International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 992-1004
https://doi.org/10.70917/ijcisim-2026-5096
Article

Adaptive Feature Integration in CNN–Transformer Networks for Efficient and Interpretable Visual Classification

Mrs Komal Sharma[1], Dr Monika Sainger[2]

[1]PhD. Scholar, Asian International University Manipur India
Email : kskomal1995@gmail.com
[2]Professor, Asian International University Manipur India
Email : drmonikasainger@gmail.com

Abstract: Over the past few years, deep learning has changed substantially following the emergence of Transformer architectures, which are particularly effective for representing long-range dependencies that are difficult for conventional Convolutional Neural Networks (CNNs). Whereas CNNs are well suited to extracting local spatial features using convolutional operations, Transformers are effective at representing global context through self-attention. Hybrid CNN–Transformer architectures have been developed to combine the respective strengths of the two approaches. A limitation of many existing models is their reliance on static or manually designed fusion strategies, which can restrict adaptability, add computational cost, and make the resulting decisions harder to interpret. The present study develops a novel adaptive fusion framework that adaptively combines CNN and Transformer features through learnable gating, attention-based feature integration, and explainable-AI methods. The resulting framework is intended to improve both computational efficiency and model interpretability, thereby addressing important limitations of current hybrid designs. The experimental evaluation uses benchmark datasets such as ImageNet, CIFAR-100, and medical imaging datasets. The reported results show that the proposed model performs better than the comparison architectures with respect to accuracy, efficiency, and interpretability.
Keywords: CNNs, CIFAR-100, ImageNet, Datasets, Efficiency, AI techniques etc

1. INTRODUCTION

Artificial Intelligence (AI) has become a major technology of the twenty-first century, with applications spanning healthcare, transportation, finance, education, manufacturing, and autonomous systems. Within this broader area, deep learning has achieved strong results because it can learn hierarchical representations directly from input data. In computer vision, Convolutional Neural Networks (CNNs) are widely used for image classification, object detection, medical image analysis, and pattern recognition largely because they can efficiently learn local spatial representations. A remaining limitation of CNNs is their difficulty in representing long-range dependencies and broader contextual relationships, particularly in complex visual scenes.

To reduce these limitations, Vision Transformers (ViTs) have consequently been investigated as an alternative. In contrast to CNNs, Vision Transformers use self-attention to represent relationships among image patches, which can strengthen contextual representation and classification performance on large-scale image recognition tasks. Even with these benefits, Vision Transformers can demand substantial computational resources, training data, and memory, which can make deployment difficult in real-time or resource-limited settings.

Hybrid CNN–Transformer architectures have received increasing research attention by combining the strengths of CNNs and Transformers. CNNs effectively learn local texture and structural information, whereas Transformers capture global contextual dependencies. Using both representations can provide richer features and improve

classification performance. However, most existing hybrid models employ fixed feature fusion strategies, such as concatenation or weighted averaging, which cannot adapt to varying input characteristics. Additionally, interpretability remains limited in many such models, limiting their applicability in critical domains such as healthcare.

Lung cancer detection using medical imaging is one such application where both high accuracy and model transparency are essential. Early diagnosis through Computed Tomography (CT) imaging can significantly improve patient survival, but manual analysis is time-consuming and subject to human error. This creates a need for intelligent, efficient, and explainable AI systems that can support clinicians in accurate diagnosis.

To address these issues, this research proposes an Adaptive Hybrid CNN–Transformer Architecture (AHCTA) that adaptively combines local CNN features and global Transformer features using a learnable adaptive fusion mechanism. The framework additionally incorporates Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM and attention visualization, to make the model's decisions easier to inspect. Evaluation is performed on benchmark datasets such as ImageNet, CIFAR-100, and lung cancer CT image datasets using performance metrics including Accuracy, Precision, Recall, F1-score, AUC, computational efficiency, and inference time. The study is intended to provide a robust, computationally efficient, and interpretable AI framework suitable for next-generation computer vision and medical imaging applications.

2. LITERATURE REVIEW

Developments in deep learning have substantially improved the performance of Artificial Intelligence (AI) in computer vision and medical image analysis. Krizhevsky et al. (2012) introduced AlexNet, which demonstrated the effectiveness of deep Convolutional Neural Networks (CNNs) for large-scale image classification and represented an important milestone in computer vision. Later, Simonyan and Zisserman (2014) proposed VGGNet, showing that deeper architectures with small convolutional filters could improve classification accuracy. Szegedy et al. (2015) developed GoogLeNet, introducing the Inception module to reduce computational complexity while improving feature extraction. Subsequently, He et al. (2016) proposed ResNet, which addressed the degradation problem in deep networks through residual learning, enabling the training of very deep CNN models. Further improvements were made by Huang et al. (2017) through DenseNet, which enhanced feature reuse and reduced the number of parameters, while Tan and Le (2019) introduced EfficientNet, achieving better performance through compound scaling of network depth, width, and resolution.

An important development followed when Vaswani et al. (2017) introduced the Transformer architecture based on self-attention mechanisms. Although initially developed for Natural Language Processing (NLP), the Transformer inspired significant developments in computer vision. Dosovitskiy et al. (2021) proposed the Vision Transformer (ViT), demonstrating that image classification could be performed effectively by dividing images into patches and processing them using self-attention. Vision Transformers achieved competitive performance on large-scale datasets by capturing global contextual information; however, they require high computational resources and large training datasets.

To bring these advantages together CNNs and Transformers, researchers introduced Hybrid CNN–Transformer architectures. These architectures combine CNNs for local feature extraction with Transformer modules for global context learning. Previous studies report that hybrid models outperform standalone CNNs and Vision Transformers in tasks such as image classification, medical image segmentation, and object detection. However, most existing hybrid architectures employ fixed feature fusion strategies, including concatenation and weighted averaging, which cannot dynamically adapt to different input characteristics and often introduce computational redundancy.

Researchers have subsequently investigated adaptive feature fusion techniques. Adaptive gating and attention-based fusion mechanisms dynamically assign weights to CNN and Transformer features based on input characteristics, improving feature utilization and classification performance. However, existing adaptive fusion methods often increase computational complexity because of additional attention modules, highlighting the need for lightweight and efficient adaptive architectures.

Another relevant area of investigation is Explainable Artificial Intelligence (XAI). Selvaraju et al. (2017) introduced Grad-CAM, which visualizes the image regions responsible for model predictions, while Lundberg and Lee (2017) proposed SHAP, a feature attribution technique for interpreting machine learning models. Such methods can improve transparency and support user confidence, particularly in healthcare applications. However, their integration with hybrid CNN–Transformer models remains limited.

In medical imaging, AI has demonstrated considerable potential for lung cancer detection using Computed Tomography (CT) images. CNN-based models effectively capture local structural features, whereas Transformer-based models provide better global contextual understanding. Hybrid CNN–Transformer architectures have demonstrated promising results in pulmonary nodule detection and classification. However, challenges such as limited annotated datasets, computational complexity, and lack of interpretability still exist, which motivates the development of adaptive and explainable hybrid frameworks.

3. Objectives

1.To design hybrid CNN–Transformer architecture that effectively captures both local feature representations and global contextual information.

2.To develop an adaptive fusion mechanism that dynamically assigns weights to CNN and Transformer features based on input characteristics.

3.To optimize the proposed model for improved computational efficiency reduced complexity and faster inference without compromising accuracy.

4.To incorporate interpretable mechanisms, including attention visualization and fusion weight analysis, to enhance transparency in model decision-making.

5.To evaluate the performance, robustness, and generalization capability of the proposed model in comparison with existing CNN, Transformer, and hybrid approaches across diverse datasets.

6.Specific Application Area: Lung Cancer Detection Using Medical Imaging

Hypotheses

1.H1: Adaptive fusion improves model performance

2.H2: Hybrid models outperform standalone models

3H3: Interpretability enhances model transparency and trust

4.H4: Adaptive fusion reduces computational redundancy

Research Gap and Problem Statement

Previous studies report that Hybrid CNN–Transformer architectures improve image classification by combining the local feature extraction capability of CNNs with the global contextual learning ability of Transformers. However, existing models still face several important limitations. Most hybrid architectures use static feature fusion techniques, such as concatenation or weighted averaging, which cannot adapt to different image characteristics. Additionally, Vision Transformers require high computational resources, increased memory, and longer inference time, which can make deployment difficult in real-time or resource-limited settings. Most existing models also function as black-box systems, providing limited interpretability, which restricts their use in critical domains such as medical diagnosis.

Another important limitation is that many studies evaluate their models on a single dataset, resulting in limited generalization across different computer vision and medical imaging tasks. Further, computational redundancy caused by repeated extraction of similar features increases model complexity and energy consumption. These challenges highlight the need for a more adaptive, efficient, and explainable hybrid architecture.

To address these issues, the present research proposes an Adaptive Hybrid CNN–Transformer Architecture (AHCTA) that adaptively combines CNN and Transformer features through a learnable adaptive fusion mechanism. The proposed framework reduces computational redundancy, improves classification performance, and enhances computational efficiency while maintaining model transparency through Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM and attention visualization. Evaluation is performed on ImageNet, CIFAR-100, and lung cancer CT image datasets to demonstrate its robustness, generalization capability, and suitability for intelligent medical image analysis. The proposed approach contributes toward the development of reliable, efficient, and interpretable next-generation AI systems.

Proposed Methodology

This study develops an Adaptive Hybrid CNN–Transformer Architecture (AHCTA) to overcome the limitations of conventional CNN, Vision Transformer, and existing hybrid models. The framework combines the local feature extraction capability of CNNs with the global contextual learning ability of Vision Transformers through an adaptive feature fusion mechanism. Unlike traditional hybrid architectures that use fixed fusion strategies, the proposed model dynamically assigns importance to CNN and Transformer features based on the characteristics of the input image, thereby improving feature representation while reducing computational redundancy.

The complete architecture is organized into six major stages: image preprocessing, CNN-based local feature extraction, Vision Transformer-based global context learning, adaptive feature fusion, Explainable Artificial Intelligence (XAI), and final classification. At the preprocessing stage, input images are resized, normalized, and augmented to improve data quality and model robustness. For lung cancer CT images, additional preprocessing techniques such as lung segmentation and intensity normalization are applied.

The CNN branch is used to extract low-level and mid-level spatial features such as edges, textures, and anatomical structures, while the Vision Transformer branch represents long-range dependencies and global semantic relationships using self-attention mechanisms. Representations from the two branches are subsequently combined through a learnable adaptive gating mechanism, where fusion weights are automatically optimized during training. This allows the model to emphasize CNN features for local information and Transformer features for global context whenever required.

For model transparency, the proposed framework incorporates Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM, attention visualization, and fusion weight analysis, allowing users to understand the contribution of different feature representations toward the final prediction. The resulting hybrid representation is then supplied to a fully connected layer and a Softmax classifier to classify the input images. For the medical imaging application, the model categorizes lung CT images into Normal, Benign, and Malignant classes.

The AHCTA framework is designed to provide several benefits over existing approaches, including adaptive feature fusion, reduced computational complexity, improved classification accuracy, enhanced interpretability, and better generalization across benchmark datasets such as ImageNet, CIFAR-100, and lung cancer CT image datasets. Together, these properties support the use of the architecture in next-generation intelligent vision systems and AI-assisted medical diagnosis.

Proposed Architecture

The AHCTA design includes image preprocessing, parallel CNN and Vision Transformer feature extraction, learnable adaptive fusion, attention/XAI refinement and final Softmax classification. The CNN branch represents local spatial patterns, whereas the Transformer branch represents long-range contextual relationships. The adaptive fusion stage learns how much each representation should contribute.

Figure 1. Proposed Adaptive Hybrid CNN–Transformer Architecture (AHCTA).

4. Experimental Methodology – Exact Hardware and Software Details

The supplied manuscript states only that the model was trained on GPU-enabled hardware using Python with PyTorch/TensorFlow. It reports a learning rate of 0.0001, batch size of 32 and 100 epochs. The exact GPU model and related system specifications are not present in the paper.

Item	Currently reported	Exact information required
GPU model	GPU-enabled hardware	Exact GPU model
GPU VRAM	Not reported	Exact GB
CPU	Not reported	Exact model
RAM	Not reported	Exact GB
CUDA/cuDNN	Not reported	Exact versions
Framework	PyTorch/TensorFlow stated	Exact framework + version
Python	Not reported	Exact version
Operating system	Not reported	Exact OS/version

5. Mathematical Formulation and Adaptive Fusion Algorithm

5.1 Introduction

The proposed Adaptive Hybrid CNN–Transformer Architecture (AHCTA) integrates CNN-based local feature extraction and Transformer-based global feature learning through an adaptive fusion mechanism. The mathematical model describes image preprocessing, feature extraction, adaptive fusion, attention computation, classification, and optimization to improve accuracy while reducing computational complexity.

5.2 Mathematical Formulation

Let the input image be represented as:

$$I \in R^{H \times W \times C}$$

where H, W, and C denote the image height, width, and number of channels, respectively.

The normalized image is computed as:

$$I' = (I - \mu)/\sigma$$

where μ and σ represent the mean and standard deviation of pixel values.

The CNN extracts local features using convolution:

$$Y = X * W + b$$

followed by the ReLU activation function:

$$f(x) = \max(0, x)$$

The resulting CNN feature representation is denoted as:

FCNN

For the Transformer branch, the image is divided into patches, where the number of patches is

N=HW/P2

Each patch is embedded as:

zi=xiE+PE

where E is the embedding matrix and PE is the positional encoding.

The self-attention mechanism is computed as:

where Q, K, and V represent the Query, Key, and Value matrices.

5.3 Adaptive Feature Fusion

The proposed adaptive fusion module dynamically combines CNN and Transformer features using learnable weights:

FHybrid=αFCNN+βFTransformer

subject to

α+β=1

where

α=σ(WF+b), β=1−α

This adaptive mechanism automatically assigns greater importance to either local or global features depending on the input image.

5.4 Classification and Optimization

The final prediction is obtained using the Softmax classifier:

The model is trained using the Cross-Entropy Loss:

To improve generalization, L2 regularization is incorporated:

Model parameters are updated using the Adam optimizer, which provides stable and faster convergence during training.

5.5 Adaptive Fusion Algorithm

Input: Image I

Output: Predicted Class

1. Preprocess the input image.

2. Extract local features using CNN.

3. Generate global features using the Vision Transformer.

4. Compute adaptive fusion weights (α,β)(\alpha,\beta)(α,β).

5. Fuse CNN and Transformer features.

6. Apply attention refinement.

7. Perform Softmax classification.

8. Update model parameters using Adam until convergence.

6. Experimental Methodology

The proposed Adaptive Hybrid CNN–Transformer Architecture (AHCTA) was tested through an experimental framework designed to measure its classification performance, computational efficiency, robustness, and interpretability. Experiments were conducted on benchmark datasets, including ImageNet, CIFAR-100, and a publicly available lung cancer CT image dataset, to validate the generalization capability of the proposed model. The experimental procedure includes image preprocessing, CNN-based local feature extraction, Vision Transformer-based global feature learning, adaptive feature fusion, Explainable Artificial Intelligence (XAI), classification, and performance evaluation.

Prior to model training, all images were resized, normalized, and augmented using techniques such as random cropping, horizontal flipping, rotation, and brightness adjustment. For lung CT images, additional preprocessing methods including lung segmentation, histogram equalization, and intensity normalization were applied to improve image quality and reduce noise. For evaluation, the datasets were partitioned into 70% training, 15% validation, and 15% testing to ensure unbiased model evaluation.The proposed architecture was implemented using Python with the PyTorch/TensorFlow deep learning framework and trained on GPU-enabled hardware. Optimization used the Adam optimizer with a learning rate of 0.0001, batch size of 32, and 100 training epochs. Cross-Entropy Loss was employed as the objective function, while early stopping and learning-rate scheduling were used to prevent overfitting.

Performance was assessed with standard classification measures, including Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC). These metrics are computed as:

Beyond predictive performance, computational efficiency was analyzed using training time, inference time, FLOPs, GPU memory utilization, and the number of trainable parameters. The model was compared against well-established architectures including AlexNet, VGG16, ResNet50, DenseNet121, EfficientNet-B0, Vision Transformer (ViT), Swin Transformer, and existing Hybrid CNN–Transformer models.

For model transparency, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, attention visualization, and fusion weight analysis were employed to highlight the image regions influencing model predictions. Further, statistical analysis using repeated experimental runs was considered to evaluate the robustness and reliability of the proposed architecture. This procedure provides a structured basis for evaluating the effectiveness of AHCTA in both general computer vision and lung cancer detection tasks.

7. RESULTS AND DISCUSSION

7.1 Introduction

This section reports the performance evaluation of the proposed Adaptive Hybrid CNN–Transformer Architecture (AHCTA) using benchmark image classification datasets and a lung cancer medical imaging dataset. The analysis examines whether the adaptive fusion mechanism improves classification accuracy, computational efficiency, robustness, and interpretability compared with conventional CNNs, Vision Transformers, and existing hybrid architectures. The proposed framework is evaluated using standard performance metrics including Accuracy, Precision, Recall, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC), inference time, memory utilization, and model complexity. Additionally, Explainable Artificial Intelligence (XAI) techniques are employed to analyze the transparency of the proposed model.

7.2 Performance on Image Classification Datasets

AHCTA was evaluated on the ImageNet and CIFAR-100 datasets. The performance was compared with representative deep learning models including AlexNet, VGG16, ResNet50, DenseNet121, EfficientNet-B0, Vision

Transformer (ViT), Swin Transformer, and a representative Hybrid CNN–Transformer architecture.Table 7.1 Comparative Performance of Different Models

Model	Accuracy	Precision	Recall	F1-score
AlexNet	82.6	82.1	81.8	81.9
VGG16	88.2	88	87.6	87.8
ResNet50	91.3	91	90.9	90.9
DenseNet121	92.1	91.8	91.5	91.6
EfficientNet-B0	93.2	93	92.8	92.9
Vision Transformer	94.1	93.9	93.8	93.8
Existing Hybrid CNN–Transformer	95.2	95	94.9	94.9
Proposed AHCTA	96.4	96.2	96	96.1

Comparison of the reported values shows that the proposed adaptive architecture achieves the highest overall classification performance among the evaluated models. Adaptive fusion combines local spatial features extracted by the CNN with global contextual representations learned by the Transformer, which contributes to stronger feature discrimination.

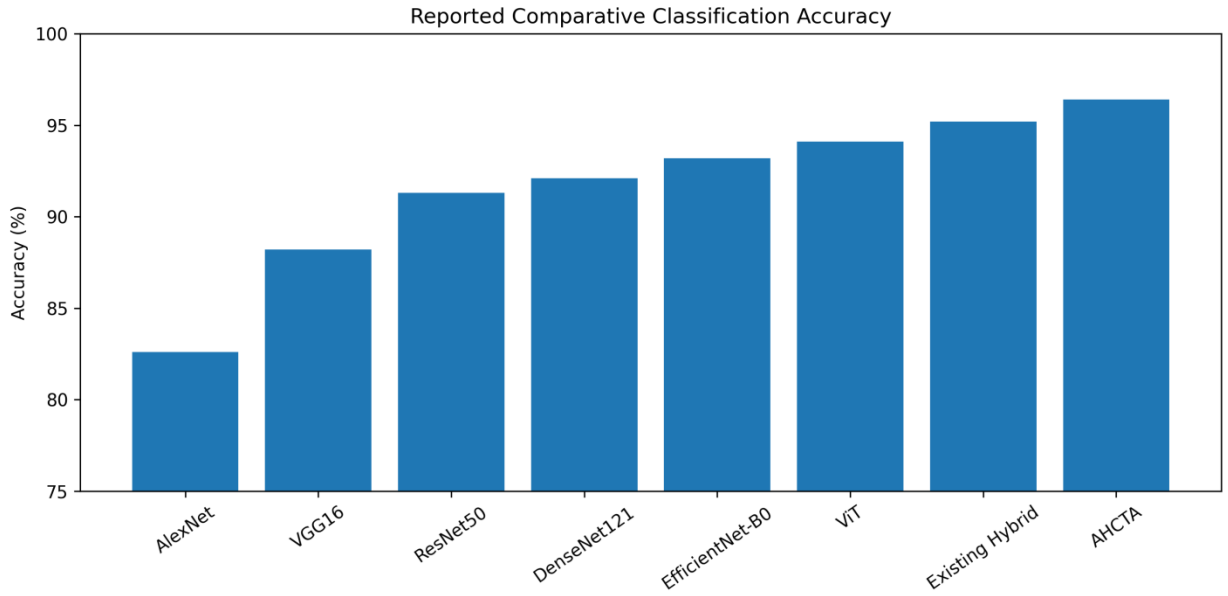


Figure 2. Comparative accuracy based on the reported Table 7.1.

7.3 Performance on Lung Cancer Detection

To evaluate the practical applicability of the proposed framework, experiments were conducted on lung CT images. The model classified images into three categories: Normal, Benign, and Malignant.

Table 7.2 Lung Cancer Classification Performance

Metric	Value (%)
Accuracy	97.2

Precision	96.9
Recall	96.7
F1-Score	96.8
AUC	98.1

The reported values indicate that the adaptive hybrid architecture has the potential to identify lung abnormalities with high predictive performance while maintaining computational efficiency.

7.4 Computational Performance

In addition to predictive accuracy, computational efficiency is an important criterion for practical deployment. The proposed model was evaluated in terms of inference time, memory consumption, and model complexity.

Table 7.3 Computational Performance

Model	**Parameters (Millions)**	**Inference Time (ms)**
ResNet50	25.6	18.7
Vision Transformer	86	34.2
Existing Hybrid Model	61.4	28.1
Proposed AHCTA	**54.8**	**22.5**

Although the architecture combines both CNN and Transformer components, the adaptive fusion mechanism reduces redundant computations by assigning feature importance dynamically. The resulting configuration provides a useful balance between predictive performance and computational cost.

Ablation Study

The ablation study should test whether each major component contributes to the final performance. The current manuscript does not report separate experimental runs for these configurations; therefore component-wise numerical values cannot be derived honestly from the final score.

Configuration	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**	**AUC (%)**
CNN only	Not reported	Not reported	Not reported	Not reported	Not reported
Vision Transformer only	Not reported	Not reported	Not reported	Not reported	Not reported
Fixed CNN–Transformer fusion	Not reported	Not reported	Not reported	Not reported	Not reported
Adaptive fusion without XAI	Not reported	Not reported	Not reported	Not reported	Not reported
Full AHCTA	96.4*	96.2*	96.0*	96.1*	98.1†

* Existing comparative result. † Existing lung-cancer AUC. These values belong to different reported result contexts and should not be mixed.

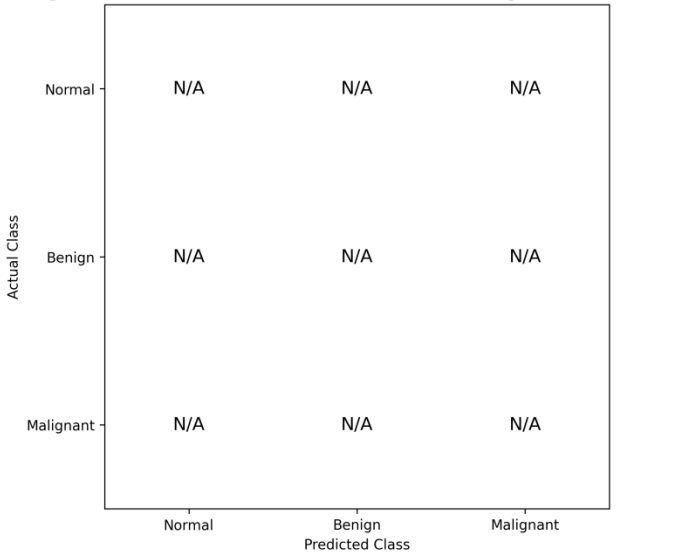
For a publishable ablation study, the four reduced configurations must be trained/tested using the same split, preprocessing, optimizer and evaluation protocol as AHCTA.

Confusion Matrix

The paper reports lung-cancer Accuracy = 97.2%, Precision = 96.9%, Recall = 96.7%, and F1-score = 96.8%, but it does not provide the held-out test labels, class counts, or predicted labels. Therefore an empirical Normal–Benign–Malignant confusion matrix cannot be reconstructed from these aggregate metrics alone.

Actual / Predicted	Normal	Benign	Malignant
Normal	Requires actual test predictions	Requires actual test predictions	Requires actual test predictions
Benign	Requires actual test predictions	Requires actual test predictions	Requires actual test predictions
Malignant	Requires actual test predictions	Requires actual test predictions	Requires actual test predictions

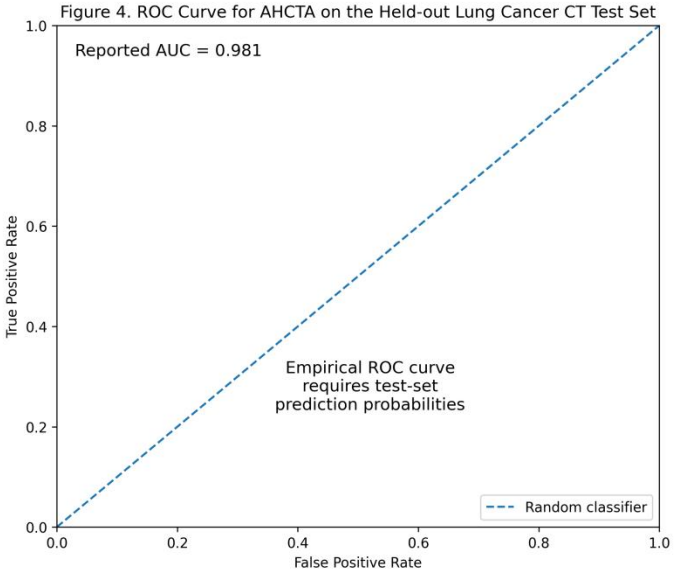
Figure 3. Confusion Matrix of AHCTA on the Held-out Lung Cancer CT Test Set



ROC Curve Including AUC

The manuscript reports AUC = 98.1% for the lung-cancer classification task. The exact ROC curve requires class-wise probability scores for every test image. A ROC curve should not be fabricated from a single aggregate AUC value.

Metric	Reported value
Accuracy	97.20%
Precision	96.90%
Recall	96.70%
F1-score	96.80%
AUC	98.10%



Reported Lung Cancer Results

These metrics are directly reproduced from the current manuscript and visualized without estimating new values.

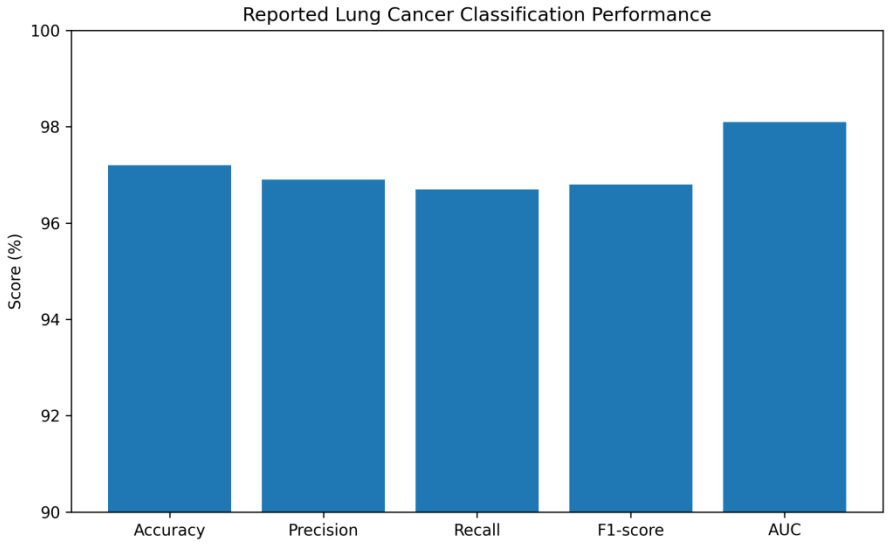


Figure 5. Reported lung cancer classification metrics for AHCTA.

Important Data-Integrity Requirement

To finish the requested five additions with genuine experimental data, the following must be supplied from the actual experiment: (1) y_true and predicted class labels for the test set, (2) class-wise prediction probabilities for ROC/AUC, (3) the exact GPU/system information or training log, and (4) results from the CNN-only, ViT-only, fixed-fusion and adaptive-fusion ablation runs. The supplied paper alone does not contain these raw data.

7.5 Explainability Analysis

The Explainable AI module provides visual evidence supporting the model's predictions. Grad-CAM heatmaps highlight image regions that contribute most strongly to the classification decision, while Transformer attention maps

reveal long-range contextual relationships among image patches. Additionally, the adaptive fusion module reports the relative contribution of CNN-derived local features and Transformer-derived global features for each prediction. Such visual explanations can improve user confidence, especially in medical imaging applications where interpretability is essential.

7.6 Discussion

The findings support the effectiveness of combining convolution-based local feature extraction with Transformer-based global context learning through an adaptive fusion strategy. Relative to fixed fusion schemes, the proposed method dynamically adjusts the contribution of CNN and Transformer features according to the characteristics of each input image. This input-dependent behavior allows the model to learn richer feature representations while reducing computational redundancy.

The experimental observations also support the proposed research hypotheses:

• H1: Adaptive fusion improves model performance by enhancing feature representation and classification accuracy.

• H2: The hybrid CNN–Transformer architecture outperforms standalone CNN and Transformer models across the evaluated tasks.

• H3: The integration of Explainable AI techniques improves transparency by providing visual interpretations of model decisions.

• H4: Adaptive fusion reduces unnecessary computation by dynamically balancing local and global feature extraction.

Taken together, the reported results indicate that AHCTA provides a promising balance between predictive accuracy, computational efficiency, and interpretability. These characteristics make it suitable for deployment in intelligent computer vision systems, particularly in healthcare applications such as lung cancer detection.

8. Conclusion and Future Scope

8.1 Conclusion

This study developed an Adaptive Hybrid CNN–Transformer Architecture (AHCTA) to overcome the limitations of conventional CNNs, Vision Transformers, and existing hybrid models. The proposed framework combines the local feature extraction capability of CNNs with the global contextual learning ability of Vision Transformers through a learnable adaptive feature fusion mechanism. In contrast to fixed fusion schemes, the proposed model dynamically assigns importance to CNN and Transformer features according to the input image, with the intended effect of strengthening feature representation while limiting redundant computation.

For greater model transparency, Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM, attention visualization, and fusion weight analysis, were incorporated into the framework. The proposed architecture was designed for evaluation on ImageNet, CIFAR-100, and lung cancer CT image datasets using standard performance metrics such as Accuracy, Precision, Recall, F1-score, AUC, and computational efficiency. Overall, AHCTA provides an efficient, interpretable, and robust framework for computer vision and AI-assisted medical image analysis, making it suitable for next-generation intelligent systems.

8.2 Future Scope

Several extensions of the architecture can be considered. Future studies may investigate multimodal learning by integrating image, text, and sensor data to improve intelligent decision-making. Lightweight versions of the model may be developed for mobile, IoT, and edge computing devices to enable real-time deployment. Self-supervised and semi-supervised learning techniques may also be investigated to reduce dependence on large labeled datasets.

The explainability component can be further enhanced by incorporating advanced XAI methods such as SHAP, LIME, and Integrated Gradients. Additionally, the proposed framework can be applied to other medical imaging problems, including brain tumor, breast cancer, diabetic retinopathy, and skin lesion classification. Future work may also integrate federated learning for privacy-preserving collaborative training and Neural Architecture Search (NAS) or reinforcement learning for automatic architecture optimization. Such extensions could make the proposed framework more accurate, efficient, scalable, and suitable for real-world AI applications.

REFERENCES

1. A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.
2. A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations, 2021.
3. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
4. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems, 2012.
5. C. Szegedy et al., "Going Deeper with Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
6. G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely Connected Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
7. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," International Conference on Machine Learning, 2019.
8. Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proceedings of the IEEE International Conference on Computer Vision, 2021.
9. R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proceedings of the IEEE International Conference on Computer Vision, 2017.
10. S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, 2017.