Université Mohamed Khider de Biskra
Faculté des Sciences et de la Technologie
Département de génie électrique

MÉMOIRE DE MASTER

Sciences et Technologies
Génie électrique
Réseaux et télécommunications

Réf. : Entrez la référence du document

Présenté et soutenu par :
LOUAFI Malek , BENAYAD Rania

Le : mercredi 4 juin 2025

Leaf disease identification using deep learning

Jury :

Dr.	Houhou Ihssane	MCB	University of Biskra	Président
Dr.	Chouchane Ammar	MCA	University Center of Barika	Encadreur
Dr.	Zehani Soraya	MCA	University of Biskra	Examinateur

Année universitaire : 2024/2025

Université Mohamed Khider de Biskra
Faculté des Sciences et de la Technologie
Département de génie électrique

MÉMOIRE DE MASTER

Sciences et Technologies
Génie électrique
Réseaux et télécommunications

Réf. : Entrez la référence du document

Leaf disease identification using deep learning

Le : mercredi 4 juin 2025

Présenté par :

LOUAFI Malek

BENAYAD Rania

Avis favorable de l'encadreur :

Signature Avis favorable du Président du Jury

Cachet et signature

Acknowledgements

First and foremost, we express our deep gratitude to God, the Almighty and Merciful, for granting us the strength, patience, and perseverance to complete this work.

We wish to express my profound gratitude to our supervisor, ***Dr. Ammar Chouchane****, for his expert guidance, insightful advice, and continuous encouragement, which have been invaluable to the successful completion of this research.*

We are deeply indebted to our co-supervisor, ***Pr. Ouamane Abdelmalik****, for his generous dedication of time, patience, and exceptional mentorship, which greatly contributed to the development and refinement of this study.*

Furthermore, We are grateful to ***Ms. Bellili Sana*** *for her unwavering support and assistance during this endeavor. We also wish to acknowledge all individuals who have offered their encouragement and support throughout this journey.*

We are also deeply grateful to Professor ***Houhou Ihssane*** *for kindly agreeing to chair our examination committee, and to Professor* ***Zhani Souraia*** *for his generous participation as a member of the jury.*

Finally, We are deeply thankful to our family and friends for their steadfast moral and practical support, which has been a source of strength and inspiration.

Dedication

I dedicate this work with all my heart to those who have stood by me with love, patience, and unwavering support.

To my dear parents *Your sacrifices, prayers, and endless love have shaped who I am today. This achievement is as much yours as it is mine. May God bless you with health and happiness.*

To my dearest mother *You are the heart of my existence the one whose prayers, silent strength, and boundless love have guided me through every hardship and every triumph. Your sacrifices, often invisible and immeasurable, have paved the way for my education and success. Mama, rest your heart your little girl has finally become what you always dreamed she would be. May God bless you with peace, joy, and a long life.*

To my precious father *Your wisdom, patience, and quiet strength have shaped the person I am today. Thank you for the countless sacrifices you've made, often in silence. I hope this accomplishment brings pride to your heart, just as your presence has always brought peace to mine.*

To my siblings Mayar, Melina, Mariem, and Mekka *For your encouragement, laughter, and faith in me. Your support has been a constant source of strength.*

To my partner Rania *Thank you for being by my side since the very start of this journey. I'm truly thankful for everything we've shared.*

To my family and friends *Thank you for your presence, your kindness, and your belief in me throughout this journey. This work is a tribute to all of you.*

Louafi Malek

Dedication

I must wholeheartedly express my gratitude to all those who supported me throughout this journey. With profound love, deep respect, and eternal gratitude, I dedicate this humble work.

To my dearest mother *the one I miss every single day. There are moments in life when we wish we could bring someone back from heaven just to hear their voice, feel their presence, or say once more how much we love them. Mom, not a day goes by without thinking of you. I hope you are proud of the woman your daughter has become. May Allah have mercy on your soul and grant you the highest ranks in Jannah.*

To my beloved father *the man who carried the weight of two roles with courage and love. Your strength, your sacrifices, and your endless support have guided me to where I am today. Words will never be enough to thank you for being my rock and my inspiration.*

To my brother Achraf *thank you for being my steadfast companion. Your presence and support have meant more to me than you'll ever know.*

To my little sisters, Aridj and Tessnim *your love and innocence have been a light in my life and a motivation to keep moving forward.*

To my precious aunt Souad *who has always been like a second mother to me. Your love, kindness, and unwavering presence in my life are treasures I will forever cherish.*

To my partner Malek *with whom I've shared this journey from the very beginning. I am truly grateful, and I sincerely wish you all the success you deserve.*

Benayad Rania

Abstract

Plant diseases pose a significant threat to global agriculture and food security, necessitating accurate and timely detection for effective management. This master's thesis explores the efficacy of advanced deep learning architectures, specifically focusing on the emerging Vision Transformers (ViTs), for robust plant disease classification. It aims to critically compare the performance of a traditional Convolutional Neural Network (CNN), AlexNet, against two prominent ViT models, Simple ViT and Distillation ViT. The strategic comparison is conducted across two critical metrics: classification accuracy and model size, providing insights into both performance and computational efficiency. Evaluation is performed on two distinct datasets: the widely recognized BANANA dataset and the newly introduced Ziban Plant Disease Dataset (ZPDD). Notably, ZPDD, collected by the LI3C research laboratory at the University of Biskra. This research contributes valuable insights into the applicability and efficiency of Vision Transformers for plant disease classification, highlighting their potential as powerful tools in precision agriculture.

Keywords: Plant Disease Classification, Deep Learning, Vision Transformers (ViT), Convolutional Neural Networks (CNN), Ziban Plant Disease Dataset (ZPDD), Self-Attention, Image Classification, Precision Agriculture.

Résumé

Les maladies des plantes représentent une menace importante pour l'agriculture et la sécurité alimentaire mondiales, nécessitant une détection précise et rapide pour une gestion efficace. Ce mémoire de master explore l'efficacité des architectures avancées d'apprentissage profond, en se concentrant plus particulièrement sur les nouveaux Vision Transformers (ViT), pour une classification robuste des maladies des plantes. Il vise à comparer de manière critique les performances d'un réseau neuronal convolutif (CNN) traditionnel, AlexNet, à celles de deux modèles ViT importants, Simple ViT et Distillation ViT. Cette comparaison stratégique est menée sur deux indicateurs clés : la précision de la classification et la taille du modèle, fournissant des informations sur les performances et l'efficacité de calcul. L'évaluation est réalisée sur deux ensembles de données distincts : le jeu de données BANANA, largement reconnu, et le nouveau jeu de données Ziban Plant Disease Dataset (ZPDD). Notamment, le ZPDD, collecté par le laboratoire de recherche LI3C de l'Université de Biskra. Cette recherche apporte des informations précieuses sur l'applicabilité et l'efficacité des Vision Transformers pour la classification des maladies des plantes, soulignant leur potentiel en tant qu'outils puissants en agriculture de précision.

Mots-clés: Classification des maladies des plantes, apprentissage profond, transformateurs de vision (ViT), réseaux de neurones convolutifs (CNN), ensemble de données sur les maladies des plantes Ziban (ZPDD), auto-attention, classification d'images, agriculture de précision.

الملخص

تُشكل أمراض النبات تهديدًا كبيرًا للزراعة العالمية والأمن الغذائي، مما يتطلب الكشف الدقيق وفي الوقت المناسب لإدارة فعالة. تستكشف أطروحة الماجستير هذه فعالية هياكل التعلم العميق المتقدمة، وخاصةً محولات الرؤية الحديثة (ViT) ، في التصنيف الدقيق لأمراض النبات. وتهدف إلى مقارنة أداء الشبكة العصبية التلافيفية التقليدية CNN (AlexNet) بشكل نقدي مع أداء نموذجين رائدين من نماذج ViT وهما Simple ViT و Disttilation ViT. وتستند هذه المقارنة الاستراتيجية إلى مؤشرين رئيسيين: دقة التصنيف وحجم النموذج، مما يوفر نظرة ثاقبة على الأداء والكفاءة الحسابية. ويركز التقييم على مجموعتي بيانات متميزتين: مجموعة بيانات BANANA معروفة على نطاق واسع ومجموعة بيانات لأمراض النبات تم تطويرها مؤخرًا ZPDD ، التي جمعها مختبر أبحاث L13C في جامعة بسكرة. يوفر هذا البحث رؤى قيّمة حول قابلية تطبيق محولات الرؤية وفعاليتها في تصنيف أمراض النبات، مسلطًا الضوء على إمكاناتها كأدوات فعالة في الزراعة الدقيقة.

الكلمات المفتاحية: تصنيف أمراض النبات، التعلم العميق، محولات الرؤية (ViT) ، الشبكات العصبية التلافيفية (CNN) ، مجموعة بيانات أمراض النبات زيبان ZPDD، الاهتمام الذاتي، تصنيف الصور، الزراعة الدقيقة.

Contents

List of Figures

List of Tables

Abbreviation List

ViT: Vision Transformer

CNN: Convolution Neural Network

DL: Deep Learning

ML: Machine Learning

NLP: Natural Language Processing

PMVT: Plant-based Mobile Vision Transformer

FNN: Feedforward Neural Network

ReLU: Rectified Linear Unit

ZPDD: Ziban Plant Disease Dataset

SVM: Support Vector Machines

PCA: Principal Component Analysis

General Introduction

Plant diseases pose a significant threat to global agricultural productivity and food security, leading to substantial economic losses and environmental challenges. These phytopathological conditions are induced by a variety of biotic agents, including fungi, bacteria, viruses, and abiotic stressors [7], which compromise plant health and yield quality. Early and accurate detection of plant diseases is paramount to implementing effective management strategies and mitigating adverse impacts on crop production [8] .

Leaves serve as primary indicators for disease diagnosis due to their early manifestation of visible symptoms, making leaf image analysis a critical component in automated plant disease detection systems. Traditional machine learning and deep learning approaches, particularly Convolutional Neural Networks (CNNs) [9], have demonstrated considerable success in image- based disease classification tasks. However, CNNs are inherently limited in capturing long-range spatial dependencies and global contextual information, which are essential for nuanced feature extraction in complex plant pathology scenarios. Recent advancements in computer vision have introduced Vision Transformers (ViTs), which utilize self-attention mechanisms to model global interactions within image data effectively. ViTs have exhibited superior performance in various visual recognition tasks by enabling comprehensive feature representation beyond the local receptive fields of CNNs. Their application to plant disease detection represents a promising frontier, offering enhanced iagnostic accu-

racy and robustness.

Central to the remarkable success of Transformer models, and subsequently ViTs in computer vision, is the innovative self-attention mechanism [10]. Unlike CNNs that process information through localized receptive fields, self-attention allows each element of an input sequence (e.g., an image patch) to dynamically weigh the importance and relevance of all other elements within the sequence [11]. This global connectivity enables the model to capture long-range dependencies and intricate relationships across the entire image directly, from the very first layers of the network. The power of self-attention lies in its ability to dynamically establish a comprehensive contextual understanding of the entire input, leading to more robust and semantically rich feature representations, which is particularly beneficial for complex visual analysis tasks such as detailed plant disease symptom recognition.

This thesis is structured into three main chapters:

- **Chapter 1**: Establishes foundational knowledge by defining key concepts, disease typologies, and the inherent challenges in plant disease detection. It also reviews conventional and deep learning-based methodologies.
- **Chapter 2**: Provides a critical review of the state-of-the-art techniques in plant disease detection, emphasizing the theoretical underpinnings and practical implementations of Vision Transformer architectures in this domain.
- **Chapter 3**: Details the system design and experimental evaluation, including dataset acquisition, preprocessing, model development using distilled ViTs, and performance assessment. The chapter concludes with a discussion of the results and their implications for precision agriculture.

Through this work, we aim to contribute to the advancement of AI driven plant pathology diagnostics, facilitating scalable, accurate, and efficient disease detection systems that support sustainable agricultural practices and global food security.

Chapter I

Basic Concepts and Terminology for plant disease detection

I.1 Introduction

Plant diseases are a major agricultural challenge, affecting crop yield and quality. These diseases are caused by various factors, including fungi, bacteria, nematodes, and environmental conditions such as temperature, humidity, and soil PH . Traditional disease detection methods rely on manual observation, which is time-consuming and prone to errors. With technological advancements, automated detection using Machine Learning (ML) and Deep Learning (DL) has gained significant attention. These approaches enhance accuracy and efficiency, allowing for early disease identification and better crop management[1].

This chapter presents a foundational overview of plant diseases, their classification, and traditional methods of detection, followed by a detailed discussion on the role of deep learning particularly CNNs in plant disease diagnosis. It concludes by highlighting key datasets used in this research area and outlining the challenges that continue to motivate the development of more advanced models such as Vision Transformers [12].

I.2 Taxonomy of Plant Diseases: Pathogen Types and Symptomatic Patterns

Understanding the taxonomy of plant diseases is fundamental to accurately diagnosing, managing, and mitigating their impacts. Plant diseases arise from a diverse range of pathogens including fungi, bacteria, viruses, and nematodes each exhibiting distinct biological characteristics and modes of infection These pathogens interact with host plants in unique ways, giving rise to a wide spectrum of symptomatic patterns such as leaf spots, wilting, chlorosis, necrosis, and deformities. Classifying plant diseases based on both the type of causal agent and the observable symptoms not only enhances our comprehension of plant-pathogen dynamics but also facilitates

the development of targeted detection and control strategies. This section provides a detailed taxonomy of plant diseases like we see in Figure I.1, mapping pathogen types to their corresponding symptomatic manifestations to support precision in both research and practical applications [1].

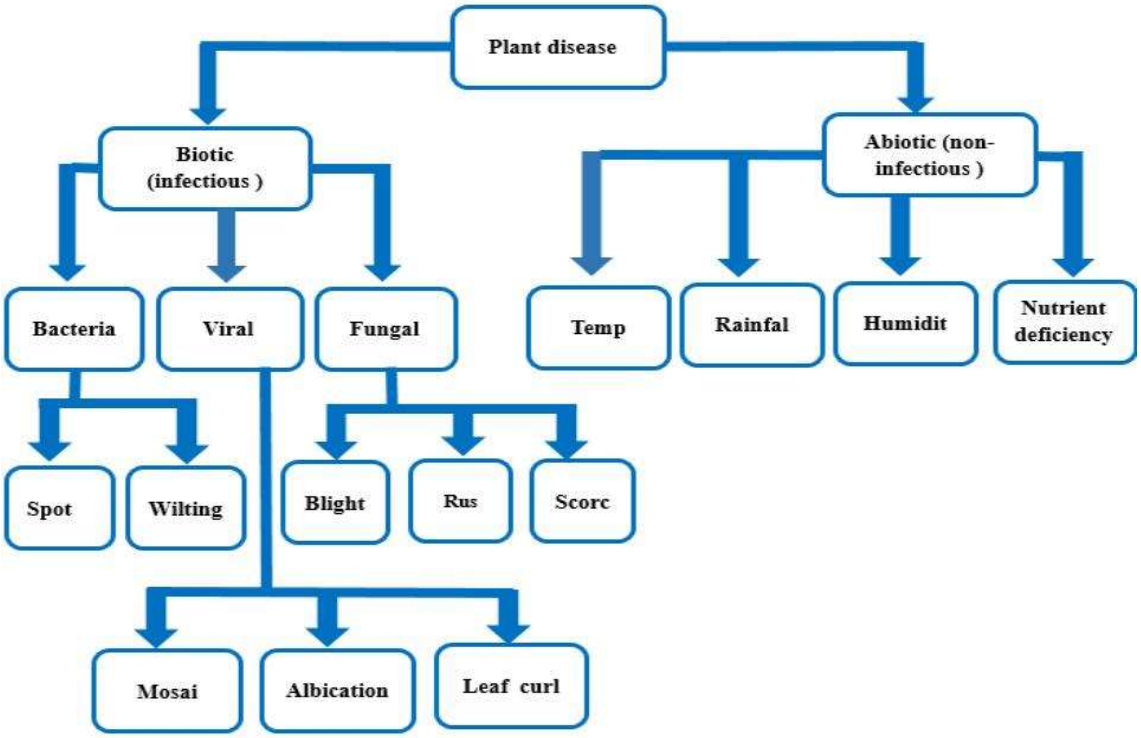


Figure I.1: Classification of plant disease in distinct categories[1]

I.2.1 Bacterial Diseases

Plant bacterial diseases manifest as water-soaked lesions, initially appearing as small green blemishes. These lesions progressively enlarge and eventually develop into dry, dead spots ,as shown in figure I.2 . A typical example is bacterial wilt in aubergine crops, where the entire plant succumbs to the disease . The leaves often exhibit water-soaked black spots, brown leaf spots, or yellow halos of uniform size, particularly under dry conditions [1].

Figure I.2: bacterial blemish

I.2.2 Viral Diseases

Viral diseases are among the most challenging to diagnose in plants. The presence of viruses in plants may show no visible symptoms or may resemble damage caused by herbicide exposure or nutrient deficiencies. Common viral diseases are transmitted by vectors such as beetles, leafhoppers, aphids, and whiteflies. An example of a viral disease is the mosaic virus, which causes green or yellow striping on the foliag ,as shown in figure I.3 [1].

Figure I.3: Viral Mosaic

I.2.3 Fungal Diseases

Fungal infections affect various plant components, leading to a range of diseases, including sclerotium wilt, common crown rot, stem rust, eyespot (on the sheath or stem), rust, blight (on leaves), ergot (on spikes), and carnal bunt or black point (on seeds). Phytophthora, the fungus responsible for late blight, causes gray-green le-

sions on older leaves . This fungus thrives under fluctuating wet and dry conditions, exacerbated by climate change. As the disease progresses, the lesions darken, and white fungal growth appears on the surface . Alternaria fungus, responsible for early blight, results in small brown lesions with a characteristic bulls-eye pattern of concentric rings ,as shown in figure I.4 . Rust fungi typically infect mature leaves, creating lesions with a yellow-green coloration that eventually turns black [1].

Figure I.4: Early Blight

I.2.4 Nematode-Related Diseases

Nematodes are microscopic, worm-like organisms, some of which are plant parasites[13] . Plant-parasitic nematodes, such as root-knot nematodes (Meloidogyne spp.), cyst nematodes (Heterodera spp.), and lesion nematodes (Pratylenchus spp.), cause significant damage to crops by feeding on roots. This results in stunted growth, wilting, reduced yield, and sometimes plant death. They also create entry points for secondary infections by fungi and bacteria, as show in figure I.5.

Figure I.5: Nematode-Related Diseases

I.2.5 Parasitic Plant Diseases

Parasitic plants, such as Striga (witchweed), Orobanche (broomrape), and Cuscuta (dodder), attach themselves to host plants and siphon off water and nutrients. These parasitic species can severely limit agricultural productivity, especially in tropical and subtropical regions. Symptoms include wilting, nutrient deficiency, stunted growth, and even plant death [14].

I.2.6 Physiological Disorders (Non-infectious)

Physiological disorders are plant problems caused by non-living (abiotic) factors, such as nutrient deficiencies, water stress, extreme temperatures, or toxic chemicals. These disorders differ from diseases caused by pathogens and include issues like blossom end rot in tomatoes (due to calcium deficiency), sunscald, frost injury, and chlorosis from iron deficiency. They can reduce crop yield and quality significantly [15] [16].

I.2.7 Diseases Affecting Tomatoes

Tomatoes are the most widely consumed vegetable globally as show in Table I.1 . In 2017, approximately 182 million tons were produced worldwide, accounting for 17% of total global vegetable production. This figure places tomatoes ahead of other commonly consumed vegetables, such as onions, which constitute around 9% of the global vegetable output. In countries where the Global Alliance for Improved Nutrition (GAIN) maintains its primary offices, the availability of tomatoes varies significantly. For instance, the average per capita supply ranges from roughly one-tenth of a medium-sized tomato (weighing approximately 60 grams) per person per week in Ethiopia, to four medium-sized tomatoes per person per week in India, and up to twelve medium-sized tomatoes per person per week in the United States .

Tomato cultivation occurs throughout all seasons, though it is most commonly

Table I.1: Tomato production by continent in 2022[6].

Continent	Production of Tomatoes in 2022 (Tons)
Africa	22.925.008
Americas	23.435.132
Asia	118.919.704
Europe	20.455.777
Oceania	372.351

grown during the winter and summer periods. The crop is sensitive to extreme frost and thrives optimally at an average monthly temperature between 21°C and 23°C. However, it can be commercially cultivated within a broader temperature range of 18°C to 27°C. Environmental factors such as temperature and light intensity significantly influence fruit pigmentation, fruit set, and the nutritional quality of the produce.

These environmental conditions also render tomato plants highly susceptible to a range of diseases caused by pathogenic fungi, bacteria, and viruses. Consequently, the early detection of plant diseases has become a critical area of research in agricultural science. Fungal infections typically manifest through morphological alterations in the leaves during the initial stages of disease development. In contrast, bacterial pathogens, which are considered more primitive than fungi and generally possess simpler life cycles, also induce detectable morphological changes in the foliage. here are the most disease that affect tomatoes [17].

1. **Early Blight :** Early blight is a common fungal disease in tomatoes that affects foliage at any growth stage. It begins as small black spots, primarily on older leaves, which enlarge and develop characteristic concentric rings resembling a bull's-eye. Surrounding tissue may yellow, and under high temperature and humidity, extensive leaf damage can occur. Stem lesions resemble those on leaves and may girdle the plant near the soil line. The fungus can also infect fruits, typically through the calyx or stem, forming large lesions with similar concentric rings that can cover most of the fruit. Infected transplants may die

shortly after being planted in the field .

2. **Bacterial Leaf Spot :** The disease thrives in moist conditions and following heavy rainstorms, which often trigger outbreaks. Infected leaves develop small, circular, water-soaked brown spots with yellow halos, primarily affecting older foliage and potentially causing significant defoliation. On green fruits, symptoms begin as small water-soaked spots that enlarge into light brown, sunken lesions with rough, scabby surfaces. Ripe fruits are not affected. The bacteria can contaminate seed surfaces and persist on alternate hosts, volunteer tomato plants, and infected plant debris .

3. **Tomato Mosaic Virus(TMV) :** TMV is characterized by a mottled appearance on the leaves, displaying alternating light and dark green patterns. This is often accompanied by wilting of young foliage, particularly under high sunlight exposure during the early stages of infection. Infected leaves typically exhibit distortion, puckering, and reduced size, with some showing characteristic "fern leaf" symptoms due to indentations. Infected plants tend to be stunted, pale green, and slender in appearance. The virus is primarily transmitted through mechanical means, including contact with contaminated hands, clothing, agricultural tools, plant debris, and through direct interaction between infected and healthy plants .

I.3 Traditional Methods for Plant Disease Identification

Traditional methods encompass long-established practices utilized by farmers, agronomists, and plant pathologists for the identification of plant diseases. These techniques typically rely on the visual assessment of symptomatic manifestations, such as leaf discoloration, wilting, and deformities; laboratory-based diagnostic procedures, including microscopy, serological assays, and molecular testing , as well as the application

of expert knowledge and experience,we see them in Figure I.6 .

Owing to their simplicity, cost-effectiveness, and ease of implementation in field conditions, traditional methods have remained fundamental in agricultural disease management for decades. However, despite their practicality, these approaches are often constrained by limitations such as subjectivity in diagnosis, the need for specialized expertise for accurate interpretation, delays in detection due to the late appearance of visible symptoms, and restricted scalability in diverse and complex agricultural environments [2].

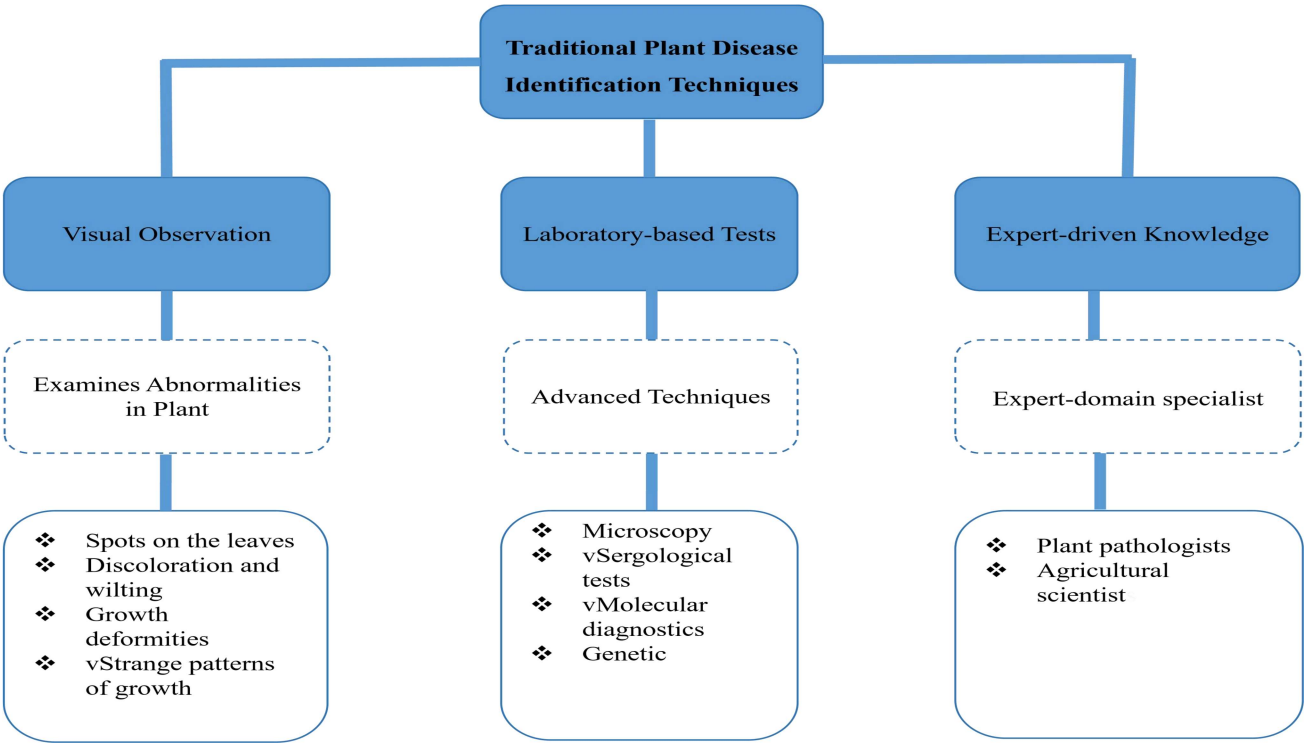


Figure I.6: Traditional Methods for Plant Disease Identification [2].

I.3.1 Visual Symptom Observation

Visual symptom observation is one of the most traditional and widely practiced methods for identifying plant diseases[8]. This approach involves the careful examination of plants for visible signs of infection, such as leaf spots, wilting, discoloration, deformities, and abnormal growth patterns. It offers several advantages, particularly its ease of application, as it does not require specialized equipment or advanced

technical skills. Additionally, it provides rapid, cost-effective, and on-site disease detection, making it highly practical for farmers and field workers.

However, visual symptom observation also presents notable limitations. The diagnosis is inherently subjective, relying heavily on human interpretation, which introduces the possibility of bias and diagnostic variability. Moreover, this method lacks precision, as many different diseases can produce similar external symptoms, making accurate differentiation challenging.

Another significant drawback is the potential delay in disease detection, as many pathogens only produce visible symptoms at advanced stages of infection. Furthermore, visual assessment alone often fails to reveal critical information about the specific pathogen involved or the underlying cause of the symptoms, necessitating further laboratory confirmation for definitive diagnosis [2].

I.3.2 Laboratory-Based Tests

Laboratory-based diagnostic methods represent a critical advancement in the identification of plant diseases [18], relying on the use of sophisticated techniques such as microscopy, serological assays, molecular diagnostics, and genetic sequencing to detect and characterize plant pathogens.

These approaches offer significant strengths, particularly their high degree of accuracy and their ability to identify infections at early stages, often before any visible symptoms manifest. Furthermore, laboratory techniques enable the precise differentiation between closely related pathogens, providing detailed insights essential for effective disease management and control strategies. Despite their advantages, these methods also present considerable limitations. Laboratory diagnostics are often expensive, requiring access to specialized equipment, reagents, and highly trained personnel, which can impose financial and logistical barriers, especially in developing or resource-constrained regions.

The procedures are also time-consuming, involving multiple steps such as sample collection, preparation, incubation, and data analysis, which can delay critical management decisions [8].

Additionally, the success of laboratory diagnostics heavily depends on the proper handling and preservation of samples; any mishandling during transportation or preparation can compromise the reliability and accuracy of the results. Consequently, while laboratory-based methods are powerful tools for plant disease identification, their accessibility, cost, and operational complexity limit their widespread application in field conditions [2].

I.3.3 Expert Knowledge-Driven Approaches

Expert knowledge-driven approaches to plant disease identification depend heavily on the expertise and experience of plant pathologists, agronomists, and agricultural scientists [19]. These methods typically involve the application of structured diagnostic frameworks, such as decision trees, symptom-based keys, or manual evaluation, allowing specialists to systematically assess plant health and determine the likely causes of disease. One of the primary strengths of this approach lies in its transparency and interpretability, as the diagnostic process and reasoning behind disease identification can be clearly understood and communicated to farmers and stakeholders [20].

Additionally, expert-based systems are highly adaptable, allowing them to be updated and refined as new diseases emerge or as scientific understanding advances.

However, despite their advantages, expert knowledge-driven methods face several notable challenges. Their scalability is limited, as access to trained experts may not be available in all regions, particularly in remote or resource-limited settings. Furthermore, the diagnostic process is susceptible to human error and subjective biases, which can compromise the consistency and accuracy of disease identification. Com-

plex cases, where multiple diseases interact with diverse environmental factors, often pose difficulties for manual evaluation, as it is challenging to model and interpret the intricate relationships involved. Moreover, these approaches may struggle to recognize newly emerging or less-studied diseases, limiting their effectiveness in rapidly evolving agricultural landscapes. As a result, while expert knowledge remains a valuable asset in plant pathology, it is increasingly complemented by data-driven and automated diagnostic systems to overcome these inherent limitations [2].

I.4 Role of Deep learning and CNNs in plant disease detection

With the growing global population placing increased pressure on the agricultural sector to produce food more sustainably, DL is proving to be a powerful tool for enhancing productivity while addressing environmental, economic, and social challenges. As a branch of artificial intelligence, DL is playing a vital role in smart agriculture by analyzing large, complex datasets to deliver valuable insights that support informed decision-making. Its ability to drive precision farming and automated field operations has shown great promise in improving crop yields, optimizing resource use, and promoting sustainable practices.

However, to fully realize its potential, key obstacles such as limited data availability, lack of model resilience, accessibility issues for end-users, and ethical considerations must be addressed. The integration of DL with other advanced digital technologies holds the key to advancing sustainable agriculture and strengthening global food security [21].Figure I.7 represents an example of Deep Learning Architecture model based CNN with different layers.

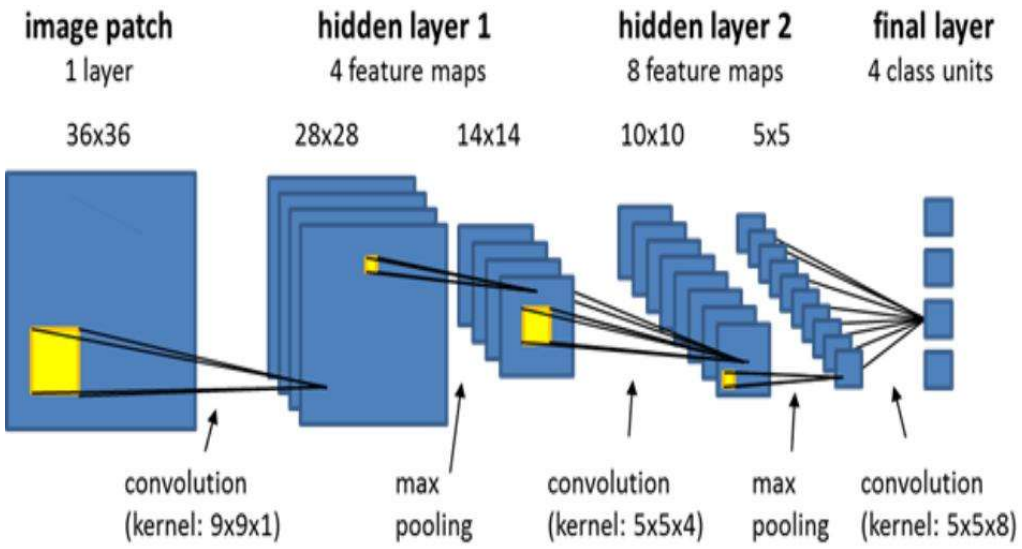


Figure I.7: Deep Learning Architecture [3].

I.4.1 Understanding Deep learning based on CNN

Convolutional Neural Networks represent a fundamental entry point to understanding deep learning, particularly in the field of computer vision. By mimicking the way humans perceive visual patterns, CNNs enable automatic feature learning, image analysis, and classification, making them a powerful and widely used architecture in deep learning-based applications.

1. **Convolutional Layer (CONV)** At the core of CNNs, convolutional layers execute crucial operations. These layers utilize kernels or filters to perform convolution operations, adjusting horizontally and vertically based on the stride rate [22] . The convolutional layer incorporates non-linear activation functions, with Rectified Linear Unit (ReLU) being the most widely used. ReLU enhances the network's ability to capture complex patterns by introducing non-linearity [23]. This figure I.8 visually demonstrates the fundamental concept of the convolution operation used in CNNs. The process starts with an input image, which is a matrix of pixel values. This patch is the same size as the kernel (or filter), which contains predefined or learnable weights. In the example, a 3×3 kernel

slides over the input image. For each position, an element-wise multiplication is performed between the kernel and the image patch, and the results are summed to produce a single output value. In this case, the computed value is 31, which becomes a pixel in the output feature map [24]. This operation enables CNNs to detect local patterns such as edges, textures, or corners [25] [26] [27]. This process forms the basis of feature extraction in CNNs, which is later passed through deeper layers for classification tasks, such as plant disease detection [28] [29] [24] .

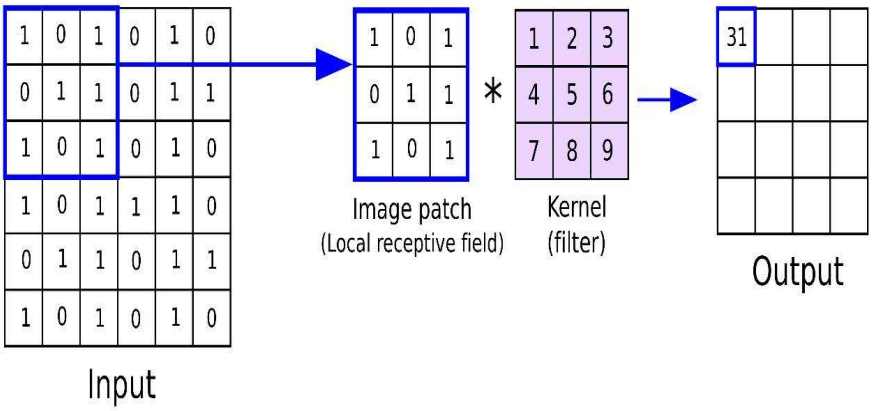


Figure I.8: an example maxpooling process [4]

2. **Pooling Layer (POOL)** After the convolution operation, the resulting feature map typically passes through a pooling layer, which serves to reduce the spatial dimensions of the data while preserving the most important information. Pooling layers help to decrease computational complexity, prevent overfitting, and enhance model robustness by making the representation more invariant to small translations and distortions in the input [30].

   The most commonly used pooling method is Max Pooling, where a filter (e.g., 2×2) slides over the input feature map and outputs the maximum value within each patch. This operation effectively down-samples the input while retaining the most dominant features. Another method is Average Pooling, which com-

putes the average of the values in the patch instead of the maximum. Although less common in modern architectures, average pooling can be useful when a smoother representation is desired. This figure I.9 illustrates an example max-pooling process [30].

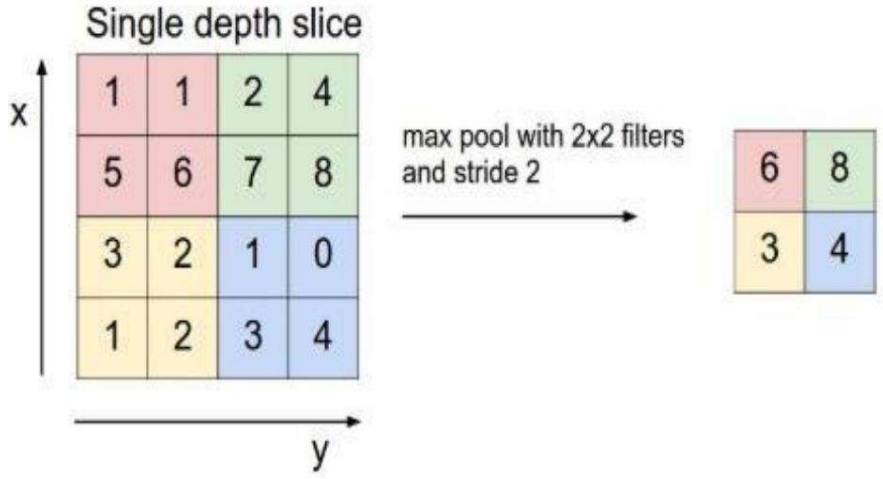


Figure I.9: Example of Maxpool operation [4]

3. **Fully Connected Layers (FC)** Operating on a flattened input, the fully connected layer connects each input to every neuron, initiating the classification process. Typically located near the network's end, FC layers perform mathematical operations, finalizing the classification task [31].

Beyond these layers, additional components contribute to CNN functionality. The activation function influences the classification outcome, especially in the last fully connected layer. The softmax function, often employed, normalizes output values to represent target class probabilities. Dropout layers prevent overfitting during training, nullify specific neurons' contributions, and promote more robust learning across different training data batches. Understanding these building blocks forms the basis for exploring popular CNN architectures such as ImageNet, VGG-16, VGG-19, etc., each tailored to specific projects within the expansive domain of deep learning algorithms [31].

I.4.2 Classification Paradigms in CNN-Based Plant Disease Detection

In the field of plant disease detection,CNNs have Emerged as a powerful tool for classifying plant leaf images by automatically identifying disease-related patterns. Unlike traditional methods that rely on handcrafted features, CNNs extract hierarchical representations directly from input images, enabling robust classification across a wide range of conditions [32]. This classification can be structured into several types based on the output labels [33] :

1. **Binary classification:** Involves distinguishing between two classes, such as healthy vs. diseased leaves. CNNs are often trained to detect the presence or absence of a specific plant disease.

2. **Multi-class classification:** Deals with more than two class labels, where a CNN can identify which disease a plant leaf suffers from among multiple possible diseases.

3. **Multi-label classification:** A single leaf image may simultaneously exhibit symptoms of multiple diseases, and CNNs can be adapted to assign multiple labels per instance.

4. **Imbalanced classification:** This denotes classification tasks where the distribution of examples across classes is unequal, as seen in medical diagnostic tests.

The classification of plant leaves typically follows a methodology wherein a classifier is initially trained using a set of images. Subsequently, this trained classifier is employed to identify diseases in new, unseen images. Machine learning methodologies are integral to this process, primarily categorized as either supervised, relying on labeled data, or unsupervised, operating on unlabeled data. Furthermore, certain approaches integrate aspects of both supervised and unsupervised learning [34].

I.4.2.1 Supervised CNN-Based Approaches:

Supervised classifiers rely on labeled data for training, where each input is associated with a known output. These models learn from this data to make predictions on unseen examples. Examples of supervised classifiers include :

- **Neural Networks:** These deep learning models are employed to classify plant images based on their health status, learning to distinguish between healthy and diseased plants [35].
- **Decision Trees:** Decision trees classify data by recursively partitioning it into subgroups according to decision rules. This method is often used to differentiate between healthy and diseased plants by evaluating various features of the data.
- **Support Vector Machines (SVMs):** SVMs work by identifying an optimal hyperplane that separates data into different classes. They are commonly applied to classify plant health, utilizing data from imaging and spectrometry to distinguish between healthy and diseased plants [36].

I.4.2.2 Unsupervised and Hybrid CNN Approaches:

Unsupervised classifiers do not require labeled data. Instead, they group data based on inherent similarities or patterns within the dataset. Examples of unsupervised classifiers include :

- **Clustering Algorithms:** These algorithms group data into clusters with similar characteristics. In plant disease detection, clustering is used to identify groups of plants exhibiting similar symptoms, aiding in disease detection.
- **Principal Component Analysis (PCA):** PCA reduces the dimensionality of large datasets by identifying the most significant features. It is commonly used in visualizing plant data, identifying patterns, and aiding in the classification of plants by health status.

- **Autoencoder Neural Networks:** These models learn to extract the most relevant features from input data. Autoencoders are particularly useful for detecting complex patterns in unlabeled plant data, facilitating disease detection through the identification of subtle abnormalities.

I.4.3 Performance Evaluation of CNN Models in Plant Disease Studies

The integration of deep learning techniques into plant disease detection and classification addresses the inherent limitations of manual feature extraction, particularly those related to the subjective selection of disease spot characteristics. By automating the feature learning process, DL models enable a more consistent, objective, and high-dimensional representation of disease patterns, thereby enhancing diagnostic accuracy and facilitating scalable, data-driven solutions. In the existing body of research, the effectiveness of these models is commonly evaluated using well-established performance metrics such as Accuracy, Precision, Recall, and F1-Score, which collectively provide a comprehensive assessment of classification robust [37].

- **Accuracy :** Accuracy evaluates the overall performance of the model by determining the proportion of correctly predicted instances relative to the total number of instances [5].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{I.1}$$

- **Precision :**It quantifies the proportion of true positive predictions among all instances that were predicted as positive and is also referred to as the Positive Predictive Value [5].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{I.2}$$

- **Recall :** It measures the proportion of correctly identified positive instances among all actual positive cases and is also known as Sensitivity or the True

Positive Rate [5].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{I.3}$$

- **F1 Score :** It represents the harmonic mean of Precision and Recall, offering a unified metric that balances the two. This measure is especially valuable in scenarios involving imbalanced class distributions [5].

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{I.4}$$

I.5 Common Datasets for Plant Disease Classification

Manual visual inspection of plant diseases is often hindered by several limitations, including low efficiency, high labor costs, and susceptibility to human error and subjectivity. To address these challenges, researchers have increasingly turned to machine learning techniques, which offer higher diagnostic accuracy, scalability, and automation in disease detection [38]. A critical component of these techniques is the availability of high-quality annotated datasets, which serve as the foundation for training, validating, and testing predictive models. Following standard practice in the field, datasets are typically divided into 80% for training, 10% for validation, and 10% for testing. The next section highlights some of the most widely adopted and benchmarked datasets used in plant.

I.5.1 PlantVillage dataset

The PlantVillage dataset is a widely used open-access resource for plant leaf disease recognition, available at www.plantvillage.org. It includes approximately 54,305 images covering 14 crop types and 20 disease types, organized into 38 classes12 healthy and 26 diseased. The images were captured under controlled conditions with a uniform background and a resolution of 256×256 pixels. In this study, a modified

version of the dataset is used, featuring segmented leaves with backgrounds removed to enhance analysis [39].

I.5.2 Taiwan tomato dataset

The Taiwan tomato disease dataset initially consisted of 622 leaf images (227 × 227 pixels) across six classes: Bacterial spot, Black mold, Gray spot, Healthy, Late blight, and Powdery mildew. To enhance model evaluation, the dataset was expanded through augmentation to a total of 4,976 images [5].

I.5.3 PlantDoc dataset

PlantDoc is a dataset comprising 2,569 images spanning 13 plant species including apple, bell pepper, blueberry, cherry, corn, grape, peach, potato, raspberry, soybean, squash powder, strawberry, and tomato and covering 30 classes representing both healthy and diseased conditions. It is designed for use in image classification and object detection tasks [40].

I.5.4 FieldPlant dataset

The FieldPlant dataset consists of 5,170 annotated leaf images captured from plantations in Cameroon, focusing on plant diseases affecting three major tropical crops: corn, cassava, and tomato. Notably, it is the first publicly available dataset to include annotated cassava images for disease detection. This dataset is well-suited for training effective plant disease detection models using field images and object detection techniques [40].

I.6 Challenges in Automated Plant Disease Detection

Despite the significant advancements in using (ML) and (DL) for plant disease detection, several challenges persist, limiting their practical implementation in real-world agricultural settings. These challenges are both technical and environmental, and they must be addressed to fully realize the benefits of automated disease detection systems.

I.6.1 Datasets Limitations and Variability

The performance of machine learning models in plant disease detection strongly depends on the availability of large, diverse, and well-annotated datasets. Collecting such datasets is challenging due to factors like seasonal changes, diverse plant species, and varying environmental conditions. This scarcity often results in models that perform well in controlled environments but fail to generalize under real-world conditions. Techniques such as data augmentation and synthetic image generation have been explored to mitigate this issue, but ensuring that these methods effectively capture real-world variability remains a significant concern [41].

I.6.2 Fine-Grained Disease Recognition

A wide range of plant diseases exhibit visually similar symptoms, posing challenges for accurate classification differences in growth stages, lighting conditions, and environmental stressors further complicate accurate identification. Advanced deep learning techniques, including attention mechanisms and multi-scale feature extraction, have been proposed to enhance the precision of disease recognition [42].

I.6.3 Environmental and Image Quality Constraints

Real-world field conditions pose numerous challenges, including inconsistent lighting, occlusions from other plant parts, and background noise, each of which can significantly degrade image quality. These factors adversely affect the robustness and accuracy of detection models. Implementing robust pre-processing techniques and developing models capable of adapting to these variations are essential steps toward reliable disease detection .

I.6.4 Computational Efficiency and Real-Time Processing

Deep learning models typically demand significant computational resources,which poses substantial challenges for deploying real-time applications in agricultural environments, particularly in resource-constrained settings. Balancing model complexity with computational efficiency is therefore essential. To address these limitations, various strategies are being investigated, including model optimization, compression techniques, and the integration of edge computing devices, all aimed at enabling real-time, on-site plant disease detection [43].

I.6.5 Generalization Across Diverse Conditions

Models trained on domain-specific datasets often exhibit limited generalization capability across diverse geographic regions, crop types, and environmental conditions. Such limitations frequently lead to reduced prediction accuracy when applied beyond their original training context. Therefore, enhancing the generalization ability of plant disease detection models is critical to ensuring their scalability and real-world applicability .

I.6.6 Integration with Existing Agricultural Practices

Integrating automated plant disease detection systems into current agricultural practices necessitates intuitive interfaces and compatibility with established workflows. However, farmers may encounter difficulties in adopting such technologies, often due to limited technical expertise or resistance to change. Thus, offering comprehensive training programs and continuous support is essential to facilitate successful adoption and long-term utilization [43].

I.6.7 Economic and Resource Constraints

The implementation of advanced plant disease detection systems can be financially burdensome, particularly for smallholder farmers. Ensuring the affordability and accessibility of these technologies is essential to guarantee equitable benefits across diverse agricultural communities. Ongoing research is focused on developing cost-effective and scalable solutions to mitigate this challenge. Effectively overcoming these constraints requires a multidisciplinary approach that combines innovations in machine learning, agronomy, and sensor technology to create robust, efficient, and widely deployable systems for automated plant disease detection [43] .

I.7 Conclusion

This chapter has laid the groundwork for understanding the complexity of plant disease identification and the evolution of techniques used to address this challenge. This field has seen significant progress. Despite the success of CNNs in image-based plant disease detection, challenges such as data variability, generalization to real-world conditions, and resource efficiency remain. These limitations pave the way for exploring more robust and adaptable models, such as Vision Transformers, which are discussed in subsequent chapters.

Chapter II

State of the art of plant disease detection and VITs

II.1 Introduction

The success of Convolutional Neural Networks in image-based classification tasks has inspired numerous applications in agriculture, particularly in plant disease detection [25]. However, CNNs face challenges in modeling global contextual information and long-range dependencies in images key factors when distinguishing between visually similar plant diseases. To address these limitations, the computer vision community has recently adopted Vision Transformers, a class of models originally developed for natural language processing [44]. ViTs bring a paradigm shift to visual recognition tasks by leveraging self-attention mechanisms that capture global relationships between image regions, without relying on local receptive fields like CNNs. This chapter explores the emergence, architecture, and application of Vision Transformers, with a focus on their role in plant disease classification. It also highlights the evolution from CNN based approaches, the key architectural elements of ViTs, and their integration into future agricultural technologies[45].

II.2 Overview of Vision Transformers

The integration of transformer architectures into computer vision has revolutionized image analysis tasks. Originally designed for natural language processing, transformers have been adapted to handle visual data, leading to the development of ViTs. This adaptation has opened new avenues for modeling complex visual patterns, particularly in fields like plant disease detection. Unlike traditional CNNs that process images using local receptive fields, ViTs treat images as sequences of patches. This approach allows the model to capture both local and global dependencies, enhancing its ability to recognize intricate disease patterns on plant leaves. Central to the ViT architecture is the self-attention mechanism, which enables the model to weigh the importance of different patches relative to each other. This mechanism facili-

tates a comprehensive understanding of the image, allowing for more accurate disease identification. The ViT architecture comprises several key components: patch embedding, which converts image patches into vector representations, positional encoding, which retains spatial information, and transformer encoders, which process the sequence of embedded patches. These components work in tandem to analyze the image holistically. The adoption of ViTs signifies a paradigm shift in image understanding, moving from localized feature extraction to a more global perspective. This shift has proven particularly beneficial in precision agriculture, where accurate and timely detection of plant diseases is crucial for crop management and sustainability [46][47].

II.2.1 Emergence of Transformers in Computer Vision

The use of transformers in computer vision began gaining momentum in 2020, when Dosovitskiy and al [48]. introduced the ViTs, demonstrating that a pure transformer architecture could outperform traditional CNNs on large-scale image classification tasks [49].This marked a pivotal shift, as ViTs leveraged self-attention to model global relationships in images, opening new directions for visual recognition research [50].

II.2.2 From Sequence to Spatial: Adapting the Transformer to Images

Transformers adapt text to images by representing both as sequences of tokens [44]. Text is tokenized and embedded, then the model generates image tokens autoregressively. These tokens are decoded into pixels, enabling the transformer to link language and visual content effectively .

II.2.3 Rethinking Visual Representation with Self-Attention

Self-attention fundamentally changes how visual information is processed by enabling models to capture relationships between all parts of an image, not just local neighborhoods. Unlike convolutions, which are limited to local receptive fields, self-attention mechanisms in Vision Transformers allow for global context modeling, leading to improved recognition of complex visual patterns and spatial hierarchies [51].Studies have shown that self-attention can outperform traditional convolutional approaches in both accuracy and robustness, and it enables perceptual grouping based on feature similarity, offering a new perspective on how visual representations are learned [52].

II.2.4 Core Principles of Vision Transformer Architecture

Vision Transformers process images by dividing them into fixed-size patches, which are then embedded into vectors. Positional encoding is added to these embeddings to retain spatial information. The sequence of patch embeddings is passed through multiple transformer layers, each comprising multi-head self-attention and feedforward networks. This structure enables the model to capture complex global and local relationships within the image for effective visual understanding [53].

II.2.5 ViTs as a Paradigm Shift in Image Understanding

Vision Transformers mark a turning point in image analysis by replacing convolutional operations with self-attention mechanisms that capture global relationships across an entire image [44]. This shift enables models to understand long-range dependencies and complex spatial patterns more effectively than traditional CNNs, leading to improved accuracy and flexibility in various vision tasks.

II.3 Key Components of Vision Transformers

Vision Transformers represent a paradigm shift in computer vision, moving away from traditional convolutional operations to a transformer-based architecture [54]. . This architecture processes images by treating them as sequences of patches, enabling the model to capture both local and global dependencies. The core components of ViTs include patch embedding, positional encoding, self-attention mechanisms, and transformer encoders figure II.1 represent them . Together, these components allow ViTs to model complex spatial relationships within images, leading to improved performance in various vision tasks figure II.1.

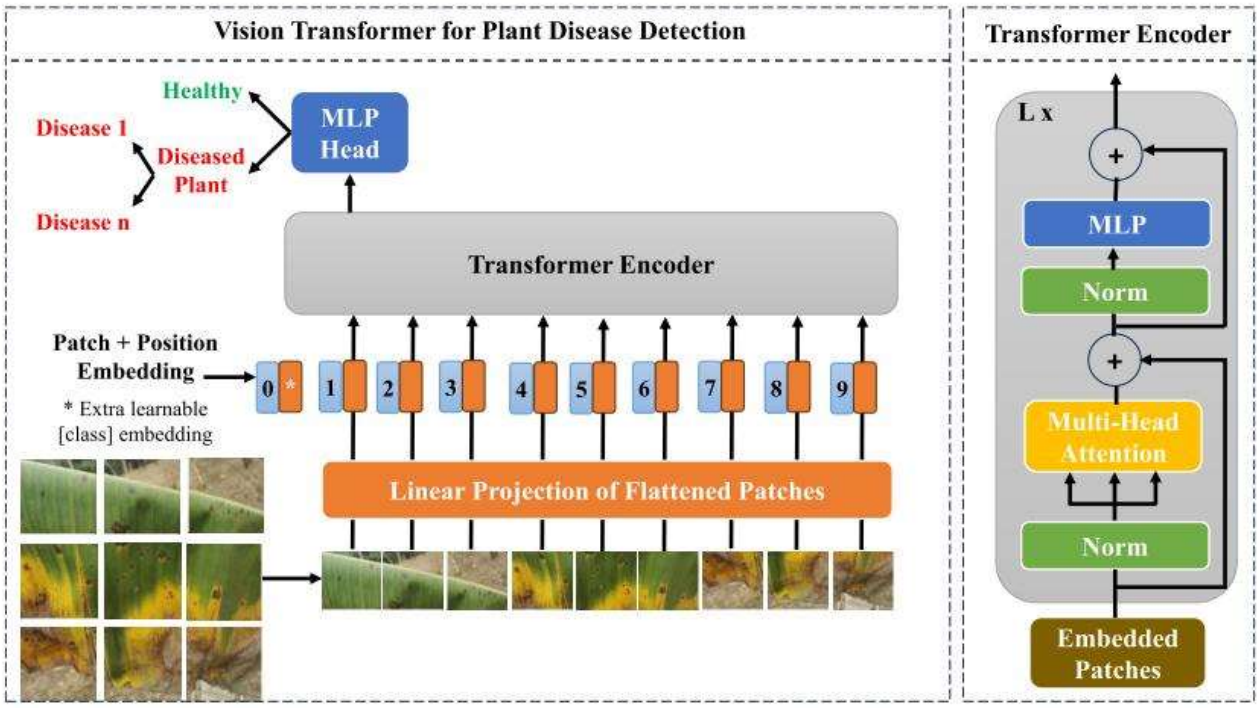


Figure II.1: Detailed Architecture of the ViT [5].

II.3.1 Patch Embedding

The first step in Vision Transformers is to convert the input image into a format that the transformer can process. Unlike CNNs that operate on the entire image with sliding filters, ViTs divide the image into smaller, fixed-size patches typically 16×16 pixels. Each patch is then flattened into a one-dimensional vector by concatenating

its pixel values across all color channels. This vector is then passed through a learnable linear projection layer that transforms it into a fixed-length embedding vector [55].

This process effectively converts the 2D image into a sequence of patch embeddings, similar to how sentences are tokenized into words for language models. Each embedding represents a small portion of the image, and the entire image is represented as a sequence of these embeddings. This transformation enables the transformer, which is designed to process sequences, to handle image data efficiently. Transforms an image into a sequence of patch embeddings, similar to word embeddings in NLP.

$$N = \frac{H \times W}{P^2} \tag{II.1}$$

The image is divided into N non-overlapping patches of size P×P.

$$x_i \cdot E, \quad E \in \mathbf{R}^{(P^2 \cdot C) \times D} \tag{II.2}$$

Each flattened patch $\mathbf{x}_i$ is projected into a D-dimensional embedding space using a learnable matrix $\mathbf{E}$.

II.3.2 Positional Encoding

Transformers are inherently permutation-invariant, meaning they treat input tokens as unordered sets. However, images have a spatial structure where the position of each patch relative to others is crucial for understanding content (e.g., the location of a leaf spot or lesion). To preserve this spatial information, positional encodings are added to each patch embedding . Positional encodings are vectors that encode the position of each patch within the original image grid. These can be fixed (such as sinusoidal functions that vary smoothly with position) or learned parameters optimized during training. By adding these encodings to the patch embeddings, the model gains

awareness of the spatial arrangement of patches, allowing it to distinguish between patches in different locations and maintain the image's structural integrity throughout processing.

$$\mathbf{z}_i = \mathbf{z}_i + \mathbf{p}_i, \quad \mathbf{p}_i \in \mathbf{R}^D \tag{II.3}$$

Positional encodings $\mathbf{p}_i$ are added to patch embeddings $\mathbf{z}_i$ to retain spatial information.

II.3.3 Self-Attention Mechanism

The self-attention mechanism is the core innovation that allows Vision Transformers to capture complex relationships within the image. Unlike CNNs, which focus on local neighborhoods, self-attention enables the model to consider interactions between all patches simultaneously. For each patch embedding, the model computes three distinct vectors: a query, a key, and a value. The query vector represents the patch's current focus, the key vectors represent all patches in the sequence, and the value vectors carry the information to be aggregated. The model calculates attention scores by comparing the query of one patch with the keys of all patches, determining how much attention each patch should pay to every other patch . These attention scores are normalized and used to weight the value vectors, producing a new representation for each patch that incorporates information from the entire image. This dynamic weighting allows the model to focus on relevant regions, capture long-range dependencies, and understand complex spatial patterns that may be critical for tasks like disease detection on leaves. Multi-head self-attention extends this by running multiple attention operations in parallel, each learning to focus on different aspects or features of the image, thereby enriching the model's representational capacity[5].

.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{II.4}$$

Where $\mathbf{Q} = \mathbf{X}\mathbf{W}^{\mathrm{Q}}$, $\mathbf{K} = \mathbf{X}\mathbf{W}^{\mathrm{K}}$, and $\mathbf{V} = \mathbf{X}\mathbf{W}^{\mathrm{V}}$ are the queries, keys, and values, respectively, and d_{k} is the dimension of the key vectors.

II.3.4 Transformer Encoder

The Transformer Encoder is the central processing unit of Vision Transformers, responsible for extracting meaningful features from the sequence of patch embeddings. It is composed of multiple stacked layers, each containing two main components:

1. **Multi-Head Self-Attention:** Enables the model to attend to different image regions simultaneously, capturing a wide range of relationships between patches and allowing for a comprehensive understanding of visual content [56].

$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\ldots,\mathrm{head}_h)W^O \tag{II.5}$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{II.6}$$

   Where h is the number of heads, and $\mathbf{W}^{\mathrm{Q}}{}_{\mathrm{i}}$, $\mathbf{W}^{\mathrm{K}}{}_{\mathrm{i}}$, $\mathbf{W}^{\mathrm{V}}{}_{\mathrm{i}}$, and $\mathbf{W}^{\mathrm{O}}$ are learnable projection matrices.

2. **Feedforward Neural Network:**Further processes each patch embedding independently, typically using two linear layers with a non-linear activation in between (such as GELU), refining the features extracted by self-attention [57].

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{II.7}$$

   Where $\mathbf{W}_1$ and $\mathbf{W}_2$ are weight matrices, and b_1 and b_2 are biases.

3. **Residual Connections and Layer Normalization:** Each sub-layer is followed by layer normalization and wrapped with residual connections to stabilize train-

ing and facilitate deeper model learning [54].. Stacking these encoder layers allows ViTs to integrate local and global information across the image, supporting robust performance in various computer vision tasks.

II.4 From CNNs to ViTs in Image processing

The evolution from Convolutional Neural Networks (CNNs) to Vision Transformers (ViTs) marks a significant shift in image processing methodologies. CNNs excel at capturing local features through convolutional operations but often struggle with modeling long-range dependencies due to their limited receptive fields. This limitation hinders their ability to understand global context within images, which is crucial for tasks requiring holistic scene comprehension . In contrast, ViTs leverage self-attention mechanisms that allow for the modeling of global relationships between image patches from the very beginning of the processing pipeline. This architecture enables ViTs to capture both local and global features effectively, leading to improved performance in complex image understanding tasks Table II.1 shows the most important differences between CNN and ViT [58].

II.4.1 Architectural Limitations of CNNs in Modeling Global Context

CNNs have been the cornerstone of image processing tasks due to their ability to capture local features through convolutional operations. However, their inherent design focuses on local receptive fields, which limits their capacity to model long-range dependencies and global context within images. This limitation becomes evident in complex scenes where understanding the relationship between distant objects is crucial. For instance, recognizing that a "boat" is likely near "water" requires analyzing the entire image, not just localized patches [58].

II.4.2 Vision Transformers as a Global Feature Modeling Alternative

Vision Transformers (ViTs) introduce a paradigm shift by employing self-attention mechanisms that allow for the modeling of global relationships between image patches from the very beginning of the processing pipeline.This global attention enables ViTs to capture long-range dependencies and holistic context within an image, leading to improved performance in tasks requiring comprehensive understanding of visual content . Moreover, the flexible and scalable architecture of ViTs allows them to handle varying input resolutions and aspect ratios effectively, further enhancing their applicability across diverse image processing tasks [58].

Table II.1: Comparison between Vision Transformer and Convolutional Neural Network

Feature	**Vision Transformer**	**Convolutional Neural Network**
Feature Extraction	Global, using non-overlapping image patches	Local, through convolutional filters
Architecture	Based on transformers	Built around convolutional layers
Data Requirement	Requires more data to perform optimally	Less data-intensive; can perform well on smaller datasets
Computational Cost	Higher, due to the need for large-scale pre-training	Relatively lower
Use Case Suitability	Excels in tasks requiring an understanding of global patterns	Ideal for tasks where local features are significant

II.5 Stat of the art: ViTs for plant Disease detection

Between 2020 and 2025, Vision Transformers emerged as a transformative approach in the field of plant disease. Their work demonstrated the early superiority of ViTs for plant pathology tasks, especially in capturing complex image patterns and long-range dependencies [59].

Sana Parez and al [45].proposed the GreenViT technique, an efficient ViT-based

model that surpassed state-of-the-art CNNs in plant disease detection. Their approach leveraged the strengths of ViTs to overcome the limitations of convolution-based models, particularly for complex and diverse agricultural datasets [45].

Wafaa H. Alwan et al [60] .developed a novel hybrid framework integrating EfficientNet-B8, ViT, and Knowledge Graph Fusion (KGF), achieving a 99.3% testing accuracy across 38 disease categories. This approach combined deep learning with semantic enrichment, providing both high accuracy and interpretability, and demonstrated the potential of combining ViTs with domain knowledge for robust, scalable diagnosis.

Yu Ruan et al [61] .focused on real-world deployment by introducing PMVT, a lightweight mobile vision transformer optimized for plant disease identification on mobile devices. Their model achieved 93.6% accuracy on wheat, 85.4% on coffee, and 93.1% on rice datasets-all with a reduced parameter count. The PMVT model's efficiency and accuracy make it well-suited for field use, and it has been successfully integrated into a mobile app for practical disease diagnosis.

Aadarsh Kumar Singh et al [62].further validated the transformative potential of ViTs by demonstrating their superior accuracy and scalability in classifying 55 plant disease classes. Their findings highlight ViTs ability to enhance early disease diagnosis, contributing to more sustainable and productive agricultural practices as illustrated in Table II.2.

Table II.2: Overview of Vision Transformer Applications in Agriculture

Category	**Year**	**Model Title**	**Dataset(s)**	**Accuracy / Metric**	**Key Contribution**
Emergence of ViTs in Agriculture [59]	2022	Plant disease and insect pest identification based on vision transformer	PlantVillage	96.71%	First demonstration of ViTs superiority over CNNs for plant disease and pest detection
Lightweight and Explainable ViTs [63]	2023	PMVT: a lightweight vision transformer for plant disease identification on mobile devices	Wheat, Coffee, Rice	Wheat: 93.6%, Coffee: 85.4%, Rice: 93.1%	Mobile-optimized ViT with attention modules for real-time, explainable diagnosis on edge devices
Hybrid Models and Specialized Architectures [64]	2023	Multispectral Plant Disease Detection with Vision Transformer–Convolutional Neural Network Hybrid Approaches	Custom Multispectral	83.3% (ViT-B16)	Combined CNN and ViT for improved multispectral disease detection and robust feature extraction

Category	Year	Model Title	Dataset(s)	Accuracy / Metric	Key Contribution
Efficient ViT Networks [45]	2023	Precision Agriculture: Exploring Plant Disease Detection via Efficient Vision Transformers	PlantVillage, others	Not specified	Fine-tuned ViT (GreenViT) outperforms SOTA CNNs, demonstrating scalability and robustness in precision agriculture
Multi-image ViT [65]	2024	Multi-ViT	Apple, Grape, Tomato	>99%	Uses attention from multiple leaves to improve recognition accuracy
Real-Time and Robust ViT Models [45]	2025	An enhanced vision transformer network for efficient and accurate crop disease detection	Not specified	Not specified	Improved ViT with attention distillation for efficiency and accuracy in real-world scenarios
Transfer Learning and Mobile ViTs [65]	2025	Attention Score-Based Multi-Vision Transformer Technique for Plant Disease Classification	Apple, Grape, Tomato	>99%	Multi-image attention and transfer learning for robust disease classification across multiple crops

II.6 Future Directions: Integrating ViTs for Smart Agriculture

Recent studies from a range of sources confirm the rapid progress and strong potential of ViTs and advanced computer vision in agriculture. For example, an enhanced ViT network was shown to deliver highly accurate and efficient crop disease detection, outperforming traditional deep learning methods [66]. Comprehensive surveys highlight how transformer models are now central to pest and disease identification and yield prediction, marking a shift from convolutional networks to transformer-based approaches in the field [67]. Lightweight ViT frameworks have also been developed for efficient maize leaf disease classification,demonstrating both high accuracy and practical scalability .

In precision agriculture, ViTs have enabled automated, scalable crop growth analysis from drone imagery, offering superior performance and flexibility compared to older image classification models. Studies integrating ViT with other deep learning models have improved the synchronous monitoring of crop growth stages and leaf area index, key indicators for crop management and yield prediction.

Furthermore, optimized Swin Transformer models have been introduced for plant disease detection, addressing efficiency and accuracy in complex field conditions. Across the sector, these advances show computer vision is becoming central to smart farming, enabling real-time monitoring, resource optimization, and sustainable practices . Collectively, these developments demonstrate that ViTs and computer vision are transforming agriculture into a more data-driven, efficient,and resilient industry.

II.6.1 Enhanced Disease Detection and Crop Monitoring

ViTs are being employed for early and accurate identification of plant diseases. For instance, a smartphone-based ViT model has been developed to differentiate between healthy and diseased tomato plants, achieving high accuracy with a dataset of over

10,000 images [68] .

Similarly, ViTs have been utilized for weed detection in crop fields, addressing challenges posed by the resemblance between crops and weeds [68] .

II.6.2 Integration of Multimodal Data for Comprehensive Analysis

The fusion of hyperspectral imaging and LiDAR data through ViTs has led to the creation of frameworks like PlantViT, which enhance vegetation classification by optimizing feature fusion across spectral and spatial dimensions [69].

Such integrations allow for more precise and holistic assessments of agricultural environments.

II.6.3 Deployment in Resource-Limited Settings

Recognizing the need for accessible solutions, researchers have developed lightweight ViT models optimized for edge devices. For example, MobilePlantViT is a hybrid ViT architecture designed for generalized plant disease classification, achieving high performance with minimal computational resources, making it suitable for deployment in resource-constrained settings [70].

II.6.4 Advancements in Precision Agriculture

The integration of ViTs with UAVs and computer vision technologies is modernizing soil sampling methodologies in agriculture.

This approach aims to enhance data accuracy, streamline sampling efficiency, and foster the adoption of precision agriculture practices [71].

II.7 Conclusion

This chapter has demonstrated how Vision Transformers represent a significant advancement in computer vision applications for agriculture. By processing images as sequences of patches and leveraging self-attention mechanisms, ViTs can capture both local details and global context more effectively than traditional convolutional methods. This capability leads to improved accuracy in detecting plant diseases,monitoring crop growth, and predicting yields. The integration of positional information ensures that spatial relationships within images are preserved, while the transformer encoder's layered architecture refines these representations to extract meaningful features. Recent research highlights the practical benefits of ViTs, including their adaptability to diverse crops and environments, scalability through drone and sensor data, and potential for deployment on edge devices for real-time field analysis. Overall, Vision Transformers are poised to transform agriculture into a more precise, efficient, and sustainable practice by enabling smarter decision-making and resource management. Continued innovation in Vision Transformers and advanced computer vision techniques, will be essential to drive smarter more efficient solutions that address global challenges in food production and sustainable farming.

Chapter III

System design and results

III.1 Introduction

Following the theoretical foundations and literature review established in the preceding chapters, this chapter transitions to the practical implementation and empirical evaluation of our plant disease classification system. We present the design and implementation of a plant disease detection system using Vision Transformer technology. It also aims to conduct a comparative performance analysis between a traditional convolutional neural network model, AlexNet, and two advanced ViT-based architectures: Simple ViT and Distillation ViT. For evaluation: the widely recognized BANANA dataset [72] and the newly introduced Ziban Plant Disease Dataset (ZPDD) are used.

The chapter also describes the hardware and software configurations, data preprocessing steps, and evaluation metrics used. Experimental results confirm the superior performance of ViT models in accurately detecting plant diseases under realistic conditions, highlighting their potential for deployment in practical agricultural scenarios.

III.2 Hardware and software

To support the development and deployment of deep learning models for plant disease detection, a robust computational infrastructure is essential. This section outlines the hardware and software components utilized in our system, detailing the specifications and tools that enable efficient processing, model training, and inference. The combination of high-performance hardware and a modern software stack ensures the system can handle the computational demands of ViT architectures and large-scale image datasets.

III.2.1 Hardware

Our system is built on a high-performance workstation featuring an Intel i9-13900KF CPU, 128 GB RAM, and an NVIDIA RTX 4080 GPU. This setup ensures efficient training and inference of Vision Transformer models, handling large-scale image datasets with high throughput and low latency.

III.2.2 Software

The software environment includes Windows 10 Professional, Python 3.11, and PyTorch 2.1 with CUDA 12.1 support for GPU acceleration. These tools provide a flexible and optimized platform for developing, training, and deploying ViT-based plant disease detection models? figure III.1 shows the software tools we used .

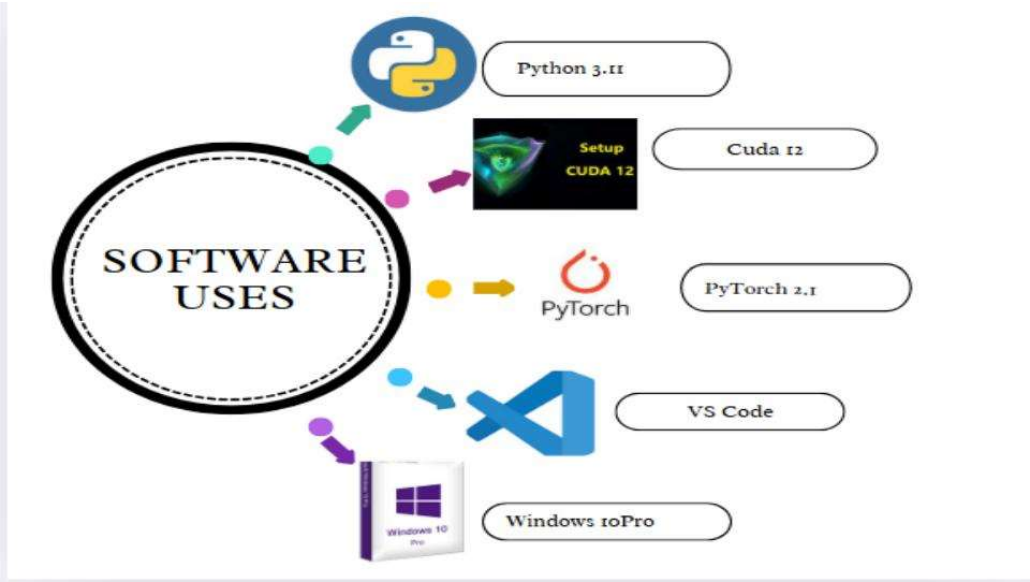


Figure III.1: Software tools

III.3 The proposed plant disease detection system based ViTs

In the proposed plant disease detection system shown in Figure III.2 , we use Simple ViT and Distilled ViT as the Vision Transformer models. Simple ViT offers a straightforward architecture with strong performance and reduced complexity, while

Distilled ViT improves accuracy and generalization by learning from a larger pre-trained model using knowledge distillation. From the Convolutional Neural Network models we use AlexNet, known for its simplicity, fast training, and strong feature extraction capabilities. These models are selected to combine the global attention power of ViTs with the local feature extraction strength of CNNs, providing a balanced solution that is both accurate and efficient for plant disease detection in real-world agricultural environments .

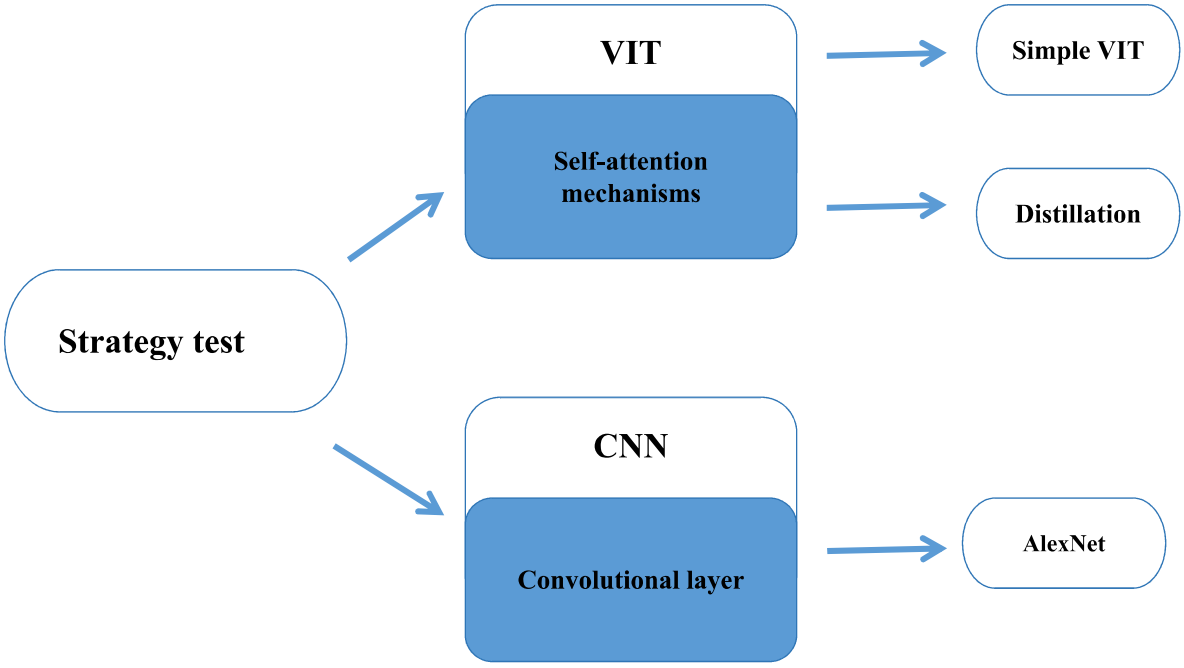


Figure III.2: Our strategy test

III.3.1 Simple ViT

1. **Overview** Simple Vision Transformer (Simple ViT) adopts a streamlined design philosophy aimed at improving computational efficiency while maintaining strong performance. This minimalist approach focuses on simplifying the standard ViT architecture by [73]:

    - Eliminating overly complex components, such as deep hierarchical layers.

- Refining the attention mechanism to enable more efficient and effective feature extraction.
- Reducing the dependency on large-scale pretraining by employing improved training techniques.

2. **Mathematical Formulation**: A Vision Transformer begins by processing an input image,where X is the input image itself.

$$\mathbf{X} \in \mathbf{R}^{H \times W \times C} \tag{III.1}$$

where R represents the set of real numbers,and H,W,and C denote where H,W,and C the height, width, and number of channels, respectively. The image is divided into a sequence of non-overlapping patches. Each patch is flattened and linearly projected into a D-dimensional embedding space, forming the initial input sequence to the transformer. The resulting sequence of embedded patches is represented as:

$$\mathbf{Z}_0 = \mathbf{x}_1\mathbf{E};\ \mathbf{x}_2\mathbf{E};\ \ldots;\ \mathbf{x}_N\mathbf{E} + \mathbf{E}_{\text{pos}} \tag{III.2}$$

Here:

- $Z_0 \in \mathbf{R}^{N \times D}$ is the sequence of embedded patches.
- $x_i \in \mathbf{R}^{(P^2C)}$ is the i-th image patch of size $P \times P$, flattened into a vector.
- $E \in \mathbf{R}^{(P^2C) \times D}$ is a learnable linear projection matrix used for patch embedding.
- $E_{POS} \in \mathbf{R}^{N \times D}$ is the positional embedding matrix, which encodes spatial information to preserve the order of patches.

These embeddings are then passed through a series of transformer encoder layers, each composed of a multi-head self-attention mechanism followed by feed-

forward layers. The self-attention operation is defined as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{D}}\right)\mathbf{V} \tag{III.3}$$

Where:

- $Q, K, V \in \mathbf{R}^{N\times D}$ are the query, key, and value matrices, respectively, derived from the input embeddings.
- $\sqrt{D}$ acts as a scaling factor to ensure numerical stability during training.

This self-attention mechanism allows the model to effectively capture long-range dependencies and global context within the image, making it a core component of the ViT architecture [74] [75] [76].

III.3.2 Distillation ViT

1. **Overview** Attention-based neural networks, particularly ViTs, have shown strong performance in image understanding tasks such as classification. However, achieving high accuracy typically requires pretraining on large-scale datasets, which demands significant computational resources. To address this, a transformer-specific teacher-student distillation strategy has been introduced, enabling efficient training on more modest datasets [77] . This method introduces a distillation token that facilitates the transfer of knowledge from a teacher model often a convolutional neural network to the student ViT through attention mechanisms.

2. **Soft Distillation** In soft distillation, the goal is to minimize the Kullback-Leibler (KL) divergence between the softmax predictions of the student and teacher models. Let Zt and Zs represent the logits from the teacher and student models, respectively. The combined loss function is defined as:

$$L_{\text{global}} = (1-\lambda)L_{\text{CE}}\big(\psi(Z_s), y\big) + \lambda\tau^2\,\text{KL}\big(\psi(Z_s/\tau), \psi(Z_t/\tau)\big) \tag{III.4}$$

Where:

- λ Balances the cross-entropy loss and the KL divergence.
- τ Is the temperature parameter used to soften the logits.
- Ψ Denotes the softmax function.
- y is the ground truth label.

3. **Hard-label Distillation** An alternative to soft distillation is hard-label distillation, where the teacher's predicted class is treated as the true label. The hard target is defined as

$$y_t = \arg\max_c Z_t(c) \tag{III.5}$$

and the objective becomes:

$$L_{\text{hardDistill}} = \frac{1}{2} L_{\text{CE}}\big(\psi(Z_s), y\big) + \frac{1}{2} L_{\text{CE}}\big(\psi(Z_s), y_t\big) \tag{III.6}$$

Hard labels can be transformed into soft labels using label smoothing, where the true label has probability $1 - \varepsilon$, and the remaining probability ε is distributed across other classes, with $\varepsilon = 0.1$ used in experiments.

4. **Distillation Token** A key innovation in this approach is the use of a distillation token appended to the patch embeddings. This token interacts with other tokens through self-attention layers and is trained using the distillation loss. Unlike a standard additional class token, which may become redundant, the distillation token develops a distinct and useful representation. Empirical results show that the cosine similarity between the class and distillation tokens remains low (averaging around 0.06), increasing across layers but never reaching full redundancy indicating complementary roles. This mechanism significantly boosts classification accuracy compared to traditional distillation methods or simply adding extra class tokens.

5. **Fine-tuning with Distillation**

   During fine-tuning at higher input resolutions, both the ground truth labels and the teacher's predictions are incorporated. When the teacher model is also adapted to handle the higher resolution, the benefits of distillation are further amplified.

6. **Classification with Joint Classifiers** At inference time, both the class token and distillation token are passed through their respective linear classifiers. Their softmax outputs are then fused using a late fusion strategy to yield the final prediction, improving accuracy and robustness .

III.3.3 AlexNet

AlexNet is a deep convolutional neural network designed for image classification. It processes RGB images resized to 227×227 pixels through a hierarchical system of layers[78] .

- Layer Composition

1. Convolutional layers for feature extraction.

   - Conv1: 96 filters (11×11×3), stride 4 aggressive spatial reduction.
   - Conv2: 256 filters (5×5×48), followed by pooling and normalization.
   - Conv3: 384 filters (3×3×256), no pooling captures complex features.
   - Conv4 and Conv5: 384 and 256 filters (3×3), further deepen feature maps; Conv5 ends with pooling.

2. Fully connected layers for classification.

- FC1 and FC2: 4096 neurons each, with dropout for regularization.
- FC3: 1000 neurons for classification (softmax output for ImageNet classes).

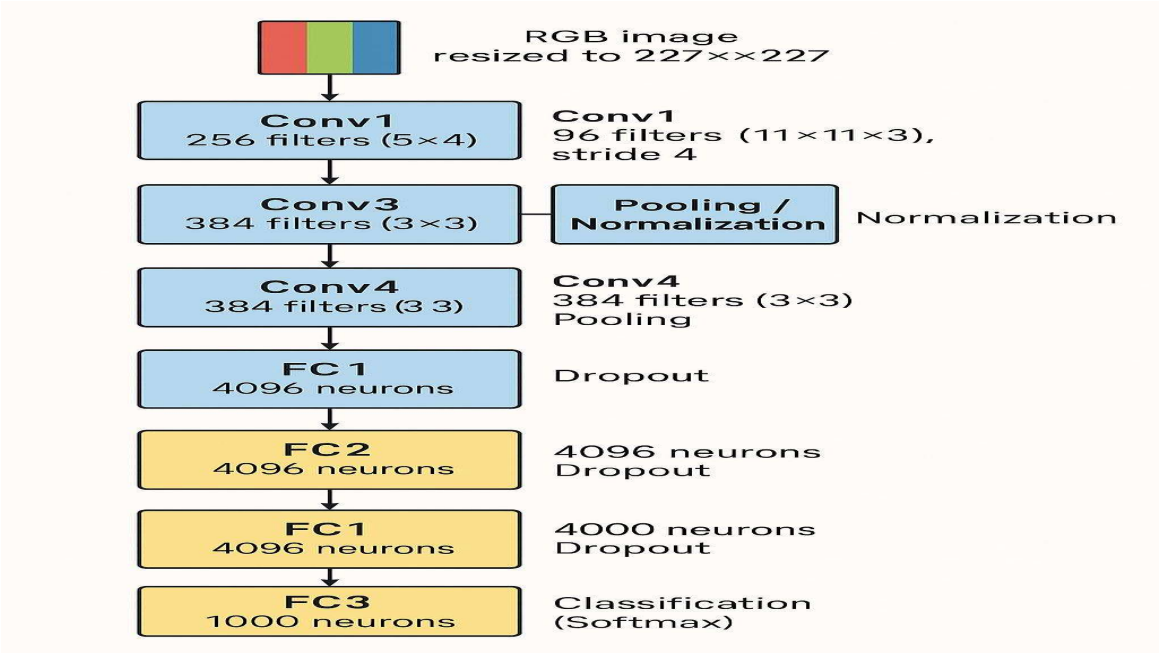


Figure III.3: AlexNet Architecture

Intermediate layers apply ReLU activations, local response normalization, max-pooling, and dropout to enhance learning and reduce overfitting AlexNet's architecture efficiently extracts and processes image features using deep layers while employing regularization techniques to generalize well, FigureIII.3 represent AlexNet Architecture.

III.4 Datasets and protocol

The datasets used in this study include a proprietary collection of plant leaf images captured by the Ziban Plant disease Dataset (ZPDD) Cropped represented in Table III.1and UNCropped represented in Table III.2 ,and research team in LI3C research laboratory university of Biskra. This dataset comprises four disease categories: Tuta absoluta, Oidium, Alternaria, and climatic incidents, with both cropped and uncropped leaf images. The cropping process was performed manually to ensure precise focus on the affected leaf areas, improving the quality of the training data.

In addition to this, we utilized a large-scale banana leaf dataset [72] containing approximately 25 million images, providing extensive variability and diversity for

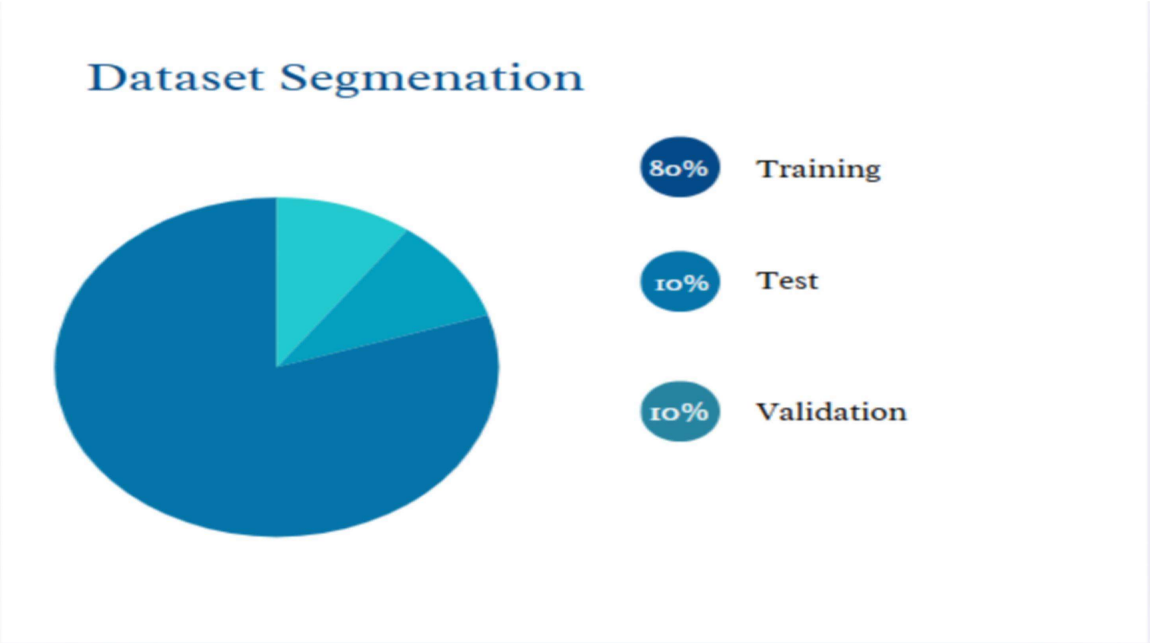


Figure III.4: Dataset Segmentation

model training and evaluation. All datasets underwent standard pre-processing steps including normalization and resizing to fit the input requirements of the Vision Transformer models. The data was split into training (80%), validation (10%), and testing (10%) subsets to ensure balanced and unbiased performance assessment as it shown in the chart below in Figure III.4

- Training set (80%): Used to train the model, containing most of the data so the model can learn underlying patterns and relationships.
- Validation set (10%): Used during training to tune hyperparameters and prevent overfitting by evaluating the model's performance on unseen data.
- Test set (10%): Used only after training to provide an unbiased, final assessment of the model's ability to generalize to new, unseen data.

This combination of real-world, manually curated images and large-scale datasets allows the model to generalize effectively across different disease manifestations and environmental conditions. The table below presents the categories of tomato leaf diseases included in our proprietary dataset, alongside representative leaf images.

Table III.1: Categories of tomato leaf diseases in cropped ZPDD

Disease	Example 1	Example 2	Example 3
Oidium			
Tuta absoluta			
Alternaria			
Climatic Incidents			

Table III.2: Categories of tomato leaf diseases in uncropped ZPDD

Disease	Example 1	Example 2	Example 3
Oidium			
Tuta absoluta			
Alternaria			
Climatic Incidents			

The banana dataset [72] represented in Table III.3 comprises expertly annotated images of healthy and diseased leaves, representing three prevalent diseases. Collected under authentic agricultural conditions, this dataset offers a comprehensive and diverse foundation for the development and evaluation of robust plant disease detection models. The table below presents the banana leaf disease categories with representative images from a publicly available, globally sourced dataset.

Table III.3: Categories of leaf diseases in Banana dataset

Healthy	Cordana	Pistalotiopsis	Sigatoka
			

III.4.1 Preprocessing

The ZPDD (Ziban Plant Disease Dataset) was manually produced in the Laboratory L13C of University Mohamed Khider Biskra. This study represents the first use of the ZPDD dataset, which currently includes images categorized into four distinct tomato plant disease classes:

- Tuta absoluta
- Alternaria
- Climatic Incident
- Oidium

III.4.2 Dataset Preparation

The original images in the ZPDD dataset include both the leaves and surrounding background, which may introduce noise and irrelevant information that can negatively impact the performance of disease recognition models. To address this, a preprocessing step involving manual cropping was performed to generate two versions of the dataset:

1. **Uncropped ZPDD:**The original images as captured, containing the full scene including the leaf and background.

2. **Cropped ZPDD:**Images manually cropped to isolate the leaf area, eliminating background noise and enhancing the clarity of the leaf features.

- **Cropping Process:** The cropping was conducted manually to precisely delineate the leaf boundaries, ensuring that only the relevant leaf area is retained in the cropped images. This step is crucial to remove extraneous elements such as soil, pots, or other environmental factors that could introduce noise and reduce the accuracy of disease classification. By focusing on the leaf area, the cropped dataset provides clearer visual information of the disease symptoms, such as spots, discolorations, or lesions, which are essential for effective disease detection and analysis. We aim to expand the dataset by including more categories and a wider variety of plant diseases, thereby increasing its applicability and robustness for plant disease detection and classification tasks across different crops and environmental conditions.

III.5 Model Parameters

Understanding the parameters of machine learning models is essential for evaluating their architecture, computational complexity, and overall performance. In this section, we provide a comprehensive overview of the key parameters utilized in the Simple ViT, Distillation, and AlexNet models. These parameters play a crucial role in determining the model structure, learning capacity, and efficiency.

1. **Vit models paramaters : Simple Vit ,Distillation**

    - **image.ize** This is the size of the input images. For example, in ViT models, images are often resized to 224x224 pixels. The image is then split into non-overlapping patches to process with the transformer architecture.
    - **patch.size** This refers to the size of each patch that the image is divided into. For instance, if patch.size = 16, it means that the image will be split

into patches of size 16x16 pixels. The patches are then flattened and passed through the transformer.

- **num.classes** This is the number of categories (classes) the model is trained to predict. For example, if you're classifying images into 10 categories, num.classes would be set to 10.
- **dim** This is the dimension of the token embeddings. It defines the size of the feature vector representing each patch. In ViT, the higher this value, the more capacity the model has to learn complex patterns, but it also increases the computational cost.
- **depth** This refers to the number of transformer layers (blocks). Each transformer block contains two main components: a multi-head self-attention mechanism and a feed-forward neural network. Increasing the depth allows the model to capture more complex dependencies in the input data but increases the computational load.
- **h**eads This defines the number of attention heads used in the multi-head self-attention mechanism. Each attention head learns a different set of attention patterns, allowing the model to focus on different parts of the input. More heads typically provide better performance at the cost of more computational resources.
- **mlp.dim** This refers to the dimension of the hidden layers in the feed-forward network that comes after the attention mechanism in each transformer block. A larger value means the feed-forward network has more capacity to process information, but this also increases computational complexity.
- **dropout** This is the dropout rate applied to the model, used to prevent overfitting during training. It randomly drops some of the neurons in the

network during training to ensure that the model doesn't rely too heavily on any one feature. A typical dropout rate is between 0.1 and 0.5.

- **emb.dropout** Similar to regular dropout, this is a dropout rate applied specifically to the embedding layer (where the image patches are mapped to vectors). This prevents overfitting during the early stages of training when the model is learning the patch embeddings.

2. **AlexNet parametres**

- **Input Size** The dimensions of the input image fed into the model. Typically written as Height × Width × Channels (e.g., 224 × 224 × 3 for an RGB image).
- **Conv Layers** Short for Convolutional Layers, these layers apply filters to extract spatial features such as edges, textures, and patterns from images. Common in CNNs like AlexNet.
- **Filter Sizes** The size of the filter (kernel) used in convolutional layers, such as 3×3, 5×5, or 11×11. Larger filters capture broader patterns, smaller filters capture fine details.
- **Pooling** A downsampling technique that reduces the spatial size of feature maps. The most common is Max Pooling, which selects the highest value in a local window (e.g., 2×2 or 3×3).
- **Fully Connected Layers** Also called Dense layers, these connect every neuron from the previous layer to each neuron in the current layer. Used at the end of the model for classification.
- **Dropout** A regularization technique that randomly turns off (drops) a percentage of neurons during training. This helps prevent overfitting and improves model generalization.

- **Activation** A function applied to the output of neurons to introduce non-linearity. Common examples include ReLU (Rectified Linear Unit), Sigmoid, and Softmax. ReLU is widely used in CNNs.

III.6 Evaluation metrics

Evaluation metrics play a vital role in assessing the performance of a classification model, especially in applications such as plant disease detection. In this work, metrics such as accuracy, precision, recall, F1 score, and loss function were used for a comprehensive evaluation. Details and explanations of these metrics are provided in Chapter 1 (pages 21-22). Together, these metrics provide a deeper understanding of a model's reliability, robustness, and generalization ability, especially when dealing with imbalanced datasets. Rather than relying on a single metric alone, combined analysis ensures a more balanced understanding of how well a model performs in accurately identifying diseases, reducing false predictions, and improving training performance through effective optimization.

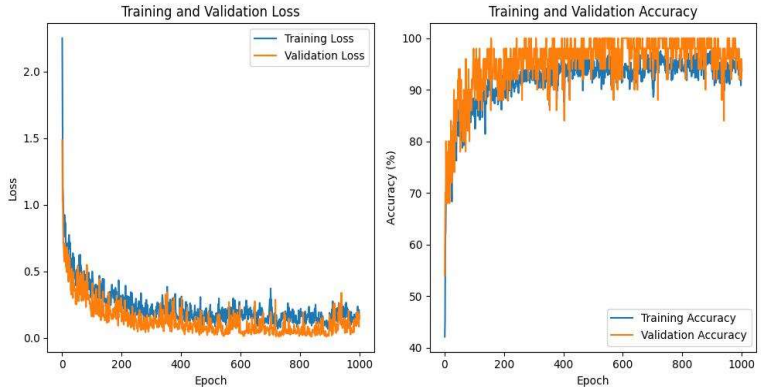
III.7 Results and discussion

We present experimental results offering valuable insights into the performance of Simple ViT, Distilled ViT, and AlexNet for plant disease classification. These models were evaluated under three conditions: using the uncropped ZPDD dataset, the cropped ZPDD dataset, and the BANANA dataset.

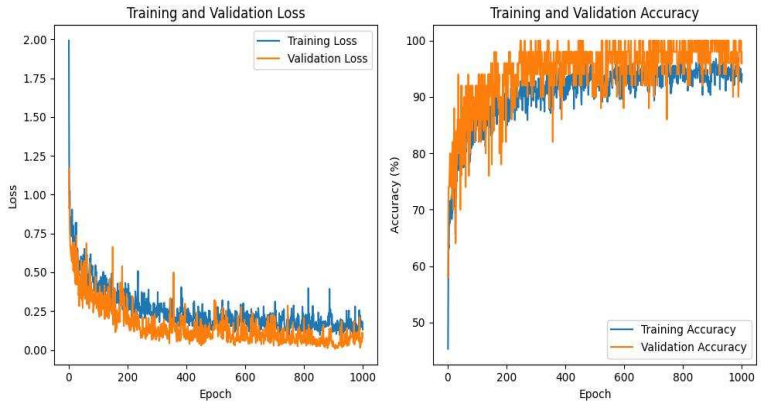
III.7.1 Performance on Uncropped ZPDD dataset

Table III.4: Performance Comparison of Models on UNCropped ZPDD Dataset

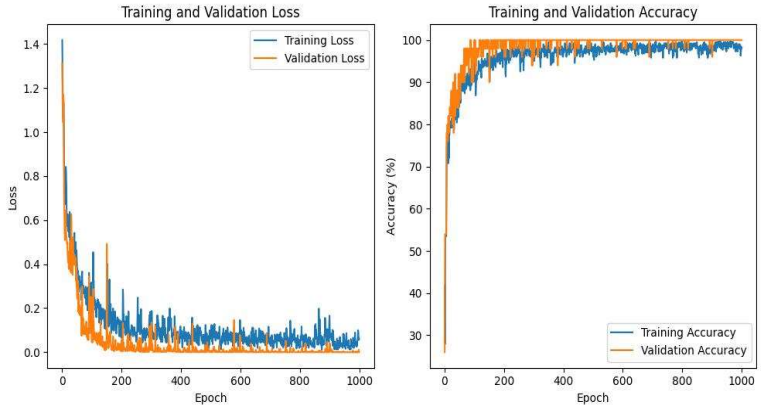
Model	**Performance Indicators**	**Train**	**Validation**	**Test**	**Parameters (M)**	**Storage (MB)**
Simple ViT	Loss	0.2119	0.1785	0.1382	13.003780	50.823
	Accuracy (%)	92.57	94.00	96.23		
	Precision (%)	87.64	94.70	96.88		
	Recall (%)	89.40	92.58	93.75		
	F1 Score (%)	88.26	93.45	94.76		
Distillation	Loss	0.1260	0.0802	0.0680	13.108228	51.235
	Accuracy (%)	96.04	96.00	98.11		
	Precision (%)	95.53	90.18	95.00		
	Recall (%)	93.44	94.51	96.88		
	F1 Score (%)	94.15	91.86	95.56		
AlexNet	Loss	0.0307	0.0001	0.0008	**57.020228**	**222.742**
	Accuracy (%)	**99.01**	**100.00**	**100.00**		
	Precision (%)	99.20	100.00	100.00		
	Recall (%)	97.70	100.00	100.00		
	F1 Score (%)	98.42	100.00	100.00		



(a)



(b)



(c)

Figure III.5: The performance of models ((a)Simple Vit ,(b)Distillation,(c)AlexNet)in Uncropped Dataset

by analyzing Table III.4 and Figure III.5 we found that:

- while all three models Simple ViT, Distillation, and AlexNet perform well on the UNCropped ZPDD dataset.
- AlexNet achieves perfect scores (100% accuracy, precision, recall, and F1).Although this is impressive, it raises concerns about potential overfitting or data leakage, especially given its large size (57M parameters, 222.74 MB).
- This concern is reinforced by the training and validation curves: as seen in the third set of graphs (Figure III.5 (c)), AlexNet reaches near-zero loss and perfect accuracy very early in training, with minimal difference between training and validation performance, a typical signature of overfitting when the model may have memorized the dataset.
- In contrast, Simple ViT and Distillation models show more natural learning behavior.
- The Simple ViT (Figure III.5 (a)) shows a steady decline in both training and validation loss with some fluctuation, and achieves validation accuracy around 95%.
- Similarly, the Distillation model (Figure III.5 (b)) exhibits slightly more stable training curves and slightly better performance than Simple ViT, with validation accuracy reaching approximately 96%.
- These models (Simple Vit ,Distillation)maintain a reasonable gap between training and validation curves, indicating better generalization.
- The slightly higher loss observed in the Simple ViT model compared to Distillation and AlexNet can be attributed to the ViT's lack of inductive biases like locality and translation invariance, which are inherent in CNNs.

- ViTs generally require more data and careful regularization to achieve optimal performance. Unlike the Distillation ViT, which benefits from teacher-guided learning, Simple ViT learns independently, which may lead to slower convergence and higher loss. Despite this, its accuracy and F1-score remain high, indicating strong predictive performance with slightly less confidence in its outputs.

- Overall, while AlexNet provides the highest scores, the evidence from the training curves suggests that Distillation offers the best balance between accuracy, generalization, and model size, making it the most suitable choice for practical deployment, especially in resource-constrained environments.

III.7.2 Performance on Cropped ZPDD dataset

Table III.5: Performance Comparison of Models on Cropped ZPDD Dataset

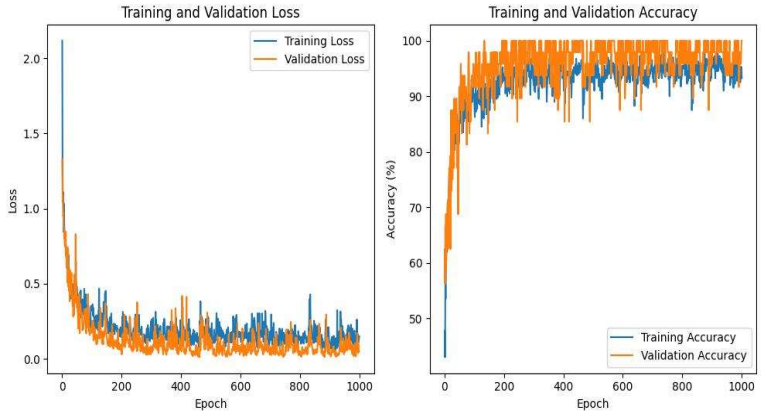
Model	**Performance Indicators**	**Train**	**Validation**	**Test**	**Parameters (M)**	**Storage (MB)**
Simple ViT	Loss	0.1230	0.0484	0.1398	13.003780	50.823
	Accuracy (%)	95.75	100.00	94.44		
	Precision (%)	93.34	100.00	94.10		
	Recall (%)	94.22	100.00	91.96		
	F1 Score (%)	93.72	100.00	92.37		
Distillation	Loss	0.1183	0.0655	0.1480	13.108228	51.235
	Accuracy (%)	96.00	97.92	90.74		
	Precision (%)	93.02	99.07	85.44		
	Recall (%)	96.08	98.08	90.62		
	F1 Score (%)	94.46	98.53	86.43		
AlexNet	Loss	0.0404	0.0000	0.0000	**57.020228**	**222.742**
	Accuracy (%)	**98.75**	**100.00**	**100.00**		
	Precision (%)	96.83	100.00	100.00		
	Recall (%)	98.11	100.00	100.00		
	F1 Score (%)	97.46	100.00	100.00		

On the Cropped ZPDD dataset Table III.5 and figure III.6 we found :

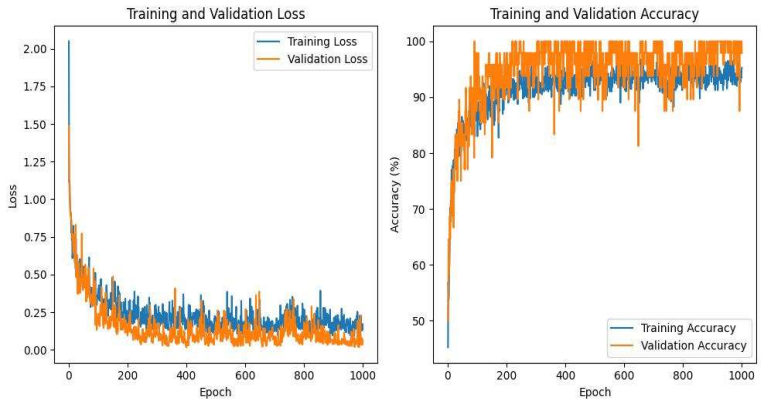
- AlexNet achieves (100%) accuracy on both validation and test sets, which, while impressive, may indicate overfitting or data leakage due to its large ca-

pacity (57 million parameters and 222.74 MB in size).

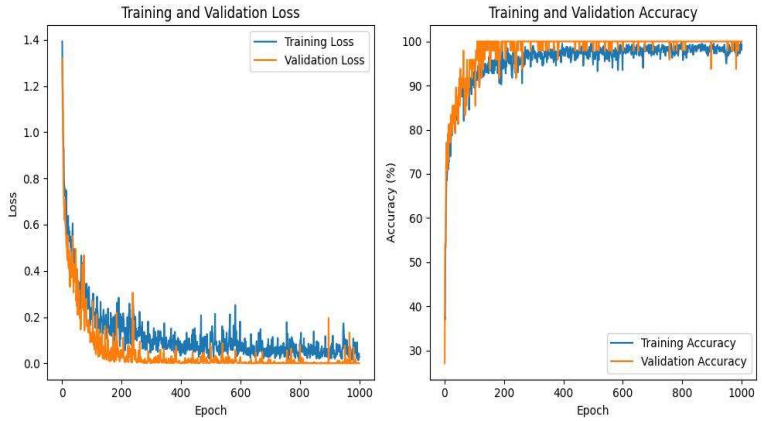
- This concern is reinforced by the training and validation curves: as seen in the third set of graphs (Figure III.6 (c)), AlexNet exhibits a very rapid decrease in training loss and a corresponding increase to perfect training accuracy early in training, with minimal difference between training and validation performance throughout the training process.
- This rapid convergence and lack of divergence between training and validation metrics is a typical signature of overfitting, where the model essentially memorizes the training data.



(a)



(b)



(c)

Figure III.6: The performance of models ((a) Simple ViT,(b) Distillation,(c) AlexNet) in cropped Dataset

- In contrast, Simple ViT and Distillation models show more natural learning behavior.
- The Simple ViT (Figure III.6 (a)) shows a more gradual decrease in both training and validation loss, although with some noticeable fluctuation, and achieves a validation accuracy around 95%.
- Similarly, the Distillation model (Figure III.6 (b)) exhibits slightly more stable training curves and slightly better performance than Simple ViT, with validation accuracy reaching approximately 97%.
- These models maintain a more pronounced gap between training and validation curves compared to AlexNet, indicating better generalization.
- the slightly higher loss for Simple ViT and Distillation models compared to AlexNet is likely due to the way these models handle features. ViTs, including the distilled version, generally need more data and training time to match the performance of CNNs like AlexNet, which have strong spatial priors.
- Overall, while AlexNet provides the highest scores, the evidence from the training curves strongly suggests that Simple Vit offers the best balance between accuracy, generalization, and model size, making it the most suitable choice for practical deployment, especially in resource-constrained environments.

III.7.3 Performance on leaf spot disease in BANANA dataset

Table III.6: Performance Comparison of Models on BANANA Dataset

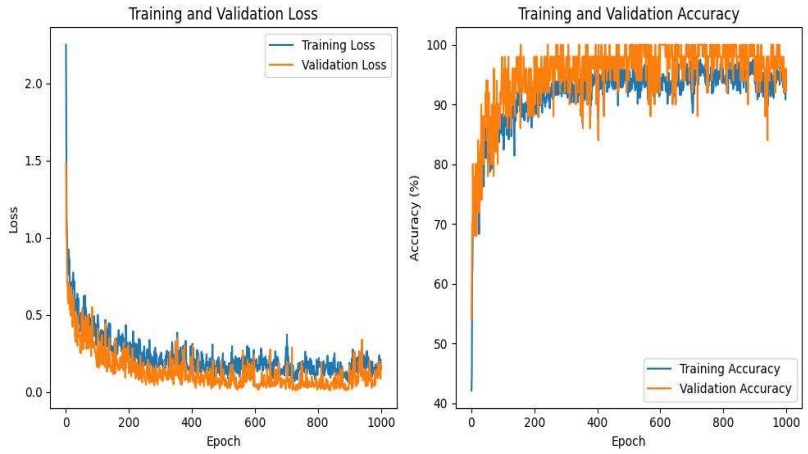
Model	Performance Indicators	Train	Validation	Test	Parameters (M)	Storage (MB)
Simple ViT	Loss	0.0767	0.4667	0.1420	13.003780	50.823
	Accuracy (%)	97.33	90.00	96.25		
	Precision (%)	96.35	91.67	96.74		
	Recall (%)	96.64	90.00	96.25		
	F1 Score (%)	96.49	89.73	96.30		
Distillation	Loss	0.1042	0.4717	0.0697	**13.108228**	**51.235**
	Accuracy (%)	**96.06**	**86.25**	**97.50**		
	Precision (%)	95.30	88.17	97.56		
	Recall (%)	94.38	86.25	97.50		
	F1 Score (%)	94.63	85.89	97.50		
AlexNet	Loss	0.0272	1.4525	0.1563	57.020228	222.742
	Accuracy (%)	98.60	85.00	95.00		
	Precision (%)	97.73	85.73	95.40		
	Recall (%)	98.07	85.00	95.00		
	F1 Score (%)	97.87	84.03	94.98		

by analyzing Table III.6 and Figure III.7 we found that :

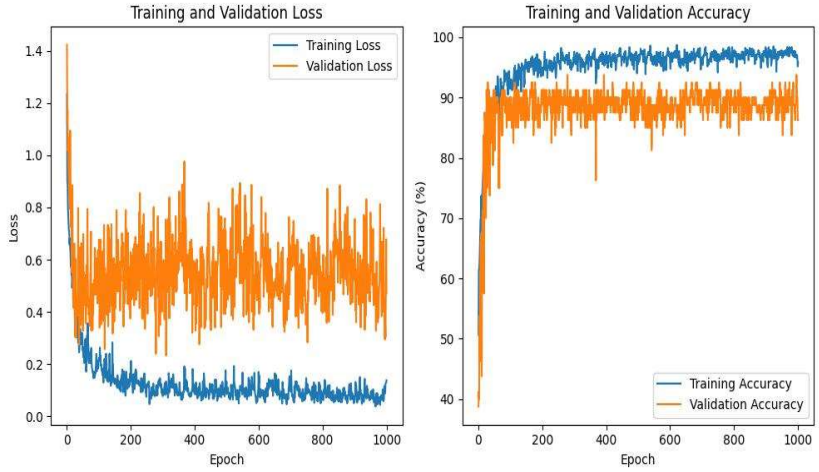
- Simple Vit demonstrates the most balanced and reliable performance, with (96.25%) test accuracy and a (96.30%) F1 score, while maintaining efficiency with only 13 million parameters and a compact size of 51.24 MB.
- Observing the training curves, as depicted in (Figure III.7 (a)), Simple Vit exhibits a consistent and stable learning process.
- Both the training loss (blue line) and validation loss (orange line) show a gradual decrease over epochs, indicating effective learning
- While there are some fluctuations in the validation loss, it generally tracks the training loss closely, suggesting good generalization.
- The training and validation accuracies also increase steadily and converge at a high level, further supporting the conclusion that Simple Vit learns well without

significant overfitting.

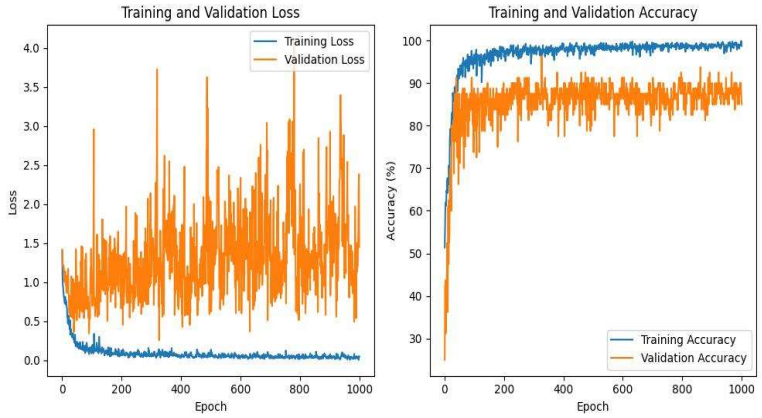
- This stable convergence of training and validation metrics reinforces the model's reliability and efficiency.
- Distillation achieves the highest test accuracy and F1 score (97.50%), indicating strong final performance, but its lower validation accuracy (86.25%) suggests potential overfitting or variability in generalization.
- In contrast to Vit1's stability, the training dynamics of Distillation, illustrated in (Figure III.7 (b)), reveal more erratic behavior.
- Although the training loss decreases smoothly and the training accuracy rises to a high level, the validation loss shows more pronounced fluctuations and even some upward spikes, particularly in later epochs.
- The validation accuracy also exhibits greater volatility and plateaus at a level considerably below the training accuracy.
- This inconsistency in the validation metrics, especially the fluctuations in validation loss and accuracy, indicates that Distillation , despite its strong test performance, may be more prone to overfitting the training data or exhibiting less stable generalization.



(a)



(b)



(c)

Figure III.7: The performance of models ((a) Simple ViT,(b) Distillation ,(c) AlexNet)on the BANANA dataset.

- In contrast, AlexNet reaches the highest training accuracy (98.60%) but underperforms on validation (85.00%) and test (95.00% accuracy, 94.98% F1 score), despite its large size (57 million parameters, 222.742 MB), which may not justify the marginal gain.
- (Figure III.7 (c)) clearly shows this overfitting tendency.
- The training loss decreases rapidly, and the training accuracy quickly approaches near-perfect levels.
- However, the validation loss plateaus early and remains significantly higher than the training loss, with substantial oscillations.
- The validation accuracy also plateaus at a much lower level than the training accuracy.
- This large disparity between training and validation performance, coupled with the rapid convergence of the training metrics, is a classic sign of overfitting, where the model essentially memorizes the training set and fails to generalize effectively.
- the higher validation loss for AlexNet (1.4525) compared to Simple ViT and Distillation, despite its strong training and test performance, suggests overfitting AlexNet fits the training data well but struggles to generalize on unseen validation data.
- Meanwhile, Simple ViT and Distillation show more balanced loss across all phases, reflecting better generalization. The lower test loss of Distillation (0.0697) indicates highly confident and accurate predictions during testing, even better than AlexNet's. This shows that while CNNs like AlexNet can perform very well on known data, ViT-based models tend to generalize better across diverse data distributions.

- Overall, Simple Vit is the most consistent and efficient, Distillation offers slightly better test performance at the cost of stability, and AlexNet appears less efficient and prone to overfitting on this dataset.
- The training curves provide crucial visual evidence to support these conclusions, highlighting the importance of evaluating not just final performance metrics but also the learning dynamics of the models.

III.7.4 Assessment of Methods

Across the three datasets Uncropped, Cropped, and the BANANA dataset the results highlight important differences in both accuracy and model size among Simple Vit , Distillation , and AlexNet. Simple Vit stands out as the most balanced model, combining strong performance with efficiency. It achieves test accuracies between (94.44%) and (96.25%) across all datasets while maintaining a compact size of approximately 13 million parameters and just 51 MB of storage. This makes it highly suitable for deployment in environments with limited computational resources, as it provides excellent accuracy without the overhead of a large model.

Distillation, with a very similar size to Simple Vit (around 13.1 million parameters and 51.2 MB), occasionally surpasses Simple Vit in terms of raw test accuracy, reaching up to 97.50%. However, its performance is less consistent, particularly on the Cropped dataset where its test accuracy drops to (90.74%). This suggests that while Distillation has the potential for high accuracy, it may suffer from generalization issues and might require further tuning to be as reliable as Simple Vit across different input conditions.

In contrast, AlexNet is significantly larger, with 57 million parameters and a storage footprint of 222.74 MB, making it the most resource-intensive model. Despite this, its test accuracy doesn't consistently outperform the smaller ViT models. Although AlexNet achieves 100% accuracy on some test sets, such perfect results raise

concerns about overfitting or possible data leakage, especially given the noticeable drop in validation or test scores on other datasets (95% test accuracy on the BANANA dataset). While AlexNet delivers strong results, its size and occasional inconsistencies limit its practicality. Distillation shows promise with high peak performance, but its generalization is less stable.Simple Vit proves to be the most reliable and efficient option, offering high accuracy with a lightweight architecture suitable for real-world applications,table III.7 compares the performance and size of three deep learning models AlexNet, Simple ViT, and Distillation ViT across three different datasets related to plant disease detection.

Table III.7: Comparison of model accuracy and size across different datasets.

Dataset	Model	Accuracy Test (%)	Model Size (MB)
Uncropped ZPDD	Alexnet	100%	222.742
	Simple Vit	96.23%	50.823
	Distillation Vit	98.11%	51.235
Cropped ZPDD	Alexnet	100%	22.742
	Simple Vit	94.44%	50.823
	Distillation Vit	90.74%	51.235
Banana leaf spot disease	Alexnet	95.00%	222.742
	Simple Vit	96.25%	51.235
	Distillation Vit	97.50%	52.742

III.7.5 Comparison with state-of-art methods

From the Table III.8 we can see the ViT model stands out with consistent excellence across all performance metrics. This consistent superiority over a wide range of models highlights not only the effectiveness of vision transformers for banana disease detection but also the crucial role of careful parameter tuning. Unlike several other approaches that perform well in only one or two metrics or lack complete evaluation data, the optimized ViT model shows strong, balanced performance, making it highly generalizable and reliable for real-world agricultural applications. These results confirm that transformer-based architectures, when properly optimized, can surpass both classical machine learning methods and traditional CNNs, paving the

way for more robust and accurate plant disease detection systems. Moreover, an added advantage of the ViT model is its relatively small storage footprint, which enhances its suitability for deployment in resource-constrained environments.

Table III.8: Performance comparison with other plant disease detection research on BANANA dataset [5].

Author	Year	Technique	Loss	Accuracy	Precision	Recall	F1 Score
Tahsin YASIN [79]	2023	Random Forest (VGG19)	/	96.8	96.80	96.80	96.80
Baki Bhuiyan et al. [80]	2023	ResNet-50	/	86.25	89.26	86.25	85.46
		Inception-V3	/	90.00	91.96	90.00	96.49
		VGG-16	/	95.00	95.45	95.00	98.27
		BananaSqueezeNet	/	96.25	96.54	96.25	98.75
Shetty et al. [81]	2024	HCA-YOLOv8	/	98.12	/	/	/
Ocimati [82]	2024	CNN Model	/	98.00	97.00	96.00	97.00
P.L. Arunima et al [83]	2024	Custom 9-layer CNN	/	95.00	/	/	/
A. Hayat, P. Baglat, F. Mendonça, et a [84]	2024	Custom CNN	/	94.00	/	/	/
Ebru Ergün [85]	2025	Vision Transformer (ViT) + Support Vector Machine (SVM)	/	**99.86**	/	/	/
Our Distilled ViT	**2025**	**Distillation**	**0.0697**	**97.50**	**97.56**	**97.50**	**97.50**

III.8 Conclusion

In this chapter, we compared distilled Vision Transformer models (Simple Vit and Distillation) with the classical CNN AlexNet on both uncropped and cropped versions of our proprietary tomato leaf dataset, as well as on a banana leaf disease dataset. Although AlexNet achieved very high accuracy and F1-scores, it required substantially more trainable parameters and storage space, making it less efficient for practical deployment. The ViT models, particularly Distillation using a patch size of 16×16 on 224×224 images, consistently demonstrated strong classification performance with competitive accuracy, precision, recall, and F1-scores. They also offered significant advantages in computational efficiency and reduced storage re-

quirements. Moreover, training on cropped images focusing on diseased areas further enhanced model performance and reduced training time. These results confirm that Vision Transformer distillation provides a powerful and efficient approach for plant disease detection, balancing high accuracy with practical resource demands. This makes ViT-based models promising candidates for scalable and deployable solutions in precision agriculture.

General Conclusion

The application of ViTs has marked a significant paradigm shift in automated plant disease detection, offering a powerful alternative to traditional convolutional neural networks (CNNs). By leveraging self-attention mechanisms, ViTs excel at capturing long-range spatial dependencies and global contextual information within plant leaf images, enabling more nuanced and accurate feature extraction critical for disease classification. Recent developments, such as lightweight and computationally efficient ViT variants), demonstrate that these models can achieve accuracy while remaining suitable for deployment on resource-constrained devices, including mobile platforms. The integration of multi-level attention modules and architectural innovations, such as inverted residual structures and inception-inspired blocks, further enhances the representational capacity and robustness of ViT-based systems. These multi-level attention mechanisms enable the models to focus dynamically on relevant features at different scales, improving their ability to discriminate subtle disease symptoms under varying conditions. Architectural enhancements contribute to more efficient feature extraction and better generalization across diverse datasets. Empirical evaluations confirm that ViTs consistently outperform conventional CNN architectures in classification accuracy, parameter efficiency, and inference speed. Unlike CNNs, which primarily capture local features through convolutional filters, ViTs model global relationships within images, providing richer contextual understanding essential for complex plant disease patterns. Additionally, ViTs' adaptability

to synthetic data augmentation and linear projection techniques facilitates improved generalization and scalability, making them well-suited for real-world agricultural applications. Despite these advances, challenges remain in optimizing ViT architectures for real-time, in-field deployment. Environmental variability such as lighting changes, leaf orientation, and background noise can affect model performance, necessitating robust adaptation strategies. Furthermore, reducing computational overhead without compromising accuracy is critical for enabling deployment on edge devices with limited resources. Key scientific challenges include:

- Designing intuitive and accessible mobile applications that enable rapid and accurate disease diagnosis directly in the field, even under resource-limited conditions.
- Enhancing model robustness and accuracy through architectural innovations, hyperparameter optimization, and expansion of diverse, high-quality datasets that capture real-world variability.
- rotocols, and accelerate the development of scalable, reliable, and user-friendly plant disease detection systems.

Future research should focus on refining model compression techniques, developing interpretable attention mechanisms to enhance transparency, and integrating ViTs with edge AI hardware for ubiquitous and efficient plant disease monitoring. In summary, Vision Transformers represent a transformative technology in precision agriculture, significantly advancing AI-driven plant disease diagnostics. Their superior ability to model complex spatial dependencies and contextual information promises enhanced early detection accuracy, reduced reliance on manual inspection, and ultimately supports sustainable agricultural productivity and global food security.

Bibliography

[1] Anuja Bhargava, Aasheesh Shukla, Om Prakash Goswami, Mohammed H Alsharif, Peerapong Uthansakul, and Monthippa Uthansakul. Plant leaf disease detection, classification, and diagnosis using computer vision and artificial intelligence: A review. *IEEE access*, 12:37443–37469, 2024.

[2] Folasade Olubusola Isinkaye, Michael Olusoji Olusanya, and Pramod Kumar Singh. Deep learning and content-based filtering techniques for improving plant disease identification and treatment recommendations: A comprehensive review. *Heliyon*, 2024.

[3] J. Abbas, N. Bibi, R. A. Naqvi, and D. Jeong. Abiotic plant stress and biotic plant diseases. In *Current Trends in Plant Pathology*, pages 57–78. Springer, 2024.

[4] Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Thipendra P Singh, and Ravi Tomar. Convolution in neural networks. In *Neuro-Symbolic Artificial Intelligence: Bridging Logic and Learning*, pages 125–139. Springer, 2024.

[5] Abdelmalik Ouamane, Ammar Chouchane, Yassine Himeur, Sami Miniaoui, Shadi Atalla, Wathiq Mansoor, and Hussain Al-Ahmad. Optimized vision transformers for superior plant disease detection. *IEEE Access*, 2025.

[6] Hannah Ritchie, Pablo Rosado, and Max Roser. Agricultural production. *Our world in data*, 2023.

[7] Prachi Pandey, Vadivelmurugan Irulappan, Muthukumar V Bagavathiannan, and Muthappa Senthil-Kumar. Impact of combined abiotic and biotic stresses on plant growth and avenues for crop improvement by exploiting physio-morphological traits. *Frontiers in plant science*, 8:537, 2017.

[8] Federico Martinelli, Riccardo Scalenghe, Salvatore Davino, Stefano Panno, Giuseppe Scuderi, Paolo Ruisi, Paolo Villa, Daniela Stroppiana, Mirco Boschetti, Luiz R Goulart, et al. Advanced methods of plant disease detection. a review. *Agronomy for sustainable development*, 35:1–25, 2015.

[9] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516, 2020.

[10] Scott Krig. Attention, transformers, hybrids, and ddns. In *Computer Vision Metrics: Survey, Taxonomy, and Analysis of Computer Vision, Visual Neuroscience, and Visual AI*, pages 551–607. Springer, 2025.

[11] Wei Zhang, Tingting Liu, Ziming Li, Yan Dai, and Xu Zhou. Multimodal knowledge graph embedding with self-attention graph neural network. *IEEE Transactions on Computational Social Systems*, 2025.

[12] Ishak Pacal, Ismail Kunduracioglu, Mehmet Hakki Alma, Muhammet Deveci, Seifedine Kadry, Jan Nedoma, Vlastimil Slany, and Radek Martinek. A systematic review of deep learning techniques for plant diseases. *Artificial Intelligence Review*, 57(11):304, 2024.

[13] Joshua Benjamin, Shaneya Miriyagalla, Oluwatosin Adebanjo, Akil Bonaparte, and Alimot Ottun. Interactions between plant parasitic nematodes and other harmful organisms. *Indian Phytopathology*, 77(3):599–614, 2024.

[14] Christopher P Randle, Brandi C Cannon, Amber L Faust, Angela K Hawkins, Sara E Cabrera, Stephen Lee, Michelle L Lewis, Amy A Perez, James Sopas, Timothy J Verastegui, et al. Host cues mediate growth and establishment of oak mistletoe (phoradendron leucarpum, viscaceae), an aerial parasitic plant. *Castanea*, 83(2):249–262, 2018.

[15] Lincoln Taiz, Eduardo Zeiger, Ian M Møller, and Angus Murphy. Plant physiology and development. 2015.

[16] William G Hopkins, Norman PA Hüner, et al. Introduction to plant physiology. 1995.

[17] Sagar Vetal and RS Khule. Tomato plant disease detection using image processing. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(6):293–297, 2017.

[18] Raju Ghosh, A Tarafdar, DR Chobe, US Sharath Chandran, S Rani, and M Sharma. Diagnostic techniques of soil borne plant diseases: recent advances and next generation evolutionary trends. In *Biological Forum–An International Journal*, volume 11, pages 1–13. Research Trend, 2019.

[19] Albert Khakimov, Ilias Salakhutdinov, Almas Omolikov, and Samad Utaganov. Traditional and current-prospective methods of agricultural plant diseases detection: A review. In *IOP Conference series: earth and environmental science*, volume 951, page 012002. IOP Publishing, 2022.

[20] Jihen Amara, Sheeba Samuel, and Birgitta König-Ries. Integrating domain knowledge for enhanced concept model explainability in plant disease classification. In *European Semantic Web Conference*, pages 289–306. Springer, 2024.

[21] Ishana Attri, Lalit Kumar Awasthi, Teek Parval Sharma, and Priyanka Rathee. A review of deep learning techniques used in agriculture. *Ecological Informatics*, 77:102217, 2023.

[22] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8:1–74, 2021.

[23] Chen-Lin Zhang and Jianxin Wu. Improving cnn linear layers with power mean non-linearity. *Pattern Recognition*, 89:12–21, 2019.

[24] Ahmed Ali Mohammed Al-Saffar, Hai Tao, and Mohammed Ahmed Talab. Review of deep convolution neural network in image classification. In *2017 International conference on radar, antenna, microwave, electronics, and telecommunications (ICRAMET)*, pages 26–31. IEEE, 2017.

[25] Jinzhu Lu, Lijuan Tan, and Huanyu Jiang. Review on convolutional neural network (cnn) applied to plant leaf disease classification. *Agriculture*, 11(8):707, 2021.

[26] Habiba N Ngugi, Absalom E Ezugwu, Andronicus A Akinyelu, and Laith Abualigah. Revolutionizing crop disease detection with computational deep learning: a comprehensive review. *Environmental Monitoring and Assessment*, 196(3):302, 2024.

[27] Abhishek Upadhyay, Narendra Singh Chandel, Krishna Pratap Singh, Subir Kumar Chakraborty, Balaji M Nandede, Mohit Kumar, A Subeesh, Konga Upendar, Ali Salem, and Ahmed Elbeltagi. Deep learning and computer vision in plant disease detection: a comprehensive review of techniques, models, and trends in precision agriculture. *Artificial Intelligence Review*, 58(3):1–64, 2025.

[28] K Anitha and S Srinivasan. Feature extraction and classification of plant leaf diseases using deep learning techniques. *Computers, Materials & Continua*, 73(1), 2022.

[29] Prameetha Pai, S Amutha, Mustafa Basthikodi, BM Ahamed Shafeeq, KM Chaitra, and Ananth Prabhu Gurpur. A twin cnn-based framework for optimized rice leaf disease classification with feature fusion. *Journal of Big Data*, 12(1):89, 2025.

[30] Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, and Sultan Almotairi. A comparison of pooling methods for convolutional neural networks. *Applied Sciences*, 12(17):8643, 2022.

[31] SH Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378:112–119, 2020.

[32] Usama Mokhtar, Nashwa El Bendary, Aboul Ella Hassenian, Eid Emary, Mahmoud A Mahmoud, Hesham Hefny, and Mohamed F Tolba. Svm-based detection of tomato leaves diseases. In *Intelligent Systems' 2014: Proceedings of the 7th IEEE International Conference Intelligent Systems IS'2014, September 24-26, 2014, Warsaw, Poland, Volume 2: Tools, Architectures, Systems, Applications*, pages 641–652. Springer, 2015.

[33] Amer FAH Alnuaimi and Tasnim HK Albaldawi. An overview of machine learning classification techniques. In *BIO Web of Conferences*, volume 97, page 00133. EDP Sciences, 2024.

[34] Sukhvir Kaur, Shreelekha Pandey, and Shivani Goel. Plants disease identification and classification through leaf images: A survey. *Archives of Computational Methods in Engineering*, 26:507–530, 2019.

[35] Min Jiang, Yuan Rao, Jingyao Zhang, and Yiming Shen. Automatic behavior recognition of group-housed goats using deep learning. *Computers and Electronics in Agriculture*, 177:105706, 2020.

[36] Louis Kouadio, Moussa El Jarroudi, Zineb Belabess, Salah-Eddine Laasli, Md Zohurul Kadir Roni, Ibn Dahou Idrissi Amine, Nourreddine Mokhtari, Fouad Mokrini, Jürgen Junk, and Rachid Lahlali. A review on uav-based applications for plant disease detection and monitoring. *Remote Sensing*, 15(17):4273, 2023.

[37] Wubetu Barud Demilie. Plant disease detection and classification techniques: a comparative study of the performances. *Journal of Big Data*, 11(1):5, 2024.

[38] Marko Arsenovic, Mirjana Karanovic, Srdjan Sladojevic, Andras Anderla, and Darko Stefanovic. Solving current limitations of deep learning based approaches for plant disease detection. *Symmetry*, 11(7):939, 2019.

[39] Imane Bouacida, Brahim Farou, Lynda Djakhdjakha, Hamid Seridi, and Muhammet Kurulay. Innovative deep learning approach for cross-crop plant disease detection: A generalized method for identifying unhealthy leaves. *Information Processing in Agriculture*, 12(1):54–67, 2025.

[40] Emmanuel Moupojou, Appolinaire Tagne, Florent Retraint, Anicet Tadonkemwa, Dongmo Wilfried, Hyppolite Tapamo, and Marcellin Nkenlifack. Fieldplant: A dataset of field plant images for plant disease detection and classification with deep learning. *IEEE Access*, 11:35398–35410, 2023.

[41] Mingle Xu, Ji-Eun Park, Jaehwan Lee, Jucheng Yang, and Sook Yoon. Plant disease recognition datasets in the age of deep learning: challenges and opportunities. *Frontiers in Plant Science*, 15:1452551, 2024.

[42] Guofeng Yang, Yong He, Yong Yang, and Beibei Xu. Fine-grained image classification for crop disease based on attention mechanism. *Frontiers in Plant Science*, 11:600854, 2020.

[43] Yao-Hong Tsai and Tse-Chuan Hsu. An effective deep neural network in edge computing enabled internet of things for plant diseases monitoring. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 695–699, 2024.

[44] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[45] Sana Parez, Naqqash Dilshad, Norah Saleh Alghamdi, Turki M Alanazi, and Jong Weon Lee. Visual intelligence in precision agriculture: Exploring plant disease detection via efficient vision transformers. *Sensors*, 23(15):6949, 2023.

[46] Poornima Singh Thakur, Pritee Khanna, Tanuja Sheorey, and Aparajita Ojha. Explainable vision transformer enabled convolutional neural network for plant disease identification: Plantxvit. *arXiv preprint arXiv:2207.07919*, 2022.

[47] Abhishek Sebastian, A Annis Fathima, R Pragna, S MadhanKumar, G Yaswanth Kannan, and Vinay Murali. Vital: An advanced framework for automated plant disease identification in leaf images using vision transformers and linear projection for feature reduction. In *International Conference on Computing and Machine Learning*, pages 31–45. Springer, 2024.

[48] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[49] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Abdul Rehman Khan, Hifsa Asif, Aqsa Asif, and Umair Farooq. A survey of the vision transformers and their cnn-transformer based variants. *Artificial Intelligence Review*, 56(Suppl 3):2917–2970, 2023.

[50] Qing Yun. Vision transformers (vits) for feature extraction and classification of ai-generated visual designs. *IEEE Access*, 2025.

[51] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10076–10085, 2020.

[52] Paria Mehrani and John K Tsotsos. Self-attention in vision transformers performs perceptual grouping, not attention. *Frontiers in Computer Science*, 5:1178450, 2023.

[53] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10033–10041, 2021.

[54] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[55] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.

[56] Sashank Sridhar and Sowmya Sanagavarapu. Multi-head self-attention transformer for dogecoin price prediction. In *2021 14th International Conference on Human System Interaction (HSI)*, pages 1–6. IEEE, 2021.

[57] Siyuan Lu, Meiqi Wang, Shuang Liang, Jun Lin, and Zhongfeng Wang. Hardware accelerator for multi-head attention and position-wise feed-forward in the transformer. In *2020 IEEE 33rd International System-on-Chip Conference (SOCC)*, pages 84–89. IEEE, 2020.

[58] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9):5521, 2023.

[59] Hang Li, Sufang Li, Jiguo Yu, Yubing Han, and Anming Dong. Plant disease and insect pest identification based on vision transformer. In *International conference on internet of things and machine learning (IoTML 2021)*, volume 12174, pages 194–201. SPIE, 2022.

[60] Wafaa H. Alwan and Sabah M. Alturfi. Multi-stage vision transformer and knowledge graph fusion for enhanced plant disease classification. *Computer Systems Science and Engineering*, 49(1):419–434, 2025.

[61] Guoqiang Li, Yuchao Wang, Qing Zhao, Peiyan Yuan, and Baofang Chang. Pmvt: a lightweight vision transformer for plant disease identification on mobile devices. *Frontiers in Plant Science*, 14:1256773, 2023.

[62] Aadarsh Kumar Singh, Akhil Rao, Pratik Chattopadhyay, Rahul Maurya, and Lokesh Singh. Effective plant disease diagnosis using vision transformer trained with leafy-generative adversarial network-generated images. *Expert Systems with Applications*, 254:124387, 2024.

[63] Yuchao Wang, Peiyan Yuan, and Baofang Chang. Pmvt: a lightweight vision transformer for plant disease. *Artificial Intelligence and Internet of Things for Smart Agriculture*, page 7, 2024.

[64] Malithi De Silva and Dane Brown. Multispectral plant disease detection with vision transformer–convolutional neural network hybrid approaches. *Sensors*, 23(20):8531, 2023.

[65] Eu-Tteum Baek. Attention score-based multi-vision transformer technique for plant disease classification. *Sensors*, 25(1):270, 2025.

[66] Md Ashraful Haque, Chandan Kumar Deb, Pushkar Gole, Sayantani Karmakar, Akshay Dheeraj, Mehraj Ul Din Shah, Subrata Dutta, MK Prasanna Kumar, and Sudeep Marwaha. An enhanced vision transformer network for efficient and accurate crop disease detection. *Expert Systems with Applications*, page 127743, 2025.

[67] Mengyao Zhang, Chaofan Liu, Zihan Li, and Baoquan Yin. From convolutional networks to vision transformers: Evolution of deep learning in agricultural pest and disease identification. *Agronomy*, 15(5):1079, 2025.

[68] Utpal Barman, Parismita Sarma, Mirzanur Rahman, Vaskar Deka, Swati Lahkar, Vaishali Sharma, and Manob Jyoti Saikia. Vit-smartagri: vision transformer and smartphone-based plant disease detection for smart agriculture. *Agronomy*, 14(2):327, 2024.

[69] Xingquan Shu, Limin Ma, and Fengqin Chang. Integrating hyperspectral images and lidar data using vision transformers for enhanced vegetation classification. *Forests*, 16(4):620, 2025.

[70] Moshiur Rahman Tonmoy, Md Mithun Hossain, Nilanjan Dey, and MF Mridha. Mobileplantvit: A mobile-friendly hybrid vit for generalized plant disease image classification. *arXiv preprint arXiv:2503.16628*, 2025.

[71] J Jayanthi and K Arun Kumar. Transformative impact of ai-driven computer vision in agriculture. In *Artificial Intelligence Techniques in Smart Agriculture*, pages 129–150. Springer, 2024.

[72] Shifat E Arman, Md Abdullahil Baki Bhuiyan, Hasan Muhammad Abdullah, Shariful Islam, Tahsin Tanha Chowdhury, and Md Arban Hossain. Bananalsd: A banana leaf images dataset for classification of banana leaf diseases using machine learning. *Data in Brief*, 50:109608, 2023.

[73] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022.

[74] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[75] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[76] Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, et al. A simple single-scale vision transformer for object detection and instance segmentation. In *European conference on computer vision*, pages 711–727. Springer, 2022.

[77] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[78] Xiaobing Han, Yanfei Zhong, Liqin Cao, and Liangpei Zhang. Pre-trained alexnet architecture with pyramid pooling and supervision for high spatial resolution remote sensing image scene classification. *Remote Sensing*, 9(8):848, 2017.

[79] Elham Yasin and Murat KOKLU. Utilizing random forests for the classification of pudina leaves through feature extraction with inceptionv3 and vgg19. In *Proceedings of the International Conference on New Trends in Applied Sciences*, volume 1, pages 1–8, 2023.

[80] Md Abdullahil Baki Bhuiyan, Hasan Muhammad Abdullah, Shifat E Arman, Sayed Saminur Rahman, and Kaies Al Mahmud. Bananasqueezenet: A very fast, lightweight convolutional neural network for the diagnosis of three prominent banana leaf diseases. *Smart Agricultural Technology*, 4:100214, 2023.

[81] Sahana Shetty and TR Mahesh. Skgdc: Effective segmentation based deep learning methodology for banana leaf, fruit, and stem disease prediction. *SN Computer Science*, 5(6):698, 2024.

[82] O. Walter, E. Sivalingam, and S. Nancy. Leveraging deep learning for early and accurate prediction of banana crop diseases: A classification and risk assessment framework. *International Journal of Computer Engineering in Research Trends (IJCERT)*, 11(4):46–57, 2023.

[83] PL Arunima, Pratheesh P Gopinath, PR Geetha Lekshmi, and M Esakkimuthu. Digital assessment of post-harvest nendran banana for faster grading: Cnn-based ripeness classification model. *Postharvest Biology and Technology*, 214:112972, 2024.

[84] Ahatsham Hayat, Preety Baglat, Fábio Mendonça, Sheikh Shanawaz Mostafa, and Fernando Morgado-Dias. Machine learning system for commercial banana harvesting. *Engineering Research Express*, 6(3):035202, 2024.

[85] Ebru Ergün. High precision banana variety identification using vision transformer based feature extraction and support vector machine. *Scientific Reports*, 15(1):10366, 2025.