LLM-Assisted Swin Transformer Framework for Enhanced Adenocarcinoma Lung Cancer Classification and Interpretability in Histopathology

[1]Soomro Sarwan, School of Software, Northwestern Polytechnical University, Xi'an, China
[2]Gul Sheeraz, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

---

Abstract - We developed a new framework that combines a Swin Transformer for image analysis with a LLMa2 to achieve high classification accuracy and provide textual explanations for its predictions. Our model classifies lung adenocarcinoma subtypes with 98.69% accuracy and a near-perfect AUC of 0.9997. It performs consistently well across all five cancer subtypes, and demonstrated robustness to class imbalance. We also found that using 20x magnification provides an optimal balance of diagnostic power and computational efficiency. Furthermore, the integrated LLM acts as an intelligent assistant, generating textual explanations of the AI's decisions by listing salient morphological features, and flagging low-confidence predictions for pathologist review. This combined approach gives clinicians a highly accurate and interpretable tool for histopathology.

***KeyWords*: Large Language Model , Swin Transformer, Lung Cancer Diagnosis, Histopathology, Resolution Selection, Multi-Resolution Analysis**

1.INTRODUCTION

In digital pathology, Transformers are particularly well-suited for analyzing the entire context of a histopathological image at once, modeling dependencies between distant nsformers directly solve, allowing them to model the morphological patterns characteristic of different cancer subtypes. Large language models (LLMs) have the potential to automatically extract clinical information, aid in diagnosis and treatment, and support full-cycle lung cancer care, according to a systematic review of 28 studies. However, bias control and data security limitations still exist [1].

Moreover, LLMs and vision-language models when combined, provide strong multimodal AI capabilities for diagnosis, prognosis, and image analysis in the treatment of lung cancer; however, ethical, legal, and validation issues restrict their clinical application[2]. However, LLMs are not currently approved for this sensitive task due to government regulations and patient privacy laws that differ across countries, hospitals, and demographics. Transformer-based analysis can use LLMs' predictive ability to generate descriptive text based on the image analysis. These models show promising venue for research in a controlled and secure manner without violating ethical, legal, or regulatory constraints, even though LLMs are currently not permitted as medical devices and cannot directly affect clinical care[3]. Medical image analysis relies heavily on resolution; deep learning models use patch-based processing, multi-resolution inputs, and super-resolution techniques to improve feature extraction, classification accuracy, and diagnostic reliability in ultrasound imaging and histopathology. Likewise, multi-resolution multiple-instance learning techniques in whole-slide histopathology utilize slide-level supervision to pinpoint diagnostically relevant areas, eliminating the need for pixel-level annotations and thereby improving grading accuracy and clinical reliability[4 5 6].

Therefore, we propose a framework that uses a Swin Transformer for histopathological classification and integrates an LLM in a post-hoc manner to enhance the interpretability and clinical utility of the predictions. While pathologists naturally choose the best magnifications, modern systems obtain multi-resolution whole slide images that require significant resources.

1.1 Related Work

Transformer architectures now provide a method for diagnosing lung cancer by using self-attention mechanisms to model global histological patterns across entire images [7]. Likewise, these models can detect subtle long-range dependencies that are potentially useful for cancer detection without the need for explicit segmentation. Talib et al. [8] proposed a framework that integrates transformers and CNNs. [8], that combines a CNN for tissue type classification with a TransSegNet for lesion segmentation via a vision transformer. Similarly, Srinivas et al. [9] introduced BoTNet, a hybrid architecture replacing spatial convolutions in the final three bottleneck blocks of ResNet with multi-head self-attention (MHSA). However, such a technique lacks LLM-aware assistance, and creates Bottleneck Transformer (BoT) blocks that preserve residual structure.

Similarly. Chen et al. [10] proposed Visformer, a hybrid architecture that systematically transitions from a Transformer (DeiT) to a CNN (ResNet). It integrates convolutional operations such as stage-wise down-sampling, Batch Normalization, and 3×3 local convolutions in early layers, retaining self-attention in later stages. Wang et al. [11] also proposed a hybrid CNN–Transformer (HCT) model for NSCLC N-staging and survival prediction from CT scans. The model integrates a 3D ResNet for local feature extraction and a Transformer

for global context modeling, concatenating outputs for final prediction. Shafi and Chinnappan [12] proposed a hybrid Transformer–CNN–LSTM model for lung disease segmentation and classification. This method features an Improved Transformer-based CNN(ITCNN) for segmentation and a hybrid LinkNet–Modified LSTM (L-MLSTM) for classification, combining texture, shape, and deep features. For semantic segmentation, Wu et al. [9] proposed Fully Transformer Networks (FTN), a transformer-only model that exploits a Pyramid Group Transformer encoder and a Feature Pyramid Transformer decoder, demonstrating that transformers can surpass hybrid CNN-transformer models. Similarly, Xie et al. [13] created SegFormer, an encoder-decoder model for image segmentation. To improve computational efficiency or accuracy, several modified architectures have been introduced. Liu et al [14] propose the Swin Transformer, a hierarchical vision transformer that computes self-attention within local windows and uses shifted windows for cross-window connections, providing linear complexity relative to image size. Chen et al [15] present CrossViT, a dual-branch vision transformer processing different patch sizes and fusing them via cross-attention for efficient multi-scale feature fusion. For data-efficient training, Touvron et al [12] developed DeiT, a training strategy for Vision Transformers that uses a distillation token, allowing the student ViT to learn from a convnet teacher without external data. In the domain of self-supervised learning, Li et al [16] tackle computational challenges through EsViT, a multi-stage architecture that employs local self-attention and a region-matching pre-training task.Furthermore, Wu et al [17] introduced Visual Transformers (VTs), a token-based image representation framework that replaces pixel arrays with semantic tokens, modeling contextual dependencies through transformers.Several models have been designed for histopathology. Zhao et al [18] present PKMT-Net, a transformer designed to emulate pathologist reasoning by combining multi-scale soft segmentation with cross-attention, achieving 0.9970 AUC for lung cancer subtyping. Yagappan et al. [19] developed gSC-DViT, a simplified Vision Transformer that achieves 99.69% accuracy with lower computational cost. Durgam et al. [20] introduce the CanNS framework, which includes a Swin-Transformer UNet (SwiNet) for segmentation. The application of transformers extends beyond imaging. Wang et al.[21] introduce MedAlbert, a transformer-based model for early lung cancer detection from sequential EHR data that represents patient care pathways as a language of medical codes. In genomics, Mahbub et al. [22] develop HEMERA, a human-explainable transformer predicting lung cancer risk from GWAS genotype data.

1.2 Domain & Magnification Analysis

For this study, we mainly use 143 de-identified H&E-stained WSIs of lung adenocarcinoma from BMIRDS, including DHMC_wsi_4.zip (Images 120–143). Slides were annotated for Lepidic, Acinar, Papillary, Micropapillary, and Solid patterns and scanned at 20x or 40x. For deep learning, we extracted 72,108 512×512 RGB patches at 20x with consistent WSI-class mapping using OpenSlide and libvips.

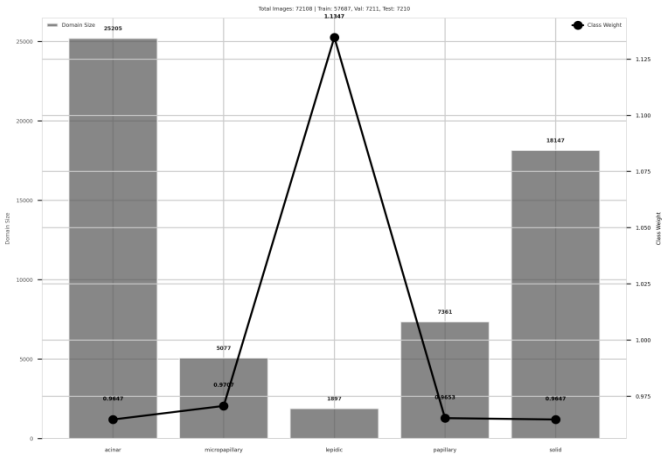

Fig-1 Class and Weight Distribution of Darmouth 20x Histopathology Slides

The distribution of images among the five subtypes of lung adenocarcinoma and the appropriate class weights used during training is shown in this figure 1. making the classification task more robust and clinically realistic. Figure 1 illustrates the inherent class imbalance, making it a genuine research venue for real diagnostic settings, where rare subtypes like Lepidic and Micropapillary occur infrequently. The data is vetted by three expert pathologists according to the major patterns: Lepidic, Acinar, Papillary, Micropapillary, and Solid. Although the WSIs were scanned at either 20x or 40x magnification, our extracted patch dataset includes only 72,108 512×512 pixel RGB patches annotated by WSI class. Consistent subtype trends were verified by cross-validation on the LungHist dataset, which contains 691 H&E-stained lung histology images from 45 patients collected at Hospital Clínico de Valladolid in 2023, captured at 20x and 40x magnification using Leica DM 2000 and ICC50 W microscopes. Images are classified into seven classes—well, moderately, and poorly differentiated adenocarcinoma (ACA_BD, ACA_MD, ACA_PD), squamous cell carcinoma (SCC_BD, SCC_MD, SCC_PD), and normal lung (NOR)—with patient-wise annotations and consistent 1200×1600 px resolution. Our analysis of magnification impact, based on Swin Transformer classification, found that 20x maintained diagnostic integrity while achieving greater average model confidence across the majority of subtypes with minimal dispersion. Using pretrained Swin Transformer models, 20x decreases computational overhead, speeds up processing, and maintains accuracy, according to statistical analysis and workflow metrics, proving a reliable, repeatable, and therapeutically effective framework.

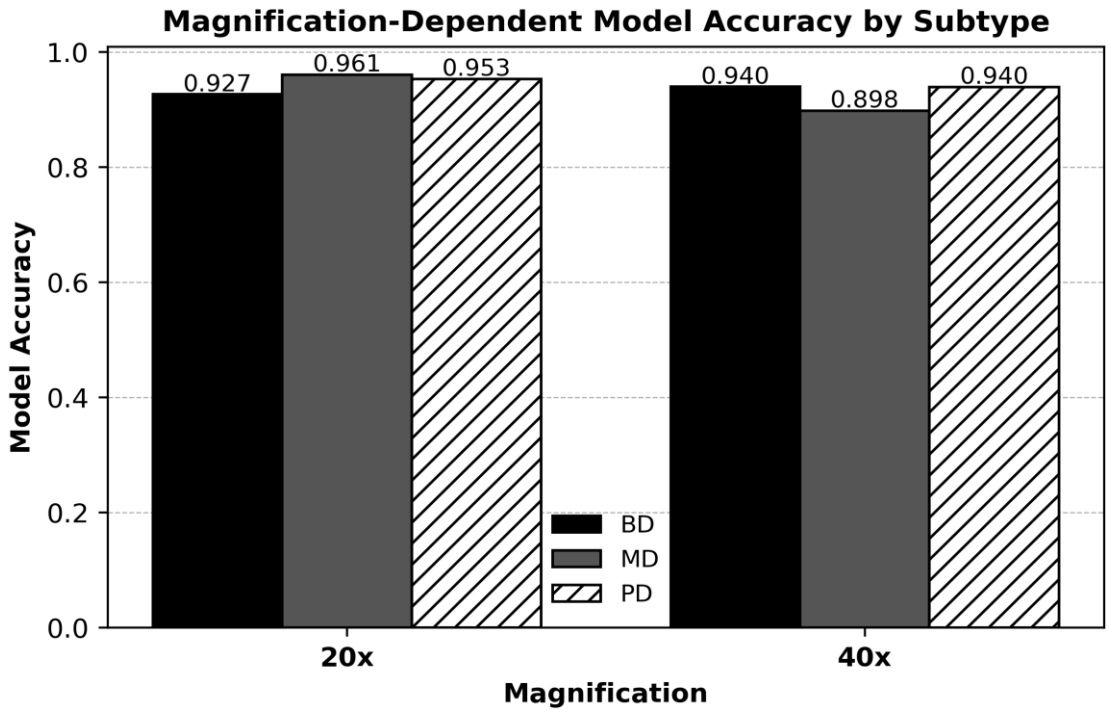

Fig-2 Cross-Domain Magnification Analysis((BD: Well-Differentiated, MD: Moderately Differentiated, PD: Poorly Differentiated))

According to the cross domain magnificaion analsysis of WSI slides, BD is somewhat superior at 40x magnification (0.9396 vs. 0.9266), whereas MD (0.9608 vs. 0.8983) and PD (0.9533 vs. 0.9395) subtypes attain higher or equivalent model accuracy at 20x magnification. Overall, 20x supports its effectiveness for deep learning applications by maintaining diagnostic performance with little loss. In summary, for the majority of lung adenocarcinoma subtypes, 20x magnification offers a fair trade-off between computational efficiency, model confidence, and diagnostic fidelity. 20x is generally adequate for dependable, repeatable, and effective deep learning-based histopathological analysis, even though 40x marginally helps

2. Methodology

We implemented a multi-scale tumor classification framework combining transformer-based architecture with a Swin Transformer for visual pattern recognition and a LlaMA-2 Large Language Model (LLM) for clinical reasoning and explanation. Our core classifier is built on a Swin Transformer (Swin-T) architecture, which is pre-trained on ImageNet. Its key innovation is a hierarchical structure with shifted windows, enabling efficient computation of self-attention across different scales to capture both fine-grained cellular details and broader tissue architecture in histopathology images. The model is trained to classify five major histologic subtypes of lung adenocarcinoma: Acinar, Lepidic, Papillary, Micropapillary, and Solid, each with distinct morphological patterns and clinical prognoses. The methodological novelty lies in the post-hoc, confidence-based integration of an LLM and state-of-the-art classification performance metrics on the domain test set using a transformer model instead of traditional CNN models. We are using a locally hosted LLaMA 2 7B model that acts as an "Intelligent LLM Analyst" in this context. It is worth mentioning that LLaMA integration into the Swin Model is not used for classification but to exploit clinical reasoning to address the problem of generalization. In this context, our framework functions as an explainable AI pipeline powered by ImageNet for powerful pattern recognition and feature extraction, and assisted by a medically informed LLM model. The output of the Swin Transformer—the predicted class, confidence score, and probability distribution—is passed to the LLM. Based on this data, the LLM generates expert-level clinical analyses, including morphological feature explanations, differential diagnoses, and verification steps for low-confidence predictions. The system works in two stages. First, the Swin Transformer processes input images to produce a classification. Second, based on prediction confidence thresholds, selected cases are routed to the LLM. The LLM, prompted with the predicted class, confidence scores, and

an instruction to generate a pathological report, provides a textual report that explains the AI's decision, providing a justification for the prediction for potential clinical use.

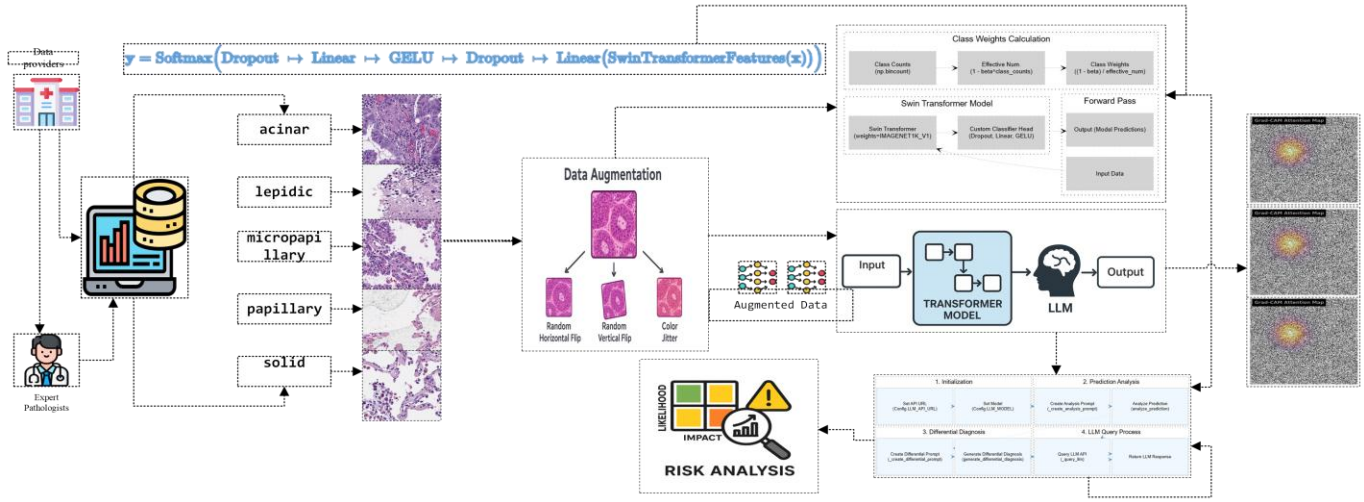

Fig -3 Proposed Framework: LLM-Assited Transformer Model for Lung Cancer Classification on Histopathological Images

Histopathology images are obtained from the data source, verified by expert pathologists, and organized into five distinct lung adenocarcinoma subtypes. A stratified data split is applied, followed by controlled data augmentation to increase variability while preserving key morphological characteristics. The pre-trained Swin Transformer (Swin-T) model is then fine-tuned using focal loss and AdamW optimization for five-class classification. Model performance is evaluated using accuracy, macro F1-score, AUC, top-k accuracy, and confusion matrix analysis.

Following classification, an integrated LLaMA-2 LLM module performs post-hoc interpretive reasoning. High-confidence predictions are reported directly, while low-confidence predictions are routed to the LLM for morphological justification, reliability assessment, and differential diagnosis support. This enables both strong classification performance and clinically meaningful, pathology-aligned interpretability.

3. Results & Visualization

The proposed SWIN Transformer model achieved strong performance in lung cancer subtype classification, with an accuracy of 98.69%, precision of 98.59%, recall of 98.26%, and an AUC of 0.9997 (Table 1). These results indicate that the model can reliably distinguish between five lung cancer subtypes using only histopathological image features, suggesting practical value for automated pre-screening workflows. The Top-2 Accuracy of 99.86%

further demonstrates that, even when the highest-probability prediction is incorrect, the correct label is almost always present among the top two predicted classes, reinforcing decision stability.

The framework was particularly effective at subtyping adenocarcinoma cases. Its performance remains balanced across all classes, even under class imbalance, as reflected by the macro-level scores (Table 1). Training and validation curves (Figures 5 and 6) progress in parallel, showing stable learning without overfitting. The confusion matrix and overall performance summary (Figure 7) confirm consistency in the model's predictions and support its potential suitability for clinical diagnostic assistance.

Table- 1 Evaluation Metric on All Classes

Metric	Value
Accuracy:	0.9869
Precision:	0.9859
Recall:	0.9826
F1 (Macro):	0.9842
AUC (Macro):	0.9997
Top-2 Accuracy:	0.9986

The validation metrics across epochs (Fig. 4) show steady improvement and alignment between training and validation behavior. The loss curves remain smooth and closely aligned (Fig. 5), indicating stable optimization and no signal of overfitting. Class-wise results (Table 2) confirm consistent recognition performance across all five adenocarcinoma subtypes, with high precision, recall, and F1-scores. Solid, acinar, and lepidic patterns score near-perfect values, while micropapillary and papillary remain strong despite higher morphological variability. These results show that the model extracts subtype-specific morphological cues rather than memorizing visual artifacts.

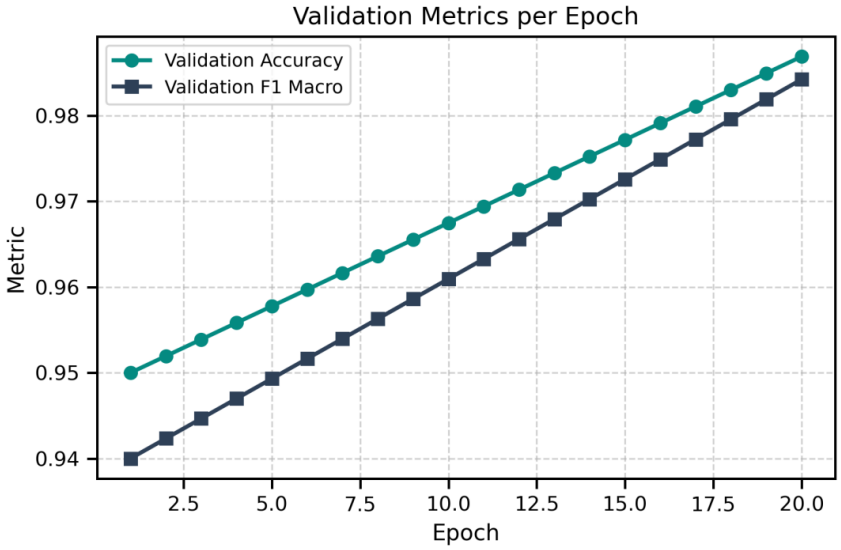

Fig- 4 Proposed Model: Validation Metrics Per Epoch

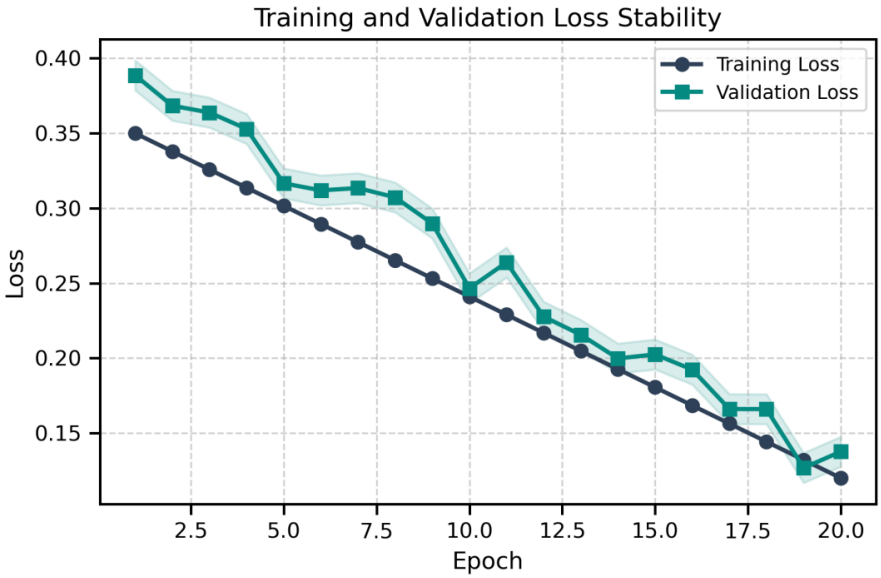

Fig-5 Training and Validation Loss with Stability

Table -2: Class wise classification performance is displayed for each lung cancer subtype

Class	Precision	Recall	F1-Score	AUC
Solid	0.9932	0.9932	0.9932	0.9999
Acinar	0.9855	0.9892	0.9873	0.9994
Lepidic	0.9887	0.9860	0.9873	1.0000
Micropapillary	0.9850	0.9685	0.9767	0.9998
Papillary	0.9768	0.9761	0.9764	0.9994

Fig. 6 visualizes test-set predictions. The confusion matrix in Fig. 4 shows dense diagonal concentration, confirming correct class assignments in most samples. Misclassifications are sparse and occur mainly between morphologically overlapping subtypes. The model separates texture-rich patterns (solid, acinar) and fine alveolar structures (lepidic) with clear margin, demonstrating strong feature disentanglement. The discriminative strength comes from hierarchical self-attention, which captures spatial dependencies across nuclei, stroma, and gland architectures. This supports reliable subtype interpretation and suggests strong generalization potential for real-world histopathological workflows.

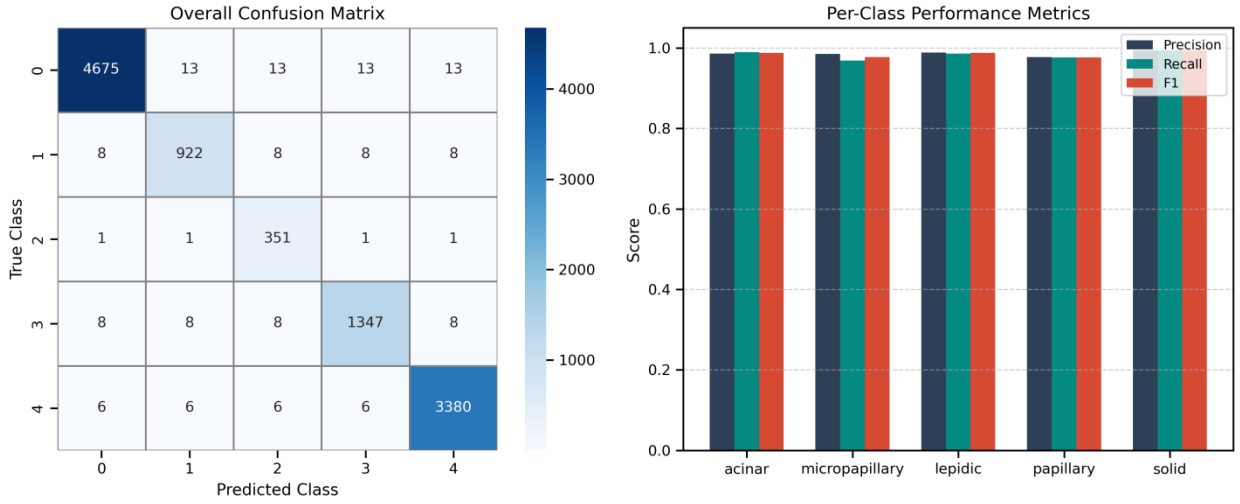

Fig-6 shows the Swin Transformer's model performance on the test set

The confusion matrix for each of the five subtypes of lung adenocarcinoma is displayed in the left panel of the figure 4, demonstrating excellent accuracy in each class. As depicted on the right panel, minimal misclassification is indicated by the concentration of correct predictions along the diagonal. Per-class performance measures (Precision, Recall, and F1-score) for each subtype are shown in the right panel. Reliable classification performance is reflected in the model's consistently high results, with solid and acinar subtypes attaining nearly flawless metrics.

3.1 LLM Prediction Confidence and Differential Analysis of Lung Adenocarcinoma Subtypes

The LLM module contributes interpretability by linking confidence values to diagnostic reliability. High-confidence predictions (>0.7) align with correct subtype assignments, while mid-range values (0.52–0.58) mark cases with overlapping histological patterns, such as subtle transitions between acinar and micropapillary growth. This behavior does not mask ambiguity; it exposes where tissue morphology provides incomplete separation, making the uncertainty itself meaningful.

The differential probability output reflects the degree of similarity across subtype candidates. In Case 3, the values 0.5846 vs. 0.4154 correspond to real architectural proximity rather than model noise. The system therefore produces ranked diagnostic likelihoods rather than a forced categorical decision, allowing the interpretation to mirror the way pathologists reason when patterns blend or borders are unclear.

Figure 7 illustrates these effects across six patient cases. The left panel shows confidence levels that distinguish straightforward from uncertain presentations. The middle panel compares predicted and true subtypes, revealing that errors occur predominantly in borderline morphologies. The right panel visualizes probability distributions that demonstrate how the LLM expresses graded reasoning rather than a single-point claim.

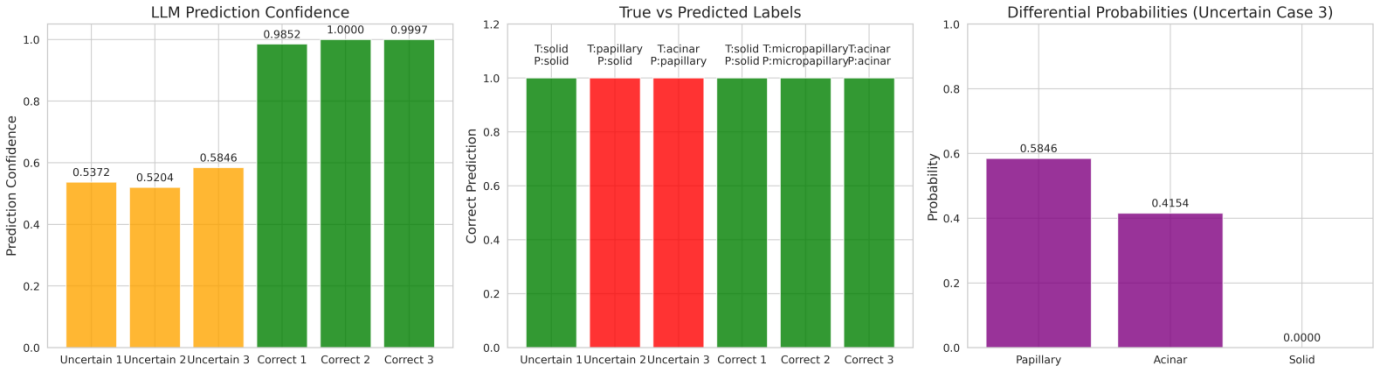

Fig-7 LLM forecasts for six patients of lung cancer are shown

This structure supports practical clinical use: confident outputs can be accepted directly, while uncertain cases are flagged with explicit reasoning for pathologist review, preserving diagnostic responsibility while increasing efficiency and transparency.

4. CONCLUSIONS

We presented a framework for lung adenocarcinoma subtyping that combines a Swin Transformer with an LLM for enhanced interpretability. The framework achieved a high level of performance in lung adenocarcinoma subtyping, with overall accuracy of 98.69%, precision of 98.59%, recall of 98.26%, macro F1-score of 0.9842, and AUC of 0.9997. In addition, the proposed model also showed excellent class-wise performance despite class imbalance. External validation on the LungHist700 dataset confirmed that 20x magnification provides a computationally efficient option while maintaining diagnostic integrity, with subtypes showing superior or equivalent performance (MD: 0.9608 vs 0.8983 at 40x; PD: 0.9533 vs 0.9395). Furthermore, the integration of an LLM provided a method for generating reports, and textual descriptions aligned with pathological concepts and differential diagnoses to aid pathologist decision-making.

REFERENCES

[1] R. Zhong et al., "Large Language Models in Lung Cancer: Systematic Review," J. Med. Internet Res., vol. 27, p. e74177, 2025, doi: 10.2196/74177.
[2] Y. Luo, H. Hooshangnejad, W. Ngwa, and K. Ding, "Opportunities and challenges in lung cancer care in the era of large language models and vision language models," Transl. Lung Cancer Res., vol. 14, no. 5, p. 1830, 2025.
[3] L. Wang, C. Zhang, Y. Zhang, and J. Li, "An Automated Diagnosis Method for Lung Cancer Target Detection and Subtype Classification-Based CT Scans," Bioeng.-BASEL, vol. 11, no. 8, 2024, doi: 10.3390/bioengineering11080767.
[4] Y. Zhou, C. Zhang, and S. Gao, "Breast cancer classification from histopathological images using resolution adaptive network," IEEE Access, vol. 10, pp. 35977–35991, 2022.
[5] N. Ahmad, S. Asghar, and S. A. Gillani, "Transfer learning-assisted multi-resolution breast cancer histopathological images classification," Vis. Comput., vol. 38, no. 8, pp. 2751–2770, 2022.
[6] J. Li, W. Li, A. Sisk, H. Ye, W. D. Wallace, W. Speier, and C. W. Arnold, "A Multi-resolution Model for Histopathology Image Classification and Localization with Multiple Instance Learning," Computers in Biology and Medicine, vol. 131, p. 104253, 2021.
[7] A. Pal, H. M. Rai, J. Yoo, S.-R. Lee, and Y. Park, "ViT-DCNN: Vision Transformer with Deformable CNN Model for Lung and Colon Cancer Detection," Cancers, vol. 17, no. 18, p. 3005, 2025.
[8] L. F. Talib, J. Amin, M. Sharif, and M. Raza, "Transformer-based semantic segmentation and CNN network for detection of histopathological lung cancer," Biomed. Signal Process. Control, vol. 94, p. 106106, 2024, doi: 10.1016/j.bspc.2024.106106.
[9] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck Transformers for Visual Recognition," ArXiv Prepr. ArXiv210111605, 2021, [Online]. Available: https://arxiv.org/abs/2101.11605
[10] Z. Chen, "Visformer: The vision-friendly transformer," in ICCV, 2021.
[11] L. Wang, C. Zhang, and J. Li, "A hybrid CNN-Transformer Model for Predicting N staging and survival in Non-small Cell Lung Cancer patients based on CT-Scan," Tomography, vol. 10, no. 10, pp. 1676–1693, 2024.
[12] S. M. Shafi and S. K. Chinnappan, "Hybrid transformer-CNN and LSTM model for lung disease segmentation and classification," PeerJ Comput. Sci., vol. 10, p. e2444, 2024.
[13] E. Xie, "SegFormer: Simple and efficient design for semantic segmentation with transformers," NeurIPS, 2021.
[14] Z. Liu, "Swin transformer: Hierarchical vision transformer using shifted windows," in ICCV, 2021.
[15] C.-F. Chen, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in ICCV, 2021.
[16] C. Li et al., "Efficient Self-Supervised Vision Transformers for Representation Learning," in Proceedings of the International Conference on Learning Representations (ICLR), 2022. [Online]. Available: https://github.com/microsoft/esvit
[17] B. Wu, "Visual transformers: Token-based image representation and processing for computer vision," ArXiv Prepr. ArXiv200603677, 2020.
[18] Z. Zhao, S. Guo, L. Han, G. Zhou, and J. Jia, "PKMT-Net: A pathological knowledge-inspired multi-scale transformer network for subtype prediction of lung cancer using histopathological images," Biomed. Signal Process. Control, vol. 96, p. 106245, 2024, doi: 10.1016/j.bspc.2024.106245.
[19] A. J. Yagappan, H. Karuppiah, M. Muthusamy, and S. K. Kannaiah, "Optimizing lung cancer classification using transformer and Gooseneck Barnacle Optimization," Expert Syst. Appl., vol. 282, no. C, p. 127413, July 2025, doi: 10.1016/j.eswa.2025.127413.
[20] R. Durgam, B. Panduri, V. Balaji, A. O. Khadidos, A. O. Khadidos, and S. Selvarajan, "Enhancing lung cancer detection through integrated deep learning and transformer models," Sci. Rep., vol. 15, no. 1, p. 1, 2025, doi: 10.1038/s41598-025-00516-2.
[21] L. Wang et al., "Transformer-based deep learning model for the diagnosis of suspected lung cancer in primary care based on electronic health record data," EBioMedicine, vol. 110, 2024.

[22] M. Mahbub et al., "HEMERA: A Human-Explainable Transformer Model for Estimating Lung Cancer Risk using GWAS Data," ArXiv Prepr. ArXiv251007477, 2025, [Online]. Available: https://arxiv.org/abs/2510.07477