Improving Weed Detection and Classification using Vision Transformers

M Anagha Ramadas, Nikita Chanalya, Jyothisha J Nair, Sreelekshmi V *

Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India.

Abstract. Chemical pesticide spraying over a large area is not only a waste of herbicides and labour, but it also pollutes the environment and compromises food quality. As a result, properly identifying weeds and spraying them are critical tactics for increasing agricultural sustainability. In this work, the models were trained using a dataset containing 17,509 labelled images of weeds native to Australia. Despite the advances of Deep learning models, it still faces challenges like high computational power, overfitting and need of a balanced labelled dataset, therefore transformers used mainly in Natural language processing tasks can be considered as a possible replacement for CNNs. This paper aims to analyse and compare Vision Transformer with CNN models for achieving good results on weed detection and classification from crop images. As the final results of the experiment, the Vision transformer gave the highest accuracy of 96.41% followed by the ResNet-50 model which gave an accuracy of 95.70% compared to 95.04% of the Xception model. The Inception V3 model gave an accuracy of 94.7% and Inception-Resnet V2 gave an accuracy of 94.15% on the dataset images.

Keywords: weed detection, segmentation, vision transformer, deep learning, classification.

1. Introduction

Agriculture is at the focus of scientific progress and innovation in order to tackle important issues such as generating high crop yield while safeguarding plant development, productivity, and quality in order to fulfil market needs. However, one of the primary issues in contemporary agriculture is the overuse of pesticides to increase agricultural output and eliminate undesired plants like weeds from the field. Plants that are undesirable and grow inadvertently on agricultural soils are known as weeds. They fight for water, nutrients, and sunlight with crops, posing a serious threat to agricultural productivity and quality if not properly managed, raising production costs and lowering the economic value of farmed regions.They are venomous, generate thorns and abrasives, and contaminate agricultural harvesting, making crop management difficult. That is why farmers spend so much money on weed treatment, sometimes without proper professional assistance, followed by appalling weed management and lower crop yields. To address these challenges, extensive research is being carried out in order to achieve a reduction in chemical use and precise herbicide application based on weed coverage. [1] The majority of these systems have been built on Deep Convolutional Neural Networks in recent years, and they have shown excellent results in weed identification and classification. However, weeds and crops, on the other hand, are difficult to distinguish from each other, even when utilising spectral data, due to their great similarities.

The rapid advancement and widespread availability of image-capturing technologies have made image capture simple. Meanwhile, computer hardware has become significantly less expensive, and GPU computational power has increased dramatically. Deep learning-based strategies for identifying weeds and classification have yielded promising results. Traditional machine learning algorithms are simple to grasp, and many advancements have been achieved, however, the majority of them are tested on low-density images. The network feature structure of deep learning is unique, and the use of various deep learning methods to obtain features is more successful than manually extracting features. Local characteristics can be learned from the bottom and then synthesised from the top to produce higher-level features. Various properties at various levels might correspond to a variety of activities.

Deep learning (DL) is a subset of machine learning that allows a computer program to learn and comprehend a dataset in terms of a concept hierarchy. In the realm of machine learning, deep Learning has ushered in a revolution. Deep Neural Networks have developed as a powerful tool in recent years, breaking records in a variety of scientific domains such as computer vision, natural language processing, and speech recognition. The reduced requirement for feature engineering (FE) is one of the most significant advantages of employing deep learning in image processing. Previously, hand-engineered features were used to perform imagery classification tasks, and their performance had a significant influence on classification accuracy. FE is a time-consuming and difficult operation that must be completed correctly. As a result, FE is a costly endeavour that relies on the expertise of specialists and does not lend itself to generalisation. Deep learning, on the other hand, does not need the use of FE, instead finds the key characteristics on its own while training.

On the contrary, in comparison to prior generations of state-of-the-art models, the attention mechanism has undergone fast progress, notably in Natural Language Processing, and has exhibited astonishing performance advances.Transformer models' amazing achievements on natural language challenges have piqued the interest of computer vision researchers to investigate their applicability to computer vision difficulties. Self-attention, large-scale pre-training, and bidirectional en-coding are key principles underpinning the success of Transformers. [2] The recently suggested vision transformer (ViT) looks to be a significant step forward in the adoption of transformer attention models for computer vision applications. ViT was found to outperform CNNs in imagery classification accuracy when given vast quantities of training data and processing resources. Time complexity is linear, feature aggregation and transformation happens in a single instance for CNN but for Transformers, these operations take place separately. So the efficiency is more for Transformers than CNN.

This paper advocates the comparative analysis of the classification of images into eight different classes of weeds and one negative class by training using the Vision Transformer model and four different architectures of CNN: ResNet-50, Xception, Inception V3 and Inception-Resnet V2. Vision Transformer outperformed these CNN models giving a better accuracy with less computational cost. In this paper, the related works section is followed with the methods and materials used, with an elaborative study on convolution-free, self attained Vision transformer, analysing the merits of transformers over CNNs, followed by experimental results and performance measures, conclusion and future work.

2. Related Works

Applications of Deep Learning in Precision Agriculture have been a popular subject of research in recent times. Weed detection and other agriculture-related research made use of hand-crafted techniques in the olden days. However, most of this relied on human experience and was time consuming. Hence deep learning was introduced into the field. But with the emerging technology in the Natural Language processing area, Transformer architectures are revolutionising NLP and Image processing. They are slowly replacing CNNs in Computer Vision, although they were originally developed to solve Machine Translation. In 2022, [3] authors of this paper reviewed the applications of CNN based supervised learning, transfer learning and few-shot learning in crop sensing. Based on a hybrid deep learning model, in 2022 [4] research provides a unique technique for human action identification. The suggested method is tested on the difficult datasets UCF Sports, UCF101, and KTH. When tested on the KTH dataset, it achieved an average of 96.3 percent accuracy. Transfer learning is one strategy employed these days to reduce overfitting. In 2021, [5] project's goal is to quantify the models' performance by optimising accuracy gained from a small dataset. Its goal is to create a system that recognises American Sign Language hand motions and detects alphabets. Both models produced good results, with VGG-16 outperforming the others. [6] In a study conducted in 2021, the authors identified three key issues that CNN-based approaches face and investigated the feasibility of conducting specialised transformer modules to solve them. They proposed a Multi-label Transformers architecture (MlTr) that combines windows partitioning, in-window pixel attention, and crosswindow attention to improve multi-label imagery classification performance. In 2021,[7] in this study, along with convolutional and recurrent neural networks, transformers have been dubbed the fourth pillar of deep learning. Transformers, on the other hand, are considerably more than that in terms of Natural Language Processing. In 2021, [8] COVID-19 impacted, Pneumonia affected, and normal were the three classifications utilised. The accuracy of deep learning models like VGG-16 and ResNet-50 were compared after the data was preprocessed. Data is classified as COVID-19, Pneumonia, or Normal by models. Results revealed that ResNet-50 gave the highest accuracy followed by VGG-16 and the other CNN model. In 2020, [9] Graph Weeds Net is a revolutionary graph based learning system presented by the authors, which was evaluated on the DeepWeeds dataset. Other common datasets include Grass-Broadleaf dataset, Flavia dataset, and Plant seedlings dataset. In 2019, [10] authors developed a big, accessible, multiclass picture collection

(DeepWeeds) of Australian rangeland weed species. This dataset depicts crops and weeds on land. Their work explored the performance of deep learning models like ResNet-50 and Inception-v3 in weed classification. In 2019,[11] authors did an Introductory Survey on attention mechanisms in Natural language Processing problems, through recent projects and perform an introductory summary of the attention mechanism in various NLP problems, with the goal of providing our readers with a basic understanding of this widely used approaches, discussing its various variants for various tasks, exploring its association with other machine learning techniques, and examining methods for assessing its efficiency. In 2019, [12] authors used deep learning to classify images to pleasant and unpleasant categories to understand how people react to some visual stimuli. In 2018, [13] authors used deep learning to detect spam images and verified that convolutional neural network approaches gave a higher accuracy than machine learning techniques. In 2018, [14] the author conducted a survey of forty research papers that used deep learning techniques to solve various agricultural challenges. Their findings showed that deep learning provided higher accuracy and outperformed commonly used handcrafted image processing techniques. One of the drawbacks in the olden times was the lack of publicly available agricultural datasets. In 2017 [15], the authors propose the Transformer, a new basic network design based purely on attention processes, with no recurrence or convolutions. Experiments on two machine translation tasks reveal that these models are higher in quality, parallelizable, and need much less training time. Another dataset that was presented in 2017 was the plant-seedlings dataset [16]. This dataset consists of around 407 images of 12 species of plants grown indoors in styrofoam boxes. A fully automated segmentation approach based on complex diffusion was proposed in 2013, [17]. The nonlinear complex diffusion approach was used to create a multiscale representation of the picture. Pixels were divided into segments using an intensity-based linkage model. In 2011 [18] the purpose was to provide farmers in India with crucial facts about crops unique to topology, geography, and climatic conditions so that they may better equip themselves to generate higher yields. This project's long-term objective was to make an area self-sufficient in the crops it can cultivate. This avoided the need to import or strain precious or overburdened resources to grow crops that weren't suited to a certain topology. This paper explained the KARSHIK system. After a part dedicated to the description of the special necessity for implementation, the article addressed the system's use in Kerala, India, for the research of crops that fit a certain terrain. The study concluded with a forecast for the system's future applications.

This research aims to explore the scope of Vision Transformers commonly used for Natural Language Processing Tasks in Computer Vision and to see if it gives superior results to the Convolutional Neural Network models commonly used in previous research works.

3. Materials and Method

3.1 Dataset

The publicly available DeepWeeds dataset was used. This dataset comprises a total of 17,509 images spreading over eight different classes of weed species native to Australia and a good number of negative class images. This dataset comprises images taken in a natural environment and hence gives a slightly less accuracy compared to other datasets like the grass-broadleaf dataset which has weeds segmented from a cluttered background. The distribution of weed species in the dataset is given in Table 1.

Table 1. Distribution of Weed Species

Name of Species	Count
Chinee apple	1125
Lantana	1064
Parkinsonia	1031
Parthenium	1022
Prickly acacia	1062
Rubber vine	1009
Siam weed	1074
Snake weed	1016
Negatives	9106

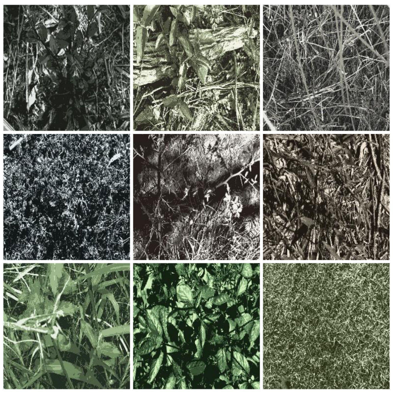

Fig. 2. Images from DeepWeeds Dataset

3.2 Data Preprocessing

The dataset was split into 80% training and 20% testing data. Except for the negative class, which is substantially bigger, stratified partitioning was used to assure equitable division of the classes inside each subset. Stratified partitioning is the process of separating data into subgroups and then randomly picking each subset to create a test group. The dataset was already available in open source in the above mentioned preprocessed form. The ImageNet dataset was used to train all models and TensorFlow libraries were used to import them. The experiment was run on an Intel i5-11400f and an Nvidia RTX 2080 super environment.

3.3 Classification

We trained the dataset on Vision transformer and 4 CNN models: ResNet-50, Xception, Inception V3, and Inception-Resnet V2, for the problem statement and performed a comparative analysis on these models based on their training statistics and outcomes to determine which model performed the best out of the five.

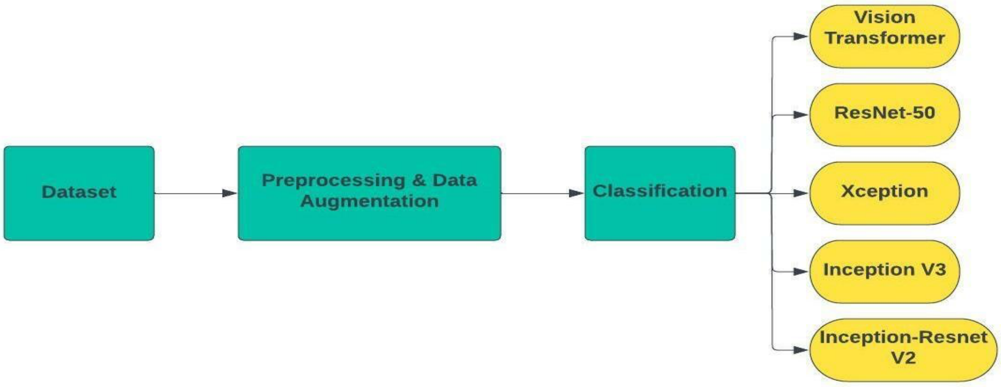

Fig. 2. Flow process of the experiment

Vision Transformers: Transformers are solely based on self-attention based architectures. The self-attention mechanism helps explicitly model the relationship between various elements of a sequence and hence have found vast application in predictive tasks. This mechanism helps predict the relevance of one element with respect to other elements in the sequence. Researchers claim that transformers work better than recurrent neural networks and feed forward neural networks even though they also make use of attention models. While recurrent neural networks can only attend to short-term contexts, transformers can learn long-range relationships.

Vision transformers were first proposed in late 2020 by the Google Brain Team. Ever since the immense success of transformers in natural language processing tasks they have been marked for experimentation in the computer vision field. Vision transformers use only the encoder part of the transformers. The images are split into 16 x 16 patches. The patches are flattened and further preprocessed. Since the model has no idea about the position of the patches or samples in the image, they are fed along with a positional embedding vector as linear sequences into the transformer encoder. Here, the image patches are counterparts of tokens in

Natural Language Processing applications. Unlike in standard transformers, the output from vision transformers are passed to a feed forward neural network instead of decoders to get the classification output.

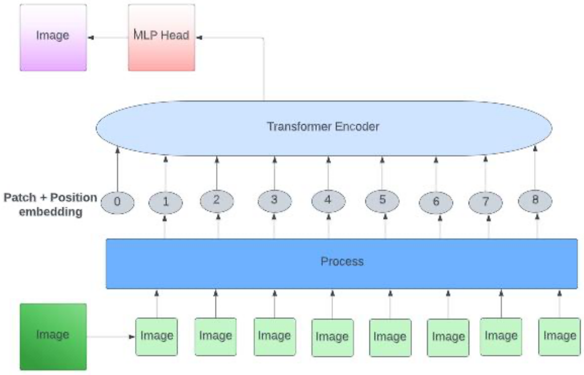

Fig 3. Vision Transformer Architecture

Resnet-50 Model: ResNet (Residual Network) is a powerful convolutional neural network model that has been dominating the computer vision field since 2015. ResNet50 is a variant of this model which is 50 layers deep (48 convolutional layers, 1 Max Pool, 1 Average Pool layer). The vanishing gradient problem is solved by ResNet by adding skip links to the network. The vanishing gradient problem is often encountered while training networks that involve gradient based learning and backpropagation. The skip connections enable skipping the learning procedure for a few specific layers and hence enable the network to carry forward the gradient throughout the extent without losing out on important information. Hence they can be used to effectively train deep neural networks. The model managed to achieve an accuracy of 80.67% on ImageNet dataset.

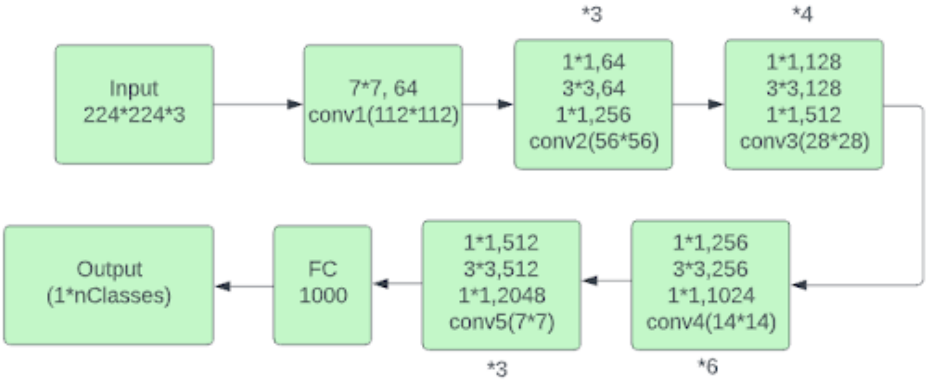

Fig 4. ResNet-50 Architecture

Xception Model: Xception, also referred to as an extreme version of the Inception module, is a deep convolutional neural network architecture introduced by Francois Chollet (google researcher). Xception architecture relies on depth wise separable convolutions which replace the standard Inception modules. In Xception, the depthwise convolutions are followed by max pooling instead of pointwise convolutions, and all are linked with residual connections as in ResNet architecture. In depthwise convolution, a convolution of size d x d x1 is applied instead of d x d x C where C is the number of channels. Depthwise convolutions are said to be more efficient in terms of computing than traditional convolutions. On the ImageNet dataset, Xception achieved a top1 accuracy of 79 percent and a top5 accuracy of 94.5 percent.

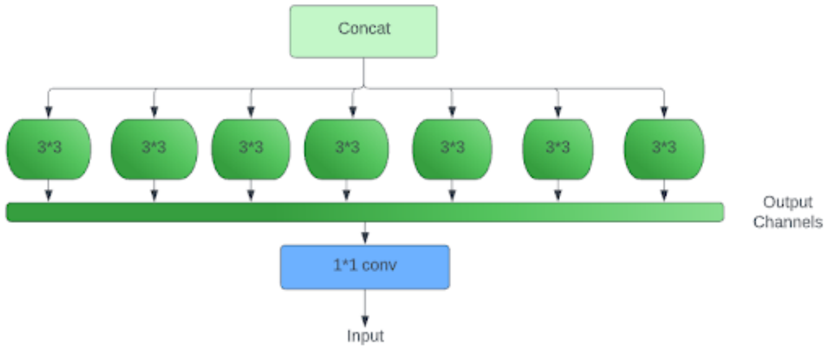

Fig 5. Xception Architecture

Inception V3 Model: Inception V3 model is a CNN used for image classification and object detection which originated from GoogleNet. The ImageNet Recognition Challenge was the first to use it and as the name suggests this is the third addition to the inception convolutional neural network by Google. The main goal here was to keep the number of parameters limited and still increase the depth of the network. The model manages to have less than 25 million parameters which is quite less compared to Alexnet which has around 60 million parameters. Inception V3 has 42 layers and a much lower error rate than Inception V1 and Inception V2. Inception V3 yields better results at the same time with lower computational costs than its predecessors. The model managed to attain an accuracy of 78.8% on the ImageNet dataset.

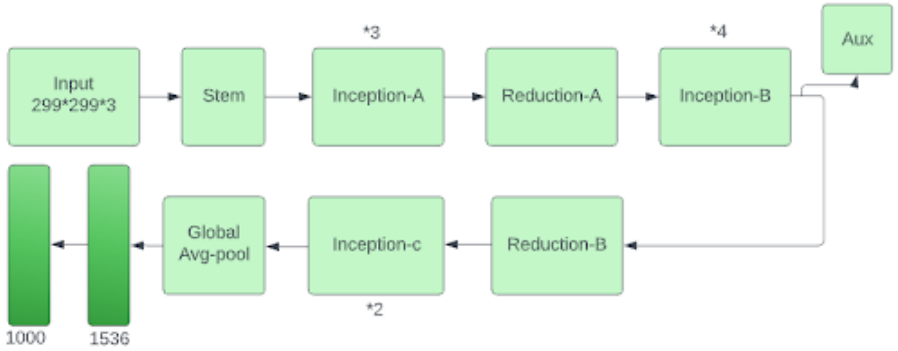

Fig 6. Inception V3 Architecture

Inception-Resnet V2 Model: The convolutional neural network Inception-ResNet V2 is an advancement on the Inception V3 architecture and a million images from the ImageNet database were used to train the model. It takes input images of size 299 x 299 pixels and produces a list of anticipated class probabilities. It is developed on a foundation of the Inception structure and the Residual connection. The residual connections help to reduce training time and avoid degradation which is caused by deep structures. The model has 164 layers and requires just double the memory and computation compared to Inception V3. On the Imagenet dataset, the Inception-Resnet V2 yields an accuracy of 80.4% compared to 78.8% of Inception V3.

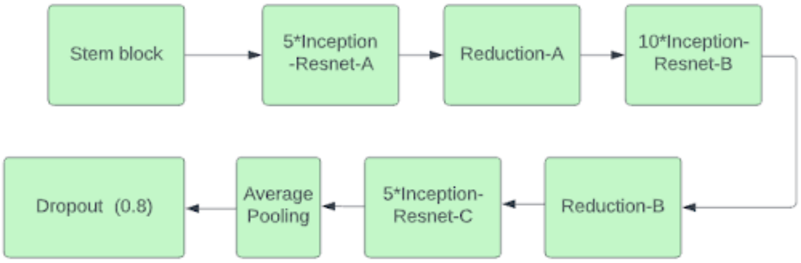

Fig 7. Inception-Resnet V2 Architecture

4. Results and Discussion

We performed the experiment on the Deep Weeds dataset using Vision Transformer and four CNN models namely Resnet-50, Xception, Inception V3 and Inception-Resnet V2. The accuracy of each of the models on the dataset are listed in Table II. This competitive analysis shows that Vision Transformer gave the highest accuracy of 96.41% followed by ResNet-50 with an accuracy of 95.70% and Xception model with an accuracy of 95.04%. The Inception V3 model gave an accuracy of 94.7% followed by the Inception-Resnet V2 model which gave an accuracy of 94.15%. Figures 7-11 show the graphical representation of Training and Validation accuracy and loss of Vision Transformer and the four CNN models: ResNet-50, Xception, Inception V3 and InceptionResnet V2. The results clearly show that transformers have great future in the field of Computer Vision as they show better performance compared to CNN models. Transformers apply the mechanism of attention and support parallel processing using similar processing blocks. Though CNNs have provided upto 99% accuracy for weed detection in the past, transformers are computationally more efficient while providing decent results and hence provides a promising path in the future.

Table 2. Model Classification Results

Models Used	Accuracy
Vision Transformer	96.41 %
ResNet-50	95.70 %
Xception	95.04 %
Inception V3	94.7 %
InceptionResnet V2	94.15 %

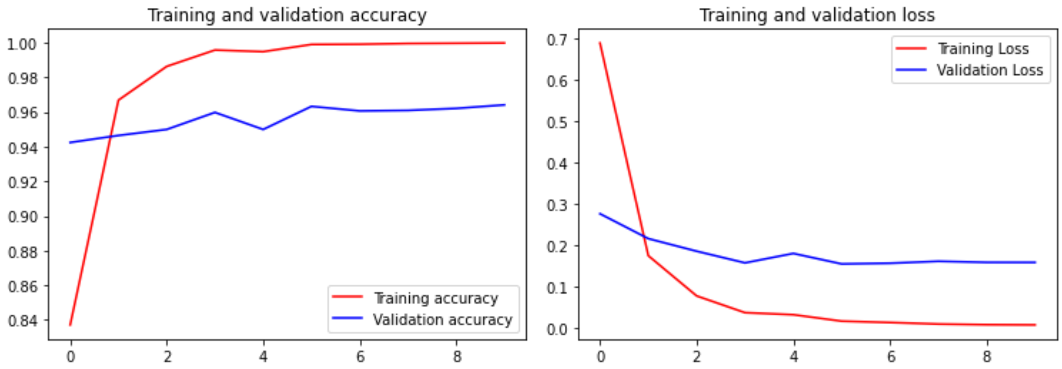

Fig 8. Training and Validation accuracy and loss of Vision Transformer

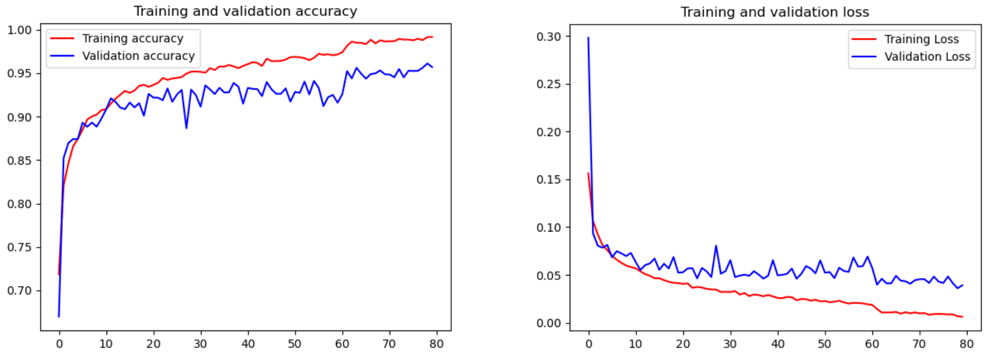

Fig 9. Training and Validation accuracy and loss of ResNet 50

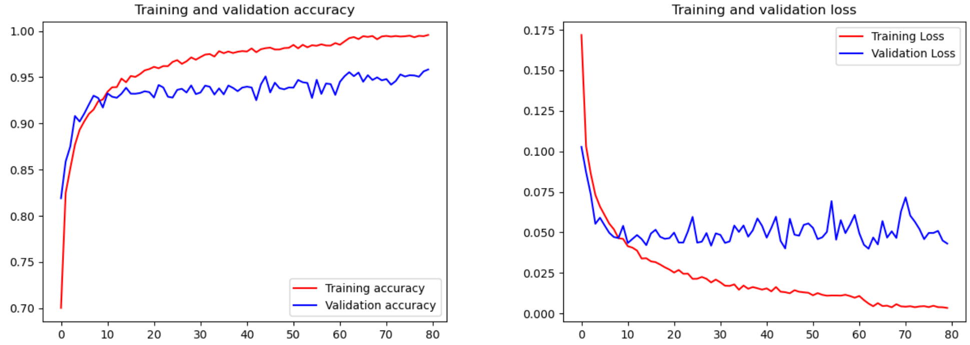

Fig 10. Training and Validation accuracy and loss of Xception

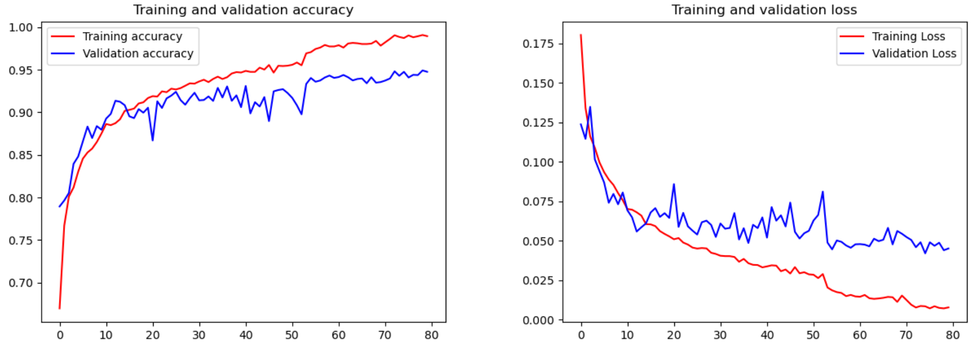

Fig 11. Training and Validation accuracy and loss of Inception V3

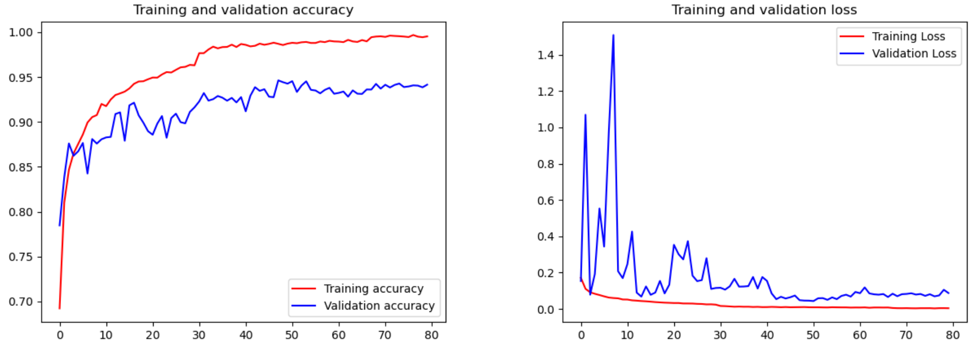

Fig 12. Training and Validation accuracy and loss of Inception-Resnet V2

5. Conclusion and Future Work

In this comparative study, the classifiers evaluated and analysed were Vision Transformers, Resnet-50, Xception, Inception-Resnet V2 and Inception V3. The results revealed that Vision Transformer outperformed the CNN models giving an accuracy of 96.41% compared to the CNN models : ResNet-50 which gave an accuracy of 95.70% and Xception gave an accuracy of 95.04%. This was followed by the Inception V3 model with an accuracy of 94.70% and Inception-Resnet V2 with an accuracy of 94.15%. Transformer excels and outperforms other common image processing approaches, according to the results. The authors want to adapt the broad principles and best practices of the emerging technologies of transformers which were earlier used in Natural Language Processing tasks to other sectors like Computer Vision where this modern method is yet to be fully utilised in the future. Transformers' overall advantages suggest that it has great potential in reducing the environmental impact on agriculture and in creating a safer and quality food supply. Hybrid models of transformer and CNN can be used for better results in future.

6. References

[1] Zhangnan Wu, Yajun Chen, Bo Zhao, Xiaobing Kang, Yuanyuan Ding,"Review of Weed Detection Methods Based on Computer Vision", May 2021.

[2] Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," arXiv, 2021.

[3] Dashuai Wang Wujing Cao Fan Zhang, Zhuolin Li , Sheng Xu and Xinyu Wu," A Review of Deep Learning in Multiscale Agricultural Sensing", 25 January, 2022

[4] Dr. Priyanka Kumar, V S Akella, C Vijayendra Sai, Kakarla Ajay Kumar Reddy,"A hybrid architecture for Action Recognition in Videos using Deep Learning",Publisher : Jadavpur University

[5] Aswathi Premkumar, Hridya Krishna R, Nikita Chanalya, Meghadev C, Utkrist Arvind Varma, Anjali T, and Siji Rani S,"Sign Language Recognition: A Comparative Analysis Of Deep Learning Models", by the Technical Program Committee (TPC) for oral presentation during Fourth IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT 2021)

[6] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, Dong Shen, Zhongyuan Wang, Nian Shi, Honglin Liu,"MlTr: Multi-label Classification with Transformer",arXiv:2106.06195 [cs.CV],11 Jun 2021

[7] James Briggs,"Multi-Class Classification With Transformers, Preprocess, train, and predict with BERT",Published in Towards Data Science Mar 25, 2021

[8] Hridya Krishna, R.; Vaishnavi, K. P.; Anagha Ramadas, M.; Chanalya, N.; Manoj, A.; Nair, J. J., "Deep Learning Approaches for Detection of Covid-19 Using Chest X-Ray Images", 4th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2021

[9] Hu et al. 2020" Graph weeds net: A graph-based deep learning method for weed recognition" in Computers and Electronics in Agriculture, July 2020.

[10] Alex Olsen, Dmitry. Konovalov, Bronson Philippa, Peter Ridd, Jake C.Wood, Jamie Johns, Wesley Banks, BenjaminGirgenti, Owen Kenny, JamesWhinney, BrendanCalvert, Mostafa RahimiAzghadi & Ronald D.White, "DeepWeeds: A Multiclass Weed Species Image Dataset for Deep Learning", 14 February 2019.

[11] Hu, "An introductory survey on attention mechanisms in nlp problems," IntelliSys, 2019.

[12] S. Tamuly, Jyotsna C, and Amudha J., "Deep Learning Model for Image Classification", Advances in Intelligent Systems and Computing, vol. 1108. Springer International Publishing, Cham, 2019.

[13] A. Dinesh Kumar, R, V., and Dr. Soman K. P., "Deepimagespam: Deep learning based image spam detection", arXiv preprint arXiv:1810.03977, 2018

[14] Andreas Kamilaris," Deep Learning in Agriculture: A Survey" in Computers and Electronics in Agriculture, April 2018.

[15] Ashish Vaswani,Noam Shazeer,Niki Parmr, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", arXiv:1706.03762v5 [cs.CL] 6 Dec 2017

[16] Thomas Mosgaard Giselsson,Rasmus Nyholm Jørgensen,Peter Kryger Jensen,Mads Dyrmann,Henrik Skov Midtiby,"A Public Image Database for Benchmark of Plant Seedling Classification Algorithms", 15 Nov 2017.

[17] Jyothisha J Nair, V K Govindan, "Multi-scale Segmentation Based on Nonlinear Complex Diffusion" Journal of Medical Imaging and Health Informatics vol 3(2), 242-245, 2013. American Scientific Publishers.

[18] Nima S. Nair, Vasudevan, A.; Benny, A.S.; Shabana, K.M.; Shenoy, A.; Dutta, M,"KARSHIK: Agricultural information monitoring and reference based on wireless networks",ACWR 2011 - Proceedings of the International Conference on Wireless Technologies for Humanitarian Relief