

Crop Pest Recognition Using Image Processing

by

Pial Ghosh

19201069

Istiak Ahmed Alin

19201087

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

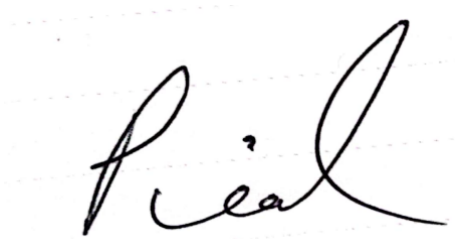
© 2024. Brac University
All rights reserved.

Declaration

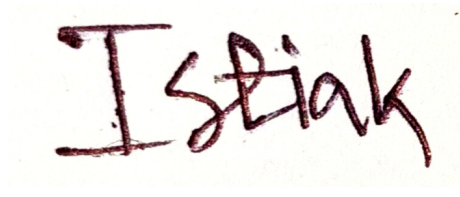
It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Pial Ghosh
19201069



Istiak Ahmed Alin
19201087

Approval

The thesis/project titled “Crop Pest Recognition Using Image Processing” submitted by

1. Pial Ghosh(19201069)
2. Istiak Ahmed Alin(19201087)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)



Md Tanzim Reza
Lecturer
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

One of the most vital aspects of a human's existence is food [15]. Each food contains several nutrients which help in growth and development of the human body. It also prevents our body from various diseases. Most of the food we consume comes from crops, trees and plants. Pests infestation is the biggest threat for the agriculture sector. It can cause various types of diseases in crops and reduce crop production. As a result, it is necessary to detect pests early and take necessary steps to stop the infestation. For decades, humans used traditional manual techniques to detect pests. However, this technique is very time consuming, laborious and less accurate. With the development of deep learning, pest detection has become easier than the traditional techniques. Although these deep learning techniques can achieve high accuracy, the techniques can be time consuming and need high computational power. These drawbacks might cause problems in real world application. The aim of this study is to propose an approach to apply ResNet50, MobileNetV2 and Vision Transformers architecture so that it consumes less time to train and needs less computational power. So, the farmer implements this model and detects pests quickly and takes necessary actions. This will increase the production of crops rapidly.

Keywords: Computational power; Human body; food; Pests; Crops; Infestation; Production; Development; Disease

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Table of Contents	1
1 Introduction	2
1.1 Problem Statement	2
1.2 Research objectives	2
2 Related Work	4
3 Methodology	13
3.1 Dataset	13
3.2 Data preprocessing and augmentation	14
3.3 Proposed model and explanation	15
3.4 Implementation	16
4 Results and Discussion	18
5 Conclusion	20
Bibliography	22

Chapter 1

Introduction

The population of the world is predicted to increase to 9.7 billion in 2050 and may peak at over 11 billion in 2100. This means that in order to feed the growing population, there has to be a sufficient supply of food while minimizing crop damage [4]. In agriculture, pests have always been a big issue for decades. Because it causes a lot of harm to the crops. So, pest monitoring has become very important. There are millions of species found in the world which is the biggest issue for pest control. As it is hard to identify an insect pest. In the past, people used to implement manual techniques to identify pests which are very time consuming, laborious and have a chance to cause errors to identify [1][2]. If somehow farmers can detect insects before attacking the crop field can help them to minimize the crop damage and increase crop production. Due to development of science and technology, computer and machine learning techniques have become very popular in the agriculture sector [3][4]. Machine learning can give accurate information about pest insects and saves a lot of time which was the biggest disadvantage of manual techniques [2].

1.1 Problem Statement

Pest attacks are one of the biggest issues in the agriculture sector. Pest attacks can cause various diseases to crops and lower the production [5]. While there are many deep learning techniques that exist for pest detection but we need more research to address the issue. There is a lack of diverse and large data to train pest detection models. As a result, many existing models rely on small datasets, hindering their ability to identify the insect pests properly.

1.2 Research objectives

Rice, sugarcane, cotton and jute are the important crops of Bangladesh. Every year many pests such as - Rice yellow stem borer, Dark headed stem borer, green leaf hopper, pink borer, jute white mite, spotted bollworm etc. attack the fields of the mentioned crops. As a result, the production of the crop is hampered and farmers have to face a huge loss. The main purpose of this paper is to enhance the pest recognition model in the agricultural sector to reduce pest infestation. So, crop production increases. The key points of our research are as follows :

- To collect images of harmful pests of rice, sugarcane, cotton and jute.

- To select a lightweight algorithm for real time processing of pest images.
- To implement state of the art algorithms in pest recognition.
- To compare the accuracy of state of the art algorithms with each other.

Chapter 2

Related Work

The Groundnut Vision Transformer (GNViT) model to detect and classify pests. Vision Transformers (ViTs) over traditional Convolutional Neural Networks (CNNs) is well articulated, laying a solid foundation for this paper [23]. The methodology section is comprehensive, detailing the dataset, the GNViT model, and the experimental setup. The use of the IP102 dataset which includes images of various pests affecting groundnut crops, provides a robust base for training and evaluation. The GNViT model uses a pre-trained ViT model for the pest classification. The authors have clearly explained the data preprocessing steps, data augmentation techniques and the model training process. The dataset has 4157 images categorized into four classes (Thrips, Aphids, Armyworms, and Wireworms). Techniques like resizing, contrast enhancement, and Error Level Analysis (ELA) ensure the dataset's integrity and improve model accuracy. Various transformations (rotation, flipping, color adjustments) are applied to increase the dataset's variability and prevent overfitting. The GNViT model adapts the ViT architecture, employing the Adam optimizer and CrossEntropy Loss function, with detailed pseudo-code provided. The results demonstrate the GNViT model's high accuracy (99.52%) and its superiority over other models such as LeNet, AlexNet, GoogleNet, and ResNet. The GNViT model's performance is evaluated using metrics like accuracy, precision, recall and F1-score, with notable improvements in training accuracy from 55.37% to 99.52%.

The paper [22] starts by highlighting the significance of tomato farming in the global agribusiness landscape and the crucial role of early disease detection in preventing yield losses. It discusses the limitations of traditional visual classification techniques and the potential of computer technologies, specifically machine learning and deep learning, to automate disease identification in tomato plants. The study uses two deep learning models, Inception V3 and Vision Transformer (ViT), to classify diseases in tomato leaves. The dataset is derived from the Plant Village dataset, which includes 10,010 images of tomato leaves across 10 different classes (nine disease classes and one healthy class). The methodology is detailed and systematic which involves data preprocessing, model training and evaluation. The use of transfer learning with the Inception V3 model and the self-attention mechanism in ViT is well-explained which provides a clear understanding of the model architectures and their implementation. The results indicate that the ViT outperforms Inception V3 in both accuracy and loss metrics. The ViT achieves a training accuracy of 97.37% and a validation accuracy of 95.76% and compared to Inception V3's training accuracy

of 89.24% and validation accuracy of 88.98%.

The paper [1] proposes an Enhanced Vision Transformer Architecture (EViTA) model which is designed for pest identification, segmentation and classification. The model leverages the capabilities of Vision Transformers (ViT) and incorporates a novel dual-layer transformer encoder to handle various segment sizes of pest images. The proposed methodology utilizes three pest datasets : Aphids (IP102 Dataset), Wireworm (IP102 Dataset) and Gram Caterpillar collected from publicly available repositories. The paper employs Moth Flame Optimization for feature extraction and StandardScaler for data normalization to improve the prediction accuracy. The model's performance is evaluated using various metrics like accuracy, precision, sensitivity, specificity, F1 score, Mean Absolute Error (MAE) and Mean Squared Error (MSE), providing a thorough assessment. The results are compared against several state-of-the-art CNN models, demonstrating the superiority of the proposed EViTA model in terms of prediction accuracy. In methodology, The process of segmenting images into tokens and the role of positional encoding should be elaborated on. More details on the MFO algorithm's implementation and how it improves feature selection are needed. The explanation of the EViTA model, including the role of the Extra Arrangement Segment (EAS) and the integration with the MFO which should be clearer.

The paper [21] addresses the critical issue of food security and crop yield by focusing on the timely and accurate identification of plant diseases. It proposes a novel approach combining Vision Transformer (ViT) for high identification accuracy and GPipe for enhanced running speed. The experiments utilize the PlantVillage dataset, and results show that ViT achieves a 93% recognition accuracy, with pipeline parallelism significantly reducing memory requirements. This research offers a low-cost and efficient plant disease identification tool beneficial for agriculture. The ViT model is chosen for its accuracy. It processes image data by dividing it into patches, mapping these to vectors, and using a Transformer encoder with self-attention mechanisms to capture image relationships. The MLP Head classifies the processed data. In terms of GPipe, This scalable training method partitions the model across multiple GPUs, allowing pipeline parallelism. This approach improves training efficiency and reduces GPU memory consumption. The study uses three sets of control experiments with different numbers of GPUs to test the impact of pipeline parallelism on model performance. Images from the PlantVillage dataset are preprocessed by resizing, normalizing, and converting pixel values to ensure uniform input for the models. ViT achieves a 93% accuracy on the color dataset, outperforming ResNet, which achieves 84%. ViT's performance drops on gray images but remains relatively high on segmented images. ResNet's performance is more stable across all datasets. Using GPipe significantly increases throughput and reduces memory usage per GPU, enhancing efficiency.

This paper[8] addresses the pressing need for precise identification and classification of plant diseases and insect pests, which is critical for the advancement of precision and smart agriculture. The proposed solution leverages the Vision Transformer (ViT) neural network, a model initially designed for natural language processing but adapted here for image recognition tasks. The paper employs the Vision Transformer,

a cutting-edge architecture known for its superior performance in image recognition tasks, particularly when dealing with large datasets. This approach is a significant advancement over traditional convolutional neural networks (CNNs) like GoogleNet and EfficientNetV2. The use of various data enhancement methods, including Histogram Equalization, Laplacian, Gamma Transformation, CLAHE, Retinex-SSR, and Retinex-MSR, is a notable strength. These techniques likely help in mitigating overfitting and improving the model's robustness and generalization capabilities. The reported test recognition accuracy of 96.71% on the Plant_Village dataset is impressive, outperforming established models such as GoogleNet and EfficientNetV2.

The paper [15] proposes using hybrid Vision Transformer (ViT) models combined with Convolutional Neural Networks (CNNs) for the early identification of plant diseases. The multispectral dataset was captured using various Kolari vision lenses covering both visible and Near-Infrared(NIR) ranges. The hybrid models achieved notable performance, with the K850 lens dataset reaching 92.83% training accuracy and 88.86% test accuracy. The approach aims to enhance precision agriculture and crop management. In the methodology section describes the hybrid ViT models, which combine CNNs for feature extraction and transformers for classification. Two specific models (ViT r26 s32 and ViT r50 l32) are discussed, with differences in their architecture and layers. The section also explains the data augmentation techniques used (light and medium) and their impact on model performance. Transfer learning and other training strategies, including early stopping and the use of the Adam optimizer, are well-documented. The dataset was split into 80% training and 20% testing. Accuracy was used as a metric for evaluating model performance. The experimental setup is logically structured, comparing different lenses, augmentation strategies, and model sizes. The paper successfully demonstrates the viability of hybrid ViT models for early plant disease identification and their potential benefits for precision agriculture.

Pest detection is critical for maintaining crop quality and productivity, but traditional methods are laborious and limited in accuracy. This study introduces Faster-PestNet, an improved Faster-RCNN model that uses MobileNet as its backbone, optimized for identifying various pest categories. The model demonstrates a huge advancement with an accuracy of 82.43% on the IP102 dataset, which has 102 insect classes. Additionally, Faster-PestNet shows high accuracy achieving around 95% precision, recall, F1-score, and accuracy. By using MobileNet's efficient depth-wise separable convolutions and Faster-RCNN's robust two-step detection process, the model effectively addresses traditional pest detection limitations, offering a powerful tool for real-time applications in agriculture to reduce pesticide usage and enhance crop management [13].

This paper explains about the Plant Based MobileViT (PMVT). It is a deep learning model for real-time plant disease detection on mobile devices. PMVT aims to balance high accuracy with computational efficiency, addressing the challenges posed by limited computing resources and the diversity of plant diseases. PMVT outperforms as a lightweight and heavyweight models on multiple agricultural datasets. Achieved notable accuracy improvements: 93.6% on wheat, 85.4% on coffee, and 93.1% on rice datasets, surpassing the performance of models like MobileNetV3 and SqueezeNet.

PMVT models (XXS, XS, S) consistently achieved higher accuracy compared to other lightweight and heavyweight models across different datasets. Confusion matrices, precision, and recall metrics show the model’s strengths and areas for improvement, particularly in identifying specific diseases like red spider mite on coffee leaves [19].

The paper [17] begins with a clear articulation of the problem: crop pests and diseases significantly impact agricultural productivity and quality, threatening both macroeconomic stability and sustainable development. Traditional manual recognition methods are described as inefficient and subjective, necessitating the use of advanced pattern recognition and deep learning methods. The study introduces an improved Vision Transformer (ViT) method for recognizing crop pests and diseases, using block partitioning to select feature-rich regions and using the self-attention mechanism of transformers to identify non-obvious lesion areas. The methodology section is detailed, outlining the technical route involving the classifier training and testing modules. The process begins with converting RGB images to a uniform size, followed by partitioning the images into non-overlapping patches to extract spatial features. The Transformer then mines the relationships between these features, and the final classification is performed using a softmax classifier. The dataset used is from the Beijing Big Data Skills Competition, comprising 4562 samples of crop leaf images with various diseases. The training set includes 3581 samples, while the test set has 981 samples. The paper explains the metrics used for evaluation, specifically precision and recall, derived from the confusion matrix. The results are presented through a confusion matrix, which shows the model’s classification accuracy and recall rates for different categories of crop diseases. The model demonstrates high accuracy overall, but with some variation across different disease categories.

The paper [20] presents a novel model. ROI-ViT which was made to improve pest identification in images with complex backgrounds and small sizes. In this paper, the challenge in pest identification lies in accurately extracting regions of interest (ROIs) from the background which often have similar textures and colors as the pests. The proposed model aims to address this by generating and updating ROIs using multiscale cross-attention fusion and enhancing robustness to background complexity and scale variations. The model incorporates ROI generators using soft segmentation and class activation maps (CAM) to produce initial ROI maps. These maps guide the model to focus on specific areas. The proposed ROI-ViT model was run on three public pest image datasets: IP102, D0, and SauTeg and IP102(CBSS) which has complex background images and small sizes. It achieved superior accuracy compared to several SOTA models. Such as EfficientNet, Swin-ViT and others. ROI-ViT maintained high accuracy on this challenging dataset whereas the accuracy of other models dropped significantly. The CAM-based ROI generator outperformed the segmentation-based generator due to its richer pixel information which provided better guidance for the model. ROI-ViT’s accuracy decreased only slightly on the complex background and small size dataset and demonstrating its robustness.

The study [18] addresses the challenge of early plant disease detection using an improved Vision Transformer network, TrIncNet. It replaces the computation expensive MLP module of traditional ViTs with a more efficient Inception module

which improves feature extraction and reduces the number of trainable parameters. The model’s architecture has skip connections to eradicate the vanishing gradient problem. TrIncNet’s performance was validated using the PlantVillage and Maize disease datasets, where it outperformed other CNN architectures and ViTs, achieving higher accuracy and efficiency. As a result, TrIncNet is suitable for integration with IoT devices for real-time plant disease detection. The TrIncNet model achieved the highest validation accuracy (97.0%) and lowest validation loss (0.035) on the Maize dataset. It also demonstrated superior results in accuracy, precision, recall, and F1-score. TrIncNet required significantly fewer trainable weight parameters (6.95 million) compared to the ViT model (7.12 million). On the PlantVillage dataset, TrIncNet attained the highest validation accuracy (99.95%) and lowest validation loss (0.02). It outperformed other models with 99.93% accuracy, 99.92% precision, 99.91% recall, and 99.91% F1-score.

The paper [10] focuses on the PlantCLEF2022 challenge, where the goal is to identify plant species from a large-scale dataset comprising millions of images and 80,000 classes. Given the limited number of images per class (average of 36), this task is treated as few-shot image classification. Instead of using the traditional convolutional neural networks (CNNs), the authors employ a self-supervised Vision Transformer (ViT) model. This approach secured first place in the challenge, achieving a Macro Averaged Mean Reciprocal Rank (MA-MRR) of 0.62692. The self-supervised ViT offers advantages over CNNs, including no inductive biases and better feature extraction for downstream tasks. The PlantCLEF2022 challenge provides a platform to address these issues using a dataset of 80,000 classes and millions of images. The PlantCLEF2022 dataset is divided into training and testing sets. The training dataset consists of web images (1.1 million images, 57,000 classes) and trusted images (2,885,052 images, 80,000 classes). The trusted dataset is preferred for training due to higher annotation quality. The testing dataset includes 55,306 images from 26,868 observations. The evaluation metric used is the MA-MRR, which accounts for multiple chances to recognize plant species, making it suitable for observation-level classification. The results demonstrate the effectiveness of the self-supervised ViT approach, which achieved first place in the challenge. The MA-MRR score of 0.62692 was significantly higher than the second (0.60792) and third place (0.51092). Extending the training by twenty epochs further improved the score to 0.64079. The pretrained model also showed promising results in plant disease recognition tasks on four public datasets, outperforming the ImageNet pretrained model.

The paper [16] effectively shows the core objectives and methodologies of the research. The focus is on plant disease classification and pest recognition using deep learning models, particularly through the lens of transfer learning (TL) with EfficientNet-V2. The paper highlights the limitations of traditional TL approaches, introduces progressive learning in EfficientNet-V2, and presents a comparative study with InceptionV3 and Vision Transformer (ViT) models. The abstract clearly outlines the datasets used and provides a succinct overview of the results, emphasizing the superior performance of EfficientNet-V2. A well-known CNN model, it is noted for its feature extraction capabilities and batch normalization, which helps in stabilizing gradients and improving convergence. A non-CNN model that uses a transformer architecture, breaking images into patches and employing self-attention

mechanisms. The section mentions the challenges of ViT with smaller datasets and longer training times. Highlighted for its progressive learning and adaptive regularization, which helps in efficiently handling datasets of varying sizes. The model's architecture and training strategies are briefly discussed. InceptionV3, The use of Keras Image Data Generator for augmentation and the specifics of the training setup are well described. ViT, The experimental setup, including patch size, optimizer, and learning rates, is outlined clearly. EfficientNetV2, The progressive learning strategy and adaptive regularization parameters are explained, along with the specific configurations for training.

The paper [6] addresses the challenges in disease detection on olive leaves using deep learning techniques, highlighting the importance of accurate and timely diagnosis in agriculture. The authors emphasize the historical significance of olive cultivation and the diversity of diseases affecting olive trees, which complicates disease detection efforts. They propose a novel hybrid deep learning model combining Convolutional Neural Networks (CNN) and Vision Transformer (ViT) models to enhance classification accuracy for olive leaf diseases. Data Preprocessing and Augmentation, The authors used median noise filtering and various data augmentation techniques to enhance the dataset. These techniques help improve model generalization and performance, especially in unbalanced datasets. Hybrid Feature Extraction, The hybrid model leverages CNN's spatial feature extraction capabilities and ViT's self-attention mechanisms. CNN models, such as AlexNet and VGG, were used alongside ViT models to test various configurations and determine the best-performing architecture. Classification Techniques: The study tested both binary and multiclass classification approaches. The models included AlexNet, VGG-16, VGG-19, and ViT, which were compared. The dataset consisted of 3,400 images categorized into healthy leaves, leaves infected with *Aculus olearius*, and leaves with olive peacock spot. The dataset was split into training (80%) and testing (20%) sets, with performance evaluated using accuracy, precision, recall, and F-measure. The hybrid model achieved high accuracy rates, with 97% for binary classification and 96% for multiclass classification. These results outperformed other individual models tested in the study. The study tested various optimization algorithms (Adam, AdaGrad, RMSProp) to minimize the loss function and improve model performance. Adam optimization yielded the best results, enhancing both binary and multiclass classification accuracy.

The paper [24] focuses on deep learning techniques to improve the detection of diseases and pests in crops, which is crucial for enhancing crop yield and quality. Two methodologies are discussed here, one for identifying diseases in cotton leaves using three images VGG16, ResNet50 and Inception V3 and another for detecting pests using VGG16 and Inception V3. The research achieves effectiveness of deep learning in this domain. In the methodology, provides a comprehensive overview of the datasets used and the preprocessing techniques employed. The use of a custom dataset for cotton leaf disease detection and a publicly available pest dataset from Kaggle ensures a robust evaluation framework. Data augmentation techniques like flipping are used to mitigate the issue of limited data. The choice of transfer learning models (VGG16, ResNet50, and Inception V3) is well-justified and given their proven effectiveness in image classification tasks. The results demonstrate the high

accuracy achieved by the models with VGG16 outperforming others in both disease and pest detection tasks. The accuracy rates of 99.58% for disease detection and 99.78% for pest detection using VGG16 are particularly notable.

In the novel hybrid model called "PlantXViT" that combines CNN and ViT for plant disease detection. The model is designed to increase the local feature extraction capabilities of CNNs and the global feature extraction capabilities of Vision Transformers to improve the accuracy. This model is very efficient in plant disease identification. PlantXViT is a lightweight model which is suitable for IoT-based smart agriculture services. The model trained on five publicly available datasets where it showed impressive performance compared to state-of-the-art methods. The combination of CNN and ViT allows efficient extraction of both local and global features, which is important for accurate plant disease detection. With only 850,500 trainable parameters, PlantXViT is very much optimized to use in IoT devices which makes it practical for real-world applications in agriculture sector. The model achieves impressive accuracy getting 93.55% on the Apple dataset, 92.59% on the Maize dataset, and 98.33% on the Rice dataset. The use of gradient-weighted class activation maps (Grad-CAM) and Local Interpretable Model Agnostic Explanation (LIME) enhances the model's predictions, which is important to trust AI decisions [9].

This paper [4] explains about a method for detecting plant leaf diseases by combining a hybrid segmentation algorithm and CNN. This method integrates hue, saturation, intensity (HSI) and LAB color spaces for effective segmentation of disease symptoms from complex backgrounds. This approach leverages the compatibility of these color spaces with human vision to enhance segmentation accuracy. Following segmentation, a modified CNN inspired by AlexNet. But optimized for computational efficiency, classifies the segmented images. The method was tested on a diverse dataset of around 1000 images which achieved a validation accuracy of 91.33% which is 15.51% higher than conventional methods. This significant improvement underscores the method's potential in practical agricultural applications and contributes to better management and prevention of plant diseases and ultimately supporting food security. Future research could expand the dataset and refine the algorithm for real-time application and include more comparative analyses with other advanced methods.

This thesis [12] addresses the global agricultural challenge of crop pest detection by proposing an advanced deep learning model. The proposed ResNet-50-PCSA model combines a parallel attention mechanism with residual blocks to enhance accuracy and real-time performance. Extensive experiments showcase its effectiveness, achieving a remarkable accuracy of 98.17% on crop pest image and demonstrating adaptability to rice leaf diseases. ResNet-50 is chosen for feature extraction and data augmentation techniques are employed for robust training. The ResNet-50-PCSA model utilizes a bottleneck-PCSA model within ResNet-50 to refine features which a carefully curated dataset comprises 10 common crop pests and contributes to the model and its accuracy is 98.17%. Comparative experiments with SENet and CBAM highlight the efficacy of the parallel attention mechanism. The model's adaptability is tested on a public dataset of rice leaf diseases, achieving a high ac-

curacy of 99.35%. Comparative experiments with other CNN models showcase the ResNet-50-PCSA model’s versatility beyond pest recognition.

The paper [5] investigates advanced image recognition technology for the identification of insect pests and crop diseases which focuses on three-dimensional image recognition and image quality classification. The approach of this paper involves enhanced three-dimensional panoramic image synthesis. This method combines object detection, coordinate ascending and inverse mapping to render a pre-prepared three-dimensional model onto estimated positions. The synthesis is facilitated by DCNN which design to address limitations of prior techniques. Acknowledging data imbalance in crop disease and insect pest images. The author introduce a quality classification model and train the DCNN-G model using Google data analysis to incorporating strategies like data enhancement and transfer learning. YOLO-V4 is also employed for testing and demonstrating effective image classification.

Likewise on other papers, Insect pest recognition addresses crucial points in agriculture and ecology in terms of variations in insect appearance. The authors propose a feature fusion network that combines ResNet and Vision Transformer and Swin Transformer backbones. They utilize Grad-CAM for localization and introduce an attention-selection mechanism to reconstruct attention areas by integrating important regions from different models. The study reflects conduct on the IP102 dataset, demonstrating superior performance compared to advanced CNN models. To make Grad-CAM applicable to attention-based models, the study reshapes their output tensors. An attention-selection mechanism which is based on the Image Fusion Convolutional Neural Network (IFCNN) and it is introduced to reconstruct attention areas and synthesize features from different models. The approach is applied to the challenging IP102 dataset which demonstrates superior classification performance compared to single-attention features and other CNN-based models [14].

Crop pest recognition in the field introduces a pioneering method, MCapsNet which is aiming to overcome the challenges posed by traditional methods and enhance crop protection. By incorporating a modified Capsule Network and an attention mechanism, the proposed approach demonstrates superior performance compared to traditional convolutional neural networks (CNNs) and CapsNet. The attention mechanism proves effective in capturing crucial classification features and MCapsNet exhibits success in classifying diverse insect types in field crops. By applying Modified Capsule Network (MCapsNet) with an attention mechanism for crop pest recognition in field images. MCapsNet leverages deep learning capabilities, specifically Capsule Networks to automatically extract invariant features from diverse pest images which points out mechanism enhances feature extraction by focusing on relevant contextual relationships and the LeakyReLU activation function accelerates model convergence. Through experiments on a dataset comprising images of common crop pests, MCapsNet demonstrates superior accuracy and recall compared to other CNN models. The key contributions lie in the effective integration of Capsule Networks, attention mechanisms and LeakyReLU activation for robust crop pest recognition which offers a promising solution for practical implementation in agriculture to enhance crop protection [11].

Unlike other papers, this paper explains crop pest and disease detection in the agriculture sector, specially sourced from local farms in Ghana. The dataset includes both raw and augmented images with a total of 22 classes covering Cashew, Cassava, Maize and Tomato crops. This paper outlines the experimental design, emphasizing the image data acquisition process and it is annotated by expert plant virologists and the subsequent verification of labels. The dataset images vary in size, providing diverse conditions for training computer vision algorithm [2].

Chapter 3

Methodology

3.1 Dataset

All pests have to go through several stages throughout their life cycle , based on the species and category of pest. As a result , it is very difficult to capture images of pests [7]. So , we used a secondary dataset IP102 which is one of the biggest insect pest dataset that is publicly available for academic usage [23]. This dataset

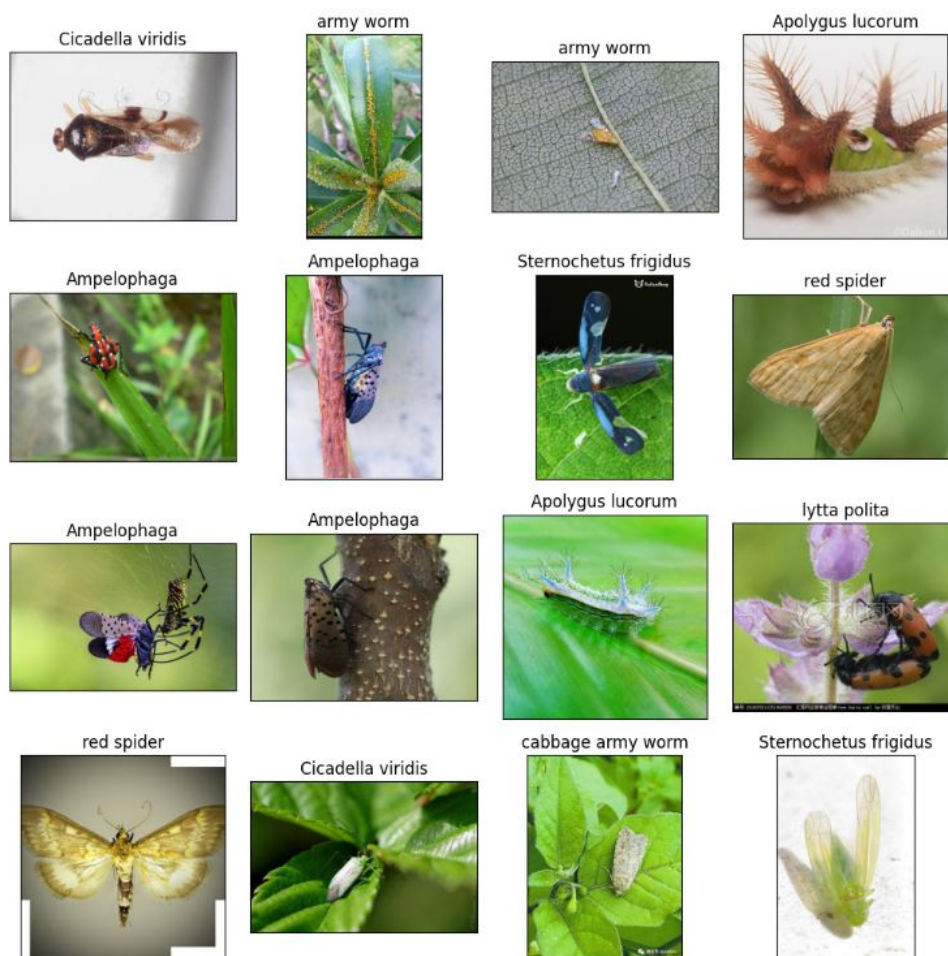


Figure 3.1: Random sample image from IP102 dataset

contains more than 75,000 images of 102 categories of insects. We used 10 largest

classes from this dataset for insect pest detection , the distribution is shown in fig.02. 17926 images were used for training the model. The sample image of the dataset is in fig.01.

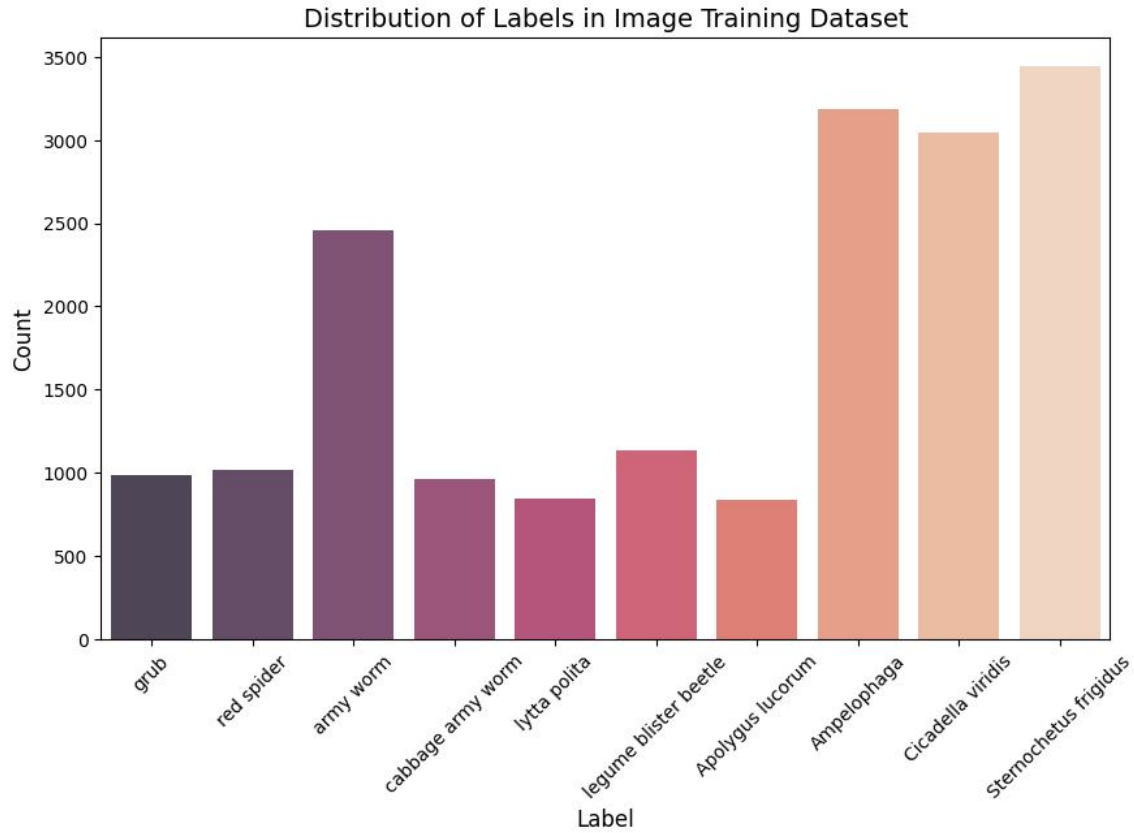


Figure 3.2: Distribution of 10 largest class from IP102

3.2 Data preprocessing and augmentation

Data processing is an essential part before using it to train the models. It is a process that removes unwanted details from the image and enhances specific features that are needed to train the model. So, it can give accurate results. Data augmentation is a technique where it creates new image data from the existing image data. Deep learning models need large amounts of data to predict anything perfectly. Hence, it was proved that augmentation has a great importance in image classification due to insufficient image data. The categories of IP102 dataset are unbalanced[7]. So, we use data augmentation to balance the data of the categories. There are some common techniques in data augmentation that can increase data without creating an overfitting problem. The techniques are rotation, flipping, zooming and scaling etc. We used a vertically and horizontally shifting technique and a zoom in and out technique within 0.8 to 1 range and shifting the color channel technique.

3.3 Proposed model and explanation

We selected CNN models like Resnet 50 , Mobilenet V2 and Vision Transformer model for insect pest detection and classification.

ResNet-50

ResNet-50 is a deep convolutional neural network architecture which is widely used in computer vision tasks. For residual learning to address the problem of vanishing gradients, which commonly hindered the training of very deep neural networks. In terms of blocks, the primary building blocks of ResNet-50 are residual blocks. Each residual block contains a short connection that skips one or more then one layers in forming a residual connection. Which allows the network to learn residual functions with reference to the layer inputs and making it more easier to optimize.

In terms of using this architecture, we can see that ResNet-50 uses residual blocks

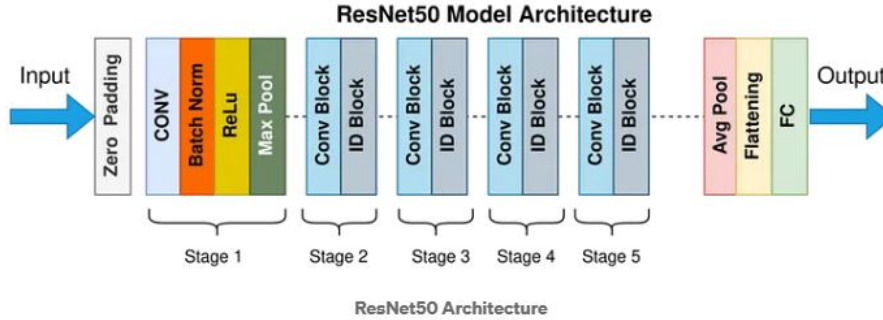


Figure 3.3: Resnet50

with skip connections that allow gradients to flow directly through the network and make it easier to train very deep networks effectively. Likewise, By enabling the training of deeper networks, ResNet-50 achieves higher accuracy on complex tasks. Deeper networks can capture more intricate patterns in data. This architecture is used in a variety of tasks including image classification, object detection, semantic segmentation and many more.

Mobilenet V2

MobileNetV2 is a convolutional neural network architecture designed specifically for mobile and edge devices which offers a balance between performance and efficiency. Unlike traditional residual blocks that use a series of convolutions followed by a shortcut connection MobileNetV2 uses inverted residuals where the shortcut connection is placed between twin bottleneck layers. In this architecture, ReLU6 activation function is used in most layers, providing better performance for mobile and edge devices by being more resistant to low precision computation.

In terms of using this architecture, we can see that the design with shortcuts connecting thin bottleneck layers helps in reducing the computational load while preserving the accuracy. Likewise, using a linear layer at the end of each epoch block helps in

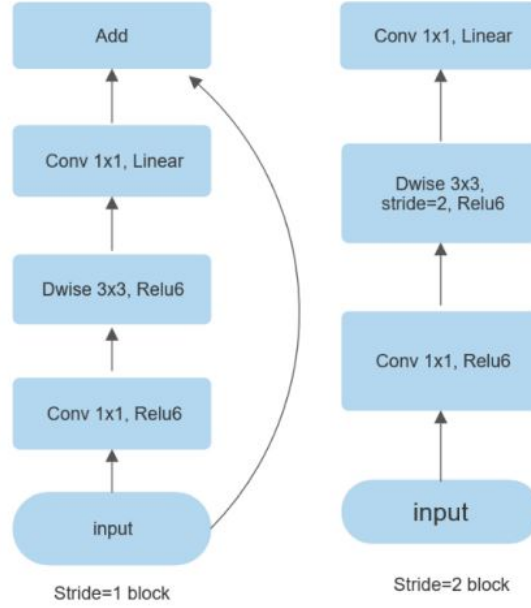


Figure 3.4: MobileNetV2

retaining more information and deducing the loss of representational power. For that reason, MobileNetV2 achieves high efficiency by extensively using depthwise separable convolutions and reducing the number of operations required.

Vision Transformer

Vision Transformer architecture (ViT) is a deep learning model which is used in images where the transformers for image recognition at scale. The ViT model applies the transformer architecture which has been highly successful in natural language processing to computer vision tasks, specifically image classification. Likewise, it works by dividing an image into fixed-size patches, embedding these patches into vectors and adding positional embeddings to retain spatial information. These embeddings are fed into a Transformer encoder composed of layers of multi-head self-attention and feed-forward neural networks. A special classification token is included to aggregate information across patched. ViT models pre-trained in large datasets and achieve competitive performance with traditional CNNs and offer scalability and flexibility for vision tasks. Key innovations include the use of self-attention to capture relationships between image patches and the ability to scale effectively with large models and datasets. ViT benefit for a uniform architecture applicable across different domains and simplifying model design and enhancing generalization.

3.4 Implementation

The dataset were split into 80% training , 10% validation and 10% training. The batch size was 32. The image size was 224*224. The step size of the training data was 560 (the total number of samples in the training data/ Batch size of training data). The step size of the validation data was 93(the total number of samples in the validation data/ Batch size of validation data). We compiled the models with Adam

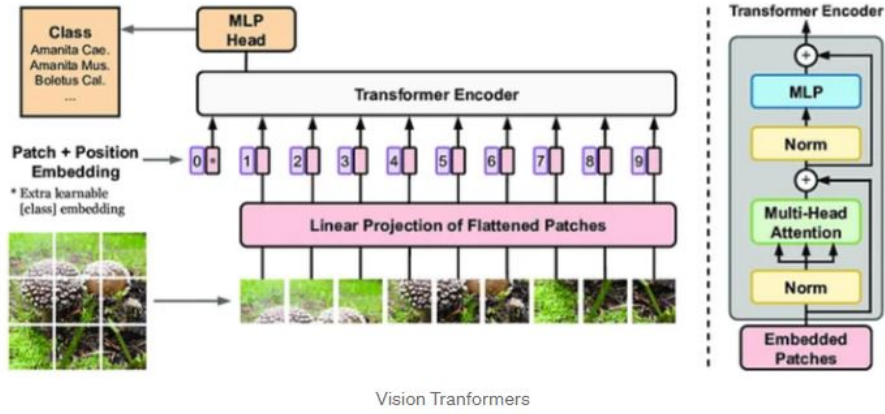


Figure 3.5: Vision Transformer

as an optimizer and the learning rate was 0.0001. We used categorical cross-entropy as a loss function and accuracy as a metrics for evaluation for all the models. After configuring the models , we trained the model for 50 epochs.

For Resnet 50 and Mobilenet V2 , we used Global average 2D to reduce the dimension of feature maps which help to prevent overfitting and reduce computational cost. Then we used dense layer with 128 units and ReLu activation function. Then We will introduce a dropout of 0.5 layer to also prevent the models from overfitting. Finally, we implement output layer where the number of units was 10 with softmax function.

For the Vision transformer , the patch size was 7 and transfer layer was 5. The projection dimension is 64. The transformer units are 128 and 64 respectively. The mlp head units are 512 and 256 respectively. The transformer will use 4 attention heads.

Chapter 4

Results and Discussion

We used Intel Xeon processor and Nvidia Tesla P100 Gpu with 29 GB ram to train these models. We used necessary python libraries like keras , tensorflow to train our model .We used Resnet50 , Mobilenet V2 and Vision Transformer as a pre-trained model. The dataset was splitted into 80% training data , 10% validation data and 10% testing data. We used Adam as an optimizer with 0.0001 learning rate.After training all of the models , we found that Resnet 50 and Mobilenet V2 achieved more than 90% accuracy and on the other hand Vision Transformer achieved almost 80% accuracy.

Mobilenet V2

We got impressive results from Mobilenet V2. It achieved 97.50% accuracy.The Validation accuracy is also impressive. It got 83.9% accuracy. Among all the models , Mobilenet V2 achieved highest accuracy in both training and validation.The model loss was a little bit higher at the start of the epoch but at the end of the training it ended a little bit above 0%

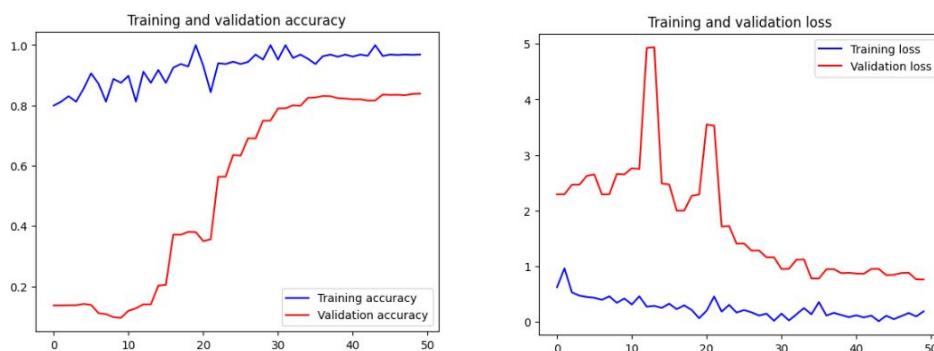


Figure 4.1: Training and Validation accuracy graph

Resnet 50

Resnet 50 performed impressively good . It achieved 95.5% in training. The validation accuracy of this model was 82% which is also impressive.

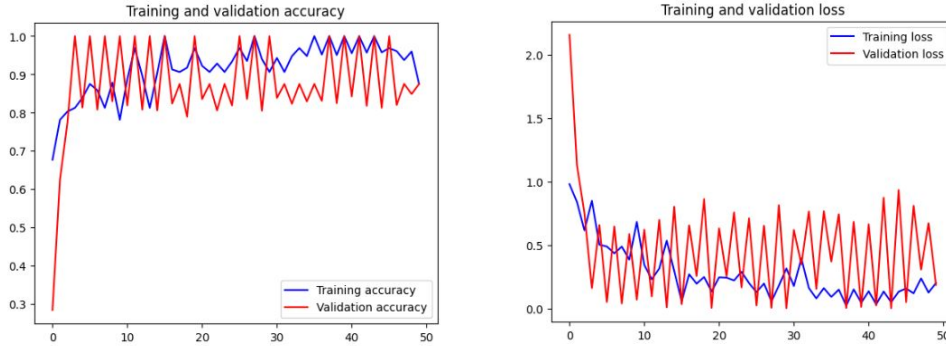


Figure 4.2: Training and Validation accuracy graph

Vision Transformer

Vision Transformer performed lowest among all the models which is 79.16%. But the validation accuracy was not good. It got 62.19%. It means that this model cannot properly predict any unseen data. So, this model will be not suitable for real life application. As we will be using models to identify the insects that will be harmful for crops.

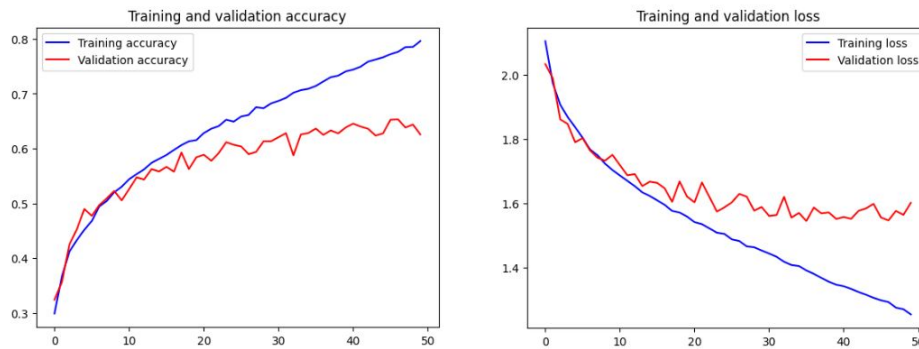


Figure 4.3: Training and Validation accuracy graph

Chapter 5

Conclusion

According to FAO, every year pests cause a lot of damage to crops [3]. Therefore, machine learning techniques are the best option for farmers to reduce their losses. In our paper we are going use various machine learning models which will be going to identify the pest and give its information as an output. We will be using IP102 dataset for the research and we will use Resnet50, MobilenetV2 and Vision Transformer to train data and get our required output. The main purpose of this research is to create a new agricultural pest classification model to identify insect pest before it destroying any crops. Thus, saving our agriculture sector from destruction.

Bibliography

- [1] D. Xia, P. Chen, B. Wang, J. Zhang, and C. Xie, “Insect detection and classification based on an improved convolutional neural network,” *Sensors*, vol. 18, no. 12, p. 4169, 2018.
- [2] L. Liu, R. Wang, C. Xie, *et al.*, “Pestnet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification,” *Ieee Access*, vol. 7, pp. 45 301–45 312, 2019.
- [3] X. Wu, C. Zhan, Y.-K. Lai, M.-M. Cheng, and J. Yang, “Ip102: A large-scale benchmark dataset for insect pest recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8787–8796.
- [4] Y. Nanekaran, D. Zhang, J. Chen, Y. Tian, and N. Al-Nabhan, “Recognition of plant leaf diseases based on computer vision,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2020.
- [5] M. Xin and Y. Wang, “Image recognition of crop diseases and insect pests based on deep learning,” *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–15, 2021.
- [6] H. Alshammari, K. Gasmi, I. B. Ltaifa, M. Krichen, L. B. Ammar, and M. A. Mahmood, “Olive disease classification based on vision transformer and cnn models,” *Computational Intelligence and Neuroscience*, vol. 2022, 2022.
- [7] N. C. Kundur and P. Mallikarjuna, “Insect pest image detection and classification using deep learning,” *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 9, 2022.
- [8] H. Li, S. Li, J. Yu, Y. Han, and A. Dong, “Plant disease and insect pest identification based on vision transformer,” in *International conference on internet of things and machine learning (IoTML 2021)*, SPIE, vol. 12174, 2022, pp. 194–201.
- [9] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, “Explainable vision transformer enabled convolutional neural network for plant disease identification: Plantxvit,” *arXiv preprint arXiv:2207.07919*, 2022.
- [10] M. Xu, S. Yoon, Y. Jeong, J. Lee, and D. S. Park, “Transfer learning with self-supervised vision transformer for large-scale plant identification.,” in *CLEF (Working Notes)*, 2022, pp. 2238–2252.
- [11] S. Zhang, R. Jing, and X. Shi, “Crop pest recognition based on a modified capsule network,” *Systems Science & Control Engineering*, vol. 10, no. 1, pp. 552–561, 2022.

- [12] S. Zhao, J. Liu, Z. Bai, C. Hu, and Y. Jin, "Crop pest recognition in real agricultural environment using convolutional neural networks by a parallel attention mechanism," *Frontiers in Plant Science*, vol. 13, p. 839 572, 2022.
- [13] F. Ali, H. Qayyum, and M. J. Iqbal, "Faster-pestnet: A lightweight deep learning framework for crop pest detection and classification," *IEEE Access*, 2023.
- [14] J. An, Y. Du, P. Hong, L. Zhang, and X. Weng, "Insect recognition based on complementary features from multiple views," *Scientific Reports*, vol. 13, no. 1, p. 2966, 2023.
- [15] M. De Silva and D. Brown, "Plant disease detection using vision transformers on multispectral natural environment images," in *2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, IEEE, 2023, pp. 1–6.
- [16] R. Devi, V. Kumar, and P. Sivakumar, "Efficientnetv2 model for plant disease classification and pest recognition," *Computer Systems Science & Engineering*, vol. 45, no. 2, 2023.
- [17] X. Fu, Q. Ma, F. Yang, *et al.*, "Crop pest image recognition based on the improved vit method," *Information Processing in Agriculture*, 2023.
- [18] P. Gole, P. Bedi, S. Marwaha, M. A. Haque, and C. K. Deb, "Trincnet: A lightweight vision transformer network for identification of plant diseases," *Frontiers in Plant Science*, vol. 14, p. 1 221 557, 2023.
- [19] G. Li, Y. Wang, Q. Zhao, P. Yuan, and B. Chang, "Pmvt: A lightweight vision transformer for plant disease identification on mobile devices," *Frontiers in Plant Science*, vol. 14, p. 1 256 773, 2023.
- [20] B. Thokala and S. Doraikannan, "Detection and classification of plant stress using hybrid deep convolution neural networks: A multi-scale vision transformer approach," *Traitement du Signal*, vol. 40, no. 6, 2023.
- [21] P. Venkatasachandranth and M. Iyapparaja, "Pest detection and classification in peanut crops using cnn, mfo, and evita algorithms," *IEEE Access*, 2023.
- [22] U. Barman, P. Sarma, M. Rahman, *et al.*, "Vit-smartagri: Vision transformer and smartphone-based plant disease detection for smart agriculture," *Agronomy*, vol. 14, no. 2, p. 327, 2024.
- [23] P. Venkatasachandranth and M. Iyapparaja, "Gnvt-an enhanced image-based groundnut pest classification using vision transformer (vit) model," *Plos one*, vol. 19, no. 3, e0301174, 2024.
- [24] H. Thakkar, A. Pingle, S. Kulkarni, R. Saraf, and R. V. Kulkarni, "Disease and pest detection in crops using computer vision: A comprehensive study,"