



## Data Article

# A comprehensive dataset of tomato leaf images for disease analysis in Bangladesh

Mangsur Kabir Oni, Tabia Tanzin Prama\*

Jahangirnagar University, Dhaka 1342, Bangladesh

## ARTICLE INFO

*Article history:*

Received 30 August 2024

Revised 14 November 2024

Accepted 20 January 2025

Available online 27 January 2025

Dataset link: [Tomato Leaf Image Dataset for Disease Analysis in Real-World Environment \(Original data\)](#)

*Keywords:*

Tomato

Healthy leaves

Diseased leaves

Computer vision

## ABSTRACT

Agriculture is the largest employment sector in Bangladesh, making up 13.4 percent of Bangladesh's GDP in 2024. For being most consumable crops, almost 9 out of 10 farmers grow tomatoes and earn their living. Tomato (*Solanum lycopersicum*) ranks fourth in respect of production and third in respect of area in Bangladesh. The tomato, a cornerstone of global agriculture, relies heavily on the health of its leaves for optimal growth and yield. These leaves are essential for photosynthesis, respiration, and transpiration, processes that directly influence the plant's overall vitality. Understanding the structure, function, and physiological characteristics of tomato leaves is crucial for developing effective agricultural strategies to maximize production and minimize the impact of environmental stressors and diseases. While tomato leaves exhibit a wide range of morphological variations across cultivars, they remain susceptible to a variety of threats. Pests, pathogens, nutrient deficiencies, temperature extremes, and environmental pollutants can all compromise leaf health. Biotic stresses, especially foliar diseases caused by bacteria, fungi, viruses, and other pathogens, are particularly devastating to tomato production. This research presents a dataset of leaves from tomato plants that are both insect-damaged and healthy. Our dataset contains 1,028 images of tomato leaves collected in Bangladesh, including 482 images of healthy leaves and 546 images of diseased leaves. The images were captured from two different tomato gardens in February, 2024 [1]. We captured the images the

\* Corresponding author.

E-mail address: [prama.stu2017@juniv.edu](mailto:prama.stu2017@juniv.edu) (T.T. Prama).

Social media: [@TabiaTanzin](#) (T.T. Prama)

in diverse backgrounds, angles, and lighting conditions. Each image is precisely annotated to mark regions as either healthy or diseased, accounting for the complex background in each image. This dataset serves as a comprehensive resource for researchers and learners to analyze and improve the health management of tomato plants through the development of advanced computational models.

© 2025 The Authors. Published by Elsevier Inc.  
This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>)

Specifications Table

Subject	Plant Pathology, Agricultural science, Horticulture, Computer science.
Specific subject area	Computer Vision, Image processing, Image classification
Type of data	Raw Images
Data collection	Two smartphones are used to manually take high-quality photos. Photos taken on two different days. The top of the leaf is more heavily inspected to determine if it is healthy or unhealthy. And they are saved in within two different folders. Images were acquired using a Xiaomi M2101K6P camera and a Samsung SM-A105F camera.
Data source location	Images were collected from two different tomato gardens. Garden's location is mentioned below. <b>Village:</b> Mohammadpur <b>District:</b> Brahmanbaria, <b>Country:</b> Bangladesh
Data accessibility	Repository name: Tomato Leaf Image Dataset for Disease Analysis in Real-World Environment Data identification number: <a href="https://data.mendeley.com/datasets/rnbsw72zb5.1">10.17632/rnbsw72zb5.1</a> Direct URL to data: <a href="https://data.mendeley.com/datasets/rnbsw72zb5.1">https://data.mendeley.com/datasets/rnbsw72zb5.1</a> Instructions for accessing these data: Datasets consist of Images and Labels folder.
Related research article	

1. Value of the Data

- Consists of 1,028 Tomato leaf images collected via mobile devices, aiding farmers in Bangladesh in making informed decisions on irrigation, fertilization, and pest management.
- Supports the development and training of machine learning models for monitoring Tomato leaf health and distinguishing between healthy and diseased leaves.
- Facilitates automated plant health assessments, enabling accurate monitoring and timely interventions.
- Essential for early disease detection in Tomato cultivation, helping to minimize crop losses and reduce reliance on chemical treatments.
- Useful for creating crop monitoring applications that assist farmers in real-time Tomato crop management, improving yield and quality.
- Serves as a valuable resource for studies on Tomato health, disease detection, and crop management, both locally and globally.

2. Background

Bangladesh's economy is fundamentally anchored in agriculture, making farming a crucial source of income for the nation. Given its dense population and various environmental challenges, ensuring food security is a critical concern. For Bangladeshi farmers, identifying tomato

leaf diseases is vital due to its significance in agricultural livelihoods and national food security. In the fertile fields of Bangladesh, where crops flourish, the health of tomato plants is essential for agricultural success. Understanding the importance of tomato cultivation for both livelihoods and economic stability, farmers are motivated to stay vigilant in monitoring and detecting diseases that could impact their yields. This vigilance extends beyond protecting individual harvests and reflects a broader dedication to sustainable agriculture and environmental stewardship. By adopting disease detection practices, Bangladeshi farmers take a proactive stance on pest management, minimizing the use of chemical pesticides and supporting integrated pest management strategies. By maintaining crop health and soil fertility, they contribute to the long-term sustainability of agricultural ecosystems, ensuring that the land remains viable for future generations. Moreover, the significance of disease detection goes beyond individual farms and is closely linked to national food security. In a nation where agriculture is both the economic backbone and the primary livelihood for millions, the health and productivity of tomato crops have major implications for community well-being and national stability. By efficiently managing diseases and safeguarding yields, farmers not only secure their own economic success but also play a crucial role in maintaining food supply stability, reducing the risk of shortages and price fluctuations. Our dataset is indirectly connected to these subjects. Several research datasets have been developed to support agricultural innovation. We can mention some of them like, Jameer Kotwal [2] introduces a soybean dataset containing 3,363 images of healthy and insect-damaged leaves from multiple farms. Kailas Patil [3] presents a dataset of 10,042 images of Lemongrass (*Cymbopogon citratus*) leaves categorized as "Dried," "Healthy," and "Unhealthy," taken in real-world conditions at Vishwakarma University in Pune. Saiful Islam [4] offers the "BDMediLeaves" dataset, featuring 2,029 original and 38,606 augmented images of medicinal plant leaves from ten species in Dhaka, Bangladesh. Fereshteh S. Bashiri [5] provides the MCIndoor20000 dataset, an extensive, annotated collection of over 20,000 images of indoor objects, accessible to researchers. Our tomato leaf dataset aligns with these research efforts. Through these dataset researchers and scientists can improve research on producing more healthy tomato plants and can take significant measures by detecting the leaves as healthy or diseased. Fig. 1 provides the



**Fig. 1.** The real tomato field from where data was collected

dataset sample from the gardens. Researchers can preprocess the dataset according to various requirements to fit their model properly. Farmers are going to be one of the largest users of this dataset and they will be benefitted at a large scale.



3. Data Description

To create a diverse dataset of tomato leaf images, we conducted extensive fieldwork in early February 2024. And we did the fieldwork with the consent of the garden’s owner. We collected 1,028 images from various tomato fields, capturing healthy leaves, diseased leaves, and those affected by environmental stresses. The images were taken under different conditions: sunny (26-29°C) and foggy (17-18°C) days.

Our dataset includes 546 images of diseased leaves and 482 of healthy leaves, providing a robust sample which is shown in Table 2. Each image will be meticulously labeled to distinguish between healthy and diseased leaves and to identify environmental stress conditions. This comprehensive dataset not only provides detailed visual data but also offers insights into the relationships between plant genetics, environmental factors, and disease vulnerability, aiding in the development of advanced disease control methods and crop optimization strategies.

The initial dataset for Tomato Leaf Disease Detection has been categorized into two groups: Healthy Leaves and Diseased Leaves. The images in the dataset are of two sizes: 3472 × 4640 pixels and 3096 × 4128 pixels. Table 1 provides a description of healthy and diseased tomato leaves within the dataset.

**Table 1**  
Description of tomato leaves (Healthy or Diseased).

Class Name	Description	Visualization
Healthy	A healthy tomato leaf is a vibrant green with a smooth, matte texture. Its size and shape can vary depending on the tomato variety, but it typically has a lobed or serrated margin and a pointed tip. Veins are visible running throughout the underside of the leaf, providing a transport network for nutrients and water. The absence of yellowing, browning, wilting, spots, or holes indicates a healthy state.	
Diseased	A diseased tomato leaf deviates from its healthy green vibrancy. Discoloration is a telltale sign, with leaves turning yellow, brown, or mottled. The smooth texture may become wrinkled, puckered, or develop raised lesions. Leaf margins might curl inwards or become distorted, and the once-proud form may wilt or develop holes. These visual disturbances signal the presence of a disease and can vary depending on the specific ailment.	

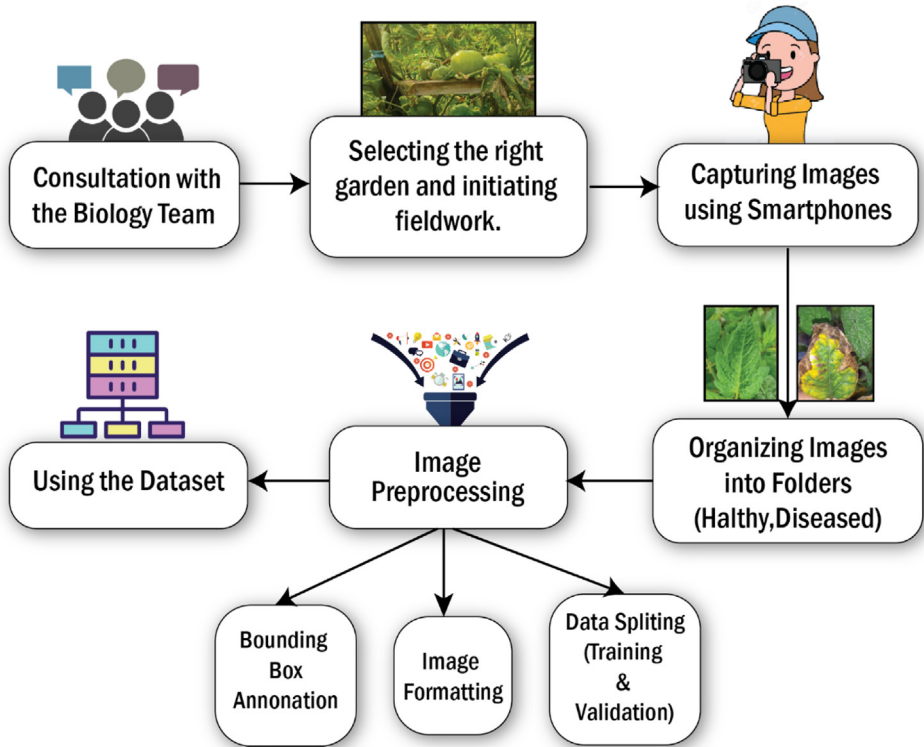
**Table 2**  
Number of Leaves.

Folder Name	Number of Images
Healthy Leaves	482
Diseased Leaves	546

4. Experimental Design, Materials and Methods

Our objective was to assist researchers and scientists by providing a dataset of tomato leaf images to facilitate their research. The creation of this dataset involved several key steps. Initially, we consulted with the Biology Team to gain insights into tomato plants, their diseases, and associated challenges. This collaboration was crucial as our background in computer science did not encompass detailed plant biology. Following expert guidance, we selected two tomato gardens in Mohammadpur, Brahmanbaria, Bangladesh, for fieldwork. After discussing our research objectives with the garden owners, we obtained permission to photograph the tomato leaves. Data collection took place in early February 2024, under varying environmental conditions including sunny days (26-29°C) and foggy days (17-18°C).

We captured images using iPhone 14 Pro Max camera (48MP Main: 24 mm,  $f/1.78$  aperture, second-generation sensor-shift optical image stabilization, seven-element lens, 100% Focus Pixels), ensuring comprehensive coverage by photographing from different angles and under various lighting conditions. The images were then systematically organized into folders to streamline the preprocessing phase.



**Fig. 2.** Workflow of Data Collection Process

**Image preprocessing:** For image preprocessing, we standardized all images to a single .JPG format and created bounding box annotations using the "MakeSense" software. These annotations, saved as text files, are designed to support effective model implementation. Originally, the images in the dataset varied in resolution, with dimensions of  $3472 \times 4640$  pixels and  $3096 \times 4128$  pixels. This variation could lead to inconsistencies during model training, as deep learning models generally require fixed input dimensions. To address this issue, we resized all images to a uniform resolution of  $224 \times 224$  pixels using Keras's ImageDataGenerator, a widely used tool for data preprocessing and augmentation. This resizing step optimizes model training, facilitates batch processing, and reduces computational load. Fig. 2 illustrates the entire data collection and preprocessing process.

## Limitations

In our dataset, we have collected a limited number of images or samples that can hinder the model's ability to generalize well to unseen data. Researchers or learners have to augment this dataset to achieve the proper result. Our dataset has 482 healthy leaves images and 546 diseased leaves images which might be considered as an imbalanced dataset. And using these dataset researchers are unable to classify the tomato leaf diseases. Rather they can only inspect the health condition of the tomato leaves. Additionally, with a higher-quality camera, we could have captured more detailed and clearer photographs of the leaves. Since the timeframe for data collection was not fixed, there may be variations in brightness and color across the images.

## Ethics Statement

Our research began with valuable guidance from advisers in the biological departments of our university. Their expertise provided us with essential knowledge about tomato plants and their diseases. Based on their recommendations, we selected suitable tomato gardens for our study. We then discussed our research objectives with the garden owners, explaining the purpose and potential benefits of our work. Appreciating the positive impact of our research, the owners granted us permission to use their gardens. We were supported by courteous staff members throughout the process, and we ensured that our photography was conducted with utmost care to avoid any harm to the tomato plants or their surroundings.

## Credit Author Statement

**Mangsur Kabir Oni:** Visualization, Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation. **Tabia Tanzin Prama:** Visualization, Conceptualization, Methodology, Investigation, Supervision, Validation, Writing- Reviewing and Editing.

## Data Availability

[Tomato Leaf Image Dataset for Disease Analysis in Real-World Environment \(Original data\)](#) (Mendeley Data).

## Acknowledgements

We would like to express our sincere gratitude to the garden owners who generously allowed us to photograph tomato leaves in their gardens, making this research possible. We also extend

our heartfelt thanks to our biological adviser for providing invaluable expertise on plant biology and for guiding us in applying computer science techniques to this field of study.

### **Declaration of Competing Interest**

We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### **References**

- [1] M.Kabir Oni, Tabia Tanzin Prama, Tomato leaf image dataset for disease analysis in real-world environment, Mendeley Data V1 (2024), doi:[10.17632/rnbsw72zb5.1](https://doi.org/10.17632/rnbsw72zb5.1).
- [2] J. Kotwal, Ramgopal Kashyap, Mohd.Shafi Pathan, An India soyabean dataset for identification and classification of diseases using computer-vision algorithms, *Data Brief.* 53 (2024) April, doi:[10.1016/j.dib.2024.110216](https://doi.org/10.1016/j.dib.2024.110216).
- [3] K. Patil, Yogesh Suryawanshi, Alimurtuza Patrawala and Prawit Chumchu, A comprehensive lemongrass (*Cymbopogon citratus*) leaf dataset for agricultural research and disease prevention, *Data Brief.* 53 (2024), doi:[10.1016/j.dib.2024.110104](https://doi.org/10.1016/j.dib.2024.110104).
- [4] S. Islam, Md.Rayhan Ahmed, Siful Islam, Md Mahfuzul Alam Rishad, Sayem Ahmed, Toyabur Rahman Utshow and Minhajul Islam Siam, BDMediLeaves: a leaf images dataset for Bangladeshi medicinal plants identification, *Data Brief.* 50 (2023), doi:[10.1016/j.dib.2023.109488](https://doi.org/10.1016/j.dib.2023.109488).
- [5] F.S. Bashiri, Eric LaRose, Peggy Peissig, Ahmad P. Tafti, MCIndoor20000: a fully-labeled image dataset to advance indoor objects detection, *Data Brief.* 17 (2018) 71–75, doi:[10.1016/j.dib.2017.12.047](https://doi.org/10.1016/j.dib.2017.12.047).