Hands-on Evaluation of Visual Transformers for Object Recognition and Detection

Dimitrios N. Vlachogiannis
Dept. of Computer Engineering and Informatics
University of Patras
Patras, Greece
st1067371@ceid.upatras.gr

Dimitrios A. Koutsomitropoulos
Dept. of Computer Engineering and Informatics
University of Patras
Patras, Greece
koutsomi@ceid.upatras.gr

2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI) | 979-8-3315-4919-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICTAI66417.2025.00125

***Abstract*—Convolutional Neural Networks (CNNs) for computer vision sometimes struggle with understanding images in a global context, as they mainly focus on local patterns. On the other hand, Vision Transformers (ViTs), inspired by models originally created for language processing, use self-attention mechanisms, which allow them to understand relationships across the entire image. In this paper, we compare different types of ViTs (pure, hierarchical, and hybrid) against traditional CNN models across various tasks, including object recognition, detection, and medical image classification. We conduct thorough tests on standard datasets like ImageNet for image classification and COCO for object detection. Additionally, we apply these models to medical imaging using the ChestX-ray14 dataset. We find that hybrid and hierarchical transformers, especially Swin and CvT, offer a strong balance between accuracy and computational resources. Furthermore, by experimenting with data augmentation techniques on medical images, we discover significant performance improvements, particularly with the Swin Transformer model. Overall, our results indicate that Vision Transformers are competitive and, in many cases, outperform traditional CNNs, especially in scenarios requiring the understanding of global visual contexts like medical imaging.**

***Index Terms*—Visual Transformers,Object Recognition, Object Detection, Medical Imaging, Data Augmentation, Swin, ChestX-ray14**

I. Introduction

The rapid development of deep learning and artificial intelligence over the last decade has transformed computer vision, enabling machines to interpret and understand images and videos with greater accuracy than ever before. Convolutional Neural Networks (CNNs) have been the default approach to visual recognition tasks for over a decade, doing both image classification and object detection due to their ability to learn local spatial hierarchies through convolutional operations. Architectures such as AlexNet [1], VGGNet [2], ResNet [3] for object recognition and Faster R-CNN [4], Yolo [5] for object detection have established CNNs as the dominant architecture in computer vision. However, CNNs often face limitations related to capturing global context effectively, which is crucial for a more in-depth insight into visual scenes and robust performance, especially in challenging conditions [6], [24].

The Transformer model, as presented by Vaswani et al. [8], has revolutionized the domain of natural language processing (NLP) by effectively capturing long-range dependencies through the self-attention mechanism, thus enabling models like BERT [7] and GPT [10] to achieve state-of-the-art results. Inspired by these successes, the Vision Transformer (ViT) [9] was devised, adapting Transformer models to vision tasks by framing images as sequences of image patches processed via self-attention. Unlike CNNs, ViTs inherently capture global image features and significantly improve their capacity to identify spatial relationships present across the entire image.

Furthermore, the robustness and application of Vision Transformers in specialized domains, notably medical imaging, have become important topics. Transformers have demonstrated enhanced resilience against adversarial perturbations compared to CNNs [21], [24], [25], attributed to their global context capabilities. Additionally, in medical imaging scenarios, ViTs have showcased superior performance, reduced sensitivity to hidden stratification, and improved generalization across diverse datasets [23], [28], [29].

Motivated by these latest developments, this paper presents a comprehensive hands-on evaluation of various Vision Transformer architectures for object recognition, detection, and medical image analysis tasks.The main contributions of this work are the following:

- Comprehensive Comparative Evaluation: We present an extensive experimental evaluation and comparative analysis of Transformer-based architecture (pure, hierarchical, and hybrid ViT models) against established CNN benchmarks on both image classification (ImageNet-1K) and object detection (COCO) datasets.
- We extend our evaluation to medical image classification, utilizing the ChestX-ray14 dataset. We demonstrate the effectiveness of Vision Transformers, particularly that of hybrid and hierarchical ViTs.
- We investigate specific data augmentation techniques (CutMix, MixUp, Random Augmentations)

2375-0197/25/$31.00 ©2025 IEEE
DOI 10.1109/ICTAI66417.2025.00125

and observe their impact on the hierarchical model (Swin). Notably, these techniques have not been previously applied to the pure Swin model on the ChestX-ray14 dataset.
- Reproducible Experimental Framework: All experiments are performed using publicly accessible tools and platforms (HuggingFace Transformers and Trainer), with publicly available code [32] and model checkpoints.

The rest of this paper is organized as follows: In the next Section II we review related work on ViTs and present a comparative evaluation of the different models based on the literature, differentiating between pure Transformer architectures and hybrid approaches. In Section III we discuss our evaluation methodology and present evaluation results for the two tasks of object recognition and detection accordingly, comparing performance between the different ViT architectures as well as traditional CNNs. In Section IV we focus on the medical imaging problem domain and evaluate the most promising architecture representatives found in the previous section on object detection in the ChestXray14 dataset. The potential of data augmentation techniques on the best model (Swin-base) is also investigated and reported. Finally, Section V summarizes our conclusions and future work.

II. Background and Related Work

A. ViT for Object Recognition

ViT [9] has demonstrated excellent results in various image recognition benchmarks (ImageNet [11], CIFAR-100 [12]) and competitive results compared with state-of-the-art CNN models. By taking advantage of the self-attention mechanism to incorporate information from the entire image from the earliest stages, it achieves a better understanding of global correlations compared to CNNs that rely mainly on local filters [9], [24], [26]. In addition, ViT offers increased flexibility in handling high-resolution images through patch processing [9], while exhibiting greater robustness to noise and corruption thanks to global feature integration [21], [24], [25]. Despite its benefits, ViT encounters some drawbacks. This includes the significant computational complexity linked to its self-attention mechanism that demands substantial computational resources [9]. Moreover, because transformers lack inherent inductive biases, ViT generally requires extensive datasets to learn efficiently [26]. Finally, ViT frequently shows deficiencies in capturing local features, impacting its effectiveness in tasks like small object detection [17].

To limit the problems that the original ViT had, certain variations have been proposed which aim to enhance the locality and the self-attention mechanism. PVT [13] is a hierarchical version of ViT, implementing multiple resolution scales, allowing for efficient modeling of details at different levels. This approach makes it particularly effective in classifying high-resolution images, improving accuracy with lower computational cost. Swin Transformer [14] introduced the shifted windows mechanism, reducing the self-attention cost from quadratic to linear. With this approach, it maintained high classification accuracy, making it one of the most efficient Transformer-based architectures for visual tasks. CvT [15] is a hybrid model that incorporates convolutional operations in the early stages of the Transformer to exploit the powerful local modeling of CNNs. This hybrid approach achieves high performance in image classification while simultaneously leveraging the advantages of CNNs and Transformers. LeViT [16] combines the advantages of CNNs and Transformers in a lightweight architecture, designed to provide efficient classification with a reduced number of parameters and computational complexity. The incorporation of convolutions in the early stages makes it ideal for applications requiring high speed and low resource usage. Table I shows a comparison between all the presented works according to their size, computational requirements and top-1 accuracy on ImageNet.

TABLE I: ImageNet result comparison of different vision transformer architectures

Model	#Params (M)	FLOPs (G)	Top-1(%)
Pure Transformer			
ViT-B [9]	86.6	17.6	83.9
ViT-L [9]	307	61.6	85.1
PvT-Medium [13]	44.2	6.7	81.2
PvT-Large [13]	61.4	9.8	81.7
Swin-B [14]	88	15.4	83.3
Swin-L [14]	197	104	86.4
Hybrid			
ViT-Hybrid-B [9]	99	49.6	85.5
CvT-21 [15]	32	7.1	82.5
CvT-21-384 [15]	32	24.9	83.3
LeViT-256 [16]	18.9	1.12	81.6
LeViT-384 [16]	39.1	2.35	82.6

B. ViT for Object Detection

Traditional object detectors are mainly based on CNNs, but after the exceptional performance shown by transformers in the field of classification, they began to be introduced into object detection. They can be categorized into two groups:

1) Detection Transformers with CNN Backbone: DETR [17] was the first fully Transformer-based model for object detection, introducing an end-to-end detection mechanism without the need for anchors or region proposal networks. Although simple in design, it exhibited relatively slow convergence and poor performance on small objects. Deformable DETR [18] is an improvement on DETR by introducing deformable attention that focuses only on a small number of critical points. This drastically reduces training time and improves performance on small objects. Swin Transformer [14] is often used as a backbone in object detection systems such as Cascade Mask R-CNN, offering hierarchical features with window-based self-attention. Its use allows state-of-the-art performance on the COCO dataset, with very good scaling in depth and resolution.

2) Detection with Pure Transformers: YOLOS [19] is an application of ViT directly to object detection, treating the input as a sequence of patches, without any special adaptation for localization. Although simple, it lacks accuracy compared

to more specialized architectures but shows that pure Vision Transformers have capabilities beyond classification. Table II shows a comparison between the above models according to their size, computational requirements and average precision (AP) for object detection on COCO 2017 validation set.

TABLE II: Performance comparison of transformer-based object detectors on the COCO 2017 validation set.

Model	AP	AP_{50}	AP_{75}	#Params (M)	FLOPs (G)
Pure Transformer					
YOLOS-Tiny [19]	0.300	–	–	6.5	21
YOLOS-Base [19]	0.420	0.622	0.445	127.8	537
CNN Backbone (Hybrid)					
DETR-R50 [17]	0.420	0.624	0.442	41.3	86
DETR-R101 [17]	0.433	0.631	0.459	41.0	187
Cond.-DETR-R50 [20]	0.430	0.640	0.457	44.0	90
Def.-DETR [18]	0.462	0.652	0.500	40.0	173

C. Special Features of ViTs

In addition to ViT performance on basic problems (classification, object detection), there is particular interest in studying their robustness against perturbation, as well as their utilization in medical imaging applications. These two features occur from the ability of ViTs to capture global dependencies through the self-attention mechanism, which has been shown to enhance stability against adversarial attacks [21] and accuracy in medical data [23].

1) Robustness: Recent studies show that Vision Transformers (ViTs) exhibit higher durability to attacks and disturbances compared to traditional CNNs, this is attributed to the self-attention mechanism, which allows the efficient exploitation of global image features [21], [24]. The study by Benz et al. [21] found that ViTs require larger distortions to be fooled by white-box attacks (such as PGD and FGSM), while CNNs show vulnerabilities even at low levels of distortion. It was also found that adversarial examples generated in CNNs do not easily transfer to ViTs. The same study showed that Transformers rely more on low-frequency features of the image, which are tolerant to attacks, while CNNs showed that they depend more on high-frequency features, which are more vulnerable. At the same time, according to Bai et al. [24] in a fair comparison between ViTs (e.g. DeiT-S [26]) and CNNs (e.g. ResNet-50 [3]) of the same size and performance, ViTs show better resilience to adversarial attacks and generalize better to out-of-distribution data (such as the ImageNet-A [33] and ImageNet-C [22] datasets). However, it was observed that the advantage of ViTs in robustness decreases when CNNs are trained with similar techniques to them. Finally, comparisons that included hybrid models that combine CNN and ViT features, namely local and global features, showed even higher robustness [25].

2) Medical Imaging: Unlike natural image recognition and object detection, medical imaging presents unique challenges. Datasets are often much smaller and imbalanced, while disease patterns may be spread across different regions of the image rather than localized in well-defined objects [34]. This makes the global context modeling of Vision Transformers particularly useful, since self-attention can capture long-range dependencies that CNNs often miss. Based on a recent study, a key finding is that ViTs tend to be less affected by hidden stratification, that is, situations where the model bases its prediction on random or irrelevant features of the image instead of essential medical indicators [28]. This is of particular importance to diagnostic accuracy, as it reduces the risk of incorrect predictions due to bias in the data. Furthermore, in the study [23], ViTs outperformed established CNNs in both accuracy and generalization for the classification of emphysema types from CT scans, demonstrating an improved ability to adapt to new and independent datasets. Similarly, in breast ultrasound image classification, ViT-B/32 achieved an accuracy of 86.7% and an AUC of 0.95, surpassing leading CNNs in some cases, while also enabling better interpretability through the use of attention maps [29].

III. Evaluation of Object Recognition and Detection

A. Configuration

The models we used for object recognition are shown in Table III, while those we used for object detection are shown in Table IV. For each model in the corresponding process, we used the same framework, dataset, evaluation hyperparameters and GPU.

1) Pre-training: For object recognition, most models are pre-trained on the ImageNet-1k dataset, except for the ViT models, which are pre-trained on the larger ImageNet-21k and then fine-tuned on ImageNet-1k, as they depend on pre-training on very large datasets followed by fine-tuning, as mentioned in [9]. In the case of object detection, all models are pre-trained on the Coco dataset, except for the YOLOS models, which are pre-trained on ImageNet-1k and then fine-tuned on Coco, as was originally proposed in [19]. For evaluation we used the corresponding datasets' subsets for the two tasks. Experiments were carried out using the HuggingFace API and Google Colab, offering a Nvidia Tesla T4 GPU with 16GB VRAM.

2) Datasets: For the Object Recognition task we used the ImageNet-1k dataset. ImageNet-1k [30] is the most well-known subset of the ImageNet-21k and it's used more widely. It includes 1,281,167 training images, 50,000 evaluation, and 100,000 test images, divided into 1,000 classes. For the purposes of this work, we only use the ImageNet-1k validation subset.

For the Object Detection task we used the Coco dataset. Coco [31] is the most well-known benchmark for evaluating models in object detection. It is a large-scale publicly available dataset that includes images with detailed annotations for object detection and provides over 328,000 images of everyday

scenes divided into 80 object categories. For the evaluation we only used the validation subset of Coco.

B. Metrics

For classification, model evaluation is done by measuring accuracy. Accuracy is the most common method for evaluating classification models. It is the ratio of the total number of correct predictions made by the model to the total number of predictions it made.

$$\text{Accuracy} = \frac{\text{(Number of correct predictions)}}{\text{(Total number of predictions)}}$$

For the evaluation in object detection, the metrics Precision, Recall, IoU and mAP are used.

Precision: Precision indicates the model's ability to make accurate positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall focuses on the model's ability to correctly identify positive samples from the entire set of positive cases.

$$Recall = \frac{TP}{TP + FN}$$

Intersection over Union (IoU): IoU is a metric that evaluates the extent of overlap between two bounding boxes, providing a measure of how well a predicted object aligns with its true counterpart.

Mean Average Precision (mAP): In object detection, mAP measures the accuracy of placing bounding boxes by comparing them to the actual ones through the IoU metric. To calculate mAP, we first calculate the Average Precision (AP) for each category, creating precision-recall curves and finding the area under the curve (AUC). The final mAP score is the average of the AP scores for all categories, providing an overall performance index.

$$mAP = \frac{1}{C} \sum_{i=1}^{C} AP_i$$

AP50 / AP75: They correspond to the average precision for specific IoU thresholds, for IoU=0.50 and IoU=0.75.

APs, APm, APl: AP metrics for small (APs), medium (APm), and large (APl) objects, based on the object's size in the COCO dataset. They provide insight into how well a model performs on objects of different scales. The same applies to ARs, ARm, ARl.

Average Recall (AR): AR is the maximum recall given a fixed number of detections per image, averaged over categories and IoUs. AR1, AR10, AR100: The average recall given 1, 10, or 100 detections per image, respectively. These metrics measure the model's ability to correctly find positive instances under varying maximum numbers of predictions.

TABLE III: CNN, Transformer, and Hybrid Models For Object Recognition Evaluation

Model	#Params (M)	Image size	FLOPs (G)
CNN			
ResNet-101 [3]	44.5	224^2	7.9
ResNet-152 [3]	60.2	224^2	11.6
EfficientNet-B0 [40]	5.3	224^2	0.4
EfficientNet-B4 [40]	19.0	380^2	4.2
EfficientNet-B7 [40]	66.3	600^2	37.1
Transformer			
ViT-B/16 [9]	86.6	224^2	17.6
ViT-L/16 [9]	304.3	224^2	61.6
PvT-medium [13]	44.2	224^2	6.7
PvT-large [13]	61.4	224^2	9.8
Swin-B [14]	87.8	224^2	15.4
Swin-L [14]	196.5	224^2	34.5
Hybrid			
CvT-21 [15]	31.6	224^2	6.6
CvT-21-384 [15]	31.6	384^2	19.5
LeViT-256 [16]	18.9	224^2	0.6
LeViT-384 [16]	39.1	384^2	2.1
ViT-Hybrid-Base [9]	99	384^2	49.6

TABLE IV: CNN, Transformer, and Hybrid Models For Object Detection Evaluation

Model	#Params (M)	Image size	FLOPs (G)
CNN			
Faster R-CNN [4]	41.5	800^2	134.7
RetinaNet [42]	33.8	800^2	151.9
SSD300 [41]	35.6	300^2	35
Transformer			
Yolos-Tiny [19]	6.5	512×768	21.4
Yolos-Base [19]	127.8	512×768	190.1
Hybrid			
Detr-ResNet-50 [17]	41.3	873×1201	102
Detr-ResNet-101 [17]	60.2	873×1201	181.4
Conditional-Detr-R50 [20]	43.2	873×1201	106.2
Deformable-Detr [18]	40	512×768	173

C. Object Recognition

Table V presents our evaluation results of various model architectures on the ImageNet dataset. As observed, the original Vision Transformer (ViT) achieves highly competitive performance compared to traditional CNNs, with ViT-Large reaching an accuracy of 82.5%, surpassing landmark models such as ResNet and EfficientNet.However, this performance comes at the cost of a significantly increased number of parameters (304.3M) and high computational cost (61.6 GFlops).

Significant superiority is observed in the Swin Transformer models, where Swin-Large records the highest accuracy (86%), surpassing all other models, both CNN and other ViT or hybrid models.This superiority is attributed to the Swin architecture, which combines the hierarchical feature extraction of CNNs with the shifted windows mechanism in self-attention, enabling the simultaneous exploitation of both global and local correlations within an image. Furthermore, Swin-Large maintains a significantly lower number of parameters (196.5M) and computational cost (34.5 GFlops) compared to ViT-Large.

The CvT (Convolutional Vision Transformer) models are also distinguished for their balance between performance and computational cost. For example, CvT-21-384 achieves an accuracy of 82.2% with only 31.6M parameters and 19.2 GFlops, values much lower than those of the corresponding ViT and Swin models. The CvT architecture is based on the combination of convolutional and attention mechanisms, drawing advantages from both approaches and making the hybrid strategy particularly effective.

TABLE V: Object Recognition results sorted by increasing Accuracy number

Model	Acc.	#Params (M)	FLOPs (G)
Efficientnet-b0	0.72578	5.3	0.4
Levit-256	0.80894	18.4	1.1
CvT-21	0.81266	31.6	6.6
PvT-Medium-224	0.81268	44.2	6.7
ViT-Base-patch16-224	0.81675	86.6	17.6
Pvt-Large-224	0.81560	61.4	9.9
Efficientnet-b4	0.81612	19.0	4.2
Resnet-101	0.81820	44.5	7.9
Levit-384	0.82034	39.1	2.1
Cvt-21-384	0.82294	31.6	19.5
Vit-Large-patch16-224	0.82512	**304.3**	**61.6**
Resnet-152	0.82616	60.2	11.6
Efficientnet-b7	0.83288	66.3	37.1
Swin-Base-patch4-window7-224	0.84820	87.8	15.5
ViT-hybrid-base-bit-384	0.84938	99.0	49.6
Swin-Large-patch4-window7-224	**0.86018**	196.5	34.5

D. Object Detection

Tables VI and VII along with Fig. 1a and 1b present the object detection results on the COCO dataset. Firstly, YOLOS—a pure Vision Transformer for object detection—manages to surpass several traditional CNN models in both mAP (39.4%) and mAR (54.6%), demonstrating the potential of transformer-based approaches in object detection. This significant result becomes even more noteworthy considering that YOLOS does not incorporate hierarchical or pyramidal feature processing (as CNNs do), but instead relies solely on a self-attention mechanism without inductive bias. This allows the model to better capture global dependencies in the image, resulting in high performance in the detection of large objects (mAPl 61.2%, mARl 81.4%). However, the lack of local processing significantly limits its ability to detect small objects (mAPs 16%, mARs 26%) as can be observed from Fig.1a, 1b. Additionally, YOLOS shows increased parameter requirements (127.8M) and computational cost (190.1 GFlops), mainly due to the need for pre-training and fine-tuning, without necessarily justifying its superiority over other, more efficient models.

The DETR and Deformable DETR models, which adopt hybrid architectures, achieve even higher performance. DETR with a ResNet-101 backbone reaches an mAP of 43.4% and an mAR of 59.0%, while demonstrating leading results in the detection of large objects (mAPl 61%, mARl 81%), surpassing both established CNNs and YOLOS. This hybrid approach leverages both the global attention capabilities of transformers and the local information captured by CNNs, thereby enhancing the overall image understanding ability. Nevertheless, DETR still lags behind in the detection of small objects as demonstrated in Fig.1a, 1b (mAPs 21%, mARs 34%).

Deformable DETR stands out as the most efficient model, achieving the highest values in almost all performance metrics: mAP44.5%, mAPs 26%, mAPl 59%, and mAR 62.9%, mARs 40%, mAPl 82%. Its success is attributed to the deformable attention mechanism, which enables the model to focus on selected regions and across multiple scales, thereby enhancing the detection of small objects without compromising performance on medium and large objects.

TABLE VI: Object Detection results sorted by increasing AP

Model	mAP	AP@50	AP@75	#Params (M)	FLOPs (G)
SSD300	0.251	0.416	0.263	35.6	35
Yolos-Tiny	0.287	0.472	0.289	6.5	21.4
RetinaNet	0.363	0.556	0.382	33.8	151.9
Faster R-CNN	0.369	0.585	0.398	41.5	134.7
Yolos-Base	0.394	0.592	0.414	**127.8**	**190.1**
Conditional-Detr-ResNet-50	0.409	0.617	0.435	43.2	106.2
Detr-ResNet-50	0.420	0.623	0.442	41.3	102
Detr-ResNet-101	0.434	**0.638**	0.462	60.2	181.4
Deformable-Detr	**0.445**	0.636	**0.486**	40	173

TABLE VII: Object detection results sorted by increasing AR

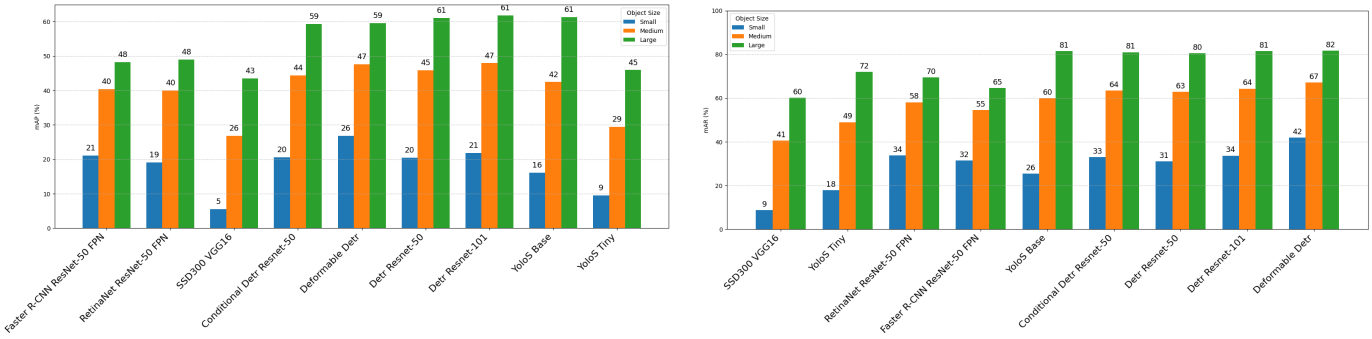
Model	mAR1	mAR10	mAR	#Params (M)	FLOPs (G)
SSD300	0.239	0.344	0.365	35.6	35
Yolos-Tiny	0.264	0.423	0.460	6.5	21.4
RetinaNet	0.313	0.499	0.539	33.8	151.9
Faster R-CNN	0.307	0.485	0.508	41.5	134.7
Yolos-Base	0.324	0.511	0.546	**127.8**	**190.1**
Conditional-DETR-ResNet-50	0.334	0.540	0.581	43.2	106.2
DETR-ResNet-50	0.333	0.532	0.574	41.3	102
DETR-ResNet-101	0.344	0.548	0.590	60.2	181.4
Deformable-DETR	**0.352**	**0.587**	**0.629**	40	173

IV. Evaluation on Medical Imaging

A. Configuration

Best models per architecture group in table V differ significantly in size. Therefore, for the experiments in this section we selected best representative models from each category that are close both in terms of performance as well as computational cost: Transformer (Swin-Base), Hybrid (CvT-21-384), and CNN (ResNet-152).

1) Datasets: For the Medical Imaging evaluation we used Chest-Xray14. The ChestX-ray14 [35] is a large dataset of chest X-rays created to enhance deep learning applications in medical image analysis, and more specifically in the classification of common chest diseases. The dataset consists of 112,120 anterior chest X-rays extracted from 30,805 patients. It is divided into 14 chest diseases, selected based on frequency and diagnostic significance. Each image can also have multiple disease labels.



(a) mAP per Object Size

(b) mAR per Object Size

Fig. 1: Results on COCO dataset: (a) mAP and (b) mAR per object size.

For the purposes of this study N-hot encoding was employed to convert text labels into arrays that represent the multiple labels each image may have. An N-hot encoded array represents the labels as a list of binary digits: "1" if the label applies to the image and "0" if it does not.

2) Metrics: For the Chest-Xray dataset we used the AUC-ROC metric.

AUC-ROC: The ROC (Receiver Operating Characteristic) curve and the AUC (Area Under the Curve) are important evaluation tools for the performance of classification models. The ROC curve is a graphical representation of the True Positive Rate (TPR) versus the False Positive Rate (FPR) for different decision thresholds. The AUC, or the area under the ROC curve, indicates how well a model can separate the categories, with a value close to 1 indicating high separation ability, while a value close to 0.5 indicates that the model does not separate the categories better than random selection. We use ROC-AUC because it is widely used to measure performance in multi-label classification tasks.

3) Fine-tuning: Fine-tuning was conducted for 3 epochs with a batch size of 16 for both training and evaluation, which provides a good balance between GPU memory usage and model convergence. Data loading was managed using 2 worker threads for increased efficiency, with the data loader set to drop the last batch if the dataset length was not divisible by the batch size.

The optimizer utilized was AdamW, which is known for its effectiveness in fine-tuning transformer-based models. The learning rate was initialized at 2×10^{-4}, a value commonly adopted for such settings. Learning rate scheduling followed a cosine decay policy with a warmup ratio of 0.25; specifically, during the first 25% of the training steps, the learning rate was gradually increased, which helps stabilize training before transitioning to the cosine decay phase. This strategy is widely used in transformer-based architectures to prevent divergence during the initial phase of training.

Logging was performed every 50 steps to monitor training progress, while model checkpoints were saved every 1,000 steps, enabling model recovery if necessary. Model selection was carried out by evaluating performance at the end of each epoch, and at the conclusion of training, the model with the highest ROC AUC score on the validation set was automatically selected and loaded. Additionally, the best-performing model was configured to be pushed to the HuggingFace Hub to ensure reproducibility.

B. Optimization

As shown in [38] ViTs exhibit limited performance when trained on small or insufficient datasets, such as those found in medical imaging as mentioned earlier in [27]. This limitation was attributed to their lack of inductive bias toward local image structures. The main technique proposed in [38] to address this issue is data augmentation. The data augmentation techniques we utilized are the following:

1) Basic Random Augmentations:

- Random Horizontal Flip: In each training epoch, every image has a 50% chance of being flipped horizontally.
- Random Rotation: Each image is rotated by a random angle within a specified range $[-15°, 15°]$.
- Color Jitter: Each image undergoes a random change in brightness and contrast.

2) CutMix: CutMix is an augmentation technique that works as follows:

- First, two images are randomly selected from the dataset.
- Then, a random rectangular patch of the same size and random position is cut from both images.
- The patch cut from the first image is placed in the empty area created in the second image, and vice versa for the first image.
- Finally, two new labels are created for the two new images, which are a linear combination of the two original labels.

3) MixUp: MixUp is an augmentation technique that creates a new sample by combining two existing ones and works as follows. Here, x represents the data and y the corresponding labels. The new sample $(\hat{x}, \hat{y})$ is constructed as:

$$\hat{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\hat{y} = \lambda y_i + (1 - \lambda) y_j$$

where $\lambda \in [0, 1]$ is the mixing coefficient. This means that the new image is intermediate between the two originals, and its label reflects the contribution of each original image.

Techniques such as CutMix and MixUp have demonstrated their effectiveness on ViT models in [39] for medical imaging tasks. For this experiment, we utilized the Swin-Base model to evaluate these techniques on a variant of the ViT architecture, the model was trained for 10 epochs using the same hyperparameters as previously described. Furthermore, the technique of *Early Stopping* was applied, terminating the training process if the ROC AUC score did not improve for three consecutive epochs.

TABLE VIII: Comparison of models on ChestX-ray14

Model	#Epoch	#Params (M)	FLOPs (G)	Eval Loss	Eval ROC-AUC
Swin-Base	3	**86.8**	15.5	0.1676	0.8174
Cvt-21-384-22k	3	31.2	**19.2**	0.1671	**0.8219**
ResNet-152	3	58.2	11.6	0.1729	0.8113

TABLE IX: Comparison of Swin-Base data augmentatio techniques on ChestX-ray14

Model Variant	#Epoch	Eval Loss	Eval ROC-AUC
Swin-Base	3	0.1697	0.8174
Swin-Base + Data Aug	10	0.1693	0.8430
Swin-Base + CutMix	9	0.1737	0.8290
Swin-Base + MixUp	10	0.1713	0.8361
Swin-Base + Data Aug + MixUp	10	**0.1690**	**0.8525**

C. Results and discussion

From the results on the ChestX-ray14 dataset (Table VIII), it is observed that Transformer-based models outperformed the CNN, with the CvT-21-384 hybrid model in particular achieving the highest performance, with a ROC-AUC of 82.1%, compared to 81.7% for Swin-Base and 81.1% for ResNet-152. This superiority of Transformer architectures aligns with recent findings [23], [28], [29], [36], which attribute their effectiveness to the modeling of global correlations in medical images through the self-attention mechanism.In such images, pathologies often manifest as diffuse or distant visual patterns that cannot be sufficiently captured by local filters alone, as used in CNNs [37]. CvT, by leveraging convolutional token embeddings in combination with self-attention, manages to combine the advantages of both CNNs and Transformers, achieving a better balance between local and global information. As a result, it performs better while having a significantly lower number of parameters and only slightly higher computational cost. This increased computational cost is due to CvT's hybrid nature, performing both attention and convolution operations, and also because it processes higher resolution inputs (384×384) compared to the other two models (224×224). Swin also demonstrates competitive performance thanks to its hierarchical shifted window structure, which mimics the architecture of CNNs in order to also capture local dependencies.

From the evaluation of data augmentation techniques applied to the Swin-Base model on the ChestX-ray14 dataset (Table IX), it becomes clear that augmentations significantly boost model performance. The combination of data augmentation techniques, specifically standard augmentation with MixUp, got the best result, by achieving a ROC-AUC of 85.25%, outperforming the baseline Swin-Base model by approximately 4%. Individual augmentation techniques, Data Augmentation alone (84.30%) and MixUp alone (83.61%), also show clear improvements, confirming the beneficial effect of each technique independently.However, CutMix demonstrates a lower performance (82.90%), suggesting that its specific augmentation approach might be less effective for this particular dataset.

V. Conclusions and future work

It becomes evident that Vision Transformers (ViTs) can perform just as well or even better than traditional CNNs in tasks like image object recognition, object detection, and especially in medical imaging. The results showed that hybrid and hierarchical models like CvT and Swin Transformer manage to combine the best of both architectures, offering high accuracy with a good balance of computational cost. Also, by trying out data augmentation techniques such as the combination of Basic Augmentations and MixUp—especially on the Swin model for medical images—we saw a noticeable improvement in performance, proving that these methods can help ViTs on smaller datasets like ChestX-ray14. Overall, our study confirms that Transformers are very competitive in computer vision and medical tasks.

In the future, it would be worth exploring even more data augmentation techniques and investigate their impact on Transformer models, especially when applied to more challenging or imbalanced medical datasets. In addition, it is particularly important to evaluate more lightweight or optimized Transformer architectures, so that they can be effectively utilized in real-time or low-resource environments, such as on mobile devices or in hospitals with limited hardware capabilities.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 1097–1105, Curran Associates, Inc., 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 91–99, Curran Associates, Inc., 2015.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[6] P. Benz, S. Ham, C. Zhang, A. Karjauv, and I. S. Kweon, "Adversarial robustness comparison of vision transformer and MLP-mixer to CNNs," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. [Online]. Available: https://arxiv.org/abs/2110.02797

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. Available: https://aclanthology.org/N19-1423/

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

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

[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

[12] A. Krizhevsky, G. Hinton, "Learning Multiple Layers of Features from Tiny Images," Technical Report, University of Toronto, 2009.

[13] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, and J. Dai, "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568–578, 2021.

[14] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[15] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing Convolutions to Vision Transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22–31, 2021.

[16] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "LeViT: A Vision Transformer in ConvNet's Clothing for Faster Inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12259–12269, 2021.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[19] Y. Fang, S. Wang, M. Li, H. Zhang, J. Yang, W. Liu, and X. Wang, "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 9719–9734, 2021.

[20] A. Meng, X. Chen, Z. Fan, G. Zeng, Y. Li, Y. Yuan, J. Sun, and C. Wang, "Conditional DETR for Fast Training Convergence," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3651–3660, 2021.

[21] S. Bhojanapalli, A. Chakrabarti, D. Glasner, T. Chanda, S. Krishnan, A. Levy, and H. Mishra, "Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 22023–22034, 2021.

[22] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. [Online]. Available: https://arxiv.org/abs/1903.12261

[23] B. Wu, X. Zhou, C. Wen, Y. Jin, W. Li, and Y. Ma, "A Vision Transformer for Emphysema Classification Using CT Images," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1040–1044, 2021.

[24] S. Paul and S. Chen, "Are Transformers More Robust Than CNNs?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23840–23850, 2023.

[25] S. B. A. A. G. Dey, L. Xiang, X. Li, R. C. Deo, and H. Madapana, "Exploring the differences in adversarial robustness between ViT-and CNN-based models using novel metrics," *Scientific Reports*, vol. 13, no. 1, pp. 1–13, 2023.

[26] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 10347–10357, 2021.

[27] S. Shen, G. Wu, and H.-I. Suk, "An Overview of Computer-Aided Medical Image Classification," *Informatics in Medicine Unlocked*, vol. 6, pp. 1–6, 2017.

[28] W. H. G. Chu, S. A. Hughes, D. D. Wang, J. E. Shelhamer, E. J. Oermann, S. Z. Seneviratne, and E. K. Oermann, "Visual Transformers and Convolutional Neural Networks for Disease Classification on Radiographs: A Comparison of Performance, Sample Efficiency, and Hidden Stratification," *Medical Image Analysis*, vol. 77, p. 102372, 2022.

[29] G. Gheflati and A. Shafiee, "Vision Transformer for Classification of Breast Ultrasound Images," *arXiv preprint arXiv:2202.09716*, 2022.

[30] Wikipedia, "ImageNet," *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/wiki/ImageNet, accessed Jun 26, 2025.

[31] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

[32] 'Vision-Transformers," GitHub repository, 2025, https://github.com/Tsomaros/Vision-Transformers.

[33] D. Hendrycks, K. Zhao, S. Basart, et al., "Natural adversarial examples," *CVPR*, 2021.

[34] D. Shen, G. Wu, and H. Suk, "An overview of computer-aided medical image classification," in *Neurocomputing*, vol. 409, pp. 394–422, 2020.

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2097–2106, 2017.

[36] M. Mancini, S. V. Vantaggiato, M. Di Benedetto, D. R. M. Rukundo, G. Castellano, and S. Tangaro, "Visual transformers and convolutional neural networks for disease classification on radiographs: A comparison of performance, sample efficiency, and hidden stratification," *Medical Image Analysis*, vol. 87, p. 102801, 2023.

[37] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Deep learning for chest radiograph diagnosis: A review," *Medical Image Analysis*, vol. 64, p. 101794, 2020.

[38] A. Steiner, A. Kolesnikov, R. Zhai, M. Wightman, B. Uszkoreit, L. Beyer, and N. Houlsby, "How to train your ViT: Data, augmentation, and regularization in vision transformers," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://arxiv.org/abs/2106.10270

[39] Y. Zhao, L. Wang, Y. Cui, Q. Luo, X. Yu, and L. Wang, "MediAug: Exploring visual augmentation in medical imaging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28956–28966, 2023.

[40] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Lecture Notes in Computer Science, vol. 9905, pp. 21–37, Cham, 2016. Springer.

[42] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988, 2017.