IMAGE BASED PLANT DISEASE IDENTIFICATION

Ryan Philip
Student# 1007731689
ryan.philip@mail.utoronto.ca

Veer Kunal Kapadia
Student# 1008817773
veer.kapadia@mail.utoronto.ca

Muhammad Abdullah Choudhary
Student# 1008767791
muhammad.choudhary@mail.utoronto.ca

Md Iftikher Zaman Chowdhury
Student# 1008890808
iftikherzaman.chowdhury@mail.utoronto.ca

ABSTRACT

Plant diseases have had devastating effects on agricultural and food security. In recent years, deep learning techniques, specifically Convolutional Neural Networks (CNNs), have shown extremely promising results in automated plant disease classification. This project aims to develop a CNN-based plant disease classification model to help farmers in detecting and identifying diseases in their crops. Through the pre-compiled Plant Village dataset of 38 different plant species, including a varying range of diseases affecting them, we aim to train a Convolutional Neural Network (CNN) to complete this botanist task. Each leaf image holds critical information, such as the plant species, the presence of diseases, and even the specific disease type. By effectively augmenting, arranging and processing this data, we aim to force the model to recognize key underlying features rather than simply memorizing the data, and thus being capable of accurate multi-class classification and disease diagnosis. Our model makes use of pre-existing techniques such as Layer Normalization, Dropout, ADAM, and the ReLU activation function in order to maximize accuracy and efficiency on test data. It also utilizes skip connections, similar to the ones in ResNet, in order to avoid vanishing and exploding gradients. To compare the performance of the developed model, a Support Vector Machines (SVM) model will be used as a baseline. SVMs are simple models suitable for multi-class classification tasks, but they lack the ability to capture complex patterns and details that neural networks do. Ethical considerations regarding data collection and usage have been taken into account too. The project utilizes publicly available data to avoid ethical issues related to data acquisition and copyright violation. However, limitations in training data availability and potential biases due to data imbalance are acknowledged and considered in our model. Finally, the project plan outlines the tasks assigned to each of the four team members and emphasizes effective communication. Meetings are conducted regularly on Tuesdays and Saturdays to discuss progress and ensure the smooth execution of the project. Risk management strategies such as clear communication, time management, and accountability are employed to limit challenges from arising in the future.

—-Total Pages: 7

1 INTRODUCTION

Plant disease is a threat to the food supply, costing around USD 200 billion annually [1] .With a rapidly growing population, the need for minimizing crop loss is abundant. By being able to identify diseases accurately, appropriate disease control measures are able to be taken to prevent these losses. In recent years, advancements in machine learning show potential in revolutionizing plant disease identification. This project aims to utilize deep learning to develop a robust system that will identify a plethora of plant diseases.

The goal of this project is to create a model capable enough of receiving an image of a plant's leaf and identifying whether the plant is healthy or diseased, while also specifically classifying the disease and species of the plant. By using deep learning algorithms, our model can be trained to recognize subtle visual patterns between diseased plant leaves which would aid in its detection of the diseases early and accurately. This would help agricultural producers to start treatment early, improving productivity and potentially stopping diseases from spreading.

Due to the recent advancements seen in deep learning and its boom in 2012, it is extremely clear that the best method for image classification is deep learning due to its ability to learn hierarchical information from complex data and identifying discrete patterns to separate data into classes. Where manual inspection is labor intensive, time consuming, costly, and prone to human error, machine learning rules out all these problems through enabling us to classify diseased and healthy plants by simply taking a picture. Additionally, this model can act as an embedding, and by being connected with other models, it can only increase the efficiency of preventing crop loss via transferred learning.

2 Background And Related Work

To explore current solutions to plant disease identification, a number of sources were explored. Below are five most important ones considered.

2.1 Using Deep Learning for Image Based Plant Disease Detection:

A deep learning convolutional neural network was trained on a dataset of 54,306 images capturing both healthy and diseased plant leaves under controlled conditions. With an impressive accuracy of 99.35%, the model successfully identified 14 crop species and 26 diseases. This achievement showcases the potential of using large, publicly available image datasets to enable smartphone-assisted crop disease diagnosis on a global scale [2].

2.2 A deep learning based approach for automated plant disease classification using vision transformer:

This idea proposes a lightweight deep learning approach using Vision Transformer (ViT) for real-time automated plant disease classification. It compares the performance of ViT, convolutional neural network (CNN) methods, and a combination of CNN and ViT on multiple datasets. The study discusses the challenges of using pre-designed architectures and highlights the importance of accuracy and prediction speed in real-time applications. Various datasets related to plant diseases are mentioned, including wheat rust, rice leaf diseases, and the Plant Village dataset. The paper also provides an overview of CNNs, ViT, and the implemented network structures [3].

2.3 Plant Diseases Detection System Using Deep Learning:

In their project, the researchers developed a system to assist farmers in detecting plant diseases by uploading leaf images. They utilized a Convolutional Neural Network (CNN) trained on approximately 4,500 leaf images from the Plant Village dataset, which had consistent backgrounds and lacked extra noise. To address variations in real-world scenarios, they integrated the YOLO algorithm to identify regions of interest (ROI) and pass them to the CNN model for prediction. The proposed CNN model consisted of 10 layers, including sequential layers for data preprocessing, convolutional and pooling layers, a fully connected layer, and an output layer with Softmax activation. The dataset was divided into 60% for training, 20% for validation, and 20% for testing. After training for 60 epochs, the model achieved a validation accuracy of 96% and a test accuracy of 93% [4].

2.4 Deep Learning-Based Leaf Disease Detection in Crops Using Images for Agricultural Applications :

This paper focuses on the importance of early plant disease detection in the agricultural sector and the potential benefits of deep learning techniques. The researchers utilized pre-trained convolutional neural network (CNN) models, specifically DenseNet-121, ResNet-50, VGG-16, and Incep-

tion V4, to identify plant diseases effectively. The experiments were conducted using the PlantVillage dataset, consisting of 54,305 image samples from 38 disease classes. The performance of the models was evaluated based on classification accuracy, sensitivity, specificity, and F1 score. The results demonstrated that DenseNet-121 achieved a significantly higher classification accuracy of 99.81%, surpassing other state-of-the-art models. This study highlights the potential of CNN-based models for accurate plant disease identification and encourages further research in this area [5].

2.5 Attention-Based Recurrent Neural Network for Plant Disease Classification:

Detecting and categorizing plant diseases are vital for ensuring global food security and agricultural stability. Recent research has shown promising outcomes in using Convolutional Neural Networks (CNN) to classify diseases based on RGB images. However, CNN models have limitations, including their inability to exclusively focus on affected areas and their tendency to consider irrelevant backgrounds or healthy parts of plants. To address these challenges, this study introduces a novel approach that utilizes Recurrent Neural Networks (RNN) to automatically identify infected regions and extract relevant features for disease classification. Experimental findings demonstrate that the RNN-based approach is more robust, exhibits superior generalization to unseen infected crops, and accurately identifies infectious diseases in plants. This method holds significant potential for enhancing the detection and classification of crop pathogens across diverse plant species [6].

3 Data Processing

The datasets for our project were collected from two primary sources: the PlantVillage Dataset (approximately 17,000 training images and 4,000 validation images) [7] and the New Plant Diseases Dataset (approximately 70,000 training images and 19,000 validation images) [8]. These datasets comprise images of leaves belonging to various plant species, captured during both healthy and diseased phases. Each disease is categorized into distinct classes (38), enabling identification of the specific ailment affecting the plant. Due to their only being 33 images for testing use, we have compiled about 200 images of our own to use testing data in order to measure the model's accuracy. To ensure unbiased machine interpretation, we combined all the images (training) into a single dataset (removing identical images through a python code) and subsequently randomized their order. To accommodate our GPU, the dataset was divided into batches of 128. Furthermore, we performed data augmentation techniques such as cropping, shifting, color variations, and image rotations within the dataset. These augmentations aim to encourage the machine to identify general image features rather than relying on memorizing specific image patterns.

4 Architecture

The project's goal is to identify whether a plant has a disease and the disease type. To accomplish this, a combination of a Convolutional Neural Networks (CNN) and a fully connected network is required. CNN's are good for analyzing images, which is the type of input data we will be using [2]. Fully connected networks work well with classification problems. A rough description of the architecture is given below [9]

- The input layer accepts images of the diseased plants leaves. The image size will be set to a fixed resolution during pre processing of the data.
- The convolution layers are responsible for learning patterns in the images. There will be multiple layers with kernels that learn through each iteration that extract the important features from the input such as edges, textures, and shapes, which are important for disease identification.
- The activation function used will be the ReLU (Rectified Linear Unit) which is the industry standard. This function will be applied after the convolution operations.
- A step size of 2 will be used instead of pooling layers in order to aggregate the information and reduce the dimensions of the data being handled, which will reduce the computational load and make the network more efficient.

- Fully Connected Network: Once the features are extracted, to classify them into the different classes (the plant type, and whether it has a disease or not) a fully connected network is implemented. The data from the CNN is then flattened into a one dimensional vector. This vector is passed through multiple layers of neurons which keep decreasing the size of the vector. These layers are responsible for combining the information learned from the features and making decisions based on the extracted information.

- The output layer will have the number of neurons equal to the number of plant species with their diseases plus the number of classes for their healthy counterpart. Using a sigmoid activation function, a probability distribution will give the likelihood of the image falling within each class, and the highest probability class will be chosen.

5 ILLUSTRATION/FIGURE

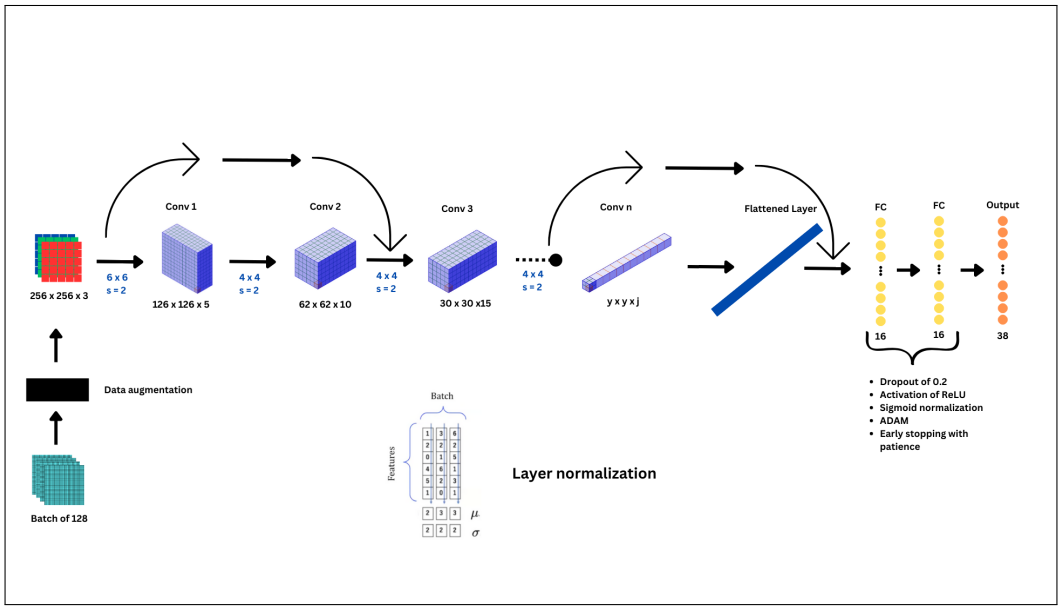

Figure 1: Project Model

6 BASELINE MODEL

The Support Vector Machines (SVM) will be used as a baseline model to compare our neural network against. The SVM was chosen over other models due to its ability of multi-class classification, as compared to other models, such as the Logistic regression, which are only capable of binary classification.

SVM's are extremely simple models, incapable of noticing patterns or details that neural networks do. As per the case of our project, they can be used for multi-class classification by one against all or one against one approaches. They separate classes by simply finding an optimal hyperplane which separates classes. This is done through kernel functions which map the data to higher dimensional spaces in order to handle non linear classification problems. They also use a term known as "C-parameters" to minimize classification errors. A greater C-value means a thinner margin, leading the model to classify the images more accurately, but potentially increasing the chances of over fitting. On the contrary, a smaller C-value results in a wider margin, which could potentially lead to many mis-classifications. [10]

7 PROJECT PLAN

The team will work on their assigned section using GitHub. Each member's task and section will be clearly defined, which will prevent confusion and complications when writing code. GitHub offers version control capabilities, so any errors can be easily undone.Additionally, The team will have two weekly meetings on Tuesday, 12:00 pm EST and Saturday, 12:00 pm EST via Zoom. Additional meetings for tasks can be held when necessary. Primary messaging platform will be using Instagram direct messaging.

Table 1 : Assigned Tasks and Deadlines

Assignment	Tasks	Assigned Person	Internal Deadline	Due Date
Project Selection	Discuss and choose topic for the project	Veer, Ryan, Muhammad, Iftikher	14/6/2023	16/6/2023
Project Proposal	Illustration + Baseline Model	Muhammad	14/6/2023	16/6/2023
	Background + Data Processing	Iftiker	14/6/2023	16/6/2023
	Architecture + Project Plan	Ryan	14/6/2023	16/6/2023
	Ethical Considerations + Risk Register	Veer	14/6/2023	16/6/2023
	Introduction	Veer, Ryan, Muhammad, Iftikher	15/6/2023	16/6/2023
Project Progress Report	Data Collection	Veer, Ryan, Muhammad, Iftikher	24/6/2023	14/7/2023
	Pre-Processing of data	Veer	28/6/2023	14/7/2023
	Baseline model implementation	Muhammad	1/7/2023	14/7/2023
	Primary model implementation	Ryan	6/7/2023	14/7/2023
	Initial model training	Iftikher	12/7/2023	14/7/2023
	Hyperparameter tuning	Veer, Ryan, Muhammad, Iftikher	12/7/2023	14/7/2023
	Project Description	Veer, Ryan, Muhammad, Iftikher	13/7/2023	14/7/2023

8 ETHICAL CONSIDERATIONS

- In terms of the collection of the data by the team, as mentioned above, all the training, testing and validation data utilized in this project have been obtained as publicly accessible data from similar projects that can be viewed on Kaggle [7][8]. Hence, the team does not face ethical issues on that front.

- Limitations to the model and training data: As mentioned under the Risk Register section, due to our large training dataset, we might be vulnerable to delays in training the model. In this case, we might have to focus on only a few classes and omit the rest. This would result in a decline in the accuracy and capabilities of our model. Another limitation is related to the data itself. For example, the team has found raw data (images) of healthy blueberry plants. However, the data does not include images of unhealthy or diseased blueberry plants, which would make it difficult to classify a sample blueberry plant as healthy or unhealthy during the final testing.
- As mentioned above, if the team does not utilize every training dataset that we have due to time constraints, our model's accuracy would decrease. Moreover, our data would be skewed and imbalanced, which would be a bias and thus an ethical issue on our behalf. The team obviously hopes to avoid this situation as much as possible and hence would aim to utilize maximally diversified data.
- An accurate and well-trained plant disease detection model can have a deep impact on agriculture and crop management. It can help detect diseases at an early stage and assists agricultural producers in implementing targeted disease control measures. This can curb the spread of diseases to other healthy crops. All this leads to increased food security and reduced environmental impact.

9 Risk Register

9.1 What if there is a lack of or decline in Communication and Coordination?

A major risk or challenge involved with the team's project is the lack of communication and collaboration. Since the team will be doing this entire project remotely, relying heavily on digital communication tools and without any physical interactions, we are at a high risk of introducing miscommunications, misunderstandings and creating fewer opportunities for spontaneous interactions. This can impact teamwork and coordination. Moreover, the team consists of 2 members who do not have any previous experience in machine learning outside of this course and hence might want to rely on their more experienced teammates when dealing with the nitty-gritties of the technical aspects of our project.

A tried and tested solution for this sort of working situation is establishing clear communication channels and utilizing collaborative tools and platforms. We communicate with each other on an instagram group and meet twice a week to discuss the project and related deliverables. Additionally, we have a shared google drive link to store important information and data. This drive also contains the rough versions of our project deliverables; we can use it to write side notes for our teammates explaining things that might be difficult to interpret at first glance. In the future, if these measures prove to be inadequate, we can try to meet more often during the week to discuss minor details that would otherwise be left for individuals to interpret. We could also have a timesheet to track individuals' time contributions which will help improve coordination and ensure that the work is distributed as fairly as possible.

9.2 What if training our model takes longer than expected?

Given that the team plans on having 38 distinct classes and approximately 91,000 images (including the overlapping images from the 2 datasets) [7][8] of training data, there is a fairly high probability of us getting delayed in preprocessing the data, training the model and tuning the hyperparameters to the point where our model is sufficiently accurate. For this purpose, we plan on focusing on only 10 classes initially. Thereafter, based on how long we take to train the model and how accurate it is, we shall add more classes and improve the capabilities of our model.

9.3 What if a team member is unable to complete their task by the team's internal deadline?

Two members of the team are based in Canada, in and around Toronto, and the other two are living abroad, approximately 10 hours ahead of Eastern Standard Time. Due to this time difference, there is a slight risk of conflicts in schedules or errors on behalf of the teammates. Thus, we may face

a situation where a team member is unable to complete their task by the internal deadline. In this case, either an extension can be given to that member, or their task shall be divided equally amongst the other members and the incident shall be recorded for future reference.

10 Link To GitHub

https://github.com/MuAbCh/CNN-Plant-Disease-Classification

References

Gupta Bhandari, Kamal Aryal, Nitesh Rijal, and Pawan Acharya. Plant diseases detection system using deep learning. pp. 1, 04 2022.

Y. Borhani, J. Khoramdel, and E Najafi. A deep learning based approach for automated plant disease classification using vision transformer. *Scientific Reports*, 12(1):1, 2022. doi: 10.1038/s41598-022-15163-0. URL https://www.nature.com/articles/s41598-022-15163-0.

Emmarex. Plant disease dataset. https://www.kaggle.com/datasets/emmarex/plantdisease, 2018.

Anna L. Frick, A. Bäckström, and E. Kuiper. Using deep learning for image-based plant disease detection, 2016. URL https://www.frontiersin.org/articles/10.3389/fpls.2016.01419/full.

Lori Tyler Gula. Researchers helping protect crops from pests. *Agriculture*, 12(10):2395, 2023. doi: 10.3390/agriculture12102395. URL https://www.nifa.usda.gov/about-nifa/blogs/researchers-helping-protect-crops-pests#:~:text=Each%20year%2C%20plant%20diseases%20cost,Organization%20of%20the%20United%20Nations.

MK Gurucharan. Understanding the basic architecture of convolutional neural networks (cnns). https://www.upgrad.com/blog/basic-cnn-architecture/, 2022. Accessed: June 16, 2023.

Ahmed Kayad and Ahmed Rady. Deep learning-based leaf disease detection in crops using images for agricultural applications. *Agronomy*, 12:2395, 2022. doi: 10.3390/agronomy12102395. URL https://www.mdpi.com/2073-4395/12/10/2395.

Sue Han Lee, Hervé Goëau, Pierre Bonnet, and Alexis Joly. Attention-based recurrent neural network for plant disease classification. *Frontiers in Plant Science*, 11:601250, 2020. doi: 10.3389/fpls.2020.601250. URL https://www.frontiersin.org/articles/10.3389/fpls.2020.601250/full.

Rushikesh Pupale. Support vector machines(svm) — an overview.

Garvit Singh. Plant disease detection dataset. https://www.kaggle.com/code/garvitsingh/plant-disease-detection/input, 2023.

Gula (2023) Frick et al. (2016) Borhani et al. (2022) Bhandari et al. (2022) Kayad & Rady (2022) Lee et al. (2020) Emmarex (2018) Singh (2023) Gurucharan (2022) Pupale