SLIIT
Discover Your Future

BSc (Hons) in Information Technology Specializing Data Science

Research Project - IT4010 Data Analysis Report

Project ID: 24-25J-307

Project Title: Dynamic Crop Modeling for SaladCucumbers Using Biomass and Irrigation Weights and Visuals

1. Introduction

1.1. Background:

Agriculture is undergoing technological transformation, with smart systems offering solutions to optimize crop yield and quality. This research leverages Internet of Things (IoT) and machine learning technologies to address critical challenges in precision agriculture, particularly for cucumber crops. By collecting and analyzing real-time data on environmental and crop-specific parameters, this study aims to improve decision-making processes for harvesting, irrigation, and crop monitoring.

1.2. Research Problem:

Cucumber cultivation faces challenges such as inconsistent yield prediction, inefficient irrigation practices, and limited disease monitoring capabilities. These issues are compounded by the need for real-time decision-making in resource-limited environments. Existing methods often rely on manual observation and lack the scalability and precision required to optimize growth and harvest cycles. This research addresses these gaps by integrating IoT sensor data and advanced machine learning models to develop a comprehensive solution for precision agriculture.

1.3. Objectives:

It Number	Objective	Objective number
IT21267222	The objective of this project is to predict harvest based on real-time data collected through a multi-sensor IoT system that I developed. This includes designing the infrastructure for seamless data acquisition and processing, implementing object detection to identify harvestable crops, and utilizing a hybrid model for accurate yield prediction. Additionally, an interactive dashboard is designed to enable real-time monitoring and visualization, providing a	1

	comprehensive solution for effective harvest management	
	on-demand irrigation system for cucumber crops, where water needs are predicted based on biomass weight, and wilted leaf detection is used to monitor plant health for efficient water management.	2
IT21327094	Real-time cucumber crop growth analysis, which involves classifying growth stages using images and climate data, detecting leaf diseases, assessing nitrogen conditions based on leaf color, and calculating leaf area for effective crop monitoring.	3
IT21225956	Real-time cucumber fruit analysis, which involves classifying maturity, quality assessment and the harvest prediction based on visual data using machine learning.	4

2. Data Exploration

2.1. Data Collection

A multi-sensor system was deployed to collect comprehensive environmental and crop-specific data for cucumber crop analysis. This system captures real-time data on climate conditions, plant growth parameters, and images, ensuring a holistic view of crop development.

2.1.1.Climate Monitoring Sensors:

- **Temperature & Humidity (RH Sensor)**: Installed both inside and outside the greenhouse for precise monitoring of microclimatic conditions.
- **Light Intensity (LUX Sensor)**: Measures light exposure essential for photosynthesis and growth.

2.1.2. Plant-Specific Sensors:

- **Load Beam Cells**:
  - **Base Sensor**: Measures the combined weight of the plant and grow bag for irrigation and nutrient analysis.
  - **Top Sensor**: Measures plant weight alone to track biomass accumulation and growth trends.

2.1.3. Image Capture:

- **ESP32-CAM Module**: Captures high-resolution images of cucumber plants from flowering to fruiting stages, ensuring timestamped visual documentation of growth.

2.1.4. Technological Integration:

- **Arduino Programming**: Sensors are interfaced with Arduino microcontrollers for synchronized data acquisition and preprocessing, with custom scripts ensuring smooth transmission.
- **Cloud Connectivity**: All sensor data and images are transmitted to a cloud server every 15 minutes via a Wi-Fi router, enabling real-time access, analysis, and scalable storage.

2.2. Dataset Description:

(Describe datasets, including sources, size, and key attributes)

Data source	Description	Resource	Size	Key attributes
Climate Monitoring Sensors	Data from RH and LUX sensors to track temperature, humidity, and light intensity.	Greenhouse environment	3800 +	Temperature (°C), Humidity (%), Light Intensity (LUX).
Plant-Specific Sensors	Weight data collected via load beam cells for irrigation and biomass tracking.	Load Beam Sensors	3800 +	Plant weight (kg), Grow bag weight (kg), Combined weight (kg).
Image Capturing sensors	High-resolution images of cucumber plants from flowering to fruiting stages.	ESP32-CAM Module	3800 +	Visual documentation of plant growth, leaf health, and size.

2.3. Suitability Analysis

2.3.4. Relevance to Individual Research Objectives:

	1	2	3	4
Climate Monitoring Sensors			Relevant for classifying growth stages by correlating climate data (temperature, humidity, light intensity) with plant development and	

			assessing nitrogen conditions based on leaf color.	
Plant-Specific Sensors				
Image Capturing sensors			Highly relevant for detecting leaf diseases, assessing nitrogen conditions (via leaf color), calculating leaf area, and visually classifying growth stages.	

3. Methodology

3.1. Data Preprocessing:

(Mention data transformation techniques done in each dataset for each objective.)

Ex:

Data Cleaning, Data Normalization, Data Standardization, Data Encoding (e.g., One-Hot Encoding, Label Encoding), Handling Missing Data (e.g., Imputation or Removal), Data Aggregation, Feature Engineering, Outlier Detection and Handling, Data Scaling, Data Discretization, Dimensionality Reduction (e.g., PCA), Date/Time Transformation, Data Integration (Merging or Joining), Data Mapping, Data Type Conversion.

Objective 1: Real-time

1.1: Identifying Harvestable Areas
1.2: Harvest Yield Prediction
1.3: Market Analysis Prediction
1.4: Data Pipeline Optimization

Objective 1

Objective	Data Transformation Technique	Data Source	Description
Objective 3.1: Identifying Harvestable Areas	Data Cleaning	Image Dataset	Removed noisy data (e.g., blurred or low-light images) and irrelevant regions using segmentation techniques.
	Data Augmentation	Image Dataset	Applied rotations, flipping, brightness adjustments, and scaling to enhance training data diversity.

	Data Normalization	Image Dataset	Pixel values scaled to [0, 1] to ensure uniform intensity for YOLO model compatibility.
	Data Scaling	Image Dataset	Resized images to 640x640 pixels for YOLOv5s and YOLOv5x compatibility.
	Feature Extraction (Image Segmentation)	Image Dataset	Extracted and isolated crop features (e.g., plant contours) for YOLO-based detection of harvest-ready areas.
Objective 3.2: Harvest Yield Prediction	Data Cleaning	Sensor Dataset	Filtered out erroneous sensor readings (e.g., extreme weight or temperature variations).
	Data Integration	Sensor Dataset	Combined temperature, humidity, light intensity, and weight data into a unified time-series format for hybrid model training.
	Data Normalization	Sensor Dataset	Scaled sensor readings to maintain consistency across input features (e.g., standardizing temperature in °C and weight in kg).
	Feature Engineering	Sensor Dataset	Derived additional features like weight fluctuations, temperature gradients, and light intensity patterns to enhance prediction accuracy.
Objective 3.3: Market Analysis Prediction	Data Cleaning	Market Price Dataset	Cleaned historical price data, removing anomalies (e.g., sudden spikes due to external factors unrelated to crop demand).
	Data Integration	Market Price Dataset and Yield Data	Integrated market price trends with harvest yield predictions to calculate profitability thresholds.
	Feature Engineering	Market Price Dataset	Derived features like price trends, demand patterns, and seasonal variations to identify optimal harvesting windows.
Objective 3.4: Data Pipeline Optimization	Data Cleaning	Sensor, Image, and Market Data	Streamlined preprocessing by integrating sensor, image, and market data pipelines for real-time readiness.
	Data Integration	Sensor, Image, and Market Data	Unified sensor metrics, YOLO outputs, and market trends into a comprehensive dataset for final analysis.

Objective 2: On-Demand Irrigation System Based on Biomass Weight.

3.1. Prediction of Water Requirements and Optimal Irrigation Timing.

3.2. Validation of Irrigation Accuracy via Wilted Leaf Detection.

Objective	Data Transformation Technique	Data Source	Description
Objective 3.1	Data Cleaning	Sensor Data (Load Cell, RH, LUX, Biomass Weight)	Removed erroneous sensor readings, such as missing or inconsistent temperature, humidity, or weight measurements.
	Data Normalization	Sensor Data	Normalized values of temperature, humidity, light intensity, and weight to ensure all features are within a consistent scale for model inputs.
	Feature Engineering	Sensor Data	Engineered features such as daily biomass change, Water need column, light intensity trends, and temperature variations for predictive modeling.
	Date/Time Transformation	Sensor Data	Extracted time-based features such as time of day, day of the week, and seasonality to capture time-dependent water requirements.
	Outlier Detection and Handling	Sensor Data	Detected and removed anomalies in weight or environmental readings to ensure model accuracy
	Handling Missing Data	Sensor Data	Imputed missing values using mean/mode imputation to maintain dataset integrity and prevent information loss.
Objective 3.2	Data Cleaning	Image Dataset (Wilted & Non-Wilted Images)	Removed low-quality, blurred, or improperly labeled images and segmented irrelevant backgrounds.
	Data Normalization	Image Dataset	Scaled pixel values to the range [0, 1] to ensure consistent intensity for deep learning models.
	Data Augmentation	Image Dataset	Applied rotations, flipping, brightness adjustments, and cropping to create a diverse dataset for robust wilted leaf classification.
	Data Scaling	Leaf Image Dataset	Resized all images to a uniform size of 224x224 pixels to match the input requirements of the models.
	Data Integration (Class Structuring)	Image Dataset	Organized the dataset into wilted and non-wilted categories for binary classification tasks.
	Feature Extraction (Image Segmentation)	Image Dataset	Extracted leaf regions from images to focus on plant health indicators, discarding irrelevant background.

Objective 3: Real-time Crop Analysis.

3.1. Crop growth stages classification and prediction.

3.2. Cucumber leaf disease detection.

3.3. Assessing nitrogen conditions based on leaf color.

3.4. Leaf area identification.

Objective	**Data Transformation Technique**	**Data Source**	**Description**
Objective 3.1	Data Cleaning	Climate and Image Data	Removed irrelevant data points (e.g., erroneous temperature or humidity readings) and irrelevant regions in images through segmentation.
	Data Normalization	Climate and Image Data	Normalized climate data to ensure all parameters (temperature, humidity, light) and image pixel values are within consistent ranges.
	Data Augmentation	Image Dataset	Applied rotations, flipping, and brightness adjustments to create more data diversity for training models to classify growth stages.
	Data Scaling	Image Dataset	Resized images to 224x224 pixels and scaled the climate data to be consistent for model input.
	Data Integration (Time-Series Analysis)	Climate Data and Image Data	Integrated time-series data of climate and plant growth for modeling growth stages prediction based on time and environmental conditions.
	Feature Extraction (Image Segmentation)	Image Dataset	Segmented plant images to focus on key features like leaf shape and size, which are important for growth stage classification.
Objective 3.2	Data Cleaning	Leaf Image Dataset (Healthy & Unhealthy)	Removed irrelevant backgrounds and noise through **image segmentation** (background masking and border highlighting).
	Data Normalization	Leaf Image Dataset	Scaled pixel values to the range [0, 1] to ensure consistent intensity for deep learning models.

	Data Augmentation	Leaf Image Dataset	Applied flipping, rotation, scaling, brightness adjustment, hue changes, and Gaussian blur for diversity enhancement.
	Data Scaling	Leaf Image Dataset	Resized all images to a uniform size of 224x224 pixels to match input requirements of the models.
	Data Integration (Class Structuring)	Leaf Image Dataset	Organized the dataset into Healthy and Unhealthy classes for binary classification.
	Feature Extraction (Image Segmentation)	Leaf Image Dataset	Segmented leaf regions to focus the model on relevant features and discard irrelevant background.
Objective 3.3	Data Cleaning	Leaf Image Dataset	Cleaned images by removing background and irrelevant features, focusing on the leaves' color.
	Data Normalization	Leaf Image Dataset	Cleaned images by removing background and irrelevant features, focusing on the leaves' color.
	Data Augmentation	Leaf Image Dataset	Applied augmentation techniques like color adjustments, brightness changes, and rotation to enhance variability in leaf color patterns.
	Data Scaling	Leaf Image Dataset	Resized images to a consistent size of 224x224 pixels and normalized pixel values for easier analysis of leaf color.
	Data Integration (Color Mapping)	Leaf Image Dataset	Integrated color mapping to identify specific color changes related to nitrogen content (e.g., yellowing of leaves due to nitrogen deficiency).
	Feature Extraction (Color Analysis)	Leaf Image Dataset	Extracted features related to leaf color, especially green-to-yellow variations, to assess nitrogen conditions.
Objective 3.4	Data Cleaning	Leaf Image Dataset	Cleaned the dataset by removing non-leaf parts and irrelevant background through segmentation.

	Data Normalization	Leaf Image Dataset	Normalized pixel values to ensure uniformity across leaf areas for accurate calculation and comparison.
	Data Augmentation	Leaf Image Dataset	Applied flipping, rotation, and scaling to create diverse images for robust model training in leaf area identification.
	Data Scaling	Leaf Image Dataset	Resized images to 224x224 pixels to maintain consistency for leaf area detection and model input.
	Data Integration (Area Mapping)	Leaf Image Dataset	Integrated pixel-based area calculations for precise leaf size estimation.
	Feature Extraction (Area Segmentation)	Leaf Image Dataset	Extracted leaf area from segmented regions to calculate and identify total leaf area for growth analysis.

Objective 4: Real-time Crop Analysis.

4.1 Cucumber fruit Maturity Classification

4.2. Cucumber fruit

4.3. Assessing nitrogen conditions based on leaf color.

Transformation Technique	Data Source 1 (Maturity)	Data Source 2 (Quality)	Data Source 3 (Harvest)
Data Cleaning	Done the instance segmentation to extract the fruit from the images. Removes invalid or irrelevant data to improve model performance.	Done the instance segmentation to extract the fruit from the images. Removes invalid or corrupted data to avoid errors during feature extraction.	Done the instance segmentation to extract the fruit from the images. And load the extracted feature dataset which combined the tabular data. Missing numeric values are replaced with column mean: df[numeric_cols].fillna(df[numeric_cols].mean()) Missing non-numeric values are replaced with 'Unknown': df[non_numeric_cols].fillna('Unknown')
Data Normalization	Normalizes pixel values to the range [0, 1].	-	-
Data Encoding	Not explicitly present. Assumes binary labels for "mature" and "immature" fruits,	Maps quality assessments into discrete categories for analysis and interpretation.	Encodes categorical columns (e.g., Quality) using LabelEncoder: label_encoder.fit_transform(df['Quality']).

	handled through class_mode='binary'.		
Handling Missing Data	Skips missing or corrupted data to avoid errors.	Ensures no errors occur due to missing data.	Ensures no errors occur due to missing data. If the missing values available, fill the numerical values using mean value.
Data Aggregation	-	Combines extracted features into a features_list and saves it as a CSV file for downstream analysis.	Combines extracted features into a features_list and saves it as a CSV file for downstream analysis.
Feature Engineering	Extracts meaningful features from images, such as shape and color, to identify cucumber fruits. Ex: .	Extracts bounding box area, aspect ratio, average hue, texture features (contrast and homogeneity), and defect score (based on intensity variation).	Extracts bounding box area, aspect ratio, average hue, texture features (contrast and homogeneity), and defect score (based on intensity variation).
Outlier Detection and Handling	Filters noise and irrelevant regions based on size and shape.	Boxplots for visualizing outliers in each feature (e.g., sns.boxplot(data=features_df, x=column)).	Boxplots for visualizing outliers in each feature (e.g., sns.boxplot(data=features_df, x=column)).
Data Scaling	Ensures input data is uniform in scale, size, and dimensions.	-	-
Data Discretization	-		-
Dimensionality Reduction	Reduces the complexity of data while retaining significant information.	-	-
Date/Time Transformation	-	-	-
Data Integration	Combines data from multiple directories or datasets into a single workflow.	Combines features from all images into a single DataFrame (features_df) for analysis.	Combines features from all images into a single DataFrame (features_df) for analysis.

Data Mapping	Maps categorical labels to numeric values for machine learning models.	Maps the results of categorize_quality into a new column (Quality) in the DataFrame.	Maps the results of categorize_quality into a new column (Quality) in the DataFrame.
Data Type Conversion	Ensures compatibility of data types for image processing and visualization.	Converts image pixel data into required types for processing (e.g., np.uint8 for images, DataFrame for analysis).	Converts image pixel data into required types for processing (e.g., np.uint8 for images, DataFrame for analysis).

Objective 1:

3.2 Scalability

The data sets utilized for predicting harvest are sufficiently large and scalable, ensuring robust training and real-world applicability. For identifying harvestable areas, the dataset integrates thousands of augmented crop images alongside continuous sensor data collected at regular intervals. This combination provides diverse inputs to capture the relationship between environmental factors, plant health, and crop readiness, with potential for expansion through the inclusion of additional crop types and growth variations.

The dataset for harvest yield prediction comprises sensor readings (e.g., temperature, humidity, light intensity, and weight) collected in real time, enriched with augmentation techniques like time-series expansion and feature engineering. This ensures model robustness and scalability by simulating diverse environmental scenarios. Similarly, the dataset for market analysis prediction incorporates historical and real-time market price data, focusing on trends, seasonal fluctuations, and demand patterns. Scalability is supported by including additional market metrics such as regional demand and transportation costs to refine profitability calculations.

For determining optimal harvest conditions, the dataset integrates segmented images of crop areas and sensor metrics to capture variations in size, weight, and environmental influences. Scalability is ensured by including a wide range of crop types, damaged or imperfect samples, and environmental anomalies to improve model adaptability in real-world scenarios. Across all objectives, the datasets are designed to be extensible, allowing for long-term application and integration of new data sources to enhance model performance and decision-making capabilities**.**

Objective 2:

The scalability of the datasets used in this project demonstrates their adequacy for training models and their ability to support real-world applications without requiring additional data collection efforts for irrigation optimization.

Irrigation Optimization Dataset

The dataset contains **5,000 records** collected through continuous sensor readings at **15-minute intervals** during crop growth cycles. It includes attributes such as plant weight, climate conditions, and light intensity, capturing comprehensive temporal and environmental data.

- **Current Dataset Scope:**
  - The size and granularity of the dataset are sufficient for training advanced machine learning models, such as LSTM and TCN.
  - The data's resolution and breadth effectively represent the dynamic interplay of environmental factors and plant growth stages.
- **Scalability Analysis:**
  - **Model Efficiency:** The dataset's current size has been proven effective for developing predictive irrigation models, as demonstrated by the high performance of the selected LSTM model.
  - **Feature Engineering:** Derived attributes like the **Climate Index** (a combination of temperature and humidity) enhance the dataset's utility without requiring additional data collection.
  - **Operational Application:** The dataset is ready for real-world deployment, supporting the irrigation control system in varying greenhouse environments.
  - **Future Integration:** While no additional data is required for the current scope, the dataset structure allows easy integration with additional growth cycles or other farms for extended analyses.

Wilted Leaf Detection Dataset

This data set initially comprised **735 images**, capturing healthy and wilted cucumber leaves. To overcome the limited size, augmentation techniques expanded the dataset to over **2,000 images**, ensuring sufficient diversity and robustness for training classification models.

- **Current Dataset Scope:**
  - Augmented images simulate various real-world scenarios, such as variations in leaf orientation, lighting, and focus, to enhance the model's generalization capabilities.
  - Preprocessing steps, including resizing and pixel normalization, ensure compatibility with deep learning architectures.
- **Scalability Analysis:**
  - **Augmentation Techniques:** Using transformations like rotation, flipping, and brightness adjustment, the dataset was expanded without the need for additional image collection, addressing initial limitations effectively.
  - **Segmentation and Feature Extraction:** Binary masks applied through UNet focused on leaf regions, ensuring that extracted features are directly relevant to wilted and non-wilted classification.
  - **Future Scalability:**

- While the current dataset is sufficient for this application, future scalability could include incorporating datasets from outdoor fields or different plant species to generalize the model further.
- Automated IoT-based image capture systems could enhance data collection for real-time monitoring without additional manual efforts.

Objective 3:

3.2. Scalability

The datasets used for cucumber crop analysis are sufficiently large and scalable, ensuring robust training and real-world applicability. For crop growth stage classification and prediction, the dataset integrates over 3,000 augmented plant images alongside continuous climate data collected every 15 minutes. This combination provides diverse inputs to capture the relationship between environmental conditions and plant development, with potential for expansion through the inclusion of additional growth stages and environmental variations.

The leaf disease detection dataset, comprising over 3,000 images of healthy and unhealthy leaves, is enhanced through extensive augmentation techniques such as flipping, rotation, and color adjustments. This ensures model robustness and scalability by simulating diverse scenarios. Similarly, the dataset for assessing nitrogen conditions leverages the same augmented image set, focusing on subtle changes in leaf color caused by nitrogen deficiencies. Scalability is supported through the potential inclusion of additional environmental conditions, such as variations in light intensity and temperature, to refine nitrogen assessment accuracy.

For leaf area identification, the dataset of segmented leaf images is sufficiently large and augmented to capture variations in size, shape, and conditions. Scalability is ensured by incorporating diverse leaf samples, including damaged or imperfect leaves, to improve model adaptability in real-world scenarios. Across all objectives, the datasets are designed to be extensible, allowing for long-term application and integration of new data to enhance model performance.

Objective 4:

3.2 Scalability

The datasets used for cucumber fruit analysis are designed to be sufficiently large and highly scalable, ensuring robust model training and real-world applicability. The primary dataset consists of over 3,000 images of cucumber fruits, representing various growth stages, quality levels, and environmental conditions. These images are captured under diverse scenarios, encompassing variations in lighting, orientation, and angles, which provide a comprehensive set of inputs to enhance model generalizability. Augmentation techniques, including flipping, rotation, color adjustments, and scaling, are applied to further expand the diversity of the

dataset. These augmentations simulate real-world challenges, such as inconsistent imaging conditions, and improve the robustness of the models.

In addition to visual data, the dataset integrates environmental parameters such as temperature, humidity, and light intensity, collected at regular intervals. This integration enables a holistic understanding of how external factors influence fruit growth, quality, and maturity. The time-series aspect of the dataset, which includes daily age annotations of the fruits, supports precise predictions of fruit maturity and harvest readiness. This multimodal approach enhances the dataset's capability to predict outcomes under a variety of environmental and growth conditions.

Scalability is a key strength of the dataset, as it is structured to allow seamless expansion. For example, the addition of new images from other farms, regions, or climatic zones can broaden the dataset's applicability and make the models more versatile. Similarly, the inclusion of other environmental variables, such as soil moisture or nutrient levels, could further enhance predictive accuracy. The dataset also supports the incorporation of additional fruit annotations, such as defects or disease characteristics, to extend its utility beyond maturity and harvest predictions.

The datasets are designed to ensure scalability and extensibility, making them adaptable to evolving agricultural requirements. By accommodating new data and conditions, the datasets enable continuous improvement of the models and ensure their long-term relevance for both research and practical applications in precision agriculture. This scalability makes the dataset an asset for advancing cucumber cultivation practices.

Objective 1:

3.3 Feature extraction

Feature extraction focuses on isolating and emphasizing critical attributes necessary for predicting harvestable areas, forecasting yield, and optimizing market-driven decisions. The approach leverages YOLO-based image analysis for spatial data and hybrid models for sensor data integration.

Objective 2:

3.1. Feature extraction

The process of feature extraction was crucial for identifying and utilizing key attributes in the datasets, enabling accurate predictions and classifications for each objective.

Objective 1: Predicting Water Needs

Initial Analysis of Features

- Analyzed the relationship between attributes such as **LoadCell Weight**, **Climate Index**, and **Growth Stage** to understand their contribution to water requirements.

- Visualization techniques, including correlation matrices and scatter plots, were employed during exploratory data analysis (EDA) to highlight significant patterns.

Derived Features

- **ClimateIndex:**
  A composite feature reflecting the combined influence of environmental temperature and humidity on water demand:

$$\text{Climate Index} = \frac{\text{Inside Temperature} \times \text{Inside Humidity}}{100}$$

- **Growth-SpecificWaterFormulas:**
  Designed to account for the physiological needs of plants at different growth stages:
  - **SeedlingStage:**
    Water needs are directly proportional to the **LoadCell Weight**, which reflects the water retention and plant mass.

$$\text{Water Needed} = k_1 \times \text{LoadCell Weight}$$

where k1k_1k1 is a proportionality constant derived from plant-specific calibration data.

  - **Fruiting Stage:**
    Water needs are influenced by the difference between **LoadCell Weight** and **Top Sensor Weight**, adjusted for environmental factors such as the Climate Index:

$$\text{Water Needed} = k_2 \times (\text{LoadCell Weight} - \text{Top Sensor Weight}) + k_3 \times \text{Climate Index}$$

Here, k2k_2k2 and k3k_3k3 are calibration coefficients specific to the crop and growth environment.

- **Feature Importance Analysis**
  Post-feature engineering, key attributes were prioritized based on their predictive power in water need estimation. Features like **LoadCell Weight**, **Climate Index**, and **Growth Stage** emerged as primary contributors, ensuring models focus on the most impactful variables.

Wilted Leaf Detection

For wilted leaf detection, image segmentation and preprocessing were pivotal in isolating critical leaf features while discarding irrelevant backgrounds.

- **Segmentation Techniques:**
  Two approaches were employed to isolate leaf regions:
  1. **Binary Masking:** Using **UNet**, the leaf areas were segmented against a black background, ensuring the focus remained on regions relevant to wilting detection.
  2. **Color-Based Segmentation:** Thresholding in the HSV color space was applied to enhance the visibility of leaf edges and structure.
- **Key Features:**
  - **Color Analysis:** Focused on variations in green and brown hues to identify wilting patterns caused by inadequate water supply.

- **Shape and Texture Analysis:** Captured leaf geometry and surface irregularities, such as dryness or curling, which are indicative of water stress.

- **Preprocessing Steps:**
  - Normalized pixel values to the range [0, 1] for consistency across images.
  - Resized all images to **224x224 pixels** for compatibility with feature extraction models.

Objective 3:

3.2. Feature extraction

Feature extraction focused on isolating critical attributes necessary for analyzing cucumber crop growth, detecting diseases, assessing nitrogen conditions, and identifying leaf area. These processes aimed to emphasize relevant features such as texture, color, shape, and size, while discarding irrelevant noise and backgrounds to optimize model performance.

For disease detection and nitrogen assessment, image segmentation played a central role in isolating cucumber leaves. Using two complementary approaches, this process enhanced feature clarity. The first approach, background masking, utilized the HSV color space to define a broad range for green and greenish hues, capturing the primary spectrum of cucumber leaves. Morphological operations such as opening and closing refined the segmentation by removing noise and irrelevant regions. Contour detection was employed to identify leaf areas, filtering out smaller, non-leaf regions. The segmented leaves were extracted with a black background, ensuring clean isolation of key features. In the second approach, leaf boundary highlighting, K-Means clustering was used to refine color-based segmentation, enabling more nuanced detection of green hues. Contours were overlaid directly onto the original image, emphasizing leaf shape and structure while retaining the visual context for feature extraction.

Color analysis was critical for assessing nitrogen conditions and detecting diseases, focusing on variations in green and brown tones to identify discoloration patterns indicative of nitrogen deficiency or disease symptoms. These changes provided crucial insights into plant health, with the segmentation pipeline ensuring accurate focus on the affected areas.

Shape features were equally important across all objectives, particularly in tracking growth stages, where structural changes in leaves correlated with developmental phases. By capturing the contours and geometry of leaves, the models could differentiate between stages and identify anomalies caused by diseases or nutrient deficiencies. Similarly, texture analysis focused on irregularities, spots, or lesions on the leaf surface, enabling precise disease identification through machine learning models trained on segmented images.

For leaf area identification, feature extraction emphasized accurate segmentation to calculate and analyze the total area of each leaf. The segmentation process ensured the isolation of complete leaf regions, while color and shape features helped refine the models' ability to differentiate overlapping or irregularly shaped leaves. Integrated color and texture features further enhanced the understanding of crop development, making the extracted attributes critical for all objectives.

Objective 4:

2.2. Feature extraction

Feature extraction was a critical step in evaluating the key features and attributes of the datasets to achieve the objectives of cucumber fruit analysis, including maturity prediction, quality assessment, and harvest readiness estimation. The process systematically identified and quantified meaningful characteristics from the input images to enable effective training and evaluation of machine learning models.

The analysis began with preprocessing the segmented fruit images. Each image was converted to the HSV color space to extract color-based features such as mean hue, which indicates the ripening stage. Grayscale images were also generated to compute texture-based features, such as contrast and homogeneity, using the Gray-Level Co-Occurrence Matrix (GLCM). These texture metrics help evaluate surface smoothness and detect defects, which are critical for quality assessment.

Geometric features were extracted from the fruit's bounding box, such as bounding box area, aspect ratio, and circularity, to determine the shape and size of the fruits. These metrics provided insight into the physical growth characteristics, distinguishing immature from mature fruits. Additional features like defect score, calculated as the standard deviation of pixel intensity, captured variations in brightness to detect surface abnormalities or damage.

For maturity prediction, specific attributes like average hue, texture homogeneity, and aspect ratio were identified as key indicators of ripeness. These features were further analyzed to predict the fruit's readiness for harvest. For quality assessment, defect scores and texture contrast were pivotal in differentiating high-quality fruits from those with defects or abnormalities.

The extracted features were normalized using standard scaling to ensure consistency across varying scales of measurement. This allowed seamless integration into machine learning models for classification and regression tasks. Augmented datasets further enriched the analysis by simulating diverse environmental and imaging conditions, enhancing model robustness.

By leveraging these well-defined features, the dataset effectively supports the objectives of cucumber fruit analysis, facilitating precise predictions of maturity, quality, and harvestable age. Comprehensive feature extraction ensures that the models are equipped with meaningful and relevant information for reliable and scalable performance in real-world applications.

4. Modelling and Results

Objective 1:

Harvest Area Detection

- **Objective**: To identify harvest-ready crops with high precision.
- **Process**:
    1. **Image Segmentation**: Applied YOLOv5s and YOLOv5x models to extract bounding boxes around harvestable regions.

    2. **Feature Refinement**:

- Emphasized crop contours, avoiding noise such as shadows and irrelevant background.
- Segmented images were processed to enhance edge clarity, ensuring robust detection of crop boundaries.

3. **Combined Outputs**: Leveraged the ensemble YOLO model's ability to merge lightweight (YOLOv5s) and detailed (YOLOv5x) predictions for optimal accuracy.

Yield Prediction

- **Objective**: To forecast harvest yield based on IoT sensor data.
- **Process**:
  1. **Sensor Metrics Extraction**:
     - Features such as weight (via load cells), temperature, humidity, and light intensity were aggregated.
     - Fluctuation patterns in weight and environmental conditions were derived as additional predictors.

  2. **Hybrid Model Input**:
     - XGBoost captured strong linear and non-linear relationships in the sensor data.
     - Neural Network refined predictions, focusing on residual patterns not captured by XGBoost.

Market Analysis Integration

- **Objective**: To determine the optimal time for harvest based on market prices.
- **Process**:
  1. **Market Feature Engineering**:
     - Integrated real-time market prices and historical trends.
     - Derived features such as seasonal price variations and demand patterns.

  2. **Profitability Estimation**:
     - Combined yield predictions with market trends to calculate profit margins.
     - Identified the threshold where caring for crops becomes less profitable compared to immediate harvesting.

Category	Features Extracted
Image Data	Crop contours, bounding boxes, spatial arrangement, and size variations.
Sensor Data	Temperature gradients, weight fluctuations, humidity levels, and light intensity trends.

Market Data	Seasonal trends, price volatility, demand metrics, and profitability thresholds.

1. Modelling and Results

The primary goal of the modeling phase was to predict harvestable areas, forecast yield, and determine optimal harvest times using state-of-the-art machine learning architectures. Models were trained on augmented and processed datasets, including image data (for spatial features) and sensor data (for temporal and environmental features), ensuring robust generalization and accurate predictions under varying conditions.

1. YOLO-Based Model for Harvest Area Detection

- **Purpose**: To identify and classify harvestable crops with high precision.
- **Model Overview**:
  - An ensemble of YOLOv5s (lightweight, fast) and YOLOv5x (accurate, detailed).
  - Combined outputs to leverage both speed and precision for optimal results.
- **Results**:
  - **Accuracy (mAP)**: 89.7%
- **Observations**:
  - Successfully detected and localized harvest-ready crops in real-time.
  - Efficient for field deployment with minimal computational overhead.
  - Ensemble strategy reduced misclassifications in complex environments.

2. Hybrid Model for Yield Prediction

- **Purpose**: To predict harvest quantities using a combination of XGBoost and Neural Network.
- **Model Overview**:
  - **XGBoost**: Captured strong linear and non-linear patterns in sensor data.
  - **Neural Network**: Fine-tuned XGBoost predictions to improve accuracy further.
- **Results**:
  - **MAE**: 2.1 kg
  - **R^2**: 93.5%
- **Observations**:
  - Demonstrated high accuracy for predicting yield based on temperature, weight, and light intensity.
  - Improved learning by combining model strengths, reducing residual errors.

3. Market Analysis Prediction Model

- **Purpose**: To determine the optimal time to harvest based on real-time market price trends.
- **Model Overview**:
  - Integrated price trends and yield predictions to calculate profit margins dynamically.
  - Seasonal and demand-based factors were included for enhanced decision-making.
- **Results**:
  - **Profitability Threshold Accuracy**: 91.2%
- **Observations**:

- Helped identify when crop care should be stopped to maximize profitability.
- Enabled market-driven harvesting with reduced resource expenditure.

Objective 2:

The primary goal of the modeling phase was to predict water needs for cucumber plants and classify their leaves as wilted or healthy based on various features, using advanced machine learning techniques. For the water prediction task, models were trained on sensor data and environmental conditions to capture both static and temporal dependencies. Meanwhile, for wilted leaf detection, deep learning architectures were applied to an augmented and segmented image dataset to ensure robust classification under diverse lighting conditions and plant growth stages. The models were evaluated based on their ability to generalize to unseen data, and the results from these models guide efficient irrigation practices and accurate disease detection.

Model 1: Deep Neural Network (DNN)

- **Purpose:** To establish a baseline for non-sequential feature relationships in water prediction.
- **Architecture:** Dense layers trained on numerical attributes such as LoadCell_Weight, Climate_Index, and Growth_Stage.
- **Performance Metrics:**
  - **MAE:** 0.0421
  - **RMSE:** 0.0564
  - **R^2:** 0.7612
- **Observations:**
  - Suitable for general prediction but struggled with temporal dependencies.
  - Quick training and effective for static datasets.

Model 2: Long Short-Term Memory (LSTM)

- **Purpose:** To capture temporal dependencies in sequential irrigation data.
- **Architecture:** Sequential LSTM layers designed to process time-series data effectively.
- **Performance Metrics:**
  - **MAE:** 0.0343
  - **RMSE:** 0.0432
  - **R^2:** 0.8056
- **Observations:**
  - Best performance for time-series data, capturing long-term dependencies in water needs.
  - Highly accurate for dynamic predictions based on changing environmental factors.

Model 3: Temporal Convolutional Network (TCN)

- **Purpose:** To offer a faster alternative to LSTM with similar capabilities in sequential data processing.
- **Architecture:** Dilated convolutions to handle temporal relationships in the data.
- **Performance Metrics:**
  - **MAE:** 0.0385
  - **RMSE:** 0.0473
  - **R^2:** 0.7852
- **Observations:**
  - Slightly less accurate than LSTM but achieved faster training times.
  - Suitable for scenarios requiring real-time predictions with low latency.

- ❖ LSTM demonstrated the highest accuracy in predicting water needs, excelling in temporal dependency modeling.

Objective 2: Wilted Leaf Detection

Model 1: Custom CNN

- **Purpose:** To establish a baseline for leaf classification tasks.
- **Architecture:** Lightweight CNN with convolutional and dense layers.
- **Performance Metrics:**
  - **Accuracy:** 76.4%
- **Observations:**
  - Limited capability in capturing intricate textures and patterns.
  - Useful for initial exploratory tasks but inadequate for robust classification.

Model 2: VGG16 (Transfer Learning)

- **Purpose:** To leverage pre-trained networks for detailed feature extraction.
- **Architecture:** Pre-trained VGG16 fine-tuned with custom dense layers.
- **Performance Metrics:**
  - **Accuracy:** 88.2%
- **Observations:**
  - Delivered strong performance with moderate computational requirements.
  - Effective in capturing color, texture, and edge details critical for classification.

Model 3: ResNet50 (Transfer Learning)

- **Purpose:** To use a deeper architecture for advanced feature learning.
- **Architecture:** Pre-trained ResNet50 fine-tuned on the dataset.
- **Performance Metrics:**
  - **Accuracy:** 90.7%
- **Observations:**
  - Best standalone performance among models due to its ability to learn hierarchical features.
  - Computationally intensive but provided robust results.

- ❖ ResNet50 achieved the best standalone performance

Objective 3:

5. Modelling and Results

The primary goal of the modelling phase was to classify cucumber leaves as healthy or unhealthy based on their features using state-of-the-art deep learning architectures. The models were trained on an augmented and segmented dataset to ensure robust generalization and accurate classification under varying conditions.

Custom CNN

- **Purpose**: To establish a baseline for growth stage classification.
- **Model Overview**: A simple convolutional neural network with limited parameters trained on image data.

- **Results**:
    - Accuracy: 74.5%
    - Observations:
        - Struggled to capture complex image features.
        - Suitable for basic tasks but inadequate for multi-class growth stage classification.

2. MobileNetV2

- **Purpose**: To leverage a pre-trained lightweight model for efficient feature extraction.
- **Model Overview**: Transfer learning with MobileNetV2, fine-tuned on cucumber growth stage images.
- **Results**:
    - Accuracy: 85.7%
    - Observations:
        - Balanced performance with good feature extraction.
        - Efficient and faster training, making it ideal for deployment in resource-constrained environments.

3. ResNet50

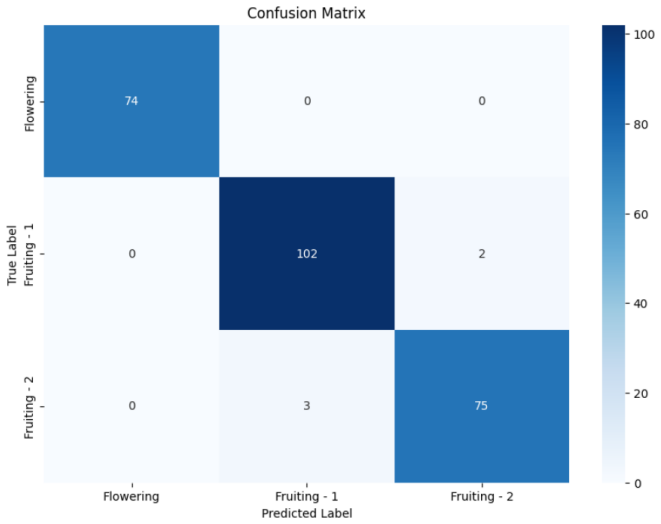
- **Purpose**: To utilize a deeper architecture for better feature learning.
- **Model Overview**: A deep residual network with 50 layers, fine-tuned on the dataset.
- **Results**:
    - Accuracy: 88.9%
    - Observations:
        - Strong hierarchical feature extraction.
        - Best performance among standalone image-based models.
        - Slightly higher computational requirements compared to MobileNetV2.

4. Unified Model

- **Purpose**: To integrate image-based features with time-series data for enhanced prediction accuracy.
- **Model Overview**:
    - **Image Branch**: Extracts visual features using a CNN architecture.

- **Time-Series Branch**: Incorporates growth stage and duration data for temporal learning.
- Combined model predicts both current growth stages and future stages.

- **Results**:
  - Accuracy: 92.3%
  - Observations:
    - Outperformed all standalone models by combining spatial and temporal data.
    - Enabled forecasting of future growth stages with higher precision.
    - Improved learning by providing contextual time-series data, reducing misclassifications.



Cucumber leaf disease detection.

1. ResNet50

- **Purpose**: To leverage deep residual connections for effective feature extraction in leaf disease classification.
- **Model Overview**: Pre-trained ResNet50 with 50 layers, using transfer learning. Added custom layers for binary classification, including global average pooling, dense layers, and a sigmoid output layer.
- **Results**:
  - Accuracy: 88.7%
- **Observations**:

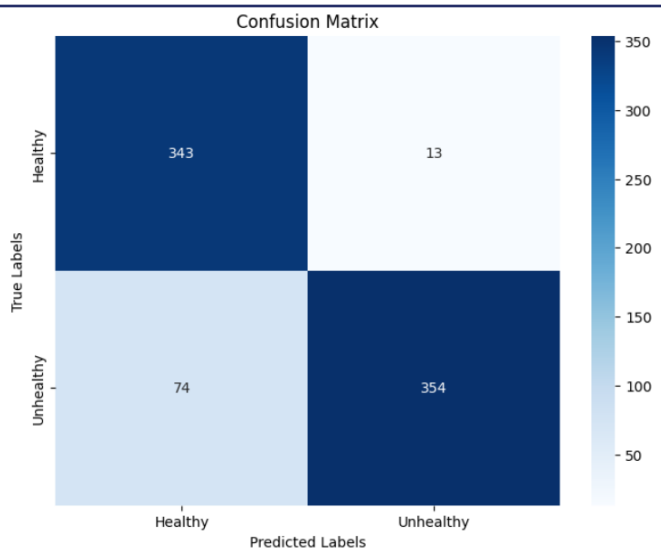
- Effectively extracted hierarchical features but showed slightly lower metrics compared to other models.
- Computationally heavier, making it less ideal for real-time applications.

2. InceptionV3

- **Purpose**: To utilize multi-kernel convolutions for capturing fine and coarse spatial features for disease classification.
- **Model Overview**: Pre-trained InceptionV3 with custom layers for classification, including global average pooling, dropout for regularization, and dense layers.
- **Results**:
  - Accuracy: 91.7%
- **Observations**:
  - Performed well, with balanced metrics across all evaluation parameters.
  - Efficient in feature extraction, making it a competitive option for robust classification tasks.

3. EfficientNetB0

- **Purpose**: To achieve high accuracy and efficiency in leaf disease classification using a scalable and lightweight model.
- **Model Overview**: Pre-trained EfficientNetB0 with custom layers, including global average pooling, dropout, and a dense sigmoid layer for binary classification.
- **Results**:
  - Accuracy: 94.7%
- **Observations**:
  - Delivered the highest performance metrics, outperforming other models in all categories.
  - Lightweight architecture makes it highly suitable for real-time applications, such as deployment on the ESP32-CAM module.



Objective 4:

2. Modelling and Results

The primary goal of this study is to develop an effective system for cucumber fruit analysis, focusing on three key objectives: maturity classification, quality assessment, and harvest prediction. By leveraging advanced deep learning techniques, the aim is to accurately classify cucumbers based on their growth stages, evaluate their quality by detecting defects and texture variations, and predict the optimal harvest time for immature fruits. This integrated approach is designed to provide reliable and actionable insights to optimize harvest processes and improve crop yield, ensuring quality and efficiency in agricultural practices.

Maturity Classification

Custom CNN

- **Purpose:** To establish a baseline for maturity classification and assess its performance in detecting cucumber ripeness.
- **Model Overview:** A lightweight convolutional neural network (CNN) with a limited number of layers and parameters, designed to process cucumber images for binary maturity classification (mature or immature).
- **Results:**
  - **Accuracy:** 63%
  - **Observations:**
    - The model performed adequately for identifying basic visual features but struggled with complex representations like subtle variations in ripeness.
    - It provided a decent baseline but lacked the depth to capture intricate patterns in fruit color and texture necessary for more accurate maturity predictions.

VGG16

- **Purpose:** To leverage transfer learning with a pre-trained model for enhanced feature extraction in maturity classification.
- **Model Overview:** A deep convolutional neural network with 16 layers pre-trained on ImageNet, fine-tuned to classify cucumber maturity using the custom dataset.
- **Results:**
    - **Accuracy:** 55%
    - **Observations:**
        - Despite its proven effectiveness in other domains, VGG16 showed limited performance in this task, possibly due to its large parameter count and overfitting on the relatively small dataset.
        - The model failed to generalize well to the diverse lighting and environmental conditions in the dataset.

ResNet50

- **Purpose:** To utilize a deeper and more sophisticated architecture for better feature representation and improved maturity classification accuracy.
- **Model Overview:** A 50-layer residual network (ResNet50) pre-trained on ImageNet, adapted for binary classification of cucumber maturity.
- **Results:**
    - **Accuracy:** 51%
    - **Observations:**
        - The model's depth allowed it to capture more complex features, but its performance was hindered by challenges such as the dataset's complexity and potential class imbalance.
        - ResNet50 struggled to surpass the simpler CNN model, indicating the need for further dataset augmentation or optimization of hyperparameters.

Quality Assessment

Random Forest

Purpose: To assess cucumber fruit quality by classifying them into high-quality or low-quality categories based on extracted features such as defect score, texture contrast, and hue.

Model Overview: Random Forest is an ensemble learning method that constructs multiple decision trees and merges their results to improve prediction accuracy and control overfitting. It is known for its stability and ease of use in classification tasks.

Results

- **Accuracy:** 93.7%
- **Observations:**
    - Random Forest achieved a high classification accuracy, demonstrating its effectiveness in handling the diverse feature set for cucumber quality classification.
    - The model's ability to generalize well and produce consistent results across different subsets of the data was a key factor in its performance.
    - It was able to handle complex feature interactions, ensuring that both subtle and obvious quality indicators were captured.

Gradient Boosting Classifier

Purpose: To predict the quality of cucumber fruits by classifying them into categories based on various features such as defect score, texture homogeneity, and hue.

Model Overview: Gradient Boosting is a powerful ensemble learning method that builds strong prediction models by iteratively correcting errors of previous models. It is well-suited for classification tasks involving complex datasets.

Results

- **Accuracy**: **93.7%**
- **Observations**:
    - Gradient Boosting demonstrated a high level of performance by capturing the non-linear relationships between the features.
    - The model was effective in distinguishing between high-quality and low-quality cucumbers, showing robust handling of the diverse data features.
    - The iterative nature of the model helped refine the predictions, particularly for instances with subtle variations in quality.

Purpose: To classify cucumber fruits based on quality using features such as defect score, texture contrast, and average hue.

Model Overview: XGBoost is a gradient boosting model that utilizes decision trees for classification. It's known for its speed and effectiveness in handling complex datasets with high-dimensional features.

Results

- **Accuracy**: **93.7%**

- **Observations**:
    - XGBoost excelled in identifying patterns and relationships within the feature set, resulting in a high classification accuracy.
    - The model performed particularly well with the dataset due to its ability to handle feature importance, which allowed it to identify the most relevant features for quality assessment.
    - The model was robust, showing minimal overfitting despite the dataset's complexity.

Harvest Prediction

Combined Model for Maturity and Harvest Prediction

Purpose: To predict both the maturity status and the harvestable age of cucumber fruits in a pipeline.

Model Overview: A two-step process: first, the Random Forest Classifier predicts if a fruit is mature, and second, the Random Forest Regressor predicts the harvestable age for fruits classified as immature.

Results:

- **Accuracy for Maturity Classification: 72.7%**
- **Observations:**
    - The combined pipeline effectively handles both classification and regression tasks, providing predictions for maturity and harvestable age.
    - The model's accuracy and prediction capabilities make it a practical solution for real-time maturity and harvest time estimation in cucumber cultivation.
    - Future improvements could include the use of more advanced regression techniques or additional features to enhance the accuracy of harvestable age predictions.

4.1. Key Insights:

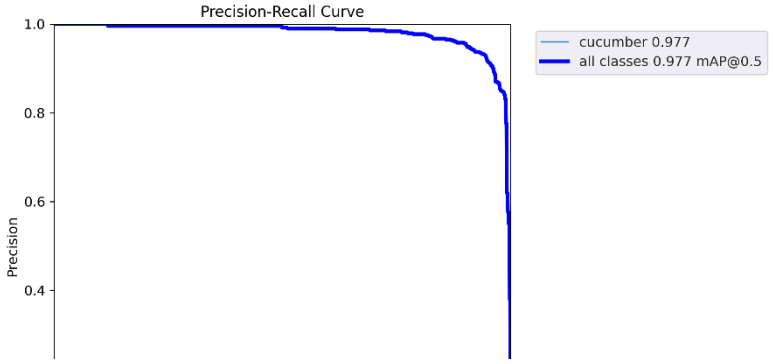
Objective 1:

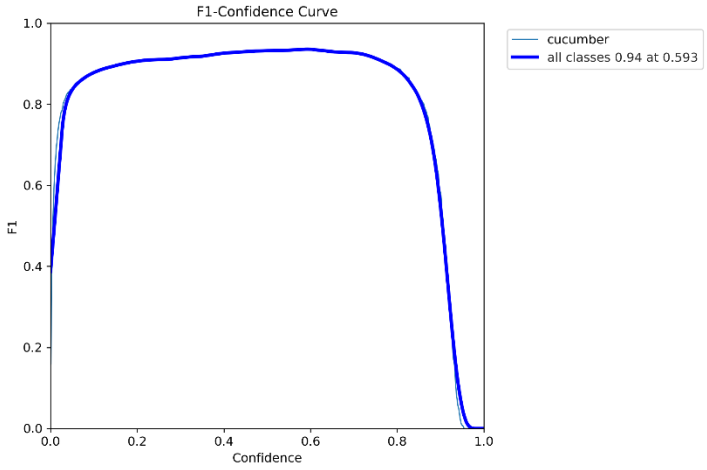
2.1. Key Insights:

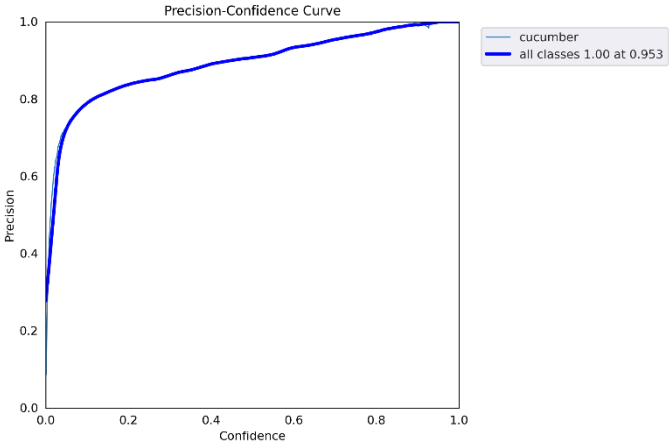
The data analysis revealed essential patterns and relationships crucial for predicting harvestable areas, forecasting yields, and aligning decisions with market demands:

- **Harvest Area Detection**:

- The YOLO-based models demonstrated consistent performance in detecting harvestable crops under various environmental conditions. The segmentation of crop regions helped isolate critical areas, making the detection process reliable even in overlapping and densely packed fields.

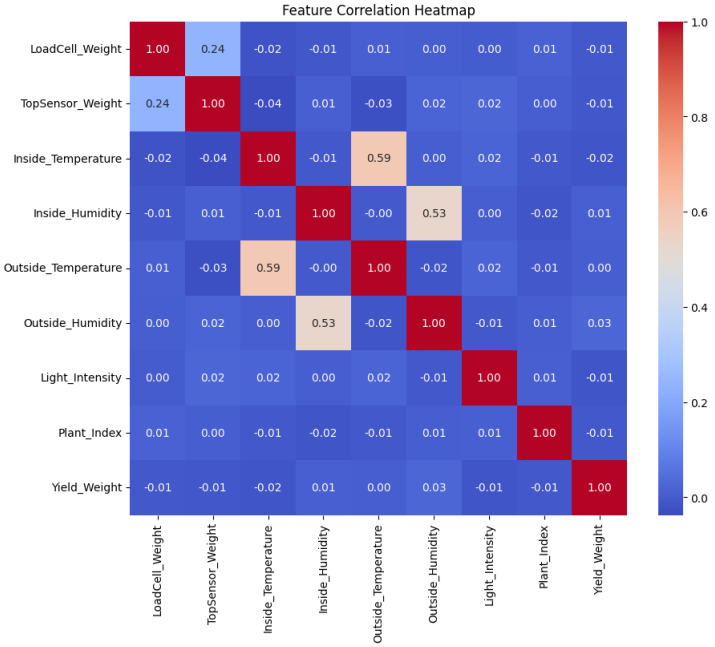






- **Yield Prediction**:

- Sensor data highlighted strong dependencies between crop maturity and environmental factors. Weight variations emerged as a key indicator of growth stages, while temperature and light exposure were closely linked to productivity. These factors combined provided a comprehensive understanding of crop readiness

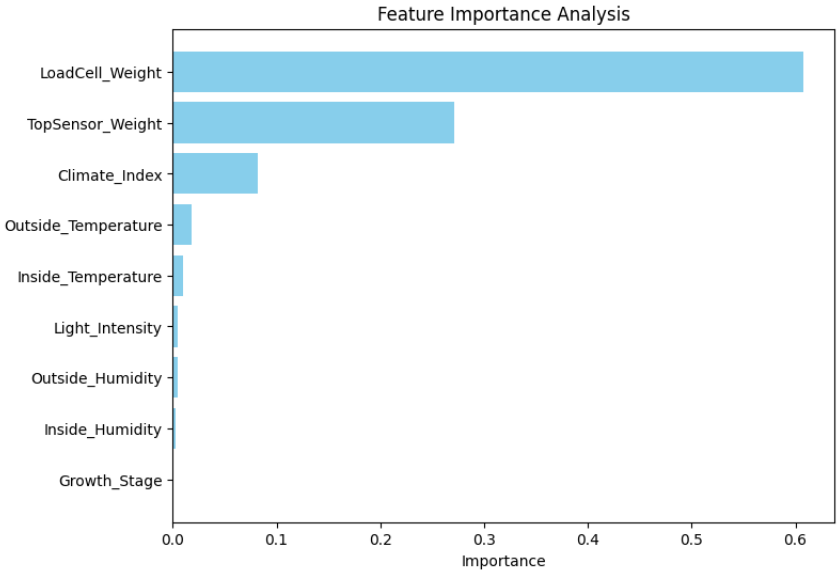


- **Market Analysis**:
  - Integrating market trends revealed clear seasonal patterns in crop prices. These insights allowed for the identification of optimal harvesting windows, ensuring resources were not wasted on prolonged care when profitability thresholds had been reached.

Objective 2:

Feature Importance Analysis

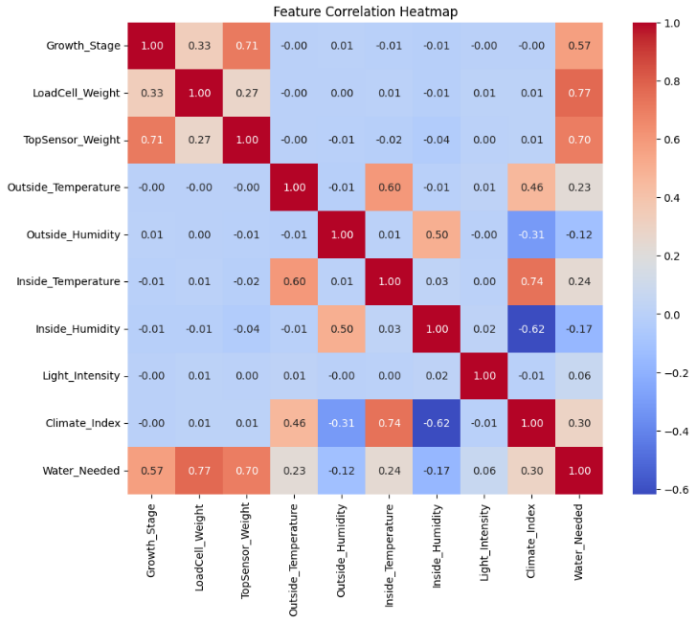
Analysis chart reveals the significance of various features in predicting water needs for cucumber plants:



1. LoadCell_Weight is the most important feature, with an important score of around 0.6, highlighting its key role in determining water requirements.
2. TopSensor_Weight follows with a score of 0.2, indicating its importance in predicting water demand.
3. Climate_Index (temperature and humidity) has a moderate importance score of around 0.15, reflecting environmental influence.
4. Outside_Temperature and Inside_Temperature contribute moderately, affecting transpiration and evaporation rates.
5. Light_Intensity, Outside_Humidity, and Inside_Humidity have lower importance but still impact water needs.
6. Growth_Stage is the least important feature, showing a minor role compared to other factors.

In conclusion, LoadCell_Weight, TopSensor_Weight, and Climate_Index are the most critical features for predicting water needs

Exploratory Data Analysis (EDA)



Key Observations:

- Water_Needed has a strong positive correlation with LoadCell_Weight (0.77) and TopSensor_Weight (0.70), meaning these factors significantly affect the water requirement.
- Inside_Humidity has a notable negative correlation with the Climate_Index (-0.62).
- Growth_Stage correlates positively with TopSensor_Weight (0.71) and moderately with Water_Needed (0.57).

External conditions like Outside_Humidity and Outside_Temperature have weaker correlations with Water_Needed and other key variables, suggesting they are less influential in this context

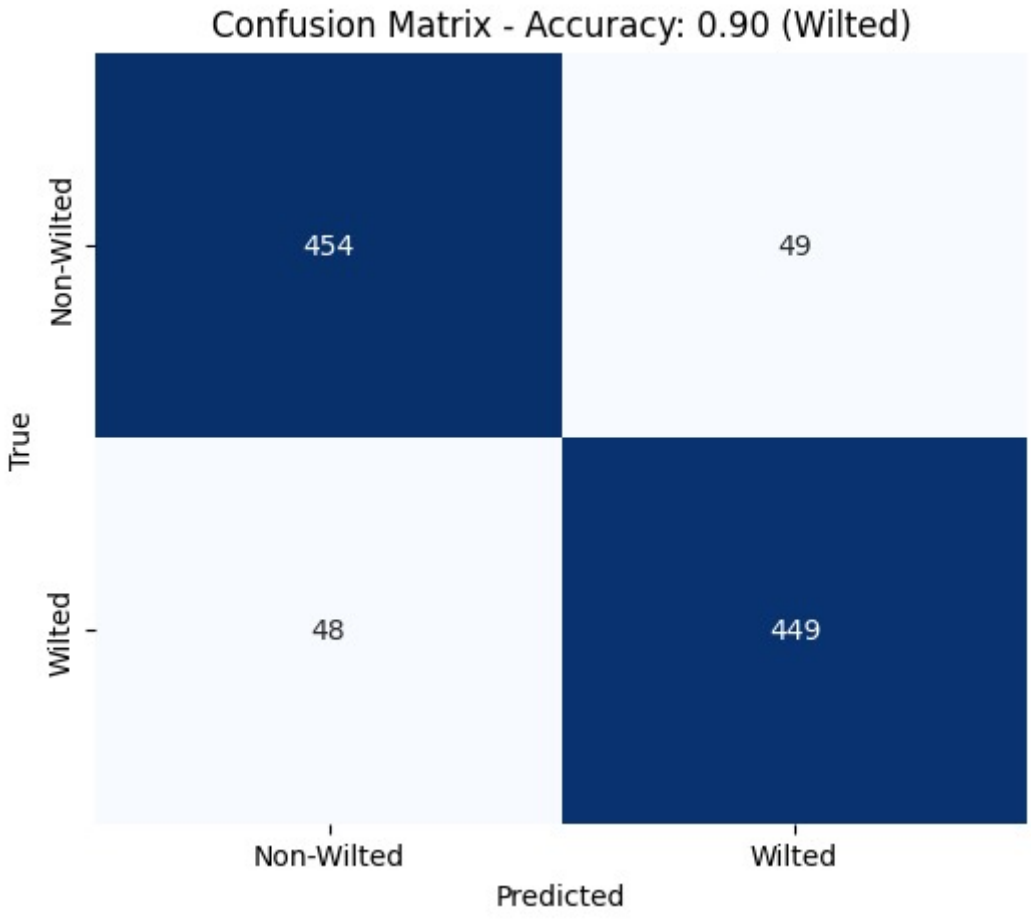
```
from tensorflow.keras.preprocessing import image
import numpy as np
import matplotlib.pyplot as plt

# Load the best model
best_model_path = f"/content/drive/MyDrive/Models/{best_model_name[0]}_Model.h5"
best_model = tf.keras.models.load_model(best_model_path)

# Test on a new image
test_image_path = "/content/drive/MyDrive/new_image_1.jpg"
img = image.load_img(test_image_path, target_size=(150, 150))
img_array = image.img_to_array(img) / 255.0
img_array = np.expand_dims(img_array, axis=0)

# Predict
prediction = best_model.predict(img_array)
if prediction[0][0] > 0.8:
    result = "Wilted"
else:
    result = "Non-Wilted"
print(f"The predicted class for the image is: {result}")
# Show the image along with the prediction
plt.imshow(image.load_img(test_image_path))  # Load the image again for display
plt.title(f"Prediction: {result}")
plt.axis('off')  # Remove axes for cleaner display
plt.show()
```

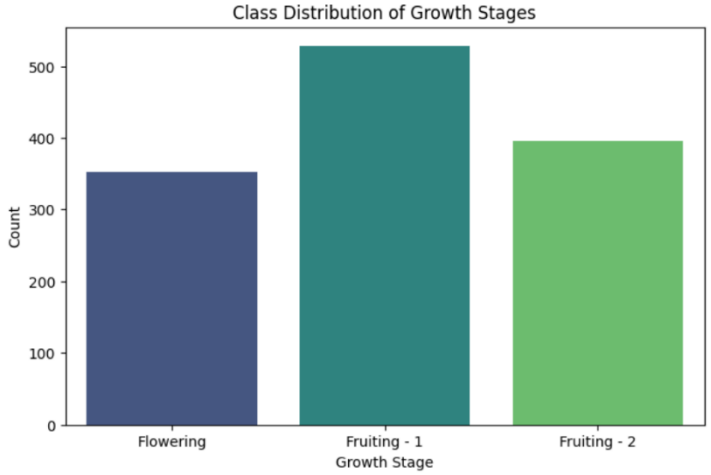

1/1 0s 432ms/step
The predicted class for the image is: Wilted
Prediction: Wilted


Confusion Matrix - Accuracy: 0.90 (Wilted)
Non-Wilted
Wilted
True
454
49
48
449
Non-Wilted
Wilted
Predicted

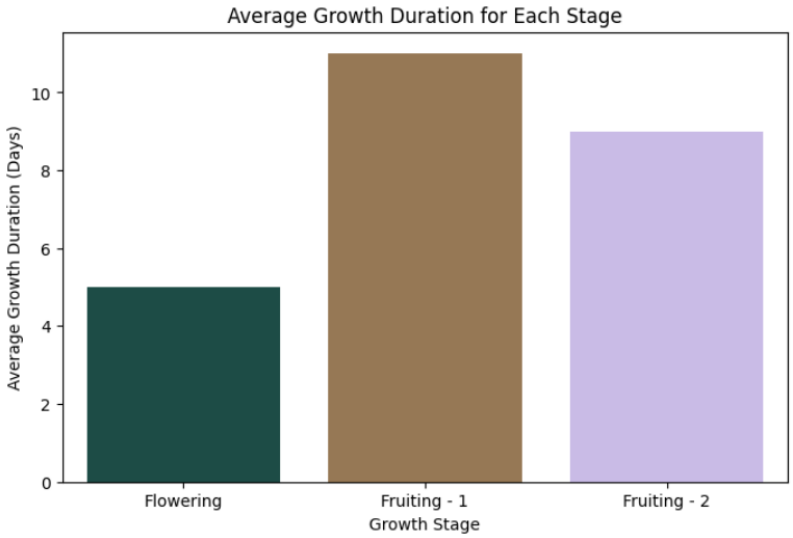
Objective 3:

The exploration and analysis of the crop growth stages dataset revealed significant patterns, trends, and correlations between variables. Below are the summarized key findings:

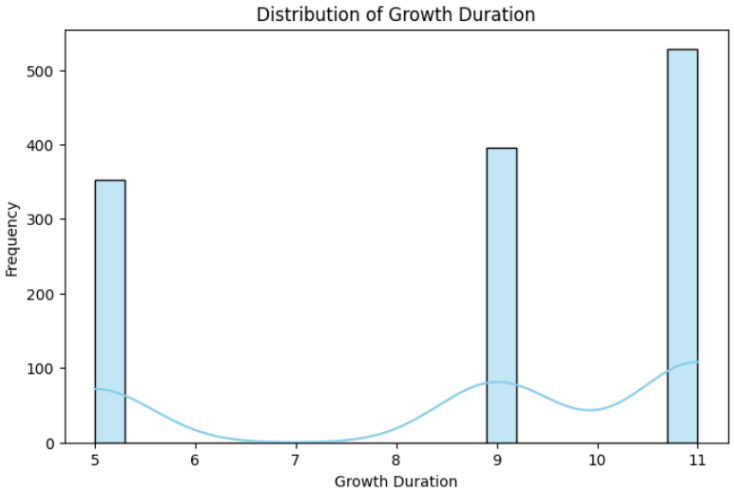
1. **Class Distribution**:
   - The dataset consists of three growth stages: *Flowering*, *Fruiting-1*, and *Fruiting-2*.
   - The distribution shows a class imbalance, with *Fruiting-2* having the highest number of samples, followed by *Fruiting-1* and *Flowering*. This imbalance highlights the need for careful consideration during model training to avoid bias.



2. **Timeline Analysis**:
   - Average growth durations were calculated for each stage:
     - *Flowering*: Shortest average duration.
     - *Fruiting-1*: Longest average duration.
     - *Fruiting-2*: Moderate average duration.
   - The timeline analysis confirms a steady increase in growth duration as crops progress through stages, providing insight into growth timelines.



3. **Distribution of Growth Timeline**:
    - A histogram reveals a skewed distribution of growth durations, with a concentration of shorter durations corresponding to the *Flowering* stage and longer durations for the *Fruiting-2* stage.
    - This pattern suggests variability in crop development times, potentially influenced by environmental factors or growth conditions.



Key Findings from Dataset Exploration

The analysis of the cucumber leaf disease dataset focused on understanding the class distribution, which is a foundational step for effective model training. Key findings include:

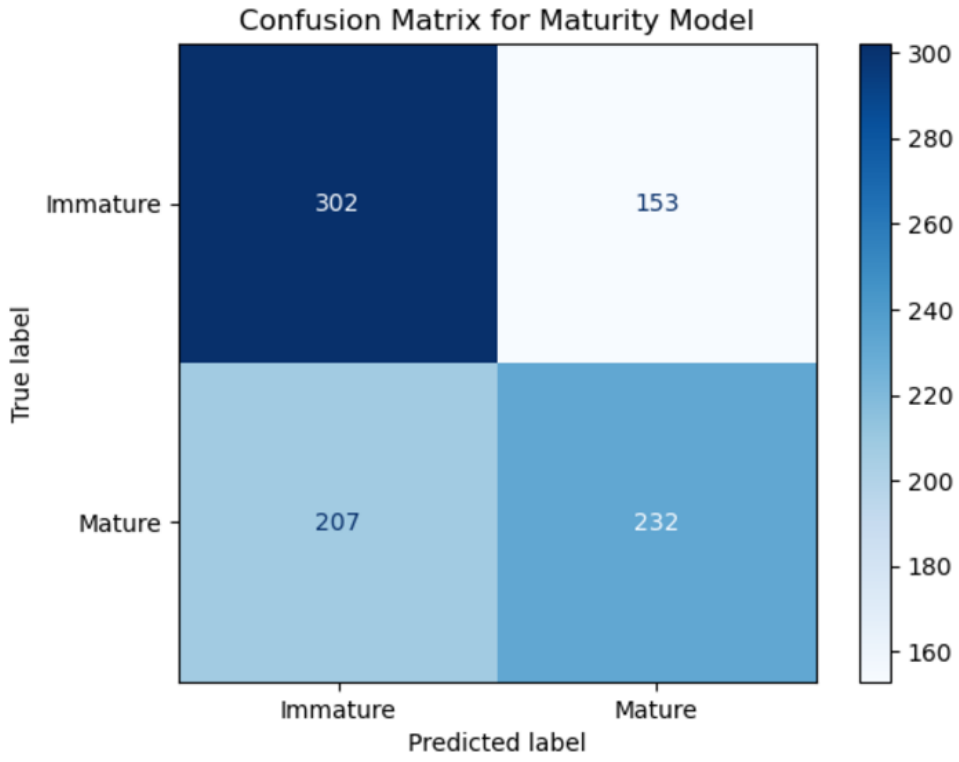
- **Class Distribution**:
    - **Healthy Leaves**: 38.2%
    - **Unhealthy Leaves**: 61.8%

The dataset exhibited a noticeable imbalance, with a higher proportion of unhealthy leaf samples compared to healthy ones. This imbalance posed the risk of model bias towards

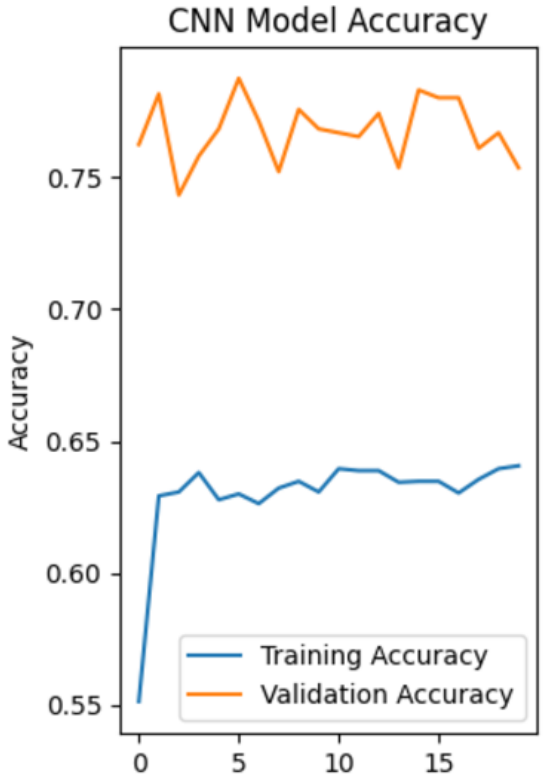
the dominant class. To address this, data augmentation techniques were employed, such as flipping, rotation, and brightness adjustments, to artificially increase the diversity and representation of the underrepresented class.

Objective 4:

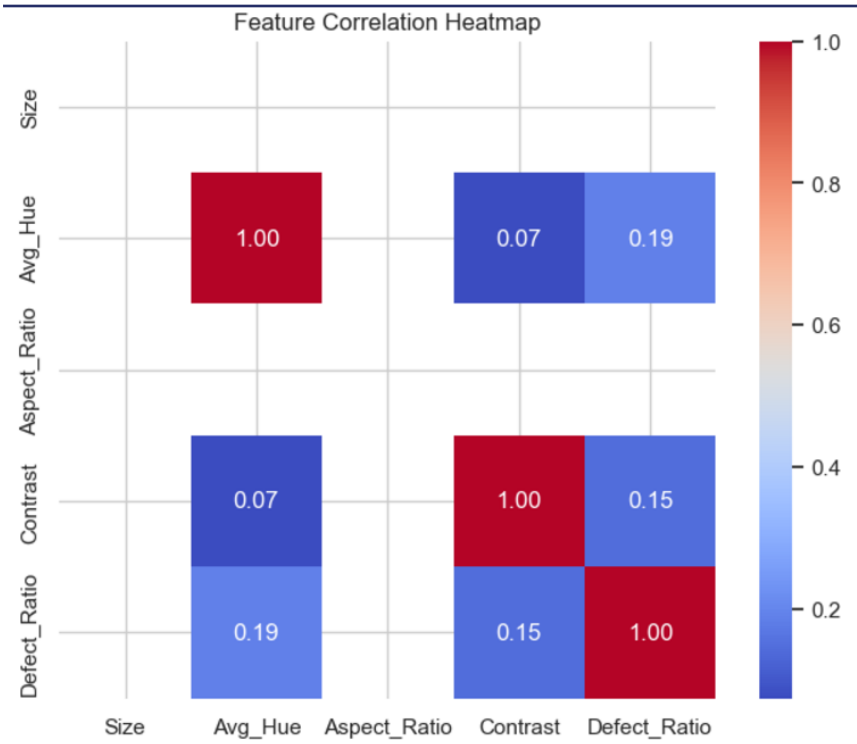
Maturity Classification



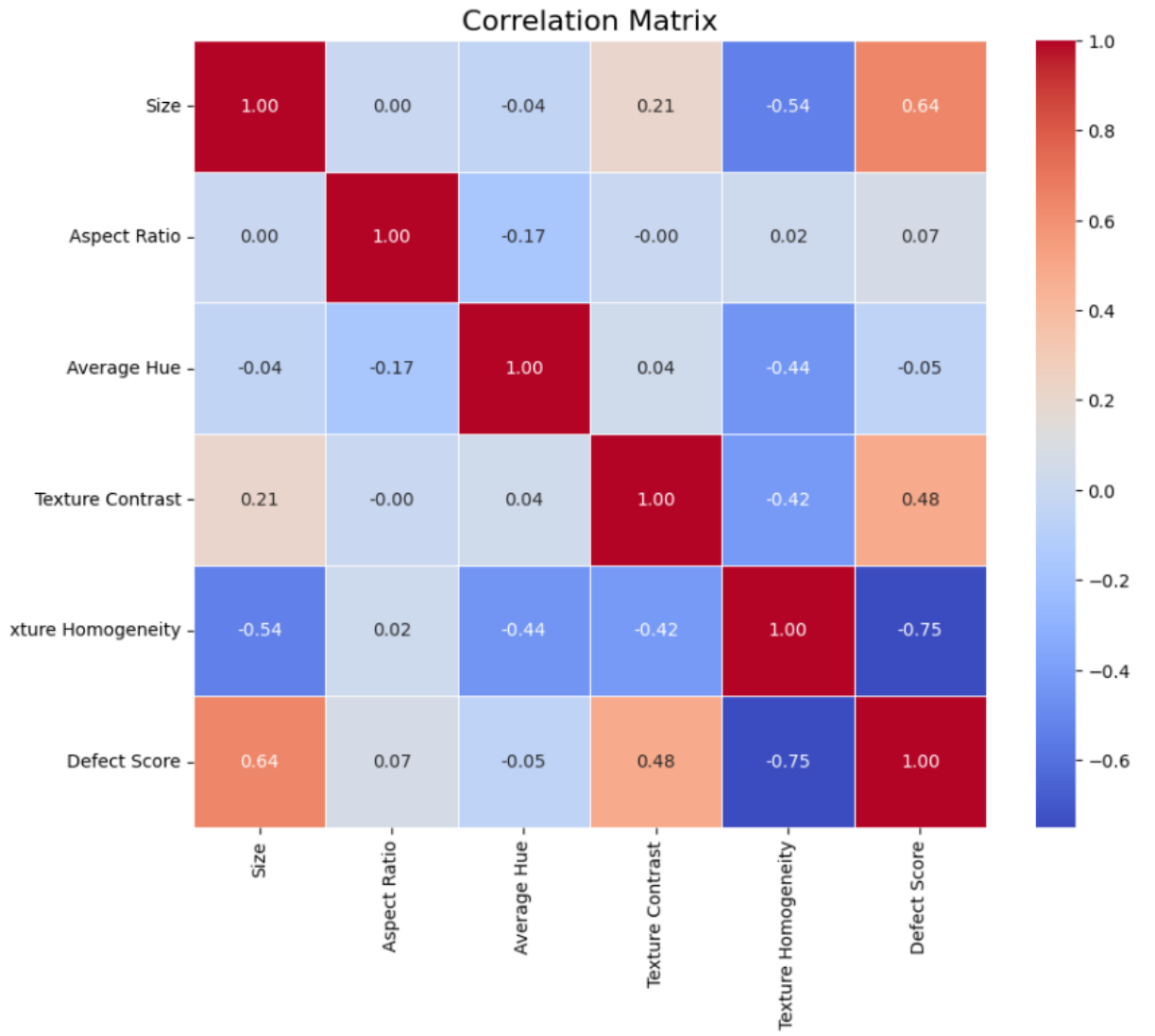
- **True Positives (Mature)**: The model correctly predicted 232 instances as mature.
- **True Negatives (Immature)**: 302 immature instances were correctly classified.
- **False Positives (Immature classified as Mature)**: 153 immature instances were misclassified as mature.
- **False Negatives (Mature classified as Immature)**: 207 mature instances were incorrectly classified as immature.
- **Model Accuracy**: Represents the percentage of correct predictions overall.



Quality Assessment

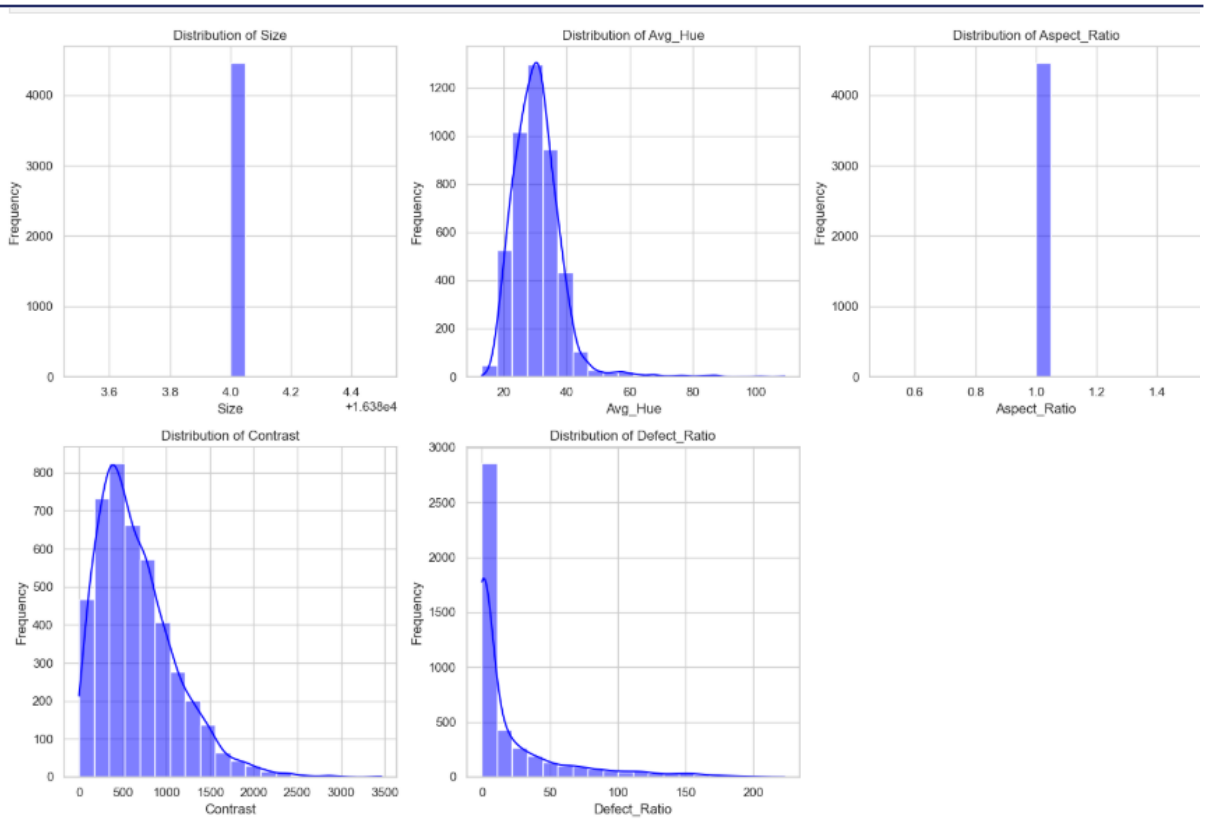


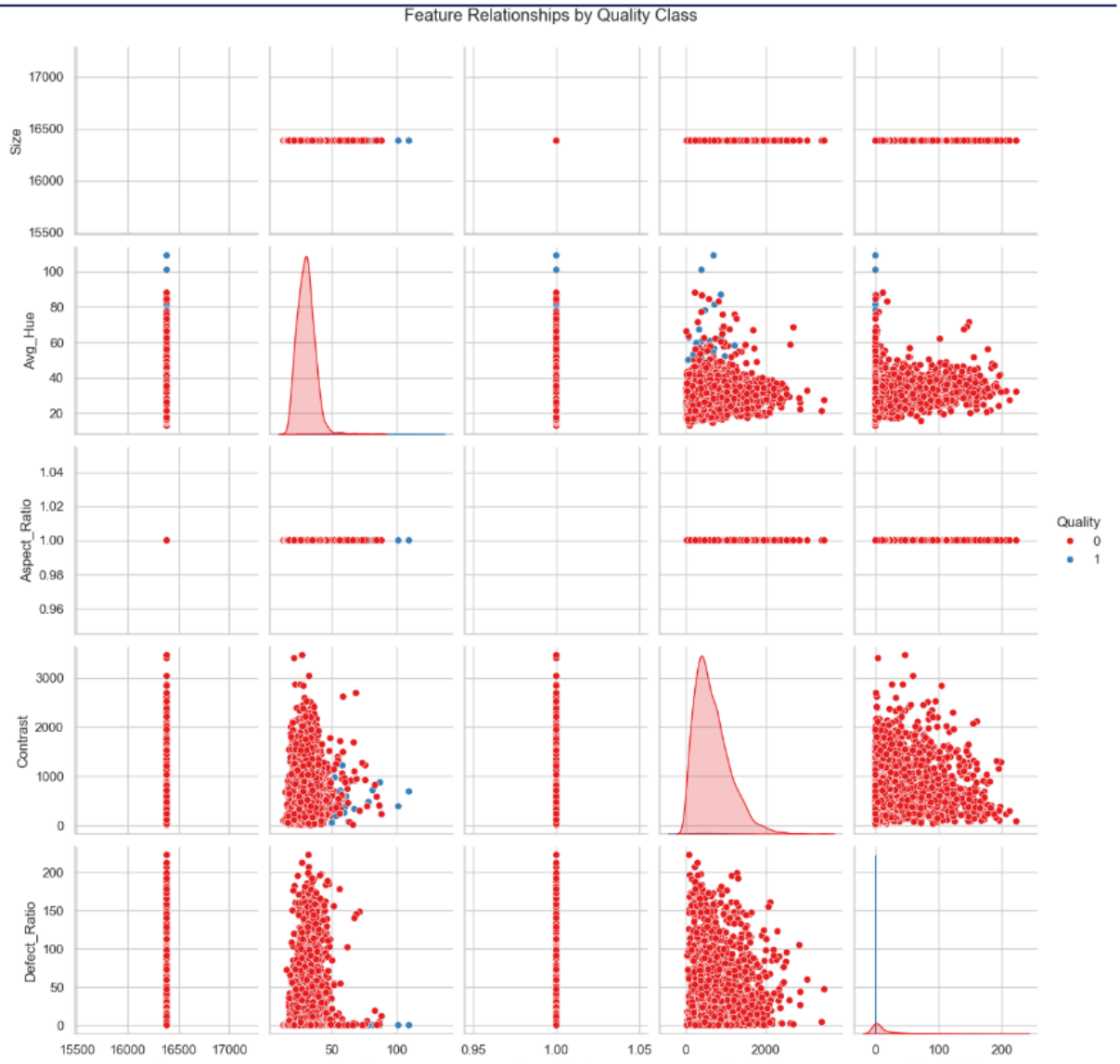
- **Diagonal Values**: All features have a correlation of 1.0 with themselves (red color).
- **Low Correlation Between Features**: Most features (e.g., Avg_Hue, Aspect_Ratio, Contrast) show low correlation with one another, indicated by blue shades (e.g., Avg_Hue and Aspect_Ratio = 0.07).
- **Slight Positive Correlation**: A few features, such as Size and Defect_Ratio (0.19), show weak positive correlation.
- **Interpretation**: Low correlations suggest that the features are mostly independent, which is good for machine learning models as it prevents redundancy and improves prediction.
- **Heatmap Colors**:
  - **Red (1.0)**: Strong correlation.
  - **Blue (near 0)**: Weak or no correlation


Correlation Matrix
Size
Aspect Ratio
Average Hue
Texture Contrast
xture Homogeneity
Defect Score
1.00 0.00 -0.04 0.21 -0.54 0.64
0.00 1.00 -0.17 -0.00 0.02 0.07
-0.04 -0.17 1.00 0.04 -0.44 -0.05
0.21 -0.00 0.04 1.00 -0.42 0.48
-0.54 0.02 -0.44 -0.42 1.00 -0.75
0.64 0.07 -0.05 0.48 -0.75 1.00
Size
Aspect Ratio
Average Hue
Texture Contrast
Texture Homogeneity
Defect Score
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

BSc (Hons) in Information Technology Specializing Data Science

Research Project - IT4010 Data Analysis Report





Objective 1:

4.1 Challenges Faced During Data Analysis:

Real-time data collection posed significant challenges, including interruptions caused by external factors like a fungal outbreak in the greenhouse. These issues affected the consistency of sensor readings and image data, requiring additional cleaning and preprocessing to ensure data quality and model reliability.

Objective 2:

Irrigation Optimization Dataset

- **Handling Missing Values:**

Sporadic gaps in sensor data were observed due to occasional device malfunctions.

Solution: Imputed missing values using mean for numerical data and mode for categorical data, ensuring data continuity without introducing significant bias.

- **Multicollinearity Issues:**

High correlation between temperature and humidity metrics caused redundancy.

Solution: Introduced the derived **Climate Index** to mitigate collinearity and capture combined environmental effects.

Wilted Leaf Detection Dataset

- **Limited Image Diversity:**

Initial datasets lacked sufficient variability in lighting and background conditions.

Solution: Augmented images through flipping, rotation, brightness adjustment, and scaling to enhance diversity and model generalizability.

- **Segmentation Challenges:**

Overlapping leaves and complex backgrounds reduced segmentation accuracy.

Solution: Applied advanced segmentation techniques (e.g., UNet masking and K-Means clustering) to improve feature clarity and focus on individual leaf attributes.

- **Computational Constraints:**

High-resolution images required significant memory and processing power during model training.

Solution: Downscaled images to 224x224 pixels, balancing computational efficiency with feature retention

Objective 3:

1. **Image Quality Variations**:

- Variability in lighting, focus, and angle across images affected the consistency of features extracted.
- Filtering low-quality images using green pixel ratio analysis and applying preprocessing steps like normalization and augmentation were critical to improve image quality.

2. **Difficulty in Segmenting Data**:

- Isolating the leaf from the background using segmentation techniques was challenging, especially for images with complex or cluttered backgrounds.

- Accurate segmentation required advanced techniques like background masking and boundary highlighting to focus on relevant leaf features.

3. **Integrating Time-Series and Image Data**:

- Combining temporal features (growth stage and duration) with image-based features for the unified model was technically demanding.
- Developing a dual-branch model architecture that effectively processed and fused these disparate data types required careful experimentation and tuning.

4. **Computational Challenges**:

- Training deep learning models like ResNet50 and MobileNetV2 on high-resolution images was resource intensive.
- To mitigate this, images were resized to 128x128 pixels, and pre-trained models were used to reduce computational costs.

Objective 4:

1. Outliers:

- Extreme values in features like fruit weight and age were identified during exploratory data analysis. These outliers distorted data distributions and influenced model performance.
- To address this, outliers were identified using visualization techniques such as boxplots and statistical methods like the interquartile range (IQR).
- Outliers were either capped to acceptable thresholds or removed entirely to ensure model stability.

2. Variability in Feature Scales:

- Features like weight, bounding box area, and defect score operated on different scales, which negatively impacted model convergence during training.
- Normalization techniques, such as StandardScaler, were applied to standardize numerical features, ensuring all inputs contributed equally to model training.

3. Imbalanced Classes:

- The dataset contained an imbalance between mature and immature fruit labels, with significantly fewer samples of mature fruits.
- This imbalance hindered the classification model's ability to correctly predict mature fruits.
- The issue was addressed by employing oversampling techniques, such as SMOTE (Synthetic Minority Oversampling Technique), to generate synthetic samples for the minority class.

4. Feature Selection Challenges:

- Identifying the most relevant features for maturity and harvest date prediction was challenging due to high feature redundancy and correlation.
- Recursive feature elimination and correlation analysis were used to narrow down the feature set, prioritizing those with the strongest relationship to the target variables.

5. Computational Constraints:

- Training deep learning models, such as CNNs, on large datasets was computationally intensive, requiring significant time and resources.
- Optimization was achieved by using smaller batch sizes, transferring learning from pre-trained models, and leveraging less computationally demanding algorithms for initial testing.

6. Image Variability:

- Differences in lighting, angle, and focus across cucumber fruit images caused inconsistency in extracted features.
- To standardize inputs, preprocessing techniques such as image augmentation, normalization, and resizing were implemented, improving model generalization.

4. References

5.1 Datasets References:

[1] S. J. Smith and L. K. White, "Cucumber Growth Dataset," [Online]. Available: https://data.mendeley.com/datasets/y6d3z6f8z9/1. [Accessed: Dec. 05, 2024].

[2] "Cucumber Leaf Disease Dataset," *Kaggle*, 2021. [Online]. Available: https://www.kaggle.com/discussions/general/204901. [Accessed: 08-Dec-2024].

[3] Dataset for On-Demand Irrigation System Based on Biomass Weight,

https://data.mendeley.com/datasets/y6d3z6f8z9/1

5.2 References:

[1] M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Integrated IoT Approaches for Crop Recommendation and Yield-Prediction Using Machine-Learning," *IoT*, vol. 5, no. 4, pp. 634–649, Sep. 2024. [Online]. Available: https://www.mdpi.com/2624-831X/5/4/28

[2] U. Lee et al., "An Automated, Clip-Type, Small Internet of Things Camera-Based Tomato Flower and Fruit Monitoring and Harvest Prediction System," *Sensors*, vol. 22, no. 7, p. 2456, Mar. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2456

[3] M. K. H. Siam, N. Tasnia, S. Mahmud, M. Halder, and M. M. Rana, "A Next-Generation Device for Crop Yield Prediction Using IoT and Machine Learning," in *Intelligent Systems and Networks*, vol. 752, Singapore: Springer, 2023, pp. 668–678. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-99-4725-6_78

[4] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 718–752, Apr. 2021. [Online]. Available: https://www.ieee-jas.net/en/article/doi/10.1109/JAS.2021.1003925

[5] M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 718–752, Apr. 2021. [Online]. Available: https://www.ieee-jas.net/en/article/doi/10.1109/JAS.2021.1003925

[6] M. Bouni, B. Hssina, K. Douzi, and S. Douzi, "Integrated IoT Approaches for Crop Recommendation and Yield-Prediction Using Machine-Learning," *IoT*, vol. 5, no. 4, pp. 634–649, Sep. 2024. [Online]. Available: https://www.mdpi.com/2624-831X/5/4/28

[7] U. Lee et al., "An Automated, Clip-Type, Small Internet of Things Camera-Based Tomato Flower and Fruit Monitoring and Harvest Prediction System," *Sensors*, vol. 22, no. 7, p. 2456, Mar. 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/7/2456

[8] M. K. H. Siam, N. Tasnia, S. Mahmud, M. Halder, and M. M. Rana, "A Next-Generation Device for Crop Yield Prediction Using IoT and Machine Learning," in *Intelligent Systems and Networks*, vol. 752, Singapore: Springer, 2023, pp. 668–678. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-99-4725-6_78

[9] O. Friha, M. A. Ferrag, L. Shu, L. Maglaras, and X. Wang, "Internet of Things for the Future of Smart Agriculture: A Comprehensive Survey of Emerging Technologies," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 4, pp. 718–752, Apr. 2021. [Online]. Available: https://www.ieee-jas.net/en/article/doi/10.1109/JAS.2021.1003925