Plant Disease Detection Using a Hybrid Approach Based on Vision Neural Network Transformers

Mohamed Walid Hajoub
SIGL laboratory, ENSATe, Abdelmalek Essaadi University, Tetouan, Morocco
mohamedwalidhajoub1@gmail.com

Hicham Touil
SIGL laboratory, ENSATE, Abdelmalek Essaadi University, Tetouan, Morocco
touil.hm@gmail.com

Mohammed Achkari Begdouri
SIGL laboratory, ENSATe, Abdelmalek Essaadi University, Tetouan, Morocco
m.achkaribegdouri@uae.ac.ma

ABSTRACT

Plant disease detection is one of the main challenges in the agricultural sector, acknowledged as a significant contributor to crop losses, negatively impacting food production and the global economy. In response to this issue, novel agricultural solutions are emerging, leveraging the synergy of deep learning and computer vision technique for early disease detection. Recently, researchers have embraced vision Transformers for plant disease identification. While this approach has shown promising results, it also presents challenges, such as high computational costs and low inductive bias for locality. In response to these challenges, we suggest a hybrid model with fewer trainable parameters, combining the power of vision Transformers with the capabilities of convolutional layers to extract relevant local features, thereby enhancing the performance of ViT. Additionally, we modify the ViT architecture to minimize the number of trainable parameters. We performed experiments on a public dataset of potatoes to assess the effectiveness of our proposed hybrid model. According to the experimental results obtained, the proposed model achieved a test accuracy of 98.27%, surpassing the original ViT and two leading CNN architectures, namely VGG16 and ResNet50. Our model has fewer parameters than the original ViT, with a reduction rate of 49%. This reduction helps to minimize the computational and memory costs of our model. Additionally, due to this parameter reduction, our proposed model requires half of training time compared to the original ViT. All these findings make our proposed model more adaptable for smart agriculture applications, particularly in real-time detection of plant diseases.

KEYWORDS

Plant Disease Detection, Vision transformers, Computer Vision, Image Processing, Convolutional Neural Network, Machine Learning, Deep Learning, Intelligent System, Smart Monitoring, Precision Agriculture

ACM Reference Format:
Mohamed Walid Hajoub, Hicham Touil, and Mohammed Achkari Begdouri. 2024. Plant Disease Detection Using a Hybrid Approach Based on Vision Neural Network Transformers. In *The 7th International Conference On Networking, Intelligents Systems and Security (NISS 2024), April 18, 19, 2024, MEKNES, AA, Morocco.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3659677.3659685

NISS 2024, April 18, 19, 2024, MEKNES, AA, Morocco

ACM ISBN 979-8-4007-0929-6/24/04
https://doi.org/10.1145/3659677.3659685

1 INTRODUCTION

In the last few years, agriculture has become a crucial source of income and development for numerous countries, playing a significant role in the global economy [1] . The exponential growth of humanity, projected to exceed ten billion within the next three decades, underscores the urgency to enhance agricultural production to meet the escalating food demand. Agriculture is presently grappling with numerous challenges that impede its capacity to fulfill the mounting requirements for food products [2].

Plant diseases pose a formidable challenge with a detrimental impact on food production. When a crop is affected by a disease, specific symptoms manifest at the leaf level, including changes in leaf shape, color, and texture [3].

The world has witnessed substantial progress in the development of artificial intelligence and computer vision tools and techniques, significantly impacting and improving agricultural practices. In this context, the early and precise identification of plant diseases emerges as a crucial solution to minimize crop losses. Various methods, particularly leveraging deep learning, such as deep convolutional neural networks, have been created to quickly detect and identify plant diseases in real time.

More recently, ViT or Vision Transformers have surfaced as a notable deep learning architecture showing promising results and robust performance, occasionally surpassing CNNs in solving image classification problems. This has captured the attention of researchers, leading them to explore the use of ViT to process images leaves and category them as healthy or diseased. Despite the exceptional performance of ViT, it exhibits significant drawbacks, such as high computational complexity and low inductive bias for locality.

The essential objective of this research is to develop and evaluate a hybrid approach that leverages the strengths of vision transformers and the powerful capabilities of convolutional neural networks in extracting relevant features and local information for plant disease detection. This integration aims to enable farmers to monitor and control their farms in real-time. Specifically, we seek to address the following research question: Can the combination of transformers and convolutional neural networks effectively identify and detect diseases from leaf images?

The main objectives of this research:

- Developing a hybrid model for identifying plant diseases using a dataset containing images of different diseased leaves.
- Assess the performance of our model using diverse evaluation metrics including accuracy score, F1-score, and number of parameters.
- Compare the effectiveness of the hybrid model in plant diseases detection with that of standard models to assess the proposed approach.

2 RELATED WORK

Precision agriculture, also known as smart farming, refers to a technological framework that integrates technologies such as computer vision, big data, visual information processing, and artificial intelligence techniques to cultivate crops in a sustainable and efficient manner [4]. Plant diseases detection is a significant research area in the field of precision agriculture. Several studies have been conducted in this field, exploring the use of these technologies for the real-time detection and monitoring of plant diseases.

One study by Mohanty et al. (2016) [5] used CNN to identify plant diseases. Evaluating the performance of GoogLeNet and AlexNet on an open-source dataset, the researchers found that GoogLeNet surpassed AlexNet, achieving an impressive test accuracy of 99.35%.

Another study by Turkoglu et al. (2022) [6] introduced a novel hybrid neural network that integrates AlexNet, ResNet, DenseNet and GoogLeNet, and Support Vector Machines for disease identification and classification. The test accuracy achieved by the proposed model was 96.83%. Another research by Biswas and Yadav (2023) [7], they presented work in which various cutting-edge CNN structures for the detection of plant diseases.

Another study by Nigam et al. (2023) [8], they conducted an experimental investigation using eight CNN architectures based on EfficientNet to identify leaf diseases. Their findings revealed that the EfficientNet-B4 CNN architecture outperformed alternatives such as VGG 19, ResNet152, DenseNet169, InceptionNetV3, and MobileNetV2, achieving a remarkable test accuracy of 99.35%.

More recently, the ViT architecture has been introduced in plant disease identification. A. Berka et al. (2023) [9] introduced CactiVIT, a system that employs the Vision Transformer to rapidly assess the health status of cactus, achieving an overall accuracy of 88.73%, which is on average +2.61% higher compared to other CNN models.

Another research by Thai et al. (2021) [10], the researchers employed fine-tuning techniques with vision neural network transformers for the purpose of detecting cassava diseases. Their reported findings highlighted an impressive F1-score of 90.3% for the ViT model. Notably, they went further to implement their model on an IoT device. Nevertheless, a drawback of their approach is the model's substantial parameter count, totaling 85.79 million, rendering it impractical for certain applications.

S.Parez et al. (2023) [1] developed a GreenViT model based on ViT to detect plant infections and diseases. The model achieved exceptional results due to the reduction in the number of encoders in the proposed architecture.

Some recent methods have integrated a fusion of convolutional blocks and vision transformers to develop efficient models for plant disease detection. In a study conducted by Shuqin Li et al. (2022) [3], they proposed a multi-stage model for Kiwi disease identification based on ViT and CNN called ConvViT, achieving a 98.78% accuracy in identifying diseases within the kiwifruit dataset. This marked an improvement of up to 4.53% in identification accuracy compared with similar-level models such as ResNet, ViT, and ResMLP.

In their study, Thakur et al. (2022) [2] designed a model combining CNN With ViT, demonstrating robust performance across diverse crops and associated diseases. The researchers conducted training and testing using two public datasets, achieving notable accuracy scores of 98.61% on PlantVillage and 87.87% on Embrapa.

However, to our knowledge, there is limited research combining the advantages of vision transformers and convolutional neural networks for plant disease detection. Our study aims to develop a hybrid model based on ViT by introducing a convolutional layer into the ViT architecture, enhancing the inductive bias for locality in the design. Furthermore, we will reduce the number of encoders in our hybrid model compared to the original architecture, aiming to minimize trainable parameters to optimize the computational complexity of the model. Finally, we will assess the precision and effectiveness of the suggested hybrid approach.

3 METHODOLOGY

3.1 Image Analysis and Computer Vision

Computer vision is a computer science discipline focused on empowering computers to interpret, comprehend, scrutinize and analyze digital information, particularly images and videos, holds potential applications in diverse domain such as agriculture, security, robotics and medicine. Computer vision has advanced rapidly due to machine and deep learning methods, delivering robust outcomes in tasks such as segmentation task, automatic detection, and image classification.

If a paper is accepted for publication, authors will be instructed on the next steps. Authors must then follow the Image processing is considered the initial step in any task of computer vision. This process includes operations such as augmentation, filtering, resizing, scaling, normalization, denoising and edge detection. After the image completes the preprocessing stage, the subsequent step involves feature extraction, which entails identifying relevant information and patterns crucial within the image for a specific task. This can be achieved through manually crafted features, demanding in-depth knowledge, or automatically learned features using deep learning [11].

CNN short for Convolutional Neural Networks, are a category of deep neural networks, have exhibited remarkable effectiveness in tasks like image classification and identification, including the detection of plant diseases. CNNs utilize a hierarchical architecture involving several layers of convolution and pooling operations.

Convolution layers are designed to extract specific features from the image at a local level, whereas pooling layers reduce the size of "feature maps" and diminish their spatial dimensions. Ultimately, the output obtained from the convolutional layers is transmitted to one or multiple fully connected layers, which are responsible for performing image classification tasks [4].

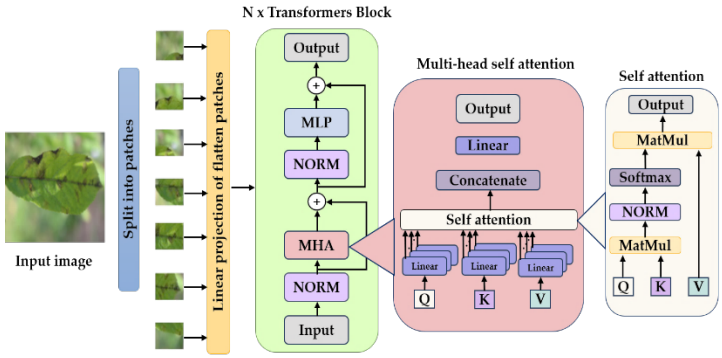

Figure 1: Vision Transformers Architecture.

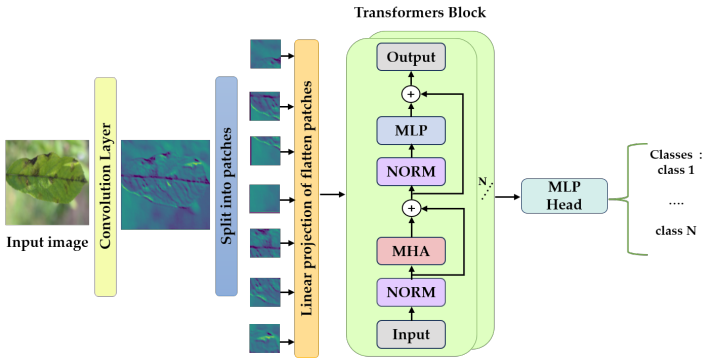

Figure 2: hybrid ViT-CNN model Architecture.

3.2 Vision transformers

Building upon the notable success of Transformers in natural language processing [12], Dosovitskiy et al. [13] introduced vision transformers, derived from the original Transformer architecture. The ViT consists of several stacked encoder blocks, each incorporating three modules within the ViT model: a multi-head attention (MHA) or self-attention, normalization layer (NORM), and an MLP layer [14]. The architecture of the ViT model is depicted in Figure 1.

To input an image into the original vision transformers, it undergoes an initial division into non-overlapping patches of a fixed size. Subsequently, these overlapped patches are compressed and converted into representations with fewer dimensions. Every flattened patch undergoes a trainable linear transformation to generate its respective "positional encoding" along with "linear projection".

The resulting vector from these operations is fed through a sequence of N ViT transformer blocks [2]. Ultimately, the final output of ViT is directed into one or multiple fully connected layers, which are responsible for performing image classification.

3.3 Model for plant diseases detection using a hybrid ViT CNN approach

The objectives of this study are to develop a hybrid model that integrates the capabilities of ViT and CNN for the identification and detection of plant diseases. Vision Transformers, equipped with a multi-head attention mechanism, adeptly capture long-distance dependencies among image patches, empowering them to discern intricate patterns within the image [13].

Vision Transformers encounters limitations in capturing local features in images. To overcome this challenge, this study employs a convolutional layer to proficiently extract local features, thereby enhancing the model's inductive bias for locality. The proposed hybrid model's architectural representation is shown in Figure 2.

All input images, initially with dimensions of 224×224×3, undergo a convolutional layer to extract local features. The feature map from this convolutional layer is then converted into patches, each with dimensions of 16×16. After flattening, these patches are linearly projected, producing a feature vector with dimensions of 196×768. Subsequently, these vectors pass through a stack of N transformer blocks intended to extract global features. The number of transformer blocks in this study is minimized to reduce the computational cost of the proposed model.

The outcome of vision neural network Transformer block is transformed into a one-dimensional vector. For classification purposes, an MLP block is introduced, comprising a dense layer with Softmax activation. This layer contains neurons corresponding to number of categories in dataset.

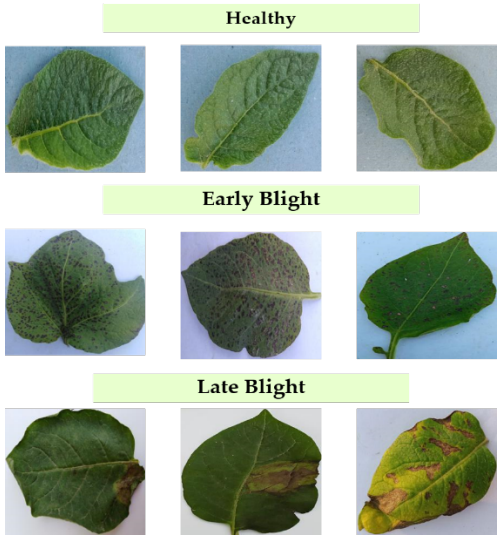

Figure 3: Examples of potatoes disease dataset.

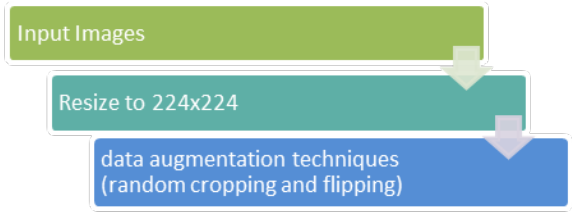

Figure 4: Pre-processing steps.

3.4 Datasets

To assess the efficacy of our model, we employ a publicly available dataset on potato plant diseases. The dataset consists of 4072 items, categorized into 3 categories based on type of potato disease. The data is devised into train set, validation set, and test set, comprising 80%, 10%, and 10% of the dataset, respectively. This allocation amounts to 3251 items for train set, 416 for validation set, and 405 for testing the model. Samples of leaves are depicted in Figure 3.

4 IMPLEMENTATION & EVALUATION

Our model underwent training using the potato dataset. Data preprocessing involved resizing each input image to 224x224x3. To address overfitting, data augmentation was introduced post-preprocessing. Augmentation techniques applied to the pre-processed images involved adjustment, counterclockwise rotation, horizontal, clockwise, zoom intensity, vertical flipping and scaling.

To expedite the learning process, we employed the fine-tuning technique. This approach involves taking a pre-trained model on a general task and adjusting it for a more specific task or a particular dataset. Fine-tuning leverages the knowledge already acquired by the pre-trained model to accelerate and improve learning on the new task [15].

4.1 Hyperparameters

In developing our model, we considered various hyper-parameters, such as the "learning rate", choice of "optimizer", size of the convolution kernel and patch, the "number of Transformer blocks", and the number of "Multi-Head Attention" (MHA) units.

Initially, we selected the "Adam optimizer", setting A low learning rate of 10-4 was used to ensure the preservation of previously learned knowledge. A 3x3 kernel size was utilized by the convolutional layer for extracting local features. Additionally, a patch size of 16 was chosen, with 6 transformer blocks or encoders, each having 4 attention heads. The model underwent training for 50 iterations, batch size equals 32 was utilized, and the loss function employed was Categorical Cross-Entropy.

4.2 Evaluation Metrics

To evaluated the effectiveness of our hybrid model, we utilized four performance metrics: accuracy score, recall metric, precision score and F1-score. In this research, true positives (TP) quantify leaves correctly identified as diseased, while true negatives (TN) express the number of leaves correctly categorized as healthy. False positives (FP) represent healthy leaves erroneously classified as diseased, and false negatives (FN) indicate the number of diseased leaves incorrectly classified as healthy.

$$Accuracy = (TP + TN)\ /(TP + TN + FP + FN) \quad (1)$$

$$Precision = TP/(TP + FP) \quad (2)$$

$$Recall = TP/(TP + FN) \quad (3)$$

$$F1 - score = \ 2/((1/precision) + (1/recall)) \quad (4)$$

5 RESULTS & DISCUSSION

The proposed model underwent training and testing on a dataset of potato images. The prediction performance was assessed on validation and test subsets, and a comparative study was conducted between the original transformers and two CNN architectures to evaluate the capabilities and performance.

5.1 Results

Our model attained a peak accuracy of 98.56% on the validation data, surpassing the original Vision Transformers model with an accuracy of 98.32%, resulting in an improvement of +0.24%. Concerning the validation loss, our model reached a minimum value of 0.044, while the original Vision Transformers exhibited a loss of 0.046.

Our model achieved a maximum validation accuracy (+1,21%) and minimum validation loss when compared with two CNN architectures. Additionally, the VGG16 and ResNet50 models demonstrated comparable accuracy rates of 97.35% and 97.11%, respectively, with similar loss rates of 0.077 and 0.095, respectively.

The original vision transformers achieved the most favorable secondary outcomes in terms of validation loss and accuracy, whereas ResNet50 obtained lower results. Figures 5 and 6 illustrate a comparison of the validation accuracy and loss of our model, original vision transformers and two CNN architectures.

Additionally, we computed several evaluation metrics including Accuracy score, F1-score, and the number of trained parameters for the proposed model in this study, as well as for the original Vision Transformers and the two CNN architectures, on the testing

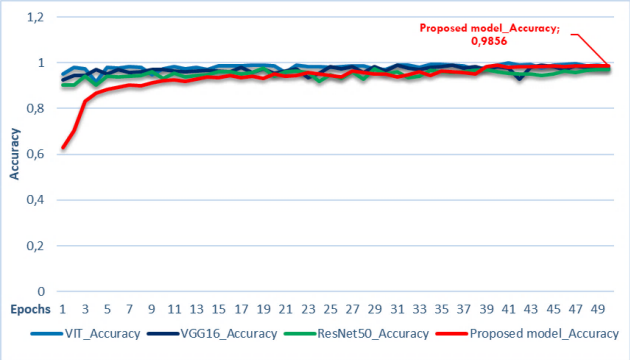

Figure 5: Evolution of validation accuracy between: our model, original ViT, and two CNN architectures.

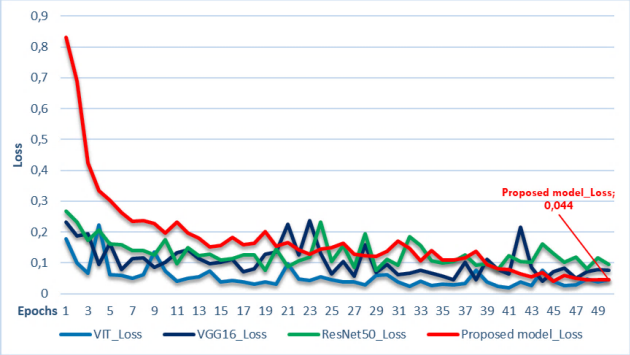

Figure 6: Evolution of validation loss between: our model, original ViT, and two CNN architectures.

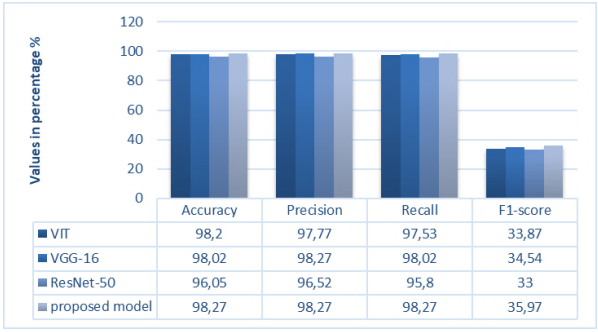

Figure 7: Comparative analysis of evaluation metrics between the proposed model, original ViT and Two CNN architectures.

subset of the potato dataset to further analyze the effectiveness of the proposed model. Figure 7 displays these results.

From Figure 7, it can be observed that our model achieved better results for each of the mentioned metrics, namely 98.27% accuracy, 98.27% precision, 98.27% recall, and 35.97% F1-score on the Potato dataset. Although vision transformers are close and similar to our model, the latter demonstrated a more robust performance than ViT.

We also evaluated the total number of parameters employed by our model on the potato dataset, comparing them with those of the ViT model. We found that our model uses 43.8 million weight parameters, while Vision Transformers use 86.3 million, indicating a big difference in the number of parameters.

5.2 Discussion

This study aims to detect plant diseases early and efficiently using digital images and artificial intelligence techniques. Most of the related studies have used different machine learning algorithms to develop efficient intelligent systems in the detection of diseases.

Recently, several researchers have adopted vision transformers for disease identification, and although this approach has demonstrated robust results, it also has certain limitations. Among those mentioned in our literature review are the high computational cost of the model [16], limited generalizability [9], and a lack of sufficient data [17].

The model proposed in this study aims to overcome these limitations and enhance the performance of vision transformers in plant disease detection tasks. To achieve this, several modifications have been made. First, a convolutional layer was added to increase the inductive bias of locality, enabling vision transformers to focus on local features as well. Additionally, a reduction in trainable parameters was implemented to reduce the complexity of the model.

The experiments carried out on the potato dataset revealed that the proposed model outperforms the original ViT in terms of accuracy during tests for detecting plant diseases using images of leaves. Even when reducing the number of transformer blocks and attention heads in each block, our model obtained results that surpass the original vision transformers, ResNet50 and VGG16.

The approach of reducing the number of transformer blocks and attention heads within the architecture of vision transformers significantly decreases the number of trainable parameters, with a reduction rate of 49%. This reduction minimizes the computational and memory costs of our model. Moreover, the reduction in the number of parameters improves the model's suitability for smart agriculture applications.

The reduced computational complexity of our model enhances its efficiency in both training and inference. In our experimental study, the original vision transformers require twice the training time compared to our proposed model, underscoring the pivotal role of number of parameters.

In this study, we assessed the performance of the proposed model on a limited dataset focused on potato diseases. Nevertheless, there are plans to further train the model on big larger disease datasets encompassing diverse images of leaves with various plant diseases.

The similarity of the patterns of diseases, colors, and textures has a negative impact on learning, explaining the low F1-score of our proposed model and the original ViT. Additionally, some diseases do not exhibit initial signs on the leaves, complicating the development of AI solutions.

6 CONCLUSION

The early identification of plant diseases is a pivotal challenge in the agricultural sector, holding the potential to minimize crop losses and maximize farmers' profits.

This study presents a hybrid approach model designed to quickly detect and identify diseases in real time. Integrating the global feature extraction capabilities of original vision transformers with the powerful capabilities of the layers convolutional neural networks in extracting relevant features and local information of convolutional neural networks enhances classification performance. A reduction in the number of trainable parameters has been implemented to mitigate complexity and meet hardware requirements.

The proposed model's performance was evaluated against both the original ViT model and two leading CNN architectures, namely VGG16 and ResNet50. Throughout the experimental study, the proposed model consistently outperformed all others models, achieving test accuracies of 98.27% on the potato dataset. Furthermore, it was observed that our proposed model utilizes a minimal number of trainable weight parameters, specifically 43.8 million, compared with ViT.

Future work involves enhancing the model's performance to address issues related to similarity. Additionally, deploying the model on IoT devices such as UAVs and Raspberry Pi will empower immediately real-time identification in agricultural sector.

REFERENCES

[1] S. Parez, N. Dilshad, N.S. Alghamdi, T.M. Alanazi, and J.W. Lee. 2023. Visual Intelligence in Precision Agriculture: Exploring Plant Disease Detection via Efficient Vision Transformers. *Sensors* 23, 15 (2023). https://doi.org/10.3390/s23156949

[2] P.S. Thakur, S. Chaturvedi, P. Khanna, T. Sheorey, and A. Ojha. 2023. Vision transformer meets convolutional neural network for plant disease classification. *Ecological Informatics* 77, (2023). https://doi.org/10.1016/j.ecoinf.2023.102245

[3] Xiaopeng Li, Xiaoyu Chen, Jialin Yang, and Shuqin Li. 2022. Transformer helps identify kiwifruit diseases in complex natural environments. *Computers and Electronics in Agriculture* 200, (September 2022), 107258. https://doi.org/10.1016/j.compag.2022.107258

[4] Abdullah Ali Salamai, Nouran Ajabnoor, Waleed E. Khalid, Mohammed Maqsood Ali, and Abdulaziz Ali Murayr. 2023. Lesion-aware visual transformer network for Paddy diseases detection in precision agriculture. *European Journal of Agronomy* 148, (August 2023), 126884. https://doi.org/10.1016/j.eja.2023.126884

[5] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. 2016. Using Deep Learning for Image-Based Plant Disease Detection. *Front. Plant Sci.* 7, (September 2016), 1419. https://doi.org/10.3389/fpls.2016.01419

[6] Muammer Turkoglu, Berrin Yanikoğlu, and Davut Hanbay. 2022. PlantDiseaseNet: convolutional neural network ensemble for plant disease and pest detection. *SIViP* 16, 2 (March 2022), 301–309. https://doi.org/10.1007/s11760-021-01909-2

[7] Barsha Biswas and Rajesh Kumar Yadav. 2023. A Review of Convolutional Neural Network-based Approaches for Disease Detection in Plants. In *2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, January 05, 2023, Bengaluru, India. IEEE, Bengaluru, India, 514–518. . https://doi.org/10.1109/IDCIoT56793.2023.10053428

[8] Sapna Nigam, Rajni Jain, Sudeep Marwaha, Alka Arora, Md. Ashraful Haque, Akshay Dheeraj, and Vaibhav Kumar Singh. 2023. Deep transfer learning model for disease identification in wheat crop. *Ecological Informatics* 75, (July 2023), 102068. https://doi.org/10.1016/j.ecoinf.2023.102068

[9] Anas Berka, Adel Hafiane, Youssef Es-Saady, Mohamed El Hajji, Raphaël Canals, and Rachid Bouharroud. 2023. CactiViT: Image-based smartphone application and transformer network for diagnosis of cactus cochineal. *Artificial Intelligence in Agriculture* 9, (September 2023), 12–21. https://doi.org/10.1016/j.aiia.2023.07.002

[10] Huy-Tan Thai, Nhu-Y Tran-Van, and Kim-Hung Le. 2021. Artificial Cognition for Early Leaf Disease Detection using Vision Transformers. In *2021 International Conference on Advanced Technologies for Communications (ATC)*, October 2021. 33–38. . https://doi.org/10.1109/ATC52653.2021.9598303

[11] Sheng Yu, Li Xie, and Qilei Huang. 2023. Inception convolutional vision transformers for plant disease identification. *Internet of Things* 21, (April 2023), 100650. https://doi.org/10.1016/j.iot.2022.100650

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Retrieved January 27, 2024 from http://arxiv.org/abs/2010.11929

[14] P. Gole, P. Bedi, S. Marwaha, M.A. Haque, and C.K. Deb. 2023. TrIncNet: a lightweight vision transformer network for identification of plant diseases. *Frontiers in Plant Science* 14, (2023). https://doi.org/10.3389/fpls.2023.1221557

[15] Andrew J., Jennifer Eunice, Daniela Elena Popescu, M. Kalpana Chowdary, and Jude Hemanth. 2022. Deep Learning-Based Leaf Disease Detection in Crops Using Images for Agricultural Applications. *Agronomy* 12, 10 (October 2022), 2395. https://doi.org/10.3390/agronomy12102395

[16] H. Alshammari, K. Gasmi, I.B. Ltaifa, M. Krichen, L.B. Ammar, and M.A. Mahmood. 2022. Olive Disease Classification Based on Vision Transformer and CNN Models. *Computational Intelligence and Neuroscience* 2022, (2022). https://doi.org/10.1155/2022/3998193

[17] G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang. 2023. PMVT: a lightweight vision transformer for plant disease identification on mobile devices. *Frontiers in Plant Science* 14, (2023). https://doi.org/10.3389/fpls.2023.1256773