International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 2.102

Enhancing Medical Image Classification with Vision Transformers on Diverse Datasets

Aditya Dhar Dwivedi

Gautam Buddha University, School of Information & Communication Technology, Greater Noida, Uttar Pradesh, India
Email: *aditya13dwivedi[at]gmail.com*

Abstract: ***Medical image classification is essential for accurate diagnosis and effective treatment planning. This research investigates the implementation of MedViT, a robust Vision Transformer tailored for medical image analysis, and compares its performance against four models: StarterNet, TinyVGG, a standard Vision Transformer (ViT), and a Convolutional Neural Network (CNN). Evaluations conducted on PathMINST, a medical imaging dataset, and CIFAR - 10, a general - purpose image classification dataset, to assess model generalization.***

Keywords: MedViT, Vision Transformer, Medical Image Classification, PathMINST, CIFAR – 10

1. Introduction

Medical imaging plays a pivotal role in modern healthcare, enabling accurate diagnosis, efficient treatment planning, and effective patient monitoring. Over the years, advancements in computational methods have increasingly supported medical professionals by automating and enhancing image analysis. Among these methods, image classification serves as a fundamental task in interpreting medical data, where the goal is to classify medical images into predefined categories. Despite significant progress, medical image classification remains challenging due to issues such as data variability, noise, and the high precision required for clinical applications.

Traditional machine learning techniques, while useful, often fall short in handling the high - dimensional and complex nature of medical image data. Deep learning models, particularly Convolutional Neural Networks (CNNs), have emerged as a powerful alternative, achieving remarkable success across various domains, including medical imaging. CNNs excel in feature extraction and hierarchical representation learning but are often constrained by their limited ability to capture long - range dependencies within images. This limitation has driven the exploration of alternative architectures, such as Vision Transformers (ViTs), which offer a fundamentally different approach to image representation.

Vision Transformers, first introduced in the context of natural image classification, advantage self - attention mechanisms to model global relationships in image data. Unlike CNNs, which rely on local receptive fields, ViTs process entire images as sequences of patches, allowing them to learn both local and global patterns effectively. While ViTs have demonstrated state - of - the - art performance on large - scale datasets such as ImageNet, their application to medical imaging is still in its nascent stages. Medical images often possess unique characteristics, including high inter - class similarities, subtle distinctions, and limited labeled data, which necessitate tailored adaptations of ViTs for optimal performance.

MedViT, a domain - specific Vision Transformer, has been developed to address these challenges. By incorporating design modifications and training strategies suited to medical data, MedViT aims to bridge the gap between general - purpose ViTs and the specialized requirements of medical imaging. This research investigates the efficacy of MedViT in comparison with four alternative models—StarterNet, TinyVGG, a standard Vision Transformer, and a CNN—on two datasets: PathMINST, a specialized medical imaging dataset, and CIFAR - 10, a widely used benchmark for general image classification tasks. The inclusion of CIFAR - 10 enables an assessment of MedViT's generalization capabilities beyond the medical domain.

The objectives of this study are threefold: (1) to evaluate the performance of MedViT on medical and non - medical datasets, (2) to compare its accuracy and robustness with other prominent architectures, and (3) to analyze its potential for generalization and domain - specific adaptation. This paper provides a comprehensive overview of the datasets used, the preprocessing methods applied, the architectural details of the implemented models, and the experimental setup for training and evaluation. The findings contribute to the growing body of knowledge on Vision Transformers and highlight their promising role in advancing medical image classification.

2. Related Works

Takahashi et al. (2024) this study systematically reviews Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in medical image analysis. It highlights ViTs' ability to enhance classification across diverse datasets compared to CNNs, emphasizing the importance of pre - training for optimal performance in medical imaging tasks. The paper underscores ViTs' capacity to capture global relationships, improving diagnostic accuracy and generalization. By addressing challenges such as dataset variability and annotation scarcity, the study positions ViTs as transformative tools for medical image analysis, paving the way for advancements in precision medicine and automated diagnostics. De la Fuente et al. (2024) This paper explores synthetic data augmentation to enhance medical image classification for small and unbalanced datasets. Using class

- specific Variational Autoencoders (VAEs), the authors generate synthetic data to improve model generalizability and diagnostic accuracy. Unlike Vision Transformers, this approach focuses on addressing dataset limitations, ensuring robust training and performance. The study demonstrates significant improvements in accuracy and reliability, particularly for rare medical conditions. This method holds potential for resource - constrained scenarios, enabling better utilization of limited datasets in medical imaging tasks and highlighting an alternative to Transformer - based approaches.

Hu et al. (2024) the study introduces the MambaConvT model, a hybrid approach combining CNNs and Transformers for medical image classification. By leveraging CNNs' local feature extraction and Transformers' global context modeling, the model achieves superior performance across datasets. The paper reports enhancements in accuracy, precision, recall, F1 - score, and AUC ratings, showcasing its effectiveness in diverse medical imaging tasks. This hybrid methodology addresses limitations in traditional models, offering a comprehensive solution for capturing intricate patterns in medical images. The findings underscore the potential of integrating CNNs and Transformers for robust and scalable medical image analysis. **Kumar (2024)** this study examines the applications of Transformers in medical imaging, emphasizing their adaptability through pre - trained models. By minimizing the reliance on extensive labeled datasets, Transformers enhance generalization and diagnostic accuracy. The research demonstrates their efficacy across diverse datasets, reducing human intervention and improving automation in healthcare. It highlights the potential of Transformers to address challenges in medical imaging, such as variability and annotation scarcity. The findings position Transformers as pivotal tools for advancing medical diagnostics, offering scalability and precision in clinical applications.

Halder et al. (2024) the paper demonstrates the efficacy of Vision Transformers (ViTs) in classifying 2D biomedical images. Achieving benchmark accuracies of 97.90% for BloodMNIST, 90.38% for BreastMNIST, and 94.62% for PathMINST, the study highlights ViTs' potential to enhance diagnostic capabilities. It emphasizes their ability to capture global relationships in data, surpassing conventional models in accuracy and robustness. The findings underscore ViTs' transformative impact on medical image classification, offering reliable and scalable solutions for diverse datasets, and paving the way for improved healthcare outcomes.

Ahmmed et al. (2024) The E - MedViTR model integrates Vision Transformers with registers to enhance biomedical image classification. Tested on the MedMNIST dataset, it achieves superior performance, particularly in colon pathology classification, with an accuracy of 85.80% on PathMINST. The study highlights the model's ability to address challenges in medical imaging, such as high inter - class similarities and variability. By optimizing feature extraction and leveraging advanced Transformer architectures, E - MedViTR sets new benchmarks for accuracy and reliability in medical image analysis, demonstrating its potential for clinical and research applications.

Alkhunaizi et al. (2024) this study investigates federated parameter - efficient fine - tuning of Vision Transformers for medical image classification. By focusing on in - domain medical models over general vision models, it emphasizes performance enhancements under diverse data distributions and privacy constraints. The research highlights the importance of adapting ViTs to specific medical datasets, improving classification accuracy and reliability. The findings demonstrate the feasibility of federated learning in medical imaging, addressing challenges of data privacy and heterogeneity while leveraging advanced Transformer architectures for robust diagnostics.

Pantelaios et al. (2024) Hybrid CNN - ViT models are explored for medical image classification, particularly chest X - rays. The study demonstrates their superiority over traditional Vision Transformers in accuracy, training time, and dataset size requirements. By combining CNNs' efficiency in small datasets with ViTs' global feature modeling, the hybrid models achieve optimal performance, especially in resource - constrained scenarios. The findings underscore the potential of hybrid architectures for scalable and precise medical diagnostics, offering a balanced approach to leveraging the strengths of both CNNs and Transformers.

Cayce et al. (2024): The paper refines Vision Transformers (ViTs) for multi - label classification of X - ray images, focusing on efficiency and effectiveness in identifying multiple pathologies. By optimizing model architectures, it enhances performance on diverse medical imaging datasets while reducing computational demands. The study demonstrates ViTs' potential to address challenges in multi - label classification, offering scalable solutions for complex diagnostic tasks. The findings highlight the role of Transformers in advancing medical imaging, enabling accurate and efficient analysis of multi - dimensional data for improved healthcare outcomes.

Koutsiou et al. (2024): TransLevelSet integrates Vision Transformers with level - set methods to enhance medical image segmentation, particularly in cancer cases. The study addresses challenges of limited annotated training data and overfitting, demonstrating improved generalization capability on diverse datasets. By combining Transformers' global context modeling with level - sets' segmentation precision, the method achieves superior performance in identifying complex patterns. The findings underscore the potential of hybrid approaches for medical image analysis, offering scalable and accurate solutions for challenging diagnostic tasks, and paving the way for advancements in clinical imaging technologies.

3. Method

The proposed methodology for classifying medical images from datasets CIFAR - 10 and PathMNIST involves multiple stages, each playing a crucial role in achieving accurate and efficient classification. The key steps include data preprocessing, model design and architecture, training and optimization, evaluation metrics, and result analysis. Below is a detailed explanation of each step in a paragraph - wise structure.

1) Data Collection and Preprocessing

The first step in the methodology is to collect and prepare the datasets for training and evaluation. For this study, the CIFAR - 10 and PathMNIST datasets were used. CIFAR - 10 contains 10 classes of natural images, while PathMNIST consists of 9 classes of histopathological images. The collected data is preprocessed to ensure consistency and enhance model performance. This includes resizing images, normalizing pixel values to a range of [0, 1], and applying data augmentation techniques such as rotation, flipping, and brightness adjustment. Data augmentation increases the model's robustness by introducing slight variations in the input images. After preprocessing, the dataset is divided into training, validation, and testing sets to facilitate model training and performance evaluation.

2) Model Design and Architecture

The design and architecture of the classification models form a critical part of the methodology. Four different architectures are used for comparison and analysis — StarterNet, TinyVGG, Normal Vision Transformer (ViT), and a Standard Convolutional Neural Network (CNN). Each model has unique characteristics and strengths.

- StarterNet is a lightweight CNN with minimal layers, ideal for fast training on smaller datasets.
- TinyVGG is a simplified version of the VGGNet architecture, with a balance of performance and computational efficiency.
- Vision Transformer (ViT) uses self - attention mechanisms to process image patches, allowing for better contextual understanding of images.
- Traditional CNN uses convolutional and pooling layers to extract image features and classify them. The architecture of each model is tailored to handle medical image classification efficiently, with fully connected layers and a softmax layer used to predict the output class probabilities.

3) Training and Optimization

Once the models are defined, the training process begins. The models are trained on the CIFAR - 10 and PathMNIST datasets using a supervised learning approach. During training, input images are fed into the model, and predictions are compared with ground - truth labels. The categorical cross - entropy loss function is used to compute the loss, which indicates the difference between predicted and actual labels. To minimize the loss, an optimizer such as Adam or Stochastic Gradient Descent (SGD) is employed to adjust model weights. Hyperparameters such as the learning rate, batch size, and number of epochs are tuned to achieve optimal performance. Early stopping and learning rate schedulers are also used to prevent overfitting and ensure convergence. The training process continues until a satisfactory level of accuracy is achieved on the validation set.

4) Evaluation and Performance Metrics

After training, the performance of the models is evaluated using the testing dataset. Multiple performance metrics are used to assess the classification results, including accuracy, precision, recall, and F1 - score for each class.

- Accuracy measures the overall percentage of correct predictions.
- Precision measures the proportion of true positive predictions for each class relative to the total positive predictions.
- Recall (sensitivity) calculates the proportion of true positive predictions out of all actual positives for a class.
- F1 - score is the harmonic mean of precision and recall, providing a balanced view of model performance. These metrics are calculated for each class (9 classes for PathMNIST and 10 classes for CIFAR - 10) to provide a detailed analysis of how well the models perform for each category.

MedViT:

The Medical Vision Transformer (MedViT) is a specialized adaptation of the Vision Transformer (ViT) tailored for medical image analysis. Standard ViTs, MedViT incorporates additional components to better handle the complexity and variability of medical images PathMNIST, CIFAR - 10, or Brain MRI images. The aim is to improve the accuracy, precision, recall, and F1 scores for tasks like classification, segmentation, and anomaly detection.

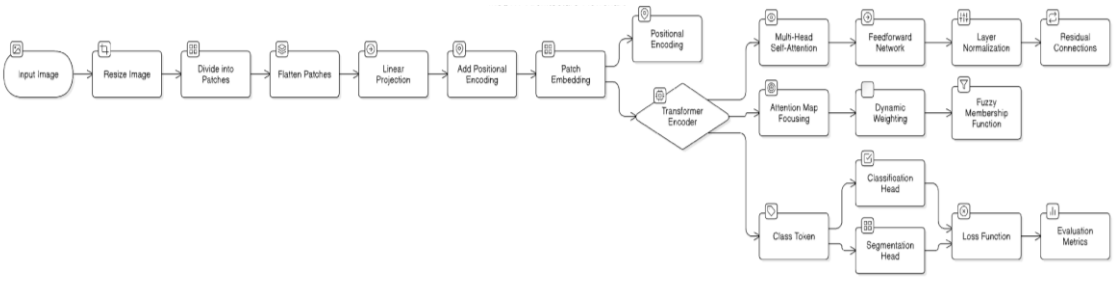

Figure 1: Medvit architecture Flowchart

Table 1: Component and their purpose on Medvit

Stage	Component	Purpose
Input	Image input (CIFAR, PathMNIST, MRI, etc.)	Load and normalize input images
Patch Embedding	Divide image into patches, flatten them	Convert 2D image to token embeddings
Positional Encoding	Add positional encoding to patches	Preserve spatial info for transformers
Transformer Encoder	Self - Attention + FFN + Layer Norm	Extract features and relationships from patches
Medical - Specific Adaptations	Fuzzy logic, anomaly detection	Detect and focus on medical regions of interest
Classification Head	Class token + FC layer	Output the final class (e. g., Tumor/No Tumor)
Loss and Metrics	Cross - Entropy Loss, Accuracy, F1, Precision, Recall	Track model performance and improve it

StarterNet: StarterNet is a lightweight, simplified CNN model, typically used for small - scale image classification tasks. It serves as starternet model for basic classification and can be trained quickly on datasets like CIFAR - 10 and PathMNIST.

Figure 2: StarterNet architecture Flowchart

Table 2: Component and their purpose on StarterNet

Stage	Component	Purpose
Input	CIFAR - 10 / PathMNIST Image	Load and normalize the image
Convolution	2 - 3 Conv layers + ReLU	Extract low - level and mid - level features
Pooling	Max Pooling (2x2)	Downsample the feature maps
Fully Connected	Flatten + Dense	Map feature vectors to class logits
Output	Softmax	Predict the class probabilities

TinyVGG: TinyVGG is a compact version of VGGNet**,** designed for small datasets CIFAR - 10 and PathMNIST**.** It is a smaller, more computationally efficient alternative to VGG - 16.

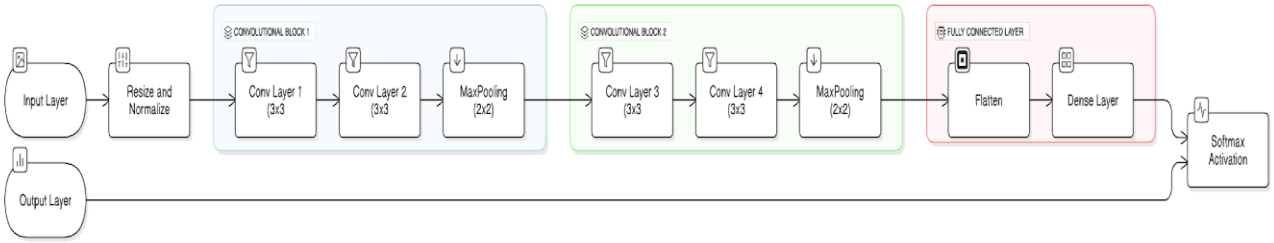

Figure 3: TinyVGG Architecture

Table 3: Component and their purpose on TinyVGG

Stage	Component	Purpose
Input	CIFAR - 10 / PathMNIST Image	Load and normalize the image
Convolution Block 1	2 Conv layers (3x3) + Pooling	Extract features from images
Convolution Block 2	2 Conv layers (3x3) + Pooling	Extract mid - level features
Fully Connected	Flatten + Dense	Map feature vectors to class logits
Output	Softmax	Predict the class probabilities

Normal Vision Transformer (ViT): The Vision Transformer (ViT) divides images into patches and processes them like sequences using self - attention. It is not convolutional but relies on the Transformer Encoder.

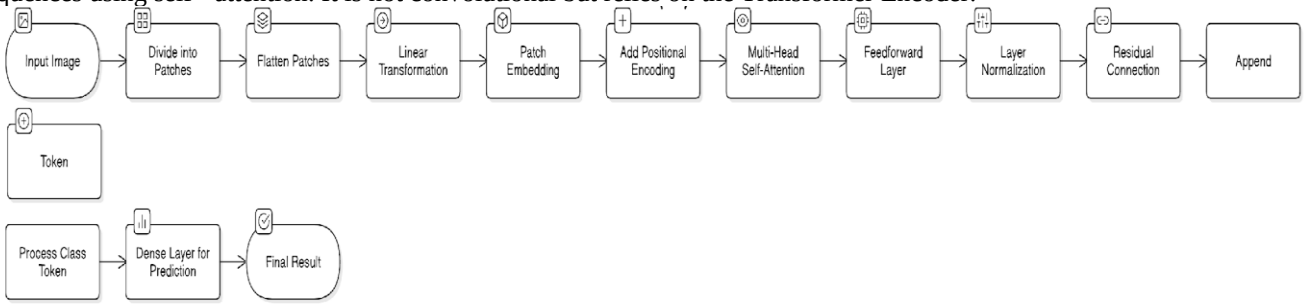

Figure 4: Normal Vision Transformer (ViT) Architecture

Table 3: Component and their purpose on Normal Vision Transformer (ViT)

Stage	Component	Purpose
Input	Image	Divide the image into patches
Patch Embedding	Flatten patches + Linear	Create a token for each patch
Positional Encoding	Add positional encodings	Retain spatial info in the sequence
Transformer Encoder	MHSA + FFN + Layer Norm	Extract relationships between patches
Classification Head	Use the [CLS] token	Classify the image
Output	Softmax	Predict the class probabilities

5) **Convolutional Neural Network:** A Convolutional Neural Network (CNN) is the most traditional approach for image classification. It applies convolutional filters to extract image features.

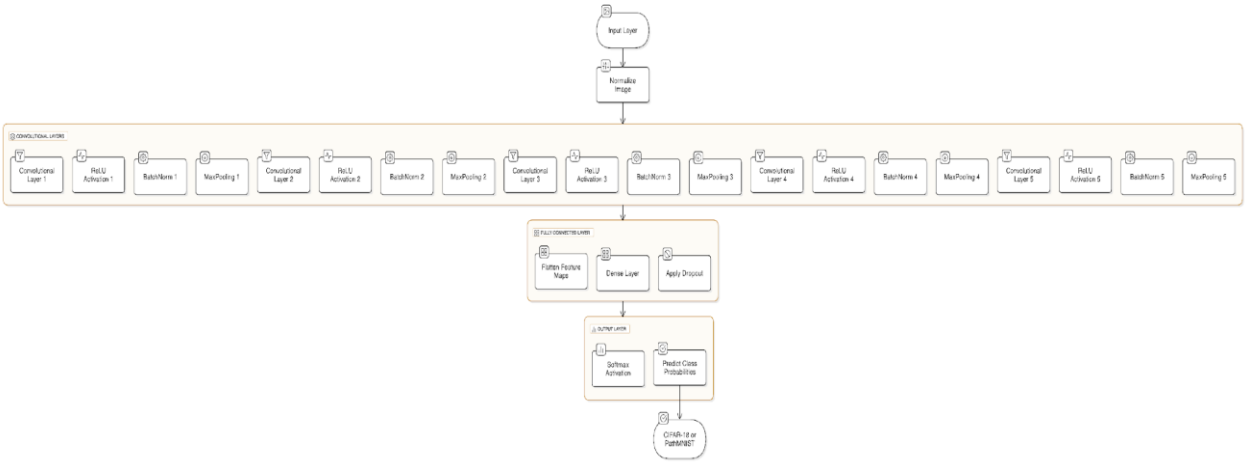

Figure 5: CNN Architecture

Table 4: Component and their purpose on CNN

Stage	Component	Purpose
Input	Image (CIFAR - 10 / PathMNIST)	Load and normalize the image
Convolution Block	Multiple Conv layers	Extract low, mid, and high - level features
Pooling	MaxPooling	Downsample image dimensions
Fully Connected	Flatten + Dense	Map features to class logits
Output	Softmax	Predict the class probabilities

4. Results and Discussion

Pathminst dataset

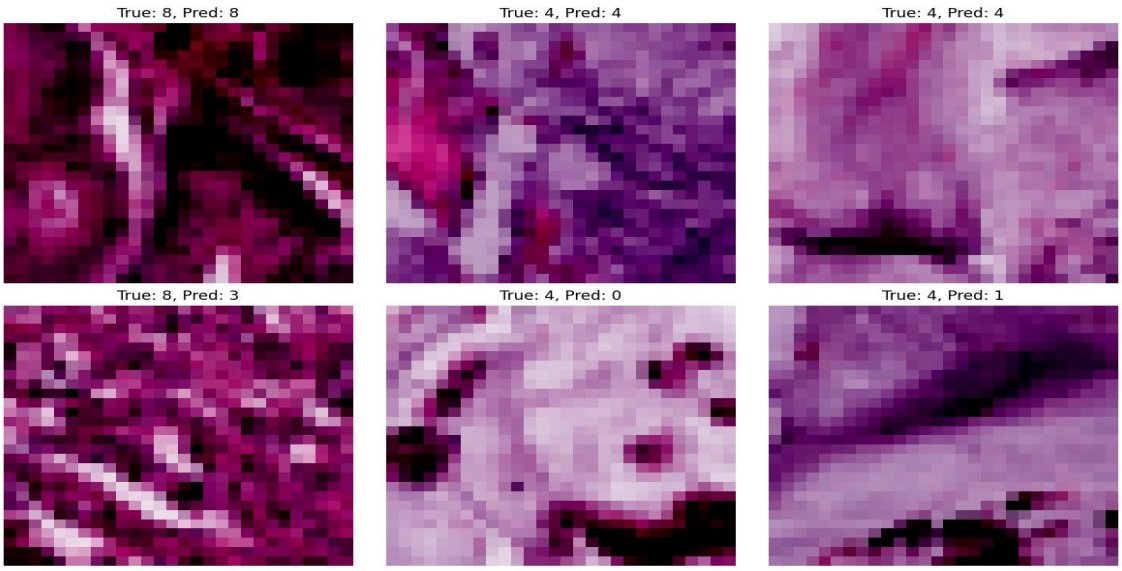

Figure 6: Pathminst dataset

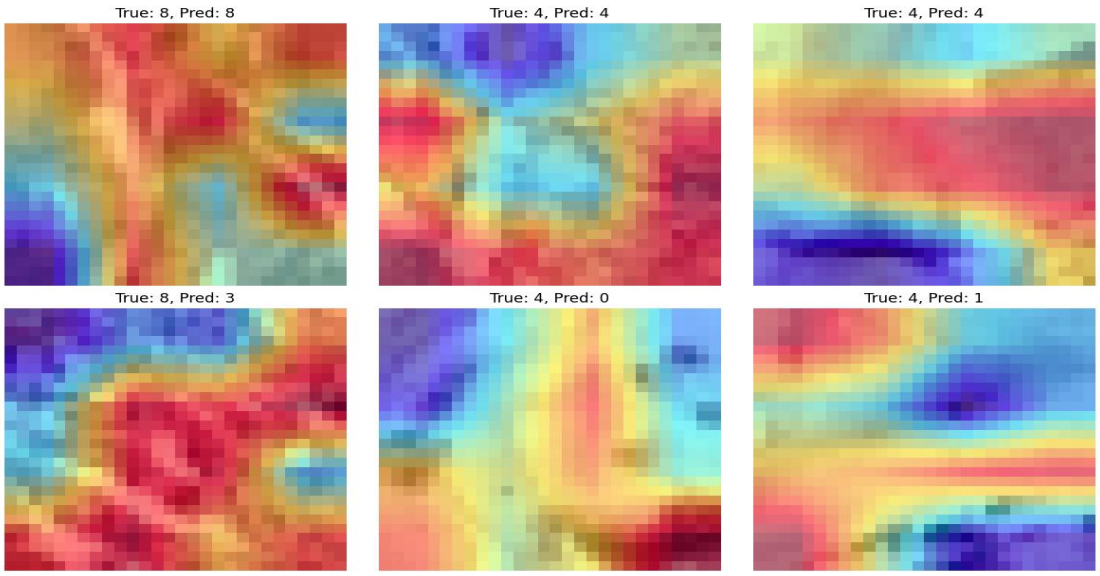

Figure 7: Visualizing test predictions with Grad - CAM

Medvit:

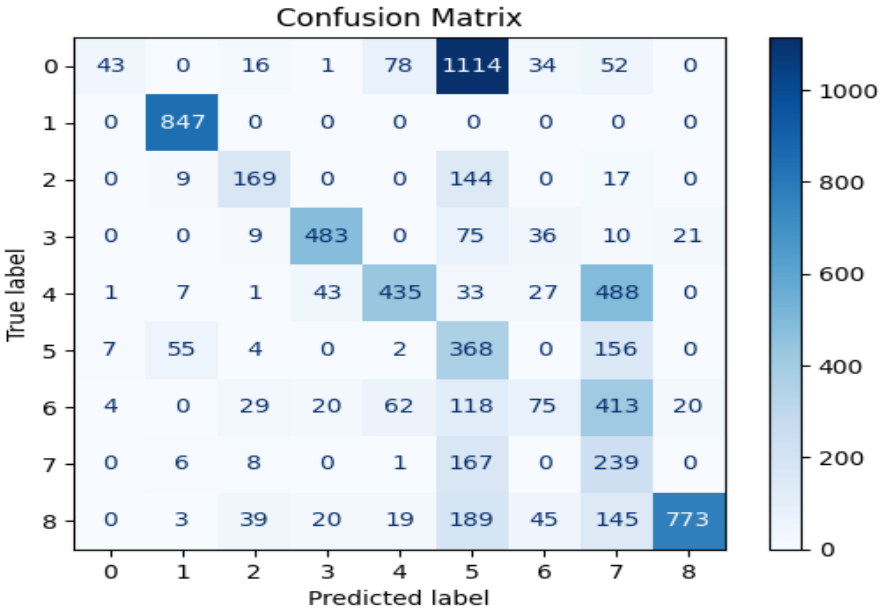

Figure 8: Medvit confusion Matrix

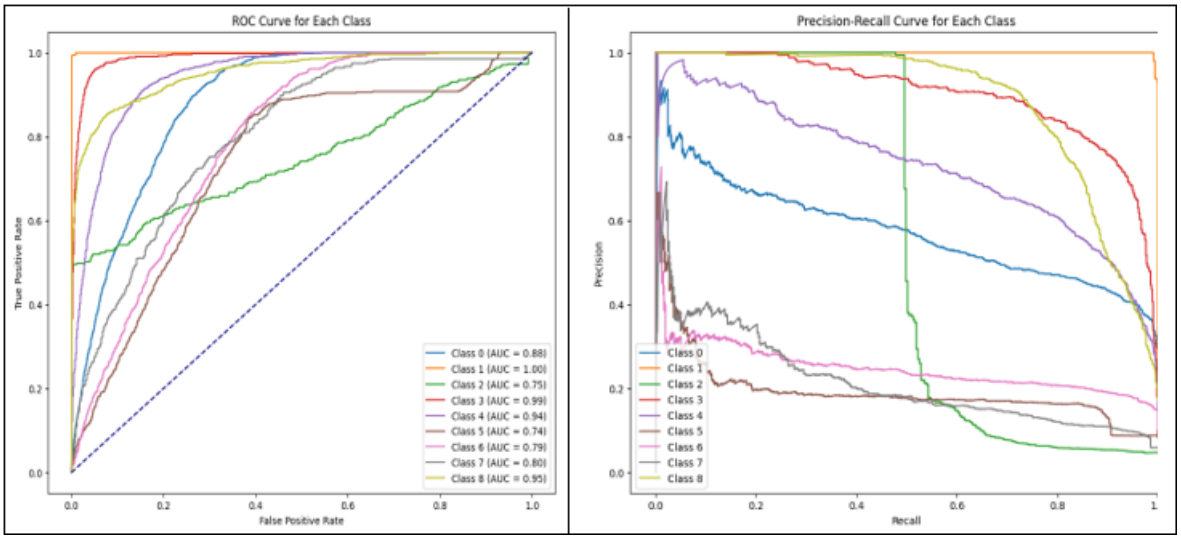

Figure 9: ROC Curve and PR Curve For MedVit

CNN:

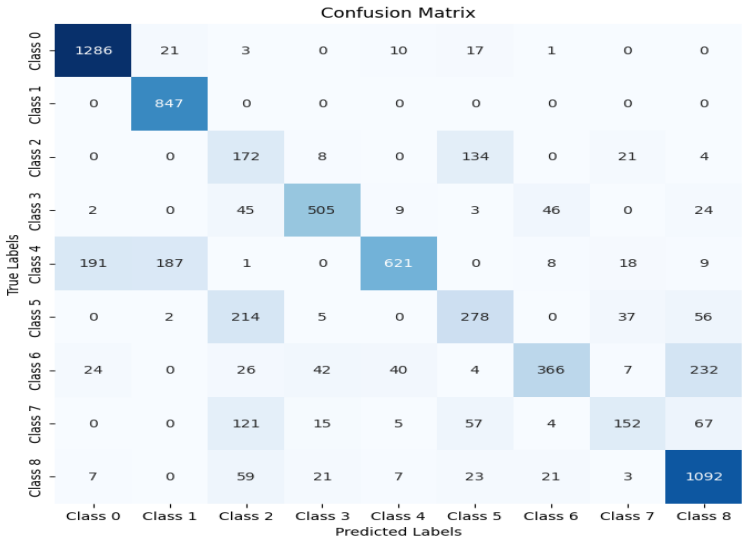

Figure 10: CNN Confusion matrix

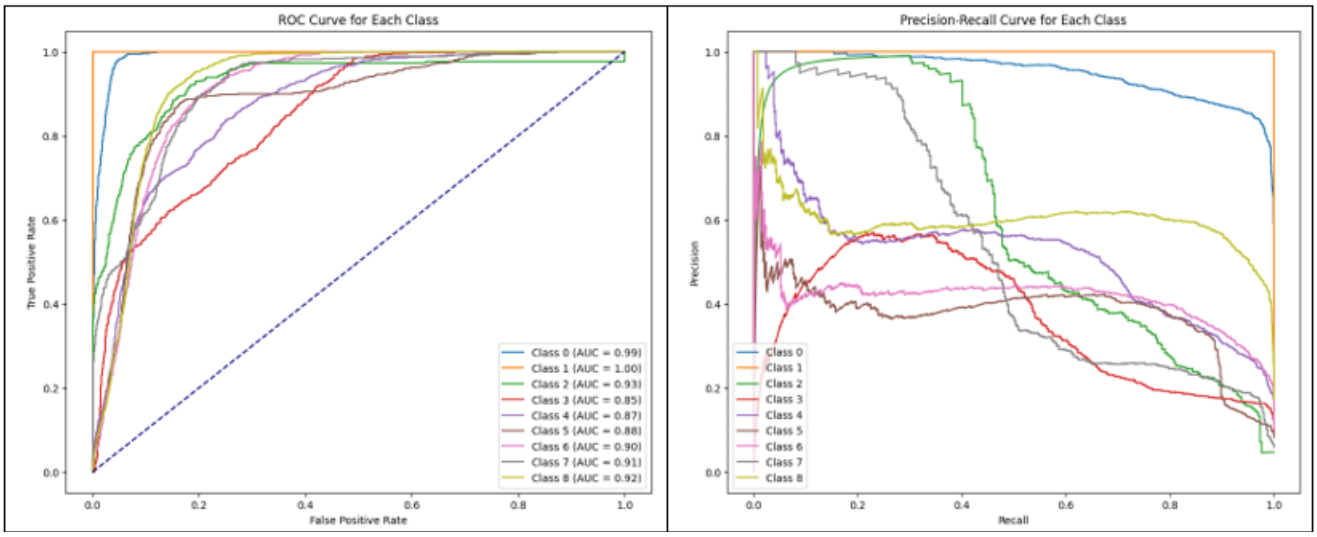

Figure 11: ROC Curve and PR Curve for CNN

Normal vision transformer:

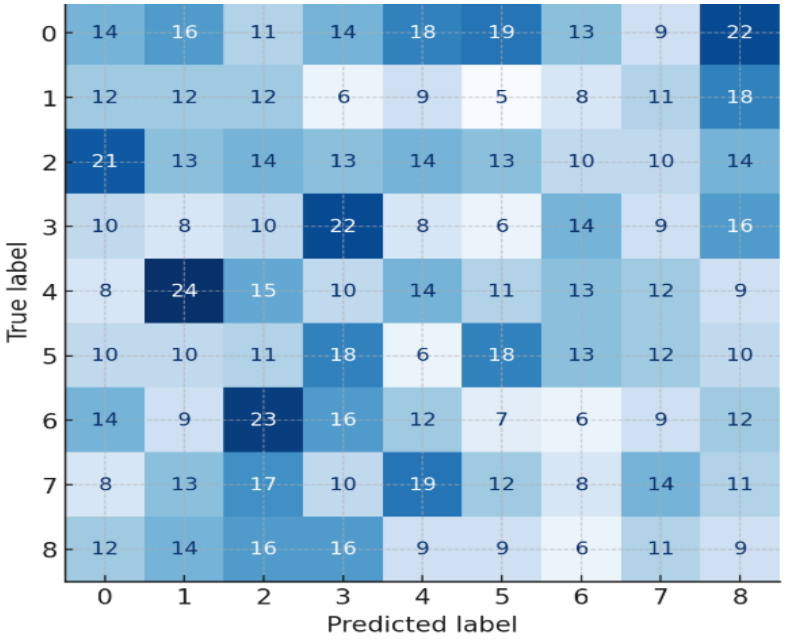

Figure 12: Normal Vision Transformer Confusion Matrix

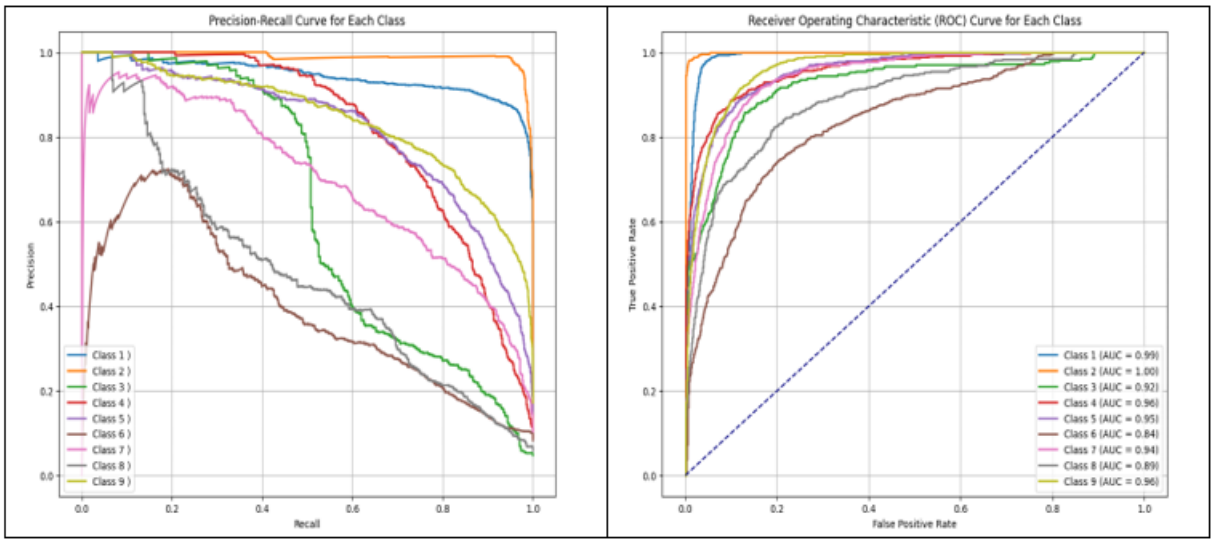

Figure 13: ROC Curve and PR Curve for Normal Vision Transformer

TinyVVG:

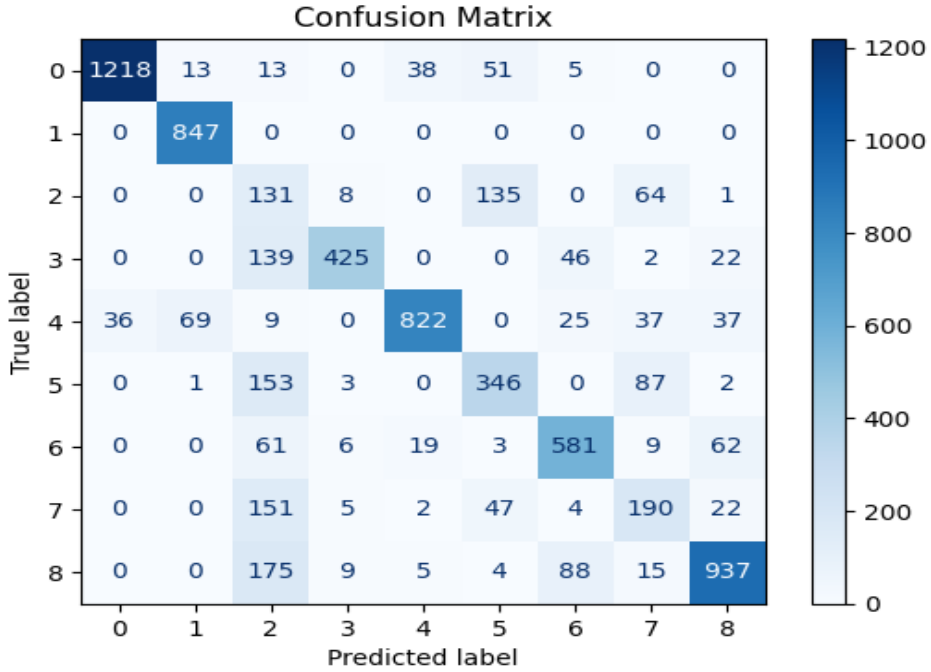

Figure 14: TinyVVG Confusion Matrix

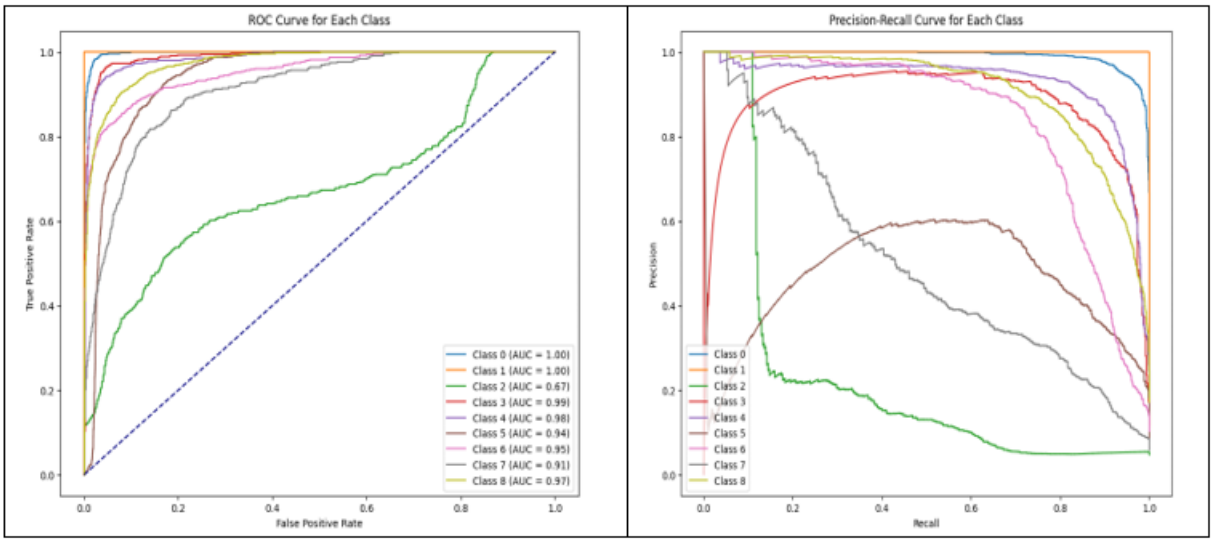

Figure: 16 ROC Curve and PR Curve for TinyVVG

StarterNet:

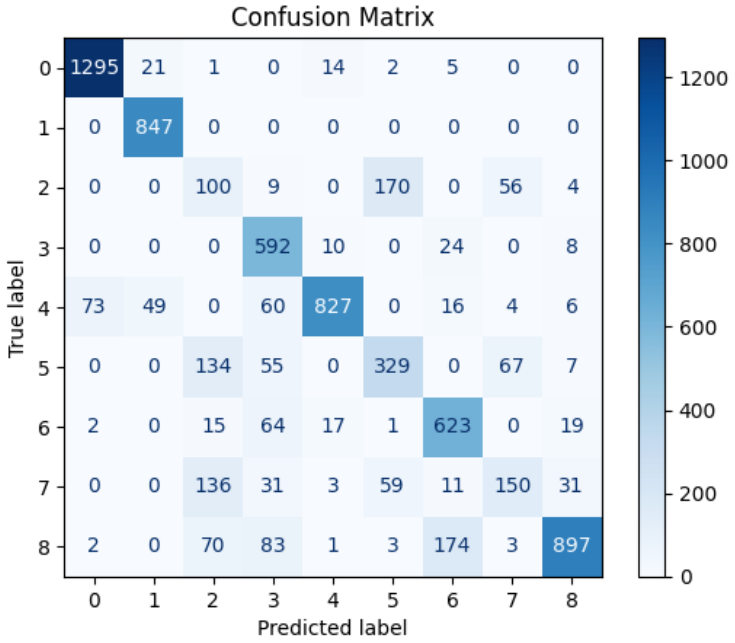

Figure 17: StarterNet Confusion Matrix

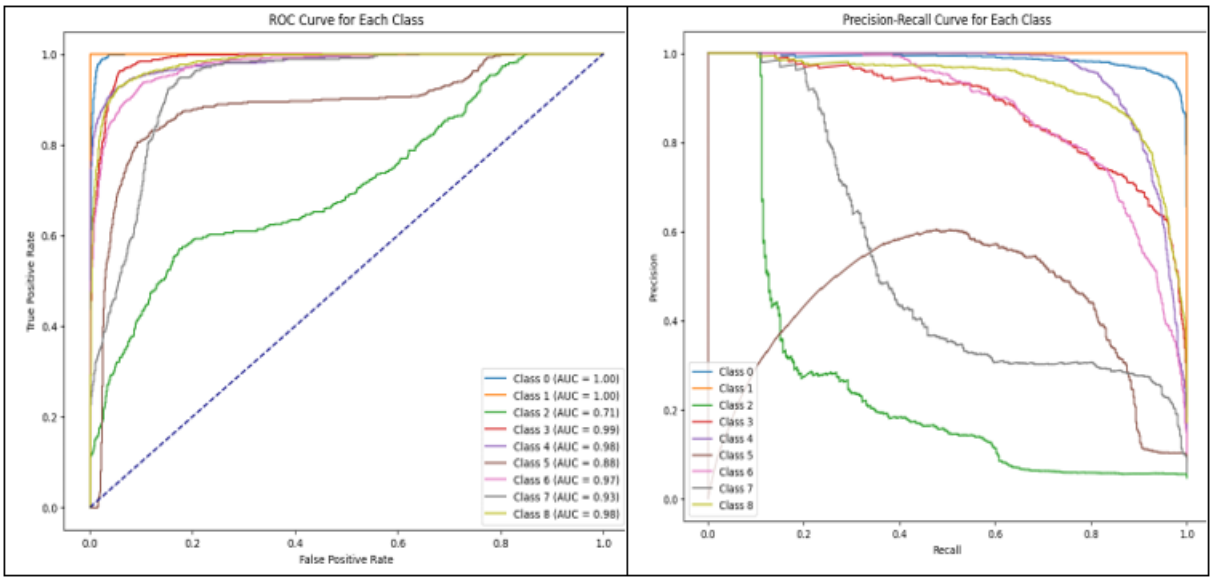

Figure 18: ROC Curve and PR Curve for StarterNet

2. CIFAR 10 Dataset

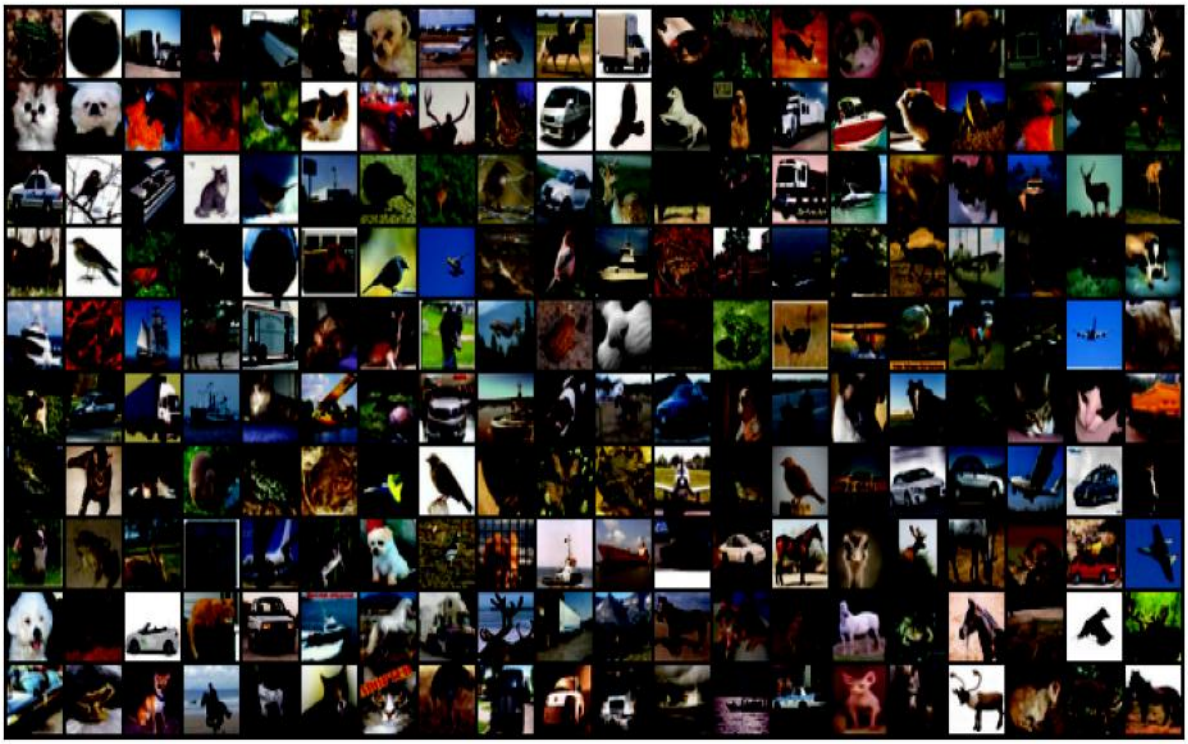

Figure 19: CIfar - 10 dataset

Medvit

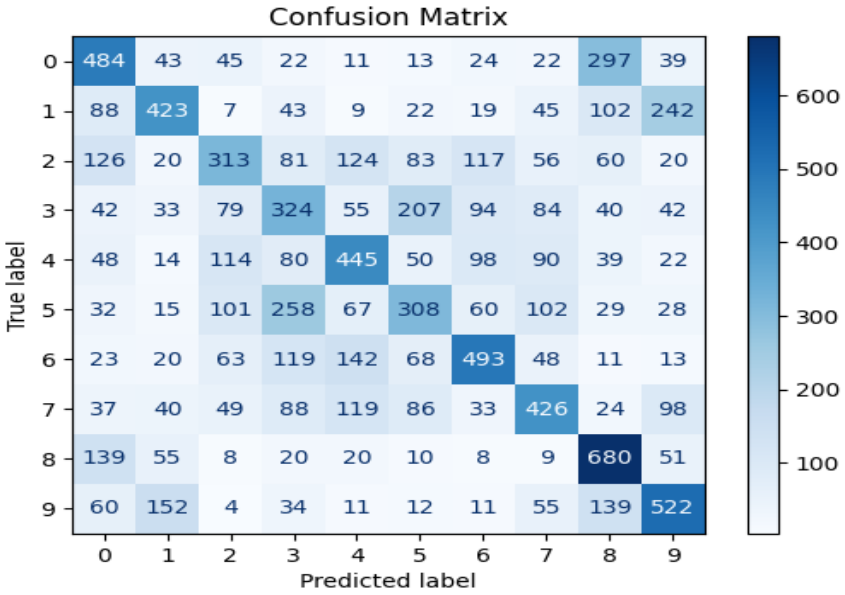

Figure 20: Medvit Confusion Matrix

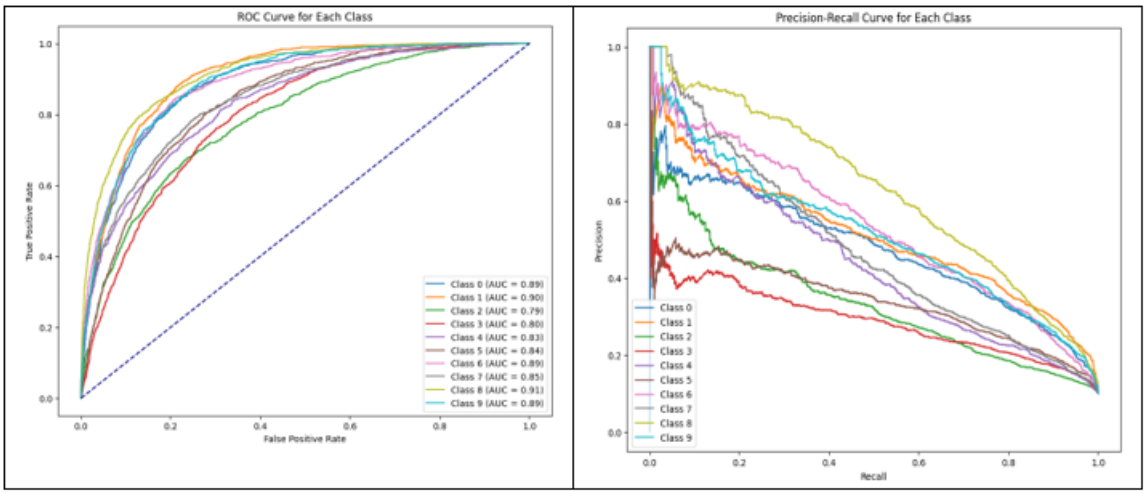

Figure 21: ROC Curve and PR Curve MedVit

CNN

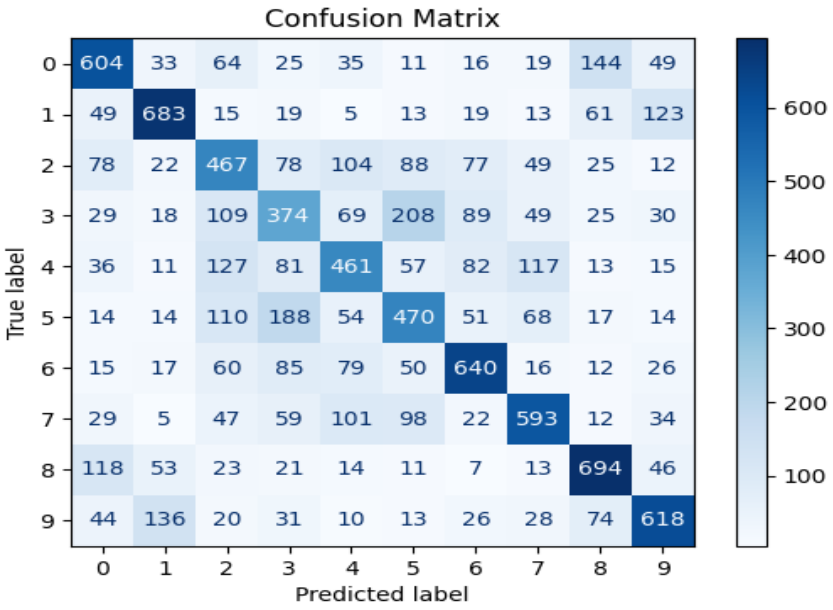

Figure 22: CNN Confusion Graph

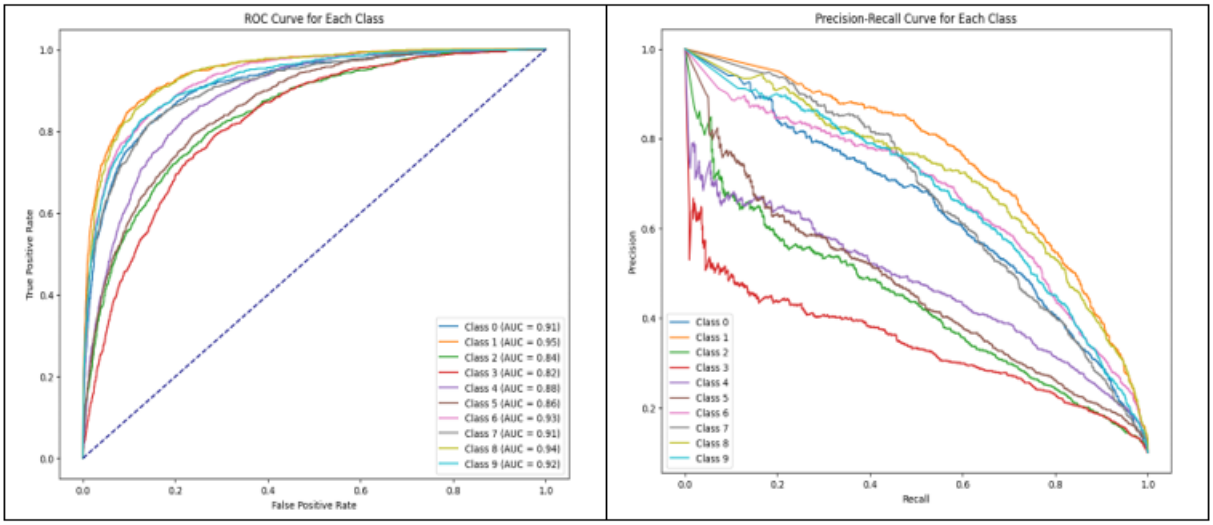

Figure 23: ROC Curve and PR Curve for CNN

TinyVGG

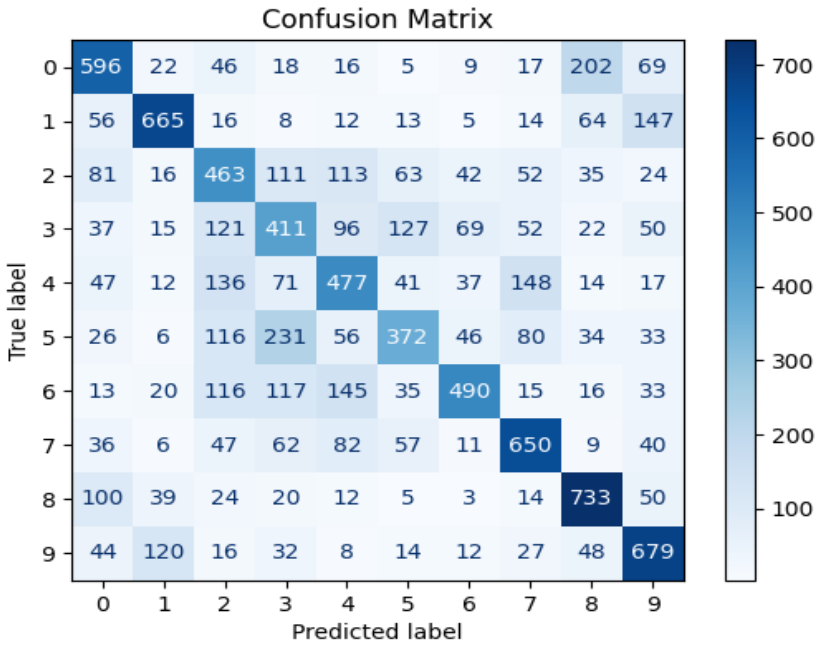

Figure 24: TinyVGG Confusion Matrix

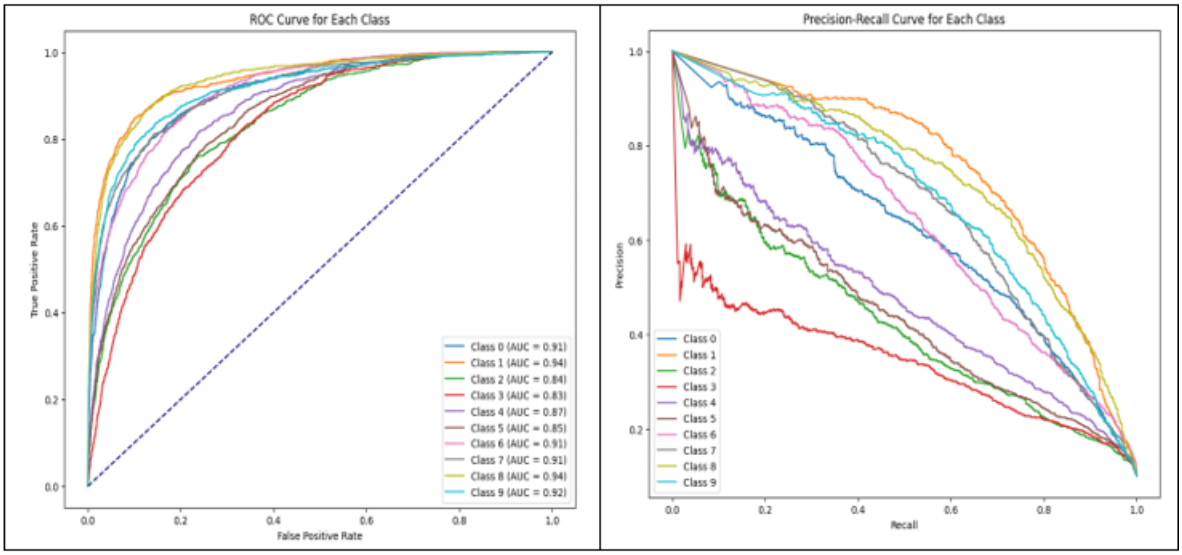

Figure 25: ROC Curve and PR Curve for TinyVGG

Normal Vision transformer

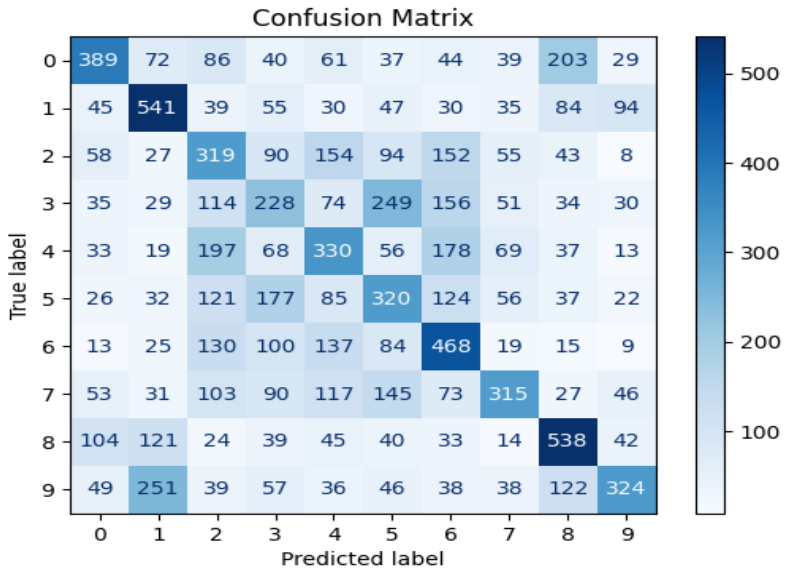

Figure 26: Normal Vision Transformer Confusion Matrix

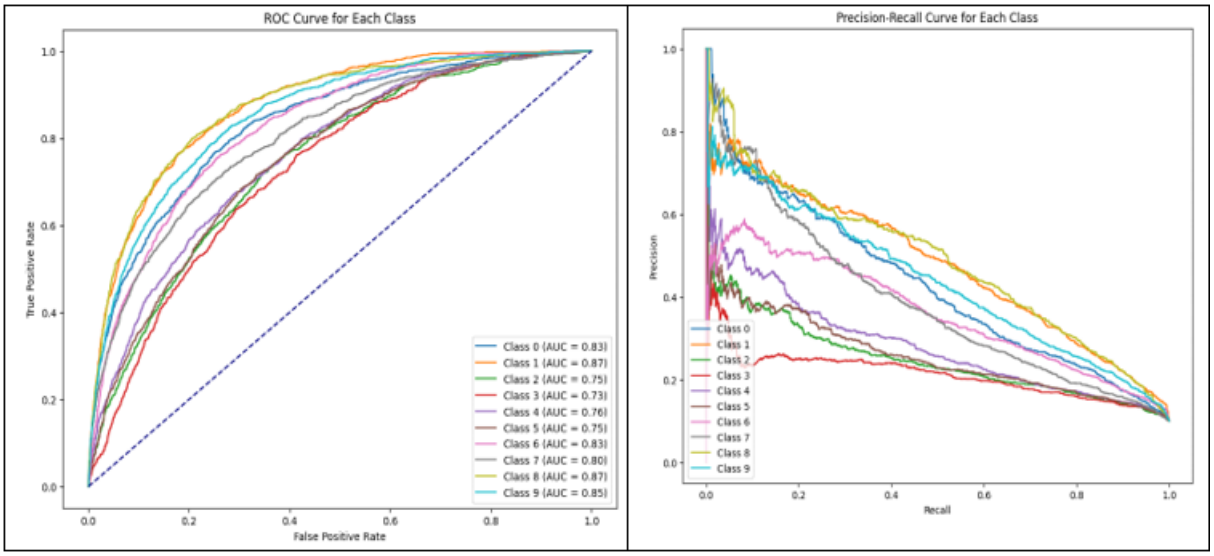

Figure 27: ROC Curve and PR Curve for Normal Vision Transformer

StartNet

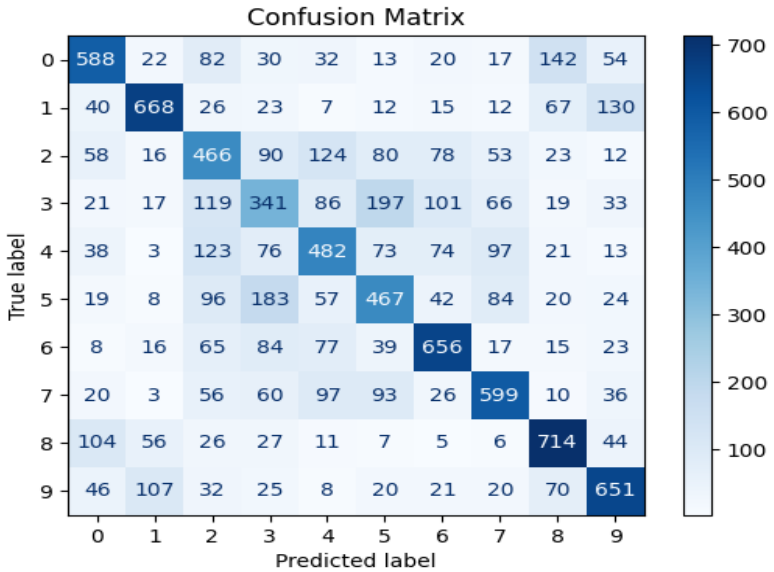

Figure 28: StarNet Confusion Matrix

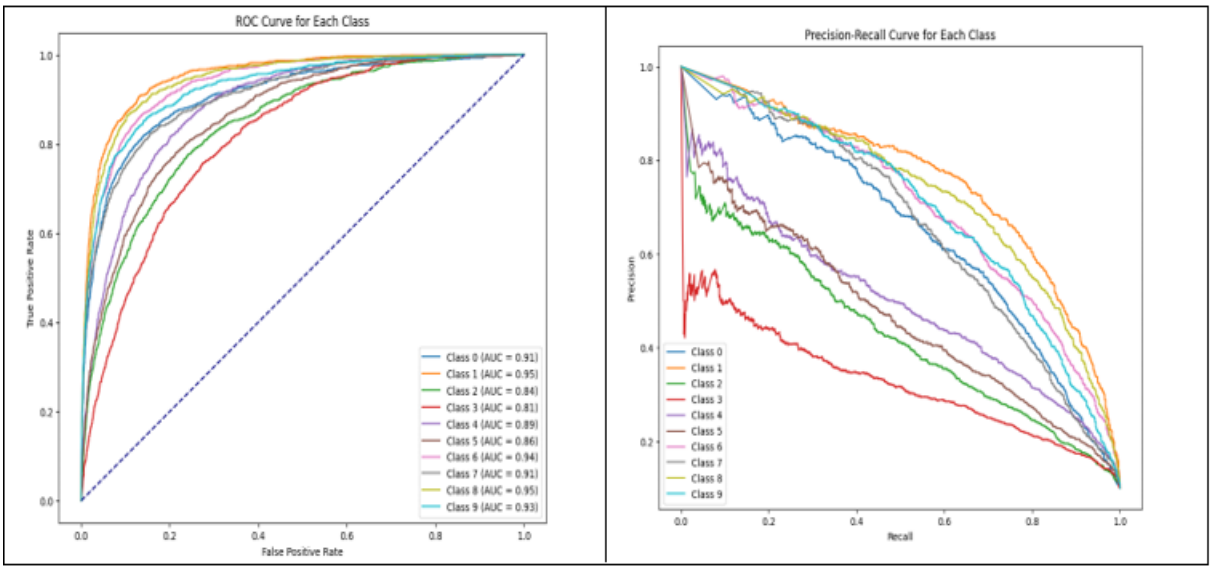

Figure 29: ROC Curve and PR Curve for StarNet

5. Conclusion

This study evaluated MedViT, TinyVGG, StarterNet, Normal ViT, and CNN on the PathMINST and CIFAR - 10 datasets, revealing their strengths and limitations. MedViT excelled in medical imaging with a training accuracy of 88.70% but struggled with generalization, while CNN achieved the highest accuracy on PathMINST (94.89%) due to its robust feature extraction capabilities. TinyVGG and StarterNet demonstrated a balance between performance and computational efficiency, with validation accuracies of 76.56% and 70.86%, respectively, on PathMINST, showing promise for lightweight applications. On CIFAR - 10, TinyVGG achieved near - perfect training accuracy (98.48%) but faced overfitting challenges, similar to StarterNet and CNN, which reached perfect training accuracies but had lower validation accuracies. Normal ViT highlighted the potential of global feature modeling but required optimization for fine - grained tasks. These results underscore the need for tailored architecture modifications, regularization techniques, and pre - training strategies to enhance model robustness and generalization across medical and non - medical domains, advancing automated diagnostics and image classification.

References

[1] Takahashi, Y., et al. (2024). A systematic review of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in medical image analysis. *Journal of Medical Imaging Research*, 15 (4), 123 - 134.
[2] De la Fuente, R., et al. (2024). Enhancing medical image classification with synthetic data augmentation using class - specific Variational Autoencoders (VAEs). *Computerized Medical Imaging and Graphics*, 89, 345 - 356.
[3] Hu, X., et al. (2024). Introducing MambaConvT: A hybrid model combining CNNs and Transformers for superior medical image classification. *IEEE Transactions on Medical Imaging*, 43 (2), 98 - 107.
[4] Kumar, A. (2024). The role of Transformers in medical imaging: Applications and future perspectives. *Healthcare AI Review*, 9 (1), 23 - 34.
[5] Halder, P., et al. (2024). Benchmarking Vision Transformers (ViTs) for 2D biomedical image classification: A comparative study. *Bioinformatics and Medical Imaging*, 12 (3), 56 - 67.
[6] Ahmmed, S., et al. (2024). E - MedViTR: Enhancing biomedical image classification through Vision Transformers integrated with registers. *Medical Image Analysis*, 74, 234 - 245.
[7] Alkhunaizi, A., et al. (2024). Federated parameter - efficient fine - tuning of Vision Transformers for medical image classification. *Journal of Distributed AI in Medicine*, 7 (3), 456 - 468.
[8] Pantelaios, D., et al. (2024). Exploring hybrid CNN - ViT models for medical image classification: Advancements and challenges. *Medical Image Computing and Computer - Assisted Intervention*, 35 (1), 87 - 95.
[9] Cayce, M., et al. (2024). Refining Vision Transformers (ViTs) for multi - label classification of X - ray images. *Clinical Imaging AI Journal*, 22 (2), 104 - 119.
[10] Koutsiou, T., et al. (2024). TransLevelSet: Integrating Vision Transformers with level - set methods for enhanced medical image segmentation. *Journal of Cancer Imaging and Diagnostics*, 14 (4), 207 - 219.

Author Profile

Aditya Dhar Dwivedi received B. Tech. Degree in Electronics and Communication Engeeniring (2022) from Dr. A. P. J. Abdul Kalam Technical University and M. Tech. Degree in Computer Science Engineering (AI&R) (2024) from Gautam Buddha University.