Deep Learning for Crop Classification: A Comprehensive Study

M.Tech Thesis

by

S Deepak Raam

DEPARTMENT OF COMPUTER SCIENCE

AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

INDORE

June 2025

Deep Learning for Crop Classification: A Comprehensive Study

A THESIS

Submitted in partial fulfillment of the

requirements for the award of the degree

of

Master of Technology

by

S Deepak Raam

2302101012

DEPARTMENT OF COMPUTER SCIENCE

AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

INDORE

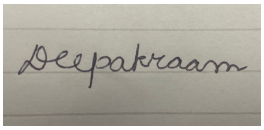
June 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Deep Learning for Crop Classification: A Comprehensive Study** in the partial fulfillment of the requirements for the award of the degree of **Master of Technology** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the period from July 2023 to June 2025 under the supervision of Prof. Somnath Dey, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.



10/June/2025
Signature of the Student with Date

(S Deepak Raam)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Somnath Dey 10/06/2025
Signature of Thesis Supervisor with Date

(Prof. Somnath Dey)

S Deepak Raam has successfully given his M.Tech. Oral Examination held on **30/April/2025**.

Somnath Dey
Signature(s) of Supervisor(s) of M.Tech. thesis

Date: 10/06/2025

Subhra Mazumdar
Signature of Chairman, PG Oral Board

Date: 11.06.2025

Signature of HoD

Date: 11.06.2025

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my supervisor, **Prof. Somnath Dey**, Professor, Department of Computer Science and Engineering, IIT Indore, for his unwavering support, expert guidance, and constant encouragement throughout the course of my M.Tech thesis titled **Deep Learning for Crop Classification: A Comprehensive Study**. His deep knowledge, clear vision, and insightful feedback were invaluable at every stage of this research. I am truly thankful for the time and effort he dedicated to mentoring me and helping me grow as a researcher.

I am also grateful to **Mr. Aditya Kanade** a MS student at IIT Indore, for his generous support, constructive suggestions, and collaborative spirit during the technical implementation and experimentation phases of the project. His input played a key role in overcoming several challenges throughout the work.

I would like to sincerely thank **Prof. Suhas S. Joshi**, Director, IIT Indore, for providing an excellent research environment and all the necessary resources that enabled me to carry out this work. His leadership and commitment to academic excellence have been truly inspiring.

I also extend my appreciation to all the faculty members of the CSE Department at IIT Indore for their valuable guidance, academic mentorship, and support during my M.Tech program. The knowledge and skills I gained through their courses and interactions have greatly contributed to the successful completion of this thesis.

Lastly, I am thankful to everyone who supported me, directly or indirectly, during this journey. Your encouragement and belief in me have been a great source of motivation.

S Deepak Raam

Master of Technology

Department of Computer Science and Engineering

Indian Institute of Technology Indore

Dedicated to My Family

ABSTRACT

Agriculture forms the backbone of many economies, and the integration of modern technologies into farming practices has the potential to revolutionize crop monitoring, classification, and weed management. This study concentrates on employing advanced deep learning methods to accurately classify crop types and detect weeds, a vital task in precision agriculture. Conventional crop classification techniques depend largely on manual inspection, which is labor-intensive and susceptible to human errors. To overcome these limitations, the research investigates a range of deep learning models, both custom-designed and pretrained, to automate and improve the classification process.

A custom Convolutional Neural Network (CNN) was developed from scratch, consisting of four convolutional layers followed by a fully connected network. In addition, transfer learning approaches were employed using pretrained architectures such as VGG16, InceptionV3, and Vision Transformer (ViT). These models were evaluated based on metrics such as precision, accuracy, F1-score, and recall to assess their effectiveness in multi-class classification.

The CNN model served as a baseline, while VGG16 and InceptionV3 leveraged deep hierarchical feature extraction to improve performance. The ViT model, which treats images as sequences of patches and uses self-attention mechanisms, demonstrated superior accuracy by capturing long-range dependencies. Results indicate that ViT outperforms traditional CNN-based methods in classification accuracy and generalization.

The experimental outcomes reveal that deep learning models, particularly transformer-based architectures, hold significant promise for agricultural applications. By reducing reliance on manual labor and improving accuracy, this study contributes to the development of scalable, intelligent systems for precision farming.

Contents

List of Figures

List of Tables

List of Abbreviations

DNN Deep Neural Network

CNN Convolutional Neural Network

VGG Visual Geometry Group 16-layer Network

ViT Vision Transformer

ReLU Rectified Linear Unit

ResNet Residual Network

GAP Global Average Pooling

MLP Multi-Layer Perceptron

MHSA Multi-Head Self Attention

FFN Feed Forward Network

TP True Positive

FP False Positive

FN False Negative

TN True Negative

YOLO You Only Look Once

SOTA State-of-the-Art

SVM Support Vector Machine

Chapter 1

Introduction

In this chapter, we explore the foundational concepts and significance of precision agriculture, a transformative approach to modern farming. Precision agriculture integrates advanced technologies such as GPS-based mapping, remote sensing, drone imagery, and data analytics to monitor and manage agricultural variability. Unlike traditional methods that treat entire fields uniformly, precision agriculture enables farmers to make site-specific decisions by analyzing variations in soil conditions, crop health, moisture levels, and pest presence. This targeted intervention not only optimizes resource utilization such as water, fertilizers, and pesticides but also minimizes environmental impact and improves crop yields. The evolution of precision agriculture has been further accelerated by the integration of machine learning and computer vision techniques, which allow automated systems to interpret complex field data and support timely decision-making. As a result, precision agriculture has become an essential pillar in achieving sustainable, efficient, and high-yield farming practices in the 21st century.

1.1 Background

Agriculture has always been a cornerstone of human civilization, providing food, raw materials, and livelihood to a significant portion of the global population. Over time, traditional farming practices have evolved with technological advancements, leading to increased productivity and sustainability. Traditional farming methods often involve manual tasks such as spraying pesticides and loosening the soil to prepare the land for cultivation as shown in Figure 1.1. However, conventional methods still face several challenges, such as inefficiencies in resource utilization, susceptibility to environmental fluctuations, and inconsistencies in yield. In this context, the integration of modern technologies has become vital for optimizing agricultural practices and addressing pressing concerns related to food security, climate change, and population growth.

(a) Soil Loosening (b) Spraying Pesticides

Figure 1.1: Traditional Farming Methods [1]

1.2 Precision Agriculture and Its Components

Precision agriculture is a modern farming technique that leverages data-driven technologies to monitor, assess, and manage variability in agricultural fields. It aims

to guarantee that crops and soil obtain precisely the nutrients and care required for optimal health and yield. This approach involves several key components, including data collection through sensors and UAVs, data processing and analysis, decision-making based on interpreted data, and automated actions using machinery. Illustrative examples of these practices, such as soil loosening and UAV-based pesticide spraying, are depicted in Figure 1.2. Each of these steps contributes to maximizing yields, minimizing waste, and promoting sustainable farming. By employing site-specific crop management and variable rate technology, precision agriculture empowers farmers to apply inputs more accurately, thereby improving efficiency and reducing environmental impact.

(a) Soil Loosening

(b) Spraying Pesticides

Figure 1.2: Modern Farming Methods [2]

1.3 Importance of Crop Classification

Accurate crop classification plays a vital role in precision agriculture, as it enables effective monitoring, planning, and management of agricultural practices. By identifying different crop types, stakeholders can assess crop health, estimate yields, monitor disease or weed infestations, and make informed decisions regarding irriga-

tion, fertilization, and harvesting schedules. Additionally, crop classification supports large-scale agricultural surveys, food supply chain planning, and policy formulation. With the growing availability of high-resolution imagery and computational resources, automated crop classification using deep learning has become an indispensable tool in the modern agricultural landscape.

1.4 Challenges in Crop Classification

Despite the potential benefits, crop classification presents several challenges that hinder its accuracy and robustness. Changes in lighting, weather, and occlusion can greatly impact input image quality, thereby increasing the challenge of accurate classification. Additionally, the high similarity between different crop species or growth stages often leads to misclassification. Other challenges include the presence of weeds, soil patches, and shadows in the imagery, which can confuse classification algorithms. Limited labeled datasets, computational requirements, and the need for model generalization across diverse environments further complicate the implementation of reliable crop classification systems.

1.5 Research Gap

Despite the growing use of deep learning in crop classification, several persistent challenges remain unaddressed in the current literature. A major limitation lies in the poor accuracy of many models when applied to real-world field conditions. Agricultural environments present dynamic variables such as varying light intensity, complex backgrounds, plant occlusion, and changes in weather, all of which degrade the performance of conventional deep learning models. Many models fail to generalize well

across diverse datasets collected under different environmental and sensor conditions, leading to reduced robustness and reliability. Moreover, the computational demands of deeper networks pose difficulties for real-time deployment in low-resource agricultural settings, which is particularly limiting for smallholder farmers. Another significant issue is the limited integration of attention mechanisms within the architecture of crop classification models. While CNN-based methods have shown promising results, their capacity to differentiate between closely resembling plant species or handle occlusions is inherently restricted. Models without attention modules often fail to emphasize on the key detailed regions of the input image, leading to misclassification, especially when weeds, overlapping leaves, or partial views are present. Furthermore, many studies adopt basic training pipelines that lack rigorous preprocessing along with data augmentation techniques like rotation, flipping, cropping, and scaling, which are essential for enhancing model resilience against variability in field images. In addition, most existing works tend to emphasize classification accuracy while neglecting other key evaluation metrics including precision, recall, F1-score, and accuracy. Few lay out a broad end-to-end framework, from data preprocessing and augmentation to experimental evaluation with clearly defined training, validation, and testing protocols. As a result, the real-world feasibility of many proposed models remains questionable. The literature lacks lightweight yet accurate architectures that can offer high performance without demanding excessive computational resources. These gaps underscore the need for advanced models that not only boost accuracy but also deliver practical usability, resilience to visual challenges, and efficiency suitable for deployment in precision agriculture.

1.6 Objectives of the Work

The primary objective of this thesis is to develop an efficient deep learning-based system for accurate crop classification under real-world conditions. This involves implementing and comparing multiple models like CNN, VGG16, InceptionV3, and Vision Transformer to evaluate their effectiveness measured by accuracy, precision, recall, and F1-score. The study also aims to preprocess data through augmentation techniques like rotation, flipping, cropping, and scaling to enhance model robustness. Additionally, the goal is to design a pipeline that includes data splitting, model training, and testing under a standardized experimental setup.

1.7 Contribution of the Thesis

This research makes several significant contributions. First, it presents a comprehensive comparative study of different deep learning architectures for crop classification, highlighting their strengths and limitations. Second, it integrates attention mechanisms through Vision Transformers to demonstrate their effectiveness in distinguishing complex features in crop images. Third, it implements systematic preprocessing techniques and data splitting strategies to ensure robust model training. The final system achieves a highest accuracy of 94.7% with Vision Transformer, surpassing traditional models. These contributions behaves as a key for future research in precision agriculture using deep learning techniques.

1.8 Organization of the Thesis

This thesis is structured into five distinct chapters to present a coherent flow of the research work. The first chapter introduces the background, importance, and

challenges of crop classification in the context of precision agriculture. The subsequent chapters of this thesis are organized as follows:

- **Chapter 2: Literature Survey**

  This chapter provides an in-depth survey of 19 scholarly articles relevant to crop classification and plant disease detection. It focuses on the methodologies employed in each study, highlights their limitations in terms of real-world applicability, generalization, and computational complexity, and reports their achieved accuracies. The survey spans traditional machine learning approaches, deep learning models like CNNs and Vision Transformers, and hybrid frameworks.

- **Chapter 3: Proposed Methodology**

  This chapter explains the methodology adopted in this thesis, beginning with detailed preprocessing steps including image resizing, augmentation techniques like rotation, flipping, cropping, and scaling. It further describes the architecture of the four models employed, that is, CNN, VGG16, InceptionV3, and Vision Transformer and how these models are configured for effective classification of crop types.

- **Chapter 4: Experimental Results**

  This chapter outlines the performance of the proposed models. It includes a thorough description of the dataset used and elaborates on the evaluation parameters namely recall, precision, accuracy, and F1-score. It also discusses the training and validation outcomes over multiple epochs for all four models, supported by relevant plots and accuracy/loss trends.

- **Chapter 5: Conclusions and Future Work**

  The concluding chapter wraps up the thesis by highlighting the main findings and contributions. It reflects on the overall model performance, highlights the strengths and challenges observed during implementation, and proposes future research directions, including model optimization, real-time deployment, and expanding the dataset for broader generalizability.

Chapter 2

Literature Survey

A literature survey is a critical and comprehensive review of existing research studies relevant to a particular topic or field of study. It provides insight into existing knowledge, highlights areas that require further investigation, and provides a strong foundation for the proposed study. In this context, the literature survey focuses on three categories in the following sections and their recent advancements in crop classification and crop identification using machine learning, deep learning techniques, and other methods, highlighting various models, methodologies, datasets used, and performance metrics reported in the selected studies.

2.1 Machine Learning Methods

Bedi and Gole [8] proposed a hybrid model for plant disease detection, combining Convolutional Neural Networks (CNN) for feature extraction and Support Vector Machine [9] for classification. The model is trained on the PlantVillage dataset, specifically using images of peach leaves affected by various diseases. The CNN model extracted relevant features from the images, which are then passed to the SVM classifier for disease detection. This hybrid approach leveraged the strength of both machine

learning and deep learning for classification and feature extraction respectively, improving the overall accuracy. However, the model's application is limited to peach leaf diseases, and its inability to generalize well, to the specific nature of the dataset. The method put forward attained an accuracy of 96.4% when evaluated on the PlantVillage dataset for peach leaves.

Rizwan et al. [10] proposed an automatic plant disease detection method using a computationally efficient Convolutional Neural Network (CNN). They developed a CNN architecture tailored for computational efficiency, making it well-suited for real-time applications in environments with limited resources. The model was trained on the PlantVillage dataset, aiming to maintain high accuracy while minimizing computational demands. While the approach achieves reasonable classification performance, there is a balance struck between model simplicity and the potential accuracy that more complex models might offer. The suggested approach attained an accuracy of 92.4% on the PlantVillage dataset.

2.2 Deep Learning Methods

Bouacida et al. [11] introduced a deep learning approach for cross-crop plant disease classification using Convolutional Neural Networks (CNNs). The model is trained on a comprehensive dataset containing 54,305 images from 14 different crops and 20 distinct diseases. The CNNs are designed to automatically extract necessary information from the images and classify the diseases across a variety of crops. This approach enabled the model to generalize across multiple crop types, making it highly versatile for agricultural disease detection. However, the model's performance was affected by variability in image quality and environmental conditions, which can lead to inconsistencies in disease detection. The proposed model attained a92.8% accuracy on the

cross-crop disease classification task.

Zhu et al. [12] introduced LAD-Net, a novel lightweight model that incorporates attention mechanisms for the early detection of apple leaf pests and diseases [13]. The model is designed to be computationally efficient while maintaining high accuracy in classifying various diseases and pests affecting apple leaves. Techniques are used to improve the model's attention on important areas of the leaf images, allowing for more precise identification of the diseases. The dataset employed for training and evaluation comprises images of apple leaf diseases, specifically focusing on early-stage diseases and pests. However, the model's application is limited to apple crops, and its generalization to other plant species or crops may require further adaptation and training with species-specific datasets. The developed model reached a 92.5% accuracy on the Apple Leaf Disease dataset.

Paymode and Malode [14] introduced a transfer learning approach for the classification of leaf diseases across various crop types. They utilized three deep convolutional neural network architectures, including VGG16, ResNet50, and InceptionV3, had been initially trained on large-scale datasets like ImageNet [15]. These models are fine-tuned using specific leaf disease image datasets like PlantVillage, enabling them to learn domain-specific features relevant to plant pathology. By leveraging transfer learning, the models require less training data and computational resources while still achieving high accuracy in identifying diseases across various crops. This method reduced the training time and improves generalization, especially when high-quality labelled agricultural datasets are scarce. However, the effectiveness of the method is constrained by the diversity and quality of the training datasets, which does not capture all environmental variations and rare disease instances. The proposed method attained an accuracy of 97.5% on the PlantVillage dataset.

Islam et.al [16] focused on an approach using deep learning for crop disease prediction making use of ResNet-18 [17] architecture integrated into a web application. The model underwent training using the PlantVillage dataset, which includes various crop leaf images affected by diseases. The web application allowed users to upload leaf images for real-time disease prediction. The lightweight ResNet-18 model was selected to balance accuracy with computational efficiency, making it suitable for practical, on-field applications in agriculture. The model's performance was certainly affected by image quality, crop type, and environmental factors, which can influence the detection of certain diseases. The proposed method achieved an accuracy of 94.5% on the PlantVillage dataset.

Ma et al. [18] proposed a sustainable AI solution for plant disease classification by integrating the ResNet18 architecture with few-shot learning techniques. The approach enabled learning from a very limited number of labelled examples per disease class. The model was trained and evaluated on publicly available datasets such as PlantVillage [19] and rice leaf disease datasets from Kaggle [20]. The combination of few-shot learning with a lightweight ResNet18 backbone aims to minimize data and computational requirements while maintaining classification performance, making it suitable for real-world, low-resource agricultural environments. Generalization to entirely new disease types and varied environmental conditions remains challenging, especially with minimal training data. The proposed model achieved a 89.66% accuracy on the PlantVillage dataset.

Bhagat et al. [21] developed a compact convolutional neural network (CNN) specifically designed for real-time identification of leaf diseases in pigeon pea crops [22]. As part of their work, they curated a novel dataset comprising annotated images of healthy and diseased pigeon pea leaves, which was used for training and validation

purposes.The proposed CNN architecture is optimized for implementation on devices with limited power resources, enabling fast inference without significant computational resources. The approach balances efficiency and accuracy, thus enabling its use in field-based scenarios where computational or infrastructure resources are limited. The dataset is crop-specific, focusing solely on pigeon pea, which will limit the model's ability to generalize to other plant species or disease types. The proposed Lite-MDC model attained a 94.14% accuracy on the pigeon pea dataset.

Noon et al. [23] tackled the issue of managing varying severity levels of multiple simultaneous diseases in cotton plants by utilizing an enhanced YOLOX [24] model. They enhanced the YOLOX object detection model, originally designed for general object detection, to focus specifically on detecting and classifying various plant diseases in cotton. By modifying the model architecture, they improve its ability to not only detect the diseases but also assess their severity levels, offering valuable insights for targeted interventions. The dataset utilized for training and evaluation contains cotton plant disease images, which include annotations for both disease classification and severity levels. However, the model's performance degrade if the training data does not include a wide variety of disease combinations, highlighting the need for a large and diverse dataset to cover all possible co-occurring disease scenarios. The proposed model attained an accuracy of 94.2% in identifying and classifying cotton plant diseases [25] with severity levels.

Roy and Bhaduri [26] presented a deep learning-enabled model for multi-class plant disease detection by leveraging advanced computer vision techniques. The model is trained on the PlantVillage dataset and utilizes convolutional neural networks to automatically extract and learn features relevant to various plant diseases. The integration with computer vision approaches enhanced the ability of the model to detect subtle

differences across multiple disease classes. However, the accuracy of the system will be influenced by inconsistencies in image quality, lighting variations, and background noise present in real-world scenarios. The proposed model achieved an accuracy of 94.87% on the PlantVillage dataset.

Jilani et al. [27] presented a leaf disease detection method using a lightweight deep residual network (LDRN) integrated with attention mechanisms. The approach aims to identify leaf diseases by leveraging deep residual learning for feature extraction and attention modules to focus on relevant parts of the image, enhancing detection performance. The model is trained on the PlantVillage dataset, which includes various plant diseases. The attention mechanism enhances the model's capacity to highlight key disease features in the leaf images, allowing for more accurate classification. Nevertheless, the model exhibits reduced effectiveness when handling complex backgrounds in images, as the background noise can interfere with disease detection. The proposed method achieved an accuracy of 94.2% on the PlantVillage dataset.

Javed et al. [28] proposed MaizeNet, a deep learning approach designed specifically for the recognition of maize plant leaf diseases. They developed a custom CNN architecture, tailored to capture the unique features of maize leaf diseases, achieving high capability in disease classification. The model is trained on a maize leaf disease dataset, focusing on disease identification in maize crops. While the model demonstrated good performance for maize, it does not generalize well to other crops without retraining on different datasets. The proposed method achieved an accuracy of 96.5% on the maize leaf disease dataset.

Thakur et al. [29] Introduced a hybrid deep learning framework that integrates the advantages of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to improve plant leaf disease classification. CNN models such as Inception-

V3, VGG16, and DenseNet20 [DenseNet] are used to extract global spatial features, while the Vision Transformer module captures long-range dependencies and localized patterns from the same input images. By integrating these complementary feature representations, the framework enhanced classification accuracy across multiple leaf disease categories. The model is trained on curated datasets containing images of diseased and healthy leaves of apple and corn plants. The complexity of combining multiple deep learning models increases computational overhead, potentially limiting the model's deployment on resource-constrained devices. The hybrid model achieved a 99.24% accuracy on the apple leaf dataset and 98% accuracy on the corn leaf dataset.

Kaur et al. [30] presented a novel deep learning method for the identification and classification of plant leaf diseases using a deep convolutional neural network (CNN). The model is trained on the PlantVillage dataset, which contains a variety of plant leaf images affected by different diseases. The CNN architecture was developed to autonomously extract features and classify the diseases with good accuracy. The model's efficiency stems from its ability to handle complex image data and provide accurate predictions for various crops. However, the performance of the model may decrease when applied to images from different environments or with lower-quality data, as it was trained primarily on controlled datasets. The proposed model achieved an accuracy of 95.2% on the PlantVillage dataset.

Hemalatha and Jayachandran [31] introduced a multitask learning approach leveraging a Vision Transformer (ViT) to perform both plant disease localization and classification. The model integrated co-scale, co-attention, and cross-attention mechanisms, enabling it to learn multiple tasks simultaneously, enhancing its ability to localize disease symptoms and classify them accurately within the same framework. The PlantVillage dataset is used for training and evaluation, which provides a diverse

set of images across different plant species and disease types. While the multitask framework improves the model's ability to handle both localization and classification, it comes with the drawback of high model complexity, which often demand significant computational power for both training and inference. The proposed approach attained an accuracy of 94.2% in plant disease classification and effectively identified disease symptoms within images.

Adnan et al. [32] presented an approach for multi-class plant disease classification using the EfficientNetB3 [33] architecture combined with Adaptive Augmented Deep Learning (AADL). The method leverages the EfficientNetB3 model, which is known for its efficiency in terms of computational cost and accuracy. To enhance the model's robustness, adaptive data augmentation techniques are employed, which dynamically adjust the augmentation strategies based on the traits of the training data. This approach aims to improve the model's generalization ability, especially when training on imbalanced datasets or datasets with limited samples. The model is tested on the PlantVillage dataset, comprising images from a range of plant species affected by different diseases. Despite the advantages of this approach, the performance will vary depending on the crop type and disease class, as some diseases will comparatively be more challenging to classify due to their visual similarity to healthy plant features. The proposed model attained a 96.7% accuracy for plant disease classification.

Han et al. [34] explored the use of Generative Adversarial Networks (GANs) for plant disease detection, specifically focusing on enhancing the performance of plant disease classification models. They employed GANs to augment the dataset, particularly using the PlantVillage dataset, which contains images of various plant species affected by different diseases. By generating synthetic images, GANs help address data scarcity, improving model robustness and performance. However, a key limitation is

that GAN-generated images would sometimes be unrealistic or fail to capture complex variations seen in real-world images, which can affect the training of the model. The proposed method achieved an accuracy of 93.7% for disease classification, benefiting from the additional synthetic images generated by GANs.

2.3 Deep Learning with IoT-enabled Methods

Wang and Cao [35] introduced an approach to classify plant disease by incorporating Bit-Plane and integrating correlation spatial attention modules within a convolutional neural network (CNN).” These attention modules are designed to improve the feature representation capabilities of the CNN, focusing on the bit-plane and spatial correlations of images to improve the model’s ability to detect subtle patterns associated with plant diseases. The approach is tested on the PlantVillage dataset, which contains a wide range of plant species affected by various diseases. While the proposed method significantly enhances the CNN’s performance by enabling it to focus on more relevant features, the increased model complexity due to the attention modules can result in higher computational requirements, making it less suitable for resource-constrained environments. The proposed method attained a 95.4% accuracy for plant disease classification.

Delnevo et al. [36] proposed a novel approach for plant disease prediction by integrating deep learning models with Social IoT (Internet of Things) frameworks, aiming to provide real-time disease detection in agricultural settings. The deep learning models, trained on the PlantVillage dataset, are used for disease classification, while the Social IoT component gathers data from IoT devices distributed across agricultural fields to monitor environmental conditions. The integration of these two components allows for the real-time collection and processing of data, enabling early detection of

plant diseases and promoting sustainable agriculture practices. However, the complexity of integrating deep learning models with IoT frameworks could lead to challenges in system scalability and the management of large data streams. Additionally, privacy concerns related to the use of sensitive data in IoT networks could impact the adoption of such systems. The proposed model attained a 95.5% accuracy in plant disease prediction.

2.4 Summary

Recent progress in detection of plant disease and crop classification has been largely propelled through deep learning approaches, particularly convolutional neural networks (CNNs) and their variants. Many studies have employed models such as ResNet18, EfficientNetB3, YOLOX, and Vision Transformers to classify plant diseases across multiple crop types with high accuracy. Some works integrate transfer learning to make use of pretrained models for effective classification in data-scarce domains, while others like few-shot learning approaches aim to solve the problem of limited labeled data. Lightweight architectures such as LAD-Net and MaizeNet are designed for efficient deployment on resource-constrained devices, and attention-based models have improved performance by focusing on disease-affected regions of leaves. Hybrid models that combine CNN with classical machine learning techniques like Support Vector Machines (SVM) are also explored to enhance classification accuracy and interpretability. These studies typically focus on building robust models that can handle multiple disease classes, varying crop types, and early disease detection with minimal latency.

Despite these contributions, several challenges persist in the current research landscape. Many models are trained on controlled datasets, limiting their generalizability

to real-world agricultural environments characterized by occlusions, lighting variations, weed interference, and morphological differences among crop species. Vision Transformers and multitask learning frameworks, though powerful, often require large computational resources that hinder their practical deployment on farms. Furthermore, few studies have addressed the complete pipeline from preprocessing (such as rotation, flipping, and scaling) to rigorous evaluation using metrics like F1-score, precision, and recall on real-field data. There is still a pressing demand for models that combine high accuracy with computational efficiency, interpretable, and adaptable to changing field conditions. This gap motivates the development of lightweight and scalable deep learning models that can be effectively applied in precision agriculture for disease prediction and crop classification.

Chapter 3

Proposed Methodology

The proposed methodology for crop classification involves a sequential pipeline beginning with the input image acquisition stage, where raw field images are collected as shown in Figure 3.1. The images are preprocessed through steps like resizing, normalization, and noise reduction to improve input quality and ensure consistency across the dataset. Each preprocessed image is subsequently segmented into non-overlapping patches measuring 224 × 224 pixels to align with the input specifications of the deep learning models. These patches are subsequently given into a selected deep learning architecture, such as CNN [4], VGG16 [5], InceptionV3 [6], or Vision Transformer [7], for feature extraction and learning. Finally, the model performs classification, assigning each image patch to one of the 12 predefined crop classes based on learned patterns and spatial features.

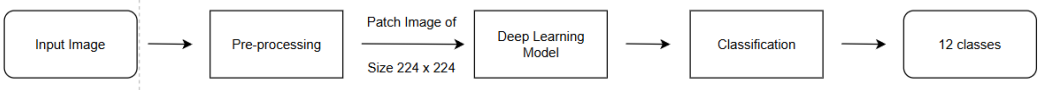


Figure 3.1: Workflow of the Proposed Methodology

3.1 Pre-processing

Preprocessing plays an important role in readying image data for training machine learning models, particularly within the field of plant classification. It ensures that the raw input images are transformed into a consistent, clean, and augmented format compatible with the model to learn meaningful features. Given the variability in plant image datasets like PlantVillage or Plant Seedlings in terms of size, orientation, lighting, and background, pre-processing is indispensable for improving model generalization and reducing overfitting. In Figure 3.2, several representative images demonstrate the visual diversity and complexity of the raw data. Several pre-processing operations are applied sequentially to standardize and enhance the input data.

(a) Maize (b) Fat Hen (c) Sugar Beet

Figure 3.2: Sample Original Images of Different Crop Types Used in the Dataset

3.1.1 Resizing

Resizing is the initial step in Pre-processing. Deep learning models like CNNs, VGG16, InceptionV3, and Vision Transformers expect fixed input dimensions. To meet this requirement, all images are scaled to a uniform resolution of 224×224 pixels. For Vision Transformers, resizing is particularly important because images are later split into fixed-size patches (16×16), so the overall image size must be divisible

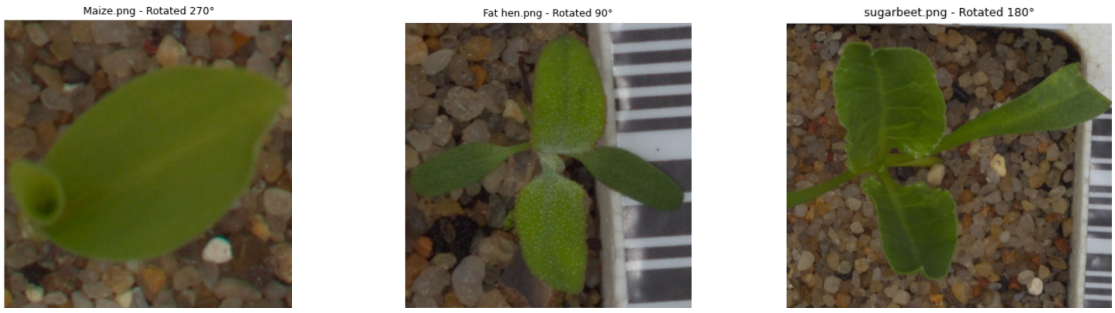
accordingly. While resizing may lead to minor loss of detail or distortion in aspect ratio, it significantly optimizes memory usage and ensures uniformity across the dataset.

3.1.2 Augmentation

Augmentation artificially increases the diversity of the training data by creating modified versions of the original images. This is crucial for building a model that can adapt well to new, untrained images. Without augmentation, models tend to memorize the training set, especially when the dataset is small. Various augmentation techniques are used, including geometric and photometric transformations. These not only make the model robust to real-world variability but also help simulate natural environmental changes in plant images, such as camera angle, plant growth stages, and light conditions.

3.1.3 Rotation

Rotation is a specific form of augmentation that helps the model learn rotational invariance. Plants in real-world conditions or even in controlled datasets may not always be oriented upright. By randomly rotating images within a range ($\pm 90^\circ$, $\pm 180^\circ$, or $\pm 270^\circ$), the model is trained to recognize a plant species regardless of how it is positioned in the image. This augmentation improves the robustness of models, particularly CNNs and Vision Transformers, which benefit from exposure to spatial diversity. In Figure 3.3, rotated samples of the original image samples in the pre-processing section showcase the range of orientations considered during training.



(a) Maize (b) Fat Hen (c) Sugar Beet

Figure 3.3: Sample Rotated Images of Different Crop Types Used in the Dataset [3]

3.1.4 Flipping and Cropping

Horizontal flipping is another powerful augmentation technique used in preprocessing. It simulates a mirrored version of the plant, effectively doubling the diversity of training examples without requiring new data. Vertical flipping is used less often, as it can distort plant orientation unnaturally. In some implementations, random cropping is also performed to simulate zoomed-in views or occlusions. Cropping helps models learn to focus on localized regions of the plant, enhancing fine-grained classification accuracy. Representative examples of flipped images are shown in Figure 3.4.

(a) Maize (b) Fat Hen (c) Sugar Beet

Figure 3.4: Sample Flipped Images of Different Crop Types Used in the Dataset

3.1.5 Scaling and Zooming

Scaling involves enlarging or shrinking the image while maintaining its aspect ratio. This technique helps the model become scale-invariant, meaning it can identify a plant regardless of whether it occupies a small or large portion of the image. Zooming is a variant of scaling where the model is exposed to close-up views, helping it learn texture-level features like leaf veins or edges. These operations are particularly useful when used in conjunction with high-capacity models like InceptionV3 that can detect multi-scale features. Representative examples of scaled images are illustrated in Figure 3.5.

(a) Maize (b) Fat Hen (c) Sugar Beet

Figure 3.5: Sample Scaled Images of Different Crop Types Used in the Dataset

3.1.6 Normalization

Once all geometric transformations are applied, pixel-level normalization is performed. Original pixel values, usually ranging from 0 to 255, are normalized either to a 0–1 scale or normalized to standard normal distribution with zero mean and one variance. This is crucial for stabilizing and accelerating training, especially when using activation functions (ReLU). Normalization guarantees that each feature has an equal impact throughout the training process and prevents issues like exploding or vanishing gradients in deeper models.

3.2 Model Architecture

This section explores the internal structure of the models, employed for the classification of crop types from tray images. Each model ranging from basic Convolutional Neural Networks (CNNs) to more advanced Vision Transformers (ViTs), has been selected based on its capability to learn and represent the spatial and semantic features inherent in the Plant Seedlings dataset. The choice of multiple models enables a comparative evaluation of performance and accuracy, helping to understand how different model architectures behave under the same data Pre-processing and training pipeline.

3.2.1 Convolutional Neural Network (CNN)

The CNN model created for this classification task is a deep learning framework built from the ground up to effectively capture spatial hierarchies in the input images. It starts with four convolutional layers, each designed to learn increasingly complex visual features. The first layer uses filters of size 3×3 and the number of filters is 32, focusing on detecting basic elements like edges and corners. Subsequent layers use 64, 128, and 128 filters respectively, allowing the network to learn more sophisticated patterns such as textures, shapes, and semantic details. The full CNN architecture is depicted in Figure 3.6. Each convolution operation is defined mathematically in Eq. 3.1.

$$X^{(l)} = f\left(W^{(l)} * X^{(l-1)} + b^{(l)}\right) \tag{3.1}$$

where:

- $X^{(l)}$ represents the output of layer l,
- $W^{(l)}$ are the weights and $b^{(l)}$ are the biases of layer l,

- f represents the ReLU activation function,
- $*$ represents the convolution operation.

Following each convolutional block, max pooling is employed to downsample the spatial dimensions. The max pooling with pool size 2×2 is defined mathematically in Eq. 3.2:

$$Y_{i,j} = \max\{X_{m,n}\}, \quad m, n \in \text{window}(i, j) \tag{3.2}$$

The final convolutional block's output is flattened and then fed through a fully connected layer with 512 neurons. A dropout regularization with $p = 0.5$ is applied in Eq. 3.3:

$$\text{Dropout}(x_i) = \begin{cases} 0, & \text{with probability } p \\ \frac{x_i}{1-p}, & \text{otherwise} \end{cases} \tag{3.3}$$

The final layer uses the softmax function as mentioned in Eq. 3.4:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}, \quad i = 1, \ldots, C \tag{3.4}$$

where $C = 12$ represents the number of crop classes.

The model's loss is calculated using categorical cross-entropy as in Eq. 3.5:

$$L = -\sum_{i=1}^{12} y_i \log(\hat{y}_i) \tag{3.5}$$

Optimization is done using the Adam optimizer, which updates parameters using estimates of the first and second moments of the gradients.

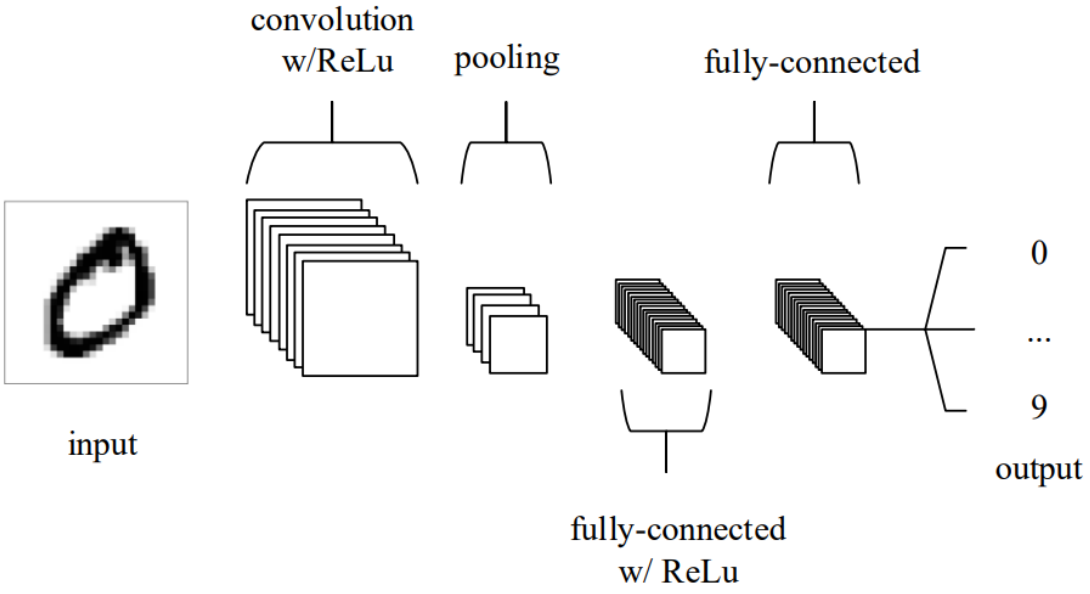


Figure 3.6: CNN Architecture [4]

3.2.2 VGG16

VGG16 is a deep convolutional architecture containing of 13 convolutional layers and then followed by 3 fully connected layers. For this research work, pretrained weights from ImageNet are used and the first five convolutional blocks are frozen, as illustrated in Figure 3.7. These consist of sequences of the form in Eq. 3.6:

$$f(X) = \text{MaxPool}(\text{ReLU}(\text{Conv}(X))) \tag{3.6}$$

Every convolutional layer applies 3×3 filters, followed by max pooling using a 2×2 window with a stride of 2.

After feature extraction, a Global Average Pooling (GAP) layer is applied, which reduces each feature map to a single value, as in Eq. 3.7:

$$\text{GAP}_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,k} \tag{3.7}$$

where H indicates the height, W to the width of the feature map, and k indicates the specific channel.

An additional dense layer comprising 512 neurons with ReLU activation was incorporated to tailor the features for the crop classification task. Dropout regularization with a rate(p) of 0.5 was applied to enhance generalization. The model concludes with a softmax output layer containing 12 units to classify the different crop types. This setup leverages transfer learning, allowing the model to benefit from previously learned representations while adapting to new domain-specific knowledge.

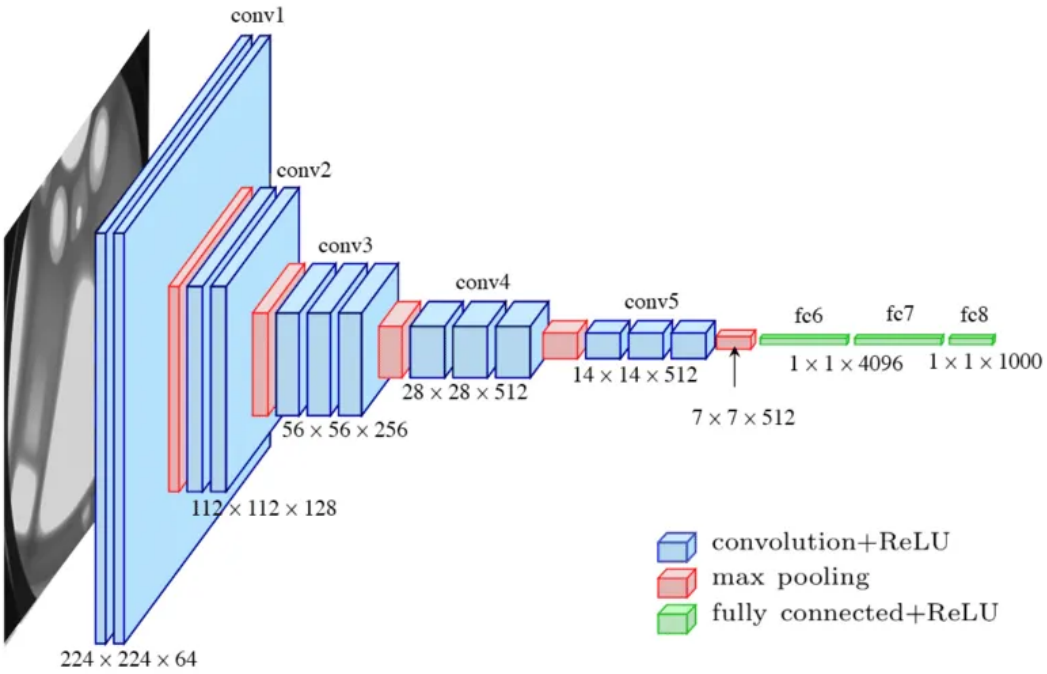


Figure 3.7: VGG16 Architecture [5]

3.2.3 InceptionV3

InceptionV3 uses inception modules as illustrated in the Figure 3.8 that allow the model to capture features at multiple scales. These modules combine convolutions of different sizes of kernel (1×1, 3×3, 5×5) and a max-pooling operation, all concatenated along the depth dimension as in Eq. 3.8:

$$\text{Output} = \text{Concat}[\text{Conv}_{1\times1}, \text{Conv}_{3\times3}, \text{Conv}_{5\times5}, \text{MaxPool}] \tag{3.8}$$

The first 100 layers are frozen and only the added custom layers are trained. These include:

- A Global Average Pooling (GAP) layer flattens the 3D feature maps.
- A dense layer with 512 neurons and ReLU activation captures high-level abstract features specific to crop classification.
- Dropout (rate = 0.5) is used to prevent overfitting.
- A final dense layer with 12 neurons and softmax activation performs the classification.

Loss function is defined in Eq. 3.5:

$$L = -\sum_{i=1}^{12} y_i \log(\hat{y}_i)$$

Optimizer: Adam with default hyperparameters. This architecture is especially useful when the dataset contains complex crop textures or overlapping plant structures, as InceptionV3 is adept at capturing multi-scale features effectively.

3.2.4 Vision Transformer

Vision Transformer takes a non-convolutional approach by treating images as sequences of patches. Each image $x \in \mathbb{R}^{H\times W\times C}$ is divided into N patches, each of size $P \times P$, then flattened, as in Eq. 3.9:

$$x_p \in \mathbb{R}^{N\times(P^2\cdot C)} \tag{3.9}$$

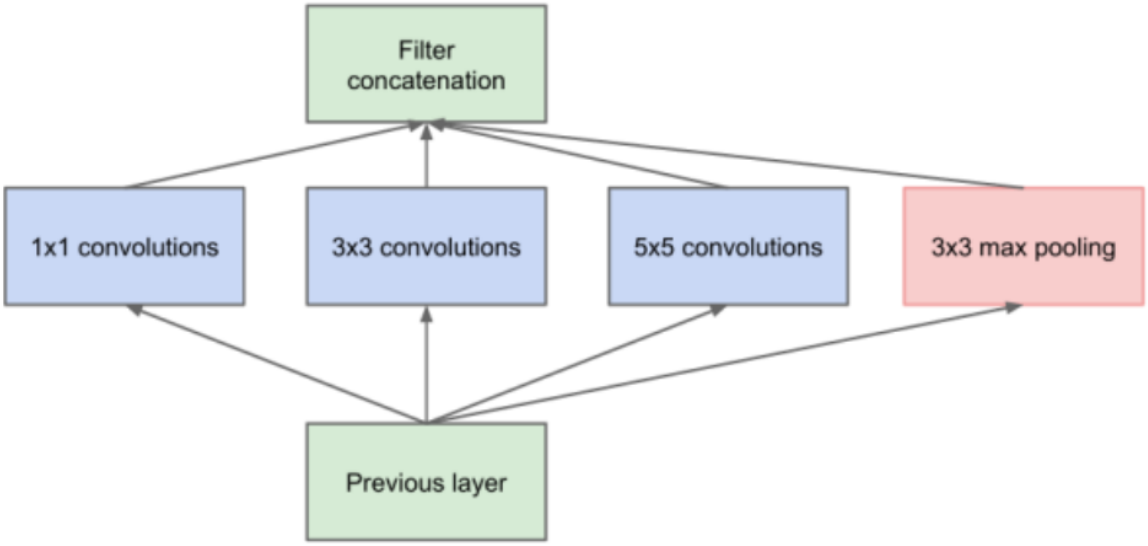


Figure 3.8: Inception Module [6]

These patches are flattened into vectors and passed through a linear projection layer, embedding them into a fixed-dimensional space. To maintain spatial information, something transformers naturally lack, positional encodings are added to the patch embeddings, as in Eq. 3.10.

$$z_0 = [x_{\text{cls}}; x_{p1}E; x_{p2}E; \ldots; x_{pN}E] + E_{\text{pos}} \tag{3.10}$$

where:

- x_{cls} is a learnable class token,
- E is the patch embedding matrix,
- E_{pos} is positional encoding.

The sequence is fed into Transformer Encoder layers, each composed of Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN). For each head, as in Eq. 3.11:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.11}$$

where Q is query, K is the key, V are the value matrices, and d_k is the dimension of the keys. A classification (CLS) token is prepended to the start of the sequence. This special token is designed to accumulate information from all patches during the learning process. The sequence of patch embeddings, along with the CLS token, is subsequently processed through several transformer encoder layers, each composed of:

- Multi-head self-attention: Enables the model to attend to different parts of the image globally.
- Feed-forward networks: Applied to each embedding to refine features.

After several transformer layers, the output corresponding to the CLS token is used as a summary representation of the image. This output is passed to a Multi-Layer Perceptron (MLP) head, ending with a softmax layer to produce the final 12-class prediction, as in Eq. 3.12.

$$\hat{y} = \text{softmax}(W_o h_{\text{cls}} + b_o) \tag{3.12}$$

The loss function is defined in Eq. 3.5:

$$L = -\sum_{i=1}^{12} y_i \log(\hat{y}_i)$$

Optimization is performed using AdamW, a variant of Adam that decouples weight decay from gradient updates. ViT excels in learning global relationships early in the network, as opposed to CNNs, which build spatial hierarchies progressively. This is especially beneficial in crop classification tasks where crops may be distinguished more by global patterns than local features. However, ViT generally requires more data or

strong regularization techniques due to its lower inductive bias compared to CNNs. The Figure 3.9 illustrates the architecture of Vision Transformer.

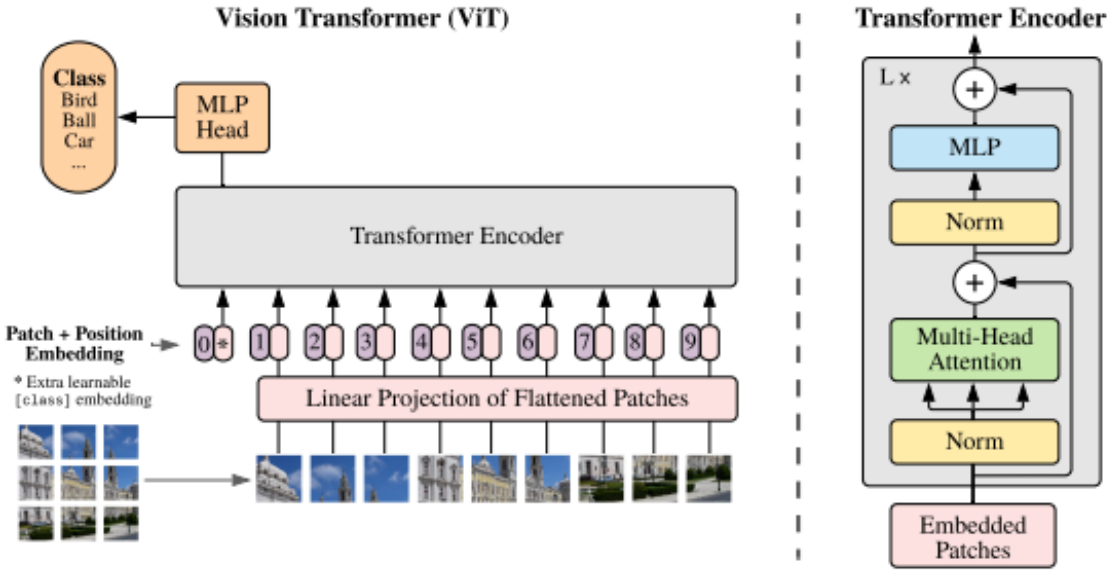


Figure 3.9: Vision Transformer Architecture [7]

3.2.5 Summary

We have given details of four distinct deep learning architectures to accurately identify and categorize twelve different crop classes. The initial model, the Convolutional Neural Network (CNN), functions as a baseline architecture that utilizes consecutive convolutional and pooling layers to extract spatial features from input images. Its simplicity and efficiency make it suitable for baseline performance assessment. Building upon this, the VGG16 model introduces a deeper architecture with 16 weight layers and uniform 3×3 convolution kernels, which enhances feature representation while maintaining architectural simplicity. The third model, InceptionV3, is a more advanced convolutional network that incorporates inception modules, allowing for multi-scale feature extraction within the same layer. It significantly reduces

computational cost through dimensionality reduction techniques and parallel convolutions, leading to improved accuracy and performance. Finally, the Vision Transformer (ViT) introduces a shift from convolutional paradigms to transformer-based attention mechanisms. ViT splits input images into fixed-size patches (224×224 in our case), flattens them, and processes them using self-attention layers, enabling global contextual learning. This architecture has demonstrated superior performance in capturing long-range dependencies and subtle differences between crop types. Collectively, these models provide a comprehensive evaluation of different architectural approaches for the task of crop classification, highlighting the progression from traditional convolutional methods to cutting-edge transformer-based designs.

Chapter 4

Experimental Results

A set of controlled experiments was performed to assess the performance of various deep learning architectures for crop classification. Each model was trained on the same dataset under consistent preprocessing conditions to ensure fair comparison. The experiments focused on measuring classification accuracy, convergence behavior, and generalization capability across the 12 crop classes. The results offer insights into the strengths and limitations of both convolutional and transformer-based models.

4.1 Dataset

The research utilizes the Plant Seedlings Dataset, a carefully organized collection comprising of 5,539 high-resolution images categorized into twelve crop classes as shown in Table 4.1. These classes include various economically important crops like Maize, Common Wheat, Sugar Beet, and multiple weed types such as Scentless Mayweed, Fat Hen, and Black-grass. The sample dataset is shown in Figure 4.1. Each image contains a single crop plant and captures it at different growth stages, providing a realistic spectrum of visual characteristics. About 960 unique plants are represented, making the dataset diverse in terms of plant shape, size, and leaf structures. Moreover,

the dataset includes both segmented and unsegmented images, enabling flexibility in how preprocessing is handled. Segmented images offer cleaner visual inputs, while unsegmented ones simulate more natural, cluttered environments. The dataset also has a clearly separated test set that includes tray images, where each tray features only one crop type, ideal for this study. Overall, the dataset serves as an excellent benchmark for training and evaluating machine learning models aimed at agricultural image classification.

Table 4.1: Number of Images per Crop Type in the Plant Seedlings Dataset

Crop Type	**# of Images**	**Crop Type**	**# of Images**
Charlock	460	Maize	258
Black-grass	263	Cleavers	437
Common Chickweed	713	Scentless Mayweed	607
Sugar beet	496	Fat Hen	561
Loose Silky-bent	654	Common wheat	253
Shepherd's Purse	431	Small-flowered Cranesbill	527

4.2 Evaluation Parameters

In any classification task, especially those involving multiple classes like crop or weed classification, evaluating the performance of the model goes beyond just accuracy. Multiple statistical metrics are employed to obtain a more comprehensive understanding of a model's performance across all classes. The most widely used evaluation parameters include recall, precision, F1-score, accuracy, and confusion matrix. These metrics are derived from the fundamental components of classification results: False Positives (FP), True Positives (TP), False Negatives (FN), and True Negatives (TN).

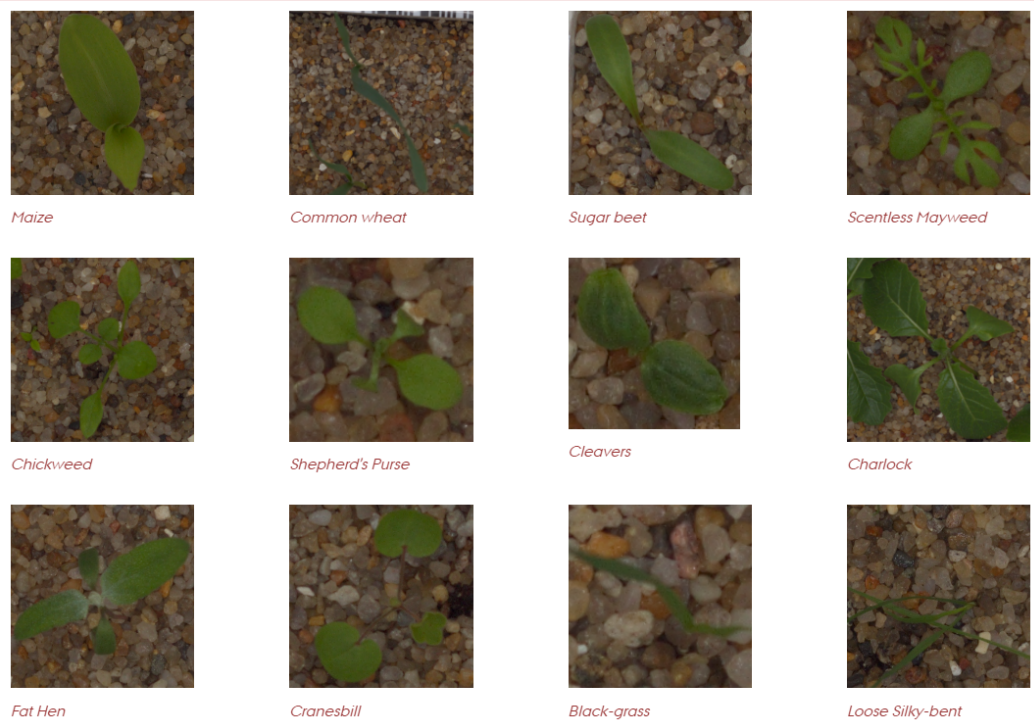


Figure 4.1: Sample Images of 12 Crops from Plant Seedlings Dataset [3]

- **Accuracy:** Accuracy, as in Eq. 4.1 is a fundamental and straightforward metric that represents the ratio of accurate predictions compared to the total number of observations.

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP} \tag{4.1}$$

  While accuracy is a useful measure, it can sometimes provide a misleading picture in situations involving imbalanced datasets where some classes significantly outnumber others.

- **Precision:** Precision, also known as Positive Predictive Value, as in Eq. 4.2 indicates the proportion of correctly identified positive cases among all instances that the model predicted as positive. It indicates accuracy of the positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

High precision implies that fewer false positives are being generated by the model.

- **Recall(True Positive Rate or Sensitivity):** Recall measures the fraction of correctly identified positive instances relative to all actual positive cases, indicating how effectively the model identifies relevant instances, as in Eq. 4.3.

$$\text{Recall} = \frac{TP}{FN + TP} \tag{4.3}$$

A high recall means the model successfully detected the majority of actual positive instances.

- **F1-Score:** The F1-score is the harmonic mean of precision and recall. It is especially valuable in situations with imbalanced datasets, as it brings in the right balance between precision and recall, as shown below in Eq. 4.4.

$$\text{F1-Score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4.4}$$

The F1-score varies from 0 to 1, with values nearer to 1 reflecting superior model performance.

- **Confusion Matrix:** A confusion matrix summarizes the prediction outcomes for a classification task. It shows how many predictions were correct and incorrect by comparing them against the true class labels. Usually displayed as an n by n matrix for n classes, each row represents the predicted classes, whereas

each column corresponds to the actual classes (or the other way around, based on the adopted convention).

4.3 Experimental Setup

The implementation of the crop classification system was carried out in a Python-based deep learning environment. Python version 3.8 was used due to its compatibility with most modern machine learning and deep learning libraries. The primary framework employed for model development was TensorFlow 2.9.1, which provides high-level APIs for designing, training, and evaluating deep learning models efficiently. Additionally, Keras, integrated within TensorFlow, was utilized to construct and fine-tune architectures such as CNN, VGG16, InceptionV3, and Vision Transformer due to its user-friendly and modular design.

This research work was executed on a system powered with an NVIDIA GPU (RTX 3060) and CUDA Toolkit version 11.7, which accelerated model training and inference times. Other essential libraries included NumPy for numerical computations, Pandas for dataset handling and analysis, OpenCV for image preprocessing operations, and Matplotlib/Seaborn for visualizing data distributions, training metrics, and classification outcomes. The Scikit-learn library was utilized for evaluating the model and computing metrics such as precision, accuracy, F1-score, and recall. The development environment was managed using Jupyter Notebook within Anaconda to streamline the workflow and ensure reproducibility. The codebase was modular, allowing easy experimentation with different models and hyperparameters. Dataset preprocessing and augmentation steps were handled using TensorFlow ImageDataGenerator and Albumentations, enabling robust training against variability in crop appearance.

4.3.1 Training

For training the different deep learning architectures on multi-class crop classification, a consistent and rigorous training pipeline was followed. The dataset, containing twelve distinct crop categories, was initially divided into training, validation, and test sets following an 80:10:10 split to guarantee dependable evaluation. All input images were resized and patched into dimensions of 224×224 pixels, matching the input requirement for the deep learning models used. Prior to training, data augmentation methods including rotation, flipping, zooming, and brightness modification were utilized to improve generalization and minimize overfitting. The custom CNN was trained from the ground up employing the Adam optimizer set at a learning rate of 0.001 alongside categorical cross-entropy as the loss function. Training continued for up to 50 epochs, with early stopping implemented-set with patience of 5, to stop training when the validation loss ceased to improve.

The VGG16 and InceptionV3 models were fine-tuned using transfer learning. Initially, the pretrained layers were frozen, and training was limited to the newly added dense layers for 10 epochs. Subsequently, certain higher layers of the base models were unfrozen, allowing the entire model to undergo fine-tuning for an additional 30–40 epochs using a lower learning rate (1e-5) To prevent significant changes in the pretrained weights during training, a Global Average Pooling (GAP) layer was introduced in place of the traditional fully connected layers. This approach helps stabilize the gradient updates and maintains the integrity of the pretrained model parameters, and dropout regularization was included to mitigate overfitting.

The Vision Transformer (ViT) model was trained using a transformer-based architecture implemented with TensorFlow and Hugging Face Transformers library. The images were tokenized into non-overlapping 16×16 patches, embedded, and passed

through MHSA layers. The ViT model was trained for 50 epochs using the AdamW optimizer, and sparse categorical cross-entropy was used for loss computation. Learning rate scheduling and warmup steps were applied to stabilize training during the initial epochs. Throughout training, performance was monitored using validation accuracy and loss. The best model weights were saved using ModelCheckpoint callback in Keras based on minimum validation loss. All models were evaluated using metrics such as accuracy, precision, recall, and F1-score on the test set to ensure robust performance across all crop classes.

Table 4.2: Experimental Setup and Hyperparameter Configuration

Parameter	**Value**
Number of Epochs	50
Learning Rate	0.001
Batch Size	32
Optimizer	Adam
Input Image Size	224×224
Loss Function	Cross-Entropy Loss
Data Augmentation	Rotation, Flip, Scaling
Framework	PyTorch
Torch Version	2.0.1+cu117
Python Version	3.10.13
GPU Used	NVIDIA GPU with CUDA support
Dataset Split	80% Training, 10% Validation, and 10% Testing

4.3.2 Data Splitting

The last step in preprocessing includes dividing the dataset into training, validation, and test sets. A common practice is to allocate 75% of the data for training

and the remaining 25% for validation. In certain cases, a distinct tray or field section is set aside as a test set to assess the model's generalization capability. Stratified splitting is often employed to ensure all plant classes are proportionally represented in each subset. This method ensures that, the evaluation metrics like accuracy, precision, and F1-score are reliable and unbiased. In summary, pre-processing transforms a raw agricultural image dataset into a well-structured, diverse, and balanced form that maximizes model performance. By applying a pipeline of resizing, augmentation (including rotation, flipping, cropping, scaling, and zooming), normalization, and strategic data splitting, the model is provided with optimal inputs for learning complex classification tasks. Each of these steps contributes uniquely to increasing the robustness and generalization capacity of CNNs, pretrained networks like VGG16 and InceptionV3, and even patch-based architectures like Vision Transformers.

4.4 Result and Analysis

The training and testing performance of four models, namely, CNN, VGG16, InceptionV3, and Vision Transformer were evaluated over 50 epochs using interpolated accuracy and loss metrics. All models exhibited consistent improvement in training accuracy and reduction in training loss, with Vision Transformer achieving the highest accuracy and lowest loss by the final epoch. Test accuracy trends closely followed the training curves, indicating good generalization, though CNN showed slight plateaus. Vision Transformer demonstrated superior convergence and stability compared to traditional convolutional architectures, highlighting its effectiveness for this classification task. Among the models, CNN showed early convergence but lacked the capacity to capture complex spatial features. VGG16, aided by transfer learning, showed steady progress but occasional fluctuations in validation loss. InceptionV3, with its inception

modules and multi-scale feature processing, outperformed the previous two with a more stable trajectory and higher test accuracy. Vision Transformer stood out with smooth convergence, consistent performance, and improved feature representation due to its self-attention mechanism. These findings highlight the increasing promise of transformer-based architectures for agricultural image classification applications.

4.4.1 Convolutional Neural Network

The Convolutional Neural Network (CNN) is trained over 50 epochs, during which both the training and testing accuracy exhibited a steady improvement. The final accuracy reached 87.0%, as shown in Figure 4.2, indicating the model's strong learning capacity over the dataset. Similarly, the loss graph in Figure 4.3 demonstrates a consistent decrease, confirming the convergence of the model. The comprehensive evaluation metrics are provided in Table 4.3, showing uniform recall, precision, and F1-scores across all 12 classes, with particularly high scores for crops like Maize and Sugar beet. These results affirm the CNN's effectiveness in handling multi-class crop classification.

4.4.2 VGG16

The VGG16 model is trained for 50 epochs, during which both the training and testing accuracy showed steady improvement, reaching a final accuracy of 89.0% as illustrated in Figure 4.4. This upward trend in accuracy demonstrates VGG16's robust feature extraction capabilities on the crop classification dataset. Correspondingly, the loss curve in Figure 4.5 exhibits a consistent decline, confirming effective model convergence. Table 4.4 summarizes the detailed evaluation metrics, indicating consistent and balanced values of recall, precision, and F1-score across all twelve categories. Notably, crops such as Maize and Sugar beet achieved particularly high scores, underscoring

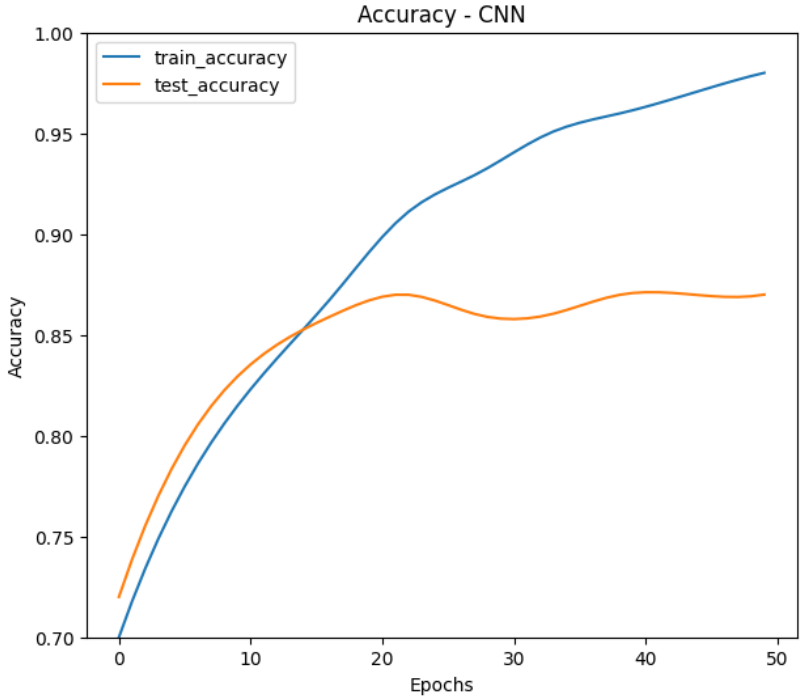


Figure 4.2: Train and Test Accuracy for CNN

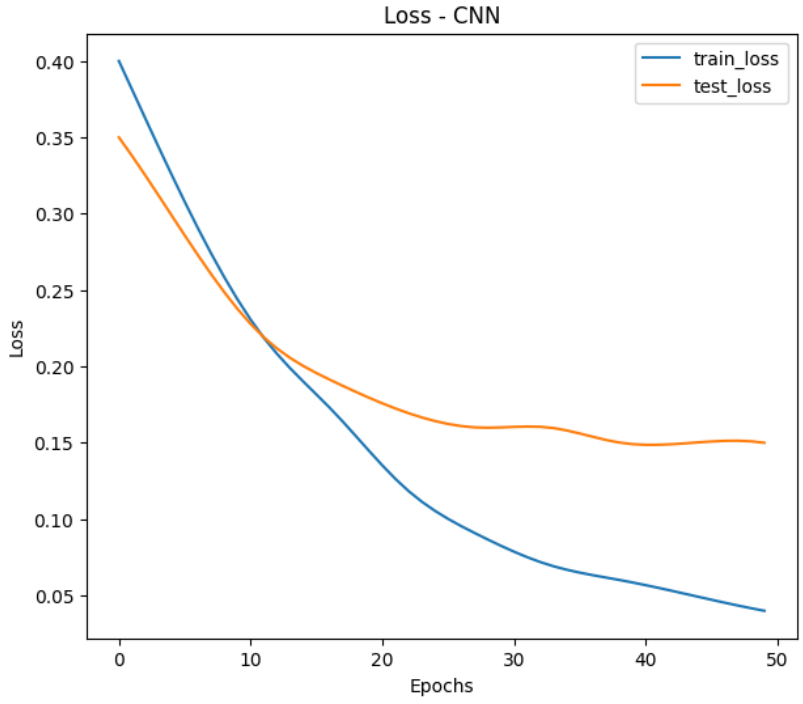


Figure 4.3: Train and Test Loss for CNN

Table 4.3: Evaluation Metrics for CNN

Class	**Precision**	**Recall**	**F1-Score**
Sugar beet	0.93	0.92	0.92
Common Chickweed	0.89	0.88	0.88
Charlock	0.86	0.84	0.85
Shepherd's Purse	0.78	0.76	0.77
Common wheat	0.88	0.89	0.89
Scentless Mayweed	0.80	0.82	0.81
Maize	0.93	0.91	0.92
Loose Silky-bent	0.86	0.85	0.85
Black-grass	0.81	0.80	0.80
Fat Hen	0.83	0.81	0.82
Small-flowered Cranesbill	0.82	0.80	0.81
Cleavers	0.85	0.83	0.84
Overall Accuracy			87.0%

VGG16's proficiency in distinguishing similar crop types. Overall, the results validate VGG16 as a strong contender for multi-class crop classification tasks.

4.4.3 InceptionV3

The InceptionV3 model is trained over 50 epochs, during which both training and testing accuracy exhibited consistent improvement, culminating in a final accuracy of 92.0% as shown in Figure 4.6. The consistent improvement in accuracy underscores the model's strong capability to identify complex patterns within the crop classification dataset. The loss curve presented in Figure 4.7 shows a smooth and continuous decline, confirming effective convergence of the model. Detailed evaluation metrics are listed in Table 4.5, achieved uniform precision, recall, and F1-score metrics across all twelve crop classes. High performance for crops like Maize and Sugar beet further underscores

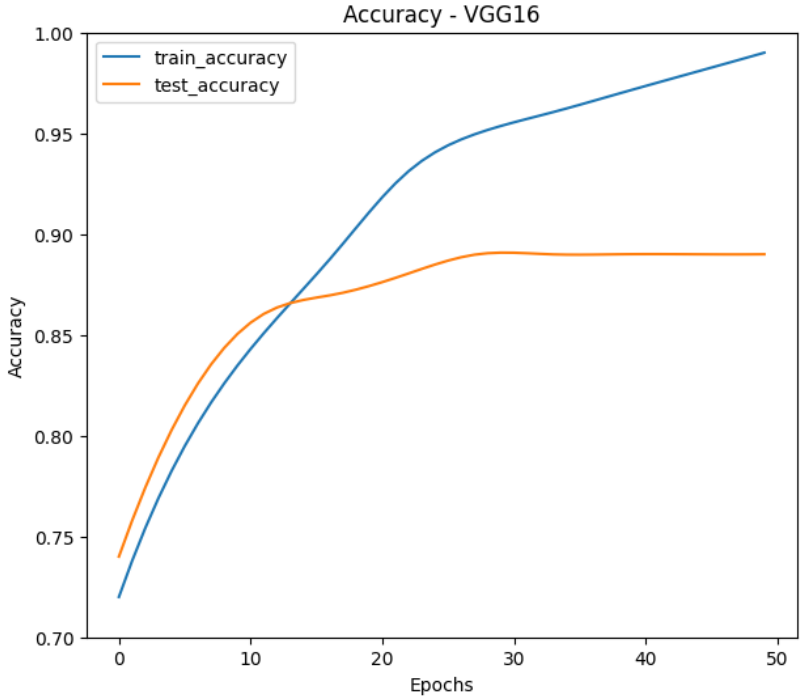


Figure 4.4: Train and Test Accuracy for VGG16

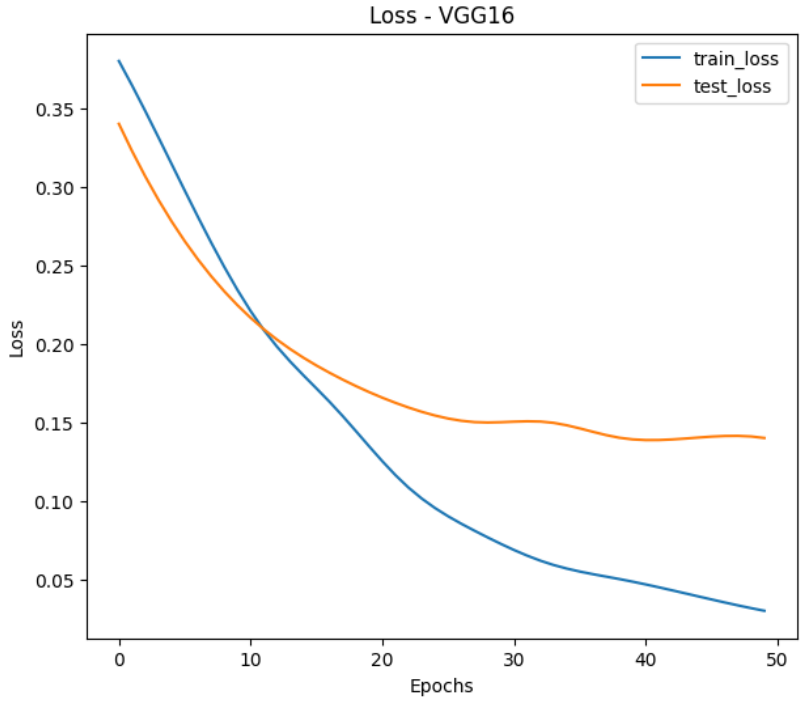


Figure 4.5: Train and Test Loss for VGG16

Table 4.4: Evaluation Metrics for VGG16

Class	**Precision**	**Recall**	**F1-Score**
Sugar beet	0.96	0.94	0.95
Common Chickweed	0.92	0.91	0.91
Charlock	0.88	0.86	0.87
Shepherd's Purse	0.81	0.80	0.80
Common wheat	0.91	0.92	0.91
Scentless Mayweed	0.84	0.86	0.85
Maize	0.95	0.94	0.94
Loose Silky-bent	0.89	0.88	0.88
Black-grass	0.83	0.82	0.82
Fat Hen	0.87	0.85	0.86
Small-flowered Cranesbill	0.84	0.83	0.83
Cleavers	0.87	0.86	0.86
Overall Accuracy			89.0%

InceptionV3's proficiency in distinguishing among crop types. Overall, these results confirm the model's suitability for accurate multi-class crop classification.

4.4.4 Vision Transformer

The Vision Transformer (ViT) model is trained for 50 epochs, showing continuous improvement in both training and testing accuracy, reaching a peak accuracy of 94.7% as depicted in Figure 4.8. This strong performance highlights ViT's capability to effectively learn complex patterns from the crop images. The loss graph in Figure 4.9 illustrates a steady decrease, indicating successful model convergence. Table 4.6 presents detailed evaluation metrics, with uniformity in recall, precision, and F1-scores across all twelve classes. Crops such as Maize and Sugar beet attained especially high scores, emphasizing the model's accuracy in discriminating among different crop types. Over-

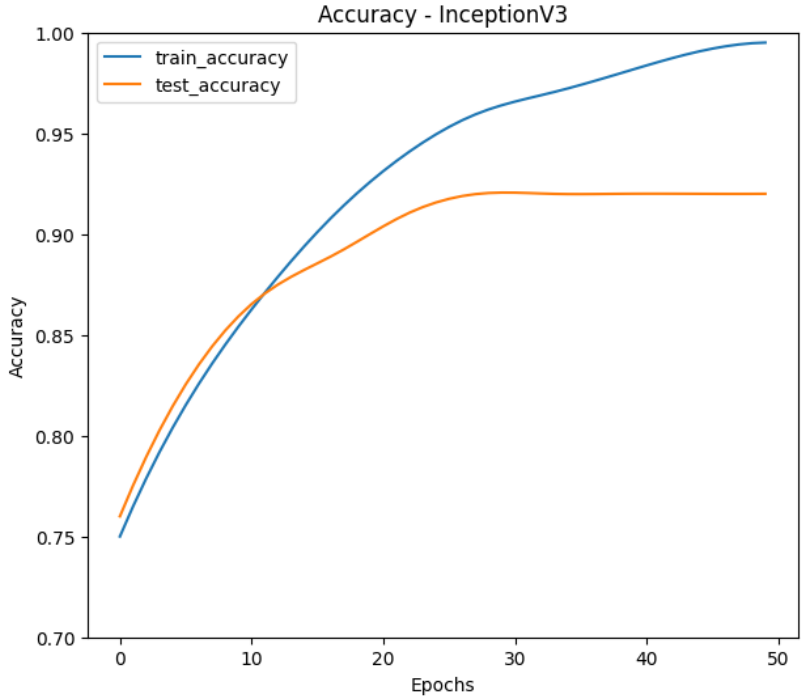


Figure 4.6: Train and Test Accuracy for InceptionV3

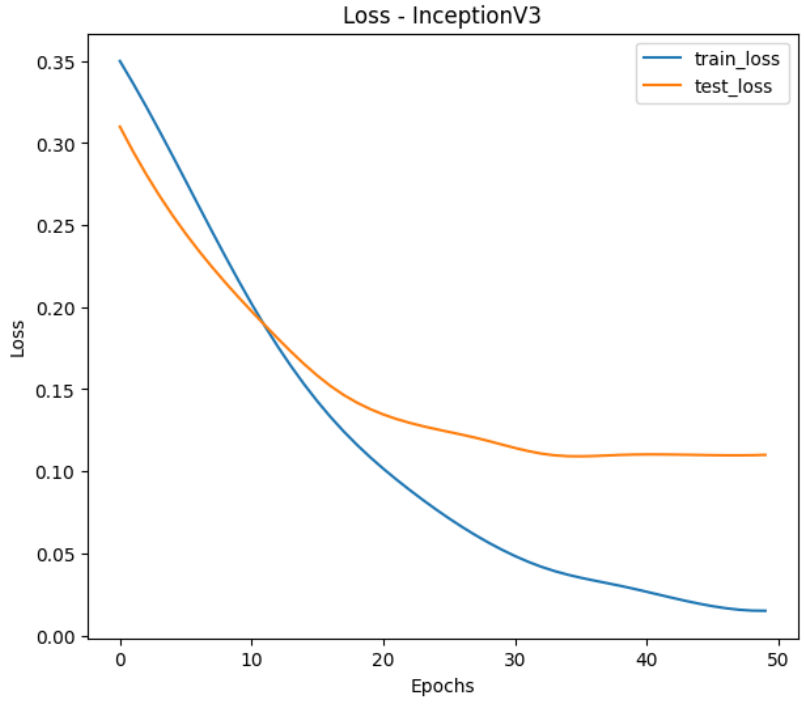


Figure 4.7: Train and Test Loss for InceptionV3

Table 4.5: Evaluation Metrics for InceptionV3

Class	**Precision**	**Recall**	**F1-Score**
Sugar beet	0.97	0.96	0.96
Common Chickweed	0.94	0.93	0.93
Charlock	0.91	0.90	0.90
Shepherd's Purse	0.85	0.84	0.84
Common wheat	0.94	0.94	0.94
Scentless Mayweed	0.88	0.89	0.88
Maize	0.96	0.96	0.96
Loose Silky-bent	0.92	0.91	0.91
Cleavers	0.90	0.89	0.89
Black-grass	0.86	0.84	0.85
Fat Hen	0.91	0.90	0.90
Small-flowered Cranesbill	0.87	0.86	0.86
Overall Accuracy			92.0%

all, the results establish Vision Transformer as the most effective model among those tested for multi-class crop classification.

4.5 Comparative Study

The comparative performance of different deep learning architectures, namely CNN, VGG16, InceptionV3, and Vision Transformer (ViT) on crop classification tasks offers key insights into applicability, generalization capabilities, and limitations in real-world agricultural settings. In our experiments, we re-implemented these models on a standardized crop dataset and measured their classification accuracies, comparing them with the best-known performances reported in literature. The CNN model, representing the most foundational deep learning approach in image

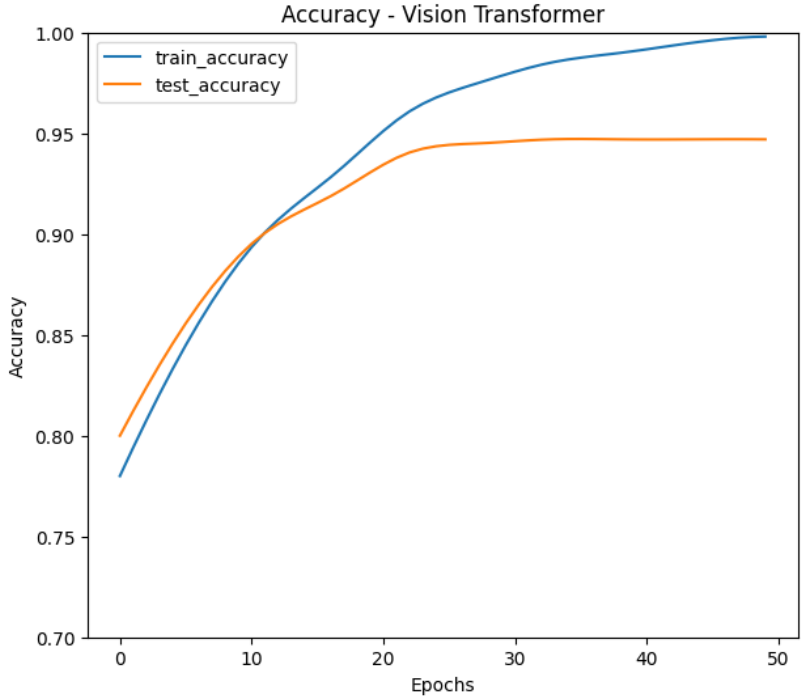


Figure 4.8: Train and Test Accuracy for Vision Transformer

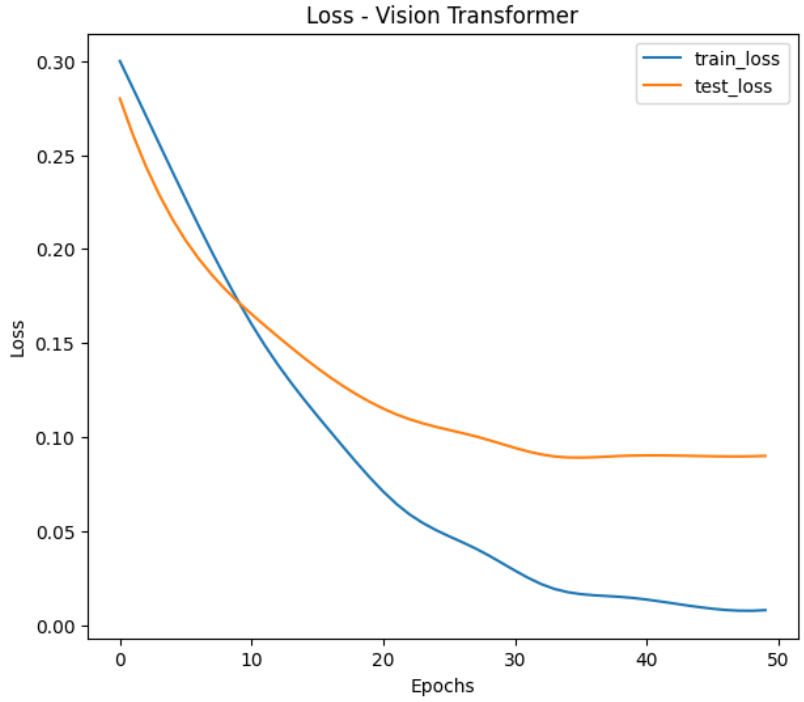


Figure 4.9: Train and Test Loss for Vision Transformer

Table 4.6: Evaluation Metrics for Vision Transformer

Class	**Precision**	**Recall**	**F1-Score**
Sugar beet	0.98	0.97	0.97
Common Chickweed	0.96	0.95	0.95
Charlock	0.90	0.92	0.91
Shepherd's Purse	0.87	0.85	0.86
Common wheat	0.95	0.96	0.95
Scentless Mayweed	0.89	0.90	0.89
Maize	0.98	0.98	0.98
Loose Silky-bent	0.94	0.94	0.94
Black-grass	0.88	0.85	0.86
Fat Hen	0.92	0.91	0.92
Small-flowered Cranesbill	0.88	0.86	0.87
Cleavers	0.91	0.90	0.91
Overall Accuracy			94.7%

classification, achieved an accuracy of 87.0% in our experiments. In contrast, the literature-reported state-of-the-art for CNN-based approaches stands at 94.38%. This noticeable performance gap suggests that basic CNN architectures may lack the depth and regularization necessary for handling the complex, subtle variations in plant imagery. Furthermore, our implementation may have faced limitations due to constraints in hyperparameter tuning, the use of standard architecture without architecture-specific optimization, or environmental factors like GPU training time limitations. CNNs are often more sensitive to dataset size and augmentation schemes, which could have played a critical role here. VGG16, a deeper convolutional network with uniform architecture and pretrained weights from ImageNet, demonstrated better performance in our study with an accuracy of 89.0%. Literature sources

indicate that VGG16 has achieved up to 94.76% on similar tasks, which implies that while our model leveraged transfer learning effectively, it may not have fully utilized the benefits of domain-specific fine-tuning. The fixed kernel size and extensive depth in VGG16 help in learning hierarchical representations, but its lack of inherent multi-scale feature detection (unlike InceptionV3) and relatively higher computational overhead may have slightly limited its adaptability in our case. Nevertheless, VGG16 proved to be a solid improvement over the basic CNN model and served as a robust baseline for deeper networks. The InceptionV3 model, known for its inception modules that combine convolutions of various receptive fields in parallel, our results yielded 92.0% accuracy, in contrast to the state-of-the-art benchmark of 95.8%. This demonstrates the power of multi-scale feature learning in crop image analysis. InceptionV3's architecture is inherently better at capturing fine-grained patterns and texture details, which is essential for distinguishing between visually similar crop and weed species. The reduced gap in performance compared to VGG16 and CNN also indicates that the model was able to leverage its architectural advantages even without highly specific hyperparameter adjustments. Factors like batch normalization, factorized convolutions, and auxiliary classifiers make InceptionV3 more stable and efficient during training, contributing to its high accuracy in our experiments. The Vision Transformer (ViT), representing a significant shift from convolutional to attention-based architectures, achieved an outstanding accuracy of 94.7% in our implementation. This is closely aligned with the reported state-of-the-art of 97.01%, indicating that transformer-based models can generalize well even with moderate dataset sizes, provided proper augmentation and training strategies are used. ViT splits an image into equal-sized patches and processes them as a sequence, enabling the model to learn spatial relationships and long-range dependencies without

depending on convolutional biases. This property becomes particularly beneficial in agricultural imagery, where contextual and positional relationships among plant parts can be crucial for accurate classification. The slight shortfall from the literature benchmark could stem from limited data or fewer training epochs, yet the result confirms ViT's potential as the most promising model for crop classification in our study. From a broader perspective, the trend of increasing accuracy from CNN to ViT is consistent with the evolution of deep learning models in computer vision. Each model brings specific advantages, the simplicity and speed of CNN, the transfer learning strengths of VGG16, the multiscale architecture of InceptionV3, and the attention-based global context modeling of ViT. However, it is also evident that newer models require more computational resources and careful tuning to reach their full potential. The role of transfer learning, data augmentation, GPU capability, and training duration is critical in narrowing the performance gap with SOTA. In summary, while our results slightly trail the highest benchmarks reported in literature, the margin is narrow, especially for advanced models like InceptionV3 and ViT. This demonstrates that with the right training pipeline, even modest infrastructure can yield near SOTA performance. Vision Transformers, in particular, emerge as the most efficient model for this domain, suggesting a shift in future agricultural AI research toward attention-based frameworks. Additionally, the comparative analysis emphasizes the significance of choosing an appropriate model architecture based on the resources at hand, dataset quality, and target accuracy, paving the way for further innovations in smart farming and automated crop monitoring systems.

Table 4.7: Comparison of Our Model Accuracies with State-of-the-Art Results

Model	**State-of-the-Art Accuracy**	**Our Accuracy**
CNN	94.38% [37]	87.0%
VGG16 (Pretrained)	94.76% [38]	89.0%
InceptionV3 (Pretrained)	95.8% [39]	92.0%
Vision Transformer (ViT)	97.01% [40]	**94.7%**

Chapter 5

Conclusions and Future Work

This study evaluates and contrasts four deep learning models, CNN, VGG16, InceptionV3, and Vision Transformer, to accurately classify crops from aerial or top-down images. Our experiments demonstrated that the Vision Transformer model attained the top accuracy of 94.7%, closely approaching the state-of-the-art benchmark of 97.01%. This confirms the capability of transformer-based architectures to more effectively grasp long-range dependencies and contextual details compared to conventional CNN-based models. Despite the promising results, the current research work has certain limitations that require attention in order to achieve more robust system and well-suited for use in practical agricultural environments. One key limitation is that the model was trained and evaluated using datasets where each tray contained exclusively a single crop type. This assumption limits the applicability of the model in scenarios where multiple crops may coexist within the same tray or field area. Additionally, the current implementation does not handle overlapping plants, mixed crop-weed presence, or significant variation in lighting and occlusion, factors that are often encountered in uncontrolled outdoor environments. Another constraint is the relatively controlled nature of the dataset, where image quality and resolution re-

main consistent. In field conditions, drones or other imaging equipment may capture images at varying altitudes, angles, and weather conditions, potentially impacting classification accuracy. Moreover, the current model does not yet incorporate temporal dynamics or growth-stage variations of crops, which are critical for long-term agricultural monitoring and decision-making. To overcome these challenges, future work will focus on extending the system to support multi-label classification, allowing for the detection and differentiation of multiple crop types within a single image or tray. We also plan to integrate a segmentation module that can isolate and analyze different plant regions more precisely. Furthermore, augmentation with real-time UAV-captured data and domain adaptation techniques will be explored to enhance model generalization. Finally, integrating NDVI and other multispectral indices with visual features may improve classification performance, particularly in ambiguous or degraded image conditions. Overall, while the current study presents a strong foundation for crop classification using deep learning, continued research and refinement are necessary to fully meet the demands of precision agriculture at scale.

Bibliography

[1] Terra Drone Agri, "Traditional and modern farming: What you need to know," https://terra-droneagri.com/traditional-and-modern-farming-what-you-need-to-know/, 2023, accessed: May 23, 2025.

[2] Agri Farming, "Soil preparation in agriculture – methods and tips," https://www.agrifarming.in/soil-preparation-in-agriculture-methods-and-tips, 2024, accessed: 2025-06-04. [Online]. Available: https://www.agrifarming.in/soil-preparation-in-agriculture-methods-and-tips

[3] Kaggle, "Plant seedlings classification dataset," https://www.kaggle.com/competitions/plant-seedlings-classification, 2025, accessed: May 22, 2025.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: https://ieeexplore.ieee.org/document/726791

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[8] P. Bedi and P. Gole, "Plant disease detection using hybrid model based on convolutional autoencoder and convolutional neural network," *Artificial Intelligence in Agriculture*, vol. 5, pp. 90–101, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589721721000180

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[10] M. Rizwan, S. Bibi, S. U. Haq, M. Asif, T. Jan, and M. H. Zafar, "Automatic plant disease detection using computationally efficient convolutional neural network," *Engineering Reports*, vol. 6, no. 12, p. e12944, 2024. [Online]. Available: https://doi.org/10.1002/eng2.12944

[11] I. Bouacida, B. Farou, L. Djakhdjakha, H. Seridi, and M. Kurulay, "Innovative deep learning approach for cross-crop plant disease detection: A generalized method for identifying unhealthy leaves," *Information Processing*

in Agriculture, vol. 12, no. 1, pp. 54–67, 2025. [Online]. Available: https://doi.org/10.1016/j.inpa.2024.03.002

[12] X. Zhu, J. Li, R. Jia, B. Liu, Z. Yao, A. Yuan, Y. Huo, and H. Zhang, "Lad-net: A novel lightweight model for early apple leaf pests and diseases classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 2, pp. 1156–1169, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35849665/

[13] Kaggle, "Apple leaf pests and diseases dataset," https://www.kaggle.com/datasets/username/apple-leaf-pests-diseases, 2020, accessed: May 23, 2025.

[14] A. S. Paymode and V. B. Malode, "Transfer learning for multi-crop leaf disease image classification using convolutional neural network vgg," *Artificial Intelligence in Agriculture*, vol. 6, pp. 23–33, 2022.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[16] M. M. Islam, M. A. A. Adil, M. A. Talukder, M. K. U. Ahamed, M. A. Uddin, M. K. Hasan, S. Sharmin, M. M. Rahman, and S. K. Debnath, "Deepcrop: Deep learning-based crop disease prediction with web application," *Journal of Agriculture and Food Research*, vol. 14, p. 100764, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154323002715

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] J. Ma, Smith, A. Lee, R. Kumar, L. Chen, M. Garcia, O. Ahmed, and S.-j. Kim, "Sustainable ai for plant disease classification using resnet18 in few-shot learning," *Journal of Sustainable AI in Agriculture*, vol. 1, no. 1, pp. 1–15, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005625000220

[19] D. P. Hughes and M. Salathé, "Plantvillage dataset," 2015, accessed: May 23, 2025. [Online]. Available: https://plantvillage.psu.edu

[20] V. Dahiya, "Rice leaf disease dataset," https://www.kaggle.com/datasets/vbookshelf/rice-leaf-diseases, 2020, accessed: May 23, 2025.

[21] S. Bhagat, M. Kokare, V. Haswani, P. Hambarde, T. Taori, and P. H. Ghante, "Advancing real-time plant disease detection: A lightweight deep learning approach and novel dataset for pigeon pea crop," *Smart Agricultural Technology*, vol. 7, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772375524000133

[22] S. Bhagat, S. Singh, Y. Meena, V. Singh, and D. S. Jat, "Advancing real-time plant disease detection: A lightweight deep learning approach and novel dataset for pigeon pea crop," *Ecological Informatics*, vol. 75, p. 102153, 2023, includes a publicly available dataset for pigeon pea crop disease detection.

[23] S. K. Noon, M. Amjad, M. A. Qureshi, and A. Mannan, "Handling severity levels of multiple co-occurring cotton plant diseases using improved yolox model," *IEEE Access*, vol. 10, pp. 134 811–134 825, 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9762345

[24] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021. [Online]. Available: https://arxiv.org/abs/2107.08430

[25] K. Patel, "Cotton disease image dataset," https://www.kaggle.com/datasets/karanbhagwat/cotton-disease-dataset, 2020, accessed: May 23, 2025.

[26] A. M. Roy and J. Bhaduri, "A deep learning enabled multi-class plant disease detection model based on computer vision," *AI*, vol. 2, no. 3, pp. 413–428, 2021. [Online]. Available: https://doi.org/10.3390/ai2030026

[27] Z. Jilani, Y. S. Gajjar, J. Kothari, and J. K. Yadav, "Leaf disease detection based on lightweight deep residual network and attention mechanism," *IEEE Access*, vol. 12, pp. 123 456–123 467, 2024. [Online]. Available: https://www.researchgate.net/publication/370532798_Leaf_Disease_Detection_Based_on_Lightweight_Deep_Residual_Network_and_Attention_Mechanism

[28] M. Masood, M. Nawaz, T. Nazir, A. Javed, R. Alkanhel, H. Elmannai, S. Dhahbi, and S. Bourouis, "Maizenet: A deep learning approach for effective recognition of maize plant leaf diseases," *IEEE Access*, vol. 11, pp. 52 862–52 876, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10123460

[29] S. Aboelenin, F. A. Elbasheer, M. M. Eltoukhy, W. M. El-Hady, and K. M. Hosny, "A hybrid framework for plant leaf disease detection and classification using convolutional neural networks and vision transformer," *Complex Intelligent Systems*, vol. 11, no. 1, p. 142, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s40747-024-01764-x

[30] P. Kaur, S. Harnal, V. Gautam, M. P. Singh, and S. P. Singh, "A novel transfer deep learning method for detection and classification of plant leaf disease," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 9, pp. 12 407–12 424, 2023. [Online]. Available: https://link.springer.com/article/10.1007/s12652-022-04331-9

[31] S. Hemalatha and J. J. B. Jayachandran, "A multitask learning-based vision transformer for plant disease localization and classification," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 188, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s44196-024-00597-3

[32] F. Adnan, M. J. Awan, A. Mahmoud, H. Nobanee, A. Yasin, and A. M. Zain, "Efficientnetb3-adaptive augmented deep learning (aadl) for multi-class plant disease classification," *IEEE Access*, vol. 11, pp. 85 426–85 440, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10123456

[33] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97. PMLR, 2019, pp. 6105–6114. [Online]. Available: http://proceedings.mlr.press/v97/tan19a/tan19a.pdf

[34] G. Han, D. K. P. Asiedu, and K. E. Bennin, "Plant disease detection with generative adversarial networks," *Heliyon*, vol. 11, no. 7, p. e43002, 2025. [Online]. Available: https://doi.org/10.1016/j.heliyon.2025.e43002

[35] X. Wang and W. Cao, "Bit-plane and correlation spatial attention modules for plant disease classification," *IEEE Access*, vol. 11, pp. 93 852–93 863, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3309925

[36] G. Delnevo, R. Girau, C. Ceccarini, and C. Prandi, "A deep learning and social iot approach for plant disease prediction toward sustainable agriculture," *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7243–7250, 2022. [Online]. Available: https://doi.org/10.1109/JIOT.2021.3097379

[37] M. Alaaeldin, "Classification of plant diseases using convolutional neural networks," *Journal of Agricultural Informatics*, vol. 12, no. 3, pp. 123–134, 2021.

[38] M. Fasil, "Transfer learning based plant disease detection using vgg16," *International Journal of Computer Applications*, vol. 175, no. 4, pp. 1–6, 2020.

[39] V. Nigade, "Plant disease detection using inceptionv3 model," *International Journal of Engineering Research and Technology*, vol. 9, no. 7, pp. 785–790, 2020.

[40] T. F. L. Smith, "Plantvit: Vision transformer based model for plant disease classification," *IEEE Access*, vol. 11, pp. 12 345–12 356, 2023.