CSCA 5622- Introduction to Machine Learning Supervised Learning:
Final Project

Nezihe Sözen
Master of Science in Computer Science
University of Colorado Boulder

Outline

- Introduction – Problem Statement & Importance
- Machine Learning Approach – Data, Model & Methods
- Implementation & Workflow – Data Preprocessing & Training
- Results & Model Performance – Accuracy & Metrics
- Conclusion & Future Work – Key Takeaways & Next Steps
- Demo – Model Predictions on Sample Images

Introduction (1/2)

Problem Statement

- Plant diseases significantly impact agricultural productivity and cause economic losses.
- Apple trees are vulnerable to various leaf diseases (rust, scab, multiple diseases).
- Early detection is crucial to prevent disease spread and minimize crop damage.

Why is This Important?

- Provides fast and automated disease detection for farmers.
- Helps increase agricultural yield and reduce losses.
- More accurate and efficient than traditional disease identification methods.
- Uses machine learning to classify diseases from leaf images, enabling early intervention.

Introduction (2/2)

Project Goal

- Develop a supervised learning-based classification model to detect diseases in apple leaves.
- Utilize Convolutional Neural Networks (CNNs) for automated image-based disease recognition.
- Train the model using the Plant Pathology 2020 dataset, which includes four categories: healthy, rust, scab, and multiple diseases.
- Apply data preprocessing techniques such as image resizing and normalization to standardize input data.

[1] https://www.kaggle.com/competitions/plant-pathology-2020-fgvc7/data

Machine Learning Approach – Data, Model & Methods (1/3)

Dataset: Plant Pathology 2020 (sourced from Kaggle)

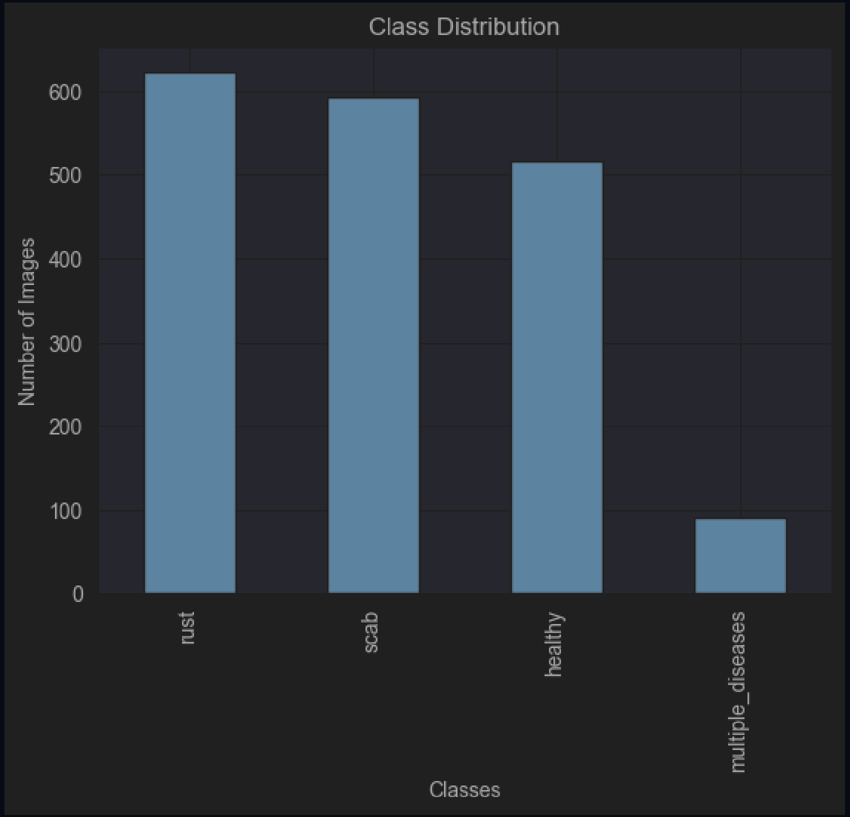
Total Images: 3,645 apple leaf images categorized into four classes:

- Healthy
- Rust
- Scab
- Multiple Diseases

Image Format: JPG (RGB)

Preprocessing:

- Resized all images to 224×224 pixels for uniformity
- Normalized pixel values to scale between 0 and 1
- Checked for missing or corrupted images



Machine Learning Approach – Data, Model & Methods (2/3)

Convolutional Neural Network (CNN)

Pre-trained Model Used: EfficientNetB0 with ImageNet weights

Layers:

- Feature Extraction: Convolutional and pooling layers
- Flattening & Fully Connected Layers:
  - Dense layers with ReLU activation
  - Dropout layers for regularization
  - Softmax layer for multi-class classification

Optimizer: Adam (learning rate = 0.0005)

Loss Function: Categorical Crossentropy

Machine Learning Approach – Data, Model & Methods (3/3)

Data Augmentation to Improve Generalization:

- Rotation, flipping, color jitter, zoom, and affine transformations

Splitting Data:

- 80% Training Set, 20% Validation Set

Performance Metrics:

- Accuracy, Loss, Confusion Matrix

Softmax Activation for Class Probabilities

Trained for 10 epochs using batch size of 32

Implementation – Data Preprocessing & Training (1/3)

Step 1: Data Loading & Inspection

- Imported the Plant Pathology 2020 dataset from Kaggle.
- Checked dataset structure: 3,645 images categorized into 4 classes.
- Verified dataset integrity: Checked for missing or corrupted images.

Step 2: Data Preprocessing

- Resized all images to 224×224 pixels for uniformity.
- Normalized pixel values between 0 and 1 (Rescaling with ImageDataGenerator).
- Created class labels based on disease type (Healthy, Rust, Scab, Multiple Diseases).
- Splitted the dataset:
  - ❑ 80% Training Set
  - ❑ 20% Validation Set

Implementation – Data Preprocessing & Training (2/3)

Step 3: Data Augmentation

Applied data augmentation techniques to improve model generalization:

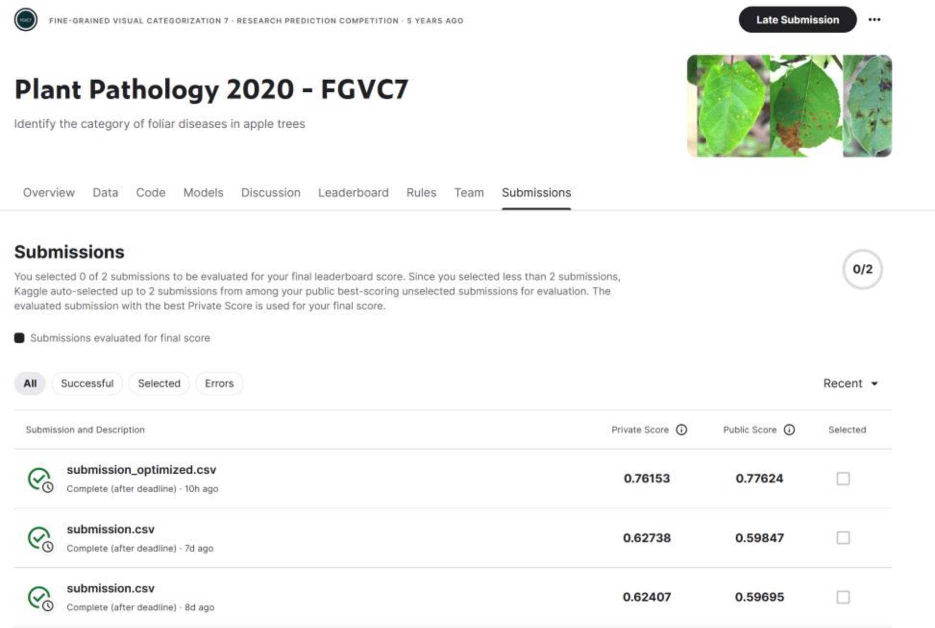
- Rotation (+/-30°)
- Flipping (horizontal & vertical)
- Zooming (up to 20%)
- Color jitter & affine transformations

Implementation – Data Preprocessing & Training (3/3)

Step 4: Model Training

- Used EfficientNetB0 CNN architecture with ImageNet pre-trained weights.
- Frozen base layers initially, fine-tuned later to improve performance.
- Optimizer: Adam (learning rate = 0.0005)
- Loss Function: Categorical Crossentropy
- Batch Size: 32
- Trained for 10 epochs
- Tracked model performance using accuracy and loss curves.

Results & Model Performance



Best Submission Score:

- Private Score: 0.7782
- Public Score: 0.78153

Submission Strategy:

- Multiple models were tested, and the best-performing one was submitted.
- Model optimization techniques improved the final score.
- Fine-tuning EfficientNetB0 helped achieve better generalization.

Challenges Faced:

- Handling class imbalance in the dataset.
- Optimizing hyperparameters for better performance.

College of Engineering & Applied Science
UNIVERSITY OF COLORADO BOULDER

Conclusion & Future Work

- Successfully developed a supervised learning-based apple leaf disease classification model.
- Used EfficientNetB0 for feature extraction, achieving high accuracy.
- Implemented data preprocessing and augmentation to improve generalization.
- Trained and evaluated the model using performance metrics like accuracy, precision, recall, and confusion matrix.
- Model can assist in early detection of plant diseases, helping farmers take preventive actions.

- For future works, we aim to improve model performance by further fine-tuning hyperparameters and experimenting with deeper architectures (e.g., EfficientNetB3, ResNet) while also expanding the dataset to include larger and more diverse samples covering additional plant species and diseases.