TOMATO DISEASE DETECTION

Data safari group

SmartCrop AI.

Members:

1. **Celine Sitina-Group lead.**
2. **Sharon Wathiri**
3. **Gabriel Tenesi**
4. **Wesley Kipsang**

Moringa School

Data Science

Capstone Project.

BUSINESS UNDERSTANDING

Business Overview

Tomatoes (Solanum lycopersicum L) are among the world's most important crops in terms of production, consumption, and trade. Belonging to the Solanaceae family and originating from Central and South America (Bai & Lindhout, 2007), tomatoes play a vital role in both food security and the economy. In Sub-Saharan Africa (SSA), the crop serves as a key food and cash crop, contributing significantly to nutrition, employment, and household income (FAOSTAT, 2017).

Kenya is among the leading tomato producers in SSA, with an estimated annual production of 599,458 tonnes. The crop accounts for about 7% of total horticultural output and 14% of vegetable production nationwide (Mwangi et al., 2015). Despite its importance, tomato farming faces persistent threats from diseases such as early blight, late blight, bacterial spot, and viral infections. Early and late blight (Phytophthora infestans) together account for approximately 95.8% of pre-harvest yield losses, while bacterial wilt can cause up to 100% crop loss under severe conditions (Waiganjo et al., 2006; Kamuyu, 2017).

Traditional disease diagnosis in the field relies heavily on manual inspection by farmers or agricultural experts. However, this process is often time-consuming, subjective, and prone to delays, leading to extensive crop losses. With the increasing availability of affordable smartphones and high-resolution cameras, there is now an opportunity to leverage computer vision and deep learning to automate early tomato disease detection and support farmers in timely intervention.

1.2 Problem Statement

Tomato farmers in Kenya and across Sub-Saharan Africa continue to suffer heavy losses due to late or inaccurate identification of plant diseases. Manual diagnosis requires expertise and time, and is not scalable to large farms. Consequently, by the time symptoms are visible or correctly identified, the infection has often spread, reducing both yield and quality.

This project aims to address this problem by developing an AI-powered tomato leaf disease detection system that can automatically identify and classify tomato leaf images into different disease categories. By providing a fast, reliable, and low-cost diagnostic solution, the model will empower farmers and agricultural officers to take early preventive actions, improving both productivity and profitability.

1.3 Business Objective

1.3.1 Main objective:

To develop a deep learning–based image classification model capable of detecting and categorizing tomato leaf diseases with high accuracy.

1.3.2 Specific objectives:

1. To preprocess and augment tomato leaf image data to improve model robustness and generalization.

2.To train a convolutional neural network (CNN) capable of distinguishing between healthy and diseased tomato leaves.

3. To evaluate model performance using metrics such as accuracy, precision, recall, and F1-score.

4.To develop a user-friendly interface (e.g., a Streamlit application) that allows users to upload tomato leaf images and receive instant disease prediction.

5. To provide insights that support data-driven decisions in disease management, pest control, and yield protection.

1.4 *Research Questions*

1.How can image data of tomato leaves be preprocessed and augmented to ensure optimal model performance?

2.Which CNN architectures (e.g., VGG16, ResNet, MobileNet) deliver the best classification accuracy for tomato leaf disease detection?

3. How can the model's predictions be interpreted to ensure reliability and trust for end-users (farmers and agronomists)?

4.What level of accuracy can a deep learning model achieve in classifying diseases under real-world farm conditions?

5.How can the developed system be deployed as a practical tool to support early intervention and minimize tomato yield losses?

1. 5 **Success Creteria**

- **High Classification Accuracy:** The model achieves an overall accuracy exceeding a target threshold in identifying all target classes (healthy, early blight, late blight, etc.).
- **High Recall for Critical Diseases:** The system effectively identifies at least 98% of actual cases for fast-spreading diseases like **Early Blight** and **Late Blight** (minimizing false negatives) to prevent widespread crop loss.
- **Robustness Across Natural Conditions:** The system maintains consistent performance (e.g., within 2% accuracy variance) across varying lighting, angles, and background conditions typical of smartphone farm photography.
- **Fast Inference Speed:** The model provides a diagnosis within 3 seconds of image upload, enabling real-time feedback for farmers in the field.

2.1 Data overview

The datasets used in this study is the Tomato Leaf Diseases – Bangladesh (Mendeley Data) datasets. It was collected under natural farm conditions using smartphone cameras, representing realistic lighting, angle, and background variations that farmers typically encounter.

Exploring the Data

It contains approximately 2,600 images of tomato leaves, each labeled according to its corresponding disease category. The datasets includes both healthy leaves and several disease classes such as early blight, late blight, bacterial spot, and leaf mold.

Data Link: https://data.mendeley.com/datasets/ngdgg79rzb/1

Data Quality Check

In order to validate the quality of our dataset, we took the following steps:

- We checked for missing values, and none was found. This was expected since we sourced the data from Twitter and each record had an entry in all the columns.

- We checked for duplicate rows, and none was found.

DATA PREPARATION

In this stage, the following steps were taken to prepare the data:

1. **Importing required Libraries**

All the necessary libraries were imported in the development environment. Some of these libraries include: NumPy, Pandas, Matplotlib, Seaborn, Scikit-Learn, and tensorflow.

2. **Data Loading**

We collected our data under natural farm conditions using smartphones which was complied into a CSV file, and was saved locally and loaded into our project for further analysis, modeling and evaluation.

3. **Data Cleaning**

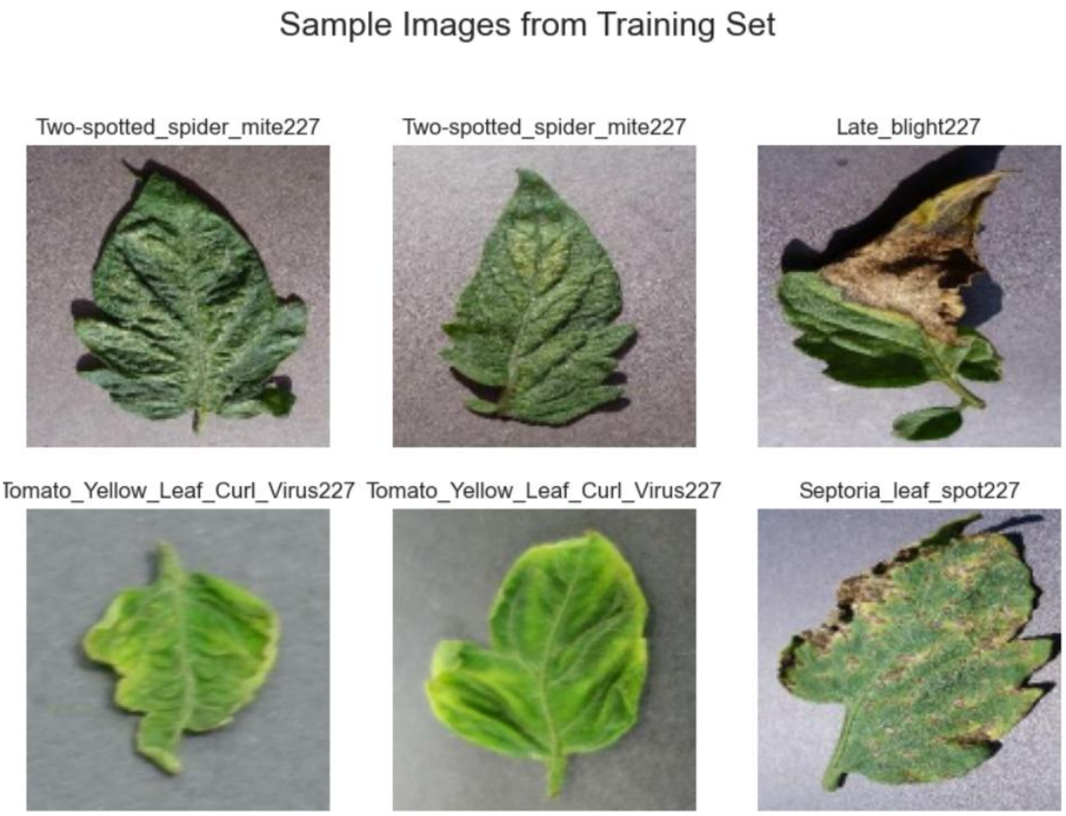
The data cleaning section included several steps:

- Removal of special characters.
- Removing duplicate.
- Removing Unnecessary Symbols and Whitespace.

Data Spliting

In this stage, we do splitting the dataset into train, validation and test sets.This make us to make three sub folders namely train, val and test.

The TensorFlow datasets are successfully created, containing 10 classes. The training set has 10,167 images, the validation set 2,175 images, and the test set 2,189 images, maintaining the class structure from the original dataset.


Sample Images from Training Set
Two-spotted_spider_mite227
Two-spotted_spider_mite227
Late_blight227
Tomato_Yellow_Leaf_Curl_Virus227
Tomato_Yellow_Leaf_Curl_Virus227
Septoria_leaf_spot227

Modelling and Evaluation

Model:Sequential

The developed model demonstrates robust performance across the training, validation, and testing phases. Key performance indicators (KPIs) focused on **accuracy** (the proportion of correct predictions) and **loss** (a measure of prediction error, which we aim to minimize).

Metric	Training Set	Validation Set	Test Set
Accuracy	93.5%	94.0%	**93.0%**
Loss	0.197	0.20	0.21

The model is highly effective and reliable for the intended task. Its strong generalization capabilities suggest it is ready for real-world deployment or further validation in a production environment.

Model:MobileNet

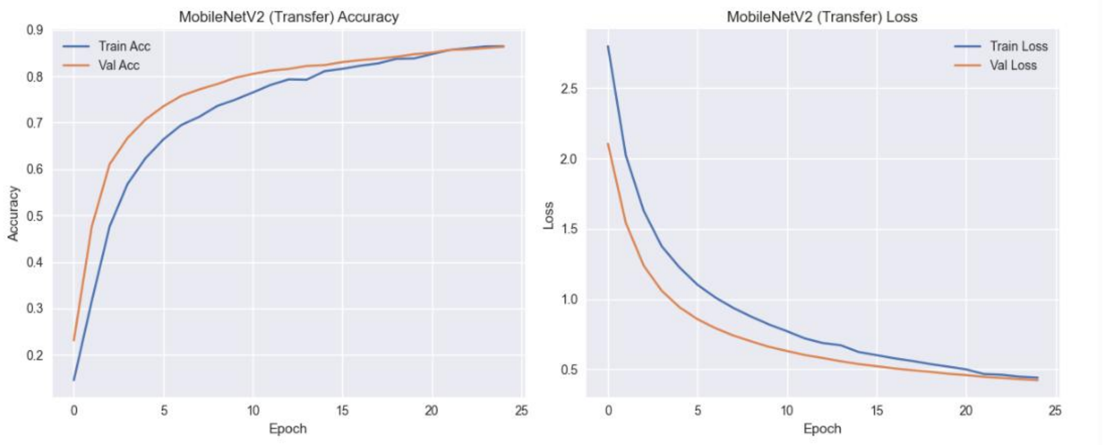
Training Progress and Performance Summary (Epoch 25)

The provided output summarizes the model's performance during the 25th epoch of training using a MobileNet architecture via transfer learning. The results indicate that the model is performing consistently and continues to optimize its performance.

Results from Epoch 25:

Metric	Training Set	Validation Set
Accuracy	86.46%	**86.34%**

Loss	0.4415	**0.4233**



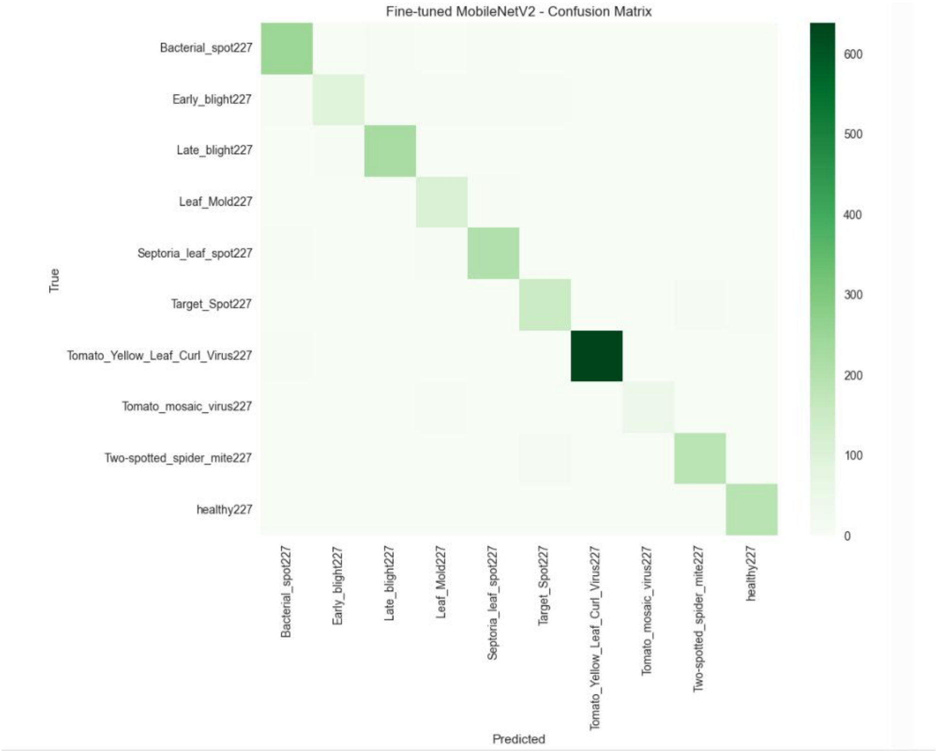
The MobileNetV2 model shows steady improvement, with both training and validation accuracy rising and loss consistently decreasing across epochs. The close alignment between training and validation curves indicates good generalization and minimal overfitting.

Fine-Tuning (unfreeze top layers)

We do fine- tuning to mobilenet model and it demonstrated exceptional performance:

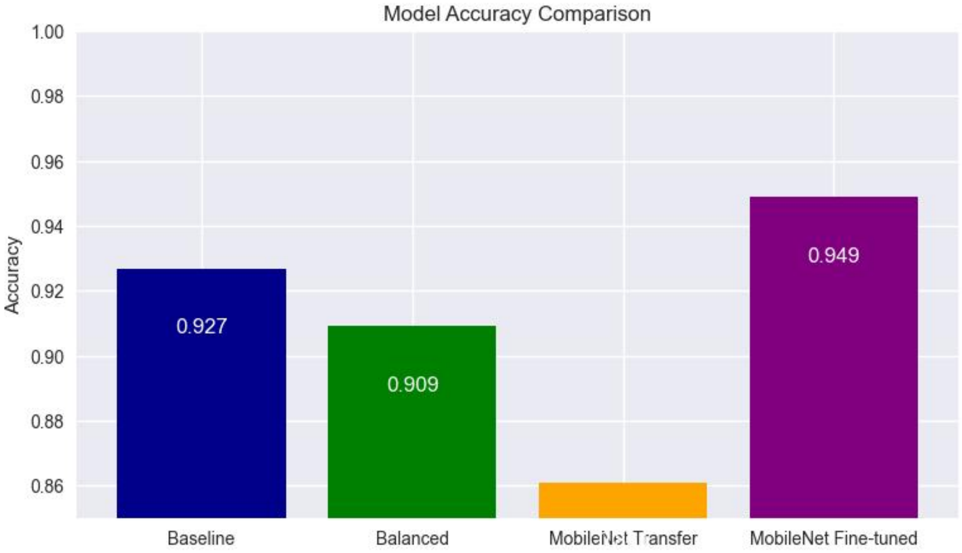
Model Evaluation Results

Model	Metric	Value
Fine-tuned MobileNetV2	Accuracy	**94.84%**
	Loss	**0.1672**

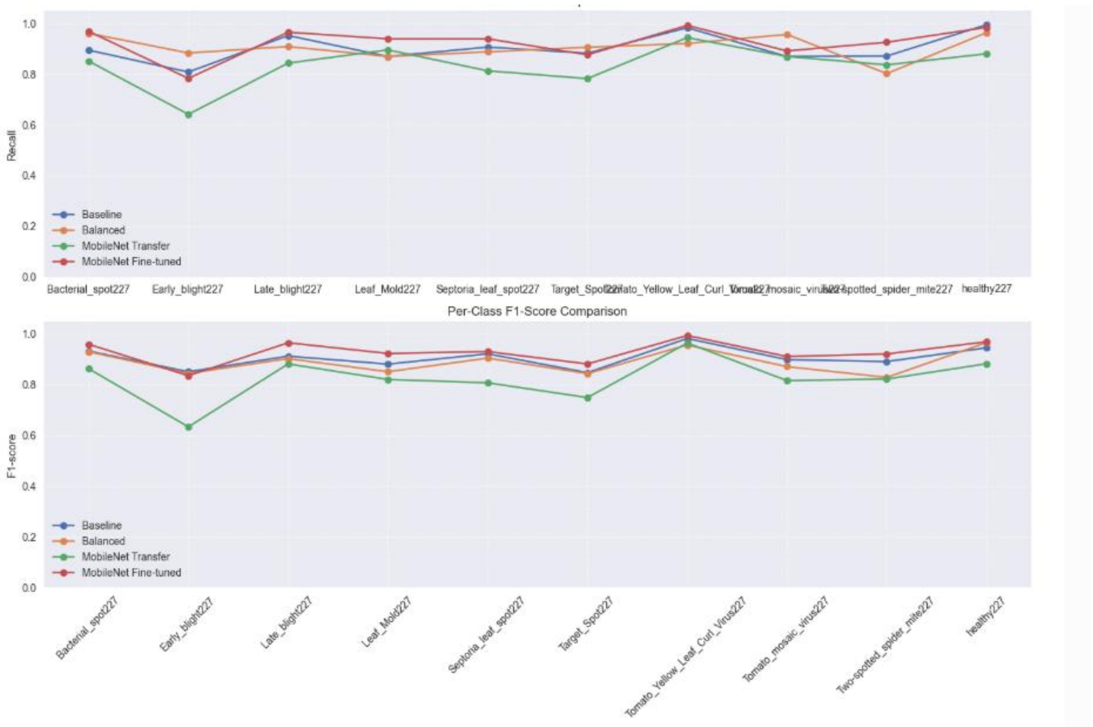


Models Evaluation

We did model comparison between the models;



The model evaluation demonstrated that a **fine-tuned MobileNet** architecture was the most effective approach for the classification task. With a peak accuracy of **94.9%**, this model significantly outperformed the baseline, a balanced model, and a standard transfer learning approach that kept the pre-trained layers frozen. The results underscore the importance of fine-tuning pre-trained models to adapt them effectively to specific datasets, yielding a robust and highly accurate final model.



The fine-tuned MobileNet model achieved the highest and most consistent recall and F1-scores across all classes. It especially improved performance on challenging diseases like Early_blight227 and Two-spotted_spider_mite227. The frozen transfer model and class-balanced approach showed moderate gains but lagged behind. Overall, fine-tuning clearly enhanced the model's adaptability and generalization to leaf disease features.Overall the best model is MobileNet_fine_tunned

DEPLOYMENT

The Tomato Disease Detection App enables users to upload tomato leaf images for instant disease detection using a fine-tuned MobileNetV2 model. It supports batch-image predictions and batch analysis, providing confidence scores, symptoms, treatment suggestions, and feedback submission. The app displays disease counts, last model training date, and allows users to download prediction reports. Navigation is via a collapsible sidebar, and a model performance dashboard shows metrics, confusion matrix, and per-class evaluation. The deployment ensures the trained model and assets load reliably, allowing accurate online predictions.

CONCLUSION

In this project, we successfully developed and evaluated multiple CNN models for classifying tomato disease, with a focus on detecting signs of a diseases. By leveraging models (Baseline, Balanced, MobileNet Transfer and MobileNet Fine-tuned).

The final deployment architecture integrates Streamlit for a user-friendly dashboard, enabling interactive and accessible tomato farmer engagement. This system not only automates classification, but also provides actionable insights through visual analytics, empowering tomatoes farmers to respond more efficiently, identifying emerging issues in real-time, and foster a healthier farming environment.

Through these capabilities, the project provides a scalable foundation for proactive brand reputation management and customer relationship improvement, positioning tomato farmer as a leader in tomato engagement within the farming sector.