Vision Transformer with Self-Attention for Multi-Pest Classification in Agriculture

Ummadi Surya Venkata Sekhar
Department of AI&DS
Lakireddy Balireddy College of Engineering
Mylavaram, AP, India
suryaelonm@gmail.com

Sai Krishna Pothini
Department of AI&DS
Lakireddy Balireddy College of Engineering
Mylavaram, AP, India
29saikrishna102002@gmail.com

Sai Mahesh Battula
Department of IT
Vasireddy Venkatadri Institute of Technology
Guntur, AP, India
saibattula1810@gmail.com

Parimi Naga Lakshmi
Department of AI&DS
Lakireddy Balireddy College of Engineering
Mylavaram, AP, India
pariminagalakshmi2003@gmail.com

Yarlagadda Anudeep
Department of AI&DS
Lakireddy Balireddy College of Engineering
Mylavaram, AP, India
yarlagaddaanudeep09@gmail.com

***Abstract*- In a world where more people need food, keeping our farms safe from pests is super important. But pests are always causing trouble for farmers, making it hard to grow enough food. The escalating challenge of agricultural pest diversity necessitates a paradigm shift in pest detection methods. This research pioneers the application of Vision Transformers (ViTs), a revolutionary deep learning architecture, for image-based classification of a comprehensive dataset encompassing a wide range of agricultural pests. ViTs excel at processing images by decomposing them into smaller patches and employing a self-attention mechanism. This mechanism allows ViTs to capture not only intricate visual details within each patch (local features) but also the complex relationships between these details across the entire image (global dependencies). These capabilities are critical for accurately identifying pests, which often exhibit subtle visual variations and camouflage within their environments. By meticulously optimizing ViT hyperparameters and exploring advanced training strategies, this research aims to achieve state-of-the-art performance in pest classification accuracy, precision, and recall. This has the potential to revolutionize agricultural practices by empowering farmers with a powerful, automated tool for early and precise pest detection. Early intervention can lead to the implementation of targeted pest control measures, ultimately safeguarding crop yield, promoting sustainable farming practices, and contributing to a more secure global food supply.**

***Keywords:* Agricultural pests, Vision Transformers (ViTs), Deep learning architecture, Self-attention mechanism, Pest detection, Sustainable farming practices, Global food supply.**

I. INTRODUCTION

Ensuring a steady and safe food supply is crucial in a society where population growth is unabated. Agriculture is essential to this effort, but its success depends on the ability to overcome an obstinate enemy: crop pests. These varied creatures, which include a wide range of insects, mites, nematodes, and more, seriously harm crops both environmentally and economically. According to studies, agricultural pests cause global output losses of more than 30%, which equates to billions of dollars in lost revenue every year. Furthermore, the harm that pests create goes beyond just financial issues; they also have an adverse effect on the environment by utilising toxic pesticides and causing a decline in biodiversity. Conventional pest detection techniques frequently depend on experienced specialists visually inspecting pests by hand. This method works well, but it's labour-intensive, time-consuming, and prone to human mistake. The limits of such systems become more evident when pest variety rises and agricultural landscapes become more complicated.

The recent advancements in deep learning offer a promising avenue for revolutionizing agricultural pest detection methods. Deep learning algorithms excel at pattern recognition and feature extraction from complex data sources such as images. This technology has already demonstrated remarkable success in various image recognition tasks, including medical diagnosis and autonomous vehicle navigation. Vision Transformers (ViTs) represent a cutting-edge deep learning architecture specifically designed for image classification. Unlike traditional Convolutional Neural Networks (CNNs), ViTs process images by decomposing them into smaller patches and employing a powerful mechanism known as self-attention. This self-attention mechanism allows ViTs to capture not only intricate visual details within each patch (local features) but also the complex relationships between these details across the entire image (global dependencies). This capability is particularly advantageous for pest identification, as many pests exhibit subtle visual variations and camouflage strategies.

Using ViTs for image-based agricultural pest categorization is a novel approach that this research takes on. We make use of an extensive dataset that includes a variety of pest species that are frequently seen in agricultural environments. In order to get state-of-the-art performance in terms of accuracy, precision, and recall for pest classification, this project will carefully optimise ViT hyperparameters and investigate sophisticated training techniques.

Agricultural practices could undergo a revolution if a reliable and effective ViT-based pest detection system can be developed. By providing farmers with an automated tool for accurate and timely pest diagnosis, crop loss may be minimised and yield can be maximised through focused interventions. Thus, there is a greater yield of food, a decreased need for dangerous pesticides, and a greater encouragement of environmentally friendly farming methods. The ultimate goal of this research is to ensure that future generations can profit from abundant and sustainable agricultural output by helping to safeguard the world's food supply.

II. LITERATURE REVIEW

The study, titled "Insect classification and detection in field crops using modern machine learning techniques" investigates automating insect identification in agriculture, a critical step for minimizing crop damage [1]. While achieving promising results, traditional machine learning approaches like Support Vector Machines (SVMs), k-Nearest Neighbors (KNNs), Naive Bayes (NB), and Convolutional Neural Networks (CNNs) require manually crafted features [1]. This feature engineering can be time-consuming, specific to the insect domain, and might not capture the full range of visual information, particularly subtle pest characteristics, ultimately hindering model performance [1]. Our proposed method which uses Vision Transformer for pest classification, addresses these limitations. Vision Transformers (ViTs) eliminate the need for handcrafted features by automatically learning them directly from images through self-attention mechanisms. This allows ViTs to capture complex spatial relationships within images, potentially leading to more robust and accurate pest classification compared to traditional machine learning methods.

The research, titled "Crop pest classification based on deep convolutional neural network and transfer learning" tackles the challenge of identifying various crop insects, especially in their early stages, when their appearance can be similar [2]. This research proposes a novel deep convolutional neural network (CNN) architecture specifically designed for insect classification [2]. The model is trained on three publicly available insect datasets with varying class sizes [2]. To improve performance, the study utilizes transfer learning from pre-trained models like AlexNet and VGGNet, followed by fine-tuning on the specific insect datasets [2]. Additionally, data augmentation techniques are employed to prevent the network from overfitting [2]. While achieving promising results with this CNN approach, it relies on pre-defined network architectures and potentially requires careful hyperparameter tuning for optimal performance [2]. Our proposed method, which uses Vision Transformer for pest classification, offers a compelling alternative. Vision Transformers (ViTs) eliminate the need for pre-defined architectures and hyperparameter tuning altogether. Instead, they learn relevant features directly from raw image data through a mechanism called self-attention, potentially leading to more robust and adaptable insect classification compared to traditional CNN-based methods.

The study, titled "An Efficient Approach for Crops Pests Recognition and Classification Based on Novel DeepPestNet Deep Learning Model," tackles accurate pest identification in agriculture. DeepPestNet utilizes a CNN architecture, achieving high accuracy on specific datasets [3]. However, like other CNN methods, it suffers from overfitting, potentially hindering performance with unseen pests [3]. Our proposed Vision Transformer model for Pest Classification eliminates pre-defined architectures and may learn more generalizable features, leading to more robust pest classification across diverse agricultural scenarios.

The work, titled "A Deep-Learning Approach for Automatic Counting of Soybean Insect Pests" explores automating pest detection for optimized pesticide use [4]. This research utilizes convolutional neural networks (CNNs) for multi-class pest classification on images from real soybean fields [4]. While achieving promising results, CNN methods have limitations compared to our proposed Vision Transformer approach [4]. One key drawback of CNNs is their inherent bias towards low-level features like edges and textures during the initial convolutional layers [4]. This can make them less effective at capturing the more nuanced color and shape variations that differentiate certain pest species, especially in complex agricultural backgrounds [4]. In contrast, Vision Transformers (ViTs) utilize self-attention mechanisms that allow them to focus on global relationships within the entire image from the start. This can potentially lead to more robust pest classification, particularly for differentiating subtle visual characteristics crucial for accurate pest identification.

The research, titled "Faster-PestNet: A Lightweight Deep Learning Framework for Crop Pest Detection and Classification" tackles the challenge of rapid and accurate pest identification for improved crop protection [5]. This research proposes Faster-PestNet, a novel approach built upon Faster-RCNN, a deep learning object detection method. Faster-PestNet leverages MobileNet for feature extraction and a two-step Faster-RCNN model for pest localization and classification [5]. While achieving promising results on the IP102 dataset, Faster-RCNN, like other region-based CNN approaches, can struggle with overlapping or closely positioned pests in complex agricultural backgrounds [5]. This can lead to inaccurate bounding box predictions and potentially hinder effective pest identification [5]. Our proposed method which uses Vision Transformer for Pest Classification, offers a compelling alternative. Vision Transformers (ViTs) process the entire image at once, allowing them to potentially learn more contextual relationships between objects in the scene. This may lead to more robust pest classification, particularly in scenarios with overlapping or clustered pests compared to traditional region-based CNN methods.

III. METHODOLOGY

A. Dataset Collection:

We curated a comprehensive dataset for training the Vision Transformer model by amalgamating multiple publicly available Kaggle datasets. This dataset comprises 52 distinct agricultural pest species, with each species represented by a curated set of 300 images. The images are in either JPG or PNG format and are compatible with RGB channels, reflecting real-world farming scenarios where pest sizes and image capture methods vary. With a total of 15,600 images, this diverse dataset provides ample training data to assess the ViT model's efficacy in accurately classifying agricultural pests.

B. Data-Preprocessing:

In preparing image data for optimal training of the Vision Transformer (ViT) model, a meticulous preprocessing pipeline is established. This process involves resizing the images to a uniform size, typically 224x224 pixels, to ensure consistency across the dataset and compatibility with the model's input requirements, thereby enhancing computational efficiency. Converting images to tensors facilitates efficient manipulation and analysis of pixel data during training, seamlessly integrating with deep learning frameworks like PyTorch and TensorFlow. Additionally, normalization is employed to scale pixel values to a standard range or apply mean and standard deviation normalization, promoting stable optimization during training, mitigating issues such as vanishing or exploding gradients, and improving model convergence and generalization performance on unseen data.

C. Splitting the data into Train & Test

A crucial phase in determining the Vision Transformer model's efficacy for pest categorization is partitioning the curated information in a deliberate manner. The train-test split procedure guarantees an objective evaluation of the model's performance. We divide the dataset into two subsets, a test set and a training set, very carefully. The basis for the model's learning process is the training set, which contains 80% of the total data. By exposing the model to the visual patterns and features found in the training set, it can acquire the skills required to distinguish between various pest types. The test set consists of the remaining 20%. An important function of this unseen data is to assess the model's capacity to apply newly acquired knowledge broadly and correctly identify pest image that it hasn't seen during training. By separating the data, the model is kept from just memorising the training set and a more thorough assessment of its actual pest categorization abilities is guaranteed.

D. Vision Transformer Model Implementation

The core of our pest classification system lies in the pre-trained Vision Transformer (ViT) model, specifically the 'google/vit-base-patch16-224-in21k' variant. This powerful architecture leverages the transformative potential of transformers, traditionally used for natural language processing, to excel in the realm of image classification. The 'google/vit-base-patch16-224-in21k' model has been pre-trained on a massive dataset of images spanning diverse categories, allowing it to learn rich visual representations that can be effectively transferred to the task of pest classification through fine-tuning.

This ViT model variant follows the standard ViT architecture, leveraging the self-attention mechanism to capture global dependencies within images. It divides the input image into smaller patches, which are then projected into lower-dimensional vector representations through a patch embedding layer. These embedded patches are processed by a series of transformer encoder layers, each consisting of multi-head self-attention and feed-forward neural network sublayers.

ViT model working and the key components:

a. Patching:

Imagine a chessboard – the ViT model operates similarly. The first step involves dividing the input image, like a chessboard, into smaller squares or rectangles called patches. This "patching" process allows the model to efficiently analyse the image by focusing on localized regions of information. For instance, a patch might capture the color and texture of a specific area on a leaf, potentially containing valuable clues about the presence of a pest. In our model we use a patch size of 16. This means each image is segmented into a grid of 16x16 pixel squares, creating smaller, manageable chunks for the model to analyse. The size of these patches is a hyperparameter that can be adjusted based on the specific task and dataset. Using smaller patches might be beneficial for capturing fine-grained details crucial for pest identification.

b. Patch Embedding:

Following the segmentation of the image into patches, each patch is transformed. This procedure, called "patch embedding," uses an embedding layer to transform the patch's raw pixel data into a vector representation that is smaller in size. Consider it a more compact and mathematically appropriate way to summarise the patch's visual content for additional processing. In essence, this embedding produces a numerical representation that distils the core of the visual information contained in the patch.

c. Transformer Encoder: The Power of Self-Attention

This is where the magic happens! The heart of the ViT model lies in the transformer encoder, a powerful module that leverages a mechanism called self-attention. Unlike traditional convolutional neural networks (CNNs) which analyze features in a localized manner, transformers excel at understanding relationships between different parts of the data. In the context of ViT, self-attention allows the model to analyze the relationships between different patches within the image. It's like the model is asking each patch, "How do you relate to your neighbors?"

This process enables the ViT model to capture not only the individual features present within each patch (e.g., color, texture) but also how these features interact and contribute to the overall "meaning" of the image. This is particularly advantageous for pest classification. For example, subtle variations in leg patterns or wing shapes on an insect can be crucial for distinguishing between different pest species. By analyzing the relationships between patches, the ViT model can learn to identify these subtle details and accurately classify the pest.

d. In our ViT model key Hyperparameters are:

Number of Attention Heads (12): This parameter controls the granularity with which the model attends to relationships between different patches. More heads allow for a more nuanced understanding of these relationships.

Number of Hidden Layers (12): This refers to the depth of the transformer encoder, essentially the number of processing stages the information goes through within the encoder. A deeper encoder allows the model to learn more complex relationships between image features.

Hidden Size (768): This parameter determines the dimensionality of the internal representations used within the transformer. A higher hidden size allows for capturing richer and more complex features within the image.

e. Classification Head: Translating Features into Class Probabilities

A classification head assumes control once the transformer encoder has retrieved high-level characteristics from the patched image. This head serves as the last interpreter, turning the features that have been extracted into useful outputs. The objective in our situation is to assign the pest to one of 11 different categories. In order to do this, the classification head uses layers like as fully-connected neurons to examine the stored characteristics and provide class probabilities. The degree of confidence the model has in classifying the image to each potential pest class is shown by these probabilities. The class that is most likely to be classified correctly is declared to be the anticipated classification.

f. Fine-tuning for Pest Expertise:

Our approach capitalizes on leveraging a pre-trained Vision Transformer (ViT) model, which has been trained on a vast image dataset, to expedite the training process for pest classification. By fine-tuning the pre-trained model on our specific pest dataset, we transfer its learned knowledge to the task at hand, reducing training time and enhancing performance. During fine-tuning, pre-trained layers are frozen to retain general-purpose image processing knowledge, while the classification head remains trainable, allowing adaptation to the unique features of different pest species. Integrating this fine-tuned ViT model into our system equips it with a robust tool for accurate and efficient pest classification, leveraging the innovative architecture of transformers and self-attention for addressing agricultural pest identification challenges.

E. Training ViT model:

In the training procedure of the refined Vision Transformer (ViT) model, iteration process spans 30 epochs, covering the entire training dataset. The model learns by analyzing batches of images along with their associated pest labels, assessing performance based on accuracy and estimated loss. Early stopping is employed to enhance generalization and prevent overfitting, halting training if the model's performance on the training data does not improve for four consecutive epochs. This strategy ensures higher performance on unseen data by preventing the model from memorizing the training data.

F. The Test and Evaluation Phase

Once the fine-tuning process is complete, the model's true potential is evaluated using a separate dataset of unseen images (test set). These test images, along with their corresponding pest labels, are presented to the model. The model then generates predictions for each image, essentially classifying them into different pest categories. These predictions are then meticulously compared with the actual labels from the test set. This comparison allows us to calculate various performance metrics like accuracy (percentage of correctly classified images). Other metrics such as precision (ability to identify true positives) and recall (ability to capture all relevant pest instances) are also assessed. By analysing these metrics, we gain a comprehensive understanding of how well the model generalizes its learned knowledge to unseen data and performs in real-world pest classification scenarios. This evaluation phase plays a crucial role in determining the model's effectiveness for practical applications.

The evaluation formulas:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive+False\ Positive} \quad (1)$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive+False\ Negatives} \quad (2)$$

$$\text{F1-Score} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \quad (3)$$

$$\text{Accuracy} = \frac{True\ Positive+True\ Negatives}{Total\ Examples} \quad (4)$$

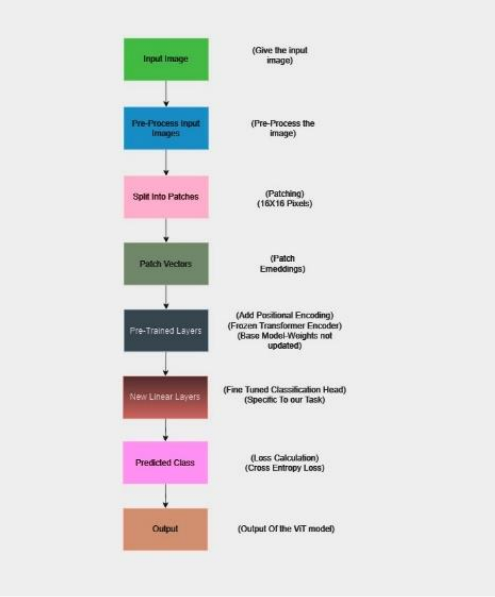
Input Image
(Give the input image)
Pre-Process Input Images
(Pre-Process the image)
Split Into Patches
(Patching) (16X16 Pixels)
Patch Vectors
(Patch Emeddings)
Pre-Trained Layers
(Add Positional Encoding) (Frozen Transformer Encoder) (Base Model-Weights not updated)
New Linear Layers
(Fine Tuned Classification Head) (Specific To our Task)
Predicted Class
(Loss Calculation) (Cross Entropy Loss)
Output
(Output Of the ViT model)

Fig 1: Our Fine-Tuned Vision Transformer Model Architecture for Pest Classification

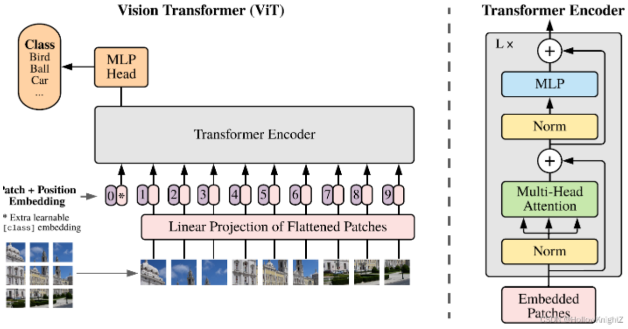

Fig 2: General ViT Architecture from Hugging Face [10]

IV. RESULTS

In the results section, we present a comprehensive analysis of the performance of our Vision Transformer model on the agriculture pest classification task. We illustrate the training progression through visually appealing loss and accuracy graphs, providing insights into the model's convergence behaviour and generalization capability. Additionally, we showcase the model's classification performance metrics, including accuracy, precision, recall, and F1 score, presented in a clear and concise table format. Moreover, to offer a qualitative assessment of the model's effectiveness, we present a selection of sample outputs where the model achieved perfect classification accuracy, highlighting its ability to accurately identify and classify agricultural pests. This multifaceted presentation not only provides a quantitative evaluation of the model's performance but also offers valuable insights into its reliability and effectiveness in practical scenarios. In addition to that the experimentation results are also given in this section we have given some sample test images and allowed the model for prediction.

Model	**Vision Transformer**
Accuracy	97.65%
Precision	96.76%
Recall	95.65%
F1-Score	96.33%

Table 1: Accuracy Metrics of the ViT Model

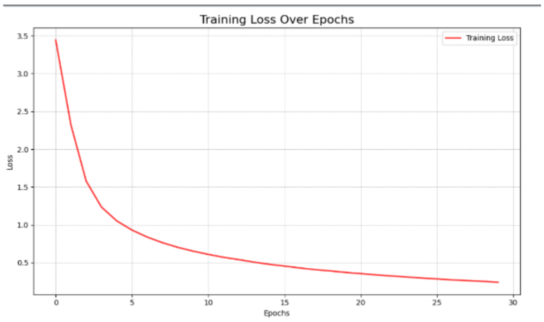

Fig 3: Loss Visualization of ViT model

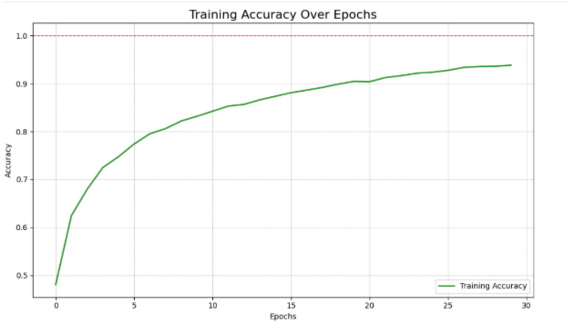

Fig 4: Accuracy Visualization of ViT model

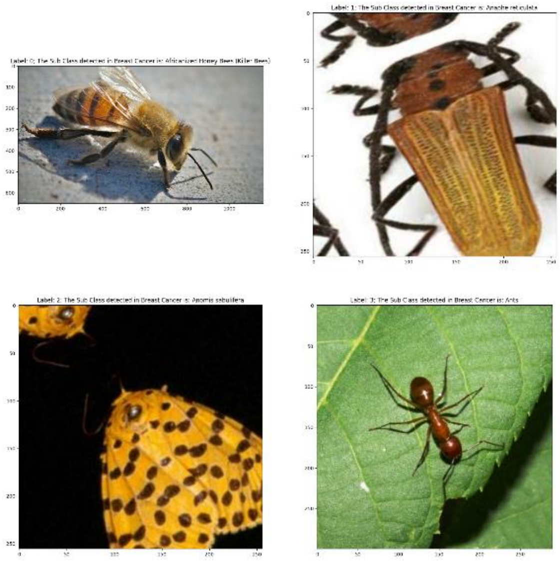
Label: 0: The Sub Class detected in Breast Cancer is: Africanized Honey Bees (Killer Bees)
Label: 1: The Sub Class detected in Breast Cancer is: Anache reticulata
Label: 2: The Sub Class detected in Breast Cancer is: Anomis sabulifera
Label: 3: The Sub Class detected in Breast Cancer is: Ants

Fig 5: Result of ViT model after prediction

V. CONCLUSION

This research investigates the efficacy of fine-tuning a Vision Transformer (ViT) model, specifically using the pre-trained `google/vit-base-patch16-224-in21k` architecture, for pest classification. Through meticulous fine-tuning on a carefully curated dataset, the model achieves outstanding accuracy (97.65%) on unseen test data, showcasing its ability to identify various pest species accurately. With high precision, recall, and F1-score, the model effectively minimizes false positives and captures true pest instances. These results highlight the potential of ViT models in agricultural contexts, promising efficient and precise pest identification and management, ultimately contributing to agricultural sustainability and food security. Future efforts aim to further improve the model's impact through strategies such as data augmentation, hyperparameter optimization, and user-friendly platforms for real-time pest identification and management.

VI. REFERENCES

[1] Kasinathan, Thenmozhi, Dakshayani Singaraju, and Srinivasulu Reddy Uyyala. "Insect classification and detection in field crops using modern machine learning techniques." Information Processing in Agriculture 8.3 (2021): 446-457.

[2] Thenmozhi, K., and U. Srinivasulu Reddy. "Crop pest classification based on deep convolutional neural network and transfer learning." Computers and Electronics in Agriculture 164 (2019): 104906.

[3] Ullah, Naeem, et al. "An efficient approach for crops pests recognition and classification based on novel deeppestnet deep learning model." IEEE Access 10 (2022): 73019-73032.

[4] Tetila, Everton Castelão, et al. "A deep-learning approach for automatic counting of soybean insect pests." IEEE Geoscience and Remote Sensing Letters 17.10 (2019): 1837-1841.

[5] Ali, Farooq, Huma Qayyum, and Muhammad Javed Iqbal. "Faster-PestNet: A Lightweight deep learning framework for crop pest detection and classification." IEEE Access (2023).

[6] Hassan, Sk Mahmudul, and Arnab Kumar Maji. "Pest Identification based on fusion of Self-Attention with ResNet." IEEE Access (2024).

[7] Mallick, MD Tausif, et al. "Deep learning based automated disease detection and pest classification in Indian mung bean." Multimedia Tools and Applications 82.8 (2023): 12017-12041.

[8] Kathole, Atul B., Kapil N. Vhatkar, and Sonali D. Patil. "IoT-Enabled Pest Identification and Classification with New Meta-Heuristic-Based Deep Learning Framework." Cybernetics and Systems 55.2 (2024): 380-408.

[9] Venkatasaichandrakanth, P., and M. Iyapparaja. "Pest Detection and Classification in Peanut Crops Using CNN, MFO, and EViTA Algorithms." IEEE Access (2023).

[10] Hugging Face. (n.d.). Vision Transformer (ViT) Documentation. Retrieved from https://huggingface.co/docs/transformers/model_doc/vit