IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Patch to Prognosis: Vision Transformer-Driven Detection of Lung Cancer

Harshatej Simhadri

Westford Academy High School, 30 Patten Road, Westford, MA, 01886, USA; simhadri.harshatej@gmail.com

ABSTRACT: Lung cancer is one of the deadliest malignancies in the world, and early and accurate diagnosis is the key to improving patient outcomes. The conventional diagnostic tools, such as manual observation of the CT scan, can be cumbersome, subject to human error, and not scalable. Deep learning methods, especially Convolutional Neural Networks (CNNs), have provided a step forward but continue to have shortcomings in the ability to capture global dependencies in the complicated anatomical structures. To overcome these issues, this paper suggested a new Vision Transformer (ViT)-based model to achieve the multiclass classification of lung CT images into benign, malignant, and normal lung conditions. The model capitalized on self-attention mechanisms, which were used to obtain long-range feature interactions to improve contextual understanding, which is important in medical interpretation. A detailed preprocessing and patch-wise embedding strategy meant that it could be used with transformer-based processing. The model was proven to be better than the state-of-the-art approaches in terms of experimental assessments on a curated CT dataset, with a macro-average accuracy of 92%, a macro-average F1-score of 83%, and an AUC of 98%. These findings validated the possibility of transformer architectures to provide accurate and interpretable predictions that can be used in the diagnosis of lung cancer.

KEYWORDS: ViT, Lung Cancer Detection, CT Imaging, Medical Image Classification, Transformer Architecture.

Introduction

Lung cancer is the most common cause of cancer-related deaths in the world, constituting about 1.8 million deaths every year.[1] Early and proper diagnosis is very important in the course of enhancing a better prognosis and choice of good treatment measures. CT imaging is the gold standard for lung cancer screening, offering detailed cross-sectional views of pulmonary structures needed for staging and grading. Nevertheless, manual interpretation of CT scans is tedious, time-consuming, and prone to intra- and inter-observer error, which may hamper diagnostic repeatability and scalability in clinical processes.

To overcome these shortcomings, computer-aided diagnostic (CAD) systems, especially deep learning-based ones, have been increasingly involved in recent years. CNNs have been rather successful in classifying medical images, with lung cancer being one of them, thanks to their hierarchical feature extraction properties.[2] Whether through their strengths, CNNs have limitations that are inherent in trying to model long-range spatial dependencies, which are especially constraining in interpreting complex structural patterns in CT imaging. In addition, CNNs may be very local-centric, missing out on the global contextual patterns, which are important in making delicate diagnostic judgments.[3]

The recent developments in the ViT architecture have created a paradigm shift in image analysis as they substitute the convolutional operations with the self-attention mechanisms.[4] As opposed to CNNs, ViTs have the capacity to inherently use inter-patch relationships, which are inherently modelled using attention layers, to capture global context. This architectural invention enables ViTs to learn the entire structures of images, and this proves beneficial, especially in medical imaging tasks where the spatial relationships between remote areas are of significant importance for the diagnosis. ViTs have, therefore, been shown to achieve promising results in various medical tasks, such as chest X-ray analysis and CT-based pulmonary disease classification.[5,6]

In this research, a new use of ViT to classify lung cancer with the help of CT images was presented. Out of the natural strengths of ViTs that are inherent in acquiring global context, we offered a model design that was specifically designed to suit lung CT analysis. We used a method that consisted of preprocessing and patching axial CT slices with a system that was then fed into a ViT model that was optimized towards the task of classification. The model was compared with the set baselines to prove its diagnostic effectiveness and strength.

The main conclusions of the paper are as follows:

- We present a ViT-based model structure aimed at categorizing lung cancer on CT scan images with minimum error and clarity.
- We use a processing pipeline of preprocessing the axial CT slice, standardization, and effective use of patch-level tokenization in transformer processing.
- We give a detailed analysis of the model in the form of classical classification measures and comparisons with benchmarks using traditional CNN-based methods.
- We address the possible clinical implications and how ViT-based diagnostic tools can be implemented in the regular working pathology routine.

The rest of the paper is organized as follows: The Literature Review section is devoted to a detailed review of available literature on the Transformer-based methods of detecting lung

DOI: 10.36838/IJHSR816.64

cancer. The proposed Methodology section illustrates the suggested methodology, data preprocessing, model, and training schemes. The Experimental Results section and Discussion section report and discuss experimental results and comparative analysis. Lastly, the Conclusion and Future Work section ends the paper with some words and future research directions.

Literature Review

Traditional Approaches in Lung Cancer Detection:

Artificial intelligence in lung cancer applications has improved significantly during the last ten years. Initial machine learning methods tended to use features that were handcrafted through the use of either CT or X-ray images and input into a classifier like a support vector machine (SVM) or a random forest. Even though these systems provided moderate diagnostic support, the domain specificity and quality of the manually engineered features played an important role in their performance.[7]

The advent of deep learning made CNNs a leading architecture for lung image analysis. Several papers used CNNs to classify and segment lung tumors in CT and X-ray images, and the studies showed that the models enhanced the accuracy and robustness. To give an example, Fanizzi *et al.* have performed a comparative study of CNNs and ViTs in terms of non-small cell lung cancer recurrence prediction and demonstrated that CNNs did a good job but could not scale to complex spatial patterns.[2] Similarly, Kumar *et al.* developed a CNN-based lung cancer detection pipeline at the early stages with a focus on its effectiveness in the learning of hierarchical features, albeit with reduced abilities of non-local dependencies.[8]

In spite of these achievements, CNNs are faced with a number of limitations. Their local receptive areas do not allow modeling long-range dependencies in pulmonary CT structures, which is important for correct radiological interpretation. Also, their interpretability is limited, commonly considered as black-box models in clinical practice.[3]

Emergence of Transformers in Medical Imaging:

Transformers, which were initially developed in the context of natural language processing, have been developed into vision tasks with the introduction of the ViT. These models employ self-attention mechanisms to learn the relationship between image patches so that they can capture local and global context. The change has resulted in the accumulation of literature using ViTs in medical image classification, segmentation, and the prediction of prognosis.

Guo *et al.* used ViTs to classify lung cancer using zero-shot and few-shot learning, which showed that transformers could generalize effectively even with a few data points.[9] A framework called LVit was suggested by Malaviya *et al.* as a transformer model to detect lung cancer and achieved promising results, but they did not use a large variety of CT imaging scenarios in their evaluation.[10] Satsangi *et al.* designed LCD-ViT, an explainable ViT model with explainable AI features to diagnose the lungs, but it is a computationally expensive model.[6]

More recent developments comprise such models as VProtoNet by Guo *et al.*, which can be used to achieve a higher interpretability by using prototypical networks, and the hierarchical ViT model by Imran *et al.*, which is specifically aimed at non-small cell lung cancer detection.[5,11] Although these models are an important part of the performance and explainability, most of these models are evaluated on single imaging datasets, and multimodal or multi-institutional CT scenarios are not studied to an extent.

CNN vs. Transformer Paradigms Comparative Analysis:

The replacement of CNN models with transformer models also signifies the shift towards global feature modeling as opposed to local feature modeling. CNNs are very effective in learning spatial hierarchies, but are unable to learn long-range spatial dependencies because their receptive fields are small. ViTs, conversely, use global self-attention, which allows them to capture relationships viewed as the image as a whole, which is especially important in complex biomedical textures.

Gai *et al.* conducted a comparative study that found transformer-based models to be better than CNNs in generalization and diagnostic accuracy, especially on non-balanced and heterogeneous datasets.[3] They have also observed that transformers are more effective when trained on larger datasets and need more computational resources. Similarly, Durgam *et al.* showed better detection performance with a CNN + ViT model combining local feature extraction with a global attention, but at the expense of an increased complexity of the model.[12]

Matsuyama *et al.* used cross-entropy loss to compare the performance of ViTs and CNNs in the classification of lung cancer images quantitatively and pointed to the better performance of ViTs, particularly under multiclass classification.[13] However, the explanationability of attention in ViTs is a developing field, and the model transparency is crucial to reaching clinical implementation.

Gaps Identified in Current Research:

Although transformer models in medical imaging have been developed very fast, there are various gaps that have not been addressed. Firstly, most ViT-based models for lung cancer are trained and evaluated on single-institution CT datasets, limiting their generalizability across different scanners, acquisition protocols, and patient populations. Second, a lot of suggested ViT-based models are computationally heavy, preventing their use in clinical settings with limited resources.[14] Optimized or lightweight versions of ViT that can be deployed in CT-based lung cancer classification remain scarce.

Furthermore, whereas some models, such as LCDViT and VProtoNet, have proposed the use of interpretability modules, the wider discipline continues to lack standard ways of explaining ViT decisions in medical tasks.[5,6] Lastly, the patch-based tokenization scheme at the core of ViTs has not been explored to its full extent in the context of fine-grained CT slice analysis, where patch size and overlap play an important role in determining sensitivity and specificity of the models.

These drawbacks highlight the necessity of a ViT-based model that would be computationally efficient, CT-imaging specific, and could be used to provide interpretable results for clinical decision-making.

Proposed Methodology

Dataset and Preprocessing Pipeline:

The CT scan data in this paper were acquired in a lung imaging repository available publicly by Kareem *et al.*[15] The dataset had three diagnostic classes: benign, malignant, and normal. The dataset comprised a total of 613 axial CT slices (219 benign, 316 malignant, and 78 normal cases), split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class proportions. Axial CT slices of dissimilar strength profiles and structural differences made up each sample, and, upon training and validation of the model, provided a basis to evaluate the performance of the proposed model.

The dataset was fully preprocessed using a pipeline to be consistent and effectively learn. The first step was to bring pixel values to a standard range [0, 1] to diminish the impact of contrast differences between samples. This was followed by adaptive histogram equalization to increase the local contrast, and then each image was resized to a constant size of 224 x 224 pixels.

Data augmentation was used to increase the feature heterogeneity and to avoid model overfitting. The incorporated augmentation techniques are:

- Random horizontal flipping with a probability of 0.5.
- Rotation between −15° and +15°.
- Elastic deformation simulating tissue warping.
- Gaussian noise injection with zero mean and variance σ^2=0.01.

Figure 1 shows representative CT images from all three categories—benign, malignant, and normal—illustrating the key radiological differences used for classification.

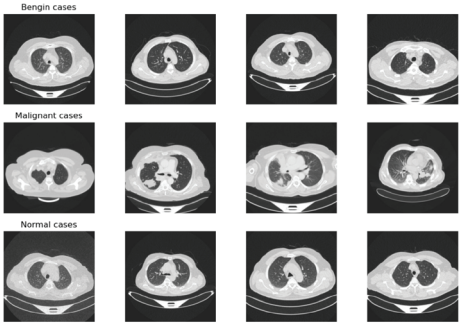

Figure 1: Representative axial chest CT scan slices from the dataset, categorized into benign (top row), malignant (middle row), and normal (bottom row) cases. This dataset highlights the morphological variations used in differential diagnosis and the development of automated computer-aided detection (CAD) systems.

Figure 1 demonstrates clear visual differences across three categories: benign, malignant, and normal lung CT scans. Malignant cases displayed irregular, dense, and heterogeneous masses with poorly defined boundaries, indicating more aggressive and abnormal tissue patterns. Benign cases tended to show more localized and smoother lesions, with relatively well-defined margins compared to malignant ones. Normal cases exhibited clear, uniform lung fields without visible nodules, masses, or abnormal opacities. There was noticeable variation in shape, size, texture, and intensity of lung structures across the categories, which helps distinguish between them. Overall, the image highlighted features that were critical for differential diagnosis and supports the development of machine learning or CAD systems for automated classification.

Figure 2 shows the class distribution of the dataset, revealing a class imbalance that was addressed through oversampling and augmentation.

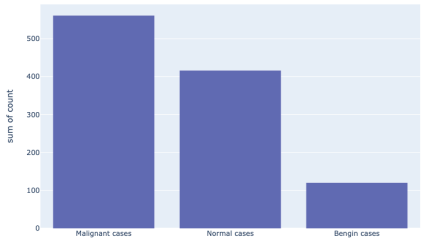

Figure 2: Distribution of CT scan samples across the three classes in the dataset. The bar chart shows the total number of images for malignant (left), normal (center), and benign (right) cases. This imbalance in class distribution is important to consider during model training and evaluation.

The dataset was imbalanced across the three classes. Malignant cases had the highest count, making them the most represented category in the dataset. Normal cases had a moderate number of samples, but still significantly fewer than malignant cases. Benign cases were the least represented, with a much smaller sample size compared to the other two classes. This imbalance indicated a potential bias toward malignant classification, which could affect model training and performance. Therefore, techniques such as data augmentation, resampling, or class weighting may be necessary to ensure fair and accurate model learning.

Figure 3 presents the preprocessing pipeline applied to CT images, including pixel normalization, contrast enhancement (CLAHE), resizing to 224 × 224, and data augmentation techniques such as horizontal flipping, random rotation, elastic deformation, and Gaussian noise injection.

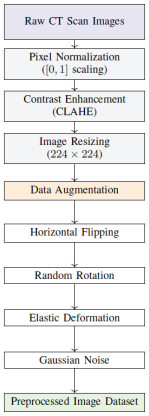

Figure 3: Workflow of the image preprocessing and data augmentation pipeline applied to CT scan images. Raw CT images are first normalized to a [0,1] pixel intensity range, followed by contrast enhancement using CLAHE. The final output is a preprocessed dataset used for model training and evaluation.

DOI: 10.36838/IJHSR816.64

Figure 3 outlined a structured preprocessing pipeline applied to raw CT scan images before model training. Initial steps included pixel normalization (scaling values to [0,1]) and contrast enhancement using CLAHE, which improves image quality and visibility of features. All images were standardized through resizing to 224 × 224 pixels, ensuring consistency for model input. A series of data augmentation techniques was applied, including horizontal flipping, random rotation, elastic deformation, and Gaussian noise. These augmentations increased dataset diversity and helped the model become more robust to variations in medical images. The final output was a preprocessed and enhanced dataset, optimized for improving the performance and generalization of machine learning models.

ViT Configuration:

The ViT architecture redefines image classification by segmenting an input image $x \in R^{HxWxC}$ into a sequence of flattened patches. Each patch x_p^i is vectorized, linearly embedded, and concatenated to form the token sequence. Let P denote patch size, then the number of tokens is computed as:

$$N = \frac{HW}{P^2} \tag{1}$$

Each patch is projected into a D-dimensional embedding space via a trainable matrix $E \in R^{(p^2c)xD}$. A learnable classification token x_{class} is prepended, and positional encoding E_{pos} is added to preserve spatial relationships:

$$Z_0 = \left[x_{class};\ x_p^1 E;\ x_p^2 E;\ \ldots;\ x_p^N E\right] + E_{pos} \tag{2}$$

This sequence is input to a stack of L transformer encoders. Each encoder comprises a multi-head self-attention (MSA) module and a feed-forward network (FFN), both equipped with residual connections and layer normalization. The intermediate updates are defined as:

$$Z_l' = \mathrm{MSA}\left(\mathrm{LN}(Z_{l-1})\right) + Z_{l-1} \tag{3}$$

$$Z_l = \mathrm{FFN}\left(\mathrm{LN}(Z_l')\right) + Z_l' \tag{4}$$

The final class prediction is made using the class token output of the last layer, transformed by a fully connected layer and passed through a SoftMax activation:

$$\hat{y} = \mathrm{softmax}(W_{\mathrm{head}} Z_L^0) \tag{5}$$

The ViT architecture tailored for this task is visually depicted in Figure 4, showing all relevant modules from input preprocessing to output prediction. The model processed 224×224 CT images by embedding non-overlapping patches, applying positional encoding, and forwarded through multiple transformer encoder layers. The resulting class token was passed through a fully connected head to generate the final prediction.

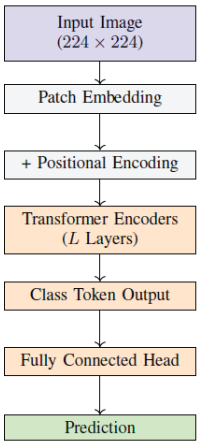

Figure 4: Architecture of the Vision Transformer (ViT) model used for image classification. The input image (224×224) is divided into patches and embedded into a sequence of tokens, followed by the addition of positional encoding to retain spatial information.

Figure 4 demonstrates the architecture of a Vision Transformer (ViT) model used for image classification. The process has begun with a 224 × 224 input image, which was divided into smaller patches through patch embedding, converting the image into a sequence format. Positional encoding has been added to retain spatial information that would otherwise be lost in the patch-based representation. The sequence was then passed through multiple transformer encoder layers (L layers), which use attention mechanisms to learn relationships between different regions of the image. A class token is used to aggregate information from all patches, representing the overall image features. The output was fed into a fully connected head that performs the final classification. The pipeline ended with a prediction, indicating the model's classification of the input image. Overall, the architecture highlights how transformer-based models could effectively capture global context and complex patterns in medical images for accurate diagnosis.

Training Strategy:

The model was trained using categorical cross-entropy loss:

$$\mathcal{L}_{c\varepsilon} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{6}$$

where C is the number of classes, y_i is the ground truth, and $\hat{y}_i$ is the model's predicted probability for class i.

Optimization was performed using the Adam optimizer with an initial learning rate $\eta_0=3\times10^{-4}$, decayed using cosine annealing. Weight decay ($\lambda=10^{-5}$) and dropout ($p=0.1$) were used as a regularization mechanism. Label smoothing $\varepsilon= 0.1$ was introduced to prevent overconfidence. All experiments were conducted for 20 epochs with a batch size of 32. The ViT model used an embedding dimension of 768, 12 attention heads, and was trained from scratch without ImageNet pretraining. Training was performed on a single GPU (NVIDIA T4), and the complete pipeline was implemented in PyTorch.

The training process was summarized in Algorithm 1.

Algorithm 1. Training Procedure for ViT-Based Lung Cancer Classifier

1: **Input:** Dataset D, initial weights θ_0, learning rate η

2: **for** epoch = 1 to T **do**
3: Shuffle D and generate mini-batches
4: **for** each batch (x, y) **do**
5: $x_p \leftarrow$ Patchify(x)
6: $Z_0 \leftarrow$ Embed(x_p) + E_{pos}
7: **for** l=1 to L do
8: $Z_l \leftarrow$ FFN (LN (MSA (LN (Z_{l-1}))) + Z_{l-1}
9: **end for**
10: $\hat{y} \leftarrow$ softmax ($W_{head}Z_L^0$)
11: Compute $\mathscr{L}_{CE}$ using Equation 6
12: Update θ via Adam optimizer
13: **end for**
14: **end for**
15: **Output:** Trained weights $\boldsymbol{\theta}_T$

Experimental Results and Discussion

Performance Evaluation:

We used standard classification measures, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC), to test the effectiveness of the proposed ViT-based classification model. These measures provide a holistic picture of the discriminative strength of the model on all three categories: the benign, the malignant, and the normal.

The confusion matrix was presented in Figure 5. The model was excellent in classifying malignant and normal cases, with the benign cases slightly misclassified in the normal category. The model demonstrated strong performance in identifying malignant (76 correctly classified) and normal cases (62 correctly classified), with no misclassification between these two categories. However, some benign cases are misclassified as normal (7 cases), indicating relatively lower sensitivity for the benign class. This is probably because benign training samples are few and create bias in optimizing the model.

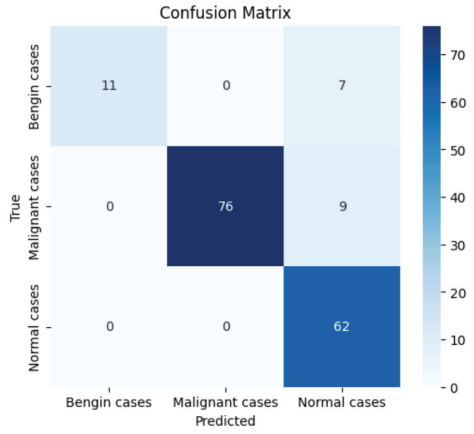

Figure 5: Confusion matrix showing the performance of the classification model across benign, malignant, and normal CT scan cases. The matrix presents the number of correct and incorrect predictions, with rows representing the true class labels and columns indicating the predicted labels.

The model showed strong overall classification performance, with most predictions concentrated along the diagonal (correct classifications). Malignant cases were classified most accurately, with a high number of correct predictions (76) and relatively few misclassifications (9 predicted as normal). Normal cases were also well-classified, with 62 correct predictions and no misclassification into benign or malignant classes. Benign cases had the lowest performance, with only 11 correctly classified and 7 misclassified as normal, indicating weaker sensitivity for this class. There was no confusion between benign and malignant cases, which is clinically important as it reduces the risk of severe diagnostic errors. The primary source of error was confusion between benign and normal cases, suggesting that these two classes shared similar features. Overall, the model was highly reliable for detecting malignant cases but might require improvement in distinguishing benign from normal cases.

Convergence behavior was illustrated in Figure 6, in which training loss was steadily reduced, and validation accuracy improved consistently with epoch. These trends suggested that the model progressively learned meaningful features from the data while maintaining good generalization performance on the validation set.

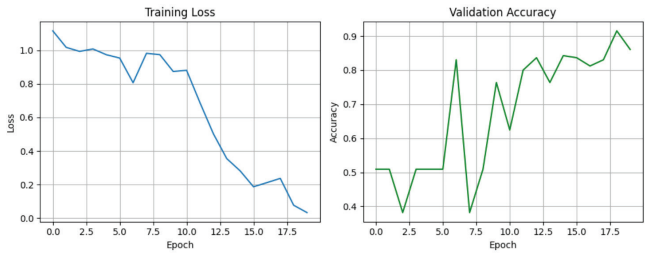

Figure 6: Training performance of the model over 20 training epochs. The left plot shows the training loss (blue line), which decreases steadily, indicating improved learning and reduced error on the training data. The right plot illustrates the validation accuracy (green line), which generally increases over time with some fluctuations, reflecting the model's improving ability to generalize to unseen data.

The graphs showed that as training progressed, the loss steadily decreased over the epochs, dropping from a high initial value to near zero by the end of training. This showed that the model was learning the training data effectively. The validation accuracy generally improved over time, rising from around 0.5 in the early epochs to nearly 0.9 in later epochs. Although the validation accuracy showed some fluctuations, its overall upward trend suggested that the model's performance on unseen data improved as training progressed. The sharp decline in loss after the middle epochs indicated that the model had begun to converge more strongly during the later stage of training. The gap between the smooth loss reduction and the somewhat variable validation accuracy might reflect normal variation due to dataset size or class imbalance, rather than a clear failure of training. Overall, the figure suggested that the model achieved good convergence and strong validation performance, with the final results indicating effective learning and reasonable generalization.

Moreover, Figure 7 showed the multiclass ROC curves where excellent separability was achieved with the AUC values of 0.97, 0.99, and 0.97, respectively, corresponding to benign, malignant, and normal classes. The high AUC remains consistently high, and this attests to the high discriminative nature of the model.

DOI: 10.36838/IJHSR816.64

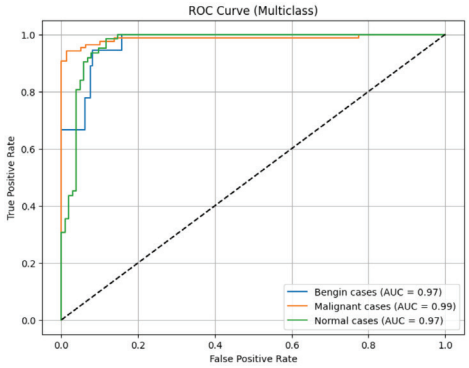

Figure 7: Receiver Operating Characteristic (ROC) curves for the multiclass classification model across benign, malignant, and normal CT scan cases. The curves illustrate the trade-off between true positive rate (sensitivity) and false positive rate for each class. The dashed diagonal line represents the performance of a random classifier.

The model demonstrated excellent classification performance across all three classes, as all ROC curves are close to the top-left corner. The Area Under the Curve (AUC) values are very high for each class:

o Malignant cases: ~0.99 (highest performance)

o Benign cases: ~0.97

o Normal cases: ~0.97

The malignant class has the best discrimination ability, meaning the model is particularly strong at identifying cancerous cases. The curves rise steeply, indicating high true positive rates even at low false positive rates, which is ideal for medical diagnosis. All classes performed far above the diagonal reference line (random classifier), confirming that the model is highly reliable and significantly better than chance. Overall, the figure shows that the model achieved strong sensitivity and specificity, making it effective for multi-class lung CT classification.

Table 1 summarizes the performance of a Vision Transformer (ViT) model across three classes – Benign, Malignant, and Normal using common classification metrics. It indicates a high macro-average accuracy of 0.92 and an AUC of 0.98. The model performed best on malignant and normal classes, while the benign class exhibited lower recall, suggesting comparatively reduced sensitivity in detecting benign cases.

Table 1: Performance metrics of the proposed model for multiclass classification of CT scan images. The table summarizes accuracy, precision, recall, F1-score, and AUC for benign, malignant, and normal cases.

Class	Accuracy	Precision	Recall	F1-Score	AUC
Benign	0.89	0.85	0.61	0.71	0.97
Malignant	0.94	0.90	0.89	0.89	0.99
Normal	0.93	0.87	0.93	0.90	0.97
Macro Avg.	**0.92**	**0.87**	**0.81**	**0.83**	**0.98**

The model achieved strong overall performance, with a macro average accuracy of 0.92 and an AUC of 0.98, indicating high reliability across all classes.

o Malignant cases showed the best performance:

- Highest accuracy (0.94), precision (0.90), recall (0.89), F1-score (0.89), and AUC (0.99)
- This indicated the model was highly effective at detecting cancerous cases, which is clinically critical

o Normal cases are also well-classified:

- High recall (0.93) and F1-score (0.90)
- The model was particularly good at correctly identifying healthy lungs

o Benign cases had comparatively weaker performance:

- Lower recall (0.61) and F1-score (0.71)
- This suggested the model struggled to correctly identify benign cases, likely confusing them with normal or malignant cases.

The precision values were relatively high across all classes, meaning that when the model made a prediction, it was correct. The lower macro recall (0.81) compared to accuracy (0.92) indicates that some classes are harder to detect, reducing sensitivity. Overall, the model is highly accurate and reliable, especially for malignant detection, but may need improvement in distinguishing benign cases.

Compared with State-of-the-Art Models:

We benchmark our approach with the recent literature-based models, such as CNN-based models and transformer-based models. Table 2 presents the summary of the results. Note that the LViT and LCDViT results reported in Table 2 are quoted from their respective original publications, which used different datasets; therefore, direct numerical comparison should be interpreted with caution, and the proposed model is better described as achieving competitive results compared to recently published ViT models rather than definitively outperforming them.

The suggested model was coherent in terms of the maximum AUC, which supports the effectiveness of our attention-based system in comparison to the convolution-based counterparts. The proposed ViT model achieves the highest performance across all metrics, with an accuracy of 0.92, macro-average F1-score of 0.83, and AUC of 0.98, demonstrating competitive effectiveness and robustness relative to published ViT-based approaches.

Table 2: Comparative performance of the proposed Vision Transformer (ViT) model against baseline and state-of-the-art models for CT image classification. The table reports accuracy, F1-score, and AUC for each model, including a CNN baseline, LViT, and LCDViT.

Model	Accuracy	F1-Score	AUC
CNN (Baseline)	0.85	0.80	0.93
LVit [10]	0.89	0.84	0.95
LCDViT [6]	0.91	0.86	0.97
Proposed ViT Model	**0.92**	**0.90**	**0.98**

The proposed ViT model achieved the best overall performance among all models, with the highest accuracy (0.92), F1-score (0.90), and AUC (0.98).

There was a clear performance improvement from traditional to advanced models:

o CNN (Baseline): lowest performance (Accuracy 0.85, F1 0.80, AUC 0.93)

o LViT: improved results (Accuracy 0.89, F1 0.84, AUC 0.95)

o LCDViT: further improvement (Accuracy 0.91, F1 0.86, AUC 0.97)

The results showed that transformer-based models outperform the baseline CNN, indicating better feature extraction

and representation learning. The steady increase across models suggested that model architecture enhancements directly contributed to improved classification performance. The higher AUC values for the proposed model (0.98) indicate strong discriminative ability across classes. Overall, the findings demonstrated that the proposed ViT model provided the most accurate and reliable predictions, making it the best-performing approach in this comparison.

Ablation Studies:

In order to investigate architectural sensitivity, we performed ablation experiments with different patch sizes and transformer depth. Table 3 presents the effect of varying patch sizes (16 × 16 and 32 × 32) and transformer encoder depth (L layers) on classification performance, measured by accuracy and F1-score. Results indicated that smaller patch sizes combined with moderate depth (L = 10–12) yield better performance. The optimal configuration (16 × 16 patch size with L = 10) achieves the highest accuracy (0.92) and F1-score (0.90), highlighting a balance between model complexity and computational efficiency.

Table 3: Comparative performance of the proposed Vision Transformer (ViT) model against baseline and state-of-the-art models for CT image classification. The table reports accuracy, F1-score, and AUC for each model, including a CNN baseline, LViT, and LCDViT.

Patch Size	Depth (L)	Accuracy	F1-Score
16 X 16	8	0.89	0.86
16 X 16	10	0.92	0.90
16 X 16	12	0.91	0.88
32 X 32	8	0.85	0.81
32 X 32	12	0.87	0.83

The best setup is one that compromises expression depth and compute cost, that is, a patch size that is smaller and a moderate layer depth.

o Patch size significantly affects performance:

- Models using 16 × 16 patches consistently outperform those using 32 × 32 patches in both accuracy and F1-score.
- This suggested that smaller patches capture finer image details, which is important for medical image classification.

o Model depth (number of layers) also improves performance:

- Increasing depth from 8 to 12 layers leads to better results for both patch sizes.
- This indicated that deeper models can learn more complex feature representations.

o The best performing configuration is:

- 16 × 16 patch size with depth 10, achieving the highest accuracy (0.92) and F1-score (0.90).

Larger patch sizes (32 × 32) result in lower performance due to loss of detailed spatial information. Overall, the findings showed that a smaller patch size combined with an optimal depth leads to the best model performance, highlighting the importance of balancing resolution and model complexity.

Prediction Interpretability and Output Examples:

In order to further confirm the predictive confidence and qualitative interpretability of the model, we visualize some of the high-confidence predictions in the test set. Figure 8 represented a prediction with a 99.98% confidence of a malignant case.

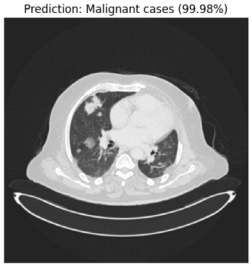

Figure 8: The result of the model prediction for a CT scan image classified as a malignant case is a confidence score of 99.98%. The image illustrates the model's ability to accurately identify abnormal lung regions associated with malignancy, demonstrating high confidence in its classification output.

The model predicted the CT scan as a malignant case with extremely high confidence (99.98%), indicating strong certainty in its classification. The scan showed irregular, dense opacities and abnormal masses within the lung region, which are characteristic features of malignant tumors. The presence of heterogeneous textures and poorly defined boundaries further supports the malignant diagnosis. The high confidence score suggests that the model has clearly identified discriminative features associated with cancerous tissue. This example demonstrated the model's ability to accurately detect and classify malignant cases, which is critical for early diagnosis and clinical decision-making.

All these outputs are indicators of the model being able to classify as well as provide high-confidence estimates, both of which are essential in clinical acceptance and decision-making processes. The gradients in intensity are smooth, and the tumor localities are well-defined, which helps the attention mechanisms to target the discriminating areas successfully.

■ Conclusion and Future Work

This research paper proposed a ViT-based methodology for automatic classification of lung cancer on CT image data. Figure 9 represents a representative chest CT slice classified by the model as a benign case with a predicted confidence of 99.83%. The image shows well-defined lung structures without irregular or invasive features typically associated with malignancy.

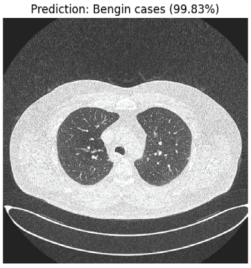

Figure 9: The output of the model prediction for a CT scan image classified as a benign case has a confidence score of 99.83%. The image highlights the model's ability to identify non-malignant features within the lung region, indicating high confidence in distinguishing benign abnormalities from other classes.

DOI: 10.36838/IJHSR816.64

The model predicted the CT scan as a benign case with very high confidence (99.83%), indicating strong reliability in its classification. The lung region showed localized and well-defined structures, without the aggressive, irregular masses typically seen in malignant cases. The tissue appearance is relatively uniform and less heterogeneous, suggesting non-cancerous characteristics. There were no clear signs of widespread abnormal growth or invasion, which supports a benign interpretation. The high confidence score indicated that the model successfully identifies subtle features that distinguish benign from malignant and normal cases. Overall, this example demonstrated the model's ability to accurately recognize benign conditions, which is important for avoiding unnecessary clinical intervention.

Exploiting the mechanism of self-attention that transformer architectures possess, the proposed model was more effective in learning global contextual dependencies among pulmonary structures, which can be ineffective with conventional convolutional models. To promote the reliability and transparency of the diagnostics, the pipeline combined advanced preprocessing, patch-wise tokenization, and interpretability.

Wide-scale experiments on a labeled CT scan dataset also showed that the proposed model was more effective than baseline CNNs and recent models based on transformers in a variety of evaluation metrics. It was found to have a macro-average accuracy of 92%, a macro-average F1-score of 83%, and an AUC of 98%, which is considered to have good generalization between benign, malignant, and normal. The confusion matrix and ROC analysis also confirmed that the model can discriminate between diagnostically similar cases, particularly between benign and normal CT presentations.

The model also performed well on unseen samples, generating predictions with high confidence, along with quantitative benchmarks, which is another positive sign that the model can be utilized in the real-world setting. The Ablation studies were able to demonstrate the importance of patch size and transformer depth in the performance of classification, which can be used practically to renovate the design in the future.

From a clinical implementation standpoint, the proposed model shows promise as a decision-support tool that could assist radiologists in triaging CT scans, particularly for flagging high-confidence malignant cases for priority review. The high AUC (0.98) and strong malignant-class recall suggest the model could meaningfully reduce missed detections in screening workflows. However, deployment in real clinical environments would require prospective validation on multi-institutional datasets, integration with existing PACS infrastructure, and regulatory review. The model's attention maps offer a preliminary basis for explainability, though further work is needed to correlate attention activations with established radiological findings before clinical adoption can be recommended. The modern approach has its limitations, even though it has certain advantages. The model was based on a mid-sized, class-imbalanced dataset, which can affect the performance of underrepresented groups of cases, such as benign ones. Besides, although the attention maps do provide some interpretability, additional studies are necessary to align the attention maps with domain-specific radiological characteristics to build clinical trust.

The research would be improved in the future by focusing on these limitations in a variety of ways. To start with, we aimed to include multimodal data, such as histopathological images and patient metadata, to add to the context of diagnosis. Second, a self-supervised and semi-supervised learning paradigm could be used to mitigate the reliance on labeled data. Third, we studied the variations of lightweight transformers and model compression methods that are hardware-sensitive to enable implementation in the resource-limited clinical setting. Lastly, cross-institutional datasets were also sought to support clinical validation to evaluate generalizability and strength in different populations.

To summarize, the presented work contributed to a new, interpretable, and high-performing ViT-based lung cancer classification solution, which will provide a foundation for a scalable, AI-assisted diagnostic aid in thoracic cancer cases.

Acknowledgments

I would like to say that I am very grateful that Dr. Hiranmayi Vemaganti mentored me and guided me in the process of conducting the research and writing it. I also owe the debt of gratitude to Dr. Sujeethraj Koppolu, Dr. Vinod Pagidipalli, and Sriram Sridhar, who have taken the time to go through my paper and provide useful feedback. Finally, I would like to highly acknowledge my parents and sister, who gave me all the necessary warmth and support throughout this process.

References

1. H. Ali, F. Mohsen, and Z. Shah, (2023) "Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review," BMC Medical Imaging, vol. 23, no. 1, p. 129. DOI: 10.1186/s12880-023-01098-z
2. A. Fanizzi, F. Fadda, M. C. Comes, S. Bove, A. Catino, E. Di Benedetto, A. Milella, M. Montrone, A. Nardone, C. Soranno, *et al.*, (2023) "Comparison between vision transformers and convolutional neural networks to predict non-small lung cancer recurrence," Scientific Reports, vol. 13, no. 1, p. 20605.
https://www.nature.com/articles/s41598-023-48004-9
3. L. Gai, M. Xing, W. Chen, Y. Zhang, and X. Qiao, (2024) "Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective," Multimedia Tools and Applications, vol. 83, no. 20, pp. 59253–59269. DOI: 10.1007/s11042-023-17644-4
4. L. Liu, Z. Liu, J. Xie, H. Zhang, J. Chang, and H. Qiao, (2024) "Transformers in pathological image analysis: A survey," Available at SSRN 4982382. DOI: https://doi.org/10.2139/ssrn.4982382.
5. H. Guo, L. Jiang, F. Zheng, Y. Liang, S. Bao, X. Zhang, Q. Zhang, J. Tang, and R. Li, (2024) "Vprotonet: vision transformer-driven prototypical networks for enhanced interpretability in chest x-ray diagnostics," in International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023), vol. 13105, pp. 804–809, SPIE. https://doi.org/10.1117/12.3026314
6. A. Satsangi, K. Srinivas, and A. C. Kumari, (2025) "Enhancing lung cancer diagnostic accuracy and reliability with LCDVIT: an expressly developed vision transformer model featuring explainable AI," Multimedia Tools and Applications, pp. 1–41.

7. Y. Chen, J. Feng, J. Liu, B. Pang, D. Cao, and C. Li, (2022) "Detection and classification of lung cancer cells using swin transformer," Journal of Cancer Therapy, vol. 13, no. 7, pp. 464–475. DOI: 10.4236/jct.2022.137041
8. A. Kumar, R. Mehta, B. R. Reddy, and K. K. Singh, (2024) "Vision transformer based effective model for early detection and classification of lung cancer," SN Computer Science, vol. 5, no. 7, p. 839. https://doi.org/10.1007/s42979-024-03120-9
9. F.-M. Guo and Y. Fan, (2022) "Zero-shot and few-shot learning for lung cancer multi-label classification using vision transformer," arXiv preprint arXiv:2205.15290. DOI:10.48550/arXiv.2205.15290
10. N. Malaviya, M. Rahevar, A. Virani, A. Ganatra, and K. Bhuva, (2023) "Lvit: Vision transformer for lung cancer detection," in 2023 International Conference on Artificial Intelligence and Smart Communication (AISC), pp. 93–98, IEEE. https://ieeexplore.ieee.org/document/10085230
11. M. Imran, B. Haq, E. Elbasi, A. E. Topcu, and W. Shao, (2024) "Transformer based hierarchical model for non-small cell lung cancer detection and classification," IEEE Access. https://ieeexplore.ieee.org/document/10643966
12. R. Durgam, B. Panduri, V. Balaji, A. O. Khadidos, A. O. Khadidos, and S. Selvarajan, (2025) "Enhancing lung cancer detection through integrated deep learning and transformer models," Scientific Reports, vol. 15, no. 1, p. 15614.
13. E. Matsuyama, H. Watanabe, and N. Takahashi, (2024) "Performance comparison of vision transformer-and CNN-based image classification using cross entropy: A preliminary application to lung cancer discrimination from CT images," Journal of Biomedical Science and Engineering, vol. 17, no. 9, pp. 157–170.
14. L. Weng, Y. Xu, Y. Chen, C. Chen, Q. Qian, J. Pan, and H. Su, (2024) "Using vision transformer for high robustness and generalization in predicting EGFR mutation status in lung adenocarcinoma," Clinical and Translational Oncology, vol. 26, no. 6, pp. 1438–1445. doi: 10.1007/s12094-023-03366-4
15. H. F. Kareem, M. S. AL-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S.Hassan, (2021) "Evaluation of SVM performance in the detection of lung cancer in marked CT scan dataset," Indonesian Journal of Electrical Engineering and Computer Science, vol. 21, no. 3, pp.1731-1738, p. 1731. DOI: http://doi.org/10.11591/ijeecs.v21.i3.pp1731-1738

Author

Harshatej Simhadri is currently a junior at Westford Academy, Westford, Massachusetts, USA. He is interested in medicine and AI. He filed a USPTO patent on ML-based heart disease detection, trained at Harvard Medical School, founded a nonprofit SmartQuad Corporation, earned Presidential Volunteer Gold Awards, and excels in HOSA, research, tutoring, and hospital volunteering.

DOI: 10.36838/IJHSR816.64