1 BananaLSD Dataset

Abstract

The **BananaLSD dataset** is a curated collection of banana leaf images intended for plant disease detection research. It contains photographs of both healthy and infected leaves taken directly from agricultural fields in Bangladesh. The dataset includes four balanced categories: *Healthy, Sigatoka, Cordana*, and *Pestalotiopsis*. Its structure makes it suitable for training deep learning models, and it has been widely used for building diagnostic systems to support farmers in detecting banana diseases.

Author Information

The dataset was introduced in 2021 by **Shifat Earman** at *Bangabandhu Sheikh Mujibur Rahman Agricultural University, Bangladesh*, and is hosted on **Kaggle**.

Dataset Overview

BananaLSD is carefully balanced, with **400 images per class**, totalling **1,600 samples**. Out of these, **937 are original photographs**, while the rest were generated using augmentation. All images are **JPEG/JPG, RGB format**, resized to **224 × 224 pixels**. Typical splits are around **70% training, 20% validation, and 10% testing**.

Feature	**Description**
Classes	Healthy, Sigatoka, Cordana, Pestalotiopsis
Images per class	400 each
Total images	1,600 (937 original + augmented)
Image type	JPEG/JPG, RGB
Resolution	224 × 224 pixels
Usage split	70% train, 20% val, 10% test

Table 1: BananaLSD Dataset Summary

Purpose

The dataset fills a gap in public resources for banana leaf disease detection. Captured under real-world conditions using smartphones, it provides realistic samples for developing **automated diagnosis tools** and **mobile applications** for farmers.

Background

Developed in **June 2021** through field surveys in Bangladesh, BananaLSD has since been used for CNN benchmarking (ResNet, DenseNet, EfficientNet), lightweight models, and explainable AI approaches.

Characteristics

- RGB images resized to 224 × 224

- Four class labels
- Metadata reflecting field conditions

Extractable Features

- Color differences (healthy green vs. brown patches)
- Texture descriptors (GLCM, LBP)
- Shape and lesion contours
- CNN feature embeddings

Disease Categories

- **Sigatoka Leaf Spot**: Black streaks spreading into necrotic patches (*Pseudocercospora fijiensis*)
- **Cordana Leaf Spot**: Brown oval lesions with pale centers (*Cordana musae*)
- **Pestalotiopsis Leaf Spot**: Gray-brown lesions with dark margins, merging in humidity (*Pestalotiopsis spp.*)
- **Healthy**: Green, lesion-free baseline leaves

Applications

- CNN benchmarking
- Real-time mobile diagnostic tools
- Explainable AI (Grad-CAM)
- Low-cost agricultural AI solutions

Limitations

- Small dataset size compared to larger repositories
- Only three diseases represented
- Augmentation cannot fully mimic real-world variety

2 Corn (Maize) Leaf Disease Dataset

Abstract

The **Corn (Maize) Leaf Disease dataset** is a benchmark resource for maize pathology research. It contains images across four categories: *Healthy, Common Rust, Gray Leaf Spot*, and *Blight*, sourced mainly from PlantVillage and similar repositories. It is frequently used in deep learning studies to support the creation of AI-based tools for disease classification and mobile diagnosis.

Author Information

The dataset was curated in 2020 by **Smaranjit Ghose**, based on PlantVillage and public datasets. It was reorganized for Kaggle and has since become one of the most cited maize disease datasets in AI research.

Dataset Overview

The dataset has **4,188 images** divided across four categories. Files are stored in **JPEG/JPG format** as **RGB color images**. While original resolutions varied, most experiments use resized images of **224 × 224 pixels**. Studies generally split data into **70% training and 30% testing**.

Feature	**Description**
Classes	Common Rust, Gray Leaf Spot, Blight, Healthy
Images per class	Rust: 1,306; Gray Spot: 574; Blight: 1,146; Healthy: 1,162
Total images	4,188
Image type	JPEG/JPG, RGB
Resolution	Resized to 224 × 224 pixels
Usage split	70% train, 30% test

Table 2: Corn Dataset Summary

Purpose

Designed as a standardized benchmark, the dataset supports machine learning studies and practical diagnostic tools for maize pathology.

Background

As a structured subset of PlantVillage, the dataset was reorganized for Kaggle. It has since been heavily cited in IEEE and Elsevier publications (2023–2025), and used for CNN benchmarking, mobile AI, and explainable AI research.

Characteristics

- RGB leaf images
- Clear disease labels

- Metadata preserved from source

Extractable Features

- Color features (orange pustules, gray-tan lesions)
- Texture differences (rough vs. smooth patches)
- Shape differences (elliptical vs. rectangular lesions)
- Deep CNN feature embeddings

Disease Categories

- **Common Rust** (*Puccinia sorghi*): Orange pustules causing early leaf death
- **Gray Leaf Spot** (*Cercospora zeae-maydis*): Rectangular lesions reducing yields by up to 50%
- **Blight** (*Exserohilum turcicum*): Large elliptical lesions in humid climates
- **Healthy**: Green leaves with no visible damage

Applications

- CNN benchmarking (VGG16, ResNet, EfficientNet, Inception)
- Model explainability (Grad-CAM, saliency mapping)
- Mobile diagnostic systems
- Reference dataset in academic studies

Limitations

- Images mostly from controlled settings
- Class imbalance (Gray Leaf Spot underrepresented)
- Limited scope (excludes other maize diseases)