


Review

Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review

José Maurício , Inês Domingues  and Jorge Bernardino * 

Polytechnic of Coimbra, Coimbra Institute of Engineering (ISEC), Rua Pedro Nunes, 3030-199 Coimbra, Portugal; a2018056151@isec.pt (J.M.); ines.domingues@isec.pt (I.D.)

* Correspondence: jorge@isec.pt

Abstract: Transformers are models that implement a mechanism of self-attention, individually weighting the importance of each part of the input data. Their use in image classification tasks is still somewhat limited since researchers have so far chosen Convolutional Neural Networks for image classification and transformers were more targeted to Natural Language Processing (NLP) tasks. Therefore, this paper presents a literature review that shows the differences between Vision Transformers (ViT) and Convolutional Neural Networks. The state of the art that used the two architectures for image classification was reviewed and an attempt was made to understand what factors may influence the performance of the two deep learning architectures based on the datasets used, image size, number of target classes (for the classification problems), hardware, and evaluated architectures and top results. The objective of this work is to identify which of the architectures is the best for image classification and under what conditions. This paper also describes the importance of the Multi-Head Attention mechanism for improving the performance of ViT in image classification.

Keywords: transformers; Vision Transformers (ViT); convolutional neural networks; multi-head attention; image classification



Citation: Maurício, J.; Domingues, I.; Bernardino, J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci.* **2023**, *13*, 5521. <https://doi.org/10.3390/app13095521>

Academic Editor: Yu-Dong Zhang

Received: 20 March 2023

Revised: 19 April 2023

Accepted: 26 April 2023

Published: 28 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Nowadays, transformers have become the preferred models for performing Natural Language Processing (NLP) tasks. They offer scalability and computational efficiency, allowing models to be trained with more than a hundred billion parameters without saturating model performance. Inspired by the success of the transformers applied to NLP and assuming that the self-attention mechanism could also be beneficial for image classification tasks, it was proposed to use the same architecture, with few modifications, to perform image classification [1]. The author's proposal was an architecture, called Vision Transformers (ViT), which consists of breaking the image into 2D patches and providing this linear sequence of patches as input to the model. Figure 1 presents the architecture proposed by the authors.

In contrast to this deep learning architecture, there is another very popular tool for processing large volumes of data called Convolutional Neural Networks (CNN). The CNN is an architecture that consists of multiple layers and has demonstrated good performance in various computer vision tasks such as object detection or image segmentation, as well as NLP problems [2]. The typical CNN architecture starts with convolutional layers that pass through the kernels or filters, from left to right of the image, extracting computationally interpretable features. The first layer extracts low-level features (e.g., colours, gradient orientation, edges, etc.), and subsequent layers extract high-level features. Next, the pooling layers reduce the information extracted by the convolutional layers, preserving the most important features. Finally, the fully-connected layers are fed with the flattened output of the convolutional and pooling layers and perform the classification. Its architecture is shown in Figure 2.

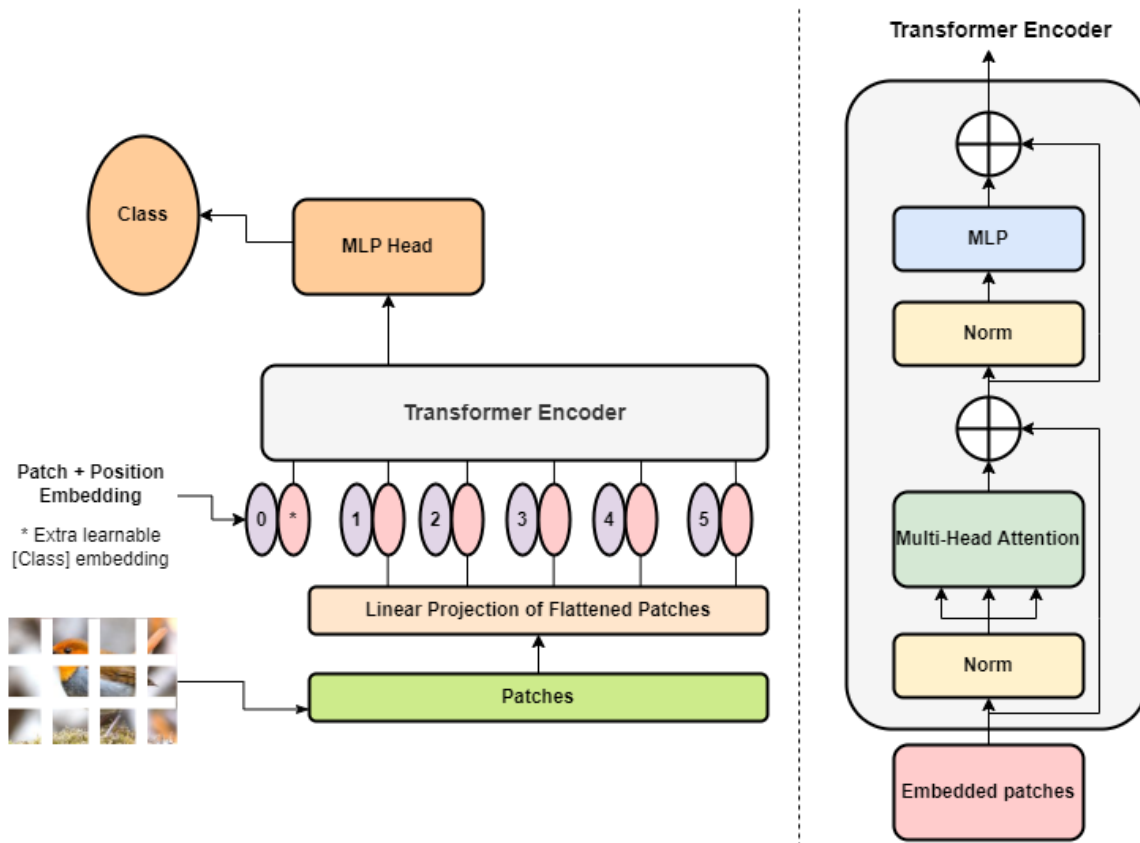


Figure 1. Example of an architecture of the ViT, based on [1].

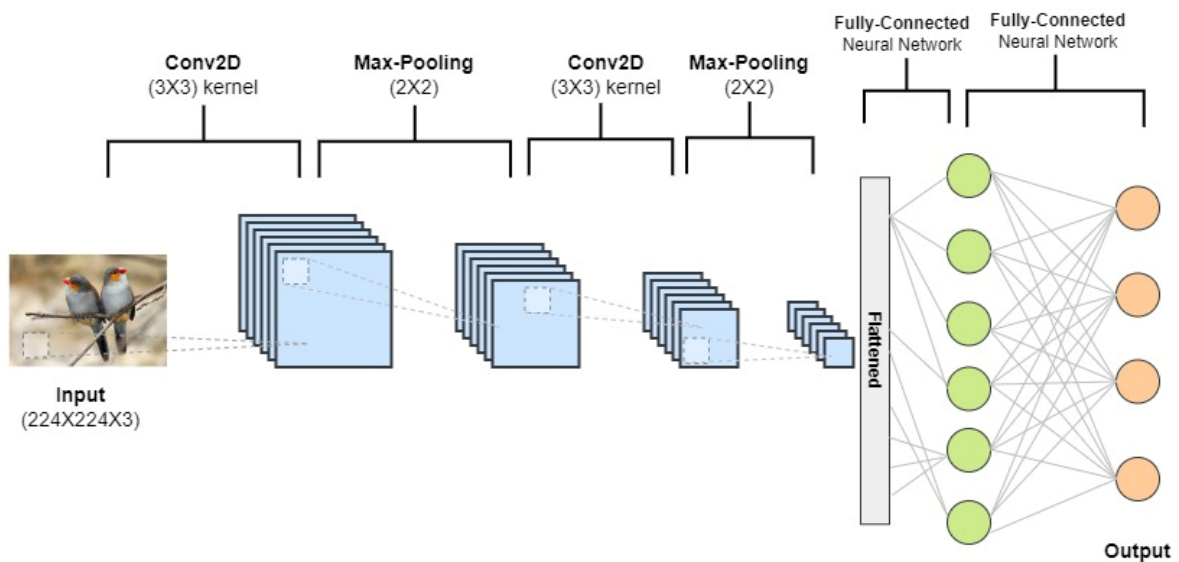


Figure 2. Example of an architecture of a CNN, based on [2].

With the increasing interest in Vision Transformers as a novel architecture for image recognition tasks, and the established success of CNNs in image classification, this work aims to review the state of the art in comparing Vision Transformers (ViT) and Convolutional Neural Networks (CNN) for image classification. Transformers offer advantages such as the ability to model long-range dependencies, adapt to different input sizes, and the potential for parallel processing, making them suitable for image tasks. However, Vision Transformers also face challenges such as computational complexity, model size, scalability

to large datasets, interpretability, robustness to adversarial attacks, and generalization performance. These points highlight the importance of comparing ViTs with older and established CNN models.

The overall goal of this work is to understand what conditions have the most influence on the performance of the two Deep Learning architectures, and what characteristics differ between the two architectures, that allow them to perform differently for the same objective. Some of the aspects that will be compared include datasets considerations, robustness, performance, evaluation, interpretability, and architecture. Specifically, we aim to answer the following research questions:

RQ1—Can the ViT architecture have a better performance than the CNN architecture, regardless of the characteristics of the dataset?

RQ2—What influences CNNs that do not to perform as well as ViTs?

RQ3—How does the Multi-Head Attention mechanism, which is a key component of ViTs, influence the performance of these models in image classification?

In order to address these research questions, a literature review was conducted by searching various databases such as Google Scholar, Scopus, Web of Science, ACM Digital Library, and Science Direct using specific search terms. This paper presents the results of this review and analyses the methodologies and findings from the selected papers.

The rest of this paper is structured as follows. Section 2 describes the research methodology and search results. Section 3 presents the knowledge, methodology, and results found in the selected documents. Section 4 provides a brief overview of the reviewed papers and attempts to answer the three research questions. Section 5 discusses threats to the validity of the research. Section 6 overviews the strengths and weaknesses of each architecture and suggests future research directions, and Section 7 presents the main conclusions of this work.

2. Research Methodology

The purpose of a literature review is to evaluate, analyse and summarize the existing literature on a specific research topic, in order to facilitate the emergence of theoretical frameworks [3]. In this literature review, the aim is to synthesize the knowledge base, critically evaluate the methods used and analyze the results obtained in order to identify the shortcomings and improve the two aforementioned deep learning architectures for image classification. The methodology for conducting this literature review is based on the guidelines presented in [3,4].

2.1. Data Sources

ACM Digital Library, Google Scholar, Science Direct, Scopus, and Web of Science, were chosen as the data sources to extract the primary studies. The number of results found after searching papers in each of the data sources is shown in Table 1.

Table 1. Data sources and the number of obtained results.

Data Source	Number of Results	Number of Selected Papers
ACM Digital Library	19,159	1
Google Scholar	10,700	10
Science Direct	1437	3
Scopus	55	2
Web of Science	90	1

2.2. Search String

The research questions developed for this paper served as the basis for the search strings utilized in each of the data sources. Table 2 provides a list of the search strings used in each electronic database.

Table 2. Data sources and used search string.

Data Source	Search String
ACM Digital Library	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Google Scholar	((ViT) AND (CNN) AND (Images Classification) OR (Comparing) OR (Vision Transformers) OR (convolutional neural networks) OR (differences))
Science Direct	((Vision Transformers) AND (convolutional neural networks) AND (images classification) AND (comparing))
Scopus	((ViT) AND (CNN) AND (comparing))
Web of Science	((ViT) AND (CNN) AND (comparing))

2.3. Inclusion Criteria

The inclusion criteria set to select the papers were that the studies were recent, had been written in English, and were published between January 2021 and December 2022. This choice of publication dates is based on the fact that ViTs were not proposed until the end of 2020 [1]. In addition, the studies had to demonstrate a comparison between CNNs and ViTs for image classification and could use any pre-trained model of the two architectures. Studies that presented a proposal for a hybrid architecture, where they combined the two architectures into one, were also considered. The dataset used during the studies did not have to be a specific one, but it had to be a dataset of images that allowed classification using both deep learning architectures.

2.4. Exclusion Criteria

Studies that oriented their research on using only one of the two deep learning architectures (i.e., Vision Transforms, or Convolutional Neural Networks) were excluded. Additionally, papers that were discovered to be redundant when searches were conducted throughout the chosen databases were eliminated. It was also defined that one of the exclusion criteria will be that the papers would have more than seven citations.

In summary, with the application of these criteria, 10,690 papers were excluded from the Google Scholar database, 89 papers from Web of Science, 53 papers from Scopus, 19,158 papers from ACM Digital Library, and 1434 papers from Science Direct.

2.5. Results

After applying the inclusion and exclusion criteria to the papers obtained in each of the electronic databases, seventeen (17) papers were selected for the literature review. Table 3 lists all the papers selected for this work, the year of publication and the type of publication.

Table 3. List of selected studies.

Ref.	Title	Year	Type
[5]	Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs	2021	Conference
[6]	Are Transformers More Robust Than CNNs?	2021	Conference
[7]	Detecting Pneumonia using Vision Transformer and comparing with other techniques	2021	Conference
[8]	Do Vision Transformers See Like Convolutional Neural Networks?	2021	Conference
[9]	Vision Transformer for Classification of Breast Ultrasound Images	2021	Conference
[10]	ConvNets vs. Transformers: Whose Visual Representations are More Transferable?	2021	Conference
[11]	A vision transformer for emphysema classification using CT images	2021	Journal
[12]	Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems	2022	Conference
[13]	Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification	2022	Conference
[14]	Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography	2022	Conference
[15]	Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection	2022	Conference
[16]	Traffic Sign Recognition with Vision Transformers	2022	Conference
[17]	An improved transformer network for skin cancer classification	2022	Journal
[18]	CNN and transformer framework for insect pest classification	2022	Journal
[19]	Single-layer Vision Transformers for more accurate early exits with less overhead	2022	Journal
[20]	Vision transformer-based autonomous crack detection on asphalt and concrete surfaces	2022	Journal
[21]	Vision Transformers for Weeds and Crops Classification of High-Resolution UAV Images	2022	Journal

Figure 3 shows the distribution of the selected papers by year of publication.

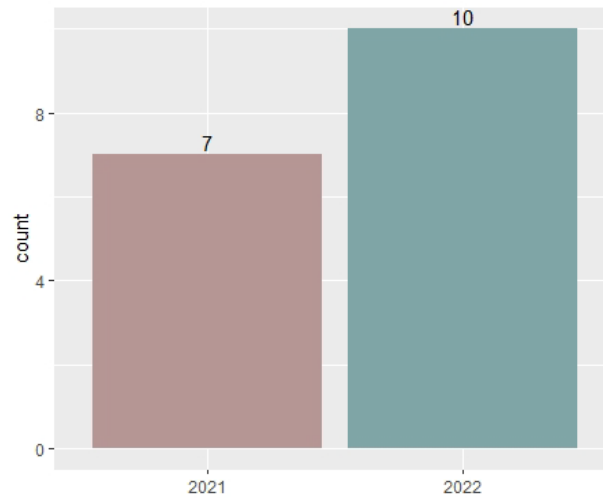


Figure 3. Distribution of the selected studies by years.

Figure 4 shows the distribution of the selected studies by application area. In the figure, most of the papers are generic in their application area. In these papers without a specific application area, the authors try to better understand the characteristics of the two architectures. For example, between CNNs and ViTs, the authors have tried to understand which of the architectures is more transferable. If architectures based on transformers are more robust than CNNs. And if the ViT will be able to see the same information as CNN with a different architecture. Within the health domain, some studies have been developed in different sub-areas, such as breast cancer, to show that ViT can be better than CNNs. The figure also shows that some work has been done, albeit to a lesser extent, in other application areas. Agriculture stands out with two papers comparing ViTs with CNNs.

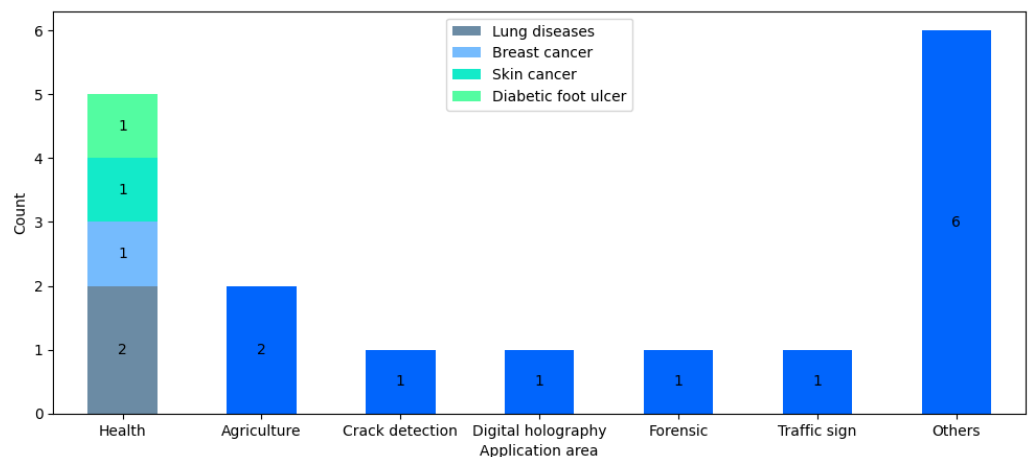


Figure 4. Distribution of the selected studies by application area.

3. Findings

An overview of the studies selected through the research methodology is shown in Table 4. This information summarizes the authors' approach, the findings, and other architectures that were used to build a comparative study. Therefore, to address the research questions, this section will offer an overview of the data found in the collected papers.

In the study developed in [12], the authors aimed to compare the two architectures (i.e., ViT and CNN), as well as the creation of a hybrid model that corresponded to the combination of the two. The experiment was conducted using the ImageNet dataset and

perturbations were applied to the dataset images. It was concluded that ViT can perform better and be more resilient on images with natural or adverse disturbances than CNN. It was also found in this work that the combination of the two architectures results in a 10% improvement in accuracy (Acc).

The work done in [14] aimed to compare Vision Transformers (ViT) with Convolutional Neural Networks (CNN) for digital holography, where the goal was to reconstruct amplitude and phase by extracting the distance of the object from the hologram. In this work, DenseNet201, DenseNet169, EfficientNetB4, EfficientNetB7, ViT-B/16, ViT-B32 and ViT-L/16 architectures were compared with a total of 3400 images. They were divided into four datasets, original images with or without filters, and negative images with or without filters. The authors concluded that ViT despite having an accuracy like CNN, was more robust because, due to the self-attention mechanism, it can learn the entire hologram rather than a specific area.

The authors in [7] studied the performance of ViT in comparison with other architectures to detect pneumonia, through chest X-ray images. Therefore, a ViT model, a CNN network developed by the authors and the VGG-16 network were used for the study which focussed on a dataset with 5856 images. After the experiments performed, the authors concluded that ViT was better than CNN with 96.45% accuracy, 86.38% validation accuracy, 10.8% loss and 18.25% validation loss. In this work, it was highlighted that ViT has a self-attention mechanism that allows splitting the image into small patches that are trainable, and each part of the image can be given an importance. However, the attention mechanism as opposed to the convolutional layers makes ViT's performance saturate fast when the goal is scalability.

In the study [21] the goal was to compare ViT with state-of-art CNN networks to classify UAV images to monitor crops and weeds. The authors compared the influence of the size of the training dataset on the performance of the architectures and found that ViT performed better with fewer images than CNN networks in terms of F1-Score. They concluded that ViT-B/16 was the best model to do crop and weed monitoring. In comparison with CNN networks, ViT could better learn the patterns of images in small datasets due to the self-attention mechanism.

In the scope of lung diseases, the authors in [11] investigated the performance of ViT models to automatically classify emphysema subtypes through Computed Tomography (CT) images in comparison with CNN networks. In this study, they performed a comparative study between the two architectures using a dataset collected by the authors (3192 patches) and a public dataset of 168 patches taken from 115 HRCT slides. In addition to this, they also verified the importance of pre-trained models. They concluded that ViT failed to generalize when trained with fewer images, because when comparing the pre-training accuracy with 91.27% on the training and 70.59% on the test.

In the work in [9], a comparison between state-of-the-art CNNs and ViT models for Breast ultrasound image classification was developed. The study was performed with two different datasets: the first containing 780 images and the second containing 163 images. The following architectures were selected for the study: ResNet50, VGG-16, Inception, NASNET, ViT-S/32, ViT-B/32, ViT-Ti/16, R + ViT-Ti/16 and R26 + ViT-S/16. ViT models were found to perform better than CNN networks for image classification. The authors also highlighted that ViT models could perform better when they were trained with a small dataset, because via the attention mechanism, it was possible to collect more information from different patches, instead of collecting information from the image.

Benz et al., in [5], compared ViT models, with the MLP-Mixer architecture and with CNNs. The goal was to evaluate which architecture was more robust in image classification. The study consisted of generating perturbations and adverse examples in the images and understanding which of the architectures was most robust. However, this study did not aim to analyse the causes. Therefore, the authors concluded that ViT were more robust than CNNs to adversarial attacks and from a features perspective CNN networks were more sensitive to high-frequency features. It was also described that the shift-variance property

of convolutional layers may be at the origin of the lack of robustness of the network in the classification of images that have been transformed.

The authors in [15] performed an analysis between ViT and CNN models aimed at detecting deepfake images. The experiment consisted in using the ForgeryNet dataset with 2.9 million images and 220 thousand video clips, together with three different image manipulation techniques, where they tried to train the models with real and manipulated images. By training the ViT-B model and the EfficientNetV2 network the authors demonstrated that the CNN network could generalize better and obtain higher training accuracy. However, ViT could have better generalization, reducing the bias in the identification of anomalies introduced by one or more different techniques to introduce anomalies.

Chao Xin et al. [17] aimed to compare their ViT model with CNN networks and with another ViT model to perform image classification to detect skin cancer. The experiment conducted by the authors used a public HAM10000 dataset with dermatoscopic skin cancer images and a clinical dataset collected through dermoscopy. In this study, a multi-scale image and the overlapping sliding window were used to serialize the images. They also used contrastive learning to improve the similarity of different labels and minimize the similarity in the same label. Thus, the ViT model developed was better for skin cancer classification using these techniques. However, the authors also demonstrated the effectiveness of balancing the dataset on the model performance, but they did not present the F1-Score values before the dataset is balanced to verify the improvement.

The authors in [19] aimed to study if ViT models could be an alternative to CNNs in time-critical applications. That is, for edge computing instances and IoT networks, applications using deep learning models consume multiple computational resources. The experiment used pre-trained networks such as ResNet152, DenseNet201, InceptionV3, and SL-ViT with three different datasets in the scope of images, audio, and video. They concluded that the ViT model introduced less overhead and performed better than the architectures used. It was also shown that increasing the kernel size of convolutional layers and using dilated convolutional caused a reduction in the accuracy of a CNN network.

In a study carried out in [20], the authors tried to find in ViTs an alternative solution to CNN networks for asphalt and concrete crack detection. The authors concluded that ViTs, due to the self-attention mechanism, had better performance in crack detection images with intense noise. CNN networks in the same images suffered from a high number of false negative rates, as well as the presence of biases in image classification.

Haolan Wang in [16] aimed to analyse eight different Vision Transformers and compare them with the performance of a pre-trained CNN network and without the pre-trained parameters to perform traffic signal recognition in autonomous driving systems. In this study, three different datasets with images of real-world traffic signals were used. This allowed the authors to conclude that the pre-trained DenseNet161 network had a higher accuracy than the ViT models to do traffic sign recognition. However, it was found in this work that ViT models performed better than the DenseNet161 network without the pre-trained parameters. From this work, it was also possible to conclude that the ViT models with a total number of parameters equal to or greater than the CNN networks, used during the experiment, had a shorter training time.

The work done in [13] compared CNN networks with Vision Transformers models for the classification of Diabetic Foot Ulcer images. For the study, the authors decided to use the following architectures: Big Image Transfer (BiT), EfficientNet, ViT-base and Data-efficient Image Transformers (DeiT) upon a dataset composed of 15,683 images. A further aim of this study to compare the performance of deep learning models using Stochastic Gradient Descent (SGD) [22] with Sharpness-Aware Optimization (SAM) [23,24]. These two tools are optimizers that seek to minimize the value of the loss function, improving the generalization ability of the model. However, SAM minimizes the value of the loss function and the sharpness loss, looking for parameters in the neighbourhood with a low loss. Therefore, this work concluded that the SAM optimizer originated an improvement in the values of F1-Score, AUC, Recall and Precision in all the architectures used. However, the authors did

not present the training and test values that allow for evaluating the improvement in the generalization of the models. Therefore, the BiT-ResNetX50 model with the SAM optimizer obtained the best performance for the classification of Diabetic Foot Ulcer images with F1-Score = 57.71%, AUC = 87.68%, Recall = 61.88%, and Precision = 57.74%.

The authors in [18] performed a comparative study between ViT models and CNN networks used in the state of the art with a model developed by them, where they combined CNN and transformers to perform insect pest recognition to protect agriculture worldwide. This study involved three public datasets: the IP102 dataset, the D0 dataset and Li's dataset. The algorithm created by the authors consisted of using the sequence of inputs formed by the CNN feature maps to make the model more efficient, and a flexible attention-based classification head was implemented to use the spatial information. Comparing the results obtained, the proposed model obtained a better performance in insect pest recognition with an accuracy of 74.897%. This work demonstrated that fine-tuning worked better on Vision Transformers than CNN, but on the other hand, this caused the number of parameters, the size, and the inference time of the model to increase significantly with respect to CNN networks. Through their experiments, the authors also demonstrated the advantage of using decoder layers in the proposed model to perform image classification. The greater the number of decoder layers, the greater the accuracy value of the model. However, this increase in the number of decoder layers increased the number of parameters, the size, and the inference time of the model. In other words, the architecture to process the images consumes far greater computational resources, which may not compensate for the increase in accuracy value with few layers. In the case of this study, the increase from one layer to three decoder layers represented only an increase of 0.478% in the accuracy value.

Several authors in [6,8,10] went deeper into the investigation and aimed to understand how the learning process of Vision Transformers works if ViT could be more transferable and better understand if the transform-based architecture were more robust than CNNs. In this sense, the authors in [8] intended to analyse the internal representations of ViT and CNN structures in image classification benchmarks and found differences between them. One of the differences was that ViT has greater similarity between high and low layers, while CNN architecture needs more low layers to compute similar representations in smaller datasets. This is due to the self-attention layers implemented in ViT, which allows it to aggregate information from other spatial locations, vastly different from the fixed field sizes in CNNs. They also observed that ViTs in the lower, self-attention layers can access information from local heads (small distances) and global heads (large distances). Whereas CNNs have access to information locally in the lower layers. On the other hand, the authors in [10] systematically analysed the transfer learning capacity in the two architectures. The study was conducted by comparing the performance of the two architectures on single-task and multi-task learning problems, using the ImageNet dataset. Through this study, the authors concluded that the transform-based architecture contained more transferable representations compared to convolutional networks for fine-tuning, presenting better performance and robustness in multi-task learning problems.

In another study carried out in [6], the goal was to prove if ViT were more robust than CNN as the most recent studies have shown. The authors developed their work comparing the robustness of the two architectures using two different types of perturbations: adversarial samples, which consists in evaluating the robustness of deep learning architectures in images with human-caused perturbations (i.e., data augmentation) and out-of-distribution samples, which consists in evaluating the robustness of the architectures in benchmarks of classification images. Through this experiment, it was demonstrated that by replacing the activation function ReLU by the activation function of transformer-based architecture (i.e., GELU) the CNN network was more robust than ViT in adversarial samples. In this study, it was also demonstrated that CNN networks were more robust than ViT in patch-based attacks. However, the authors concluded that the self-attention mechanism was the key to the robustness of the transformer-based architecture in most of the experiments performed.

Table 4. Overview of selected studies.

Ref.	Datasets	Images Size	Number of Classes	Hardware	Evaluated Architectures	Best Architecture	Best Results
[5]	ImageNet-1K (more than 1.431 M images) for training and ImageNet-C for validation	224 × 224	2	N/A	ViT-B/16, ViT-L/16, Mixer-B/16, Mixer-B/16, RN18 (SWSL), RN50 (SWSL), RN18 and RN50	ViT-L/16	82.89% of Acc
[6]	ImageNet-A; ImageNet-C and Stylized ImageNet	224 × 224	N/A	N/A	ResNet50 and DeiT-5	N/A	N/A
[7]	5856 images collected by X-ray	250 × 250 to ViT 224 × 224 to CNN	2	Intel Core i5-8300H 2.30 GHz	ViT, CNN and VGG16	ViT	96.45% Acc, 86.38% val. Acc, 10.92% loss and 18.25% val. Loss
[8]	ImageNetILSVRC 2012 (1.78 M images)	224 × 224	1000	N/A	ViT-B/32, ViT-L/16, ViT-H/14, ResNet50 and ResNet152	N/A	N/A
[9]	Public dataset 1 [25] with 780 images; Public dataset 2 [26] with 163 images	224 × 224	3	N/A	ViT-S/32, ViT-B/32, ViT-Ti/16, R26 + S/16, R + Ti/16, VGG, Inception and NASNET	ViT-B/32	86.7% Acc and 95% AUC
[10]	Flower 102 (4080 to 11,016 images); CUB 200 (11,788 images); Indoor 67 (15,620 images); NY Depth V2 (1449 images); WikiArt; COVID-19 Image Data Collection (700 images); Caltech101 (9146 images); FG-NET (1002 images)	384 × 384; 224 × 224; 300 × 300	40 to 102	N/A	R-101 × 3, R-152 × 4, ViT-B/16, ViT-L/16, and Swim-B	N/A	N/A
[11]	3192 images collected by CT and 160 images of a public dataset with CT biomarkers	61 × 61	4	Intel Core i7-9700 3.0 GHz, 326 GB RAM; NVIDIA GeForce RTX 2080 Ti (116 GB DDR6)	AlexNet, VGG-16, InceptionV3, MobileNetV2, ResNet34, ResNet50 and ViT	ViT	95.95% Acc
[12]	ImageNet-C benchmark	224 × 224	2	NVIDIA Quadro A6000	ViT-L/16, CNN, hybrid model (BiT-M + ResNet152 × 4)	Hybrid model	99.20% Acc
[13]	Dataset provided in DFUC 2021 challenge (15,683 images)	224 × 224	4	NVIDIA GeForce RTX 3080, 10 GB memory	EfficientNetB3, BiT-ResNeXt50, ViT-B/16 and DeiT-S/16	BiT-ResNeXt50	88.49% AUC, 61.53% F1-Score, 65.59% recall and 60.53% precision
[14]	3400 images collected by holographic camera	512 × 512	10	NVIDIA V100	EfficientnetB7, Densenet169 and ViT-B/16	ViT-B/16	99% Acc
[15]	ForgeryNet with 2.9 M images	N/A	2	N/A	EfficientNetV2 and ViT-B	N/A	N/A
[16]	German dataset (51,830 images); Indian dataset (1976 images); Chinese dataset (18,168 images)	128 × 128	15, 43 and 103	AMD Ryzen 7 5800H; NVIDIA GeForce RTX 3070	DenseNet161, ViT, DeepViT, MLP-Mixer, CvT, PiT, CaiT, CCT, CrossViT and Twins-SVT	CCT	99.04% Acc

Table 4. Cont.

Ref.	Datasets	Images Size	Number of Classes	Hardware	Evaluated Architectures	Best Architecture	Best Results
[17]	HAM10000 dataset (10,015 images); 1016 images collected by dermoscopy	224 × 224	3	Intel i7; 2x NVIDIA RTX 3060, 12 GB	MobileNetV2, ResNet50, InceptionV2, ViT and Proposed ViT model	Proposed ViT model	94.10% Acc, 94.10% precision and 94.10% F1-Score
[18]	IP102 dataset (75,222 images); D0 dataset (4508 images; Li's dataset (5629 images)	224 × 224; 480 × 480	10 to 102	Train: Intel Xeon; 8x NVIDIA Tesla V100, 256 GB. Test: Intel Core; NVIDIA GTX 1060 Ti, 16 GB	ResNet, EfficientNetB0, EfficientB1, RepVGG, VGG-16, ViT-L/16 and hybrid model proposed	Proposed hybrid model	99.47% Acc on the D0 dataset and 97.94% Acc on the Li's dataset
[19]	CIFAR-10 and CIFAR-100 (6000 images); Speech commands (100,503 1-second audio clips); GTZAM (1,00,030-second audio clips); DISCO (1935 images)	224 × 224 px; 1024 × 576 px. Spectrograms: 229 × 229 samples; 512 × 256 samples	10 to 100	4x NVIDIA 2080 Ti	SL-ViT, ResNet152, DenseNet201 and InceptionV3	SL-ViT	71.89% Acc
[20]	CrackTree260 (260 images); Ozegenel (458 images); Lab's on dataset (80,000 images)	256 × 256; 448 × 448	2	N/A	TransUNet, U-Net, DeepLabv3+ and CNN + ViT	CNN + ViT	99.55% Acc and 99.57% precision
[21]	10,265 images collected by Pilgrim technologies UAV with Sony ILCE-7R-36 mega pixels	64 × 64	5	Intel Xeon E5-1620 V4 3.50 GHz with 8 processor, 16 GB RAM; NVIDIA Quadro M2000	ViT-B/16, ViT-B/32, EfficientNetB, EfficientNetB1 and ResNet50	ViT-B/16	99.8% Acc

4. Discussion

The results can be summarized as follows. In [12], ViTs were found to perform better and be more resilient to images with natural or adverse disturbances compared to CNNs. Another study [14] concluded that ViTs are more robust in digital holography because they can access the entire hologram rather than just a specific area, giving them an advantage. ViTs have also been found to outperform CNNs in detecting pneumonia in chest X-ray images [7] and in classifying UAV images for crop and weed monitoring with small datasets [21]. However, it has been noted that ViT performance may saturate if scalability is the goal [7]. In a study on classifying emphysema subtypes in CT images [11], ViTs were found to struggle with generalization when trained on fewer images. Nevertheless, ViTs were found to outperform CNNs in breast ultrasound image classification, especially with small datasets [9]. Another study [5] found that ViTs are more robust to adversarial attacks and that CNNs are more sensitive to high-frequency features. The authors in [15] found that CNNs had higher training accuracy and better generalization, but ViTs showed potential to reduce bias in anomaly detection. In [17], the authors claimed that the ViT model showed better performance for skin cancer classification. ViTs have also been shown to introduce less overhead and perform better for time-critical applications in edge computing and IoT networks [19]. In [20], the authors investigated the use of ViTs for asphalt and concrete crack detection and found that ViTs performed better due to the self-attention mechanism, especially in images with intense noise and biases. Wang [16] found that a pre-trained CNN network had higher accuracy, but the ViT models performed better than the non-pre-trained CNN network and had a shorter training time. The authors in [13] used several models for diabetic foot ulcer image classification and compared SGD and SAM optimizers, concluding that the SAM optimizer improved several evaluation metrics. In [18], the authors showed that fine-tuning performed better on ViT models than CNNs for insect pest recognition.

Therefore, based on the information gathered from the selected papers, we attempt to answer the research questions posed in Section 1:

RQ1—Can the ViT architecture have a better performance than the CNN architecture, regardless of the characteristics of the dataset?

The literature review shows that ViT in image processing can be more efficient in smaller datasets due to the increase of relations created between images through the self-attention mechanism. However, it is also shown that if ViT trained with little data will have less generalization ability and worse performance compared to CNN's.

RQ2—What influences the CNNs that do not allow them to perform as well as the ViTs?

Shift-invariance is a limitation of CNN that makes the same architecture not have a satisfactory performance because the introduction of noise in the input images makes the same architecture unable to get the maximum information from the central pixels. However, the authors in [27] propose the addition of an anti-aliasing filter which combines blurring with subsampling in the Convolutional, MaxPooling and AveragePooling layers. Demonstrating through the experiment carried out that the application of this filter originates a greater generalization capacity and an increase in the accuracy of CNN. Furthermore, increasing the kernel size in convolutional layers and using dilated convolution have been shown as limitations that deteriorate the performance of CNNs against ViTs.

RQ3—How does the Multi-Head Attention mechanism, which is a key component of ViTs, influence the performance of these models in image classification?

The Attention mechanism is described as the mapping of a query and a set of key-value pairs to an output, the output being the result of a weighted sum of the values, in which the weight given is calculated, through the query with the corresponding key by a compatibility function. Multi-head Attention mechanism instead of performing a single attention function will perform multiple projections of attention functions [28]. This mechanism improves the ViT architecture because it allows it to extract more information from each pixel of the images that have been placed inside the embedding. In addition, this

mechanism can have better performance if the images have more secondary elements that illustrate the central element. And since this mechanism performs several computations in parallel, it reduces the computational cost [29].

Overall, ViTs have shown promising performance compared to CNNs in various applications, but there are limitations and factors that can affect their performance, such as dataset size, scalability, and pre-training accuracy.

5. Threats to Validity

This section discusses internal and external validity threats. The validity of the entire process performed in this study is demonstrated and how the results of this study can be replicated in other future experiments.

In this literature review, different search strings were used in each of the selected data sources, resulting in different results from each source. This approach may introduce a bias into the validation of the study, as it makes it difficult to draw conclusions about the diversity of studies obtained by replicating the same search. In addition, the maturity of the work was identified as an internal threat to validity, as the ViT architecture is relatively new and only a limited number of research projects have been conducted using it. In order to draw more comprehensive conclusions about the robustness of ViT compared to CNN, it is imperative that this architecture is further disseminated and deployed, thereby making more research available for analysis.

In addition to these threats, this study did not use methods that would allow to quantitatively and qualitatively analyse the results obtained in the selected papers. This may bias the validity of this review in demonstrating which of the deep learning architectures is more efficient in image processing.

The findings obtained in this study could be replicated in other future research in image classification. However, the results obtained may not be the same as those described by the selected papers because it has been proven that for different problems and different methodologies used, the results are different. In addition, the authors do not describe in sufficient detail all the methodologies they used, nor the conditions under which the experiment was performed.

6. Strengths, Limitations, and Future Research Directions

The review made it possible to identify not only the strengths of each architecture (outlined in Section 6.1), but also their potential for improvement (described in Section 6.2). Future research directions were also derived from this and are presented in Section 6.3.

6.1. Strengths

Both CNNs and ViTs have their own advantages, and some common ones. This section will explore these in more detail, including considerations on the Datasets, Robustness, Performance optimization, Evaluation, Explainability and Interpretability, and Architectures.

6.1.1. Dataset Considerations

CNNs have been widely used and extensively studied for image-related tasks, resulting in a rich literature, established architectures, and pre-trained models, making them accessible and convenient for many datasets. On the other hand, ViTs can process patches in parallel, which can lead to efficient computation, especially for large-scale datasets, and allow faster training and inference. ViTs can also handle images of different sizes and aspect ratios without losing resolution, making them more scalable and adaptable to different datasets and applications.

6.1.2. Robustness

CNNs are inherently translation-invariant, making them robust to small changes in object position or orientation within an image. The main advantage of ViTs is their ability

to effectively capture global contextual information through the self-attention mechanism, enabling them to model long-range dependencies and contextual relationships, which can improve robustness in tasks that require understanding global context. ViTs can also adaptively adjust the receptive fields of the self-attention mechanism based on input data, allowing them to better capture both local and global features, making them more robust to changes in scale, rotation, or perspective of objects.

Both architectures can be trained using data augmentation techniques, such as random cropping, flipping, and rotation, which can help improve robustness to changes in input data and reduce overfitting. Another technique is known as adversarial training, where they are trained on adversarial examples: perturbed images designed to confuse the model, to improve its ability to handle input data with adversarial perturbations. Combining models, using ensemble methods, such as bagging or boosting, can also improve robustness by exploiting the diversity of multiple models, which can help mitigate the effects of individual model weaknesses.

6.1.3. Performance Optimization

CNNs can be effectively compressed using techniques such as pruning, quantization, and knowledge distillation, reducing model size and improving inference efficiency without significant loss of performance. They are also well-suited for hardware optimization, with specialized hardware accelerators (e.g., GPUs, TPUs) designed to perform convolutional operations efficiently, leading to optimized performance in terms of speed and energy consumption.

ViTs can efficiently scale to handle high-resolution images or large-scale datasets because they operate on the entire image at once and do not require processing of local receptive fields at multiple spatial scales, potentially resulting in improved performance in terms of scalability.

Transfer Learning for pre-train on large-scale datasets and fine-tune on smaller datasets can potentially lead to improved performance with limited available data and can be used with both architectures.

6.1.4. Evaluation

CNNs have been widely used in image classification tasks for many years, resulting in well-established benchmarks and evaluation metrics that allow meaningful comparison and evaluation of model performance. The standardized evaluation protocols, such as cross-validation or hold-out validation, which provide a consistent framework for evaluating and comparing model performance across different datasets and tasks, are applicable for both architectures.

6.1.5. Explainability and Interpretability

CNNs produce feature maps that can be visualized, making it possible to interpret the behaviour of the model by visualizing the learned features or activations in different layers. They capture local features in images, such as edges or textures, which can lead to interpretable features that are visually meaningful and can provide insight into how the model is processing the input images. ViTs, on the other hand, are designed to capture global contextual information, making them potentially more interpretable in tasks that require an understanding of long-range dependencies or global context. They have a hierarchical structure with self-attention heads that can be visualized and interpreted individually, providing insights into how different heads attend to different features or regions in the input images.

6.1.6. Architecture

CNNs have a wide range of established architecture variants, such as VGG, ResNet, and Inception, with proven effectiveness in various image classification tasks. These architectures are well-tested and widely used in the deep learning community. ViTs can be

easily modified to accommodate different input sizes, patch sizes, and depth, providing flexibility in architecture design and optimization.

6.2. Limitations

Despite their many advantages and the breakthroughs made over the years. There are still some drawbacks to the architectures studied. This section focuses on these.

6.2.1. Dataset Considerations

CNNs can be susceptible to biases present in training datasets, such as biased sampling or label noise, which can affect the validity of training results. They typically operate on fixed input spatial resolutions, which may not be optimal for images of varying size or aspect ratio, resulting in information loss or distortion. While pre-trained models for CNNs are well-established, pre-trained models for ViT are (still) less common for some datasets, which may affect the ease of use in some situations.

6.2.2. Robustness

CNNs may struggle to capture long-range contextual information, as they focus primarily on local feature extraction, which may limit the ability to understand global context, leading to reduced robustness in tasks that require global context, such as scene understanding, image captioning or fine-grained recognition.

Both architectures can be prone to overfitting, especially when the training data is limited or noisy, which can lead to reduced robustness to input data outside the training distribution. Adversarial attacks can also pose a challenge to the robustness of both architectures. In particular, ViTs do not have an inherent spatial inductive bias like CNNs, which are specifically designed to exploit the spatial locality of images. This can make them more vulnerable to certain types of adversarial attacks that rely on spatial information, such as spatially transformed adversarial examples.

6.2.3. Performance Optimization

CNNs can suffer from reduced performance and increased memory consumption when applied to high-resolution images or large-scale datasets, as they require processing of local receptive fields at multiple spatial scales, leading to increased computational requirements. Compared to CNNs, ViTs are computationally expensive, especially as the image size increases or model depth increases, which may limit their use in certain resource-constrained environments. Reduced computational complexity can sometimes result in decreased robustness, as models may not have the ability to learn complex features that can help generalize well to adversarial examples.

6.2.4. Evaluation

As mentioned above, CNNs are primarily designed for local feature extraction and may struggle to capture long-range contextual dependencies, which can limit the evaluation performance in tasks that require understanding of global context or long-term dependencies. ViTs are relatively newer than CNNs, and as such, may lack well-established benchmarks or evaluation metrics for specific tasks or datasets, which can make performance evaluation difficult and less standardized.

6.2.5. Explainability and Interpretability

Despite well-established methods for model interpretation, CNNs still lack full interpretability because the complex interactions between layers and neurons can make it difficult to fully understand the model's decision-making process, particularly in deeper layers of the network.

While ViTs produce attention maps for interpretability, the complex interactions between self-attention heads can still present challenges in accurately interpreting the model's behaviour. ViTs can have multiple heads attending to different regions, which can make it

difficult to interpret the interactions between different attention heads and to understand the reasoning behind model predictions.

6.2.6. Architecture

CNNs typically have fixed, predefined model architectures, which may limit the flexibility to adapt to specific task requirements or incorporate domain-specific knowledge, potentially affecting performance optimization. For ViTs, the availability of established architecture variants is still limited, which may require more experimentation and exploration to find optimal architectures for specific tasks.

6.3. Future Research Directions

As future research, meta-analysis or systematic reviews should be conducted within the scope of this review to provide the scientific community with more detail on which of the architectures is more effective at image classification, in addition to specifying under what conditions a particular architecture stands out from the others. It is therefore necessary to facilitate the choice of the deep learning architecture to be used in future image classification problems. This section aims to provide guidelines for future research in this area.

6.3.1. Dataset Considerations

The datasets used in most studies may not be representative of real-world scenarios. Future research should consider using more diverse datasets that better reflect the complexity and variability of real-world images. As an example, it would be interesting to study the impact that image resolution might have on the performance of deep learning architectures. That is, it would be important to find out in which of the architectures (i.e., ViT, CNN, and MLP-Mixer) the resolution of the images will influence their performance, as well as what impact it will have on the processing time of the deep learning architectures.

6.3.2. Robustness

As documented in [5], deep learning models are typically vulnerable to adversarial attacks, where small perturbations to an input image can cause the model to misclassify it. Future research should focus on developing architectures that are more robust to adversarial attacks (for example by further augmenting the robustness of ViTs), as well as exploring ways to detect and defend against these attacks.

Beyond that, most studies (as the ones reviewed in this work) have focused on the performance of deep learning architectures on image classification tasks, but there are many other image processing tasks (such as object detection, segmentation, and captioning) that could benefit from the use of these architectures. Future research should further explore the effectiveness of these architectures on these tasks.

6.3.3. Performance Optimization

Deep learning architectures require substantial amounts of labelled data to achieve high performance. However, labelling data is time-consuming and expensive. Future research should explore ways to improve the efficiency of deep learning models, such as developing semi-supervised learning methods or transfer learning (following up on the finding in [10]) that can leverage pre-trained models.

In addition, the necessity of large amounts of labelled data requires significant computational resources, which limits the deployment on resource-constrained devices. Future research should focus on developing architectures that are optimized for deployment on these devices, as well as exploring ways to reduce the computational cost of existing architectures. It should explore the advantages of the implementation of the knowledge distillation of deep learning architectures to reduce computational resources.

6.3.4. Evaluation

The adequacy of the metrics to the task and problem at hand is also another suggested line of future research. Most studies have used standard performance metrics (such as accuracy and F1-score) to evaluate the performance of deep learning architectures. Future research should consider using more diverse metrics that better capture the strengths and weaknesses of different architectures.

6.3.5. Explainability and Interpretability

Deep learning models are often considered as black boxes because they do not provide insight into the decision-making process. This may prevent the usage of the models in certain areas such as justice and healthcare [30], among others. Future research should focus on making these models more interpretable and explainable. For example, by designing transformer architectures that provide visual explanations of their decisions or by developing methods for extracting features that are easily interpretable.

6.3.6. Architecture

In future investigations, it will be necessary to study the impact of the MLP-Mixer deep learning architecture in image processing, what are the characteristics that allow it to have a performance superior to CNNs, but inferior to the performance obtained by the ViT architecture [5]. Future research should also focus on developing novel architectures that can achieve high performance with fewer parameters or that are more efficient in terms of computation and memory usage.

7. Conclusions

This work has reviewed recent studies done in image processing to give more information about the performance of the two architectures and what distinguishes them. A common feature across all papers is that transformer-based architecture or the combination of ViTs with CNN allows for better accuracy compared to CNN networks. It has also been shown that this new architecture, even with hyperparameters fine-tuning, can be lighter than the CNN, consuming fewer computational resources and taking less training time as demonstrated in the works [16,19].

In summary, the ViT architecture is more robust than CNN networks for images that have noise or are augmented. It manages to perform better compared to CNN due to the self-attention mechanism because it makes the overall image information accessible from the highest to the lowest layers [12]. On the other hand, CNN's can generalize better with smaller datasets and get better accuracy than ViTs, but in contrast, ViTs have the advantage of learning information better with fewer images. This is because the images are divided into small patches, so there is a greater diversity of relationships between them.

Author Contributions: Conceptualization, J.B.; Methodology, J.M. and J.B.; Software, J.M.; Validation, J.M., I.D. and J.B.; Formal analysis, J.M., I.D. and J.B.; Investigation, J.M.; Resources, J.M.; Data curation, J.M.; Writing—original draft preparation, J.M.; Writing—review and editing, J.M., I.D. and J.B.; Supervision, J.B. and I.D.; Project administration, J.B. and I.D.; Funding acquisition, J.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929. [[CrossRef](#)]
2. Saha, S. A Comprehensive Guide to Convolutional Neural Networks—The ELI5 Way. Available online: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> (accessed on 8 January 2023).

3. Snyder, H. Literature Review as a Research Methodology: An Overview and Guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [[CrossRef](#)]
4. Matloob, F.; Ghazal, T.M.; Taleb, N.; Aftab, S.; Ahmad, M.; Khan, M.A.; Abbas, S.; Soomro, T.R. Software Defect Prediction Using Ensemble Learning: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 98754–98771. [[CrossRef](#)]
5. Benz, P.; Ham, S.; Zhang, C.; Karjauv, A.; Kweon, I.S. Adversarial Robustness Comparison of Vision Transformer and MLP-Mixer to CNNs. *arXiv* **2021**, arXiv:2110.02797. [[CrossRef](#)]
6. Bai, Y.; Mei, J.; Yuille, A.; Xie, C. Are Transformers More Robust Than CNNs? *arXiv* **2021**, arXiv:2111.05464. [[CrossRef](#)]
7. Tyagi, K.; Pathak, G.; Nijhawan, R.; Mittal, A. Detecting Pneumonia Using Vision Transformer and Comparing with Other Techniques. In Proceedings of the 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, Coimbatore, India, 2 December 2021; pp. 12–16.
8. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do Vision Transformers See Like Convolutional Neural Networks? *arXiv* **2021**, arXiv:2108.08810. [[CrossRef](#)]
9. Gheflati, B.; Rivaz, H. Vision Transformer for Classification of Breast Ultrasound Images. *arXiv* **2021**, arXiv:2110.14731. [[CrossRef](#)]
10. Zhou, H.-Y.; Lu, C.; Yang, S.; Yu, Y. ConvNets vs. Transformers: Whose Visual Representations Are More Transferable? In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), IEEE, Montreal, BC, Canada, 17 October 2021; pp. 2230–2238.
11. Wu, Y.; Qi, S.; Sun, Y.; Xia, S.; Yao, Y.; Qian, W. A Vision Transformer for Emphysema Classification Using CT Images. *Phys. Med. Biol.* **2021**, *66*, 245016. [[CrossRef](#)]
12. Filipiuk, M.; Singh, V. Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems. *AAAI Workshop Artif. Intell. Saf.* **2022**, *3087*, 1–5.
13. Galdran, A.; Carneiro, G.; Ballester, M.A.G. Convolutional Nets Versus Vision Transformers for Diabetic Foot Ulcer Classification. *arXiv* **2022**, arXiv:2111.06894. [[CrossRef](#)]
14. Cuenat, S.; Couturier, R. Convolutional Neural Network (CNN) vs Vision Transformer (ViT) for Digital Holography. In Proceedings of the 2022 2nd International Conference on Computer, Control and Robotics (ICCCR), IEEE, Shanghai, China, 18 March 2022; pp. 235–240.
15. Coccomini, D.A.; Caldelli, R.; Falchi, F.; Gennaro, C.; Amato, G. Cross-Forgery Analysis of Vision Transformers and CNNs for Deepfake Image Detection. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, 27–30 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 52–58.
16. Wang, H. Traffic Sign Recognition with Vision Transformers. In Proceedings of the 6th International Conference on Information System and Data Mining, Silicon Valley, CA, USA, 27–29 May 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 55–61.
17. Xin, C.; Liu, Z.; Zhao, K.; Miao, L.; Ma, Y.; Zhu, X.; Zhou, Q.; Wang, S.; Li, L.; Yang, F.; et al. An Improved Transformer Network for Skin Cancer Classification. *Comput. Biol. Med.* **2022**, *149*, 105939. [[CrossRef](#)]
18. Peng, Y.; Wang, Y. CNN and Transformer Framework for Insect Pest Classification. *Ecol. Inform.* **2022**, *72*, 101846. [[CrossRef](#)]
19. Bakhtiarnia, A.; Zhang, Q.; Iosifidis, A. Single-Layer Vision Transformers for More Accurate Early Exits with Less Overhead. *Neural Netw.* **2022**, *153*, 461–473. [[CrossRef](#)]
20. Asadi Shamsabadi, E.; Xu, C.; Rao, A.S.; Nguyen, T.; Ngo, T.; Dias-da-Costa, D. Vision Transformer-Based Autonomous Crack Detection on Asphalt and Concrete Surfaces. *Autom. Constr.* **2022**, *140*, 104316. [[CrossRef](#)]
21. Reedha, R.; Dericquebourg, E.; Canals, R.; Hafiane, A. Vision Transformers for Weeds and Crops Classification of High Resolution UAV Images. *Remote Sens.* **2022**, *14*, 592. [[CrossRef](#)]
22. Bottou, L.; Bousquet, O. The Tradeoffs of Large Scale Learning. In *Advances in Neural Information Processing Systems*; Platt, J., Koller, D., Singer, Y., Roweis, S., Eds.; Curran Associates, Inc.: Vancouver, BC, Canada, 2007; Volume 20.
23. Foret, P.; Kleiner, A.; Mobahi, H.; Neyshabur, B. Sharpness-Aware Minimization for Efficiently Improving Generalization. *arXiv* **2020**, arXiv:2010.01412. [[CrossRef](#)]
24. Korpelevich, G.M. The Extragradient Method for Finding Saddle Points and Other Problems. *Ekonom. Mat. Metod.* **1976**, *12*, 747–756.
25. Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; Fahmy, A. Dataset of Breast Ultrasound Images. *Data Brief* **2020**, *28*, 104863. [[CrossRef](#)]
26. Yap, M.H.; Pons, G.; Marti, J.; Ganau, S.; Sentis, M.; Zwiggelaar, R.; Davison, A.K.; Marti, R. Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J. Biomed. Health Inform.* **2018**, *22*, 1218–1226. [[CrossRef](#)]
27. Zhang, R. Making Convolutional Networks Shift-Invariant Again. *arXiv* **2019**, arXiv:1904.11486. [[CrossRef](#)]
28. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *Neural Inf. Process. Syst.* **2017**, *30*, 3762. [[CrossRef](#)]
29. Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; Feng, J. DeepViT: Towards Deeper Vision Transformer. *arXiv* **2021**, arXiv:2103.11886. [[CrossRef](#)]
30. Amorim, J.P.; Domingues, I.; Abreu, P.H.; Santos, J.A.M. Interpreting Deep Learning Models for Ordinal Problems. In Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 25–27 April 2018.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.