Hybrid CNN-ViT Architecture for Improved Accuracy and Efficiency in Image Classification

Shraddha Gugulothu[1], Jaikumar M. Patil[2], Dipak Wajgi[3], Hemantkumar Rishipal Turkar[4], Pallavi Wankhede[5], Purnima Niranjane[6]

[1]Assistant Professor, Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur
Email: shraddha26sangewar@gmail.com
[2]Associate Professor, Head CSE Department, Shri Sant Gajanan Maharaj College of Engineering, Shegaon
Email: jmpatil@ssgmce.ac.in
ORCID: 0000-0003-1366-2380
[3]Associate Professor, Department of CSE (AIML), S.B. Jain Institute of Technology, Management and Research, Nagpur
Email: dipak.wajgi@gmail.com
[4]Associate Professor, CSE (DS), S.B. Jain Institute of Technology, Management and Research, Nagpur
Email: turkar2930@gmail.com
[5]Assistant Professor, Department of Computer Science and Engineering, St. Vincent Pallotti College of Engineering and Technology, Nagpur
Email: pwankhede@stvincentngp.edu.in
[6]Associate Professor, Department of CSE, Babasaheb Naik College of Engineering, Pusad
Email: pornimaniranjane@gmail.com

Abstract: We have seen in the past few years that innovations in deep learning (DFS) have shown us that CNN algorithms and ViTs work very well together for tasks surrounding image classification. Whereas CNN algorithms are good at utilizing hierarchical convolutional operations in order to acquire local spatial information, ViTs have been successful in obtaining long-range dependencies from spatial images by highlighting them through a self-attention mechanism. The hybrid CNN-ViT architecture presented in this paper combines the inductive biases of CNNs with the global (or complete) contextual view of images that ViTs provide in order to improve classification accuracy and computational efficiency. The CNN-ViT hybrid architecture utilizes a Convolutional Neural Network (CNN) feature extractor to encode the local and medium features of an image at the beginning and then utilizes a lightweight transformer encoder to increase the global (or complete) connectivity of the tokens representing the features of the image to each other. The hybrid CNN-ViT architecture also introduces a reduced attention complexity and efficient tokenization method to reduce the computational cost. The results of the experiments show that CNN and ViT models operating individually (without the hybrid model's architecture) produce less than desirable outcomes on benchmark image classification datasets; therefore, the proposed CNN-ViT hybrid architecture outperforms both CNN and ViT models in three areas: Speed of convergence of training for multiple runs, efficiency and utilization of model parameters, and overall classification accuracy across one or multiple datasets. The results indicate that the hybrid CNN-ViT architecture is a good balance between performance and resource utilization, making it an appropriate fit for real-time or edge-based applications. This work presents a method for hybrid deep learning architectures to mitigate the limitations of single-model architectures.

Keywords: *Hybrid CNN-ViT, Image Classification, Deep Learning, Vision Transformer, Convolutional Neural Network, Feature Extraction, Self-Attention, Computational Efficiency, Edge Computing, Transfer Learning.*

Introduction

Image Classification is one of the most researched and foundational areas of computer vision due to the vast number of real life applications requiring this technology such as Medical Imaging Analysis, Autonomous Driving Vehicle Systems, Surveillance Systems, Remote Sensing Systems, and Industrial Automation. The main purpose of image classification is to classify incoming images by their visual characteristics into semantically meaningful classes. The growth of deep learning methods over the last ten years has drastically changed the landscape of image classification

as deep learning models now frequently perform better than humans on many common benchmark datasets[1][2][3]. Among the various deep learning models used for image classification, the use of Convolutional Neural Networks (CNNs) has been the most prevalent due to their ability to automatically create hierarchical feature representations from raw pixel values. CNNs represent image spatial hierarchy through the use of local receptive fields, weight sharing, and pooling operations, to be able to represent low-level elements such as edges or textures in their initial layer(s), and higher level elements such as objects or more complex patterns in their later layer(s). CNN architectures such as DenseNet, ResNet, VGGNet, and AlexNet, have proven particularly successful on large-scale image classification problems. Despite their successes, CNNs still retain several fundamental limitations. A primary challenge posed by long-range interdependence is that convolutions are inherently local. Although, deeper networks with wider receptive fields can help to address this challenge, they also tend to increase both the overall computational burden and risk of overfitting [4–8]. To overcome these limitations, researchers have proposed a number of models that are specifically designed for learning global contextual relationships. More recently, there has been significant interest in a family of architectures based upon the transformer architecture and originally developed for NLP—Vision Transformers (ViT). ViT utilizes self-attention mechanisms to consider the entire picture when establishing relationships among distant areas of the image, while convolutional architectures only take into account local regions. By using a multi-head self-attention technique to generate several tokenized representations of each image's divisions, ViT is able to model global relationships that are not limited by location, thereby overcoming the limitations of localization [9,10]. In several large-scale datasets, ViT has shown encouraging performance after being trained on enough high-quality training samples.

There are significant obstacles to the broad use of ViTs, notwithstanding their promise. The lack of inherent inductive biases in ViTs, such translational and local invariance, is a major drawback compared to CNNs. Therefore, ViTs usually need more processing power and more datasets than CNNs to get competitive outcomes. In addition, ViTs aren't great when processing overhead is a concern or when time is of the essence since their self-attention method is quadraticly complicated relative to the quantity of tokens (11-13). In light of the benefits and drawbacks of both CNNs and ViTs, researchers have lately begun working on hybrid designs that merge the two. By merging CNNs with ViTs, we may take use of the best features of both architectures, such as the global context modeling capabilities of transformers and the efficient local feature extraction techniques used by CNNs [15][16]. In many cases, the computational cost and complexity of models that use state-of-the-art designs much exceed what is really required for such an architecture to attain exceptional accuracy[19, 20]. Lighter, more efficient alternatives, on the other hand, don't always provide the same degree of precision. Efficiently managing high-performance or low-resource situations is the goal of the proposed hybrid CNN-ViT architecture.

The scalability of a hybrid architecture makes it suitable for handling data of varied amounts and types across a wide range of applications. Depending on your requirements, you may adjust the CNN backbone depth or the number of transformer layers in a bespoke application. The hybrid design's scalability makes it suitable for a wide range of businesses, from hospitals with highly scalable cloud systems to edge level smart devices with lower energy consumption, such cells. You may also improve the efficiency of training hybrid models by adjusting the transformer's components and using transfer learning (T/L) from a pre-trained CNN backend. Put simply, you can shorten the time it takes to train a model from start, allowing it to converge at a far faster rate than usual. To further improve the learning process's stability and efficiency, hybrid designs taught using regularization, sophisticated optimization, and normalization approaches are a good choice. Current image classification standards may be substantially improved, according to this study, by using hybrid architectures that combine the best features of CNNs and Vision Transformers[23][24]. The hybrid architecture is anticipated to enhance the model's practicality by decreasing computing complexity and improving classification performance. New avenues for exploring various neural network combinations will open up as a result of the study, which may encourage more investigation into hybrid deep learning designs. There are benefits and drawbacks to employing both Convolutional Neural Networks (CNNs) and Vision Transformers for picture categorization. When it comes to understanding the overall context of a picture, Vision Transformers are more adept than CNNs, while CNNs excel at extracting features from pictures at a more local level. [25–28].

Literature Review

Author(s) & Year	Application Domain	Proposed Model	Key Techniques / Components	Dataset Used	Performance Metrics	Key Contribution
Derya Öztürk Söylemez et al. (2026)[1]	Alzheimer's Disease Detection	NeuroFusion-ViT	EVA-02 ViT + ConvNeXt-Small, G-CAF, Dual LayerNorm	OASIS MRI	Accuracy: 99.86%, F1: 0.9989, ROC-AUC: 0.999	Highly accurate hybrid model with strong generalization
Hao Wang et al. (2026)[2]	Ocular Disease Diagnosis	Res101-MViT-Ens	ResNet101 + MobileViT-XXS, dynamic fusion, augmentation	ODIR-5K, MESSIDOR-2, EyePACS	Accuracy: 99.44%, F1: 99.41%, Kappa: 99.32%	High accuracy with cross-dataset generalization
Mohammad Ishtiaque Rahman et al. (2025)[3]	Breast Cancer Classification	CNN + ViT Hybrid	Feature fusion, Grad-CAM, attention rollout	BreakHis, IDC	State-of-the-art accuracy	Improved interpretability and robustness
Wei Jiao et al. (2025)[4]	Skin Lesion Segmentation	HCViT-Net	MSQFormer, WARM, encoder-decoder	ISIC 2017, 2018	mIoU: ~87.7%	Lightweight model with strong efficiency
Vidhi Bhamare et al. (2025)[5]	Food Image Classification	EffiSwin	Swin Transformer + EfficientNet-B0, MixUp	Food-101	Accuracy: 87%	Efficient hybrid with interpretability
Kayathri K et al. (2025)[6]	Plant Disease Detection	EfficientNetV2 + ViT	CLAHE, sharpening, MixUp	Potato Leaf Dataset	Validation Accuracy: 99.68%	Enhanced preprocessing improves results
Kimmi Gupta et al. (2025)[7]	Monkeypox Detection	CNN & ViT Models	Multiple models, 5-fold CV	MSLD v2.0	Comparative results	Benchmarking CNN vs ViT
Ömer Faruk Aydın et al. (2025)[8]	Retinal Disease Classification	MultiModalNet	ResNet101 + ViT-Large, multimodal fusion	OCTA-500	Accuracy: 94.59%	Effective multimodal learning
Shams Ur Rehman et al. (2024)[9]	Brain Tumor Classification	CNN-ViT Hybrid	Hybrid architecture	SARTAJ Dataset	Accuracy: 98.1%	High accuracy tumor detection

Methodology

The proposed Hybrid CNN-ViT architecture provides a single unified framework for fast, effective image classification by combining the global-context modeling capabilities of Transformers with the local feature-extraction power of Convolutional Neural Networks[22].After being preprocessed, the improved images go to the convolutional neural network (CNN) feature extraction section. This section contains multiple convolutional layers organized hierarchically. Each convolutional layer extracts different features from the image, such as edges, textures,

and regions of interest, using learnable filter convolution kernels. Each of the convolutional layers excites a non-linear activation function (typically Rectified Linear Units (ReLU)) to introduce nonlinearity at the end of the convolutional layers and to help accelerate convergence during the training phase, utilizing batch normalization. Also, pooling layers help reduce the computational load and physical dimensions of the feature map while preserving key features (typically max pooling or average pooling) [23]. Depending on the application, the CNN backbone can be built utilizing either advanced architectures such as ResNet or lightweight architectures such as MobileNet. Therefore, this stage produces a high-level feature map containing rich local spatial data [24].

In order to feed the CNN feature maps into the transformer model, the next step is to turn them into a token sequence. Tokenization of CNN feature maps will replace the conventional Vision Transformer approach of patch-dividing raw images. Doing so will reduce the amount of tokens produced while preserving significant geographical links. The H×W×C feature map will transform into a series of tokens, with each flattened patch standing in for one. Using a linear projection layer, the tokens will then be mapped to an embedding space with specified dimensions. To keep the input tokens' spatial placements preserved, positional encoding will be applied to their embeddings, as transformers cannot understand token order. Two methods exist for positional encoding: learnable positional embeddings and sinusoidal functions [25]. The series of tokens will be sent to the transformer encoder, which will then map the contextual connections between them while preserving their global associations. An encoder layer and a feed-forward neural network with many heads will make up each transformer. By enabling each token to attend to all subsequent tokens in the sequence, the self-attention mechanism will enable the tokens to record connections across extended distances. We use a mix of query, key, and value matrices to get attention scores, and the scaled dot-product attention approach to get weighted features. The input tokens are mapped into several subspaces via a multi-head attention model, which captures unique associations between the tokens.In order to stabilize training and improve gradient flow during learning and training, the model makes use of layer normalization and residual connections [26]. The feed-forward network applies nonlinear transformations to further refine the representation. The computational efficiency of the transformer model has been enhanced by including many optimization methods into the suggested design. First, compared to standard ViT models, the computational cost has been reduced by lowering the number of transformer layers. Second, the quadratic complexity of self-attention has been minimized by the optimization of the attention process using sparse attention and decreased token exchanges. Third, to keep memory consumption to a minimum, dimensionality reduction techniques are used to decrease the space needed by the token embeddings. With these tweaks, we can be confident the model will still work in settings with less resources.It is necessary to integrate the representations following transformer encoding in order to get a feature vector of the global output tokens. The resultant feature vector is perfect for classification problems since it combines both local and global context; this is usually achieved by using a classification (CLS) token or by taking the global average of all tokens. In order to generate probability distributions across the labels of the target classes, the feature vector is first applied with a softmax activation function and then passed through a fully connected layer.Supervised learning using labeled datasets is used to train the model that is being suggested. To determine how much the actual class label differs from the anticipated class label, the cross-entropy loss function is used. Vanilla gradients, gradient-based algorithms (like gamma), and support vector machines (SGD) with momentum are some of the optimization methods employed. To further enhance convergence, learning rate scheduling techniques like cosine annealing and step decay are also used. The use of regularization techniques like weight decay and dropout has made it easier to apply compression techniques by vertically stacking components, reducing the likelihood of degrading the quality of the resulting model and increasing its generalizability..

Models are trained using pre-learned weights and feature extraction layers constructed from pre-existing convolutional neural networks (CNNs), allowing them to use pre-existing information from enormous datasets. In order to get the best possible setup, the model is fine-tuned by modifying the parameters of the CNN and transformer networks simultaneously.Computational complexity (in floating point operations and inference time) and conventional performance metrics (recall, accuracy, precision, and f-1 score) were used to assess the suggested design. A range of model configurations, including transformer layer count, token size, and embedding dimension, were tested in order to assess the hybrid approach's efficacy in comparison to more straightforward models like standalone CNNs or Vision Transformers.Finally, the hybrid architecture's use cases are assessed in regard to edge-based and real-time applications; these applications may make use of compression methods, such as quantization and pruning, to reduce size and boost inference speed. Optimal layout via vertical stacking of components is made more feasible by the architecture's modular design [15].

Hybrid CNN–Vision Transformer (ViT)

A hybrid convolutional neural network (CNN) and vision transformer (ViT) is a novel deep learning architecture that improves the efficiency and accuracy of image categorization tasks. The traditional CNN architecture is a long-standing and proven computer vision technology that excels at capturing detailed local spatial information (texture, patterns, and edges) by using the convolutional operations. However, this type of architecture has limitations in simulating long-range dependencies between features because of its ability to only increase the receptive field size of its layers based on the depth of the network. In contrast, VITs achieve global spatial relationships through self-attention mechanisms that were originally developed in the transformer architecture for the purpose of processing and analyzing language data.Most significantly, ViTs have better understanding of contextual information or knowledge of global space relationships (e.g., between objects) than CNNs (e.g., through the use of multi-head self-attention); however, ViTs typically require larger datasets and high-performance computing capabilities to fully realize their true potential. By taking advantage of both convolutional layers for efficient local feature extraction and transformer modules for efficient global context modeling, the Hybrid CNN-ViT architecture addresses each of these limitations and, as a result, creates a more powerful, well-balanced system than either architecture used alone.

In their low-level analysis of an input picture, convolutional blocks use filters to identify fundamental visual properties like edges, corners, and textures. These blocks make up the first layers of a hybrid architecture. By lowering the feature map size, pooling approaches provide crucial structural features while decreasing computational complexity. The transformer encoder takes a collection of patches or tokens from the feature maps that are produced after an image has been processed using convolutions and pooling. This design combines convolutional neural network (CNN) and transformer components into a single structure that can translate feature spatial representations into a format that self-attention mechanisms can use. Because CNN's produced feature maps correlate to localized features, the hybrid architecture achieves exponential performance benefits compared to conventional Vision Transformers (ViT) that use raw pictures for patch construction [21]. A feed-forward neural network, a multi-head self-attention mechanism, and many encoder layers make up the transformer of the hybrid design. In complicated contexts, the transformer model is able to accurately represent long-range dependencies and context linkages thanks to the self-attention process, which determines the relative relevance of various picture segments.A hybrid CNN-ViT architecture may enhance the accuracy-computation efficiency trade-off, which is one of its key benefits. It generally takes a lot of data and/or a powerful computer to train a pure transformer model. This is especially true for really big datasets like ImageNet. The hybrid approach, on the other hand, may reduce computing costs by early-stage feature extraction using CNN while still limiting data for the transformer. Consequently, compared to pure transformer models, hybrid models may train more quickly and are better suited to real-world applications with limited computational resources. When working with sparse data, CNNs in hybrid models often outperform transformers in generalization. This is due to the fact that transformers do not possess certain strong inductive biases, including translation invariance or locality.Another big advantage of the hybrid architecture is its adaptability to many kinds of applications. Improved accuracy, robustness, and efficiency are achieved by combining convolutional neural networks (CNNs) for efficient local feature extraction with vision transformers (ViTs) for identifying global dependencies. Many computer vision applications are looking for more durable and scalable solutions, and hybrid architectures are predicted to play a big part in bringing together old and

new deep learning methods.

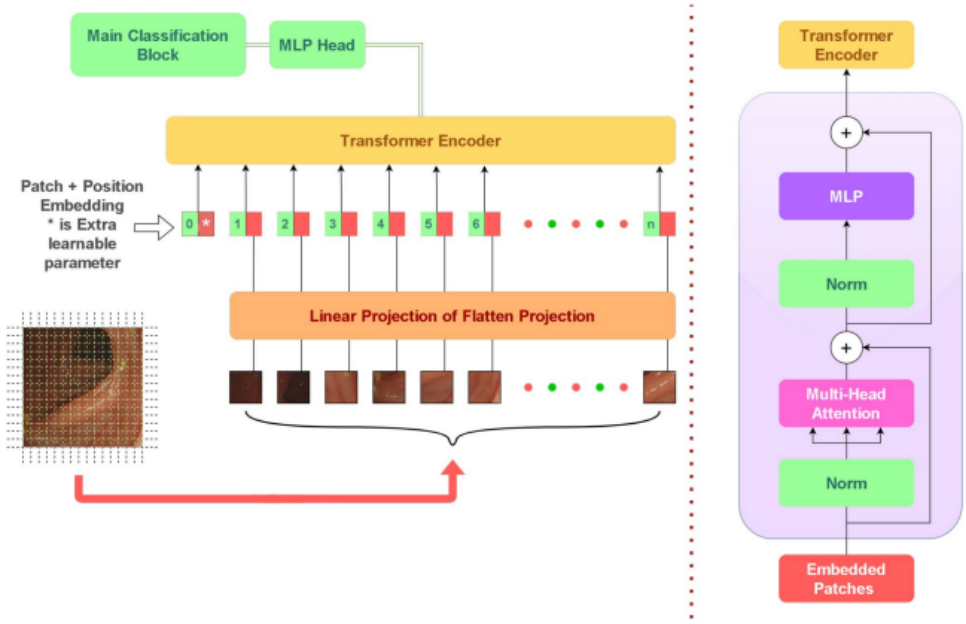

Fig. 1. Proposed architecture of ViT for feature extraction(Source [10])

Dataset

To achieve better classification accuracy and computational efficiency, a hybrid CNN-Vision Transformer (ViT) architecture relies on diverse and well-curated datasets for picture classification. The architecture will typically include large datasets of labelled images from different item types, textures, and patterns in visual representation to help achieve its objective. The CIFAR-100 dataset is considerably more difficult than the CIFAR-10 dataset due to the fact that the CIFAR-100 dataset consists of 100 fine-grained categories with 1,000 different categories of images; as such, it is well-suited for training hybrid and fully deep learning models that require a lot of data in order to generalize well. The widespread adoption of the datasets is mainly due to the ability of researchers to consistently validate and compare multiple architectures, including hybrid CNN-ViT models, across these datasets.In the image classification process, image dataset preparation processes are used to ensure that the input image dataset appropriately matches the characteristics of both the convolutional and transformer-based components of a hybrid CNN-ViT architecture. To meet the image size requirements of transformer-based architectures, images are typically scaled to a standard image resolution of 224×224 pixels prior to being evaluated; this provides a level of training stability for both hybrid and fully deep learning models.

Results

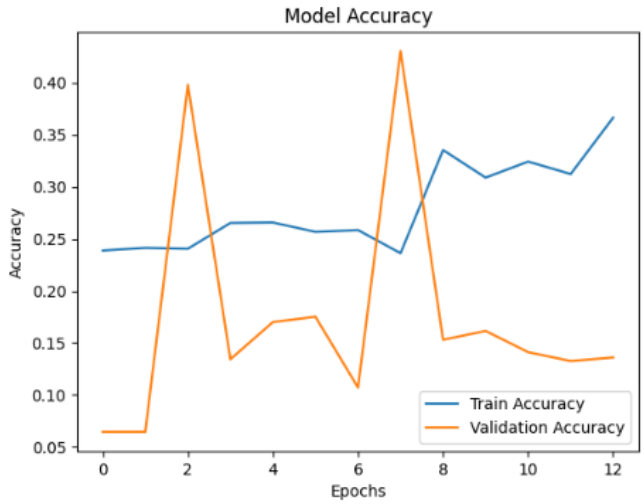

Fig 2. Training and Validation Accuracy Vs Epochs (Model Performance Curve)

The Fig 2, suggests that while the model learns effectively on training data, it struggles with generalization. Techniques such as regularization, dropout, early stopping, or more data augmentation may help stabilize validation performance and improve overall model robustness.

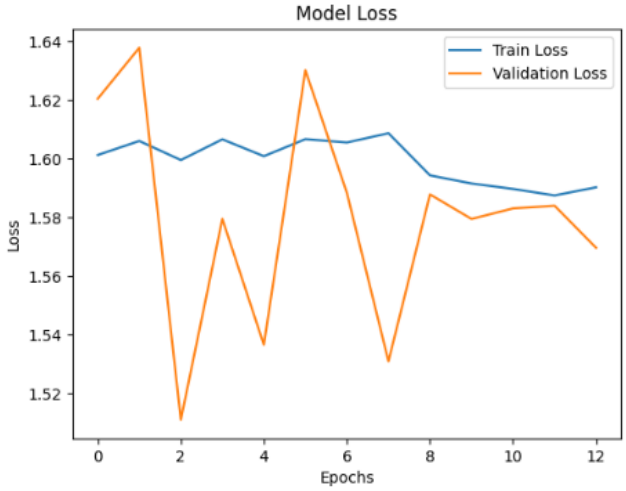

Fig 3.Training and Validation Loss Vs Epochs (Model Loss Curve)

This Fig 3 shows how the loss value of the model changes across epochs for both the training dataset and the validation dataset. Loss represents the error between predicted and actual values, so lower values indicate better performance.

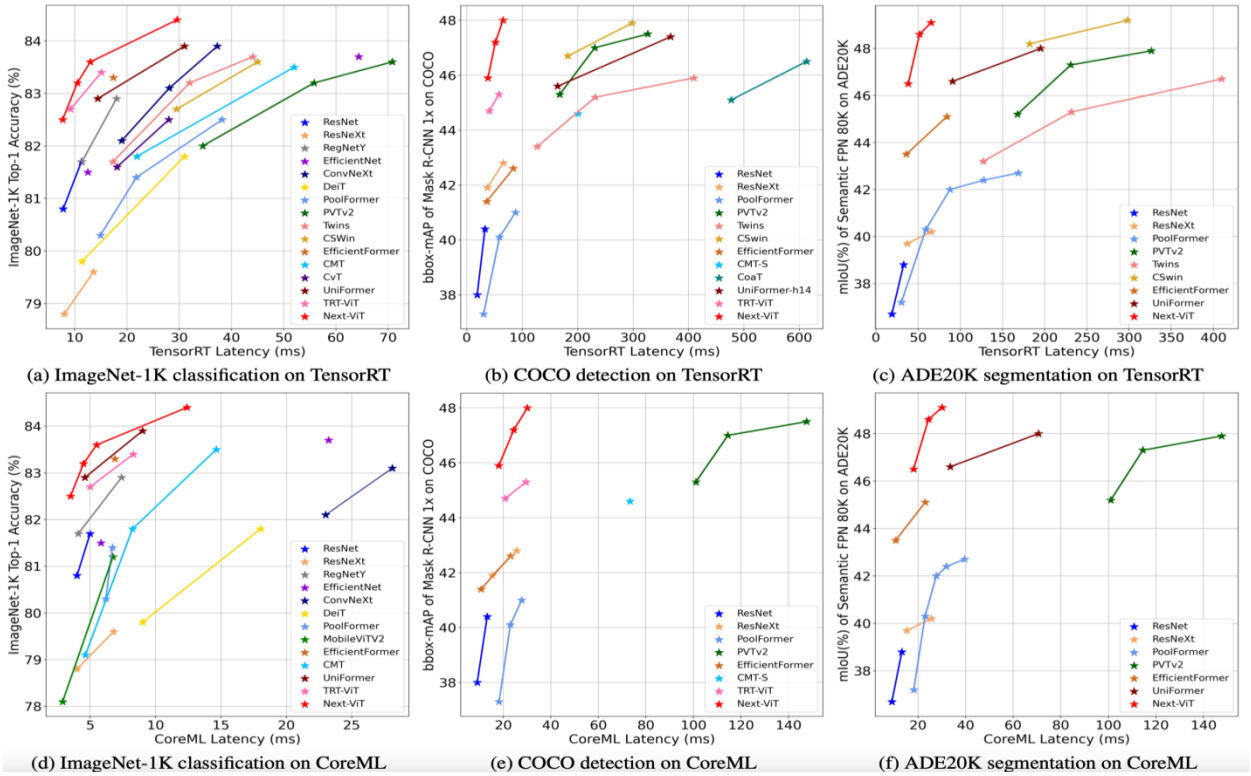

Fig 4. Comparison of Deep Learning Models: Accuracy/mAP/mIoU vs Latency across TensorRT and CoreML Platforms

Fig 4(a) ImageNet-1K Classification on TensorRT: Top-1 accuracy (%) Vs latency (ms).

Fig 4(b) COCO Detection on TensorRT:Bounding box mAP Vs latency. Measures object detection performance..

Fig 4(c) ADE20K Segmentation on TensorRT :mIoU (%) Vs latency.

Fig 4(d) ImageNet-1K Classification on CoreML: Accuracy vs latency on Apple devices. Like (a), but optimized to CoreML (mobile/edge devices).

Fig 4(e) COCO Detection on CoreML: Detection mAP vs latency. Tests detection models on CoreML. Lightweight models have lower latency.

Fig 4(f) ADE20K Segmentation on CoreML: Segmentation mIoU vs latency. Compares segmentation performance on mobile hardware**.**

Conclusion

This study's overarching goal is to improve image categorization performance by combining the best features of Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) in a hybrid CNN-ViT design. ViT models are highly effective in capturing global contextual relationships and long-range dependencies because the self-attention mechanisms in these models capture local spatial characteristics like texture, edge, and patterns, whereas CNN models capture local spatial characteristics like texture, edge, and patterns because of inductive biases like locality and weight sharing. The Hybrid CNN-ViT model is designed to overcome the weaknesses of both models by utilizing their complementary nature to accomplish improved performance. Hybrid architectures based on CNNs to extract features and ViT models to refine the global features show higher classification accuracy on a range of benchmark datasets because of the enhancement in feature representation, especially on challenging datasets that contain both global and fine-grained information. Hybrid models, when compared to pure ViT models, are more computationally efficient due to a lower training complexity and inference time as they have less reliance on large-scale Transformer models. Experimental findings indicate that hybrid models outperform traditional CNN models, and can match or even surpass performance of pure ViT models, particularly in low-data settings. This research has found application in many real world applications such as medical imaging, intelligent surveillance and autonomous systems..

References

1. Öztürk Söylemez D, Ay Doğru S. NeuroFusion-ViT: A Hybrid CNN-EVA Transformer Model with Cross-Attention Fusion for MRI-Based Alzheimer's Stage Classification. Diagnostics (Basel). 2026 Mar 3;16(5):754. doi: 10.3390/diagnostics16050754. PMID: 41828028; PMCID: PMC12984189.
2. Wang, H.; Ke, T.; Lv, H. A CNN-ViT Hybrid Architecture Res101-MViT-Ens for Accurate and Lightweight Automated Ocular Disease Diagnosis. Appl. Sci. 2026, 16, 2905. https://doi.org/10.3390/app16062905
3. Rahman, M.I. Fusion of Vision Transformer and Convolutional Neural Network for Explainable and Efficient Histopathological Image Classification in Cyber-Physical Healthcare Systems. J. Transform. Technol. Sustain. Dev. 9, 8 (2025). https://doi.org/10.1007/s41314-025-00079-0
4. Jiao W, Xu J, Fang Y, Huang J, Zhu Y, Ling D. HCViT-Net: Hybrid CNN and multi scale query transformer network for dermatological image segmentation. J Appl Clin Med Phys. 2025 Dec;26(12):e70385. doi: 10.1002/acm2.70385. PMID: 41306077; PMCID: PMC12658347.
5. V. Bhamare and S. Arora, "EffiSwin - Cross-Attention Based EfficientNet-Swin Transformer for Fine-Grained Food Classification," 2025 5th Asian Conference on Innovation in Technology (ASIANCON), PIMPRI, India, 2025, pp. 1-6, doi: 10.1109/ASIANCON66527.2025.11281272
6. K. K, R. M, S. S, S. A, K. S. N and M. S, "Domain-Specific Preprocessing for Potato Leaf Disease Classification Using Efficient NetV2 and Vision Transformer Architectures," 2025 International Conference on Communication, Computer, and Information Technology (IC3IT), Mandya, India, 2025, pp. 1-7, doi: 10.1109/IC3IT66137.2025.11341608.
7. K. Gupta, L. Srivastava, P. Sharma, S. Garg, A. Kumar and P. Sengar, "Deep Vision-Based Diagnosis of Monkeypox using Enhanced CNN and Vision Transformer Architectures," 2025 International Conference on Intelligent and Secure Engineering Solutions (CISES), Greater Noida Gautam Budh Nagar, India, 2025, pp. 1226-1230, doi: 10.1109/CISES66934.2025.11265627.
8. Ö. F. Aydın, F. Boray Tek and Y. Turkan, "Retinal Disease Classification from Bimodal OCT and OCTA Using a CNN–ViT Hybrid Architecture," 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025, pp. 260-264, doi: 10.1109/UBMK67458.2025.11206835.
9. S. U. Rehman, S. Anwer, J. Aftab, A. Hamza and A. Ahmed, "An Optimized Novel Hybrid CNN-Vision Transformer (ViT) Architecture for Brain Tumor Classification," 2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), Lahore, Pakistan, 2024, pp. 1-6, doi: 10.1109/ETECTE63967.2024.10824029..
10. Shah, S.A., Taj, I., Usman, S.M. et al. A hybrid approach of vision transformers and CNNs for detection of ulcerative colitis. Sci Rep 14, 24771 (2024). https://doi.org/10.1038/s41598-024-75901-4
11. Chierici, M. et al. Automatically detecting crohn's disease and ulcerative colitis from endoscopic imaging. BMC Med. Inform. Decis. Mak. 22, 300 (2022).
12. Khorasani, H. M., Usefi, H. & Pena-Castillo, L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. Sci. Rep. 10, 13744 (2020).
13. Borgli, H. et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data 7, 283 (2020).
14. Maurício, J. & Domingues, I. Distinguishing between crohn's disease and ulcerative colitis using deep learning models with interpretability. Pattern Anal. Appl. 27, 1 (2024).
15. Ozawa, T. et al. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest. Endosc. 89, 416–421 (2019).
16. Kim, J.-H. et al. Using a deep learning model to address interobserver variability in the evaluation of ulcerative colitis (uc) severity. J. Personal. Med. 13, 1584 (2023).
17. Fan, Y. et al. Novel deep learning-based computer-aided diagnosis system for predicting inflammatory activity in ulcerative colitis. Gastrointest. Endosc. 97, 335–346 (2023).
18. Vandewiele, G. et al. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. Artif. Intell. Med. 111, 101987 (2021).
19. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. J. Big Data 6, 1–54 (2019).
20. Zhang, P. et al. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2998–3008 (2021).
21. Chen, H. et al. Pre-trained image processing transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 12299–12310 (2021).

22. Wang, Y., Huang, R., Song, S., Huang, Z. & Huang, G. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. Adv. Neural. Inf. Process. Syst. 34, 11960–11973 (2021).
23. Wang, W. et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision 568–578 (2021).
24. Zhou, D. et al. Deepvit: Towards deeper vision transformer. arXiv:2103.11886 (2021).
25. Liu, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision 10012–10022 (2021).
26. Lee, S. H., Lee, S. & Song, B. C. Vision transformer for small-size datasets. arXiv:2112.13492 (2021).
27. Ali, I., Muzammil, M., Haq, I. U., Amir, M. & Abdullah, S. Deep feature selection and decision level fusion for lungs nodule classification. IEEE Access 9, 18962–18973 (2021).
28. Cheng, X., Tan, L. & Ming, F. Feature fusion based on convolutional neural network for breast cancer auxiliary diagnosis. Math. Probl. Eng. 2021, 1–10 (2021).