Scientific Research Publishing

Open Journal of Medical Imaging, 2025, 15(2), 57-72
https://www.scirp.org/journal/ojmi
ISSN Online: 2164-2796
ISSN Print: 2164-2788

A Wavelet-Based Two-Stage Vision Transformer Model for Histological Subtypes Classification of Lung Cancers on CT Images

Eri Matsuyama[1], Haruyuki Watanabe[2], Noriyuki Takahashi[3]

[1]Faculty of Informatics, The University of Fukuchiyama, Kyoto, Japan
[2]School of Radiological Technology, Gunma Prefectural College of Health Sciences, Gunma, Japan
[3]School of Health Sciences, Fukushima Medical University, Fukushima, Japan
Email: matsuyama-eri@fukuchiyama.ac.jp

How to cite this paper: Matsuyama, E., Watanabe, H. and Takahashi, N. (2025) A Wavelet-Based Two-Stage Vision Transformer Model for Histological Subtypes Classification of Lung Cancers on CT Images. *Open Journal of Medical Imaging*, **15**, 57-72.
https://doi.org/10.4236/ojmi.2025.152005

Received: March 17, 2025
Accepted: April 19, 2025
Published: April 22, 2025

Abstract

Accurate histological classification of lung cancer in CT images is essential for diagnosis and treatment planning. In this study, we propose a vision transformer (ViT) model with two-stage fine-tuning using wavelet transformation to improve classification performance. In the first stage, feature extraction is enhanced using wavelet-transformed images, and in the second stage, the model is fine-tuned with the original CT images. This method improves classification accuracy and enhances model robustness. Experimental results show that the proposed method outperforms conventional ViT and CNN fine-tuning methods. It achieves a classification accuracy of 0.971, surpassing the 0.953 obtained with conventional ViT fine-tuning and 0.945 with ResNet50 fine-tuning. Moreover, the proposed method reduces classification uncertainty, with particularly significant improvements in the classification of large cell lung carcinoma. These results demonstrate the effectiveness of incorporating wavelet-based feature extraction into ViT fine-tuning for lung cancer classification. Future research will focus on developing optimization techniques, applying the method to multimodal medical imaging, and integrating explainable AI technologies to further improve its applicability in clinical settings.

Keywords

Lung Cancer, Vision Transformer, Fine-Tuning, Wavelet Transformation, Medical Imaging

1. Introduction

Lung cancer remains one of the leading causes of cancer-related mortality world-

wide, with approximately 2.5 million new cases diagnosed in 2022 [1] [2]. It is broadly classified into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). The latter includes subtypes such as lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and large cell carcinoma (LULC), which are classified based on cytological characteristics [3] [4]. Early identification of histological subtypes is crucial for effective treatment and improved prognosis. Although chest X-rays and CT scans are considered the gold standard in health screenings and examinations, they have limitations in histological differentiation, making accurate diagnosis challenging.

The advent of vision transformers (ViTs) has significantly influenced computer vision, leveraging the self-attention mechanism from natural language processing. Since their introduction, ViTs have achieved state-of-the-art performance in image recognition, object detection, segmentation, and classification, surpassing convolutional neural networks (CNNs) in several benchmark tasks [5]-[9]. Their ability to capture long-range dependencies enables a more comprehensive understanding of spatial relationships, which is particularly beneficial for medical imaging applications, including lung cancer detection [3] [10].

ViT-based methods have been widely applied in medical image analysis, including cancer classification, tumor segmentation, nodule detection, and survival prediction [9]-[14]. While numerous studies have demonstrated promising performance in lung cancer classification, several challenges remain. One of the primary limitations is the reliance on positional embedding and the lack of locality in self-attention. Since ViT encodes spatial information using positional embeddings, it may be less effective in capturing fine-grained local features compared to CNNs.

Moreover, although self-attention considers relationships across all patches, it exhibits a limited capacity to capture local patterns, particularly in the early layers. These limitations significantly affect the detection of small abnormalities, such as pulmonary nodules in lung CT scans. Therefore, further improvements in ViT-based approaches are necessary to enhance diagnostic accuracy.

Wavelet-based ViT models have recently emerged as a promising solution for improving local feature extraction [15]-[19]. Unlike conventional ViT models, these methods leverage wavelet transformation to decompose images into multiple frequency components, capturing both global (low-frequency) and local (high-frequency) features. This enables better differentiation of fine-grained structures, such as tumor textures and subtle intensity variations, while preserving the overall structural integrity of the image. For example, LUAD typically exhibits fine glandular structures, whereas LUSC tends to have coarser tissue patterns. To effectively capture these morphological differences, leveraging multi-scale information is essential, making wavelet transformation a well-suited approach for this task.

This study proposes a wavelet-based two-stage fine-tuning ViT model that integrates feature extraction using wavelet transform and applies stepwise fine-tuning to enhance adaptability to lung CT images. In our previous study [9], ViT models pretrained on ImageNet 2012 were fine-tuned for four-category classification of LUAD, LUSC, LULC, and normal cases; however, challenges persist in

improving classification accuracy. In particular, pretraining on small-scale datasets resulted in insufficient utilization of local information, limiting classification performance improvement. ViT models are typically pretrained on large-scale datasets such as ImageNet-21k (14 million images) and JFT-300M (300 million images), enabling high accuracy even with a limited number of samples during fine-tuning. However, these large-scale datasets are not publicly available.

This study aims to achieve improved classification performance while maintaining the same scale of pretraining and fine-tuning datasets as in our previous study. To this end, we propose a two-stage fine-tuning method incorporating wavelet transform, which effectively leverages local information even in small-scale datasets.

Our method consists of two fine-tuning stages: Stage 1: The model is fine-tuned using wavelet-transformed images to enhance feature learning. Stage 2: The fine-tuned model is further trained on the original CT images to refine the learned representations.

To evaluate model performance, we employ cross-entropy as an additional performance evaluation metric, along with standard classification metrics, to quantify uncertainty in image classifiers.

The key contributions of this study are as follows:

1) We propose a novel wavelet-based two-stage fine-tuning ViT model that combines wavelet transformation with progressive transfer learning. This approach enhances classification accuracy and generalization by leveraging multi-scale feature extraction and adaptive learning.

2) We use cross-entropy not only as a cost function in CNN training but also as a performance evaluation metric to quantify classifier uncertainty.

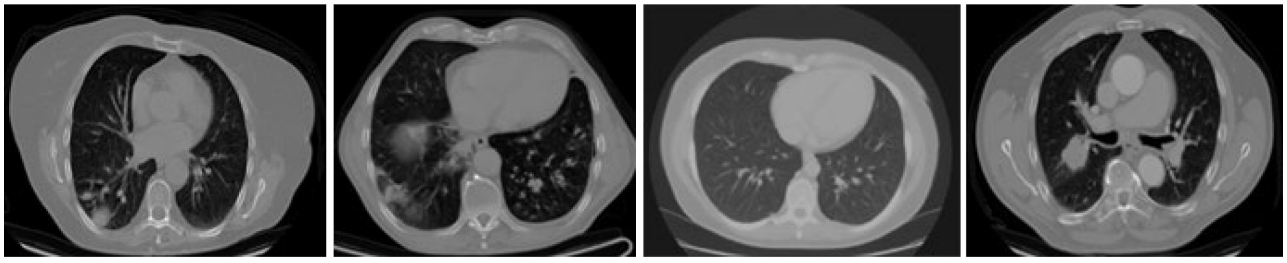
3) We demonstrate that the proposed model achieves higher classification accuracy and robustness than conventional ViT-based and CNN-based fine-tuning models, as evaluated on the same pretraining and lung CT datasets used in our previous study.

2. Materials and Methods

This study proposes a method to improve histological classification performance of lung cancer in CT images during pretraining with a small-scale dataset and fine-tuning with a limited amount of data.

2.1. Dataset

The dataset used in this study consists of lung CT images publicly available on Kaggle [20] and intended for non-profit research purposes. Therefore, this study does not raise ethical concerns, and informed consent is not required. The dataset is categorized into four classes: LUAD, LULC, LUSC, and normal lung tissue. An example is shown in **Figure 1**. For the experiments, each class contained 187 images, resulting in a total of 748 images. A 10-fold cross-validation approach was applied, with 90% of the data used for training and the remaining 10% for validation.



(a) (b) (c) (d)

Figure 1. An example of image data: (a) LUAD (lung adenocarcinoma); (b) LUSC (lung squamous cell carcinoma); (c) Normal (healthy lung); (d) LULC (large cell carcinoma).

2.2. Proposed Approach

2.2.1. The 2D Discrete Wavelet Transform

Wavelet transform is a mathematical technique for analyzing data across multiple scales. The discrete wavelet transform (DWT) applies wavelet decomposition in a discrete manner, making it particularly effective for image processing and compression. In medical imaging, the two-dimensional discrete wavelet transform (2D-DWT) is widely used for data compression, image enhancement, and noise reduction [21] [22].

The 2D-DWT starts at decomposition level 0 and, at level 1, decomposes the image into four frequency sub-bands: a low-frequency component (LL) and three high-frequency components—low-high (LH), high-low (HL), and high-high (HH). The LL component provides a smoothed approximation of the image, while the high-frequency components capture structural details. At higher decomposition levels, further transformations are applied exclusively to the LL component, progressively reducing resolution while preserving key multi-scale features. This hierarchical process efficiently represents complex image structures while retaining crucial details. The overview of 2D-DWT decomposition is shown in **Figure 2**.

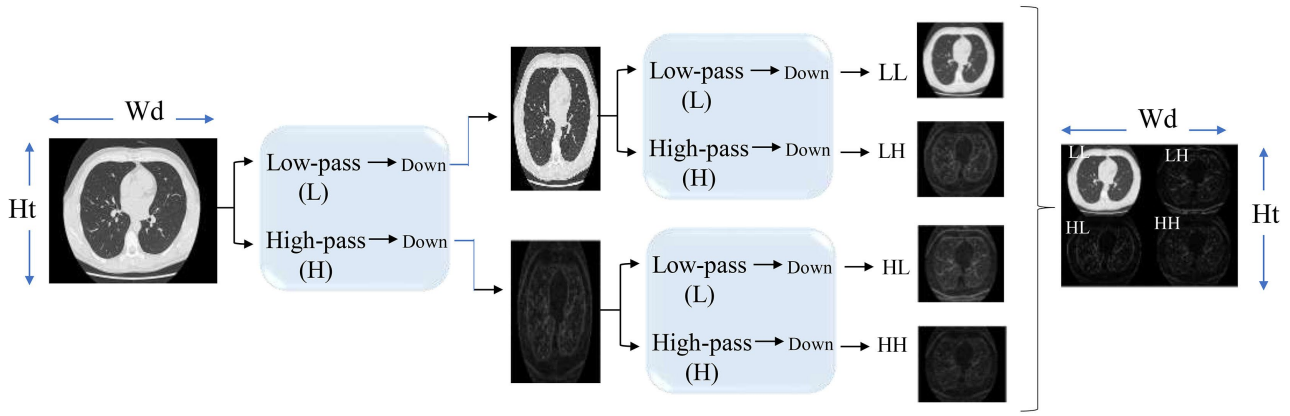


Figure 2. Level-1, two-dimensional DWT. A one-dimensional DWT is applied along the rows of the image, followed by another one-dimensional DWT along the columns. The four generated components (LL, LH, HL, and HH) are arranged to maintain the original image size.

DWT employs various wavelet basis functions for decomposition, including

Haar, Daubechies, Coiflet, and Meyer wavelets. In this study, the Daubechies wavelet of order 2 (db2) was chosen for its optimal balance between computational efficiency and feature preservation. Additional details on 2D-DWT can be found in the literature [23]-[25].

2.2.2. Two-Stage Fine-Tuning Strategy

To adapt the ViT model for lung cancer classification, this study employs a two-stage fine-tuning strategy. This approach consists of the following three steps:

Step 1: Pretraining Using ImageNet

First, the ViT model undergoes pretraining on the ImageNet 2012 dataset, which consists of approximately 1.3 million natural images with a resolution of 384 × 384 for a 1000-class classification task. This pretraining method is widely used and learns general feature representations. However, at this stage, the model does not incorporate features specific to CT images or lung cancer.

Step 2: Fine-Tuning Using Wavelet-Transformed CT Images

In this step, wavelet decomposition is applied to lung CT images, and the resulting wavelet coefficients (Level 1) are used as input to fine-tune the pretrained ViT model.

ViT models are known to have weak inductive biases and tend to prioritize extracting low-frequency components over high-frequency components [26] [27]. This step leverages the self-attention mechanism of ViT, which accounts for correlations across all features, while enhancing its ability to capture the relationships between high- and low-frequency components—an aspect where ViT typically underperforms. By incorporating medical image-specific features and utilizing multi-scale decomposition, this approach enables the model to capture texture variations more effectively compared to using raw CT images for training alone.

During this step, the multi-layer perceptron (MLP) head of the pretrained ViT model is removed and replaced with a feedforward layer tailored to the number of target classes. Additionally, since this stage involves processing images with a different resolution from that used during pretraining, the original patch embedding and position embedding are not suitable for direct use. Consequently, all embeddings are retrained, and the linear transformation layer preceding the scaled dot-product attention in the transformer encoder is fine-tuned as well. This process can be regarded as a form of deep fine-tuning.

Step 3: Fine-Tuning Using Original CT Images

In this final step, the ViT model, which has already learned wavelet-based features, undergoes additional fine-tuning using original CT images. This process aims to adapt the model to high-resolution CT data while balancing texture-based and structural feature learning. As in the previous step, embeddings are updated, and the linear transformation layer preceding the scaled dot-product attention in the encoder is retrained.

This two-stage fine-tuning approach first emphasizes texture enhancement through wavelet feature learning in the first stage and subsequently refines the global representations in the second stage using original CT images.

2.2.3. Overall Network Framework

Figure 3 illustrates the overall network framework of the proposed method. This approach utilizes a Base-sized ViT model consisting of 12 stacked transformer encoder blocks, each with 12 attention heads. The model is pre-trained on the ImageNet 2012 dataset. Although the standard ImageNet 2012 dataset uses an image resolution of 224 × 224 pixels, this study adopts high-resolution training at 384 × 384 pixels to capture finer details.

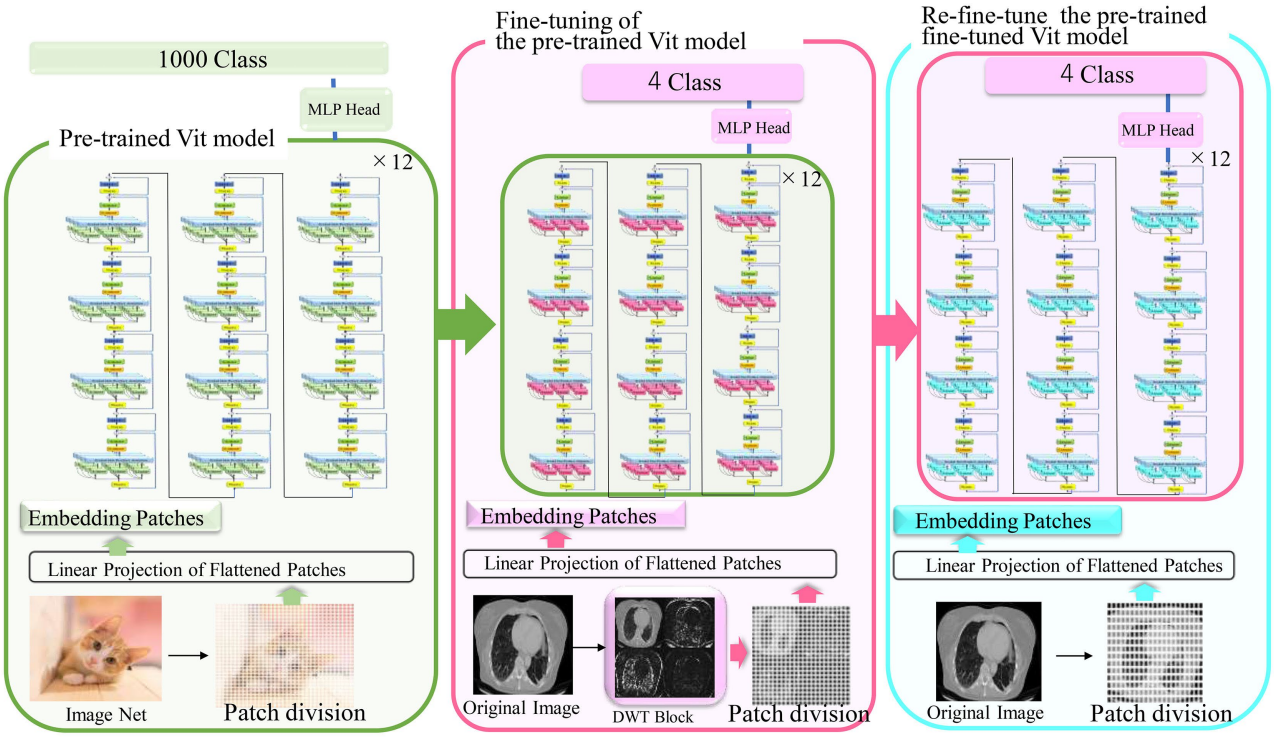


Figure 3. Overall network framework of the proposed two-stage fine-tuning method. From left to right: Pretraining on ImageNet, first-stage fine-tuning using wavelet coefficients, and second-stage fine-tuning using raw images.

The training data are divided into 16 × 16 patches, which are then transformed into 768-dimensional patch embeddings. A class token is then prepended, and position embeddings are added to each patch before they are fed into the transformer encoder. After initial training, the existing MLP head is replaced with a new one, and a novel feature map generation mechanism is integrated into the first transformer encoder block. Specifically, a DWT block is added before patch partitioning. This block applies a level-1 DWT to the image, producing four wavelet coefficient components (LL, LH, HL, and HH) that form a 384 × 384 × 3 array. The array is then partitioned into patches, and after initializing the pre-trained embeddings, the model undergoes additional training.

During retraining, the 36 linear layers preceding the 12 scaled dot-product attention mechanisms in each block (totaling 432 layers) are updated. For training, Adam is used as the optimizer, with a learning rate of 0.0001, a mini-batch size of 12, and a maximum of 50 epochs. In other words, this process corresponds to the fine-tuning of the pre-trained ViT model.

For retraining, 10-fold cross-validation is conducted to generate ten subset-trained models. The model with the highest accuracy is selected, after which the DWT block is removed and raw CT images are fed directly into the model. The model is then retrained under the same learning conditions, updating the embedding layer and the 432 linear layers preceding the scaled dot-product attention mechanism.

2.3. Advantages of the Proposed Two-Stage Fine-Tuning Approach

1) Enhanced Feature Representation and Inductive Bias Reinforcement in ViT with Wavelet Transform:

Wavelet decomposition captures features in both spatial and frequency domains, making it easier for ViT to learn the fine structures of lung cancer tissues. Compared to raw CT images, wavelet-transformed images emphasize texture variations and provide critical information for classifying lung cancer subtypes. ViT has a weaker inductive bias than CNNs and struggles with learning local information effectively. However, by using DWT to decompose images into LL, LH, HL, and HH components and arranging them to match the original image size, local information is naturally embedded within patches, allowing ViT to learn both local patterns and long-range relationships more effectively. Specifically, LL preserves the coarse structure of the image, aiding in the learning of long-range dependencies, while LH, HL, and HH retain edge and local features, compensating for ViT's weakness in capturing local information. In this way, ViT can learn both global structures and local details.

2) Improved Domain Adaptation from ImageNet to Medical Imaging:

The wavelet-based fine-tuning phase acts as a transition between ImageNet pretraining and lung CT classification, allowing the model to gradually adapt to the medical imaging domain. This progressive learning approach enhances generalization compared to directly fine-tuning on CT images.

3) More Effective Feature Learning:

In conventional fine-tuning, only the final fully connected layer is primarily updated. In contrast, our method retrains the linear transformations in the encoder to refine the embedding representations. This approach enables the optimization of feature representations tailored to the medical imaging domain while retaining useful prior knowledge learned from ImageNet.

4) Higher Classification Accuracy:

The wavelet-based fine-tuning stage enhances texture feature learning, while subsequent fine-tuning on original CT images refines global structural representations. This balanced approach improves classification accuracy and reduces catastrophic forgetting during the transition from ImageNet pretraining to lung CT image classification.

5) Greater Generalization and Robustness:

Fine-tuning directly on small medical datasets often leads to overfitting. Introducing an intermediate wavelet-based stage enhances generalization to unseen data while normalizing variations in contrast and texture, making the model more robust to differences in CT image quality.

6) Reduction of Computational Cost:

DWT decomposition reduces data redundancy and improves the efficiency of feature extraction while compressing information. As a result, it reduces the computational burden of ViT while also improving accuracy.

2.4. Performance Measurement

In this study, in addition to standard performance metrics, we use cross-entropy to assess uncertainty in image classifiers.

2.4.1. Standard Metrics

A confusion matrix is essential for performance evaluation, as it serves as the basis for calculating standard metrics [28] [29]. It comprises four possible outcomes: true positive (TP), false negative (FN), false positive (FP), and true negative (TN). In this study, the following standard metrics were used: accuracy, precision, recall, specificity, and F1-score.

Accuracy, a widely used metric, accounts for all values in the confusion matrix, measuring the proportion of correctly classified instances among all cases (TP, FN, FP, and TN). Precision evaluates the proportion of TP cases among all predicted positive instances. Recall, also called sensitivity, assesses the proportion of TP cases among all actual positive instances. Specificity measures the proportion of correctly classified negative cases. The F1-score, the harmonic mean of precision and recall, provides a balanced measure of both metrics.

2.4.2. Cross Entropy

We use cross-entropy as a performance metric to quantify classifier uncertainty.

Cross-entropy measures the difference between the two probability distributions [30]-[32]. In machine learning and deep learning, it is commonly used to assess how closely a model's predicted probabilities align with the true probabilities. Essentially, cross-entropy quantifies the discrepancy between the two distributions.

The cross-entropy between two distributions is defined mathematically as follows:

$$H(p,q) = -\sum_x p(x)\log_e q(x) \tag{1}$$

where p represents the true distribution, q denotes the predicted distribution, and x sums over all possible outcomes.

Cross-entropy also quantifies the information loss when approximating the true distribution using the predicted one. It is particularly useful for evaluating classification models that output probabilities in the range [0, 1]. In simple terms, a lower cross-entropy value indicates a closer match between the predicted and true distributions, while a higher value signifies greater divergence.

As a performance metric, cross-entropy enables model comparison by comparing entropy values. A lower cross-entropy suggests greater confidence in predictions, often correlating with higher accuracy, whereas a higher value indicates higher uncertainty and lower accuracy. A numerical example illustrating the use

of cross-entropy in multi-class classification is provided in [33].

3. Results and Discussion

In this study, we proposed a two-stage fine-tuning method and evaluated its performance in classifying lung cancer histological subtypes using CT images. The proposed method is designed for a four-class classification task, distinguishing three lung cancer histological subtypes (LUAD, LULC, and LUSC) and normal lung tissue. In the experiments, we used the same number of training samples as in a previous study [9] and assessed classification performance using five metrics: accuracy, precision, recall, F1-score, and specificity. Cross-entropy was used to evaluate uncertainty.

Tables 1-3 present the performance evaluation results for three different fine-tuning approaches: first-stage fine-tuning only (Table 1), the proposed two-stage fine-tuning method (Table 2), and conventional fine-tuning (Table 3). A comparison of Tables 1-3 indicates that the proposed method (Table 2) achieved better performance than the other approaches across all metrics (mean values).

Table 1. Evaluation results when only first-stage fine-tuning is applied (average values of 10-fold cross-validation).

Category	Precision	Recall	F1	Specificity
LUAD	0.926	0.866	0.895	0.977
LULC	0.957	0.941	0.949	0.986
Normal	0.964	0.995	0.979	0.988
LUSC	0.918	0.963	0.940	0.971
Average	0.941	0.941	0.941	0.980

Table 2. Evaluation results when second-stage fine-tuning is applied following first-stage fine-tuning (proposed method) (average values of 10-fold cross-validation).

Category	Precision	Recall	F1	Specificity
LUAD	0.962	0.947	0.954	0.988
LULC	0.973	0.952	0.962	0.991
Normal	0.984	1.000	0.992	0.995
LUSC	0.963	0.984	0.974	0.988
Average	0.971	0.971	0.970	0.990

Table 3. Evaluation results of the conventional fine-tuning in our previous study using original image training (average values of 10-fold cross-validation).

Category	Precision	Recall	F1	Specificity
LUAD	0.921	0.941	0.931	0.970
LULC	0.959	0.907	0.932	0.991
Normal	0.995	1.000	0.997	0.998
LUSC	0.947	0.947	0.947	0.980
Average	0.955	0.949	0.952	0.985

On the other hand, in the first-stage fine-tuning (Table 1), the metric values were slightly lower than those of conventional fine-tuning (Table 3). This may be because the pre-trained ViT model is optimized for pixel-level information and thus may not effectively capture feature representations from wavelet-transformed images. However, the improvement in the metrics after the second-stage fine-tuning suggests that wavelet-based features learned in the first stage emphasized local information and facilitated the ViT's ability to capture global relationships. In other words, this suggests that the proposed method reinforces the inductive bias of the ViT.

Furthermore, when comparing the classification metrics for normal images (indicated as "Normal" in the tables), conventional fine-tuning (Table 3) achieved slightly higher scores than the proposed method (Table 2). This result suggests that normal images, which have simpler and more consistent structures compared to lesion images, can be classified with high accuracy even without the additional inductive bias introduced by the proposed method. Specifically, normal tissue in CT images tends to have a relatively homogeneous pixel distribution, allowing the model to distinguish it based on simple features such as brightness, shape, and texture. Therefore, biases introduced by model design and training methods are likely to have a smaller impact on the classification of normal images.

Tables 4-6 present the accuracy and cross-entropy for each subset in the 10-fold cross-validation. Table 4 shows the results of first-stage fine-tuning, where the ViT model is fine-tuned using wavelet coefficients, representing the intermediate stage of the proposed algorithm. Among the subsets in Table 4, the model with the highest accuracy (subset No. 4) was selected for fine-tuning on the original CT image set. The final results of the proposed method, applying second-stage fine-tuning following first-stage fine-tuning, are presented in Table 5. Table 6 shows the results of conventional fine-tuning, where the pre-trained ViT model was fine-tuned using original lung CT images.

The final column of the first row in Table 4 and Table 5 shows the mean accuracy for first-stage fine-tuning and second-stage fine-tuning, which are 0.941 and 0.971, respectively. A statistically significant difference was observed between the two ($P < 0.05$). This result suggests the effectiveness of the proposed two-stage fine-tuning method. In contrast, our previous study reported that the accuracy of a fine-tuned ResNet50 model using original CT images was 0.945 [9]. Additionally, the accuracy of the pre-trained ViT model in this study was 0.953 (final column of the first row in Table 6), showing no significant difference compared to ResNet50 ($P = 0.26$). This finding supports the known limitation that the ViT model does not fully demonstrate its potential when pre-trained with a small dataset. Despite the use of a small-scale pre-training dataset, as in the previous study, the proposed method achieved an accuracy of 0.971, demonstrating a statistically significant improvement over ResNet50 (accuracy: 0.945, $P < 0.05$). This result suggests that the proposed method enables a ViT model pre-trained on a small dataset to outperform a high-performance CNN model (ResNet50) in classification accuracy.

Table 4. Cross entropy and accuracy when only first-stage fine-tuning is applied.

	Accuracy	0.960	0.893	0.920	0.987	0.960	0.920	0.947	0.960	0.946	0.919	0.941
	Subset No.	1	2	3	4	5	6	7	8	9	10	Average
Cross entropy	LUAD	0.509	1.000	0.119	0.131	0.458	0.613	0.327	0.277	0.519	0.502	0.429
	LULC	0.050	0.476	0.652	0.021	0.003	0.053	0.076	0.263	0.093	0.317	0.204
	Normal	0.000	0.021	0.197	0.003	0.028	0.032	0.006	0.008	0.000	0.000	0.028
	LUSC	0.047	0.336	0.103	0.000	0.017	0.244	0.593	0.181	0.086	0.158	0.179
	Average	0.152	0.458	0.268	0.039	0.127	0.235	0.251	0.180	0.174	0.244	0.209

Table 5. Cross entropy and accuracy when second-stage fine-tuning is applied following first-stage fine-tuning. The results were obtained from fine-tuning on the original images using the Subset No. 4 model from Table 4.

	Accuracy	0.987	0.947	0.987	0.973	1.000	0.960	0.920	0.987	0.973	0.973	0.971
	Subset No.	1	2	3	4	5	6	7	8	9	10	Average
Cross entropy	LUAD	0.192	0.546	0.017	0.058	0.024	0.235	0.036	0.002	0.061	0.129	0.156
	LULC	0.030	0.461	0.105	0.041	0.001	0.000	0.222	0.062	0.130	0.070	0.114
	Normal	0.013	0.003	0.027	0.000	0.039	0.002	0.004	0.000	0.001	0.001	0.008
	LUSC	0.002	0.012	0.000	0.003	0.010	0.211	0.369	0.034	0.014	0.042	0.072
	Average	0.059	0.256	0.037	0.026	0.019	0.112	0.238	0.024	0.051	0.060	0.087

Table 6. Cross entropy and accuracy for the conventional fine-tuning in our previous study using original image training.

	Accuracy	0.947	0.960	0.947	0.974	0.974	0.974	0.960	0.974	0.946	0.879	0.953
	Subset No.	1	2	3	4	5	6	7	8	9	10	Average
Cross entropy	LUAD	0.418	0.018	0.113	0.181	0.373	0.137	0.201	0.048	0.303	0.410	0.22
	LULC	0.143	0.511	0.776	0.307	0.051	0.062	0.216	0.031	0.378	0.195	0.267
	Normal	0.028	0.002	0.16	0.000	0.145	0.026	0.000	0.000	0.000	0.197	0.056
	LUSC	0.275	0.024	0.004	0.008	0.037	0.240	0.224	0.414	0.088	0.155	0.147
	Average	0.216	0.139	0.263	0.124	0.152	0.117	0.160	0.123	0.192	0.240	0.172

Rows 3-6 in Tables 4-6 present the uncertainty (cross-entropy) for each disease category across different subsets, whereas the final row shows the mean cross-entropy value. A cross-entropy value of 2.0 indicates that the model is entirely unable to classify the four categories (*i.e.*, it has extremely low confidence in its predictions). A cross-entropy value of 1.0 suggests that the model cannot distinguish between two of the four categories and assigns similar confidence levels to them. When the cross-entropy value is 0.5, the model has slightly higher confidence in one category while maintaining lower confidence in the other three. A cross-entropy value of 0.2 indicates high confidence in one category, while a value of 0 implies absolute certainty in one category, meaning there is no uncertainty in the model's prediction.

From the results in Tables 4-6, it was confirmed that cross-entropy values var-

ied even when accuracy remained the same. This suggests that models with identical classification accuracy can exhibit different levels of uncertainty in their predictions. For instance, subsets with an accuracy of 0.960 include subset No. 1, 5, and 8 in Table 4, subset No. 6 in Table 5, and subset No. 2 and 7 in Table 6. Among these, subset No. 6 in Table 5 had the lowest mean cross-entropy value, at 0.112 ($P < 0.05$). This result indicates that even when accuracy is the same, the level of uncertainty differs, and the predictions made by the model trained with two-stage fine-tuning are closer to the target distribution—meaning they are more precise and exhibit lower uncertainty—compared to other models. Conversely, subset No. 5 in Table 5 achieved an accuracy of 1.00, yet its mean cross-entropy value was not zero. This implies that even when the model achieves 100% accuracy, it still retains a certain level of uncertainty. This finding suggests that evaluating model uncertainty can help identify models with higher reliability.

When evaluating each disease category, the proposed method exhibited lower cross-entropy values across all categories than conventional fine-tuning. For example, in conventional fine-tuning, the category with the highest uncertainty was LULC, with a mean cross-entropy value of 0.267 (final column of Table 6). In contrast, the mean cross-entropy value for LULC using the proposed method was 0.114 (final column of the fourth row in Table 5), which was statistically significantly lower ($P < 0.05$). These results demonstrate the superiority of the proposed method and suggest that in evaluating deep learning model performance, assessing uncertainty alongside conventional metrics provides a more comprehensive assessment of model reliability. In particular, cross-entropy as an evaluation metric facilitates a more holistic assessment of model performance.

This study has several limitations. First, the proposed two-stage fine-tuning method increases computational costs compared to conventional fine-tuning approaches. Deploying this model in resource-constrained environments may necessitate optimization techniques such as model pruning and quantization. Second, while the proposed method improves classification accuracy, the interpretability of ViT-based models remains a challenge. Future research should explore techniques for visualizing learned features and elucidating the decision-making process to facilitate clinical applications. Third, this study compared the proposed model with ViT and CNN-based ResNet50 models. In our next study, we will conduct a comprehensive comparison with more advanced architectures, including Swin Transformer, ConvNeXt, and other relevant hybrid models.

4. Conclusions

In this study, we proposed a ViT model with wavelet-based two-stage fine-tuning for histological classification of lung cancer using CT images. In the first stage of fine-tuning, wavelet-transformed images were utilized to enhance feature extraction. In the second stage, the model was further refined using the original CT images to improve learned representations. This approach integrates both global and local features, enhancing classification accuracy and model robustness.

Experimental results demonstrated that the proposed method outperformed conventional fine-tuning methods based on CNNs and ViTs, achieving a classification accuracy of 0.971. This significantly surpasses the accuracy of 0.953 achieved by conventional ViT fine-tuning and 0.945 achieved by ResNet50 fine-tuning. Moreover, the two-stage fine-tuning approach significantly reduced cross-entropy loss, indicating enhanced model confidence in classification decisions. Notably, the classification uncertainty associated with large cell carcinoma (LULC), which had been higher in previous methods, was effectively mitigated, leading to improved classification accuracy across all lung cancer subtypes. These findings suggest that incorporating wavelet-based feature extraction in the initial fine-tuning stage enhances the performance of ViT for lung cancer histological classification.

Furthermore, an evaluation of uncertainty revealed that models with similar accuracy levels exhibited variations in prediction confidence, emphasizing the importance of assessing uncertainty alongside conventional accuracy metrics. However, several challenges remain, including dataset size, computational cost, and model interpretability. While the proposed method achieved high classification accuracy even with a relatively small dataset, further validation on larger and more diverse datasets is necessary to establish its generalizability.

Future research will focus on optimizing fine-tuning strategies through the incorporation of adaptive learning rates and alternative feature extraction techniques. Additionally, we will explore extending this method to multimodal medical imaging data, such as PET-CT fusion images, to further enhance classification performance and clinical applicability. Furthermore, improving model interpretability through explainable AI techniques will be essential for integrating it into clinical practice.

The findings of this study indicate that the wavelet-based two-stage fine-tuning ViT model improves accuracy in automated lung cancer diagnosis and has the potential to enhance the reliability of histological classification in medical imaging.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2024) Global Cancer Burden Growing, Amidst Mounting Need for Services. https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services

[2] Luo, G., Zhang, Y., Rumgay, H., Morgan, E., Langselius, O., Vignat, J., *et al.* (2025) Estimated Worldwide Variation and Trends in Incidence of Lung Cancer by Histological Subtype in 2022 and over Time: A Population-Based Study. *The Lancet Respiratory Medicine*, **13**, 348-363. https://doi.org/10.1016/s2213-2600(24)00428-4

[3] Mannepalli, D., Kuan Tak, T., Bala Krishnan, S. and Sreenivas, V. (2025) GSC-DVIT:

A Vision Transformer Based Deep Learning Model for Lung Cancer Classification in CT Images. *Biomedical Signal Processing and Control*, **103**, Article ID: 107371. https://doi.org/10.1016/j.bspc.2024.107371

[4] Luna, H.G.C., Severino Imasa, M., Juat, N., Hernandez, K.V., May Sayo, T., Cristal-Luna, G., *et al.* (2023) Expression Landscapes in Non-Small Cell Lung Cancer Shaped by the Thyroid Transcription Factor 1. *Lung Cancer*, **176**, 121-131. https://doi.org/10.1016/j.lungcan.2022.12.015

[5] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., *et al.* (2024) Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review. *Medical Image Analysis*, **91**, Article ID: 103000. https://doi.org/10.1016/j.media.2023.103000

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al.* (2021) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929. https://arxiv.org/abs/2010.11929

[7] Zhu, X., Su, W., Lu, L., *et al.* (2020) Deformable DETR: Deformable Transformers for End-to-End Object Detection. arXiv: 2010.04159. https://arxiv.org/abs/2010.04159

[8] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., *et al.* (2021) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. arXiv: 2102.04306. https://doi.org/10.48550/arXiv.2102.04306

[9] Matsuyama, E., Watanabe, H. and Takahashi, N. (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*, **17**, 157-170. https://doi.org/10.4236/jbise.2024.179012

[10] Ali, H., Mohsen, F. and Shah, Z. (2023) Improving Diagnosis and Prognosis of Lung Cancer Using Vision Transformers: A Scoping Review. *BMC Medical Imaging*, **23**, Article No. 129. https://doi.org/10.1186/s12880-023-01098-z

[11] Kumar, A., Mehta, R., Reddy, B.R. and Singh, K.K. (2024) Vision Transformer Based Effective Model for Early Detection and Classification of Lung Cancer. *SN Computer Science*, **5**, Article No. 839. https://doi.org/10.1007/s42979-024-03120-9

[12] Martin, O.A. and Sanchez, J. (2025) Evaluation of Vision Transformers for Multi-Modal Image Classification: A Case Study on Brain, Lung, and Kidney Tumors. arXiv: 2502.05517v1. https://arxiv.org/html/2502.05517v1

[13] Xiong, Y., Du, B., Xu, Y., Deng, J., She, Y. and Chen, C. (2022) Pulmonary Nodule Classification with Multi-View Convolutional Vision Transformer. 2022 *International Joint Conference on Neural Networks* (*IJCNN*), Padua, 18-23 July 2022, 1-7. https://doi.org/10.1109/ijcnn55064.2022.9892716

[14] Yang, L., Li, B., Dong, T. and Wang, L. (2024) ViTR-SP: A CT-Based Vision Transformer Model for Prediction of Pneumonitis in Patients with Non-Small Cell Lung Cancer Who Received Thoracic Radiotherapy and Immunotherapy. *Journal of Clinical Oncology*, **42**, e20034-e20034. https://doi.org/10.1200/jco.2024.42.16_suppl.e20034

[15] He, C., Diao, Y., Ma, X., Yu, S., He, X., Mao, G., *et al.* (2024) A Vision Transformer Network with Wavelet-Based Features for Breast Ultrasound Classification. *Image Analysis and Stereology*, **43**, 185-194. https://doi.org/10.5566/ias.3116

[16] Ding, M., Qu, A., Zhong, H., Lai, Z., Xiao, S. and He, P. (2023) An Enhanced Vision Transformer with Wavelet Position Embedding for Histopathological Image Classification. *Pattern Recognition*, **140**, Article ID: 109532. https://doi.org/10.1016/j.patcog.2023.109532

[17] Yao, T., Pan, Y., Li, Y., Ngo, C. and Mei, T. (2022) Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 328-345. https://doi.org/10.1007/978-3-031-19806-9_19

[18] Wu, F., Wu, J., Shu, H., Carrault, G. and Senhadji, L. (2024) Spatial-Enhanced Multi-Level Wavelet Patching in Vision Transformers. *IEEE Signal Processing Letters*, **31**, 446-450. https://doi.org/10.1109/lsp.2024.3350811

[19] Yang, D. and Seo, S. (2023) Discrete Wavelet Transform Meets Transformer: Unleashing the Full Potential of the Transformer for Visual Recognition. *IEEE Access*, **11**, 102430-102443. https://doi.org/10.1109/access.2023.3316144

[20] Chest CT-Scan Images Dataset. https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images

[21] Chui, C.K. (1992) An Introduction to Wavelets. 2nd Edition, Academic Press.

[22] Daubechies, I. (1992) Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970104

[23] Mallat, S.G. (1989) A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 674-693. https://doi.org/10.1109/34.192463

[24] Shahbahrami, A. (2012) Algorithms and Architectures for 2D Discrete Wavelet Transform. *The Journal of Supercomputing*, **62**, 1045-1064. https://doi.org/10.1007/s11227-012-0790-x

[25] Abdulazeez1, A.M., Zeebaree, D.Q., Zebari, D.A., MustafaZebari, G., Adeen, I.M.N. (2020) The Applications of Discrete Wavelet Transform in Image Processing: A Review. *Journal of Soft Computing and Data Mining*, **1**, 31-43. https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/7215/3935

[26] Park, N. and Kim, S. (2022) How Do Vision Transformers Work? arXiv: 2202.06709. https://doi.org/10.48550/arXiv.2202.06709

[27] Bai, J., Yuan, L., Xia, S., Yan, S., Li, Z. and Liu, W. (2022) Improving Vision Transformers by Revisiting High-Frequency Components. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, 1-18. https://doi.org/10.1007/978-3-031-20053-3_1

[28] Powers, D.M. (2020) Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. arXiv: 2010.16061. https://doi.org/10.48550/arXiv.2010.16061

[29] Müller, D., Soto-Rey, I. and Kramer, F. (2022) Towards a Guideline for Evaluation Metrics in Medical Image Segmentation. *BMC Research Notes*, **15**, Article No. 210. https://doi.org/10.1186/s13104-022-06096-y

[30] Shan, B. and Fang, Y. (2020) A Cross Entropy Based Deep Neural Network Model for Road Extraction from Satellite Images. *Entropy*, **22**, Article 535. https://doi.org/10.3390/e22050535

[31] Mannor, S., Peleg, D. and Rubinstein, R. (2005) The Cross Entropy Method for Classification. *Proceedings of the* 22*nd international conference on Machine learning—ICML*'05, Bonn, 7-11 August 2005, 561-568. https://doi.org/10.1145/1102351.1102422

[32] Mao, A., Mohri, M. and Zhong, Y. (2023) Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *Proceedings of the* 40*th International Conference on Machine Learning*, Honolulu, 23-29 July 2023, 23803-23828. https://proceedings.mlr.press/v202/mao23b/mao23b.pdf

[33] Matsuyama, E., Nishiki, M., Takahashi, N. and Watanabe, H. (2024) Using Cross Entropy as a Performance Metric for Quantifying Uncertainty in DNN Image Classifiers: An Application to Classification of Lung Cancer on CT Images. *Journal of Biomedical Science and Engineering*, **17**, 1-12. https://doi.org/10.4236/jbise.2024.171001