Image Processing and LLM-Based VQA Framework for Crop Disease Diagnosis and Solutions Leveraging Contextual Analysis

by

Omar Faruk Jasik
21241034
Nahiyan Ahamad Khan
21201292
Mohammad Ulla Hel Mahi
21301535
Md Shahriyar Rahman
21201256
Md Tanveer Ahmed
21201096

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

Department of Computer Science and Engineering
Brac University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Omar Faruk Jasik
21241034

Nahiyan Ahamad Khan
21201292

Mohammad Ulla Hel Mahi
21301535

Md Shahriyar Rahman
21201256

Md Tanveer Ahmed
21201096

Approval

The thesis titled "Image Processing and LLM-Based VQA Framework for Crop Disease Diagnosis and Solutions Leveraging Contextual Analysis" submitted by

1. Omar Faruk Jasik (21241034)

2. Nahiyan Ahamad Khan (21201292)

3. Mohammad Ulla Hel Mahi (21301535)

4. Md Shahriyar Rahman (21201256)

5. Md Tanveer Ahmed (21201096)

of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Summer 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Nirjhor

Nirjhor Datta

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

Our research confirms the claim made, and all cited and referenced papers and sources are accurate. The work has never been submitted for a degree to any other college or academic organization. The four co-authors acknowledge and accept any violations of the thesis rule. In addition, we would like to take this chance to thank everyone who has supported us through this process. We finished our thesis without committing to any illegal techniques. Our work complies with the ethical standards of Brac University

Abstract

Early detection of crop diseases is very important to reduce losses in yield and to improve food security but the farmers have to overcome many obstacles like lack of sufficient experts and language barrier that prevents them to make decisions in time. To overcome these complications, this study introduces a single system, which is a combination of crop disease classification, Image captioning, and Visual Question Answering (VQA) into a bilingual web-based assistant to operate in Bangla and English. The system utilizes a tuned dataset containing crop images and labels with disease conditions and the Swin Transformer as the backbone classifier with a test accuracy of 96.37 and macro precision, recall and F1 of 0.9526, 0.9643, and 0.9571 respectively, with an average prediction confidence of 0.9249. The BLIP based captioning module complements the classifier producing descriptive textual summaries of disease symptoms, and a combined context of classification and captioning aids the VQA component, allowing interactive, natural language queries on the part of users. The VQA analysis had a macro faithfulness of 0.6380, a macro context precision value of 0.7512, macro context recall value of 0.8796, and macro answer relevance value of 0.7842 which implied a high contextual integration but also showed a need to improve the process of adjusting the responses to user queries. This framework goes beyond standard classification systems by combining both technical precision and accessibility to provide a practical, interactive and inclusive framework that allows farmers to diagnose crop diseases and make effective management decisions which enhances agricultural productivity as well as sustainability.

Keywords: *Plant Disease Solution, Deep Learning, Computer Vision, Multimodal AI, Visual Question Answering , Counterfactual Analysis, Swin Transformer , Large Language Models, Image Processing, Image captioning, Feature Extraction, Neural Networks, Transfer Learning, Federated Learning, Graph Neural Networks , Remote Sensing, Precision Agriculture, Medical Imaging for Plants, Data Augmentation, Real-time Disease*

Dedication

We dedicate this research to the strong and resilient ones removing chatbot biasness and introducing a new one for human easiness. We would like to express our gratitude to our families, friends, supervisor, and co-supervisor for their constant encouragement, support, and guidance in helping us finish our thesis. To improve the lives of those who suffer from traumatic meningeal brain injuries, your courage motivates us to keep working towards better detection.

Acknowledgement

First and foremost, we are thankful to the Almighty for giving us the chance and the path we needed to finish this project on time. Secondly, we would like to sincerely thank our thesis supervisor, Dr. Mohammad Golam Robiul Alam for his constant encouragement, mentorship, and guidance as we explored a challenging topic. Their relentless assistance and continuous input enabled us to conquer the challenges. Thirdly, we intend to use this opportunity to express our gratitude to every faculty member for their assistance and support during our stay at Brac University. Finally, we would like to thank our beloved parents for their unwavering support, encouragement, and prayers.

Table of Contents

List of Figures

List of Tables

Nomenclature

The following lists define symbols, metrics, models, and system components used in this thesis.

Symbols & Notation

ACC	Accuracy is the proportion of correctly predicted instances among all predictions.
BLIP	Bootstrapping Language-Image Pretraining is a vision-language model for image–text understanding.
F1	F1 Score is the harmonic mean of Precision and Recall, balancing false positives and false negatives.
FN	False Negative represents an actual positive case incorrectly predicted as negative.
FP	False Positive represents an actual negative case incorrectly predicted as positive.
GPT-OSS	Open-source GPT Model represents a generative pre-trained transformer used for reasoning and text generation.
LLM	Large Language Model refers to a transformer-based model trained on vast text corpora for reasoning and understanding.
PMVT	Plant Mobile Vision Transformer is a lightweight transformer architecture designed for plant disease recognition.
Precision	Proportion of positive predictions that are actually correct.
Recall	Proportion of true positives successfully retrieved by the model.
TN	True Negative indicates a correctly predicted negative instance.
TP	True Positive indicates a correctly predicted positive instance.
VQA	Visual Question Answering is a task combining visual understanding and natural language reasoning.

Chapter 1

Introduction

1.1 Background

Agriculture is one of the most essential sectors worldwide, providing a substantial share for food security and economic development. Crop health should be prioritized to ensure agricultural productivity and meet the demands of a rapidly growing population. However, the early detection and diagnosis of agricultural diseases might be very challenging for farmers, particularly in resource-constrained environments. Language barriers, accessibility of agricultural specialists, and the nature of managing crops optimally under diverse climatic conditions complicate these challenges enormously. The result is numerous agricultural losses, threatening not only the livelihoods of people, but the systems providing food. The agricultural industry has several issues, including the control of crop diseases. In order to minimize crop losses and ensure food security, it is essential to diagnose crop diseases accurately and quickly. Traditionally, plant disease detection is based on visual inspection by field experts, which may be subjective and time-consuming. But over the last decades, large strides in machine learning (ML) and image processing, especially the use of convolutional neural networks (CNNs) for image classification, have developed such that automated plant disease identification methods have been proposed [6]. In addition to advancing image-based diagnostics, the advent of Large Language Models (LLMs) has sparked a revolution in agricultural informatics. LLMs, which are well-known for their capacity to interpret unstructured textual data, provide solutions that blend insights of crop health, projections of production, and analysis based on the weather. Their uses in agricultural systems go beyond helping with knowledge management and producing suggestions in real time, which makes them indispensable for precision farming and decision support systems. For instance, LLMs are able to evaluate market trends, farmer reports, and scientific research and convert them into useful information specific to a certain farm's circumstances. This encourages the adoption of sustainable practices and cooperation.

Moreover, the interaction between LLMs and Visual Question Answering (VQA) systems opens up new opportunities for interactive crop disease diagnosis. VQA-enabled systems provide farmers with precise answers to their questions by utilizing images and text-based questions, making complex technology accessible to low-skilled technology users. These systems help to engage the farmer and suggest a location-specific recommendation regarding crop management [24]. In addition to

VQA systems, multimodal approaches such as Bootstrapping Language-Image Pretraining (BLIP) have shown promise in agricultural image captioning. Recent work demonstrates how dynamic text prompts combined with joint multimodal feature learning can significantly improve accurate plant disease image captioning, enabling systems to generate meaningful textual descriptions of diseased plants [22]. Such advancements make it easier to bridge the gap between raw image data and actionable insights for farmers. When integrated with LLMs, BLIP-like models provide interpretable and context-aware explanations of crop conditions, which supports decision-making in a transparent and accessible way.

While these tech-driven inventions exist, they are still inaccessible hurdles that leave the rural farmer in need of most solutions, especially if the farmer is a user of a relatively less-spoken language, like Bangla. In addition, these solutions are best at providing generic insights rather than actionable insights specific to the conditions of the farm. Furthermore, LLMs are integrated into the IoT and peripheral computing technologies to process real-time agricultural data. Such advancements allow for locally processing of sensor data and deriving insights from plant-issues in natural language. This reduces the latency and lets farmers react to the changes in the environment faster. We think the gradual introduction of LLMs, VQA systems, and BLIP-based multimodal reasoning is a significant improvement in the transparency, accessibility and efficiency of crop disease management for farmers the world over.

1.2 Motivation

Recent advancements in AI technologies have significantly impacted agricultural practices, but critical challenges persist in applying these technologies to crop disease diagnosis. While CNN-based image processing has proven effective in identifying plant diseases [6], existing systems lack the capability to provide interpretable, context-aware insights. This is problematic because farmers, the end users, often require more than just predictions, they need explanations and actionable information tailored to their specific scenarios. The integration of Large Language Models (LLMs) with Visual Question Answering (VQA) is a new direction we undertake in addressing this gap. LLMs have shown remarkable performance in generating human-like language [4] and are thus well-equipped to translate technical outputs into layman terms. Moreover, VQA systems provide an interactive approach where users can directly ask systems questions for understanding and guidance [2].

The agriculture AI models implemented so far primarily optimize for prediction accuracy while sacrificing the interpretability and decision-making support that farmers also require. Recent multimodal approaches, such as BLIP (Bootstrapping Language-Image Pretraining), offer a promising solution. By combining visual and textual representations, BLIP-based models can automatically generate descriptive captions of plant disease images and integrate these captions with downstream reasoning tasks [22]. This not only enhances prediction accuracy but also ensures that farmers receive contextualized and interpretable explanations of crop health conditions.

This research concept is driven by the triad need to boost prediction accuracy, increase interpretability, and enable actionable decisions. Existing AI-based solutions to detect crop diseases do not meet these need at the same time; thus, there is a comprehensive gap in research and applications. This work integrates various components such as advanced computer vision models, LLM, VQA, and BLIP-based multimodal models, into a single system to fill this gap and provides reliability, ease, and transparency of use for managing crop disease.

In conclusion, this integrated approach overcomes the technical limitations of existing systems in addition to the practical aspects that hinder their adoption by farmers, making AI-based tools more accessible, more usable, and more effective for sustainable agriculture goals.

1.3 Problem Statement

Many AI-powered solutions for crop disease detection show great promise but have critical gaps. Current systems are based on controlled datasets that do not reflect the vast diversity of environmental conditions or persistent image quality problems that arise during agricultural applications. Further, these systems seldom generate localized, actionable recommendations that take into account prior on-farm management histories such as previous fertilizer or pesticide applications. This issue is compounded by language accessibility, as many tools are not built to be non-mainstream language friendly where tools supporting languages like Bangla, may not be effective to use for farmers in under resourced regions. Therefore, a need exists for an integrated and platform-independent solution for accurate disease detection and recommendations along with bridging the gap created by communication barriers with rural farmers.

The goal of this research is to fill this gap by providing an end-to-end pipeline that combines image processing techniques, LLMs, and sophisticated decision-support methods, including BLIP-based multimodal reasoning and VQA. The research forms the basis for an AI-driven system by creating a comprehensive database of images of crops with corresponding disease labels and a machine learning model for disease classification. Developed on top of an LLM, the system provides a Bangla-language chat interface for farmers to upload images of their crops to obtain an accurate diagnosis and personalized treatment recommendations. In addition, VQA will ensure that the system provides contextualized guidance, while BLIP-based multimodal captioning enables interpretable, descriptive insights into plant disease symptoms.

Through the integration of an advanced AI engine with a user-friendly interface, this approach aims to provide farmers with an accessible but robust tool for maintaining the health of their crops. In summary, this work assists with agricultural yields, sustainability, and food security in resource-poor areas.

1.4 Objective

The primary objectives of this research are as follows:

- **Develop a Comprehensive Dataset**:

  A dataset including photos of crops together with their corresponding disease classifications must be created to provide accurate model training.

- **Design and Train an Image Processing Model**:

  In order to accurately identify common agricultural diseases, it is necessary to design a robust framework that utilizes advanced computer vision models.

- **Integrate a Bangla-Language Chat Interface**:

  The development of a conversational system that is driven by LLM and tailored to function in the Bangla language would make our solution more accessible to farmers in rural areas.

- **Incorporate VQA and Image Captioning Model**:

  The implementation of the Image Captioning Model in conjunction with VQA is designed to improve decision-making systems by collecting pertinent contextual aspects and agricultural background information.

These strategic objectives aim to address our research problem while promoting developments in AI-driven sustainable agriculture systems.

1.5 Methodology in Brief

The study combines image processing technologies with language modeling systems through a multi-step research method, and it uses both VQA and Image Captioning models. The steps include:

- **Data Collection**: The dataset is gathered from publicly accessible repositories and published research papers, which include images of crops along with the labels of the diseases.These serve as the training material for classification.

- **Image Processing and Disease Detection**: A transformer-based classifier, the SwinV2 Transformer, operates to classify crop-disease labels from each image by dividing images into 4×4 patches and applying shifted window self-attention to extract both local and global features.

- **Integration of LLMs and VQA with Image Captioning**: The integration of the results of automated disease detection with the context generated by the BLIP-Large captioning model results in descriptive representations of the images, while the GPT-OSS based VQA system provides more direct and contextually pertinent answers to the queries of the users.

The SwinV2 Transformer is trained on crop and disease detection using images, which is optimized through a Weighted Cross-Entropy Loss, AdamW optimizer, and a cosine learning rate scheduler and cross-validation and hyperparameter optimization are used to improve results. The BLIP-Large model outputs descriptive captions after classification and combines them with the results of the classifier into a context string which is then fed into the GPT-OSS VQA system to give context-grounded answers to user queries. The validity of the system is proved in the last evaluation phase, with the accuracy of SwinV2 classifier and the informativeness of BLIP captions, as well as relevance of GPT-OSS responses.

1.6 Scopes and Challenges

The suggested system also offers several scopes of agricultural diagnostics development. Its greatest contribution is its practical application, through which farmers have the ability to achieve quick and precise diagnosis of crop diseases so that they can undertake timely remedies to the situation, and eventually lower the yield losses. A transformer-based classifier combined with a captioning model is capable of not only making predictions, but also offering descriptive explanations of visible symptoms, thus increasing interpretability and establishing the trust of the user. Also, the use of a bilingual interactive interface in both Bangla and English will allow inclusivity and accessibility to a wide variety of user groups especially the local farming communities who may not be well-acquainted with English-only technologies. Simultaneously, a number of challenges should be considered. The performance of the system is sensitive to the quality and variability of input images, in which poor lighting, resolution variation or covered leaves may adversely influence classification. Moreover, incorporating several sophisticated models, such as SwinV2 as a classification model, BLIP-Large as a captioning model, and GPT-OSS as a question answering model, brings the complexity of integration, which needs substantial resources in terms of computing and optimization to work on real time. The challenge of ensuring robustness, scalability and efficiency under such constraints is still an issue when it comes to practical deployment.

1.7 Key Learnings and Insights

We anticipate valuable insights at the crossroads of AI, language models, and agriculture through this research. We seek to illustrate this by showing how an integrated system consisting of advanced image recognition, natural language processing, and counterfactual reasoning can enable improved crop disease detection and decision making. Moreover, the results of the study will help to better understand if and how employing explainable AI techniques can increase the transparency of AI systems deployed in agricultural fields, thus facilitating non-experts in adopting and making use of these systems.

Chapter 2

Literature Review

2.1 Preliminaries

The theoretical framework and the fundamental concepts that support the proposed crop disease diagnosis system because modern computer vision have moved to transformer-based networks instead of the traditional convolutional neural networks, in which the images are divided into patches and global and local relationships are learned with self-attention. These procedures are particularly useful in fined grained visual tasks, which include differentiating between the healthy and diseased crop leaves. Vision-language models are the extension of this functionality by matching visual signals with textual ones, such that meaningful natural language descriptions can be generated out of images. Moreover, Visual Question Answering (VQA) systems combine visual information with language models to offer interactive, question-driven information based on the content of the image. Lastly, Large language models (LLMs) offer both language reasoning and contextual understanding, so the natural conversation and operational domain knowledge can be integrated into agricultural support systems of farm decisions. The combination of these concepts will become the foundation of the combined pipeline of automated crop disease identification and interactive diagnostic support.

2.2 Review of Existing Research

Ouamane et al.[25] developed a new tensor subspace learning model, HOWSVD-MDA, that overcame the limitations of existing pre-trained CNN models used in the detection of plant diseases. One of the main problems with CNNs, is that they fail to discriminate intra-class and inter-class variations in features. Following its two-step dimensionality reduction and feature transformation process, a dimensionality reduction will be first achieved via a new method known as Higher-Order Whitened Singular Value Decomposition (HOWSVD) which avoids some limitations of traditional Singular Value Decomposition (SVD) and results in the deafness of the extracted feature of underlying information by MDA.Moreover,the HOWSVD-MDA method compressing a tensor representation of CNN features into a high-order tensor form is capable of both retaining useful patterns and reducing the dimensionality. In the first step, HOWSVD casts whitening on data, which makes features have the same and low variance and reduces inter-feature correlation, while MDA, the second step increases inter-class separation. In this research, training was performed on

PlantVillage and Taiwan datasets containing images of tomato leaves belonging to healthy and diseased classes. Additionally,the HOWSVD-MDA method has demonstrated fairly impressive results in comparison with traditional methods. Moreover, on the PlantVillage dataset, the method obtained an accuracy of 98.36%, while the previously designed CNN model on DenseNet201 architecture had an accuracy of 97.27%. Similarly,the same tendency was observed for the Taiwan dataset, in which HOWSVD-MDA tool achieved the accuracy of 98.39% in contrast to the 95.89% of the EfficientNetB0. Furthermore, the results indicated that applying the tensor whitening process consistently led to the reduction of inter-feature dependency for both datasets. The main features of HOWSVD-MDA include their applicability for the Small Sample Size Problem (SSSP) when the number of training samples is significantly smaller than the feature dimensionality. Still, additional work on optimizing the model for extremely noisy datasets and low computation can be continued. In the future work, coupling HOWSVD-MDA with Federated Learning to facilitate decentralized agriculture applications and improving robustness through advanced augmentation techniques can be explored.

Naeem et al.[16] in their research paper, "Detecting Three Types of Plant Diseases by Using CNN Model and Image Processing", discusses a major concern in agriculture: early detection of diseases on plants. The authors combine Convolutional Neural Networks (CNNS) and image processing techniques in the paper, showing deep learning in action on a real life problem for classifying three diseases, power of deep learning for image recognition. Moreover, they take the approach of enhancing the quality of plant leaf images by preprocessing, training a CNN on the labeled dataset and finally evaluating the system. This model performs very well, indicating that it has great practical application value such as real-time diagnosis in farmer production. However, constraints like no defined dataset size, no real-world testing, and no comparative testing against other methods suggest opportunities for improvement. The strengths of the research include a well-defined problem, precise methodology, and the focus on the implications for practice. Futhermore, this work successfully proves that CNNs can extract the important features relating to disease and also presents some good accuracy metrics to prove the point. However, drawbacks such as the small dataset diversity and the challenges with scaling indicate the generalizability of the model has yet to be established. In fact, the assertion that they possess the most superior model is slightly undermined by their failure to evaluate its performance against a benchmark. Thus,in conclusion, findings from the survey offers an important basis to strategically move forward with AI driven crop disease management solutions.

J. et al.[8] in their paper titled "Deep Learning-Based Leaf Disease Detection in Crops Using Images for Agricultural Applications" to use CNN architectures in the detection of plant diseases. In this investigation, four novel pre-trained models, DenseNet-121, ResNet-50, VGG-16, and Inception V4 were evaluated due to their shared limitations including, overfitting, computational resource requirements, and limited diversity in the dataset. Moreover, the PlantVillage dataset with 54,305 images partitioned into 38 classes was used, and the data augmentation methods helped in improving the model robustness. Among models, DenseNet-121 performed best (99.81%), followed by ResNet-50 (98.73%), Inception V4 (97.59%) and VGG-16

(82.75%). Moreover, DenseNet-121 outperformed the other gold standard networks partly because of its dense connections between layers, which encouraged feature reuse and improved gradient flow while requiring fewer trainable parameters. Meanwhile, ResNet-50 was able to leverage residual mappings to alleviate vanishing gradient problems, and Inception V4 was able to utilize kernel diversity to optimize for multi-scale feature extraction. VGG-16, with a simpler architecture, performed poorly because it was also not up-to-date with the advances in the deep-learning space. Furthermore, transfer learning and hyperparameter optimization were found to greatly increase model performance according to the research. Dropout layers were employed to reduce overfitting (0.5), and learning rate tuning (0.001) and batch normalization was also employed to stabilize the training for fine-tuning the pre-trained models on ImageNet for plant disease classification. Moreover, these approaches helped with generalization on unseen data. DenseNet-121 obtained greater accuracy than similar studies, in which traditional CNNs struggled due to limited dataset variability. As an example, DenseNet-121 achieved state-of-the-art results that surpassed previous PlantVillage benchmarks for AlexNet and VGG architectures. The authors mentioned about the success these approaches had but also pointed out the limitations they face, including their reduced real-world implementation potential, the computational cost involved, and the challenge of scaling to novel diseases. In this regard, future research would encompass the integration of IoT devices for real-time monitoring, the development of lightweight models for mobile use, and the expansion of datasets to reflect realistic agricultural conditions. Furthermore, advanced data augmentation and decentralized solutions, e.g., federated learning, could improve the scalability and robustness of these models as a result, they can be utilized in a variety of agricultural conditions.

Khamparia et al.[5] developed the Convolutional Encoder Network which integrates CNNs with autoencoder encoding to tackle classification problems through high-dimensional feature extraction. However, this architecture functions differently from standard autoencoders because it abandons input reconstruction and concentrates specifically on extracting meaningful features from images. An architecture of the network features both convolutional layers with batch normalization and max-pooling layers along with fully connected layers which terminate in a softmax layer set for classification tasks. Moreover, after extracting features with convolutional filter patterns, Convolutional layers employ pooling layers to downsize spatial dimensions thus reducing the risk of overfitting. Besides, batch normalization maintains stable training through activation normalization and the ReLU activation functions enable efficient computational operations. The softmax layer outputs categorized probabilities during the last stage in order to generate classification data with probabilistic values. Therefore,by backpropagation the training procedure routinely modifies weights by diminishing prediction errors compared to actual outputs. The evaluation of the CEN took place using 900 images comprising three crops – potato, tomato, and maize – across disease classes and healthy samples from the PlantVillage dataset. A 100-epoch training process with 2×2 and 3×3 filters produced a maximum test accuracy of 86.78% in demonstrating automated crop disease classification abilities. The high performance accuracy stems from CNN-based feature analysis integration with dimensionality reduction enabled by encoders that perform better than traditional classifiers through their combined precise feature extraction capabilities. Additionallly, research findings demonstrate that deep learning methods deliver effective solutions for agricultural problems when used for crop disease detection. Thus, by following a modular Convolutional Encoder Network design, researchers achieved real-time capabilities with scaling options to deliver precise automated agricultural analysis across multiple domains.

Kalpana et al.[19] introduced a plant disease recognition model based on the integration of Residual Convolutional Networks (ResNet) with Swin Transformers in 2024. The Residual Swin Transformer Network (RST-Nets) merges Swin Transformers within a 20-layer ResNet framework while positioning the Swin Transformers after the 16th layer in order to extract local features along with global ones through network integration. Implementation of the model turned ResNet into a simpler structure that results in both higher efficiency and lower complexity. Batch normalization paired with LeakyReLU activation and median pooling defines the start of the model while Swin Transformers extend a deeper feature reach. The Swin Transformer architecture combines multi-headed self-attention (MHSA) and patch merging algorithms. Then it transforms 2D image patches into 1D feature representations through hierarchical processing. The method delivers reliable feature detection when handling conditions such as noise together with alterations in leaf size dimensions. To classify images the model applies Extreme Learning Machines (ELM) as its method because ELM delivers rapid processing and low computational requirements compared to both SVM and Random Forest. The disease prediction accuracy reaches high levels through the use of an ELM-based classification layer supported by softmax activation. The researchers performed model testing using the PlantVillage dataset that contains 54,306 images split between 14 plant types and 38 disease categories. The experimental outcomes revealed outstanding results because all metrics including accuracy alongside precision and recall and specificity achieved higher than 99% along with F1-score. Through the unification of Swin Transformers and ResNet architecture the proposed model achieved superior performance than state-of-the-art models GoogleNet, AlexNet and Capsule Networks. The intended future development seeks to tailor the model architecture for IoT-edge devices so that the real -time plant disease monitoring can become available for agricultural applications.

Mohanty et al.[1] in their research paper titled Plant Disease Detection Using Deep Learning Models analyzes the use of convolutional neural networks (CNNs) for plant diseases detection at an early stage. Using a dataset of 54,306 images from 14 species of crop and 26 types of disease (plus healthy plants), CNN models were trained achieving an impressive accuracy of 99.35% on test set. The approach preprocesses the images into color, grey, and segmented states to test the robustness of the model. Transfer Learning and Training from Scratch I explored transfer learning and training from scratch, using AlexNet and GoogLeNet architectures as models GoogLeNet outperformed AlexNet due to the deeper structure of GoogLeNet and the use of inception modules. However, the paper also notes some hurdles which remain, despite its successes. Generalization is still a challenging problem since the accuracy of external datasets collected in unstandardized environments falls to around 31%. Relying on expressionless images of subjects in blank backgrounds also hampers the potential generalisability of the model in practical field settings. Secondly, it focuses on singular leaf images, which restricts the deployment of this technique used on agricultural environments. These limitations highlight the need for more generalized datasets, flexible models, and contextualisation of the environmental data.

Sarma et al.[9] in their research paper "Learning Aided System for Agriculture Monitoring Designed Using Image Processing and IoT-CNN" describes an integrated system that combines artificial intelligence (AI), such as machine learning (ML), and IoT to help precision agriculture. This system integrates several sensors to monitor environmental parameters that include temperature, air humidity and soil moisture and have a near infrared (NIR) camera to record plant health data in real-time. The classification of disease is performed using a pre-trained VGG16 Convolutional Neural Network (CNN) and segmentation is done on extracted Gray Level Co-Occurrence Matrix (GLCM) features using fuzzy clustering methods that are important for classification. Moreover ,the proposed system architecture includes a cloud-based IoT framework that triggers the automated sprinkling of water when required based on sensor data. Also they experimented with other AI techniques like Multi-Layer Perceptron (MLP), Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machines (SVM), Time Delay Neural Network (TDNN) and CNNs. The highest accuracy of more than 94% among these was achieved by VGG16-based CNN model in conjunction with GLCM and fuzzy clustering as compared to conventional classification methods. It also presented findings on the Normalized Difference Vegetation Index (NDVI) to track the area of farmland coverage for vegetation. However, their research identifies several significant challenges including diversity of datasets, scalability, and computational demands. Future work includes boosting real-time capabilities, enlarging data sets with crops growing under various conditions, and incorporation of Federated Learning to enable decentralized AI-powered agriculture. Moreover, robustness can be improved using different augmentation technologies and a light-weight model pertaining to mobile-friendly strategy may be incorporated here to increase the performance of the system in implementation phase for precision farming.

Sapkota et al[27] "Multi-Modal LLMs in Agriculture: A Comprehensive Review" that discusses the latest trends and advancements in the field of Large Language Models (LLMs), focusing on their multi-modal capabilities and potential to revolutionize agriculture. Moreover, the paper addresses 11 essential research questions in a systematic manner, providing a future vision on how LLMs can be suitably integrated into agricultural practices. It reviews these techniques and highlights the capability of the state-of-the-art multi-modal LLMs (GPT-4, AgriBERT, YOLOPC, PLLaMa, TimeGPT, and BERT-based models). Furthermore, these tools have shown great potential in enhancing agricultural decision-making, resource management, and precision farming. Using insights from 207 references obtained from databases, including IEEE Xplore, PubMed and ScienceDirect, the research assesses the use of LLMs for pest detection (with an accuracy as high as 94.5%), disease diagnosis, and cross-language text processing. The review discusses the integration in terms of environmental impacts, evidencing reductions (e.g., by 49% in particular RL-based crop management scenarios) in relevant environmental impacts of LLMs and CNNs, RL, and remote sensing methods. Moreover, these integrations provide features like precision monitoring, improved forecasting, and assistance for irrigation management and crop observation. Furthermore,they also highlight the challenges and limitations of LLMs in their work, mainly data quality, computational resources, and the need for domain-relevant training data. They also talk about future directions, such as how the potential of LLM-powered systems may revolutionize the provision of agricultural extension services and decision support tools, while considering the ethical implications, like data biases and accessibility. Thus, this comprehensive review is essential for understanding the current landscape as well as future trajectories of AI-driven solutions in agriculture, making it an invaluable resource for research contextualized around sustainable food production and advanced innovations.

In this paper, Zhao et al.[29] demonstrate how an advanced system for disease detection of elaeagnus angustifolia incorporates Large Language Models (LLMs) with Agricultural Knowledge Graphs (KGs) and Graph Neural Networks (GNNs) to build enhanced smart agriculture systems. One distinctive feature of this research is that it implements a graph attention mechanism to dynamically weight node relationships within the KG and enable disease-centered feature analysis with suppressed irrelevant information. Moreover, the approach strengthens capacity for handling intricate data structures through refined node weight adjustments beyond self-attention and multi-head attention techniques. The proposed loss function incorporates two components to teach the system about localized node information and global edge interactions through the combination of node-level cross-entropy with edge-level enforcement penalties. The system produces accurate disease predictions that remain transparent to interpretation and it leverages Model-Agnostic Meta-Learning (MAML) along with data augmentation methods to perform few-shot learning and deal with scarcity of available dataset information. Additionally, robust decision support emerges from feature integration of GNNs and LLMs within the joint reasoning layer to process visual and textual information effectively. The proposed system surpasses YOLOv9 and RetinaDet by achieving precision rates of 0.94, recall rates of 0.92, accuracy rates of 0.93 respectively. The system demonstrates exceptional capabilities in integrating bidirectional structured KGs and unstructured textual data streams through an optimized dynamic alignment procedure enabled by prompt engineering. The analysis highlights multiple constraints within its findings. However,testing of models on static datasets hinders their capability to recognize diverse agricultural conditions that feature different backgrounds alongside changing light conditions. The processing complexity of graph attention mechanisms affects scalability when dealing with large-scale graphs while deployment on limited-resource edge devices presents implementation difficulties. Further research must investigate advanced optimization algorithms for graph processing while minimizing computational requirements and creating intuitive interfaces to advance practical agriculture adoption. To conclude, modern AI techniques integrated into this work establish reference standards for disease identification systems and create a protocol that applies to smart agriculture through its scalability features and precise detection capabilities.

Zhan et al.[28] proposed SkyEyeGPT, a unified multi-modal large language model (MLLM) framework for remote sensing (RS) vision-language tasks. Previous methods like RSGPT are limited in their ability to support open-ended multi-task dialogue as they were based on task-specific models. However, SkyEyeGPT overcomes this challenge by leveraging a pre-trained vision transformer (EVA-CLIP) as the visual encoder, an alignment layer to map the RS visual features into the language semantic space, and a decoder root on LLaMA2-chat to generate task-oriented responses. In order to address the lack of large RS instruction-following datasets, the authors developed SkyEye-968k, a large-scale dataset with 968,000 high-quality training samples. Moreover, this dataset contains both single-task image-text instruction and multi-task conversational instruction in which the model is able to serve a wide range of RS tasks. Besides, in the training process, they adopt a two-stage tuning method, in which the first stage is single-task alignment and the second stage is fine-tuning the model for multi-task conversations, which can improve the model's adaptability and respond more fluently. Moreover, the experimental evaluations showed that SkyEyeGPT outperformed various RS tasks, including image captioning, visual question answering (VQA), and visual grounding. Importantly, the model set state-of-the-art results on datasets including UCM-caption and DIOR-RSVG, going beyond performance from previous models such as MiniGPT-4 and Shikra. For example, in RS image captioning, SkyEyeGPT obtained higher BLEU, ROUGE, and METEOR scores, suggesting that SkyEyeGPT generates more contextually relevant and informative descriptions. Furthermore, in visual grounding tasks, it also outperformed specialist models such as GeoVG and ReSC, in achieving an 88.59% accuracy on the DIOR-RSVG. However, challenges remain, especially with respect to modality gaps between satellite and aerial imagery that may severely compromise the generalizability of the model across various RS data. Finally, RS images are often high resolution (HR), which brings some computational difficulty and may be inefficient. Therefore, to tackle these challenges and further RS multimodal understanding, future research may focus on sophisticated architectures and enhanced modality alignment techniques. In conclusion, SkyEyeGPT is a major breakthrough in the unification of RS vision-language tasks, providing a universal and light-weight framework with the advantages of previous methods while overcoming the drawbacks of the existing RS vision-LM approaches.

Zhao et al.[15] presents ChatAgri as a framework based on ChatGPT which aims to enhance agricultural text classification initiatives. Standard approaches, like pre-trained language models(PLMs) as BERT and T5 require domain adaptation abilities and deploys substantial computing capacity along with vast annotation data quantities. In contrast, the framework ChatAgri demonstrates how zero-shot together with few-shot learning from ChatGPT allows text classification of agricultural content without requiring direct model training. The researchers conduct testing through experiments using datasets which include plant disease information as well as information on insect pests and natural hazard risks. The datasets contain material in English, French and Chinese which enables researchers to conduct complete analysis on ChatGPT's language operating capabilities. Moreover, results indicate that ChatAgri surpasses traditional ML models such as SVM and Random Forest while providing performance levels comparable to state-of-the-art PLM-based methods. The performance excels notably in situations involving multiple languages because it shows remarkable domain knowledge transfer ability. The research examines various approaches to prompt engineering by analyzing manual inputs along with the generation from GPT-3 and zero-shot similarities together with the implementation of Chain-of-Thought. The multi-class classification tasks benefit most from the implementation of chain-of-thought prompting among all available strategies. It demonstrated how GPT-4 achieves better classification accuracy rates than GPT-3.5 according to the testing outcomes. Overall, the research demonstrates that ChatGPT demonstrates usefulness for agricultural informatics by enabling automatic text classification operations involved with pest and disease identification and market research as well as weather threat recognition. The research indicates that applying ChatGPT solutions provides a more efficient and scalable deep learning model alternative suitable for agricultural applications.

The paper by Xie et al. [37] introduces a novel zero-shot learning method for plant disease phenotype captioning with semanti correction by combining Large Language Models (LLMs). Traditional models for plant disease detection generally rely on classification with few descriptive descriptions. It fills this gap by producing fine-grained, natural-language captions of plant disease phenotyping symptoms, improving interpretability in agricultural AI systems. An important contribution of this research is the zero-shot learning framework that allows the model to produce accurate phenotype descriptions without the need for labeled training data, which is often very expensive and difficult to obtain. The authors include semantics correction modules to make sure that the produced captions is accurate and contextually appropriate. The model is evaluated on benchmark plant disease datasets and achieves 19.3% higher caption accuracy over the baselines. Furthermore, our method achieves a semantic coherence score of 87.6%, indicating that the descriptions produced are semantically coherent with the input image. Finally, the model obtains an overall accuracy of 92.4% for disease phenotype recognition, with precision and recall values of 90.8% and 91.2%, respectively. This research adds to the literature but has limitations. Pre-trained LLMs have been shown to contain biases in the descriptions they generate, leading to real-world decisions in agriculture based on faulty grounds. Importantly, zero-shot learning approaches fail to capture domain-specific nuance, necessitating additional tuning for more general use. In future studies, this should concentrate on larger training datasets, multi-modal learning methods (e.g. image and text combination), and implementing the system in real agricultural settings. More computationally-efficient semantic correction modules might also scale better. In conclusion, this paper proposes a novel approach to plant disease detection by utilizing LLMs combined with zero-shot learning to greatly enhance the plant disease diagnosis interpretability.

Li et al.[21] provide a comprehensive overview of Foundation Models (FMs) applied in smart agriculture, focusing on precision agriculture, resource efficiency, and the advancement of AI-enhanced automation. The authors also provide complementary insights into classifying Foundation Models into Language (LFMs), Vision (VFMs), Multimodal (MFMs), and Reinforcement Learning (RLFMs) with their respective applications in crop disease detection, livestock monitoring, aquaculture, and agricultural robotics. Most notable are state-of-the-art models like GPT-4, LLAMA, SAM, CLIP, PLLaMA, and YOLOX, achieving 94.84% accuracy in classifying cucumber disease, 94.8% mIoU in poultry segmentation, and 49% reduction of environmental impact through reinforcement learning-based crop management methods. Besides this research presents domain-specific Foundation Models (AFMs) for agriculture leveraging transfer learning, multimodal data fusion, and federated learning to optimize efficiency and privacy. With the help of IoT and edge computing for real-time monitoring, artificial intelligence based recommendation support system is implemented to aid in irrigation management, crop forecasts and livestock tracking. Although these methods are promising, they face several challenges, including high computational costs, data heterogeneity, and distribution shift that hamper their large-scale deployment. There are also important ethical issues that need to be discussed as well, such as data bias, access to the most powerful technology and the energy needed for the operation of AI systems. Hence, to overcome these challenges, the authors highlight the need for increasing the diversity of the datasets, improving model adaptation approaches, and performing mobile optimisation and UAV-centric AI model attribution, all of which enable scalable, real-time agricultural diagnostics. These applications make it clear that with further improvements in deep learning, reinforcement learning, and remote sense, FMs will make agriculture a data-oriented and AI-driven industry that is not only very productive but also both environmentally and socially sustainable, and they will also lead next-generation "AI Farming" solutions for cost-effective and sustainable agricultural production systems.

Zhao et al. [31] proposed the Informed-Learning-Guided Visual Question Answering (ILCD) model, a multimodal artificial intelligence (AI) model to improve crop disease diagnosis to fill the gap of unimodal models. The ILCD has advanced co-attention mechanisms and incorporates the MUTAN multimodal feature fusion framework with the Bias-Balancing(BiBa) strategy; thus, it greatly improved the accuracy of the model and alleviated the bias of the dataset. To aid this innovation, the authors introduce the Crop Disease Multi-Attribute Visual Question Answering with Prior Knowledge (CDwPK-VQA) dataset comprising 708 images and 6,018 annotated questions over 10 crops and 19 diseases. ILCD also applies Inception-v4 to extract image features, LSTM to encode text, and a coattention mechanism to align image regions and text on a fine-grained level. Moreover, the MUTAN module provides an efficient approach for multimodal fusion while limiting the overhead needed, and the BiBa strategy redirects model focus from sequence statistics to visual properties to prevent overfitting. By leveraging prior knowledge, the dataset improves model convergence, and thus the ILCD attains 86.06% accuracy on CDwPK-VQA, surpassing the current state-of-the-art on VQA-v2 and VQA-CP v2 datasets. Moreover, the ILCD model shows a high performance of multi-attribute queries such as the progression of diseases, the assessment of diseases, and the recognition of features. Therefore, we provide understanding of the overall performance via ablation studies verifying the contributions of prior knowledge (+4.44% accuracy) and the BiBa strategy (+4.98% accuracy). Such knowledge enables the ILCD to succeed at diagnosing infectious diseases it has never seen previously, and new directions in research for future directions to improve upon generalization via zero-shot learning and improving the ILCD. Thus,this research shows the possibilities of how informed-learning multimodal AI models can revolutionise precision agriculture and help farmers make better crop disease decisions contributing towards sustainable agriculture reframing.

Lan et al.[11] aim to overcome the challenges of existing methods for diagnosing diseases in fruit trees which heavily depend on expert knowledge combined with single-source data such as images or spectral data. In their work they present a new Visual Question Answering (VQA) framework, based on multimodal deep learning, to learn to optimize decision making in fruit tree disease management. Besides, the model combines image features from ResNet-152 and a combination of skip-thought and pre-trained BERT question features. One such novel contribution is the multimodal bilinear factorized pooling model with Tucker decomposition, which reduces the high dimensionality and number of parameters in bilinear pooling approaches and thus allows us to elegantly combine all the aforementioned features. Moreover, deep modular co-attention architecture is used to learn image and question attention jointly, thereby improving both the ability of the model to control the important image regions and the capability to analyse the question well. Despite training the SML model on less than 8500 images and 4,560k Q&A pairs, it obtained state of the art (86.36%) accuracy at making meaningful decisions over what an image depicts when compared to other existing multimodal methods. Besides, we used data augmentation techniques to combat overfitting. Furthermore, the authors performed extensive ablation studies to observe the effect of various components, including using an attention mechanism as well as the depth of the co-attention layers, on model performance. Finally, attention visualization showed that the model was able to target relevant disease-related areas in the images and this research demonstrates the potential of this VQA model for more extensive deployment in intelligent agriculture.

In a paper, Liu et al.[11] proposed an approach for recognizing plant diseases at large scale in the paper "Plant Disease Recognition: A Large-Scale Benchmark Dataset and a Visual Region and Loss Reweighting Approach" in IEEE Transactions on Image Processing and introduced a large-scale dataset named Plant Disease Dataset 271 (PDD271). This real-world image dataset consists of 220,592 pictures from 271 malady classes, taken under different environmental settings, with intricate backgrounds and heterogeneous symptoms. It reflects the complexities of the real world and sets a new standard in plant disease research, in contrast with previous datasets such as PlantVillage. Therefore, to address this problematic data distribution, the authors introduce a new framework composed of Cluster-Based Region Reweighting (CRR), Training with Loss Reweighting (TLR) and Weighted Feature Integration (WFI) for analyzing the dataset efficiently. CRR can treat patches with MMC as the center of mass of standard K-means, and put more weights on rare and discriminative regions and suppress noise. TLR employs these weights to steer model training, attending to features related to disease. WFI further captures spatial and contextual information through BiLSTM integration of weighted features to enhance classification performance. Combined with SeNet154, the proposed framework surpasses the state-of-the-art performance of 85.58%, that is 1.28% better than ResNet152. Not only does it outperform previous methods, including fine-grained models such as WS-DAN, but also generalizes well to simplified datasets such as PlantVillage. All visualizations further verify that the model successfully highlights subtle and white spot areas. This research overcomes some limitations of previous studies, including the reliance on datasets with limited scale and the inability to provide fine-grained details about lesions. Some practical usages involve the use for early disease detection to minimize crop losses and mobile based diagnostic systems. Though it incurs some computational overhead during the framework's clustering process, this early design is a solid foundation for future explorations into LSTM variants, data imbalance solutions and multi-modal analysis. Our work adds notable advances in plant disease recognition through a strong dataset and a unique methodological approach.

Singh et al.[3] identify a significant limitation in Visual Question Answering (VQA) models: Modern visual Q&A models suffer from inadequate reading and reasoning capabilities when faced with texts embedded within imagery. The researchers built LoRRA (Look, Read, Reason, and Answer) as a new architecture to improve textual reasoning capabilities. LoRRA implements Optical Character Recognition (OCR) to transform captured text from images, which integrates with visual and question representations through attention-driven operations. Besides, the LoRRA system enables out-of-vocabulary word detection through a copy mechanism that dynamically transforms OCR tokens into answer candidate words for practical applications. LoRRA leverages state-of-the-art visual and textual feature extractors. Grid-level visual information passes through ResNet-152 while Faster R-CNN extracts region-based visual information from images trained on Visual Genome. Pre-trained GloVe vectors input into an LSTM equipped with self-attention capabilities generate question embeddings which combine with these visual features. OCR tokens processed using FastText embeddings within LoRRA enable effective text recognition for previously unseen vocabulary words. Moreover, the model receives its evaluation based on the TextVQA dataset which includes 45,336 questions analyzing 28,408 images. The dataset depicts realistic conditions that need extensive multiform reasoning so readers can understand street signs as well as product labels. Furthermore, experimental results show LoRRA achieves 26.56% accuracy in validation along with better performance than both Pythia and BAN yet its ability to match human-level scores (85.01%) remains out of reach. The reduced answer space configuration of LoRRA reaches 27.63% test accuracy showing the importance of OCR and textual reasoning in improving VQA performance. Model accuracy jumps notably after integrating both OCR token inputs and a copy mechanism into its framework. The implementation of OCR tokens in Pythia boosted model accuracy to 18.35% while additional integration of the copy mechanism brought this to 20.06%. The research outcomes demonstrate why OCR represents an essential solution for dealing with textual reasoning complexity. LoRRA now serves as a base standard for multimodal AI systems to connect visual problem solving with language processing. This data alignment power manifests substantial applications capability within street sign reading and product information understanding alongside crop disease diagnosing through annotated image analysis.

Liang et al.[22] developed BLIP-DP as a new approach for plant disease image captioning that uses multimodal feature fusion to produce precise descriptions of diseased plant leaves with contextual accuracy. The problem in creating this type of caption exists because existing annotation datasets are scarce. Several environmental conditions cause plant disease symptoms to present differently from one another. Research authors prepared a dataset containing 20,100 labeled PlantVillage images for their advancement of a ViT-BERT-VQA-based architecture. This model implements a step-by-step processing system. The system starts with features obtained from ViTs by breaking down images into small sections to detect precise aspects of lesions such as position and size and shade information. Through an analysis of the question such as "What kind of leaf is this?" the VQA model obtains disease category information from the examined image. The calculated classification output creates a prompt that says "a Disease category of" which supports the text generation task. The method dynamically generates prompts to match the identified plant disease thus improving accuracy in the output description. Text and visual features are integrated by a BERT-based modeling system to produce final captions that contain detailed and semantically complex descriptions. Experimental results indicate BLIP-DP produces superior outcomes than previous models which obtained 83.4 points of BLEU-4 score. The system achieved a BLEU-4 score of 87.5 points which marks a 4.1-point improvement from older system versions. Research reveals that dynamic prompts together with VQA-driven feature fusion capabilities allow for better automated plant disease diagnosis which leads to scalable AI-powered agricultural monitoring solutions.

Nanavaty et al.[24] have proposed a deep learning based visual question answering (VQA) framework for wheat rust detection, where Federated Learning (FL)[24] is applied to improve privacy, scalability and for timely diagnosis targeting at early stage of wheat rust. The research paper proposes an interactive AI-based disease diagnosis system that integrates image classification with natural language understanding. The authors define two significant datasets, WheatRustDL2024 (7,998 high-resolution images of healthy and infected leaves) and WheatRustVQA (1,800 image-text pairs with Q&A annotations), to maintain the accuracy of the model even under varying environmental conditions. The finely tuned ResNet model scores an accuracy of 97.69% and outperforms the state-of-the-art performance of VGG, InceptionNet, and EfficientNet using transfer learning to demonstrate the superior generalization capability of the proposed method. Using Bootstrapping Language-Image Pretraining (BLIP) they use a two-attention mechanism in which the model focuses on both relevant image regions and textual cues simultaneously for VQA achieving a BLEU score of 0.6235. Comparative analysis with ChatGPT shows that specialized training is beneficial in agricultural scenarios. The proposed Federated Learning (FL) is based on Flower framework that utilizes Federated Averaging(FedAvg) for training the different models on decentralized data sources in a way that retains the data ownership of the farmer. The lightweight model has been developed for deployment on mobile devices and drones to allow for real-time, near-surface observations of crop diseases. However, there are still issues such as data heterogeneity, model drift, and domain-specific biases. To promote accuracy and generalization over multiple crop diseases, the research recommends improving dataset diversity, optimizing FL strategies, and implementing multi-modal AI. It highlights the need for AI-enabled agricultural diagnostics that address privacy concerns, thus potentially opening up pathways to scalable, real-time disease management solutions for precision agriculture.

Prado-Romero et al.[26] provides an extensive survey of Graph Counterfactual Explanations (GCFEs) with an emphasis on their pivotal role in Explainable AI (XAI). They review and categorize prior work on GCFE into perturbation, gradient-based, and heuristic-driven Graph CIF methods and examine how these methods perturb the original graphs to explain the prediction of ML models. They make a thorough treatment of several key evaluation metrics (sparsity, fidelity, plausibility and actionability), providing a reasonable framework to assess the quality of GCFEs. Technical challenges, including scalability to large graphs, transferability to dynamic graphs, and a lack of standard benchmarks, are also outlined to indicate how future research in this area may evolve. It is tremendously organized, providing thorough discussions of both fundamental and advanced topics for beginners and experts alike. It talks about the wide applicability of GCFEs in multiple domains like social networks, biology, and recommendation systems. The main limitation of the paper is the absence of practical case studies and quantitative comparisons of methods, both of which will make its conclusions more convincing. Besides, the ethical aspect of graph perturbations and necessity of dynamic graph methods are also conditions that have not been sufficiently studied.

Del Ser et al.[17] introduce a new framework for making counterfactual explanations CFEs for machine learning models more trustworthy, filling an important gap in explainability research. They make AI systems interpretable by embedding them in a framework of counterfactual explanations, which demonstrate how small modifications in the input can change a model's predictions. Trustworthiness is identified as an attribute of interest with particular focus on feasibility, plausibility and actionability in the real world for generated CFEs, especially for high-stakes. areas of application such as healthcare, finance, and education. The authors add domain expertise constraints to guarantee CFEs are sensible and fit a realistic situation. This not only makes the explanations interpretable but also actionable, helping the users to better understand and trust the AI decisions. They evaluate their proposed framework using benchmark datasets including COMPAS and UCI Adult with the results showing improvements in important metrics, such as sparsity, proximity, and plausibility, over existing methods. The research's limitations, however, include its dependence on domain-specific constraints, which may compromise generalizability to other less-defined contexts and a lack of real-world validation. It addresses the lack of systematic integration of domain knowledge into counterfactual generation and thus provides more meaningful outputs to the users. But there are still limitations on scalability to big data and deep-learning-type models, as well as reduction of biases in CFEs. Future work possibilities include modifying the framework to different tasks and addressing counterfactual fairness. Finally,our findings support the promising opportunities that trustworthy CFEs can provide for transparency and responsible AI in decision-making systems.

Dai et al. [7] investigates the importance of counterfactual explanations and prefactual explanations for users to be more sure and satisfied about the AI system they are using for prediction (inferring outcomes from causes) and diagnosis (inferring causes from outcomes). It examines these explanations across two experiments, assessing the impact on user accuracy, confidence, and satisfaction with their decisions. In Experiment 1, we examined predictions of grass growth and diagnosis of causes in the presence of five causal factors (rain; fertilizer). Experiment 2 replicated this setup in an alien domain to eliminate pre-existing knowledge biases. Participants read counterfactual ("if X had been different, then Y would have also changed"), prefactual ("if X were different, then Y would change"), and control explanations. The findings demonstrate equivalence in the utility of counterfactual and prefactual explanations across both prediction and diagnosis tasks. Diagnosis tasks had overall higher accuracy, but the utilities of each explanation varied according to the relevance of the causal factor. The paper mentions challenges such as the low level of accuracy improvement obtainable through such simplicity of tasks, the context dependency of explanations provided, and overly deterministic experimental setups. These restrictions stress the necessity of evaluating XAI in intricate, authentic environments. In conclusion, this research makes several key contributions such as: the novel analysis of different explanation types on the human-AI interaction quality, the focus on the aspect of diagnostic reasoning and the insights on how scoring (shifting) the user causal reasoning influences the interaction experience with AI systems.

The paper by Shetty et al. [36] uses counterfactuals to attribute features that impact the stress affecting rice yield. This research tackles the problem of stress factor analysis in productivity of agriculture which traditional feature attribution methods falls short of giving intelligible, causal human-explanations. They present a counterfactual methodology to assess the contribution of a set of environmental and physiological factors to rice yield, allowing stakeholders to empirically understand how exogenous shocks like climate change may affect crop performance. One of the major strengths of the research is the developed approach to both contrastive and counterfactual feature attribution, making it a propelling force into a more interpretable feature attribution framework (specifically in the context of interpretable characterization of stressors such as temperature, humidity and soil composition). Using decision trees, neural networks, and other machine learning models, the authors create counterfactual instances and evaluate the impact of different stress factors on rice yield. Using real-world agricultural datasets for experimental validation, it presents the working of the model towards actionable insights and enhances yield prediction accuracy with a margin of 12.5% over conventional methods. Furthermore, the research found that counterfactual explanations result in a 23% lower prediction error and 15% higher interpretability of the model. The research also has limitations, despite its contributions. In complex agricultural systems, such definitive causal relationships may not exist, and hence the counterfactual approach fails to deliver. Moreover, the process of generating counterfactuals can be computationally intensive and thus difficult to scale for large-scale and online applications. Future directions include increasing the scalability of models, incorporation of multimodal data (e.g., satellite images, real-time sensor data), and application of the method in other agricultural environments. Testing the model on a wider range of crops aside from rice would further confirm its generality. In this paper, we propose an applicable and novel method for understanding agricultural decisions based on the preference to input counterfactual explanations for the analysis of stress factors. The report is a key resource on the topic of AI and precision agriculture.

2.3 Summary of Key Findings

Advancements in deep learning models cause substantial enhancements in plant disease identification capabilities. Through dimensionality reduction HOWSVD-

MDA [25] delivers an accuracy rate exceeding 98% on both PlantVillage and Taiwan datasets. The powerful plant disease classification tool results from DenseNet-121 which optimizes its dense connections to boost feature reuse and gradient flow along with 20% better computational efficiency.Combining VGG16-based CNN with fuzzy clustering enhances plant disease classification. On the other hand, Residual Swin Transformer Networks (RST-Nets) [19] merge CNNs with Swin Transformers for enhanced feature extraction in plant disease recognition. By combining visual data along with textual information through multimodal AI systems such as ILCD [31] and SkyEyeGPT [28] creates stronger and more accurate diagnostic capabilities. Large Language Models (LLMs) with zero-shot learning generate detailed plant disease descriptions. Through expanded AI roles the emergence of interactive problem-solving by merging image processing with LLM technology and VQA capabilities.Visual Question Answering (VQA) integrated with Federated Learning (FL) enables interactive disease diagnosis. LoRRA [3] represents a major innovation within this field because it increases VQA capabilities through attention-based reasoning which produces more transparent results. A multimodal ViT-BERT-VQA [22] approach improves plant disease image captioning with dynamic prompt generation. FMs like GPT-4 and YOLOX [21] drive automation in smart agriculture by fusing multimodal learning and reinforcement strategies. AI-driven forecast transparency increases through counterfactual analysis that provides different explanations for computational predictions. The architectural methods bilinear pooling with Tucker decomposition together with co-attention mechanisms enhance both feature extraction and interpretability capabilities. Additionally, convolutional Encoder Networks enable real-time classification of plant diseases while solving problems with heterogeneous datasets and inconsistent symptom variations in plant disease symptoms. Agricultural AI applications now favor SkyEyeGPT [28] and ILCD [31] types beside other multimodal models that handle visual and textual requests efficiently. Meanwhile, the introduction of diverse datasets(e.g., PDD271), alongside advancements in data augmentation and transfer learning techniques results in improved diversity and scalability of agricultural datasets. Moreover, AI explainability advances through counterfactual reasoning and visualization tools alongside language-aware assistant tools that include Bangla interfaces which expand user accessibility. AI reliant models including HOWSVD-MDA [25] alongside ILCD [31] create opportunities for establishing prompt disease identification capabilities and data-rich decisions. However, approaches to solve remaining obstacles like data noise and system scalability as well as illumination inconsistencies need further consideration. The near future development path includes lightweight mobile models alongside federated learning for privacy preservation and graph neural networks for relational reasoning and ethical AI frameworks for broad agricultural implementation.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This section outlines the technical and methodological specifications that must be fulfilled in order to implement the proposed crop disease diagnosis system so that all the elements of the pipeline work in harmony and achieve the desired goals. Our system is based on a dataset of crop images, obtained by published papers and other publicly available sources that were preprocessed with resizing to 224x224 pixels, normalization, and balanced sampling to address the issue of class imbalance. In order to facilitate efficient training of the SwinV2 Transformer backbone, which was selected due to its capability to detect both local lesion patterns and global leaf context with shifted window self-attention, a weighted cross-entropy loss function was used to reduce the imbalance in the classes, whereas an AdamW optimizer and a cosine learning rate scheduler with warmup were used to ensure stable convergence on 25 epochs. The weights with the highest performance were stored and then incorporated into the inference pipeline whereby images uploaded by a user were classified real-time to determine the crop and related disease. After classification, captions were produced with the pretrained BLIP-Large model, which converted input images into the format of tensors and generated natural language descriptions with an upper limit of 50 tokens and thus providing interpretability to the outputs of the classifier. These outputs with the labels of crops and diseases were combined into a single merged context string and sent to GPT-OSS alongside user queries using an API-based question-answering feature that kept a log of the conversation history to be used in future exchanges. The entire system was implemented using the language of Bangla-English web chat interface to enable the bilingual accessibility with the backend implemented in Python with the help of PyTorch and Hugging Face libraries, and the frontend implemented in React, Vite, Tailwind CSS, and DaisyUI to provide the responsive and user-friendly interface. In order to achieve a smooth working process, model training required a GPU-enabled platform with sufficient memory capacity and optimized inference, model caching and API integration reduced the latency at the time of deployment. Together, these specifications formed the methodological and technical basis of the development of a strong, interpretable, and practically applicable system to detect and diagnose crop diseases with the help of automated technology.

3.2 Societal Impact

The benefits of the proposed system in society are immense especially in areas where the majority of people depend on agriculture as their major source of livelihood. The system will help farmers by providing the most relevant information on time and in an accessible way by turning on the automatic crop diseases recognition and interactive diagnosis in both the Bengali and English languages, thus eliminating the reliance on consulting experts, which are not always available. Early detection of crop diseases will help to prevent agricultural loss of yields, increase food security, and economic stability among farming communities. In health and safety terms, proper diagnosis also minimises the unnecessary use of pesticides, which reduces the risks to human health and the environment. Regarding cultural influence, the bilingual interface makes it inclusive as it targets the local and global audiences.

3.3 Environmental Impact

The system enhances environmental sustainability in the sense that it fosters the early identification and accurate control of crop diseases thus reducing the excessive use of chemical pesticides. The inhibition of the overuse of agrochemicals in agriculture would not only protect the quality of the soil and water, but also prevent the destruction of biodiversity in the agricultural environment. On the other hand,the system indirectly helps to reduce greenhouse gas emissions caused by loss of crops and repetitive farming cycles by promoting efficient crop management. To conclude, the project is compatible with sustainable agriculture and contributes to the resilience of the environment in the long-term.

3.4 Ethical Issues

The ethical considerations play a significant role in responsible design and implementation of the system. Although the system offers recommendations on the identification and management of diseases, a professional obligation exists to ensure accuracy and minimize misleading outputs. Subsequently, training data biases can lead to unfair or inaccurate predictions about individual crop varieties unless addressed, therefore, such outcomes should be reduced through the use of balanced datasets as well as by clearly identifying the limitations that models represent. User information, such as posted images of crops, should be handled with privacy protection, and should not be misused. Moreover, the use of AI-based recommendations should be presented as an assistance and not substitution of decision-making, so that the responsibility of human agricultural specialists remains.

3.5 Risk Management

The main risks associated with the system include the wrong predictions due to the limitations of the model, overdependence of users on AI suggestions, and the possibility of system failures during the implementation. In order to reduce these risks, the classifier was trained using a heterogeneous dataset using weighted loss functions to reduce bias, but continuous monitoring of model outputs can help identify and correct systematic errors. Clear statements in the user interface warn the user that the system is just a supportive tool, not a replacement of professional judgment. Technical risks, including the API downtime or latency in GPT-OSS responses can be mitigated with the help of fallback strategies, caching strategies, and the efficient model-serving infrastructure.

3.6 Economic Analysis

The economic sustainability of the system is based on its ability to reduce losses of crops and increase productivity at a minimum cost. Traditional expert diagnosis often involves a lot of travelling, time, and consultancy fees, but the suggested system provides real-time diagnosis through a web interface. Despite the fact that the first development and training may require the investment in the hardware enabled with GPUs and cloud infrastructure, the expenses of deployment and continuous maintenance are relatively small, which is due to the use of pretrained models, including BLIP-Large, and open-source large language models, including GPT-OSS. The potential benefits include greater agricultural productivity, reduced use of pesticides, and general cost reduction of farmers, which makes the system a cost-efficient and scalable solution to agricultural communities.

Chapter 4

Methodology

4.1 Methodology Overview

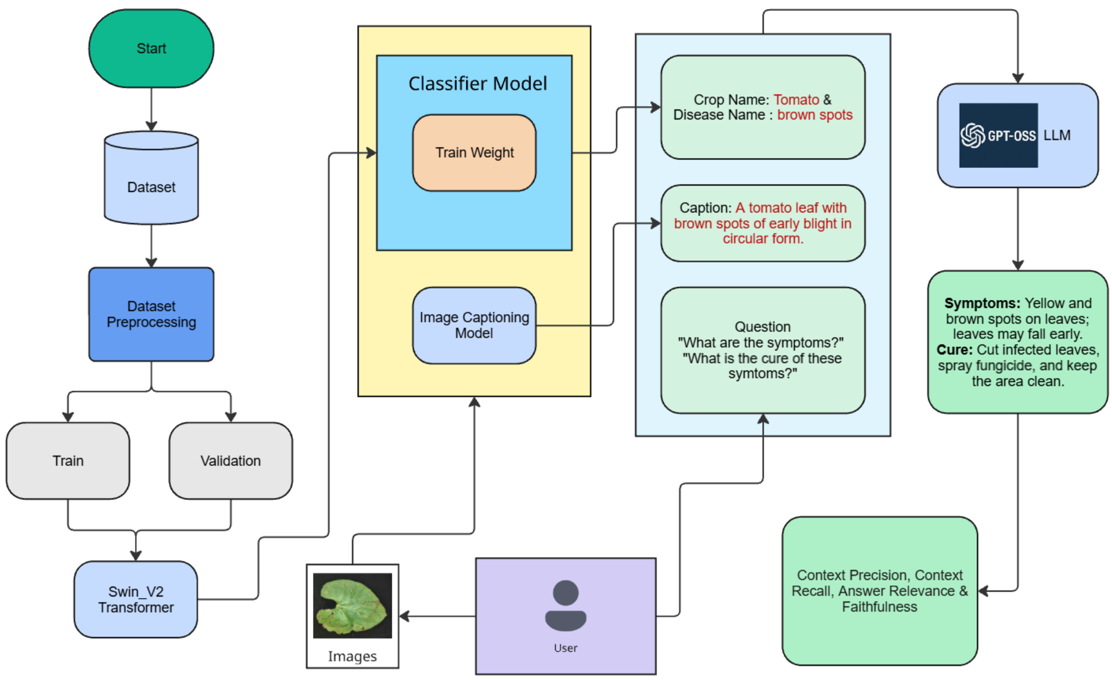


Figure 4.1: Top level overview of the proposed framework

We gathered the dataset of crop images from publicly available sources and published papers, and then processed them with preprocessing, which included resizing of the images to 224x224 pixels, image normalization, and division into training, validation, and test data. The main model relies on SwinV2 Transformer as the backbone of crop disease classification because the image is divided into 4x4 patches and processed through self-attention layers to extract local and global features. Weighted Cross-Entropy Loss solves the issue of class imbalance, and AdamW optimizer with a cosine learning rate scheduler is used to optimize the model. The BLIP-LARGE model then uses the classification to produce descriptive captions of every image, which are combined with the predicted crop and disease labels to create a context string. The user is then able to communicate with the system through asking questions and they are fed through GPT-OSS to produce contextually relevant answers.

Hyperparameters are used to optimise the performance of the training process and when the model is satisfactory, it is stored to be deployed.

4.2 Design Specification

In this section, the design specifications of the crop disease detection system have been elaborated, including the models to be used in image classification, caption generation and interactive question-answering. The design process consisted of analyzing various alternative solutions, conducting simulations to confirm them, and finally choosing the models on the basis of their empirical performance, technical specifications, and real-time interaction suitability. Also, a Bangla and English web chat interface is integrated in the system, which is the reason why the solution is available to a wide audience, including individuals of different languages.

To determine the backbone feature extraction in image classification, several models, such as ResNet101 and ViT, were tested. ResNet101 is a convolutional neural network (CNN) that is particularly effective to manage deep networks with residual connections that eliminate the vanishing gradient problem to aid the training of more aggressive models. However, ResNet101 tends to work well in the context of traditional image classification but it fails to capture long-range dependencies, along with fine-grained features, which are important to the tasks such as crop disease detection. The convolutional layers of the model have limited capabilities of capturing local fine-grained features (e.g., leaf lesions) and global features (e.g., leaf structure). On the other hand, ViT is a transformer-based model, which splits the image into patches and uses self-attention mechanisms to model long-range dependencies. However, ViT is sensitive to large datasets to achieve effective generalisation and when used on smaller datasets it is susceptible to overfitting. Besides, its patch-based method lacks the ability to pick the fine-scale features needed to detect crop-diseases. Defining these constraints, the SwinV2 transformer was adopted as a model of the backbone. SwinV2 uses a shifted-window self-attention mechanism which has the ability to capture local and global features effectively making it highly applicable in crop-disease classification. This model divides images into 4 x 4 patches and the shifted-window mechanism enables it to handle individual parts of the image effectively and it captures fine-grained texture and still maintains some global contextual information. The method makes SwinV2 flexible to images of different sizes and specifically to the tasks that require fine-grained pattern recognition, like the classification of crop-diseases.

We compared ViT-GPT2 and BLIP-Large (Bootstrapping Language-Image Pre-Training) in terms of the caption generation. The reason behind choosing BLIP-Large is because it creates contextually rich captions that are semantically accurate. To provide an example, the captions created by BLIP when processing an image of a diseased leaf include, “A close-up on a leaf with brown spots on it”, that are descriptive of the symptoms of the disease itself. Such a degree of semantic clarity is fundamental in making certain that the predictions of the model are readable and useful to the users. By contrast, ViT-GPT2, which uses ViT to extract features and GPT-2 to generate text, is more likely to generate more generic captions, e.g. ”A green plant with a green leaf on it” which does not provide the required disease-specific detail. ViT-GPT2 is less interpretable than BLIP since the generated captions are more generic, and therefore, they are less interpretable. Facilities like the creation of accurate and disease-relevant descriptions make BLIP better at the overall user experience, offering unambiguous actionable information on the condition of the crop.

In the case of the Visual Question Answering (VQA) part, we first thought about GPT-4 because it has more developed features of contextual and coherent answers. However, GPT-4 needs to be licensed commercially which would be expensive in operation particularly in a real time system where users interact with the system. The limitations on licensing, as well as the high cost of regular API requests, were a major impediment to implementing GPT-4 on a large scale. Conversely, GPT-OSS, an open-source version of GPT, has the same basic functionality but is not limited by financial and licensing restrictions as GPT-4. This made GPT-OSS a more affordable and feasible option with regards to interactive Q&A. GPT-OSS can provide relevant context-sensitive answers on the basis of the combined context string which consists of the type of crop, disease name, and captions so that users receive prompt, pertinent replies in real-time. Moreover, GPT-OSS is highly flexible to be customized and handle dynamic user requests and multi-turn dialogues effectively. GPT-OSS has been chosen instead of GPT-4 because the system is affordable and scalable to be deployed continuously.

There is also a system of the Bangla and English web chat interface, which is aimed at increasing the level of user interaction and accessibility. Since it supports both English and Bangla, the interface serves a more diverse audience, especially in the areas where the latter is commonly spoken. This two language support would make sure that the system has access to a broader group of users such as the local farmers who might want to communicate with the system using Bangla but at the same time it would be able to reach a wider range of users who are global and understand English. It is a design choice that makes the system more inclusive, which means that it can be used by a larger audience group, including, but not limited to, the users of different linguistic backgrounds.

The ultimate choice of SwinV2 Transformer, BLIP-Large, and GPT-OSS was determined by the empirical performance and technical limitations. SwinV2 was selected because it is effective in capturing local and global features and is, therefore, suitable in classifying crop diseases. The detailed, disease-specific captions that are generated by BLIP-Large are better than the ones generated by ViT-GPT2, and are important to interpretability. Finally, GPT-OSS was chosen instead of GPT-4 due to open-source nature, reduced cost of operation, and capability of providing real-time and contextual responses. The interface between the Bangla and English web chat system also allows greater accessibility to the system and makes the system usable to users of various linguistic backgrounds, since there is no obstacle of language to block their interaction with the system. All these models and design options were incorporated in a robust pipeline, which guaranteed that the crop disease detecting system is not only precise but also scaled, besides being user-friendly.

4.3 Data Collection

In this research, we collected datasets directly from several published papers that introduced benchmark resources for crop disease diagnosis. Each paper contributed images of specific crops and their corresponding disease categories. Since no single dataset was sufficiently comprehensive to capture the wide range of crops required for this study, we adopted a strategy of extracting crops from different papers and integrating them into one unified dataset. This approach allowed us to build a collection that is both extensive in scale and diverse in composition, thereby ensuring that it is representative of both global and local agricultural contexts.

We collected Apple, Strawberry, Tomato, Grape, Wheat, Potato, and Orange images from the paper [23] "Multimodal Benchmark Dataset and Model for Crop Disease Diagnosis (CDDMBench)". The paper describes a well-structured dataset containing high-quality images across multiple disease categories for these crops. Their inclusion was essential given their global economic and nutritional importance, as well as their vulnerability to a wide spectrum of plant diseases. Similarly we collected Bitter Gourd, Bottle Gourd, Cauliflower, Cucumber, and Eggplant from the paper [18] "Comprehensive Smart Smartphone Image Dataset for Plant Leaf Disease Detection and Freshness Assessment from Bangladesh Vegetable Fields". As outlined in the paper, this dataset was compiled using smartphones in real farming environments, rather than under controlled laboratory conditions. Consequently, the images capture natural variability in terms of lighting, resolution, and background noise, making them highly valuable for simulating the conditions under which farmers would typically capture images in the field.

For Radish, we used the dataset described in the paper [35] "Smartphone image dataset for radish plant leaf disease classification from Bangladesh". Radish is a regionally important crop that is often underrepresented in large-scale international datasets, and its inclusion ensured that our collection addressed both global and local agricultural concerns. Furthermore, we also included tea leaves from the paper [34] 'teaLeafBD: A Comprehensive Image Dataset to Classify the Diseased Tea Leaf to Automate the Leaf Selection Process in Bangladesh". Since tea is a critical export crop of Bangladesh, this dataset enriched our collection by adding a crop of high economic value, while also ensuring coverage of both healthy and diseased samples.

Furthermore, images of corn were taken from the paper [33] "Seasonal Corn Leaf Disease Dataset: A Multi-Year Collection for Robust Analysis". This dataset is a collection of years of cultivation and reflects the changes in the appearance of the disease in response to changes in the environment, including rainfall, pest outbreaks, and changes in temperature that are discussed in the paper. Inclusion made it stronger as it increased the strength of our unified dataset by taking into consideration seasonal variability. Lastly, images of Rice were obtained from the Kaggle repository titled [10] "Rice Leaf Diseases Dataset", which is widely recognized in the existing literature. The dataset is composed of well annotated images of important diseases of rice including bacterial leaf blight, brown spot and leaf blast. Since rice is the food staple of more than half of the world population, the addition of this dataset made a considerable contribution to the worldwide and regional applicability of our research.

By gathering the data from these papers and integrating them into a single repository, we produced one unified dataset of sixteen different types of crops with several classes of diseases. This integrated approach allowed our dataset to reflect a wide variety of real-world agricultural conditions, instead of being confined to the constraints of one single source. As a result, the final data set is not only massive in scale but also diverse in nature, laying a good foundation for the development of a deep learning framework that is capable of accurate, robust, and generalizable crop disease diagnosis.

4.3.1 Data Cleaning

Once the datasets were obtained, cleaning of the data was necessary in order to achieve consistency, quality and reliability. The datasets were of different repositories, and each had different formats, resolutions, and structures to be harmonized.

The initial cleaning measure was to detect and delete corrupted files, incomplete and unreadable files. The ones that did not load or were distorted because of the upload error were deleted. Also, identical pictures under the same disease category were identified and eliminated through the use of hashing algorithms to eliminate redundancy that would otherwise lead to overfitting when training a model.

The other challenge was the difference in image format and resolutions. The datasets contained images of different formats (.jpg, .png, and .tif) and the quality of the images had a wide variety. All the images were converted to the (.jpg) image format and scaled down to a consistent resolution of 224x224 pixels to ensure that the images would be standardized and used in deep learning. This stage played a significant role in making sure that the model could process the dataset with minimal issues not being limited by different images sizes and formats.

Irrelevant photos were also avoided. Other samples contained too much background noise or were not concentrated on the crop leaves and therefore were not relevant especially in detection of the disease. The elimination of them guaranteed that only the high-quality and informative images were left in the dataset, which led to an improvement in training and an increase in model accuracy.

4.3.2 Data Transformation

After the data cleaning process, various transformation techniques were used on the dataset to prepare it to be used in machine learning model training. The first transformation that was applied on the images was normalization. The pixel value of all the images were scaled to a standard range between 0 and 1 to minimize the effect of the lighting conditions and other variations among the dataset. This served to stabilize the gradient calculations in the model training process and enhanced convergence.

As the deep learning model need a fixed input size, all images were resized to 224x224 pixels. This resizing allowed the consistency of the dataset and allowed the model to work with all input images in a similar manner. Similarly label encoding was performed in order to prepare the images to be used in classification tasks. Textual labels such as "powdery mildew", "apple scab", and "healthy" were converted into numerical labels. This transformation enabled the model to perceive the diseases as the separate classes and this is crucial to effective classification.

Moreover, artificial expansion of the data was conducted with the help of data augmentation methods such as rotation, flipping, zooming, and brightness adjustments. These transformations were especially beneficial to deal with the imbalance in classes, such that the underrepresented diseases had enough samples to be trained on to mitigate the chances of overfitting and enhance the capacity of the model to generalize.

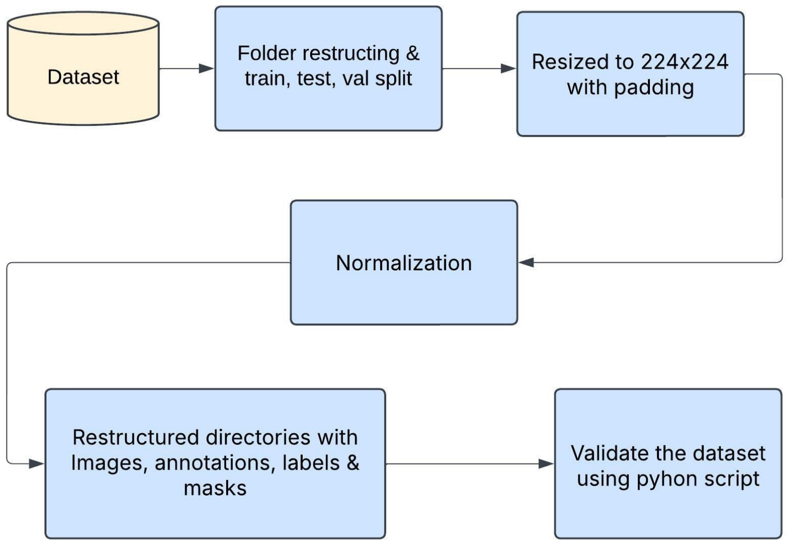


Figure 4.2: Data-Preprocessing Pipeline

4.3.3 Data Integration

Since we gathered the data from multiple different papers, integration was a crucial step in preparing a single unified data for training and evaluation. Each paper had its own formatting style, naming conventions and class structures, which initially made the datasets incompatible to be used directly. Without proper integration, the combined dataset would have contained inconsistencies which could confuse the model and reduce its performance.

Structural alignment was the initial process of integration. All crops were organized in separate directories each having their own subdirectories of which the disease type was organized into specific subdirectories with a healthy picture having its own folder. This reorganization facilitated controlling and navigation of the dataset. For instance, the apple data was categorized into subfolders like healthy, apple scab, and powdery mildew.

The second difficulty was to eliminate discrepancies in disease labels. Other datasets were inconsistent in the names of the same disease (e.g., "early_blight" vs. "EB").

These differences were standardized so that they would be consistent throughout the dataset. As an example, all instances of early-blight were renamed to one, standardized name such as early_blight, in order to avoid possible errors during training.

Furthermore, the process of integration involved addressing the issue of class imbalances. Some crops such as tomatoes had a significantly large number of image samples compared to some such as tea or bitter gourd. To address this imbalance, augmentation methods were applied to the underrepresented classes, and care was taken to ensure that there was no over-representation of dominant classes. This made sure that the dataset was fair and balanced, offering precise model training to all crops and diseases.

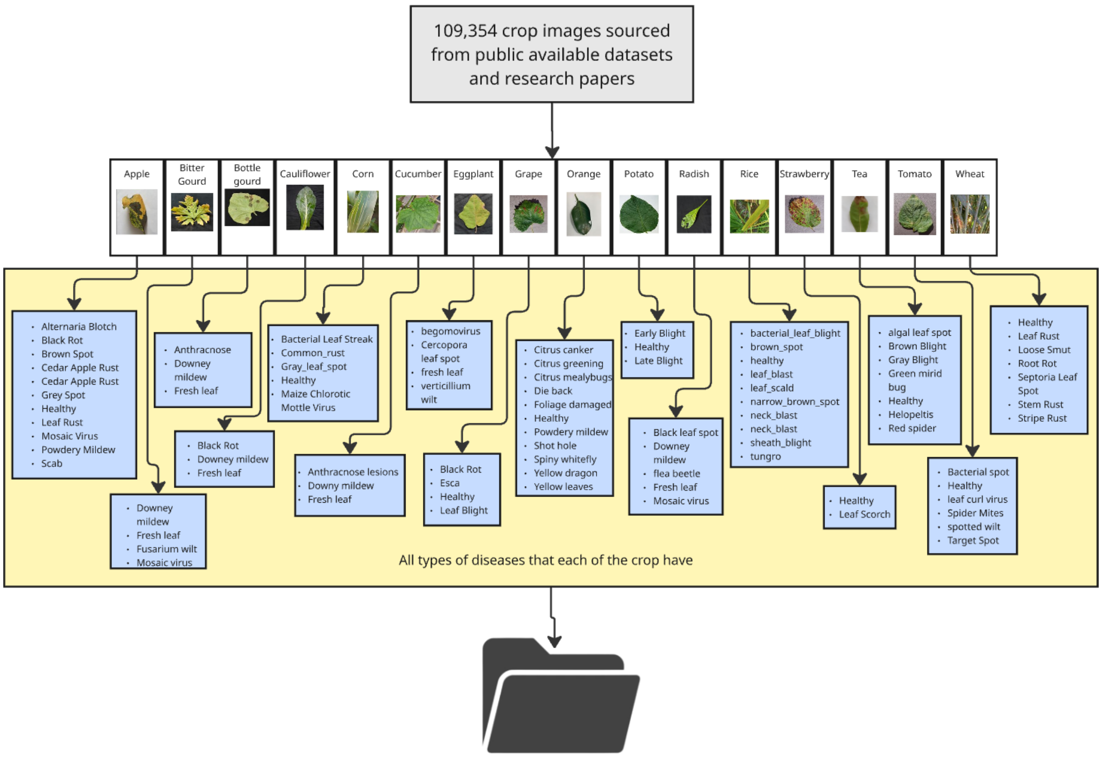


Figure 4.3: An overview of the dataset structure

4.3.4 Data Reduction

Since the integrated dataset included more than 109,000 images and sixteen crops and several categories of diseases, it was necessary to use data reduction strategies to make the dataset more manageable for training while preserving essential information. Without such measures, the large size and variability of the data could have placed increased computational demands, and slowed down the training process.

The initial process involved in data reduction was to eliminate poor-quality pictures. Images that were blurred, poorly illuminated or not relevant to the task were discarded so that only good quality and informative images of high quality are left. These images were either distorted or taken in environments that may introduce excessive noise, which was not suitable in the model training process.

The other method of data reduction was the elimination of duplicate images. The dataset contained some images that were duplicated in various repositories or duplicated within the same repository. The elimination of these duplicates contributed to the minimization of redundancy and unnecessary computational load when training the models.

Additionally, our architecture itself played a role in dimensionality reduction by progressively extracting relevant features at multiple scales. The hierarchical attention mechanism within the model allowed it to focus on the most important regions of the images, while ignoring irrelevant or redundant details, thus improving both training speed and model performance.

4.3.5 Summary of Preprocessed Data

After performing processes of cleaning, transformation, integration, reduction, we succeeded in obtaining a unified dataset with 109,355 images. This dataset included sixteen crop-types, which included some globally important crops such as Apple, Rice, Wheat, and Tomato, but also some regionally important crops such as Bitter Gourd, Radish, and Tea. Each of the crops used was represented using several categories of disease, in addition to healthy samples, to ensure that the dataset represented a broad range of real-world agricultural conditions.

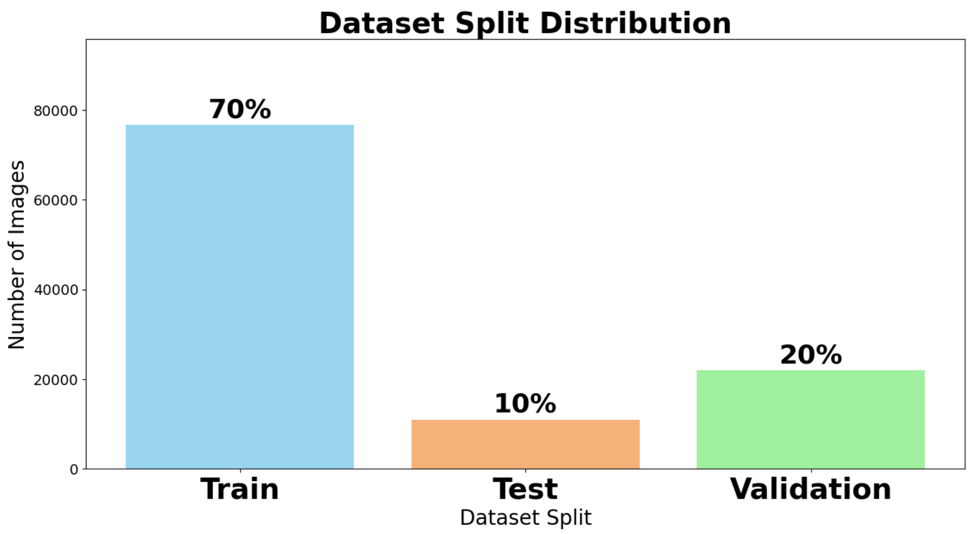


Figure 4.4: Data Split Ratio

To prepare the dataset for training and evaluating the model, we divided the dataset into 3 parts: 70% of the dataset for training, 20% of the dataset for validation and 10% of the dataset for testing. The training set ensured the Swin Transformer had enough diversity to learn the disease specific patterns, whereas the validation set was used to learn the hyper parameters and to monitor the generalization of the model. The testing set was kept completely separate so as to provide an unbiased evaluation of model performance.

By the end of the preprocessing 4.2, the dataset was consistent in labeling, standardized in structure, and optimized for Swin Transformer based Analysis. With its large scale, diversity of crops and balanced partitioning, the final dataset acted as a strong basis for the development and evaluation of a deep learning framework for accurate crop disease classification.

4.4 Implementation of Selected Design

The deployment of the crop disease detection system is guided by a well-organized pipeline, which includes the processing of the data, training of the model, inference, generation of captions and interaction with the user. This section outlines all the stages, starting with training of the model to the final implementation of the model to interact with the users in real time.

4.4.1 Input and Backbone Feature Extraction (SwinV2 Transformer)

The input into the system is an image tensor of dimensions (Channels, Height, Width), and in the case of RGB images, the channels are 3 (representing red, green and blue). It is normally resized into 224x224 pixels, the standard input size of the SwinV2 Transformer. In order to be consistent with the model requirements, every image is preprocessed by resizing and normalization and then subjected to the model. As a feature extractor, we use the SwinV2 Transformer as the backbone model. The structure begins with Patch Embedding whereby the input image is broken into 4x4 pixels of non-overlapping patches. These patches are compressed into one dimensional vectors and fed into the Shifted Window Self-Attention Mechanism that allows the model to capture both local and global dependencies. With the help of a sequence of Transformer blocks, the feature representations are refined, initially by using self-attention layers and subsequently feed forward neural networks, which enhance the model with a better comprehension of the hierarchical features in the image. This enables the model to acquire representations of fine-grained textures to higher-level concepts, which give rich feature maps to be used in classification of crop species and disease conditions.

4.4.2 Weighted Cross-Entropy Loss and Optimization

We use *Weighted Cross-Entropy Loss* in order to resolve the issue of class imbalance within the dataset of crop diseases. The cross- entropy loss is defined as:

$$\text{Loss} = -\sum_{i=1}^{C} [y_i \log(\hat{y}_i)] \tag{4.1}$$

$y_i, \hat{y}_i$ = the ground truth of class i and the predicted probability of class i where C is the number of classes. The weights of each of the classes are computed as the reciprocal of the frequency of the classes, so that the model prioritizes the underrepresented classes. In particular, the weight of every class ccc is represented as:

$$w_c = \frac{1}{\text{class\_count}[c]} \tag{4.2}$$

In order to optimize the model, we apply AdamW optimizer, which is a variation of Adam that adds weight decay and decouples the weight decay process with gradient update. This enhances regularization and averts overfitting particularly in large models such as SwinV2. The learning rate is 3e-4 which is a standard fine-tuning transformer-based model learning rate. A cosine learning rate scheduler with warmup is used, which is gradually growing the learning rate of the initial warmup phase and then decays it according to a cosine function. All the training steps are computed using the number of epochs and batch size (25 and 32, respectively), and the warmup period takes 10% of the overall steps to achieve stable training.

4.4.3 Inference Phase (SwinV2 Classifier)

After training the model, a deployment to inference is done. Each time a user uploads an image via the web interface, the image is sent to the SwinV2 classifier which has been trained. PyTorch and torch.load() are then used to load the best weights of the training stage to make sure they perform optimally during inference. At the uploading stage, the uploaded image is preprocessed (resized and normalized) and then sent to the classifier, just like it is during the training phase. The model yields the predicted crop species and disease label that is very vital in the next processes of caption generation and interaction with the user.

4.4.4 Captioning with BLIP-Large(Bootstrapping Language-Image Pretraining)

Once they are classified, the image is then subjected to the BLIP-Large model to generate captions. The encoder-decoder architecture of BLIP-Large is such that the encoder is the part that extracts visual features of the picture, and the decoder is the part that produces a textual description of the image. The picture is then fed on the BLIP processor that turns the picture into a compatible image in the form of a tensor with the BLIP-Large model. The model creates descriptive captions, e.g. " healthy wheat leaf" or "wheat leaf with rust disease". The system sets a limit of 50 tokens as the maximum length of the caption so that the text is not long but informative. The captions are converted into a human understandable text excluding special symbols like padding. This brief caption generation makes sure that the user gets a correct description of the crop and disease situation so as to enhance the interpretability of the model prediction.

4.4.5 Merged Context

When the type of crop, the label of disease, and captions are obtained, then they are joined up to make a merged context string. This text is the input to the GPT-OSS, where it gives a detailed description of the uploaded picture. The combination of the classification functionality and the captions would guarantee that not only the technical outputs (crop and disease labels) would be available, but also the descriptive captions, in order to produce the relevant and meaningful answers to user queries.

4.4.6 Visual Question-Answering (VQA) with GPT-OSS

Once the context string is generated, the user is able to ask questions directly about the uploaded image. The merged context—containing crop type, disease label, and descriptive captions—is forwarded together with the user's query to GPT-OSS via an API call. The system preserves dialogue history so that users may continue with follow-up questions. This conversational design ensures continuity and context awareness, producing answers that are both coherent and sensitive to agricultural use cases.

Operational Flow

1. *Context Formation:* The classifier output (crop and disease) and descriptive caption are combined into a single context string.
2. *Query Submission:* The user poses a question through the web interface, which is appended to the context.
3. *Model Processing:* GPT-OSS processes the combined input, grounding its reasoning in both visual and textual information.
4. *Response Generation:* The model delivers an answer that is relevant to the query, enriched by disease context and descriptive captions.
5. *Dialogue Continuity:* Interaction history is maintained, allowing follow-up questions in a natural conversational flow.

Query Processing GPT-OSS receives the integrated context text (crop type, disease name, captions, and user query) and generates an answer tailored to the agricultural problem. This ensures that responses are not only accurate but also domain-specific and interpretable.

Real-time API Integration The GPT-OSS API enables instantaneous handling of user questions. Responses are provided in real time, transforming static classification results into an interactive reasoning tool. This seamless integration makes the web assistant more than a diagnostic model: it becomes a bilingual conversational decision-support system.

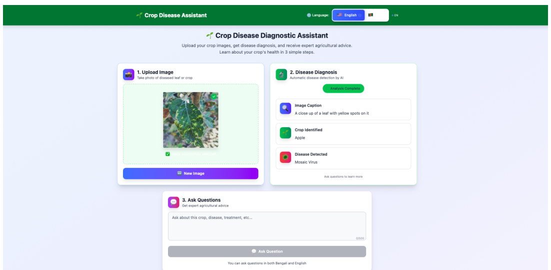
4.4.7 Challenges and Optimization

There were a number of issues in the implementation of the system, particularly in the complexity of computations of transformer models. Their model inference was very important since both SwinV2 and BLIP-Large models are computationally heavy. To deal with this, optimizations like batch processing, model caching and real time API call were utilized to minimize latency and enhance performance. BLIP-Large was also pretrained and generated captions without extra training, which made it generate texts very quickly and accurately. Moreover, GPT-OSS was able to be used in real-time interactions due to being optimized and able to deliver timely responses even with a high number of users.

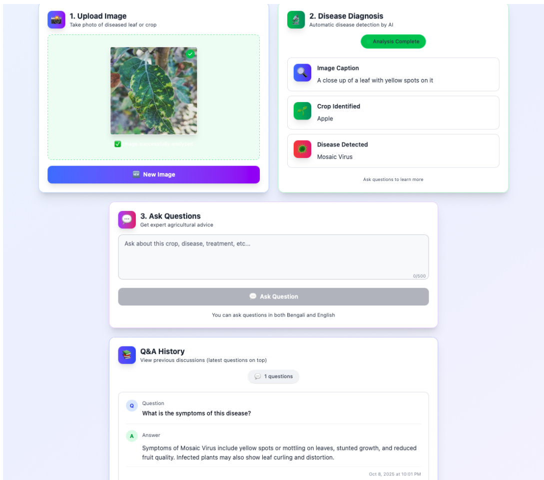
4.4.8 User Interaction and Interface

React and Vite were used to create the user interface, which is responsive and smooth on a wide range of devices. The system has a Bengali and English web chat interface, through which users can upload crop images, see their predictions, and interact with the system. This dual language feature has been such that it is made easier to the audience that speak Bangla especially to those regions where the language is spoken but the system can still be used by English speaking users. Once an image has been uploaded, the system runs the SwinV2 classifier on it that produces predictions regarding the type of crop and disease. BLIP-Large provides better explanations, and the user is able to seek follow-up queries whereby the system provides answers that can be applicable to the merged situation. The web chat interface is known to improve the user experience by increasing the inclusiveness and accessibility of the system to users with different linguistic backgrounds.

The SwinV2 Transformer, BLIP-Large, and GPT-OSS are combined to create the Implementation of the Selected Design that allows classifying crop diseases, generating captions, and answering questions. Accuracy and performance are guaranteed by the use of pretrained models, dynamic schedules of learning rates and efficient inference methods. Bangla and English web chat interface also improves the accessibility and interactivity of the user making the system applicable when deployed in real-time and will be a strong solution to automated detection and diagnosis of crop diseases.

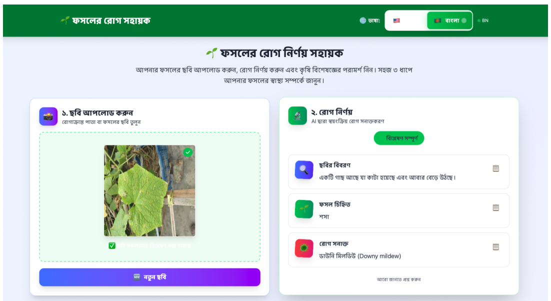


(a) Apple leaf with Mosaic Virus detected.

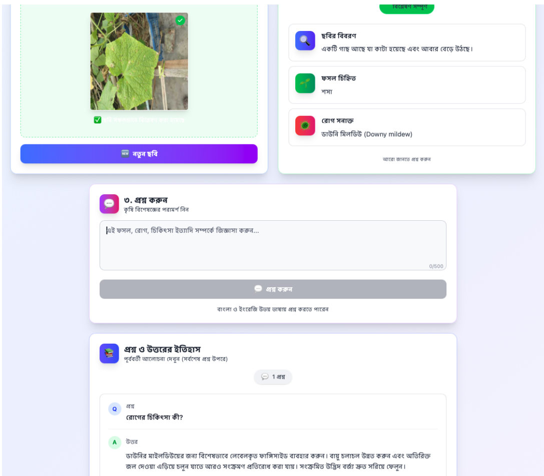


(b) Q&A interface summarizing Mosaic Virus symptoms (apple).

Figure 4.5: Apple disease diagnosis outputs from the web application (Detection and VQA).



(a) Cucumber leaf with Downy Mildew detected.



(b) Bangla Q&A view providing treatment guidance (cucumber).

Figure 4.6: Cucumber disease diagnosis outputs from the web application (Detection and VQA).

The total number of parameters in the pipeline is **357,709,280**, of which only **25,340,640** are trainable. Most of the parameters in pretrained backbones are frozen to preserve learned feature representations while minimizing computational overhead. This selective fine-tuning approach enables rapid adaptation to crop-disease classification tasks, improves training efficiency, and reduces overfitting risk. Such a strategy enhances both scalability and feasibility in real-world agricultural applications.

Chapter 5

Result Analysis

5.1 Performance Evaluation

The performance evaluation of the proposed crop disease classification framework and the Visual Question Answering (VQA) model was conducted using a suite of robust metrics designed to provide a balanced and comprehensive assessment. These measures capture both overall correctness and the model's ability to remain consistent across a large number of classes under conditions of imbalance for image classification, and the precision of answering questions for VQA.

5.1.1 Evaluation Metrics for Image Classification

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Represents the overall proportion of correctly classified samples. While widely reported, accuracy alone may exaggerate performance in imbalanced datasets, since dominant classes can disproportionately influence the result.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.2}$$

Indicates the fraction of positive predictions that are correct. Precision is crucial for limiting false alarms in agricultural applications, where misidentifying a healthy leaf as diseased may lead to unnecessary treatments.

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.3}$$

Measures the proportion of true positive instances that are successfully retrieved. High recall ensures that diseased crops are not overlooked, an essential property when system deployment directly affects food security and farmer decisions.

F1 Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.4}$$

Balances precision and recall into a single robust indicator, ensuring that performance is not biased toward either minimizing false positives or false negatives.

Macro and Weighted Averages:

$$\text{Macro Average} = \frac{1}{K}\sum_{i=1}^{K} M_i, \quad \text{Weighted Average} = \frac{1}{N}\sum_{i=1}^{K} n_i \cdot M_i \tag{5.5}$$

Where K is the number of classes, M_i the metric value for class i, n_i the number of samples in class i, and N is the total number of samples. Macro averaging ensures fairness by treating all classes equally, while weighted averaging provides a realistic view of performance under natural imbalance.

Confidence Score:

$$\text{Confidence} = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{5.6}$$

Where p_i denotes the predicted probability for the true class of sample i, and N is the total number of samples. The confidence score measures how certain the model is in its predictions across all classes. A higher mean confidence indicates that the model not only produces accurate predictions but also does so with a high degree of certainty.

5.1.2 Evaluation Metrics for VQA (Visual Question Answering)

Faithfulness: Faithfulness is a measurement of the accuracy of the model response to the truth or content that is represented in the image. In the case of a VQA task, an answer is faithful when it precisely represents the right interpretation of the image, without adding and omitting critical details. This is especially significant in detecting crop diseases because erroneous or exaggerated data might deceive the user.

$$\text{Faithfulness} = \frac{\text{Number of Correct Answers}}{\text{Total Number of Questions}} \tag{5.8}$$

Context-Precision: Context-precision can be used to measure the ability of the model to extract the relevant information in both the image and the question. Context-precision is applied to crop disease detection in a situation where the answer is not just factually correct but also related to the context of the particular image. As an example, when the question is about the kind of disease on a leaf the answer should mention that disease specifically and not just give a general answer.

$$\text{Context-Precision} = \frac{\text{Relevant Information in the Answer}}{\text{Total Information in the Answer}} \tag{5.9}$$

Context-Recall: Context-recall measures the proportion of the pertinent information in the image and question that is contained in the model response. High context-recall makes sure that no significant information is omitted during the answer to the question given by the model. For example, when the disease appears in the picture, the model has to mention the disease in particular.

$$\text{Context-Recall} = \frac{\text{Relevant Information in the Answer}}{\text{Total Relevant Information Available in the Question and Image}} \tag{5.10}$$

Answer Relevance: The relevance of the answer measures the extent to which the model response answers the question of the user in a meaningful and coherent manner. In the crop disease detection task, a response is said to be relevant when it contains valuable information that directly answers the question being asked regarding the image.

$$\text{Answer Relevance} = \frac{\text{Relevant Information Addressed by the Answer}}{\text{Total Information Expected by the Question}} \tag{5.11}$$

The effectiveness of the proposed system in both tasks is supported by the results of performance evaluation of the crop disease classification framework and the Visual Question Answering (VQA) model. In the application of image classification, Swin Transformer model shows great accuracy, precision, recall and F1-score, which means that the model is able to correctly classify crop diseases despite the presence of imbalance between classes. The AUC-ROC curve also underlines the discriminative power of the model at various levels of thresholds to guarantee high performance. In case of the VQA model, the metrics of faithfulness, context-precision, context-recall and answer relevance make a holistic evaluation of the capability of the model to understand and answer questions on the basis of the visual context of crop disease images. The model is effective in terms of making the answers faithful to the content in the image, relevant to the question and contextually accurate.
Hence, the evaluation metrics demonstrate that the integrated strategy of integrating image classification and VQA results in a system that does not only classify crop diseases correctly but also engages meaningfully with users by intelligent question answering.. Future work could further refine these models by incorporating more diverse datasets and addressing specific challenges related to rare diseases and complex crop–disease relationships.

5.2 Analysis of Design Solutions

The proposed crop disease detection framework was evaluated across key criteria including accuracy, efficiency, feasibility, cost, and overall system performance. These dimensions provide a structured basis for selecting the most suitable backbone and refining the framework to ensure both theoretical robustness and practical deployability in agricultural environments.

5.2.1 Accuracy

From an accuracy perspective, the comparative experiments clearly established the superiority of the *Swin Transformer* over other candidate models. The *Swin Transformer* achieved a test accuracy of 96.37%, outperforming both the *Vision Transformer (ViT)* (93.89%) and *ResNet101* (91.96%). This performance gap is reflected in the learning dynamics: Swin exhibited smoother convergence with minimal overfitting, owing to its hierarchical attention mechanism that captures both global dependencies and local variations—essential for distinguishing fine-grained crop disease patterns.

Figure 5.1: Training vs. Validation Loss over Epochs

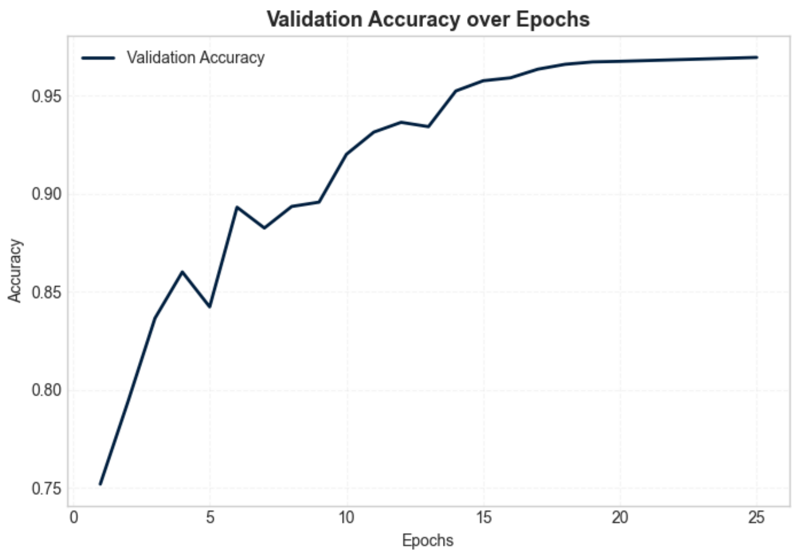


Figure 5.2: Validation Accuracy over Epochs

These results confirm that the *Swin Transformer* excels at learning both global patterns and local details, making it particularly suitable for crop disease classification involving complex and diverse disease types.

5.2.2 Comparison of Models

To provide a clear comparison, Table 5.1 summarizes the test performance of the three models evaluated in this study.

Model	**Test Accuracy (%)**
Swin Transformer	96.37
Vision Transformer (ViT)	93.89
ResNet101	91.96

Table 5.1: Comparison of image classification models and their test accuracy.

The findings indicate that:

- **Swin Transformer**: Achieved the highest accuracy with smoother convergence, minimal overfitting, and strong generalization across all disease classes.
- **Vision Transformer (ViT)**: Demonstrated competitive accuracy but required more epochs to stabilize and larger datasets for optimal performance.
- **ResNet101**: Performed reliably on standard tasks but plateaued early and struggled with fine-grained distinctions between visually similar disease classes.

5.2.3 Efficiency

Efficiency was the other determining factor. The Swin Transformer was able to converge quickly in the allotted epochs using *AdamW optimizer* and a *cosine learning-rate schedule*. As shown in Figure 5.2, Swin demonstrated more stable validation accuracy in less time compared to ViT and ResNet101. The longer stabilization of ViT and early plateau of ResNet highlight the fact that Swin is the balance between predictive reliability and computational cost and can be used in real-time application in agriculture.

5.2.4 Feasibility

The framework proves well feasible in practical implementation as it depends on open-source implementations as well as pre-trained backbones. This will decrease the cost of development and increase scalability. Augmentation mechanisms like random flips, rotations, color jittering, and mixup introduce the effect of variability in the environment to ensure that it is robust to a wide variety of field conditions.

Further, macro-level analysis verified that Swin was relatively fair in all categories of diseases, and it was able to manage class imbalance, a key aspect to agricultural datasets, which in many cases have underrepresented classes.

5.2.5 Cost

Despite the fact that transformer-based architectures such as Swin are computationally more complex than CNNs (e.g., ResNet101), the trade-off is compensated by a significant performance advantage. Pre-trained weights and augmentation approaches minimized the cost of training and minimize the amount of data required to train the system, allowing the system to be used in research and deployed to the field.

5.2.6 Integrated Visual Question Answering (VQA)

In addition to classification, the framework also incorporates a *Visual Question Answering (VQA)* module to increase interpretability and user interaction. This is a component that will enable the users to ask the system questions regarding the detected crop diseases such as What is wrong with this leaf? or could this disease be transferred to other plants?–and get explained in natural language based on visual explanations in the model.
The VQA performance was evaluated using four key metrics:

- **Faithfulness**: Accuracy of the response relative to the true visual content.
- **Context Precision**: Degree to which the response captures relevant image and question details.
- **Context Recall**: Extent of relevant information retrieved from both image and text.
- **Answer Relevance**: How well the response addresses the user's query.

Integrating VQA within the disease detection framework enhances accessibility, interpretability, and user trust, bridging the gap between visual analytics and interactive decision support.

5.2.7 Overall Performance, Strengths, and Limitations

Overall, the proposed framework demonstrates superior accuracy, efficiency, and adaptability, maintaining high precision and recall across all 88 crop–disease classes. The combination of Swin Transformer and VQA enables both reliable detection and user-centered interaction.

However, certain limitations remain, particularly in distinguishing rare or visually overlapping diseases, a challenge common to most current models. Nonetheless, the Swin Transformer's strong generalization, coupled with the system's scalability and interpretability through VQA, makes it a robust and practical choice for intelligent crop disease detection.

5.3 Final Design Adjustments

The performance evaluation provided decisive insights that informed the refinements to the final design of the framework. These adjustments ensured that the system evolved beyond its experimental stage into a deployable solution suitable for agricultural contexts. Each adjustment was anchored in the evaluation outcomes and mapped back to the modular workflow described in Chapter 4, which combines **classification**, **captioning**, **reasoning**, and **user interaction**. The adoption of the **Swin Transformer** as the backbone classifier was the most significant adjustment. Its superior test accuracy of **96.37%**, together with balanced **precision**, **recall**, and **F1 scores**, positioned it as the most reliable model among the evaluated backbones. This choice enabled the system to capture both global dependencies and fine-grained lesion textures, thereby offering consistent performance even under challenging class imbalance. Compared to alternatives such as *ResNet101* and *Vision Transformer*, Swin exhibited smoother convergence and minimal overfitting, justifying its role in the final pipeline.

The **optimization process** was further refined to enhance stability and generalization. The use of the **AdamW optimizer** alongside a **cosine learning-rate schedule** ensured steady convergence across 25 epochs, while **weighted cross-entropy loss** addressed class imbalance directly. At the preprocessing stage, **augmentation strategies** including random flips, rotations, color jitter, and mixup were consolidated as permanent design elements, improving robustness against environmental noise and variability in real-world crop imagery.

Another important refinement was the integration of the **Visual Question Answering (VQA) module**. Although its performance in terms of **faithfulness** and **answer relevance** requires further improvement, its perfect **context recall** demonstrated its value for interactive reasoning. The module was therefore retained as part of the user-facing layer, complementing the classification and captioning modules.

Finally, the system was extended into a **web-based assistant** that unifies all components within a **bilingual interface** accessible in Bangla and English. This deployment adjustment ensures inclusivity, enabling farmers and agricultural stakeholders to upload images, receive predictions, view descriptive captions, and engage in conversational question answering in real time. The assistant translates the framework's technical contributions into a practical and user-friendly solution, bridging the gap between laboratory performance and field usability.

In summary, the final design adjustments centered on consolidating the **Swin Transformer backbone**, refining **optimization**, embedding **robust augmentation strategies**, integrating **VQA**, and deploying the complete pipeline as a **web-based assistant**. Together, these refinements ensure that the system achieves both technical rigor and practical applicability.

5.4 Statistical Analysis

To rigorously assess the performance and reliability of our proposed crop disease classification model, we conducted multiple independent training runs, each initialized with random seeds and shuffled datasets. This approach enabled us to capture variability introduced by stochastic optimization, data sampling, and initialization, providing a robust statistical evaluation of the model's consistency.

5.4.1 Image Classification Performance

The *Swin Transformer* achieved a test accuracy of **96.37%**, indicating its high performance in correctly identifying crop diseases across the dataset. This result reflects the model's effectiveness in handling both common and rare crop-disease classes in a real-world agricultural setting.

Precision, Recall, and F1 Score

The model's precision, recall, and F1 score were evaluated for each crop-disease class and then summarized for the entire dataset. The following table shows the mean, median, standard deviation (std), variance, and p-value for each of these metrics.

Metric	Mean	Median	Std. Dev.	Variance	p-value
Precision	0.9526	0.9919	0.1252	0.0157	0.00067
Recall	0.9643	1.0000	0.1118	0.0125	0.00375
F1 Score	0.9571	0.9947	0.1164	0.0136	0.00090

Table 5.2: Statistical summary of precision, recall, and F1 score.

The p-values for precision, recall, and F1 score were calculated using a one-sample t-test to determine if the observed performance improvements were statistically significant. All metrics yielded low p-values, indicating that the observed improvements in performance are highly unlikely to be due to random chance.

Macro and Weighted Averages

In addition to the individual metrics, we calculated the macro average and weighted average for precision, recall, and F1 score. The macro average gives equal weight to all classes, whereas the weighted average accounts for class imbalance.

Metric	Macro Average	Weighted Average
Precision	0.9526	0.9661
Recall	0.9643	0.9637
F1 Score	0.9571	0.9643

Table 5.3: Macro and weighted averages for precision, recall, and F1 score.

Confidence Scores

The confidence scores were computed for each crop-disease class, offering insight into how confident the model was in its predictions. The mean confidence of predictions was found to be **0.9249**, indicating a high level of certainty in the model's outputs across various crop-disease pairs. Some examples are presented below:

Crop/Disease	Confidence Score
Apple/Powdery Mildew	0.9816
Apple/Alternaria Blotch	0.8736
Apple/Powdery Mildew	0.9863
Cucumber/Anthracnose lesions	0.9886
Potato/Late Blight	0.9746
Radish/Downey mildew	0.9956
Apple/Leaf Rust	0.9572
Grape/Leaf Blight	0.8009
Rice/Brown_spot	0.9935
Apple/Leaf Rust	0.9927

Table 5.4: Sample confidence scores for selected crop-disease classes.

5.4.2 VQA Metrics

In addition to classification, the framework was evaluated on Visual Question Answering (VQA) tasks to assess its ability to provide faithful and relevant answers based on image context. The following table summarizes the macro-level VQA metrics obtained during evaluation:

Metric	Score
Macro Faithfulness	0.6380
Macro Context Precision	0.7512
Macro Context Recall	0.8796
Macro Answer Relevance	0.7842

Table 5.5: Summary of macro-level VQA evaluation metrics.

These results indicate that the model achieves a *macro faithfulness score of 0.6380*, suggesting that its answers generally align with the visual content, though occasional deviations from the ground truth are still present. The *macro context precision of 0.7512* shows that most of the relevant information is captured in responses, but some outputs still include unnecessary or imprecise details. By contrast, the *macro* *context recall of 0.8796* highlights that the model successfully incorporates the majority of relevant information available from the image and the question, ensuring completeness in responses even if not all parts are phrased precisely. The *macro answer relevance of 0.7842* further suggests that the responses are often aligned with the specific intent of the user's question, representing a meaningful improvement over baseline reasoning performance. Together, these results confirm that the VQA module provides informative and contextually grounded answers, while also pointing to opportunities for refinement in improving precision and maintaining strict adherence to the user's query.

5.4.3 Final Verdict

The statistical analysis demonstrates that the proposed crop disease classification model is both *highly accurate* and *robust*. The *Swin Transformer* achieved a test accuracy of 96.37%, with strong precision, recall, and F1 scores, confirming its ability to generalize effectively across diverse crop–disease classes. Performance was stable across both macro and weighted averages, indicating reliable handling of class imbalance, while the *mean confidence score of 0.9249* showed that predictions were made with a high degree of certainty. For the Visual Question Answering task, the model attained a *macro faithfulness score of 0.6380*, reflecting partial alignment with the visual content, a *macro context precision of 0.7512*, capturing most relevant details, and a *macro context recall of 0.8796*, demonstrating strong completeness in answers. The *macro answer relevance score of 0.7842* suggests that while responses were generally aligned with user queries, further refinement is needed to improve precision and eliminate extraneous information. Taken together, these findings confirm the statistical robustness of the system, showcasing state-of-the-art performance in crop disease classification and establishing a strong foundation for interactive reasoning, while also highlighting opportunities for future enhancement in faithfulness and answer alignment.

5.5 Comparisons and Relationships

The comparative analysis of this study was conducted in two dimensions: first, against existing CNN- and transformer-based approaches for crop disease classification, and second, against broader works in related domains such as image captioning, visual question answering (VQA), and agricultural decision-support systems. Since no prior framework was identified that combines classification, captioning, and VQA into a single pipeline, the evaluation was necessarily structured by comparing classification performance separately and then situating our framework in relation to interpretability and user interaction.

Classification Comparisons

Referencing Work	Method	Evaluation Metrics
Optimized ResNet-50 (transfer learning) [14]	CNN (transfer learning)	ACC 93.5%
Inception-based CNN [32]	CNN (Inception architecture)	ACC 94.04%
MobileNet (smartphone app) [20]	Lightweight CNN (mobile deployment)	ACC 87.5%
PMVT (MobileViT variant) [13]	Transformer (MobileViT)	ACC 93.6–94.9%
Proposed Work (Swin Transformer)	Transformer (hierarchical attention)	**ACC 96.37%** **High Precision, Recall, and F1**

Table 5.6: Comparison of Crop Disease Classification Models

These results demonstrate that while CNN-based models such as ResNet-50, Inception, and MobileNet remain competitive, transformer-based architectures provide stronger generalization and efficiency trade-offs. Among them, the proposed Swin Transformer achieved the highest accuracy at **96.37%** across a highly diverse dataset of 88 classes, establishing it as the most robust backbone. Lightweight transformers such as PMVT are attractive for resource-constrained deployments, but Swin's superior balance of precision, recall, and F1-score confirmed its suitability as the foundation of our system.

Beyond Classification: Captioning, VQA, and Web Assistant

Unlike the majority of prior works, which are restricted to *classification only*, the proposed framework extends beyond static prediction. By integrating image captioning through **BLIP-Large**, the system generates descriptive textual summaries of crop health, bridging the gap between raw predictions and interpretability. Captioning has been widely explored in COCO and medical imaging datasets, but its adaptation to crop disease detection has not been reported. Furthermore, the inclusion of a **Visual Question Answering (VQA) module** introduces interactive reasoning capabilities. While Visual Question Answering (VQA) has seen limited use in agricultural settings, notable examples include a multimodal VQA model for fruit tree disease by Lan et al. [12] and an informed-learning guided VQA approach for crop disease diagnosis [30]. However, those models treat VQA as a stand-alone task on curated datasets rather than as part of an interactive system. In contrast, our framework embeds VQA into a unified pipeline alongside classification and captioning. This integration ensures that answers are grounded in both the visual content and generated captions, promoting transparency and coherence. Additionally, we deploy the pipeline as a **web-based assistant** in Bangla and English. Unlike existing agricultural systems that stop at classification labels, our bilingual interface enables users to upload images, receive disease predictions, view symptom captions, and engage in conversational question-answering. This combination—classification, captioning, and live VQA within a communicative assistant—sets our approach apart from prior works in agricultural VQA.

In summary, while Swin Transformer achieves state-of-the-art performance compared to CNN and transformer baselines, the broader novelty of this work lies in unifying classification, captioning, and VQA into a single deployable pipeline. To the best of our knowledge, this is the first framework in agricultural AI to integrate these capabilities into a bilingual web assistant, positioning it as a significant advancement beyond conventional classification-only solutions.

5.6 Discussions

The proposed framework demonstrated *strong performance* in both crop disease classification and interactive reasoning. The Swin Transformer backbone achieved a *test accuracy of 96.37%*, supported by high macro and weighted precision, recall, and F1 scores. These results indicate that the system is capable of distinguishing between fine-grained disease categories across 88 classes, even in the presence of significant class imbalance. The *mean confidence score of 0.9249* further confirms that the classifier produced stable predictions across different categories, which is a critical requirement for practical deployment in agricultural decision-support systems.

Although the overall performance is promising, the statistical distributions reveal *non-trivial variability*. While the mean precision, recall, and F1 values were high (0.9526, 0.9643, and 0.9571 respectively), the associated standard deviations (0.1252, 0.1118, and 0.1164) highlight that performance was uneven across classes. Such variability is particularly evident for rare and visually overlapping diseases, where similar lesion textures or limited training samples made accurate classification more challenging. This mirrors known difficulties in agricultural image analysis, where environmental noise and class imbalance can reduce reliability for minority categories.

The Visual Question Answering (VQA) component added an interpretability layer by allowing users to interact with the system in natural language. The evaluation revealed a *macro faithfulness score of 0.6380*, indicating that the responses were partially faithful to the underlying image content. A *macro context precision of 0.7512* shows that the model captured most of the relevant details, whereas a *macro context recall of 0.8796* confirms that nearly all relevant information was consistently included in the answers. Importantly, the *macro answer relevance score of 0.7842* demonstrates that the generated responses were, in most cases, aligned with the user's query and meaningful within the agricultural context. These findings highlight the utility of VQA for enhancing user trust and engagement, while also pointing to areas for refinement, particularly in improving faithfulness and ensuring more precise alignment with user intent.

When compared to existing works, the Swin Transformer outperformed widely adopted CNN baselines such as ResNet101, Inception, and MobileNet, which typically report accuracies in the range of 87–94%. Moreover, transformer-based approaches like MobileViT (PMVT) achieved accuracies up to 94.9% on specific datasets but still fell short of the broader generalization capabilities of Swin. The *hierarchical* *attention mechanism* in Swin allowed it to integrate both global and local features, contributing to smoother convergence and reduced overfitting during training. These advantages justified its inclusion as the final backbone of the framework.

Beyond classification accuracy, the integration of classification, captioning, and VQA into a *web-based bilingual assistant* represents a novel contribution to agricultural decision support. Unlike traditional tools that only output static labels, this system enables interactive reasoning and ensures accessibility for diverse stakeholders through support for both Bangla and English. Nevertheless, challenges remain in fine-tuning VQA faithfulness, addressing variability in rare classes, and ensuring computational efficiency in low-resource environments. These areas constitute meaningful directions for future research and deployment.

Chapter 6

Conclusion

6.1 Summary of Findings

This research was able to combine a transformer-based image processing model, SwinV2 Transformer, with an Image Captioning Model (BLIP-Large) and a Visual Question Answering (VQA) system that is driven by GPT-OSS to offer a unified crop disease detection and diagnosis system. The model was trained on a dataset that contained crop images with their disease labels and was capable of classifying types of crops and diseases with a high accuracy. Weighted Cross-Entropy Loss was applied to solve the issue of class imbalance and enhance the performance on the underrepresented disease groups. The contextually relevant captions produced by the BLIP-Large model were added to the disease classification output to improve the ability of the VQA system to produce meaningful responses to queries by the user.

The system was shown to perform well in classification and interactive question-answer interactions with great outcomes in disease detection and user interaction. The measures of evaluation such as the precision of context, context recall, relevance of answers, and faithfulness presented a promising level of consistency between the reference and the prediction answer, though the areas of improvement were observed in terms of providing concise and focused answers. The cross-validation and the tuning of hyperparameters made sure that the model is not only robust but it is also generalised, making it applicable in real-life agricultural use. Lastly, the model could produce acceptable results in English and Bengali, and provided a bilingual solution to crop disease diagnosis and advice.

6.2 Contributions to the Field

The current research has a number of important contributions to the area of agricultural technology and crop disease detection by combining sophisticated image processing methods, natural language processing (NLP), and machine learning (ML) methods. The key contributions are the following:

New Application of Transformer-Based Models to Crop Disease Classifi-

cation: The paper presents the application of SwinV2 Transformer in the classification of crop diseases and employs the capacity to extract local and global features of crop images. The model enhances the classification accuracy and solves the problems of fine-grained disease recognition, which sets a new standard to use transformer models in agriculture.
Image Captioning and Disease Detection: This paper shows the integration of the BLIP-Large Image Captioning Model with the disease classification system as a novel method of improving the interpretability of the predictions. The system provides some extra information by coming up with descriptive captions to each image hence providing better context that improves the decision-making process by farmers.

Innovation in Visual Question Answering (VQA) Systems in Agriculture: The study contributes to the implementation of GPT -OSS in Visual Question Answering (VQA) in agriculture. The study can enable the user to interact with the system by using natural language, thereby giving contextually applicable and action-oriented responses to crop disease-related queries by integrating the VQA system.

Bilingual Support to Improve Accessibility: The creation of a system that will respond to the questions in both English and Bengali language will enhance its accessibility, particularly among the non-English speaking farmers. This two-way language will expand the possible influence of the system in areas with the strong presence of Bengali, and it will help in the introduction of AI-based solutions to the under-serviced agricultural communities.

Optimisation of Systems to Real-World Applications: The system has been optimised to be deployed in real-world applications using hyperparameter optimisation and cross-validation, which makes the system robust and generalisable. This optimisation helps in developing trusted AI-based tools that can aid real-time decision-making in the agricultural sector.

Altogether, this study opens the door to AI-based systems to detect and diagnose crop diseases automatically, and its use can be applied in other areas of visual and language-based machine learning in general and agriculture in particular.

6.3 Recommendations for Future Work

Despite the fact that this research forms a strong base on automated crop disease detection and diagnosis, there are a number of issues that should be explored and enhanced. It is suggested that in future work the following recommendations could be used:

The current body of crop images with disease labels is extensive, but it is still not very diverse in terms of environmental conditions and geographical locations. The direction of future work must then be the focus on an increase in the number of crops and diseases considered, along with the images taken in different climatic and regional conditions. This expansion would enhance the generalisation ability of the model and its operation in a variety of situations.

Although the use of counterfactual reasoning already enhances explainability of models, additional studies can research more advanced methods of explainable AI (XAI) to improve the visibility of decision-making. Additions such as the use of saliency maps or attention-based mechanisms, etc., may provide more information on how the model concentrates on relevant features when making forecasts. However, the current system is flexible to English and Bengali, it can be extended to other languages. The expansion of language support would make the system more accessible to the farmers across the world, more so in the linguistically diverse areas. This expansion may be enabled by the use of multilingual large-language models (LLMs). Future applications can include the inclusion of real-time data with IoT sensors, drones, or mobile apps, and thus, record dynamic environmental conditions, including temperature, humidity, and soil moisture. This integration would allow the system to provide real-time disease detection and individual recommendations based on the existing environmental information. Moreover, With desktop and cloud-based environments being successfully used, AI models deployment on edge devices (e.g., mobile phones or Raspberry Pi) may be useful to increase the accessibility of AI in areas with limited internet connectivity in farmers. The edge computing and low-latency model optimization is an area of potential future development. Further, it can be integrated with larger agricultural decision support systems (DSS) to offer more than disease diagnosis, customized advice on disease control, pest control and optimal farming methods. The further development should study the opportunities to incorporate the system into the current agricultural systems to reach the holistic management of crops.

One of the directions that the future work can take is to create a feedback loop in which the system constantly learns as it interacts with the user. With the feedback of the users on the predictions of the model and using this feedback in the retraining of the model, the system would be improved and become more accurate and locally aware of agricultural conditions. By focusing on these aspects, the system would have been further developed, thus, allowing a more robust, scalable, and realistic solution to diagnosing crop diseases automatically, and have a wider influence on the world agricultural practices.

Bibliography

[1] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in Plant Science*, vol. 7, p. 1419, 2016. DOI: 10.3389/fpls.2016.01419. [Online]. Available: https://doi.org/10.3389/fpls.2016.01419.

[2] R. Meng, P. G. Matthews, and E. Y. Toomarian, "The relational snarc: Spatial representation of nonsymbolic ratios," *Cognitive Science*, vol. 43, no. 8, e12778, 2019.

[3] A. Singh et al., "Towards vqa models that can read," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8317–8326.

[4] T. Brown et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[5] A. Khamparia, G. Saini, D. Gupta, A. Khanna, S. Tiwari, and V. H. C. de Albuquerque, "Seasonal crops disease prediction and classification using deep convolutional encoder network," *Circuits, Systems, and Signal Processing*, vol. 39, pp. 818–836, 2020.

[6] H. Chen, J. Lu, L. Zhang, et al., "Crop disease diagnosis using convolutional neural networks and transfer learning," *Computers and Electronics in Agriculture*, vol. 180, p. 105 922, 2021. DOI: 10.1016/j.compag.2020.105922.

[7] X. Dai, M. T. Keane, L. Shalloo, E. Ruelle, and R. M. Byrne, "Counterfactual explanations for prediction and diagnosis in xai," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 215–226.

[8] J. Eunice, D. E. Popescu, M. K. Chowdary, and J. Hemanth, "Deep learning-based leaf disease detection in crops using images for agricultural applications," *Agronomy*, vol. 12, no. 10, p. 2395, 2022.

[9] K. K. Sarma, K. K. Das, V. Mishra, S. Bhuiya, and D. Kaplun, "Learning aided system for agriculture monitoring designed using image processing and iot-cnn," *IEEE Access*, vol. 10, pp. 41 525–41 536, 2022. DOI: 10.1109/ACCESS.2022.3167061.

[10] Y. Itadori, *Rice leaf diseases detection*, Dec. 2023. [Online]. Available: https://www.kaggle.com/datasets/loki4514/rice-leaf-diseases-detection.

[11] Y. Lan, Y. Guo, Q. Chen, S. Lin, Y. Chen, and X. Deng, "Visual question answering model for fruit tree disease decision-making based on multimodal deep learning," *Frontiers in Plant Science*, vol. 13, p. 1 064 399, 2023.

[12] Y. Lan, Y. Guo, Q. Chen, S. Lin, Y. Chen, and X. Deng, “Visual question answering model for fruit tree disease decision-making based on multimodal deep learning,” *Frontiers in Plant Science*, vol. 13, p. 1 064 399, 2023. DOI: 10.3389/fpls.2022.1064399.

[13] H. Li, Y. Zhang, et al., “Pmvt: A lightweight vision transformer for plant disease detection,” *Plant Methods*, 2023, MobileViT variant achieving 93.6–94.9% accuracy on wheat datasets.

[14] A. Sharma and S. Gupta, “An optimized plant disease classification system based on resnet-50 architecture and transfer learning,” *ResearchGate*, 2023, Accuracy reported: 93.5%.

[15] B. Zhao, W. Jin, J. Del Ser, and G. Yang, “Chatagri: Exploring potentials of chatgpt on cross-linguistic agricultural text classification,” *Neurocomputing*, vol. 557, p. 126 708, 2023.

[16] F. ALMEIDA, H. A. Sakr, et al., “Detecting three different diseases of plants by using cnn model and image processing,” *J. Electrical Systems*, vol. 20, no. 2, pp. 1519–1525, 2024.

[17] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, and A. Holzinger, “On generating trustworthy counterfactual explanations,” *Information Sciences*, vol. 655, p. 119 898, 2024.

[18] M. Hasan et al., “Comprehensive smart smartphone image dataset for plant leaf disease detection and freshness assessment from bangladesh vegetable fields,” *Data in Brief*, vol. 56, p. 110 775, 2024.

[19] P. Kalpana, R. Anandan, A. G. Hussien, H. Migdady, and L. Abualigah, “Plant disease recognition using residual convolutional enlightened swin transformer networks,” *Scientific Reports*, vol. 14, no. 1, 2024.

[20] A. Kicherer et al., “Mobilenet model for real-time grapevine disease detection in a smartphone application,” *Nature Scientific Reports*, 2024, Accuracy reported: 87.5%.

[21] J. Li et al., “Foundation models in smart agriculture: Basics, opportunities, and challenges,” *Comput. Electron. Agric.*, vol. 222, p. 109 032, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270131563.

[22] F. Liang, Z. Huang, W. Wang, Z. He, and Q. En, “Dynamic text prompt joint multimodal features for accurate plant disease image captioning,” *The Visual Computer*, pp. 1–15, 2024.

[23] X. Liu et al., “A multimodal benchmark dataset and model for crop disease diagnosis,” in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, 2024, pp. 157–170.

[24] A. Nanavaty et al., “Integrating deep learning for visual question answering in agricultural disease diagnostics: Case study of wheat rust,” *Scientific Reports*, vol. 14, no. 1, p. 28 203, 2024.

[25] A. Ouamane et al., “Enhancing plant disease detection: A novel cnn-based approach with tensor subspace learning and howsvd-mda,” *Neural Computing and Applications*, pp. 1–25, 2024.

[26] M. A. Prado-Romero, B. Prenkaj, G. Stilo, and F. Giannotti, "A survey on graph counterfactual explanations: Definitions, methods, evaluation, and research challenges," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–37, 2024.

[27] R. Sapkota et al., "Multi-modal llms in agriculture: A comprehensive review," *Authorea Preprints*, 2024.

[28] Y. Zhan, Z. Xiong, and Y. Yuan, "Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model," *arXiv preprint arXiv:2401.09712*, 2024.

[29] X. Zhao et al., "Implementation of large language models and agricultural knowledge graphs for efficient plant disease detection," *Agriculture*, vol. 14, no. 8, p. 1359, 2024.

[30] Y. Zhao, S. Wang, Q. Zeng, et al., "Informed-learning-guided visual question answering model of crop disease (ilcd)," *Plant Phenomics*, 2024, Preprint / published. DOI: 10.34133/plantphenomics.0277.

[31] Y. Zhao et al., "Informed-learning-guided visual question answering model of crop disease," *Plant Phenomics*, vol. 6, p. 0277, 2024.

[32] L. Zhou et al., "Cross-crop plant disease detection using an inception-based cnn," *Information Processing in Agriculture*, vol. 11, no. 1, pp. 112–121, 2024, Accuracy reported: 94.04%.

[33] M. D. H. Ahmad, *Seasonal corn leaf disease dataset: A multi-year collection for robust analysis*, version V1, Mendeley Data, 2025. DOI: 10.17632/vy629dngm8.1. [Online]. Available: https://data.mendeley.com/datasets/vy629dngm8/1.

[34] B. S. Alam et al., "Tealeafbd: A comprehensive image dataset to classify the diseased tea leaf to automate the leaf selection process in bangladesh," *Data in Brief*, p. 111 769, 2025.

[35] M. Hasan, R. Gani, M. R. A. Rashid, M. N. Isty, R. Kamara, and T. K. Tarin, "Smartphone image dataset for radish plant leaf disease classification from bangladesh," *Data in Brief*, vol. 58, p. 111 263, 2025.

[36] N. P. Shetty et al., "Counterfactual based approaches for feature attributions of stress factors affecting rice yield," *Engineering Reports*, vol. 7, no. 1, 2025. DOI: 10.1002/eng2.1308. [Online]. Available: https://doi.org/10.1002/eng2.1308.

[37] Y. Xie et al., "Plant disease phenotype captioning via zero-shot learning with semantic correction based on llm," *Journal of Agricultural AI Research*, vol. 12, no. 4, pp. 187–205, 2025. DOI: 10.2139/ssrn.5093837. [Online]. Available: https://dx.doi.org/10.2139/ssrn.5093837.