A Deep Dive into Vision Models: Comparing CNN and Transformer-Based Approaches for Pneumonia Detection

Md Saiful Islam Sajol
Dept. of CSE
Louisiana State University
Louisiana, USA
msajol1@lsu.edu

Raisa Islam
Dept. of CS
New Mexico Inst. of Mining and Tech.
NM, USA 87801
raisa.islam@student.nmt.edu

A S M Jahid Hasan
Dept. of ECE
North South University
Dhaka, Bangladesh
jahid.hasan12@northsouth.edu

Mainul Kabir
Dept. of CS
New Mexico Inst. of Mining and Tech.
NM, USA 87801
mainul.kabir@student.nmt.edu

***Abstract*—Image classification is a fundamental computer vision task crucial for interpretation of visual data. Significant advancements have been observed in image classification models over recent years. In this paper, a taxonomy is presented categorizing image classification models into CNN, Transformer, and Hybrid models, in addition to an evaluation of these models for pneumonia detection using a chest X-ray (CXR) dataset. Key metrics such as accuracy, precision, recall, F1-score, and average inference time per image were analyzed to understand trade-offs between performance and efficiency for most feasible model selection concerning this task. Transformer-based models, particularly SwinV2, proved their robustness for pneumonia detection, achieving the highest accuracy (95.03%), recall (94.66%), and F1-score (94.70%). However, the longer inference time (133.58 ms) presents a considerable disadvantage. Although, CNN models like EfficientNetB0 and ResNet50 achieved the highest precision, they underperformed in other metrics, highlighting the need for balanced evaluations. Hybrid models, such as CvT, consistently performed well across all metrics, showcasing the potential of transformers integrated with convolutional layers. However, no single model category demonstrated unanimous superiority, as CNN models exhibited much faster temporal efficiency than Transformer and Hybrid models. The source code for this work can be accessed here: https://github.com/MdSaifulIslamSajol/pneumonia-detection-with-chest-x-ray-images**

***Index Terms*—Vision Classifier, Image Classification, Transformer, CNN**

I. Introduction

Image or Vision classification is a fundamental task in computer vision, essential for interpreting and understanding visual data [1, 2]. It involves categorizing images into predefined classes, enabling various applications such as object recognition, facial recognition, medical imaging, and autonomous driving. Accurate classification facilitates advancements in machine learning (ML) and artificial intelligence, enhancing automation and decision-making processes [3]. Moreover, it offers a basis for complex tasks like image segmentation and object detection, making it crucial for developing intelligent systems that interact with and interpret visual data. The application of image classification models in the medical imaging domain, particularly for disease detection using X-Ray images, can be highly beneficial. Developing highly accurate image classification models to assist in disease diagnosis through X-Ray imaging has become a growing area of interest among researchers.

The evolution of image classification models has made significant advancements, beginning with simple linear classifiers and advancing to sophisticated deep learning architectures. Initially, models like k-Nearest Neighbors (kNN) and Support Vector Machines (SVM) were employed for basic image classification tasks. The introduction of Convolutional Neural Networks (CNNs) revolutionized the field by enabling automatic feature extraction and hierarchical learning of image features. Models such as LeNet, AlexNet, and VGG showcased the power of deep learning in achieving high accuracy in image classification. Subsequently, the development of more complex architectures like ResNet, which introduced residual connections, addressed the issue of vanishing gradients and allowed for much deeper networks. The recent emergence of transformer-based models, originally from natural language processing, brought further innovation by providing greater flexibility for processing multi-modal inputs including images [4]. Further, hybrid models were proposed combining the strengths of CNNs and transformers, pushing the boundaries of image classification capabilities and opening new avenues for research and application.

Khalil et al. [1] provided a chronological overview of transformer-based vision classifiers, but their study did not categorize the models. This research offers a comprehensive overview of models developed for vision classification by establishing a taxonomy of the existing methodologies. Unlike

existing researches that predominantly concentrate on assessing the accuracy of individual models [2, 5, 6] or comparing models of similar types [7, 8, 9, 10] or only a few models [11, 12, 13], our study provides a broader comparative analysis. We utilized Chest X-Ray images as a case study to evaluate and compare at least one representative model from each category within the proposed vision classifier taxonomy. This approach not only highlights the distinct characteristics of each model type but also provides insight of their performances.

Introduction of newer vision classification models within the scope of Chest X-Ray leads us to the following research questions:

- RQ#1. What are the prominent models used for image classification? Is there a taxonomy for categorizing these models?
- RQ#2. Which classification model performs best for diagnosing Pneumonia through Chest X-Ray images?

Rest of the paper is organized as following: in Section II we answer RQ#1 by providing taxonomy of existing vision classifiers. Section III describes the dataset and experiment, and result of these experiments are presented in Section IV. Finally, in Section V, we conclude our paper.

II. Taxonomy of Vision Classifiers

CNNs have been utilized for image classification for many years, revolutionizing the field of computer vision [14]. Instead of requiring preprocessing to derive features like textures and shapes, CNNs use raw pixel data as input and "learn" to extract these features, ultimately identifying the objects within the images. In 2014 Google introduced GoogLeNet [15], with Inception module, which allows the network to capture multi-scale features by applying multiple convolution filters of different sizes simultaneously within the same layer. This innovative design enhances the network's ability to recognize complex patterns while maintaining computational efficiency. Visual Geometry Group (VGG) [16] employs a series of straightforward CNN architectures, consisting primarily of 3x3 convolutional layers stacked on top of each other with increasing depth, interspersed with max-pooling layers to reduce spatial dimensions. This simple and uniform design allows VGG networks to capture intricate patterns and hierarchical features effectively. Residual Network (ResNet) [17] is another popular deep CNN architecture designed to address the degradation problem in deep neural networks. By incorporating residual learning through shortcut connections, ResNet allows for the training of extremely deep networks without the issues of vanishing gradients. These shortcut connections effectively skip one or more layers, enabling the network to learn identity mappings that preserve information across layers. MobileNet [18] proposed to use depthwise separable convolutions, which significantly reduce the computational cost and number of parameters compared to standard convolutions. Further, inverted residuals and linear bottlenecks were introduced in MobileNetV2 [19], improving the performance and efficiency. This bottleneck convolution block concept was extended by EfficientNet [20] family with an addition of squeeze-and-excitation optimization to improve channel-wise feature recalibration.

Transformers have achieved state-of-the-art performance in a wide range of tasks from Natural Language Processing (NLP) [21] to image classification [1]. Dosovitskiy et al. [22] extended Transformer concept to visual images by introducing the pioneering Vision Transformer (ViT), a global vision transformer, which differs from traditional CNN models by employing a self-attentive mechanism. ViT requires global feature reasoning by computing self-attentions among all tokens, making it computationally inefficient when processing images with many visual tokens [4]. To address this issue, local vision transformer, such as Swin Transformer [23] and PvT [24], have been proposed. In contrast to global vision transformers, local vision transformer architectures utilize more convenient attention windows that can remain fixed or adjusted during fine-tuning phase. Later, Swin Transformer V2 [25] was proposed to address the limitations and scaling challenges of Swin by incorporating several key innovations, including a hierarchical architecture with shifted windows for local attention, which improves computational efficiency and performance. Additionally, Swin V2 introduces post-normalization and a scaled cosine attention mechanism to stabilize training and enhance generalization. Pyramid Vision Transformer (PvT) V2 [26], iGPT [27] are notable among other transformer based visual classifiers.

Transformers lack the inductive biases inherent to CNNs, which results in lower performance compared to similarly sized CNN counterparts when trained on smaller datasets [22, 28, 29]. To overcome this issue, a hybrid CNN-Transformer model, Convolutional Vision Transformer (CvT) [28] was proposed that incorporates convolutional layers within the transformer architecture. Recently another model, FastViT [30] proposed to use large convolutional kernels to substitute self-attention layers in early stages. Alternatively, Local Relation Network (LR-Net), BoTNet e.t.c., introduces self-attention mechanism in CNNs. LR-Net [31] enhance CNNs by incorporating self-attention mechanisms to capture local relational features between neighboring pixels or regions. Similarly, EfficientFormer [32], a successor of EfficientNet, employs factorized self-attention with linear approximation and sparse attention pattern. ConvNext V2 [33, 34], building upon its predecessor ConvNeXt, maintains the architectural simplicity and scalability of traditional CNNs while incorporating modern enhancements inspired by vision transformers, such as improved self-attention mechanisms and hierarchical feature extraction.

Based on the architectures used in vision classification, classifier models can be categorized into three main types: (a) CNN based model, (b) transformer based model, and (c) hybrid model. Hybrid models integrate features from both CNN and transformer architectures, either by incorporating self-attention mechanisms from transformers into CNNs or by adding convolutional layers to transformer models. CNN based models can be sub-divided furthermore based on the number of layers used in their architecture, i.e., classic, deep,

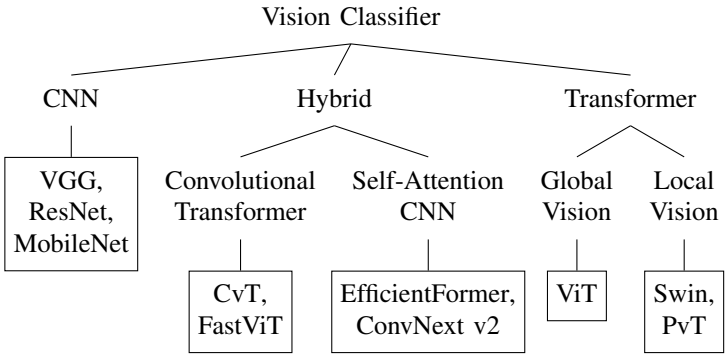
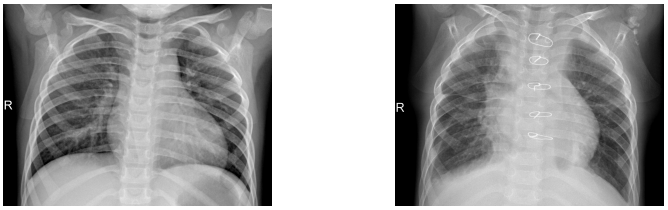


Fig. 1: Architectural Taxonomy of Vision Classifier Models

very deep e.t.c, however for simplicity those divisions were not included. Fig. 1 represents the architectural taxonomy of visual classifiers where leaf nodes represents popular models.

III. Experiment

This research aims to find the best method for Pneumonia detection using various methods.



(a) Normal (b) Pneumonia

Fig. 2: Chest X-rays

A. Dataset

The Chest X-Ray Image dataset (CXR) [35] for Pneumonia Detection is a publicly available collection of chest radiographs aimed at facilitating the development and evaluation of machine learning models for diagnosing pneumonia. This dataset contains thousands of X-ray images categorized into two classes: (a) normal, and (b) pneumonia. Table I depicts the number of samples in the test and train dataset used in our experiment. The images are typically grayscale and vary in resolution, reflecting the real-world variability encountered in clinical settings. Each image is labeled by expert radiologists, providing a reliable ground truth for training and validating diagnostic models.

TABLE I: CXR Dataset Description

	Normal	Pneumonia
train	1341	3875
test	234	390

B. Preprocessing

Data preprocessing helps to improve model accuracy and enhance efficiency. Initially, training set images are resized to 256x256 pixels. Data augmentation is an important step in training ML models [7], hence images were randomly rotated by upto 20 degrees and zoomed-in by centre-cropping. The images are center-cropped to 224x224 pixels for two reasons: a) to match the input shape of the VGG16 and ResNet50 pre-trained model used [36], and b) some images might be larger or in a different aspect ratio, recropping will ensure an uniform aspect ratio. The final preprocessing step involves normalizing the images using the mean and standard deviation values derived from the dataset. This normalization step ensures that the input data has a consistent scale, which aids in the convergence of the model during training.

For the test phase, the preprocessing pipeline is simplified. Images are directly resized to 224x224 pixels and normalized using the same mean and standard deviation values as in the training phase. This consistency in preprocessing across different phases ensures that the model evaluates data under similar conditions as it was trained.

C. Training

This phase involves training the ML models for 30 epochs with a batch size of 100. The `AdamW` optimizer is utilized with a learning rate set at 0.0001, and the learning rate is adjusted using a `StepLR` scheduler. This optimizer combines the benefits of Adam and weight decay regularization, which helps in achieving better generalization. The learning rate scheduler adjusts the learning rate at predefined steps, aiding in fine-tuning the model's convergence during training. The learning rate selection is done by varying it from 0.1 to 0.0001 and observing the best outcome.

The Cross-Entropy loss function is employed as the objective function, which is suitable for classification tasks as it measures the performance of the model's output probabilities against the true class labels. Training is conducted on NVIDIA GPUs with CUDA support, which significantly accelerates the computation, allowing for more efficient training of deep learning models. The authors would like to mention that, although the training process is resource-intensive, the models

do not need to be trained during deployment. Once trained, they can be deployed for inference.

IV. Result and Analysis

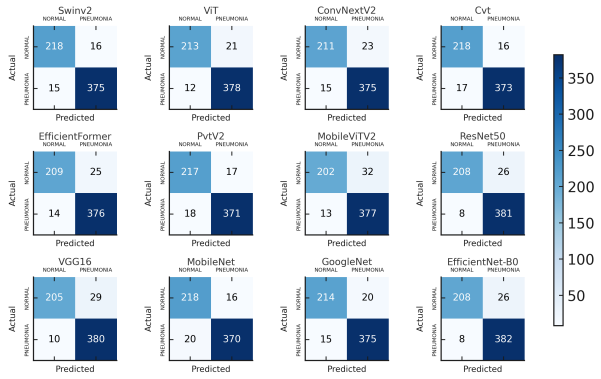


Fig. 3: Confusion Matrices of different models for test dataset

Results obtained from the experiments performed are presented and analyzed in this section. Fig. 3 presents the confusion matrices for each model on the test dataset. It can be observed that all methods perform reasonably well. However, for a more thorough analysis, we need to compare the models across different performance metrics. Figure 4 compares the performance of ML models from the main categories of our taxonomy: CNN, Transformer, and Hybrid. The x-axis represents different ML models, while the y-axis shows performance in percentages, with bars ordered in descending order by metric value. For performance evaluation we compare four most prominent metrics for classification task: accuracy, precision, recall and F1-score.

Figure 4a shows that SwinV2 achieves the highest accuracy at 95.03%, while ViT, CvT, and EfficientNetB0 each reach 94.71%. EfficientFormer and VGG16 perform the poorest, with an accuracy of 93.75%.

Figure 4b presents the precision values of the trained models. The CNN models EfficientNetB0 and ResNet50 achieve the highest precision at 95.19% and 94.96%, respectively. SwinV2 and ViT, both Transformer-based models, follow with precision values of 94.74% and 94.70%, respectively. Most Hybrid models, except CvT, show lower precision.

Figure 4c indicates that the SwinV2 model has the highest recall at 94.66%, followed by CvT at 94.40% and PvTV2 at 94.06%. CNN models exhibit recall rates from 92.52% to 94.02%. Figure 4d shows similar trends, with SwinV2 (94.7%) and CvT (94.36%) leading, followed by ViT (94.31%). For both recall and F1-score, ConvNextV2, EfficientFormer, and VGG16 are the lowest performers.

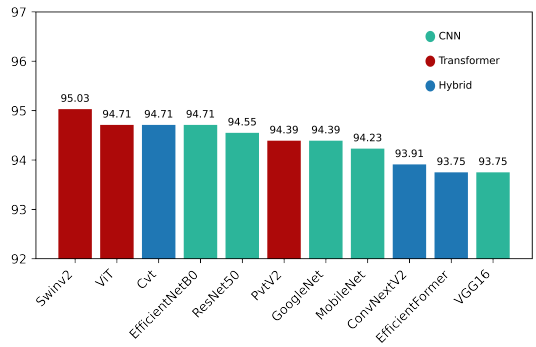
The comparison of results highlights SwinV2 as the best-performing model, achieving the highest scores in three out of four metrics: accuracy, recall, and F1-score. In the case of precision, it ranks third. CvT demonstrates the second-best overall performance among the models. Similar to SwinV2, it excels in accuracy, recall, and F1-score, ranking second, but falls short in precision. Among other transformer models, ViT demonstrates decent but inconsistent performance. In contrast, PvTV2 does not deliver noteworthy results. Excluding CvT, the hybrid models ConvNextV2 and EfficientFormer exhibit poor performance across all metrics, ranking third-to-last and second-to-last, respectively. Regarding CNN models, EfficientFormer and ResNet50 display the best precisions but are average or below average in other metrics. From this performance analysis, it can be inferred that while certain architectures show strong capabilities for diagnosing pneumonia from chest X-rays, there is no visible advantage of one architectural taxonomy over another for this specific task.

Figure 5 illustrates the average inference time for image classification across various models. Swinv2 demonstrates the longest average inference time at 133.58 ms, significantly exceeding the other models. CvT follows with an average inference time of 41.25 ms, though still substantially lower than Swinv2. Both Swinv2 and CvT are top-performing models, and their higher inference times suggest a trade-off between accuracy and speed. The remaining models have average inference times ranging from 12.44 ms to 22.55 ms, with VGG being an exception at 4.59 ms. Despite its low inference time, VGG is among the poorest performers based on the evaluation metrics, indicating that its speed advantage could be leveraged only by compensating its lower accuracy. The results also indicate that CNN models demonstrate better temporal performance compared to transformer and hybrid models, likely due to the heavier architecture of transformers. The exception is EfficientFormer, which is expected as it is specifically designed as a lightweight, low-latency vision transformer optimized for faster performance.

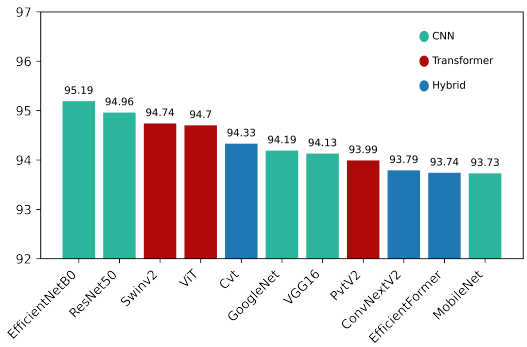
V. Conclusion

Vision classification is essential for the interpretation and understanding of visual data, making it one of the prominent research field of computer vision. The ability to accurately classify images facilitates advancements in automation and decision-making processes across various fields, including medical imaging applications. In recent years, image classification models have undergone significant advancements, driven by the development of sophisticated deep learning architectures. This paper identifies the popular image classification models and develop a comprehensive taxonomy to categorize these models systematically. Following this, a thorough evaluation of various ML models for pneumonia detection using the CXR dataset is conducted, comparing them based on the architectural taxonomy. Our study compared the performance of three categories of models: Convolutional Neural Networks (CNNs), Transformer models, and Hybrid models. We focused on key performance metrics such as accuracy, precision, recall, and F1-score. Temporal efficiency is also assessed by comparing the inference times of the models.

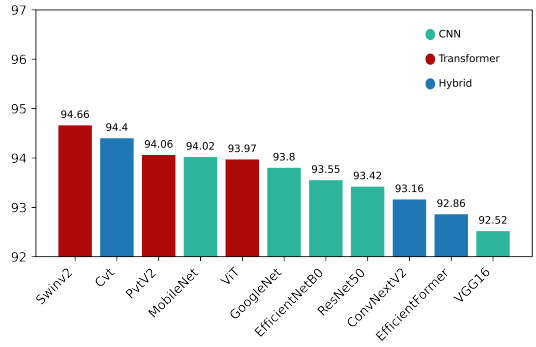
The performance analysis demonstrates that Transformer-based models, particularly SwinV2, consistently outperformed other models across multiple metrics. SwinV2 achieved the



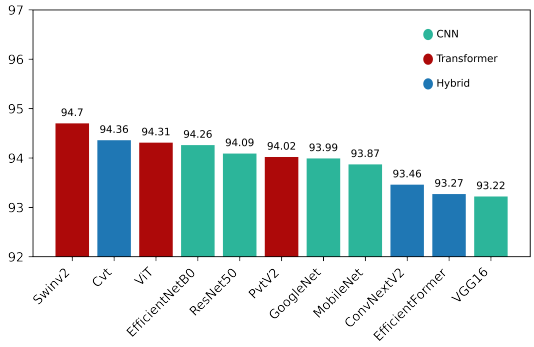
(a) Accuracy



(b) Precision



(c) Recall



(d) F1-score

Fig. 4: Performance comparison (in descending order)

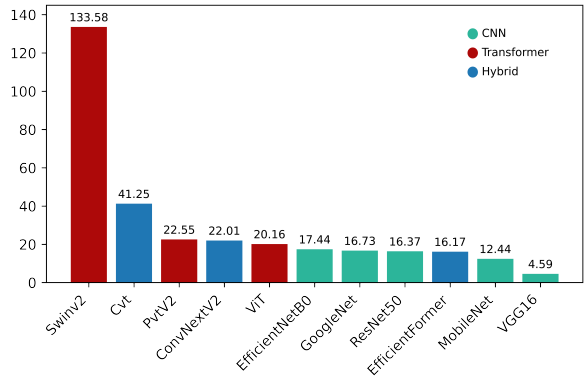


Fig. 5: Inference Time Comparison (in milliseconds)

highest accuracy (95.03%), recall (94.66%), and F1-score (94.70%), indicating its robustness and reliability in pneumonia detection. However, the inference time is too high for SwinV2 compared to rest of the models. Although CNN models like EfficientNetB0 and ResNet50 showed the highest precision and median inference time, they did not perform as well in other metrics. CvT, a hybrid model also shows good performance in classification task, although with a higher time. These illustrate the compromise between performance versus the speed of the models. While no evidence of a clear advantage is observed among the categories when performance is taken into account, the CNNs clearly exhibit superiority in speed compared to other models.

In future research, we plan to expand our investigation by exploring a wider variety of datasets to determine whether Transformer-based models consistently demonstrate superior performance across diverse applications.

REFERENCES

[1] M. Khalil, A. Khalil, and A. Ngom. A Comprehensive Study of Vision Transformers in Image Classification Tasks, 2023.

[2] E. Goceri. Analysis of Deep Networks with Residual Blocks and Different Activation Functions: Classification of Skin Diseases. In *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 2019.

[3] O. M. Moushi, N. Ara, M. Helaluddin, and H. S. Mondal. Enhancing the accuracy and explainability of heart disease prediction models through interpretable machine learning techniques. In *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 1–6, 2023.

[4] J. Li, Y. Yan, S. Liao, X. Yang, and L. Shao. Local-to-Global Self-Attention in Vision Transformers, 2021.

[5] P. R. Jeyaraj, E. R. Samuel Nadar, and B. K. Panigrahi. ResNet Convolution Neural Network Based Hyperspectral Imagery Classification for Accurate Cancerous Region Detection. In *2019 IEEE Conference on Information and Communication Technology*, 2019.

[6] K. Zhao, Y. Huo, L. Xue, M. Yao, Q. Tian, and H. Wang. Mushroom Image Classification and Recognition Based on Improved Swin Transformer. In *2023 IEEE 6th International Conference on Information Systems and Computer Aided Education (ICISCAE)*, 2023.

[7] D. Say, S. Zidi, S. M. Qaisar, and M. Krichen. Automated Categorization of Multiclass Welding Defects Using the X-ray Image Augmentation and Convolutional Neural Network. *Sensors*, 23:6422, 14, 2023.

[8] A. Sharma, A. Zehra, A. Das, K. Rastogi, M. Agarwal, S. Mascarenhas, J. J, V. M, and D. S. Brain Tumor Classification: A Comparison Study CNN, VGG 16 and ResNet50 Model. In *2023 International Conference on Data Science and Network Security (ICDSNS)*, 2023.

[9] J. Wongbongkotpaisan and S. Phumeechanya. Plant Leaf Disease Classification using Local-Based Image Augmentation and Convolutional Neural Network. In *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2021.

[10] V. R. N. Murthy Teki, R. Anandha Ragaven, N. Manoj, V. V, and S. S. A Comparison of Two Transformers in the Study of Plant Disease Classification. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2023.

[11] I. Shiri, Y. Salimi, N. Sirjani, A. H. Aval, Z. Mansouri, M. Amini, A. Saberi, G. Hajianfar, M. Pakbin, M. G. Oghli, M. Oveisi, and H. Zaidi. Deep Vision Transformers for Prognostic Modeling in COVID-19 Patients using Large Multi-Institutional Chest CT Dataset. In *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–3, 2022.

[12] N. S. Kumar and B. Ramaswamy Karthikeyan. Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures. In *2021 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2, 2021.

[13] M. S. Islam, M. A. T. Rony, and T. Sultan. Gastrovrg: enhancing early screening in gastrointestinal health via advanced transfer features. *Intelligent Systems with Applications*, 23:200399, 2024.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 1989.

[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, Sept. 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In June 2016.

[18] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, 2017.

[19] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[20] M. Tan and Q. V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *ArXiv*, abs/1905.11946, 2019.

[21] M. S. I. Sajol, A. S. M. J. Hasan, M. S. Islam, and M. S. Rahman. Transforming social media analysis: tweeteval benchmarking with advanced transformer models. In *2024 8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6, 2024.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

[23] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[24] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[25] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[26] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. PVT v2: Improved baselines with Pyramid Vision Transformer. In volume 8, 2022.

[27] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[28] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang. CvT: Introducing Convolutions to Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[29] M. S. I. Sajol and A. S. M. J. Hasan. Benchmarking cnn and cutting-edge transformer models for brain tumor classification through transfer learning. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, pages 1–6, 2024.

[30] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan. FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization, 2023.

[31] H. Hu, Z. Zhang, Z. Xie, and S. Lin. Local Relation Networks for Image Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Los Alamitos, CA, USA. IEEE Computer Society, Nov. 2019.

[32] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren. EfficientFormer: Vision Transformers at MobileNet Speed, 2022.

[33] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] M. S. Islam Sajol, A. S. M Jahid Hasan, M. S. Islam, and M. S. Rahman. A convnext v2 approach to document image analysis: enhancing high-accuracy classification. In *2024 IEEE 3rd Conference on Information Technology and Data Science (CITDS)*, pages 1–6, 2024.

[35] D. Kermany, K. Zhang, and M. Goldbaum. Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification, 2018.

[36] A. Abubakar, M. Ajuji, and I. U. Yahya. DeepFMD: Computational Analysis for Malaria Detection in Blood-Smear Images Using Deep-Learning Features. *Applied System Innovation*, 4:82, 4, 2021.