International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Lung Vision: Leveraging CNNs and Vision Transformers for Accurate and Scalable Lung Cancer Diagnosis

Zeeshan Ahmad[1], Mamta Sharma[2]

[1]Research Scholar, Department of Computer Science and Engineering, Jayoti Vidyapeeth Women's University, Jaipur

[2]Research Supervisor and Associate Professor, Department of Computer Science and Engineering, Jayoti Vidyapeeth Women's University, Jaipur

Abstract- Proposed framework-LungVision is a deep learning framework for automated lung cancer detection using CT and histopathological images. It integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). CNNs are good at extracting local features while ViTs utilize self-attention mechanisms to capture global context. This combination helps to extract local and global feature extraction. The sole CNN application achieved a detection accuracy of 94.063%, while the best performing ViT model achieved 84.211%. The study evaluates both models using balanced and unbalanced datasets under supervised and transfer learning setups. Performance is assessed using standard metrics including accuracy, precision, recall, and F1 score. With CNN integrated ViTs, cancerous fea- tures are extracted with precision 97%. The proposed framework is also deployed as a real-time web application, offering an accessible solution for early lung cancer diagnosis in case of cancer detection.

Keywords- Vision Transformers (ViTs), Convolutional Vision Transformer (CVT), Lung Cancer Detection, Medical Image Analysis, Parallel Vision Transformers (P-ViT).

I. INTRODUCTION

Cancer remains the leading cause of death across the globe, and lung cancer alone contributes meaningfully to cancer- induced mortality in general, reports the World Health Orga- nization (WHO) [1]. Early detection of lung cancer is critical to improve survival rates. Conventional diagnosis methods like CT scan interpretation by radiologists are often time- consuming, subject to human error, and influenced by inter- observer and intra-observer differences [2,3]. As imaging technology continues to evolve, there is a mounting need for rapid, automatic diagnosis systems that minimize diagnostic errors and yield standardized interpretation for different patients and clinicians.

Artificial Intelligence (AI), and Deep Learning (DL) in particular, has made significant strides in overcoming these obstacles. Vision Transformers (ViTs) have more recently emerged as a very promising alternative, borrowed from Natural Language Processing (NLP) methods applied to Computer Vision. Through the application of self-attention mechanisms, ViTs are able to efficiently capture global contextual relations in image data, and are therefore well-suited to application in challenging domains such as histopathological and radiological image classification.

Researchers are progressively investigating hybrid architectures that combine the advantages of Vision Transformers with convolutional neural networks (CNNs) to get better results. These models provide a

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

balanced approach to challenging medical imaging tasks by combining the global dependence modeling of ViTs with the local feature extraction capability of CNNs. Fig. 1 illustrates the Vision Transformer model for lung cancer detection. The input CT image is partitioned into fixed-size patches, each partition is linearly embedded with positional encoding, and a learnable [class] token which is added at the beginning of the sequence for classification.

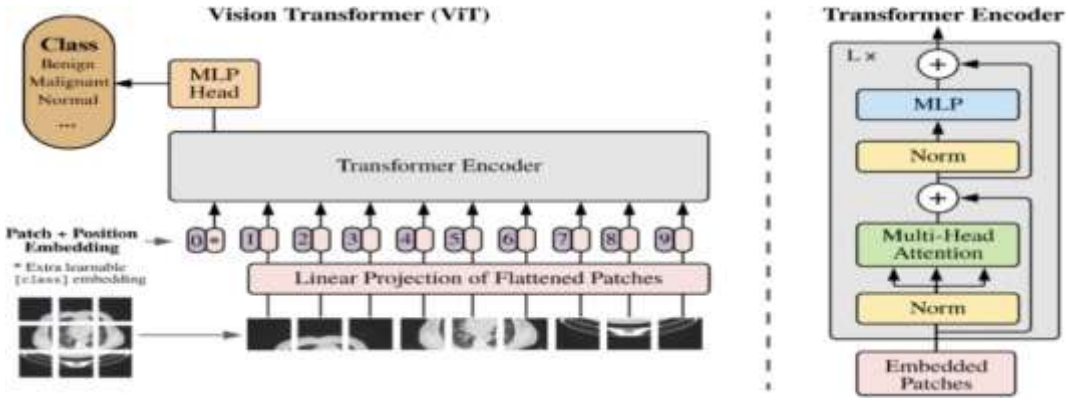


Fig. 1: Vision Transformer Architecture for Lung Cancer Classification using CT image patches.

This paper evaluates the performance of CNN and Vision Transformer models, a deep learning models, for lung cancer detection on both balanced and imbalanced datasets [23][24] in supervised and transfer learning settings. Accuracy, precision, recall, and F1-score are employed to measure performance. Several variants—CVT, CCT ViT, Parallel ViT, and Efficient ViT [4-7] are also explored, and evaluated with metrics such as accuracy, precision, recall, F1-score, and AUC-ROC [8-11]. Cost-sensitive metrics, including weighted loss and cost matrices, were also applied to reflect the clinical impact of misclassification, particularly false negatives [12]. The solution is implemented as a real-time web application to enable accessible and efficient lung cancer detection in clinical and remote settings.

II. LITERATURE REVIEW

Deep learning has significantly advanced medical image analysis, particularly in automating cancer detection work- flows. ViT architecture is based on the transformer architecture originally developed for Natural Language Processing (NLP). Differently from convolutional models operating over local receptive fields, ViT uses global self-attention over the whole image. The image is split into fixed-sized patches, flattened, and linearly projected into an embedding space. Positional embeddings are added to maintain spatial context, and a learnable [class] token is appended to the sequence prefix for classification.

The encoded sequence is passed through a sequence of Transformer encoder blocks, which consist of multi-head self-attention and feedforward layers. The final output is generated by a classification head. Recent works have proposed enhancements to ViT architectures which was initially proposed by Dosovitskiy et al. [13,14]. Several reviews and studies have explored ViTs in medical imaging. Ayana et al. [15] combined spatial and vision transformers for colorectal cancer, improving local and global feature extraction. Khan et al. [16] demonstrated ViTs' strength in handling high-resolution data for lesion diagnosis.

Convolutional Vision Transformer (CVT) is a transformer- hybrid model that integrates convolutional layers and self- attention mechanisms to draw benefits from both global and local feature extraction [17]. It is hierarchical, multi-stage in design with convolutions used to down-sample and embed the input features. Depth-wise convolutions are used to produce queries, keys, and values for the attention mechanism, ensuring the model remains computationally efficient while retaining spatial context. This

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

integration allows CVT to benefit from convolutional inductive biases while retaining transformer scalability.

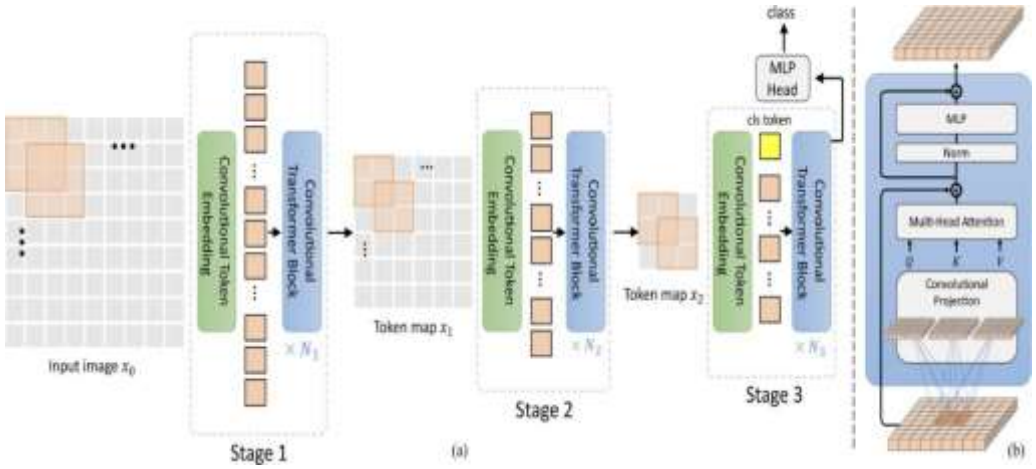


Fig. 2: Pipeline of the CVT architecture [17]

In the illustration, the pipeline of the CVT architecture comprises two parts, (a) part highlights the hierarchical multi-stage structure facilitated by the convolutional token embedding layer while (b) part highlights the convolutional transformer block, where the convolution projection serves as the initial layer.

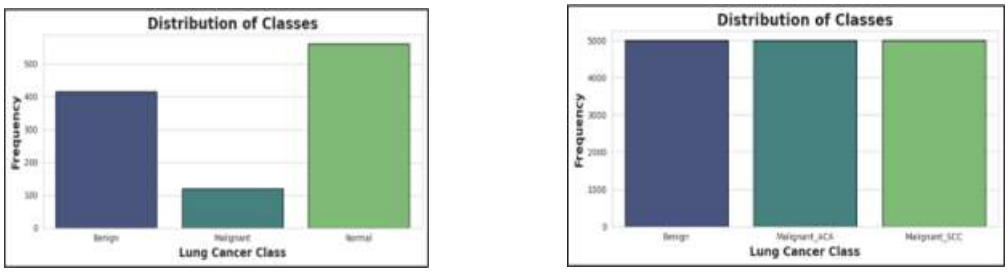
Touvron et al.[18] proposed a model that can parallelise multi- headed self-attention and residual feedforward networks, making the ViT model more cost-efficient during training . Lee et al. [19] in their work introduced Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA)—two lightweight add-on techniques for injecting locality inductive bias into Vision Transformers, thereby enabling 3% (approx.) average accuracy boost on Tiny-ImageNet and approximately 4% on Swin Transformer when trained from scratch on small datasets.

By combining CNN-based local feature extraction, attention-driven global context learning, and transfer learning, Gandhi et al. [20] proposed a Hybrid Attention Vision Transformer (HViT) architecture that achieves 95.6% accuracy, 97.3% recall, and 0.98 AUC in early-stage lung cancer detection from CT scans, with improved interpretability and clinical applicability. Tran et al. [21] showed promise in ViTs for lung cancer detection but omitted augmentation and modality diversity. While improvements in efficiency, locality awareness, and interpretability have enhanced performance, challenges such as limited data availability, lack of augmentation strategies, and modality diversity remain open research directions.

III. METHODOLOGY

Datasets used:
Two datasets provided the training data for each model. IQ- OTH/NCCD, provided by Alyasriy et al., was used to train most models due to the imbalanced ratio of classes it offers [23]. IQ- OTH/NCCD dataset provides CT scan images of three classes: normal, malignant, and benign (Fig. 3). The second dataset is a perfectly balanced large dataset using histopathological lung cancer images with three defined classes provided by Borkowski et al. [24].

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752



- Fig. 3: Imbalanced & balanced dataset [23], [24]

Image Preprocessing the data:
Normalization and Standardization: Normalisation and standardisation limits the range of each pixel value in each image contained in the dataset. These techniques ensures that the pixel values are similar and contribute equally to training the model.

Image augmentation: Rotation, Cropping, Scaling and flipping are utilized to create image size and orientation variations, this helps the model to generalize better on unseen data.

Training Approaches:
- Supervised Learning: This learning approach associates each image in the lung cancer detection training set with a label that the model will be learning.
- Transfer learning: This training approach applies knowledge gained from a pre-trained model on a source task and configures it to a specific task.

Evaluation Metrics:
Each model was evaluated using established evaluations such as accuracy, recall, precision, F1-score, and area under the ROC (AUC-ROC).

Cost-sensitive Evaluation Metrics: In lung cancer detection, misdiagnosing a patient who has cancer (False Negative) is far more critical than misdiagnosing a patient without cancer (False Positive). Cost-sensitive metrics will help analyse which models minimised False Negatives. False negatives may occur due to class imbalance when a class has a higher frequency ratio in the corpus. For Cost-Sensitive Evaluation on Multi-class Classification, the following metrics will be used:

Real-world weight cross-entropy loss:
This evaluation is the real-world weight that models the cost of misclassification, allowing the model to account for the class with a lower representation in the dataset [22]. The standard weighted binary cross-entropy loss function is given by:

$$J_{wbce} = -\frac{1}{M}\sum_{m=1}^{M}[w \times y_m \times \log(h_\theta(x_m))] + (1-y_m) \times \log(1-h_\theta(x_m))$$

Where M is the number of training samples, w is weight, ym is target label for training sample m, xm is the input for training sample m, and hθ is the model with neural network weights Θ.

Cost-sensitive matrix: The Cost-Sensitive Matrix measures the probability of mis-classification given a class. It helps understand the trade- offs between errors, which may entail real-world consequences [12].

IV. RESULT AND DISCUSSION

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Table 1: Evaluation Metrics Vs. Patch Size

Patch Size	Accuracy	Loss	F1 Score	Precision	Recall	AUC-ROC
16px	0.66	0.5	0.44	0.47	0.47	0.83
32 px	0.74	0.41	0.52	0.5	0.54	0.87

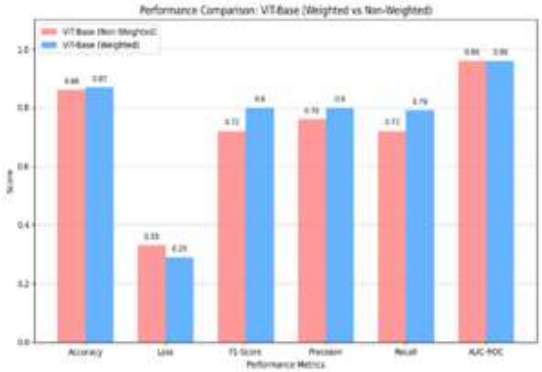


Fig. 4: Performance Comparison of ViT- Base with weighted and unweighted loss function

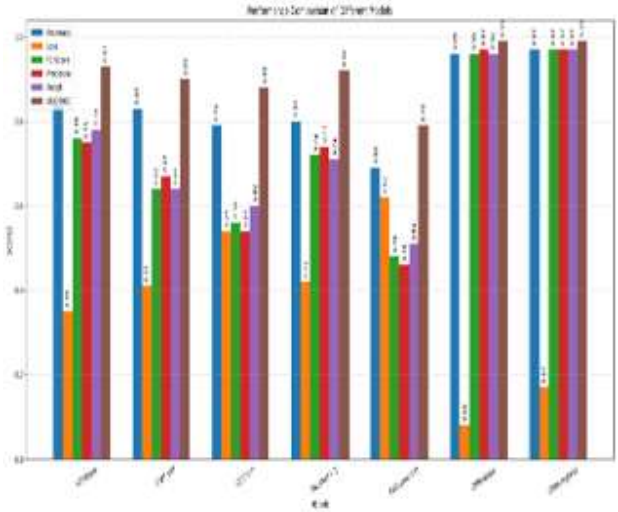


Fig. 5: CNN and ViT Models Performance Comparison

The first result computed was the hyperparameter, the patch size, which contributed to the overall performance of the Vision Transformer model. Vision Transformers use global attention, which requires the input image to be divided into patches [13,14]. Experiments were conducted for 16px as well as 32 px patch size. Using a larger patch size, 32px, takes better advantage of the global self-attention mechanism.

While training a model, loss functions can be weighted or unweighted to handle misclassification. A weighted loss function can give higher importance to minority classes, accounting for the above imbalance. The weights are calculated by taking the inverse of each class's frequency, ensuring that classes with fewer samples receive higher weights and, hence, higher importance, and then averaging

the weights across training, validation, and test sets to create a universal weight. The results below show the difference between weighted and unweighted loss functions.

Vision transformers and state-of-the-art architecture results:

Fig.5 illustrates the results of the ViT-Base, CvT-ViT, CCT-ViT, Parallel-ViT, Efficient-ViT, CNN-Base, CNN-Hybrid models trained on the IQ-OTH/NCCD dataset from the ground up [25]. Each Vision Transformer model achieved an average accuracy of 80 percent with 20 epochs, while CNN achieves same accuracy with the 50 epochs. This demonstrates that the multi-head self-attention architecture could generalise and converge. Bar chart compares the performance of seven different models across six metrics: Accuracy, Loss, F1-Score, Precision, Recall, AUC-ROC. It can be observed from the results that CNN-Hybrid performs better than other models by a clear margin. Better performance can be pinned due to leveraging of the strength local precision and global context.

V. CONCLUSION

In this work, we explored the Vision Transformer, the first sequence transduction model based entirely on attention. It replaces the recurrent layers most used in encoder-decoder architectures with multi-headed self-attention. The CNN-Hybrid model performs better than the others, obtaining the highest scores in accuracy, F1-score, precision, recall, and AUC-ROC, as well as the lowest loss, according to the performance comparison across several evaluation criteria. This suggests that it has a high capacity for generalization and efficient dataset learning.

On the other hand, the ViT-based models (such as ViT-Base, CvT-ViT, and CCT-ViT) perform rather poorly, which could be explained by their reliance on extensive training data and their limited capacity to identify local features in smaller datasets. Future work will focus on understanding the embedding of Vision Transformer Models. These models can be more efficient by employing algorithms that can optimise the quadratic self-attention mechanism.

REFERENCES

1. B. Zhang, H. Shi, and H. Wang, "Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach," Frontiers in Oncology, vol. 13, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC10312208
2. K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," Genome Medicine, vol. 13, no. 1, 2021, Art. no. 152. https://pmc.ncbi.nlm.nih.gov/articles/PMC8477474
3. H. Al-Khawari, R. P. Athyal, O. Al-Saeed, P. N. Sada, S. Al-Muthairi, and A. Al-Awadhi, "Inter- and intra-observer variation between radiologists in detecting abnormal parenchymal lung changes on high-resolution computed tomography," Annals of Saudi Medicine, vol. 30, no. 2, pp. 129–133, 2010. https://pmc.ncbi.nlm.nih.gov/articles/PMC2855063
4. H. Wu, J. Xiao, H. Codella, Y. Liu, and Y. Dai, "CvT: Introducing convolutions to vision transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 22–31. doi: 10.1109/ICCV48922.2021.00011.
5. A. Hassani, S. Walton, J. Li, and H. Shi, "CCT: Compact convolutional transformers," in Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2022, pp. 1829–1839. doi: 10.1109/WACV51458.2022.00191.
6. Y. Yang, S. Tang, C. Xu, and Y. Song, "Parallel vision transformer: Multi-branch attention for efficient image recognition," Pattern Recognition, vol. 138, 2023, Art. no. 109440. doi: 10.1016/j.patcog.2023.109440.

7. L. Liu, Z. Li, L. Kuang, H. Lin, J. Zhou, and Y. Wang, "EfficientViT: Memory efficient vision transformer with cascaded group attention," arXiv preprint arXiv:2205.14756, 2022. https://arxiv.org/abs/2205.14756
8. J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, H. Luo, and D. Tao, "A survey on self-supervised learning: Algorithms, applications, and future trends," arXiv preprint arXiv:2301.05712, 2023. https://arxiv.org/abs/2301.05712
9. M. Cowan, G. Tesauro, and V. de Sa, "Learning classification with unlabeled data," 1997.
10. M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 2, pp. 1–11, 2015. doi: 10.5121/ijdkp.2015.5201
11. K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Caspian Journal of Internal Medicine, vol. 4, no. 2, pp. 627–635, 2013.https://pmc.ncbi.nlm.nih.gov/articles/PMC3755824
12. I. D. Miyenye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," Informatics in Medicine Unlocked, vol. 25, Art. no. 100690, 2021, doi: 10.1016/j.imu.2021.100690.
13. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017
14. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020. arXiv:2010.11929, 2020. https://arxiv.org/abs/2010.11929
15. G. Ayana, H. Barki, and S. W. Choe, "Pathological insights: Enhanced vision transformers for the early detection of colorectal cancer," Cancers, vol. 16, no. 4, p. 1441, 2024. doi: 10.3390/cancers16041441.
16. S. Khan, H. Ali, and Z. Shah, "Identifying the role of vision transformer for skin cancer—A scoping review," Frontiers in Artificial Intelligence, vol. 6, 2023, Art. no. 1202990. doi: 10.3389/frai.2023.1202990.
17. H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), Montreal, QC, Canada, Oct. 2021, pp. 22–31. doi: 10.1109/ICCV48922.2021.00011.
18. H. Touvron, M. Cord, A. El-Nouby, J. Verbeek, and H. Jégou, "Three things everyone should know about vision transformers," in Computer Vision – ECCV 2022, vol. 13683, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds. Cham: Springer, 2022, pp. 497–515. doi: 10.1007/978-3-031-19830-4_29.
19. S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," arXiv preprint arXiv:2112.13492, 2021. https://arxiv.org/abs/2112.13492
20. J. N. Gandhi, K. V. Guru, Y. Varun, K. Vasanth, P. Vimal, and K. Madheswaran, "Hybrid Attention Vision Transformer for Enhanced Lung Cancer Detection," in Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024), Atlantis Press, 2025, pp. 671–682. doi: 10.2991/978-94-6463-718-2_58.
21. K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," Genome Medicine, vol. 13, no. 1, p. 152, 2021. doi: 10.1186/s13073-021-00998-y.
22. Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," IEEE Access, vol. 8, pp. 4806–4813, 2020, doi: 10.1109/ACCESS.2019.2962617.
23. H. Alayasiry and M. Al-Huseiny, "The IQ-OTH/NCCD lung cancer dataset," Mendeley Data, vol. 4, 2023. doi: 10.17632/bhmdr45bh2.4
24. A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (LC25000)," arXiv preprint arXiv:1912.12142, 2019. https://arxiv.org/abs/1912.12142