Detecting Lunc Cancer in Histopathological Images Using CNN and Visual Transformers

David Caspers
University of Syracuse, ISchool
IST691, Applied Deep Learning
Austin, Texas
dcaspers@syr.edu

***Abstract*— This study explores the use of deep learning models for detecting lung cancer in histopathological images, comparing a convolutional neural network (CNN) and a Vision Transformer (ViT). While the CNN demonstrates strong performance, the ViT significantly outperforms it by minimizing false negatives and providing interpretable attention maps, making it a more suitable choice for medical diagnostics. The findings highlight the transformative potential of advanced AI tools in improving early cancer detection and patient outcomes.**

I. Introduction

Lung cancer is a leading cause of cancer-related deaths worldwide, responsible for over 1.8 million deaths annually. Early and accurate diagnosis is critical, as detecting lung cancer in its early stages can significantly improve survival rates [1]. Histopathological analysis of tissue samples, where stained biopsy specimens are examined under a microscope, is a primary diagnostic method. However, this process is time-intensive, subjective, and prone to variability among pathologists, leading to inconsistencies in outcomes [2].

Machine learning, particularly deep learning, offers a promising solution by automating and improving the diagnostic process. This study evaluates two advanced deep learning models for detecting lung cancer in histopathological images: a convolutional neural network (CNN) inspired by the YOLO architecture [3] and a pre-trained Vision Transformer (ViT), google/vit-base-patch16-224-in21k [4, 5]. While CNNs efficiently extract spatial features, ViTs leverage attention mechanisms to capture global context and enhance interpretability. By comparing the performance of these models, this paper aims to identify their effectiveness in reducing diagnostic variability and improving outcomes in clinical settings.

A. Traditional Lung Cancer Detection Mechanisms

Lung cancer is traditionally detected through imaging techniques such as chest X-rays and CT scans, followed by histopathological examination, where tissue samples from a biopsy are stained and analyzed under a microscope to identify abnormal cell patterns indicative of cancer. This process, while essential for diagnosis, is time-intensive and prone to variability due to subjective interpretations by pathologists.

B. Convolutional Neural Networks and Transformers

Artificial intelligence (AI) plays a growing role in lung cancer detection, leveraging models like convolutional neural networks (CNNs) and, more recently, transformers. While CNNs are well-established for extracting spatial features from medical images, transformers are relatively new in this domain and offer significant advantages through their use of attention mechanisms. These mechanisms allow the model to detect dependencies across different regions of an image, making transformers particularly effective at capturing complex, global patterns within histopathological slides. This capability enhances the accuracy and consistency of cancer detection, providing a promising direction for AI-driven diagnostics.

II. Methodology

A. Dataset

The dataset used in this study is the Lung and Colon Cancer Histopathological Images dataset, sourced from Kaggle [6]. It contains 15,000 histopathological images equally divided into three classes: lung adenocarcinoma (lung_aca), lung squamous cell carcinoma (lung_scc), and normal lung tissue (lung_n). The dataset was already augmented, with variations in rotation, scale, and other transformations, ensuring a diverse set of samples that reduces the need for additional data augmentation.

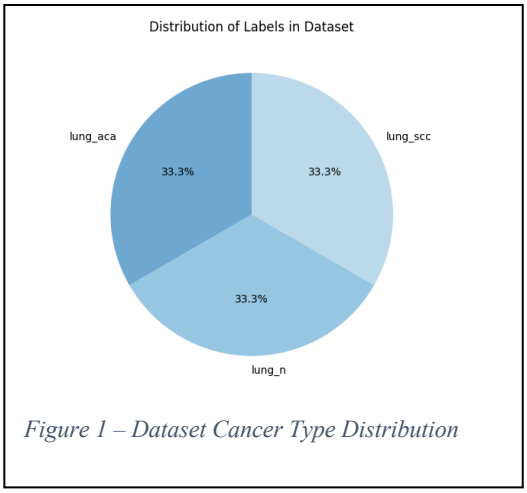

Figure 1 – Dataset Cancer Type Distribution

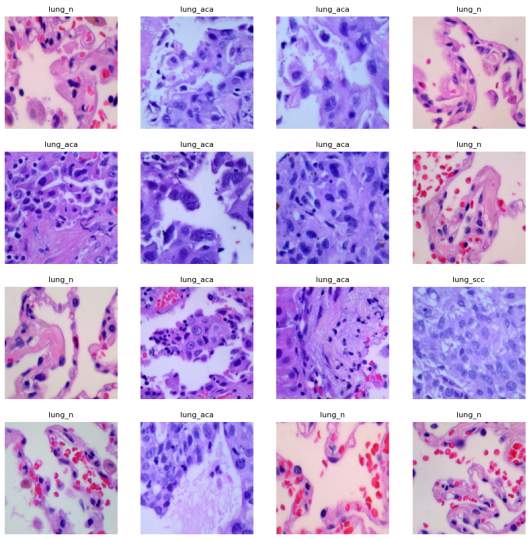

Figure 2 – Sample of Training Images

B. Model Architectures:

The convolutional neural network (CNN) architecture used in this study, loosely inspired by the YOLO design, consists of four convolutional layers with max-pooling, followed by two dense layers for classification. Designed to balance computational efficiency with accuracy, this model ensures faster training and inference while effectively classifying histopathological images.

The Vision Transformer (ViT) model ('google/vit-base-patch16-224-in21k') uses an attention-based approach by splitting images into 16x16 patches, linearly embedding each of them with added position embeddings, and feeding the resulting sequence of vectors to a standard Transformer encoder. Finally, a feed-forward network is used for classification. The technique of splitting the image into discrete patches helps reduce the quadratic computational complexity of attention mechanisms. The attention mechanism is particularly useful as it can easily be used to visualize specific areas of the image that the classifier finds useful in making a categorization.

C. Preprocessing

Very little data preprocessing was required. For the CNN, images were resized to 224x224 pixels. Data augmentation was not required as the dataset included augmented images already. For the Vision Transformer (ViT), images were resized to 224x224 pixels and normalized using the mean and standard deviation specified by the pre-trained model. No additional preprocessing was required aside from formatting transformations required for the respective models.

D. Training and Evaluation

The models were trained using the training set, with 70% of the dataset allocated for training, while 10% was used as the validation set to monitor performance during training. The remaining 20% of the dataset was reserved as the test set for final evaluation. Training samples were split evenly across sets. Both models were trained for five epochs, using sparse categorical cross-entropy as the loss function and the Adam optimizer. Evaluation metrics, including accuracy, precision, recall, and F1-score, were calculated on the test set to assess model performance and robustness, which can be found in the next section.

III. Experiments and Results

A. Model Performance

Model	**CNN Model Performance Statistics**			
	Precision	Recall	F-1	Support
ACA	0.85	0.83	0.84	912
Normal	0.99	0.98	0.99	933
SCC	0.86	0.88	0.87	1035

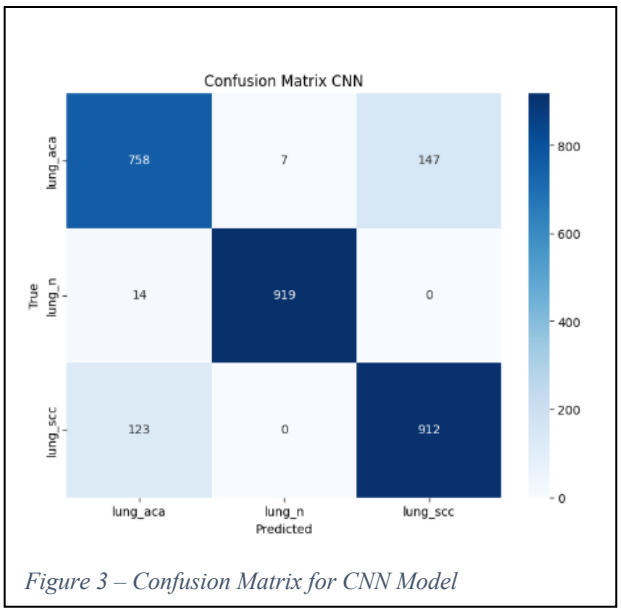

Figure 3 – Confusion Matrix for CNN Model

Model	**Transformer Model Performance Statistics**			
	Precision	Recall	F-1	Support
ACA	0.91	0.99	0.95	1002
Normal	1.00	1.00	1.00	992
SCC	0.99	0.91	0.95	1006

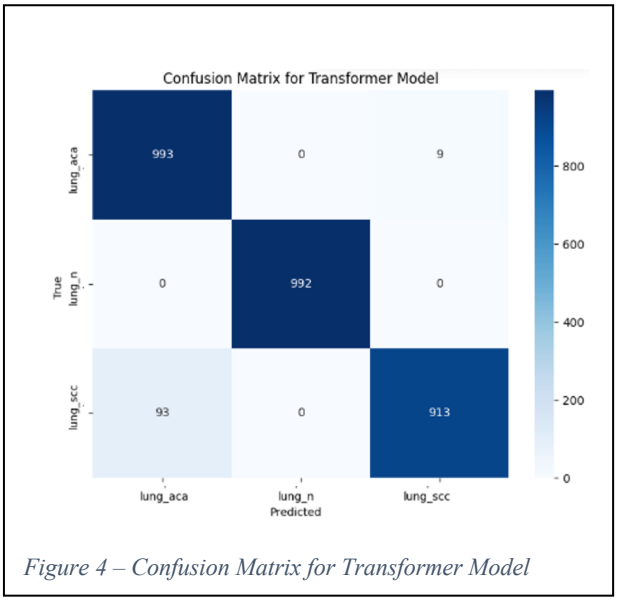

Figure 4 – Confusion Matrix for Transformer Model

B. Visualization

The attention mechanism in the Vision Transformer provides a critical advantage in interpretability over the CNN. Attention maps generated by the ViT highlight specific regions of histopathological images that are deemed important for classification decisions. This capability is particularly valuable in medical diagnostics, as it allows clinicians to understand which areas of the image the model focused on, providing an additional layer of trust and transparency in high-stakes decisions.

In contrast, the CNN relies on convolutional filters that effectively capture local patterns but lack the explicit visualization of decision-making regions. The ViT's attention maps can also be used to verify the relevance of the identified regions, ensuring that the model's predictions align with known pathological indicators. This enhanced interpretability, coupled with its superior classification performance, makes the Vision Transformer a promising tool for assisting medical professionals in the diagnostic process.

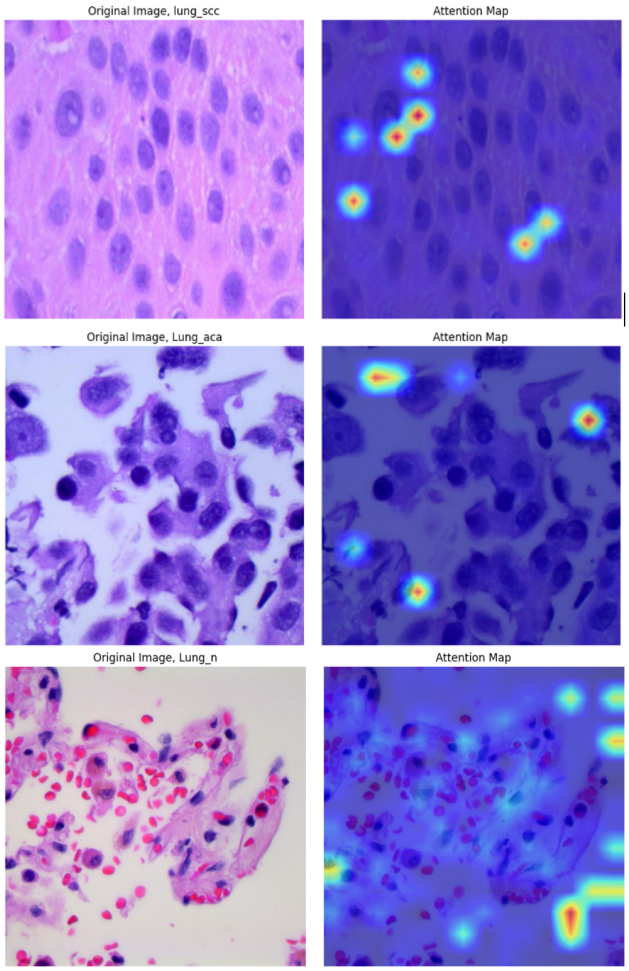

Figure 5 - ViT Attention Mechanism

IV. Discussion

A. Interpretation of Results and Insights Gained

In the context of lung cancer detection, minimizing false negatives is of paramount importance. False negatives—where malignant conditions such as adenocarcinoma (ACA) or squamous cell carcinoma (SCC) are incorrectly classified as "normal"—pose significant risks by delaying diagnosis and treatment. In contrast, false positives, while less ideal, often result in additional testing rather than missed opportunities for early intervention.

The CNN model, while computationally efficient, demonstrated limitations in distinguishing between cancer types. For example, it achieved a recall of 0.83 for ACA and 0.88 for SCC, suggesting its sensitivity is insufficient for clinical applications where accuracy is critical. Conversely, the Vision Transformer (ViT) excelled in reducing false negatives, achieving a recall of 0.99 for ACA and perfect recall for normal tissue. These results underscore the ViT's potential for improving diagnostic accuracy and reliability in high-stakes healthcare scenarios.

B. Comparison

When weighing false positives against false negatives, the Transformer model is clearly superior. It dramatically reduces false negatives for ACA and SCC compared to the CNN model, aligning better with the preference for minimizing missed diagnoses in high-stakes medical contexts. Although both models struggle to differentiate between cancers, the Transformer's ability to correctly classify nearly all malignant cases makes it a safer and more effective choice for this application. Overall, the Transformer's focus on minimizing false negatives makes it far better suited for medical image analysis, where the cost of a missed diagnosis can be life-altering.

C. Limitations and Future Considerations

Despite the Vision Transformer's superior performance in reducing false negatives, its deployment in real-world clinical settings is hindered by practical limitations. The model's computational complexity, requiring significant memory and processing power, may make it unsuited for resource-constrained environments such as rural hospitals or clinics. While the CNN is less accurate in reducing false negatives, its computational efficiency and faster inference times make it a viable alternative for settings with limited hardware resources. Future research should focus on optimizing the Vision Transformer for lower-resource environments, such as through efficient model architectures or techniques like knowledge distillation.

V. Conclusion

This study highlights the potential of artificial intelligence in improving the accuracy and efficiency of lung cancer detection. While the Vision Transformer model proved particularly effective in identifying cancer cases and providing insights into its decision-making process, it is worth noting that larger versions of traditional models, such as Convolutional Neural Networks (CNNs), could rival its performance. However, the addition of the attention mechanism in the Vision Transformer makes it especially intuitive and useful for medical practitioners, offering transparency that is equally important as accuracy in high-stakes healthcare settings.

By integrating tools like these into healthcare, the process of identifying cancer could become faster, more reliable, and less dependent on individual interpretation, leading to better outcomes for patients. Future advancements in these technologies hold great promise for improving early detection and treatment, ultimately saving lives and enhancing the quality of care.

VI. References

[1] World Health Organization. (n.d.). *Cancer.* Retrieved December 20, 2024, from https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Sarvamangala, D. R., & Kulkarni, R. V. "Convolutional neural networks in medical image understanding: a survey." *Evolutionary Intelligence*, 15(1), 1–22 (2022). doi: 10.1007/s12065-020-00540-3. Epub 2021 Jan 3. PMID: 33425040; PMCID: PMC7778711.

[3] V7 Labs. (2023, January 17). *YOLO Object Detection: What it is and how it works.* Retrieved December 20, 2024, from https://www.v7labs.com/blog/yolo-object-detection

[4] Dosovitskiy, A., et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint arXiv:2010.11929v2* (2021). Available: https://arxiv.org/abs/2010.11929v2

[5] "Vision Transformer (ViT): An Overview." *Viso.ai*. Available: https://viso.ai/deep-learning/vision-transformer-vit/

[6] Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019

[7] "ViT Base Patch 16 224 Model Card." *Hugging Face*. Available: https://huggingface.co/google/vit-base-patch16-224

[8] "Fine-Tuning Vision Transformers." *Hugging Face Blog*. Available: https://huggingface.co/blog/fine-tune-vit