Machine Learning Based Leaf Disease Detection

Submitted by: Galimuthy Elia Anusha, Keerthana H, Muhammad Uwais, Tiya Rose, Vaidhav Adit

Problem Statement & Relevance

Agriculture is vital to India'seconomy, employingover 45% oftheworkforce. However, climate change, irregular rainfall, and newplant diseasespose significant challenges, especially for small farmers lacking access to expert diagnostic services.

Late disease detection leads to crop failure and financial losses. There's a critical need for accessible tools to identify plant diseases early. Machine learning (ML) offers a practical solution by analyzing leaf images for instant feedback on crop health, enabling timely decisions and reducing reliance on external services.

Project Objectives & Expected Outcomes

Train ML Model

Develop amodelto predict plant type and disease status (healthy or specific disease) from a single leaf image.

Efficient & Reliable

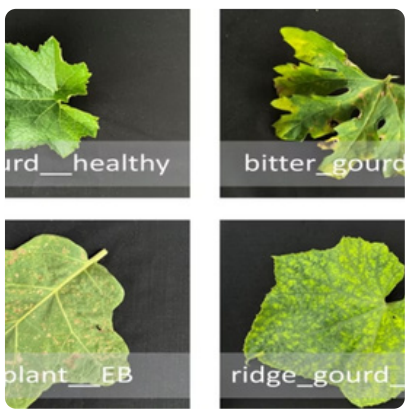
Build amodelthatis efficient, consistent, and reliable for early disease detection by farmers.

Transparent AI

Ensure model transparency by explaining its decision-making process, key features, training, and predictions.

Data Collection: Diverse Sources for Robustness

Toaddress dataimbalanceandensurecomprehensivecoverage,wecombineddatafrommultiplesources,withtheOLID-Idataset forming the coreofour pipeline.



OLID I (Open Leaf Image Dataset)

Collected under natural field conditions in Bangladesh, this dataset includes leaf images of various crops and stress conditions (biotic and abiotic). Images are annotated with primary (crop type) and secondary (stress/disease) class labels, totalling 4,749 images.

Proposed Layered BG Dataset

From Roboflow Universe, this dataset contains 2,430 augmented Bitter Gourd images across nine leaf deficiency classes, supporting deep learning tasks.

Cucumber Leaf Disease Dataset

From Kaggle, this dataset features 3,754 images with three classes: Healthy, Powdery Mildew, and Downy Mildew, all with white backgrounds and augmented.

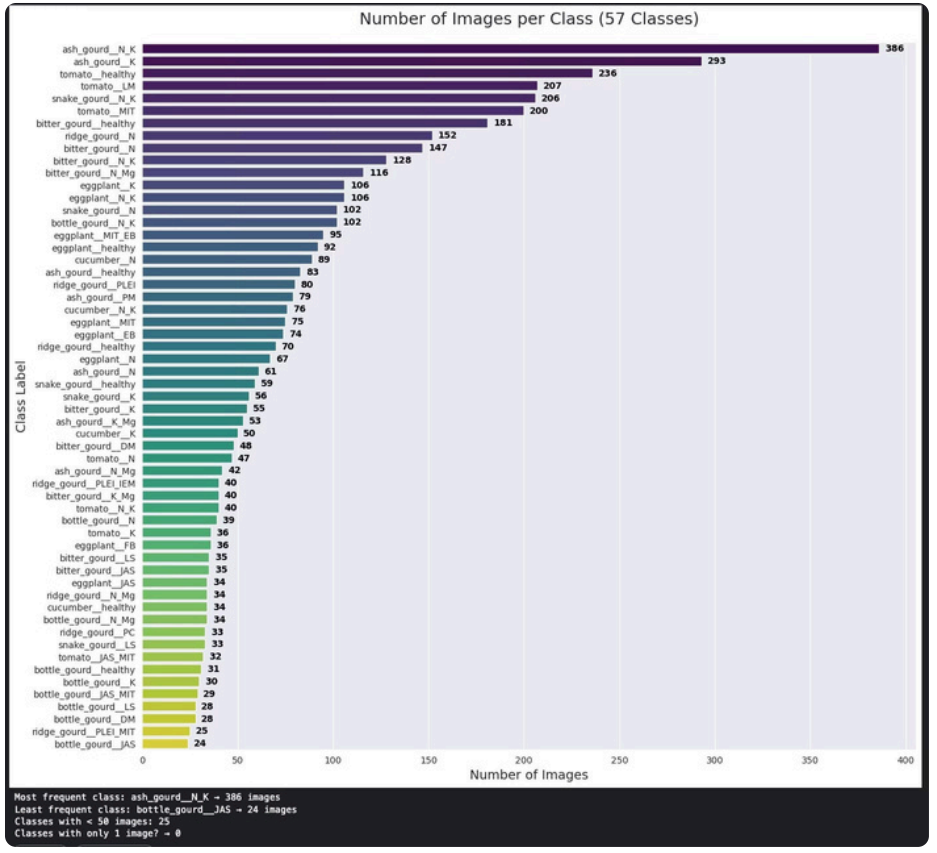
Bottle Gourd Dataset

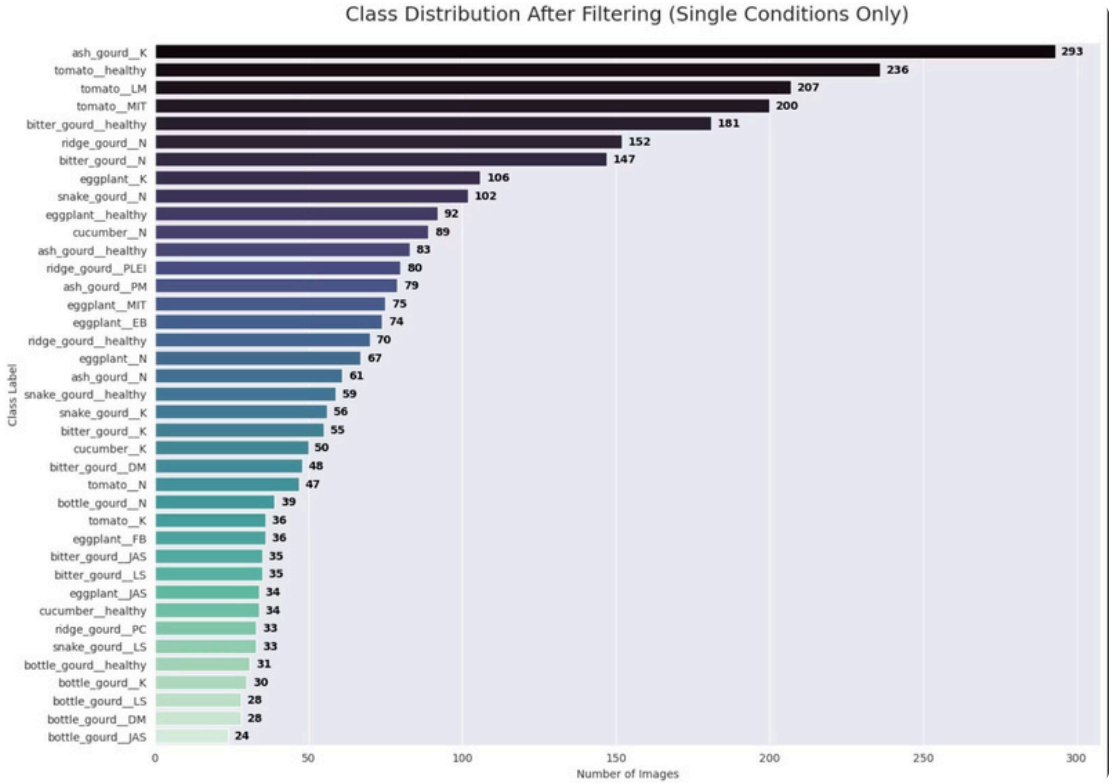
Also from Kaggle, this dataset has 1,819 images across three classes: Healthy, Downy Mildew, and Anthracnose, augmented with white backgrounds.

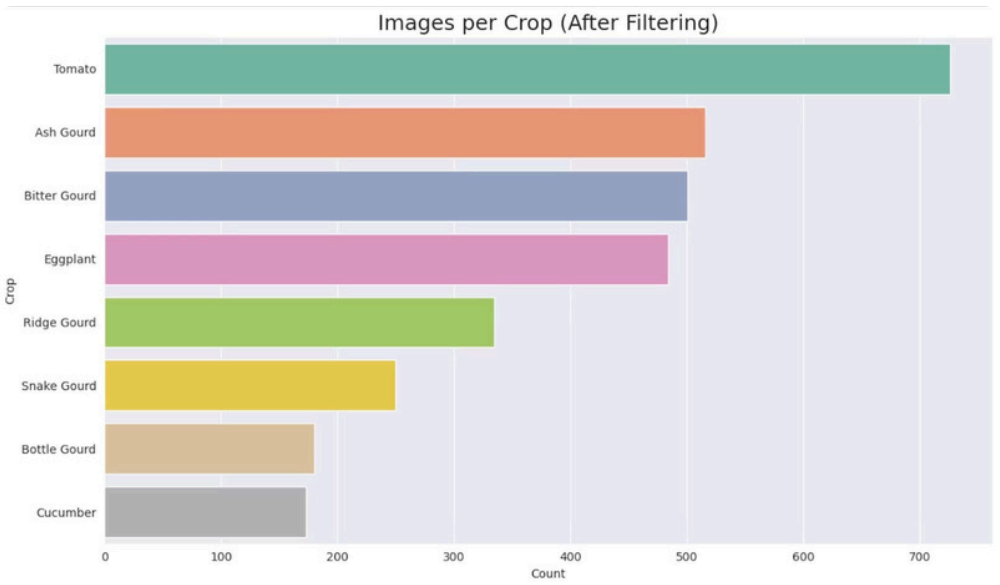
Tomato Dataset

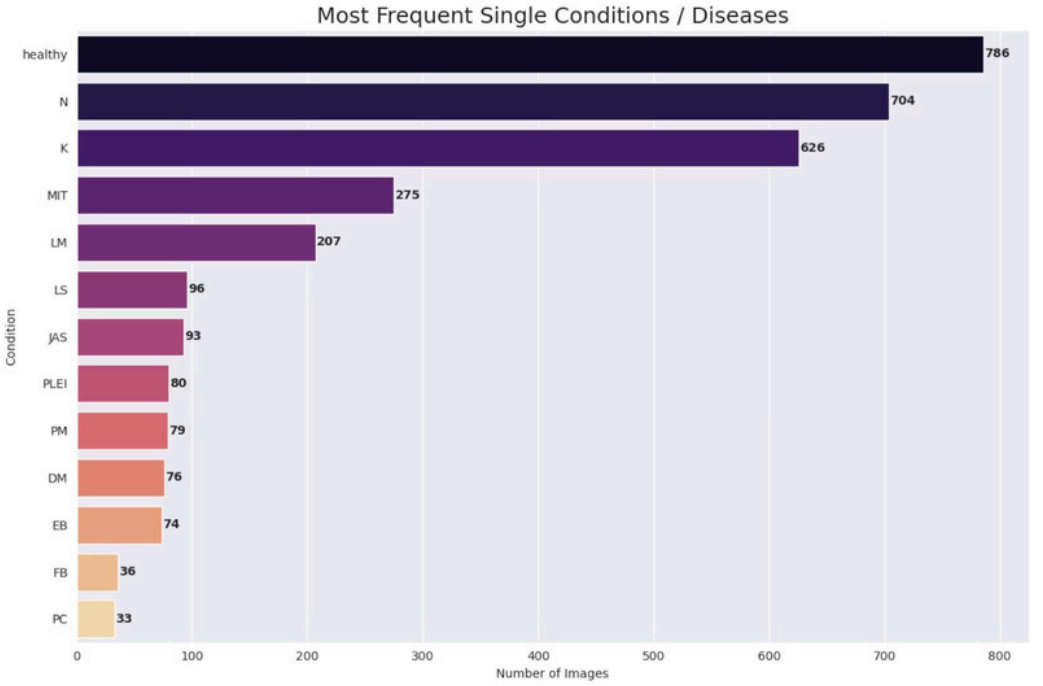
This Kaggle dataset includes 2,610 images across four classes: bacterial spot, fresh leaf, fresh leaf virus, and spotted wilt, all augmented with white backgrounds.

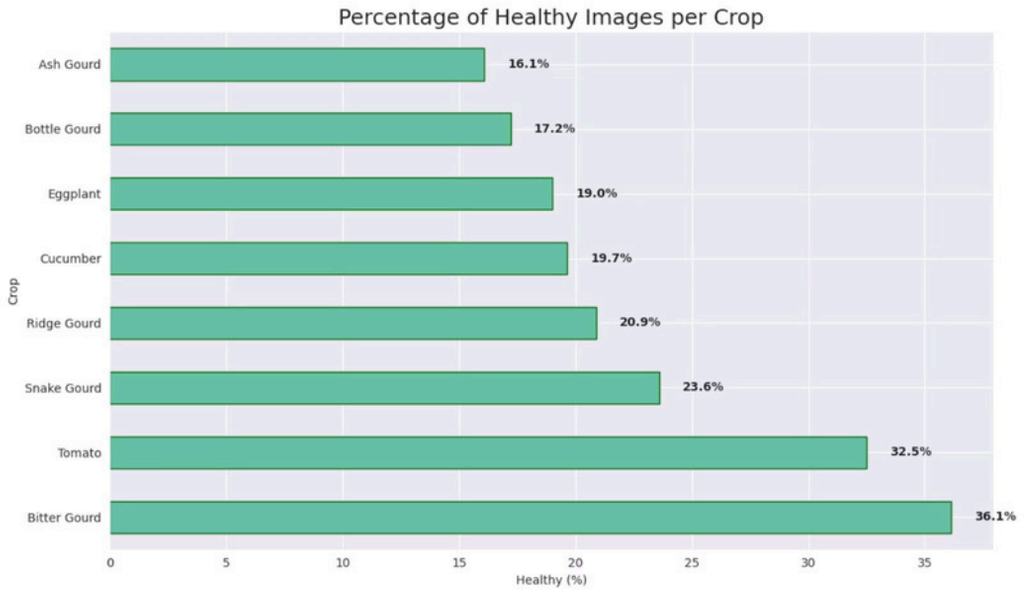
Exploratory Data Analysis: Unveiling Data Patterns

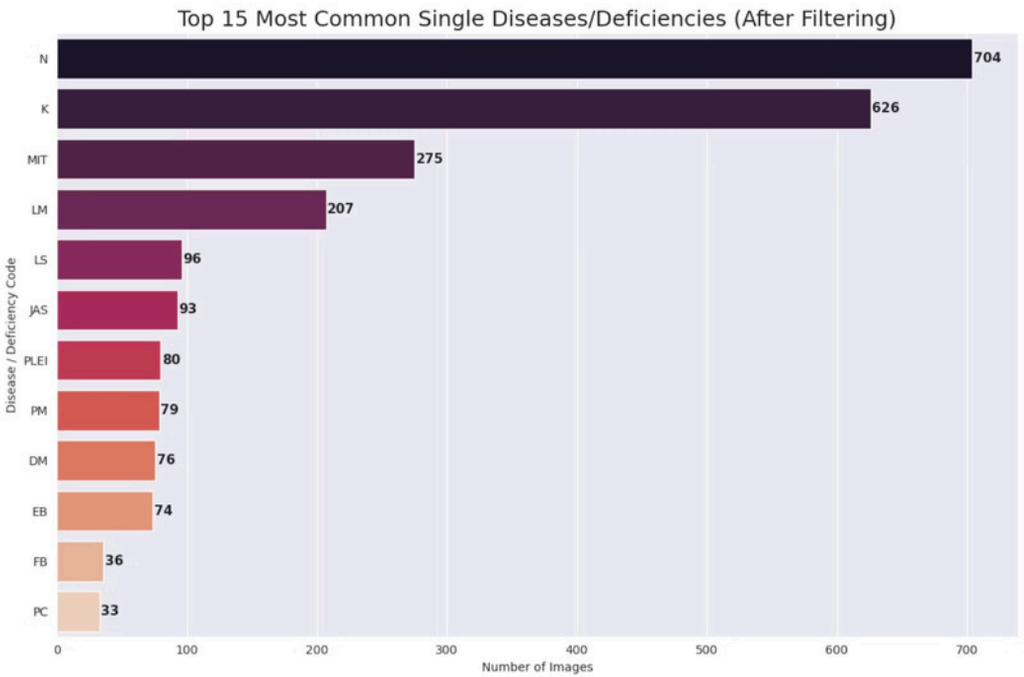


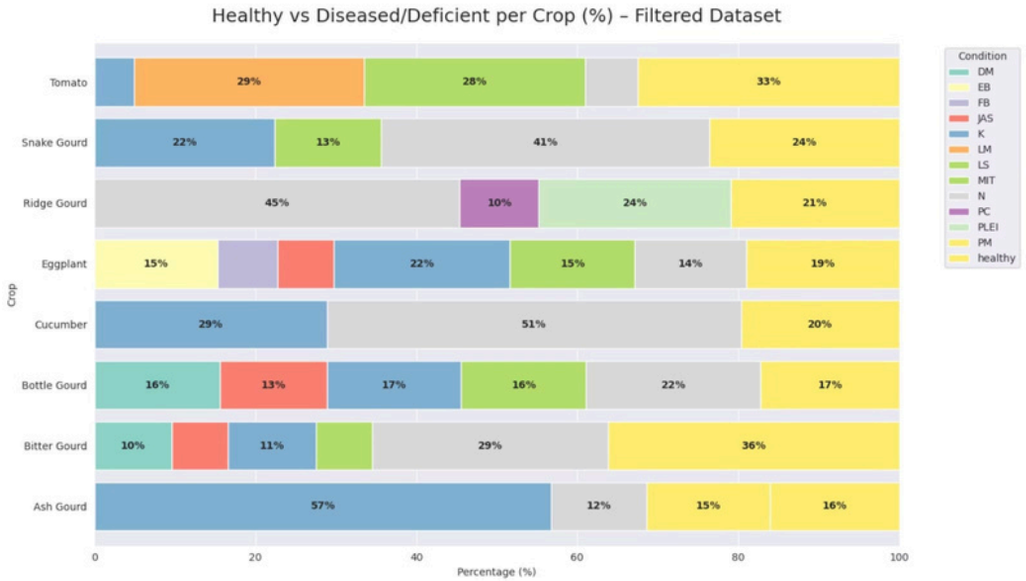



Images per Crop (After Filtering)
Tomato
Ash Gourd
Bitter Gourd
Eggplant
Ridge Gourd
Snake Gourd
Bottle Gourd
Cucumber
Crop
0
100
200
300
400
500
600
700
Count


Most Frequent Single Conditions / Diseases
healthy
N
K
MIT
LM
LS
JAS
PLEI
PM
DM
EB
FB
PC
786
704
626
275
207
96
93
80
79
76
74
36
33
Condition
0
100
200
300
400
500
600
700
800
Number of Images







Raw counts:

condition	DM	EB	FB	JAS	K	LM	LS	MIT	N	PC	PLEI	PM	healthy
crop													
Ash Gourd	0	0	0	0	293	0	0	0	61	0	0	79	83
Bitter Gourd	48	0	0	35	55	0	35	0	147	0	0	0	181
Bottle Gourd	28	0	0	24	30	0	28	0	39	0	0	0	31
Cucumber	0	0	0	0	50	0	0	0	89	0	0	0	34
Eggplant	0	74	36	34	106	0	0	75	67	0	0	0	92
Ridge Gourd	0	0	0	0	0	0	0	0	152	33	80	0	70
Snake Gourd	0	0	0	0	56	0	33	0	102	0	0	0	59
Tomato	0	0	0	0	36	207	0	200	47	0	0	0	236

```
============================================================
        IMBALANCE SUMMARY — FILTERED DATASET
============================================================
Total images              : 3,165
Total classes             : 39
Avg images per class      : 81.2
Top 10 classes contain    : 1,716 images → 54.2% of data
Bottom 20 classes contain: 771 images → 24.36% of data
Classes with ≤30 images   : 4 (10.3% of classes)
Classes with ≤20 images   : 0 (very rare!)
Imbalance ratio (top10/bottom20): 2.2x
============================================================

Rarest classes (≤25 images):
  • Bottle Gourd – JAS: 24 images
```

Image Dimension Statistics (sample):

	Height	Width
count	1000.000000	1000.000000
mean	3023.645000	3023.645000
std	7.935613	7.935613
min	2843.000000	2843.000000
25%	3024.000000	3024.000000
50%	3024.000000	3024.000000
75%	3024.000000	3024.000000
max	3024.000000	3024.000000

3.5 Feature Engineering: Transforming Images into Data

1

Leaf Segmentation

Images were resized to 256x256pixels and converted to HSV color space. Color thresholds were applied to create a binary mask, isolating leaf tissue from the background. Morphological operations cleaned the mask, resulting in segmented color and grayscale leaf images. This removes clutter, ensuring features represent only the leaf.

2

Color Feature Extraction (HSV Histogram)

A3D HSV histogram (8 Hue,12 Saturation,3 Value bins) generated 288 color features per image. These quantify color distribution, crucial for detecting symptoms like yellowing or browning in nutrient deficiencies and fungal infections.

3

Texture Feature Extraction (GLCM + LBP)

GLCM (Gray LevelCo-occurrence Matrix) extracted20features by analyzing pixel intensity relationships at four orientations. LBP (Local Binary Patterns) generated 26 features, capturing minute texture characteristics like lesions and roughness. Together, these provided 46 texture features.

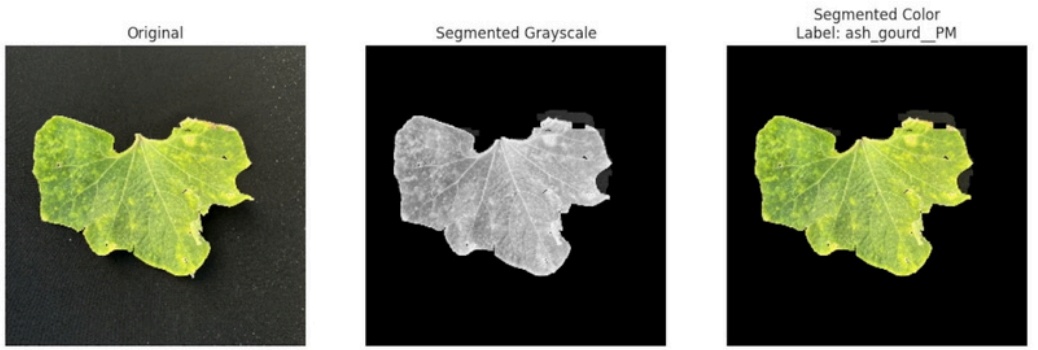
4

Shape Feature Extraction (HOG)

Histogram ofOriented Gradients(HOG)extracted 1,764 shape features, capturing edge structures, contours, and gradient variations indicative of disease symptoms like curling or boundary roughening.

The final feature vector combined 288 color, 46 texture, 1,764 shape features, and 1 encoded plant type value, totaling 2,099 features per image. This comprehensive vector formed the input for our machine learning models.

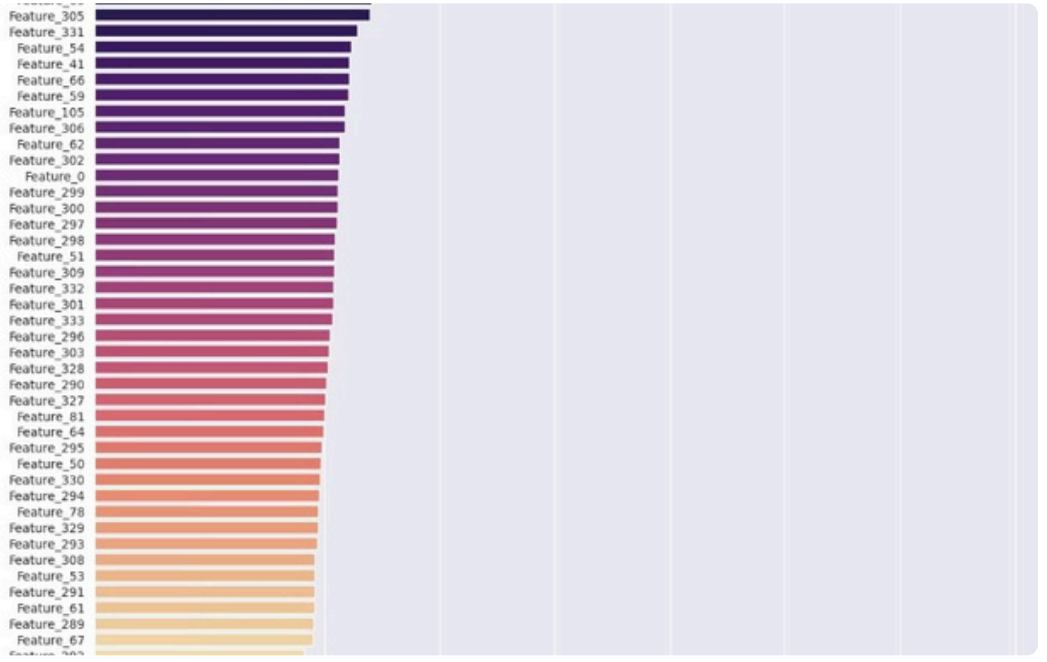
```
--- Dataset Build Complete ---
X (feature matrix) shape: (3165, 2099)
y (target vector) shape: (3165,)

Number of features per image: 2098
Number of 'plant_type' features: 1
Total features: 2099
```

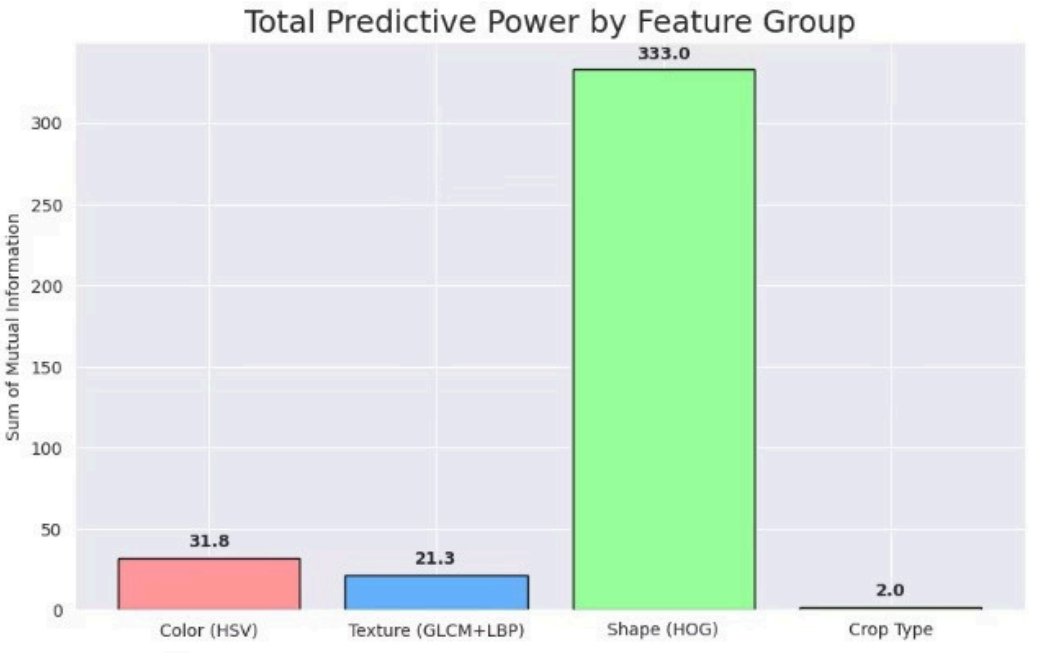


Final Dataset Shape

Segmentation

Feature Analysis: Understanding Predictive Power



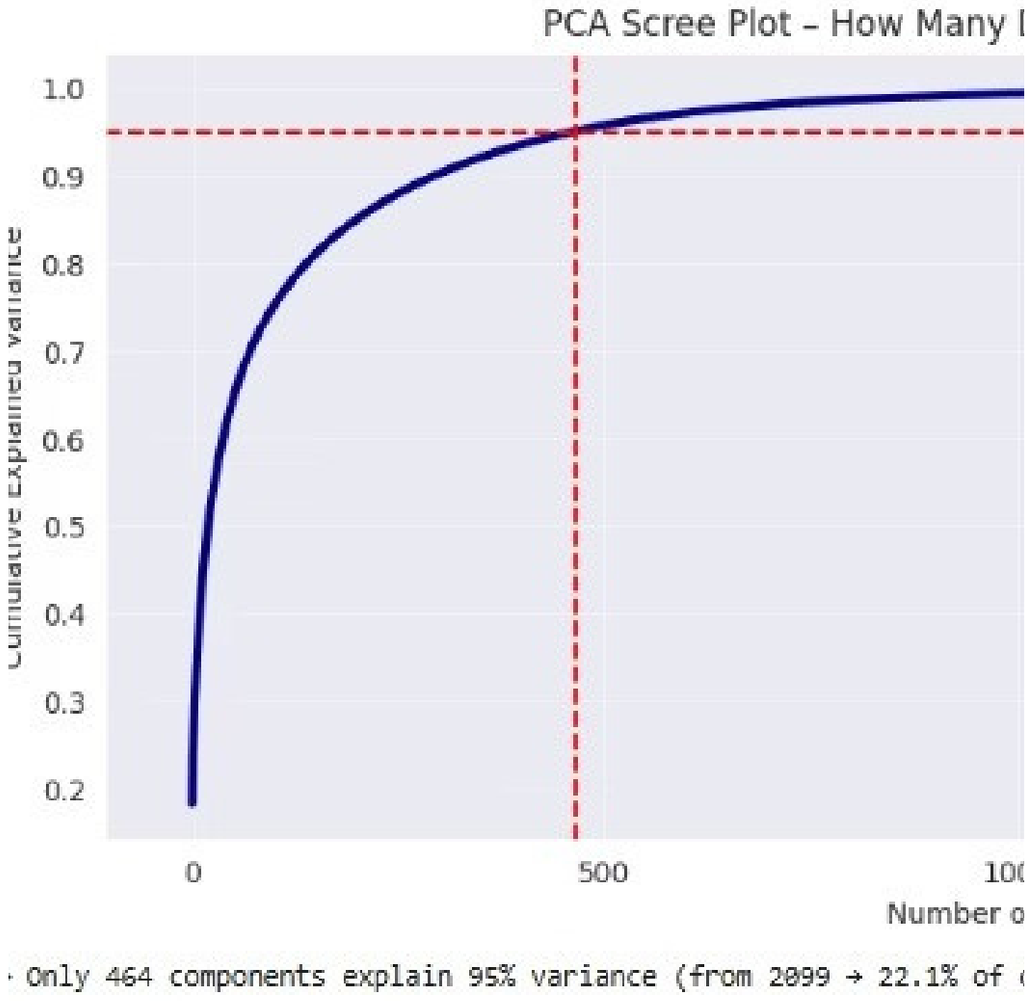


Mutual Information Ranking

Mutual Information (MI) quantified thedependency between features and class labels. The plant type encoding had the highest MI, followed by HOG, GLCM, and LBP features, confirming the academic relevance of our feature engineering. Over 1,980 features showed MI values greater than 0.01, indicating meaningful signals.

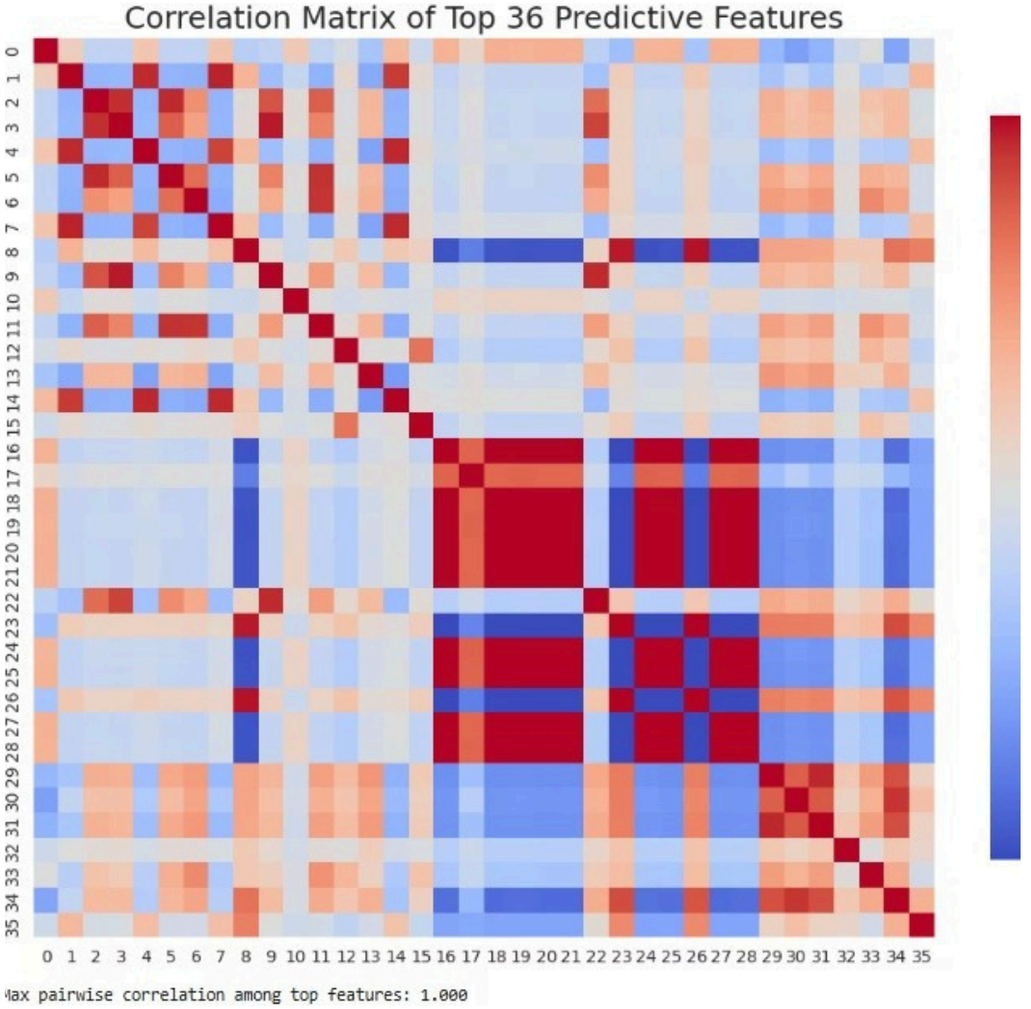
Predictive Power by Group

Shape features (HOG) contributed over85% of the total predictive power (333.0 MI units), significantly outweighing color (31.8) and texture (21.3) features. This highlights the critical role of structural patterns in disease identification.



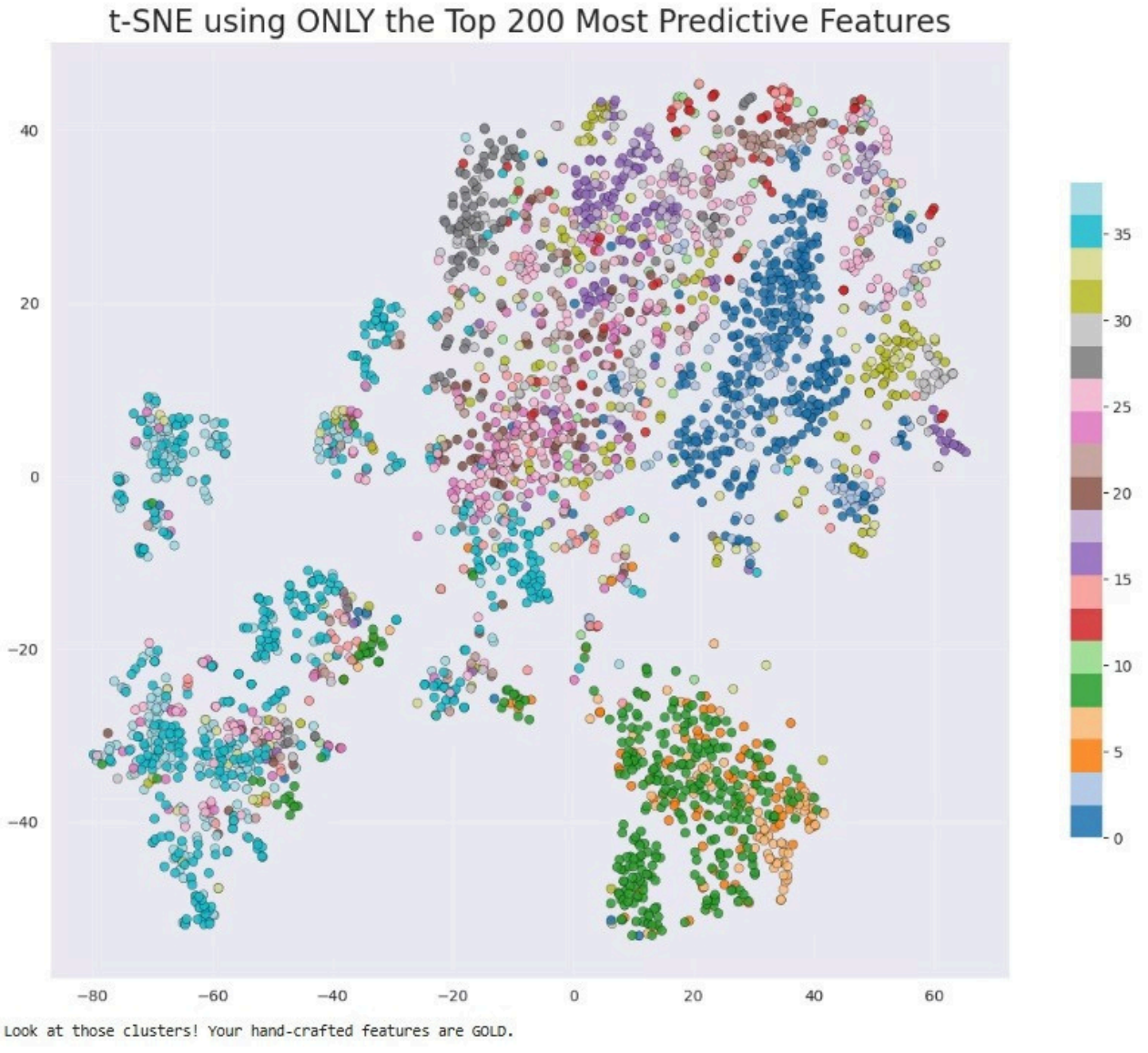
PCA for Dimensionality Reduction

PrincipalComponent Analysis (PCA) showed that only464 principal components explained 95% of the total variance, reducing the original 2,099 features by 78% while retaining nearly all meaningful information.



Correlation Analysis

Acorrelation matrix of thetop36 features revealed high correlation among many HOG features, indicating redundancy. This validated the need for dimensionality reduction to address multicollinearity.



t-SNE Visualization

t-SNEvisualizationof the top 200 predictive features revealed clear clustering for many classes, especially diseases with strong structural signatures. This confirmed that our handcrafted features captured disease specific biological patterns and that the dataset is inherently separable in high dimensional space.

Classical ML Models 4 Without PCA

KNN

- 78.9% plant accuracy
- Weak Status (macroF1=0.32)
- Struggles with overlap+imbalance

Logistic Regression

- Strong plant accuracy
- Status weak (macro F1 j 0.42)
- Linear model [3] limited separability

Decision Tree (Weighted)

- Overfitting visible
- Status macro F1 j 0.447
- Sensitive to noise

Random Forest (Weighted)

- More stable than single trees
- Status macro F1 j 0.476
- Still biased to frequent classes

SVM

- Good plant classification
- Weak on Status
- High training cost

AdaBoost (Best Model)

- Bestmacro F1for Status
- Handles imbalance better
- SHAP: few strong features dominate

Key Issue: Class imbalance + overlapping deficiency symptoms limited all models.

Classical ML Models 4 With PCA (95% Variance)

PCA Summary

- Reduced 2099 ^{3} 464 features
- Retained 95% variance
- Lower noise + faster training

KNN + PCA

- Slightly improved stability
- Minor Status gains
- Better generalization

Logistic Regression + PCA

- Faster training
- Slight performance improvement
- Cleaner decision boundaries

Decision Tree + PCA

- Less overfitting
- No major Status gain
- Still prone to noise

Random Forest + PCA

- More consistentsplits
- Small improvement in minority classes
- Reduced variance

SVM + PCA

- Best improvement in training time
- Slight macro F1 gains
- Faster inference

AdaBoost + PCA

- Minimal change
- Still the strongest classical model
- Robust to dimensionality reduction

PCA improved stability more than accuracy; AdaBoost remained the top model.

Stage 2: Dataset Enhancement & Model Refinement

1. Collected more data
   - Proposed LayeredBGDataset
   - Cucumber Leaf Disease Dataset
   - Bottle Gourd Dataset
   - Tomato Dataset

   Added **real-world images** but we did not conduct augmentation. Reason: Augmentation + preprocessing would take **a lot of time** becausewe wantedtoupscale to 51000
2. Split data into train (75%) - test (25%) SMOTE applied only on the train data. Applied PCA To reduce dimensionality
3.
4.

Stage 2: Dataset Enhancement & Model Refinement

Why SMOTE?

- Dataset still **imbalanced after collection**
- Models like **KNN & SVM bias toward majorityclass**
- SMOTE [3] generates synthetic minority pointsinfeaturespace
- But may introduce **unrealistic samples** [3] **riskofoverfitting**

Limitation

Can create **noise when classes are not well separated**

May exaggerate patterns [3] misleading learning

Test data still imbalanced [3] **real-world mismatch**

Hidden Bottlenecks

SMOTE can generate unrealistic samples, leading to overfitting, especially with complex models like Random Forest. It also increases training complexity and may introduce synthetic noise risk.

Model Evaluation

KNN

- Hightrain accuracy, strong drop in test [3] **overfitting**
- **Sensitive to local patterns**, SMOTE synthetic clusters mislead
- PCA reduces noise but may lose subtle disease features
- **Low bias, very high variance**

Metric	Train	Test
Plant Type Accuracy	**0.9964**	**0.9055**
Status Accuracy	**0.9852**	**0.7311**
Macro F1 (Status)	**0.9890**	**0.5493**

Decesion Tree

- **Near-perfecttrain**,weaktest [3] memorised data
- Overfits to SMOTE-generated structured samples
- Highly sensitive to feature space distortions
- **Very high variance, low generalisation**

Metric	Train	Test
Plant Type Accuracy	**1.0000**	**0.6077**
Status Accuracy	**0.9999**	**0.4653**
Macro F1 (Status)	**0.9999**	**0.2779**

RANDOM FOREST

Extreme overfitting

- **Perfect train**, sharp test performance drop
- Handles major classes, **fails minority (F1 = 0.00)**
- SMOTE causes deep trees to learn synthetic noise
- **Low bias, very high variance**

Metric	Train	Test
Plant Type Accuracy	**1.0000**	**0.8409**
Status Accuracy	**0.9999**	**0.6733**
Macro F1 (Status)	**0.9999**	**0.3783**

Model Evaluation

ADABOOST

- **Very poor train &test [3] underfitting**
- Synthetic SMOTE samples amplified as errors [3] noise boosting
- Weak learners struggle after PCA compression
- **High bias, low variance**

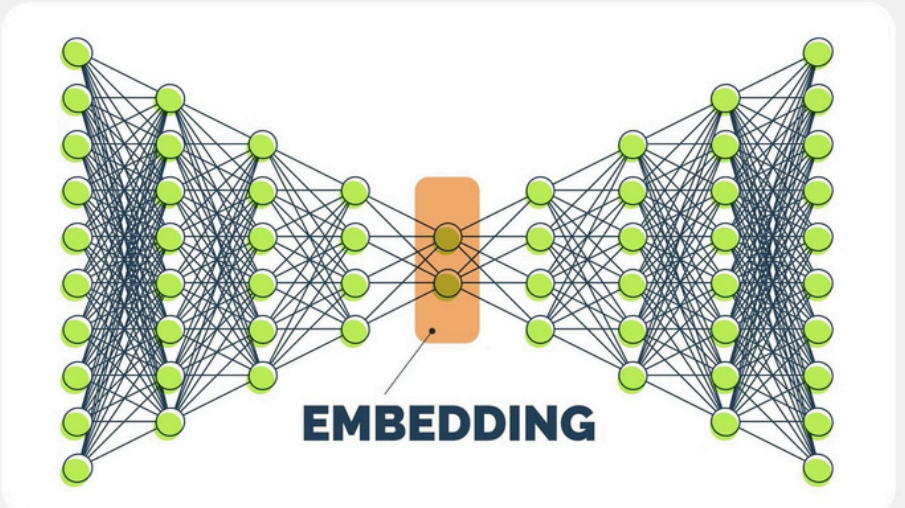
Metric	Train	Test
Plant Type Accuracy	0.5462	0.5100
Status Accuracy		
Macro F1 (Status)	0.2277	0.2847
	0.1267	0.1009

Logistic Regression

- Stable butlower accuracy on complex diseaseclasses
- Works betterfor **PlantType (linear), notStatus(non-linear)**
- SMOTE improved balance but oversimplified decision boundary
- **High bias, low variance**

Metric	Train	Test
Plant Type Accuracy	1.0000	0.9485
Status Accuracy		
Macro F1 (Status)	0.9385	0.6696
	0.9731	0.4879

Auto Encoder



Convolution Autoencoder

The Encoder: This acts as our feature extractor. We passed the image tensor through sequential layers of Convolution (to find patterns) and Pooling (to shrink the size). So this process transformed the input from a high dimensional tensor into a compact vector of just 128 values.

The Decoder: This layer attempts to reconstruct the image. We used UpSampling, which creates new pixels to increase the image size, effectively reversing the pooling operation.

Methodology:

- Step 1: Preprocessing and Masking
- Step 2: Self Supervised Training
- Step 3: Feature Extraction
- Step 4: Hybrid Classification

Auto Encoder Results

```
========================================
  EVALUATION: RF + AE (Split First)
========================================

[ Target 1: Plant Type ]
Accuracy: 0.9317

[ Target 2: Status ]
Accuracy: 0.6187
Macro F1: 0.3759

Detailed Status Report:
              precision    recall  f1-score   support

          DM       0.77      0.73      0.75       233
          EB       0.42      0.26      0.32        19
          FB       0.00      0.00      0.00         9
         JAS       0.29      0.42      0.34        43
           K       0.39      0.45      0.42       185
          LM       0.71      0.46      0.56        52
          LS       0.38      0.43      0.40        47
         MIT       0.53      0.41      0.46        69
           N       0.41      0.42      0.42       291
          PC       0.00      0.00      0.00         8
        PLEI       0.25      0.10      0.14        20
          PM       0.42      0.25      0.31        20
     healthy       0.75      0.77      0.76       908

    accuracy                           0.62      1904
   macro avg       0.41      0.36      0.38      1904
weighted avg       0.62      0.62      0.62      1904
```

```
========================================
  EVALUATION: RF + AE (Split First)
========================================

[ Target 1: Plant Type ]
Accuracy: 1.0000

[ Target 2: Status ]
Accuracy: 0.9999
Macro F1: 0.9999

Detailed Status Report:
              precision    recall  f1-score   support

          DM       1.00      1.00      1.00      1900
          EB       1.00      1.00      1.00       950
          FB       1.00      1.00      1.00       950
         JAS       1.00      1.00      1.00      2850
           K       1.00      1.00      1.00      6650
          LM       1.00      1.00      1.00       950
          LS       1.00      1.00      1.00      2850
         MIT       1.00      1.00      1.00      1900
           N       1.00      1.00      1.00      7600
          PC       1.00      1.00      1.00       950
        PLEI       1.00      1.00      1.00       950
          PM       1.00      1.00      1.00       950
     healthy       1.00      1.00      1.00      7600

    accuracy                           1.00     37050
   macro avg       1.00      1.00      1.00     37050
weighted avg       1.00      1.00      1.00     37050
```

Real-World Testing & Future Directions

SL NO	Image Type	KNN	Logistic Reg	Decision Tree	Random Forest	Ada Boost	SVM
1	Tomato Healthy	Tomato Healthy	Bottle Gourd Healthy	Tomato Healthy	Tomato Healthy	Egg Plant Healthy	Bottle Gourd Healthy
2	Tomato N	Tomato K	Tomato K	Tomato K	Tomato K	Tomato Healthy	Bottle Gourd Healthy
3	Bitter Gourd N	Bitter Gourd DM	Bitter Gourd LS	Bitter Gourd DM	Bitter Gourd DM	Bitter Gourd Healthy	Bitter Gourd JAS
4	Bottle Gourd DM	Bottle Gourd DM	Bottle Gourd DM	Bottle Gourd DM	Bottle Gourd DM	Bottle Gourd K	Bottle Gourd DM
5	Ash Gourd Healthy	Ash Gourd Healthy	Ash Gourd Healthy	Ash Gourd Healthy	Ash Gourd Healthy	Ash Gourd Healthy	Ash Gourd Healthy
6	Ash Gourd K	Ash Gourd K	Ash Gourd K	Ash Gourd Healthy	Ash Gourd Healthy	Eggplant K	Ash Gourd K
7	Cucumber Healthy	Cucumber Healthy	Cucumber Healthy	Cucumber Healthy	Cucumber Healthy	Ridge Healthy	Cucumber Healthy
8	Cucumber DM	Cucumber Healthy	Cucumber Healthy	Eggplant N	Cucumber Healthy	Snake Gourd N	Cucumber K
9	Ridge Healthy	Ridge Gorud N	Ridge Gorud N	Bitter Gourd Healthy	Bitter Gourd Healthy	Cucumber N	Ridge Gorud K
10	Eggplant Healthy	Cucumber Healthy	Eggplant DM	Ash Gourd PM	Cucumber Healthy	Bitter Gourd PLEI	Eggplant MIT
11	Bottle Gourd Healthy	Bottle Gourd Healthy	Bottle Gourd Healthy	Bottle Gourd Healthy	Bottle Gourd Healthy	Bottle Gourd Healthy	Bottle Gourd Healthy

Key Insights

- Strong modelbias toward **correct plantspecies identification** and **Healthy cases** (e.g., Cases 5 & 11 show near-perfect agreement).
- **Major struggle in disease differentiation**, especially between **similar nutrient deficiencies** like *Tomato N vs. Tomato K*.
- **Feature overlap** in leaf symptoms confuses models and blurs decision boundaries.
- **KNN and Random Forest** perform most consistently able to detect plant type even when disease prediction fails.
- **AdaBoost is the most unstable**, often misclassifying even the plant species (e.g., Cucumber ³ Snake Gourd).
- Accuracy can improve by **engineering better distinguishing features** (e.g., yellowing vs. spotting patterns).
- Current dataset issues: **SMOTE-based synthetic samples** improved balance but **reduced real-world reliability**.
- Real performance suggests need for **more real, diverse, high-quality leaf images** instead of artificial oversampling.

Limitations of Our Models

- **Models couldnot"see"realvisual patterns**
  Classical MLmodels(KNN, AdaBoost, RF) dependonly on**hand craftedfeatures**. Sincewe only gave them color/shape statistics, the model treated different yellowing patterns (virus vs nutrient deficiency) as mathematically identical.
- **Critical disease details were lost during preprocessing**
  Downsizing images to **128×128** reduced noise but also removed early-stage spots and fine textures that are essential for accurate diagnosis.
- **Segmentation + preprocessing removed useful information**
  Green screen segmentation cleaned backgrounds but also simplified leaf edges and minor discoloration patterns that could have helped in classification.
- **SMOTE caused overfitting instead of improving learning**
  For rare diseases, SMOTE created synthetic images that the model memorized. This explains the **very high training accuracy** but poor performance on real test data.
- **Feature extraction was too limited for complex plant diseases**
  Color histograms and shape features cannot capture vein distortions, mosaic patterns, patch textures, or irregular spotting which are **key indicators** for real-world disease detection.