the precomputed class weights, gradient clipping at a max norm of 1.0 to prevent exploding gradients, a ReduceLROnPlateau scheduler that halves the learning rate if validation loss does not improve for two consecutive epochs, and early stopping with patience of 3 epochs.

2.6 Models 2a and 2b: Two-Phase Transfer Learning

Models 2a and 2b apply an identical two-phase transfer learning strategy but differ in their starting checkpoint, allowing us to isolate the effect of initialization.

In Phase 1, the entire backbone is frozen and only the classification head is trained for 5 epochs at a learning rate of 10^{-4} . This warm-up step trains the head to produce reasonable outputs before any backbone weights are modified.

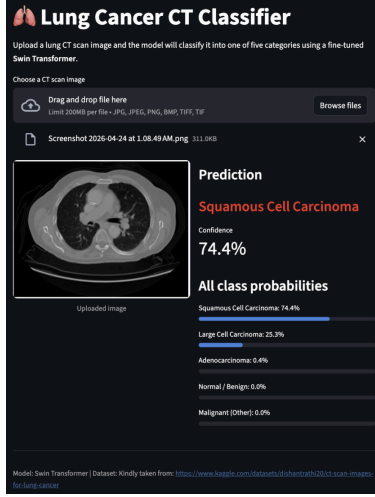
In Phase 2, the backbone is partially unfrozen, specifically the last two Swin stages (`layers.2` and `layers.3`), and trained with differential learning rates: 10^{-4} for the head and 2×10^{-5} for the backbone. The lower backbone learning rate preserves general low-level features learned during pretraining while allowing the deeper, more task-specific layers to adapt to CT scan data. Early stopping with patience of 3 is applied throughout Phase 2.

Model 2a initializes from Model 1’s best saved checkpoint, meaning Phase 1 begins from already domain-adapted weights. Model 2b re-initializes from the original ImageNet pretrained weights, so Phase 1 begins from scratch. This controlled comparison lets us measure how much of Model 2a’s performance comes from the training strategy versus the quality of the starting point.

2.7 Streamlit Application

To make our model more interactive and easier to use, we developed a simple web application through Streamlit. This app allows users to upload an image of a CT scan and receive a prediction from our trained model. This app provides a practical demonstration of how deep learning could be used in real world settings. The user would be able to quickly upload new images without having to run the full training pipeline. The app loads the best performing trained model and applies the steps used during training. After processing the input image, the model outputs a prediction along

with its confidence score, allowing users to see not only the result but also how confident the model is in its prediction. The application is publicly accessible [here](#). This app helps to highlight the potential for machine learning models as solutions for modern problems, acting as a support tool for medical imaging.



The application provides insight into how the model makes its predictions by displaying class probabilities for different diagnoses. Rather than returning a single label, the model presents a distribution over many different categories, which allows users to better understand uncertainty in the prediction. This is important in medical contexts, where decisions often require careful interpretation rather than absolute answers.

From a technical perspective, the application integrates the trained model directly into a deployment pipeline, which enables efficient inference without requiring retraining. This is a good demonstration into how machine learning models can be transitioned from development to practical use.

Figure 1: Streamlit application interface

3 Experiments

3.1 Baseline: PCA + Logistic Regression

The baseline achieved 50.9% test accuracy and a macro F1 of 0.49. Results within classes show that the model performed well on normal and malignant cases with F1 of 0.72 and 0.73, respectively. However, it performed much worse on adenocarcinoma with a recall of only 0.07, missing 93% of the most common lung cancer subtype. This shows us that the patterns needed to distinguish cancer subtypes are not necessarily captured by linear classifiers operating on flattened pixel features.

3.2 Model 1: Swin Transformer Full Fine-Tuning

Training ran for all 15 epochs, reaching a peak validation accuracy of 95.8%. On the test set, Model 1 achieved 90.5% accuracy and a macro F1 of 0.90, which is a 40 percentage point improvement over the baseline. Large cell carcinoma achieved perfect recall, and normal cases reached 0.97 recall, indicating the model reliably avoids missing non-cancerous cases.

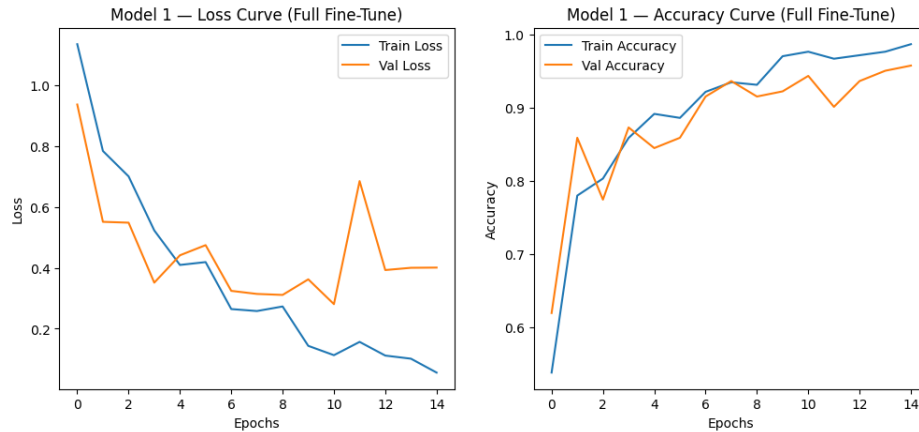


Figure 2: Model 1 Swin Transformer Loss and Accuracy Curves

3.3 Model 2a: Transfer Learning from Model 1 Weights

Phase 1 training started from Model 1’s already adapted weights, so validation accuracy began at 93.7% and remained stable throughout. Phase 2 fine-tuning ran for all 10 epochs. Model 2a achieved the best overall performance with 91.2% test accuracy and a macro F1 of 0.91. All per-class metrics improved or held steady relative to Model 1, with large cell carcinoma maintaining recall and normal cases at 0.97 recall.

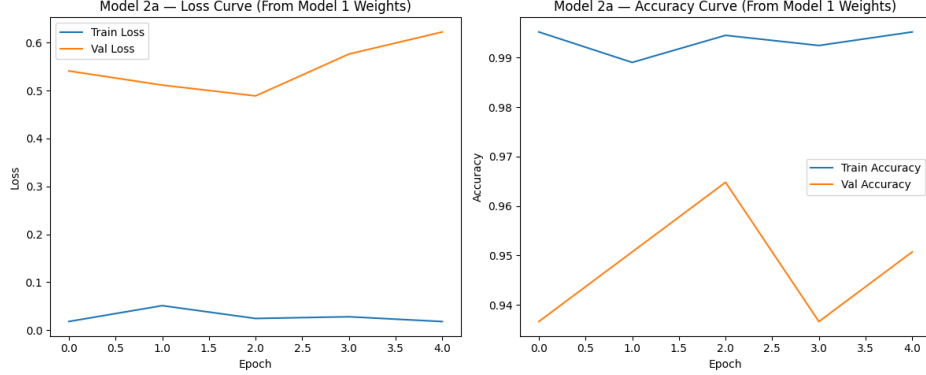


Figure 3: Model 2a Transfer Learning (Adapted) Loss and Accuracy Curves

3.4 Model 2b: Transfer Learning from Fresh ImageNet Weights

Phase 1 training began from random classifier weights over a newly loaded ImageNet backbone, reaching only 45.1% validation accuracy after 5 epochs, far below Model 2a’s starting point. Phase 2 improved performance but early stopping triggered at epoch 9 with a best validation accuracy of 92.3%. On the test set, Model 2b reached 83.6% accuracy and a macro F1 of 0.82. The confusion matrix shows squamous cell carcinoma had the lowest recall at 0.59, suggesting the frozen backbone in Phase 1 failed to develop sufficient discriminative power before fine-tuning began.

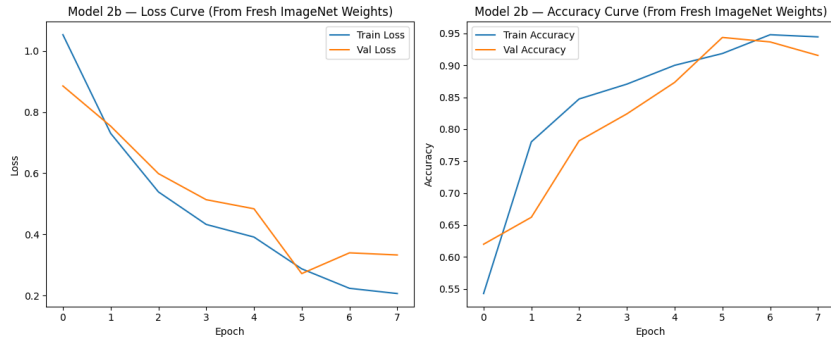


Figure 4: Model 2b Transfer Learning (Fresh) Loss and Accuracy Curves

Table 1 summarizes all four models across the three primary metrics. Figure 5 shows the confusion matrices for all four models on the held-out test set.

Table 1: Test Set Performance Across All Models

Model	Test Accuracy	Macro F1	Macro Recall
Baseline (PCA + Logistic Regression)	0.510	0.491	0.551
Model 1: Swin Full Fine-Tune	0.905	0.899	0.896
Model 2a: Swin TL from Model 1	0.912	0.905	0.897
Model 2b: Swin TL from Scratch	0.836	0.822	0.825

Three findings are shown from these results. First, deep learning is far better at such a task than classical methods, as shown by the 41 percent gap between the baseline and Model 2a. This confirms that spatial feature extraction centered around attention finds relevant patterns that PCA and logistic regression cannot. Second, initialization matters

significantly for transfer learning in that Model 2a outperforms Model 2b by 7.6 percent, despite using identical training procedures. Such a finding indicates that the quality of the starting checkpoint may be as important as the fine-tuning strategy itself. Third, full fine-tuning like we achieved with Model 1 performs comparably to two-phase transfer learning from a strong checkpoint shown via Model 2a, suggesting that for this dataset size, the primary bottleneck is representation quality rather than training strategy. Squamous cell carcinoma was the most challenging class across all models, likely due to its visual similarity with other carcinoma subtypes in CT imaging. The malignant cases category also showed lower recall in some models, likely due to label ambiguity since this category overlaps with other cancer subtypes also present in the dataset.

3.5 Comparison and Discussion

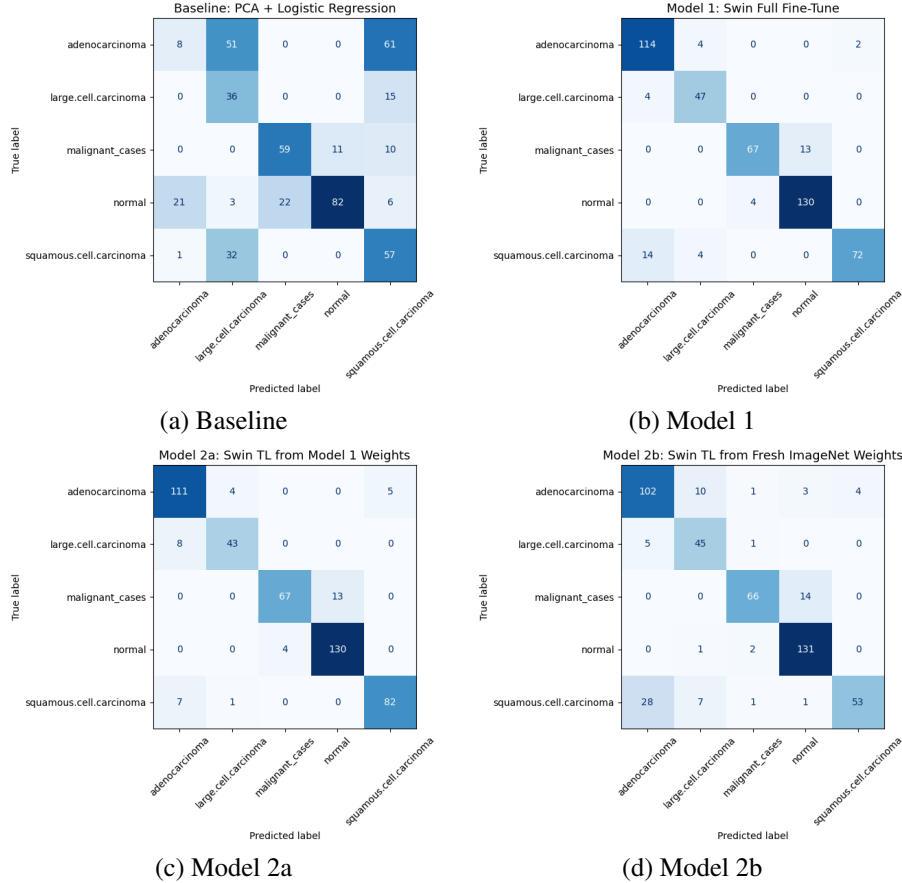


Figure 5: Confusion matrices for all four models on the held-out test set.

4 Conclusion

In this project, we compared four approaches to automated lung cancer classification from CT scan images: a PCA + Logistic Regression baseline, a fully fine-tuned Swin Transformer, and two variants of two-phase transfer learning differing only in their initialization. Our best model, Model 2a, achieved 91.2% test accuracy and a macro recall of 0.897, a 41 percentage point improvement over the classical baseline. The results show that deep learning is necessary for this task, as linear classifiers operating on pixel features fail to capture the spatial patterns that distinguish cancer subtypes. More importantly, we find that initialization strategy has a larger impact than training procedure for transfer learning on small medical imaging datasets: starting from an already domain-adapted checkpoint substantially outperformed starting from generic ImageNet weights under identical fine-tuning conditions. Limitations of this work include the relatively small dataset size, label ambiguity in the malignant cases category, and the use of a single dataset source. Overall, this project shows that deep learning has strong potential for assisting in medical image analysis, particularly for problems like cancer detection. While not a replacement for medical professionals, these models can serve as valuable tools for more consistent diagnoses. Future work could focus on improving generalization through larger and more diverse datasets, as well as integrating additional clinical information to enhance prediction reliability.

References

- [1] Armato, S. G., et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI). *Medical Physics*, 38(2):915–931, 2011.
- [2] Esteva, A., et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.
- [3] Ardila, D., et al. End-to-end lung cancer screening with deep learning on low-dose CT. *Nature Medicine*, 25:954–961, 2019.
- [4] Liu, Z., et al. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV 2021*.
- [5] Rathi, D. CT scan images for lung cancer. Kaggle, 2022. <https://www.kaggle.com/datasets/dishantrathi20/ct-scan-images-for-lung-cancer>