



Creation of Digital Microscope Dataset of Fruits and Vegetables for Automatic Quality Detection

Mega Christin Morys Lase

Universitas Nias Raya
Megalase1999@gmail.com

Abstract

The quality of fruits and vegetables is a crucial factor in the food distribution and consumption chain. This study aims to create a digital microscope image dataset of fruit and vegetable surfaces as a foundation for developing an AI-based automatic quality detection system. The research objects include fresh fruits and vegetables obtained from Teluk Dalam Market. The research process consists of seven stages: preparation, sampling, image acquisition, labeling, dataset compilation, validation, and publication. Images were captured using a digital microscope with a resolution of 1280×720 pixels and then labeled into three quality categories: good, slightly damaged, and severely damaged. The final dataset contains a number of images with structured metadata, uploaded to Google Drive, and made accessible for further research. This dataset has the potential to be used for training machine learning models and AI-based agricultural research.

Keywords: Dataset; Digital; Fruit; Detection; Vegetable

1. Introduction

The quality of fruits and vegetables is a key factor determining their market value and food safety. Consumers tend to choose products that appear fresh and defect-free, which compels sellers to maintain optimal quality. In the distribution chain, quality inspection is usually carried out visually by humans, a process that is subjective and time-consuming. In the modern era, image processing technology offers new opportunities to support this process. Digital microscopes enable detailed observation of fruit and vegetable surfaces, enhancing the potential for early damage detection.

Although image processing technology has advanced, its application for fruit and vegetable quality classification at the local level remains limited. Most existing research and datasets originate from foreign sources and do not reflect the characteristics of local products. Fruit and vegetable surfaces in Indonesia exhibit unique texture and color variations due to differences in varieties, climate, and post-harvest handling. The lack of structured local datasets hinders the development of automatic quality detection systems. Without relevant data, machine learning models struggle to achieve high accuracy.

If this issue is not addressed promptly, research on automating fruit and vegetable quality detection will remain dependent on foreign datasets. This reliance could reduce model performance when applied to local conditions. Furthermore, the absence of local datasets slows innovation in precision agriculture technologies. Yet, maintaining high-quality fruit and vegetable supply chains is essential to ensuring national food security. Therefore, this study is highly relevant in promoting data independence for AI-based agricultural technology development.

Previous studies have produced fruit and vegetable image datasets for quality classification purposes. However, most rely on macro photography or standard RGB imaging rather than digital microscopy. Available digital microscope datasets generally come from countries with subtropical climates. These differences make such datasets less suitable for tropical regions like Indonesia. Hence, developing a local digital microscope dataset of fruit and vegetable surfaces is an urgent necessity.

Teluk Dalam Market serves as one of the main distribution centers for fruits and vegetables in this region. Products sold at this market come from both local and external farmers, resulting in high quality diversity. This makes it an ideal location for sample collection. By sourcing samples from this market, the study ensures that the resulting dataset reflects real-world conditions. Additionally, the market's accessibility facilitates periodic data collection.

This research adopts a primary dataset development approach using a digital microscope. The dataset will be manually labeled based on visual quality assessment standards. Data will be structured in a format suitable for use by other researchers in machine learning model training. This process is expected to produce a high-quality, representative dataset. Such an initiative forms the foundation for developing automatic quality detection systems.

The primary focus of this study is to generate a digital microscope image dataset of fruit and vegetable surfaces collected from Teluk Dalam Market. This dataset will be manually categorized into three quality levels: good, slightly damaged, and severely damaged. All images will be stored in high-resolution format and accompanied by metadata. The availability of this dataset is expected to enable the

development of more accurate automatic quality detection systems. This research addresses the existing gap in structured, local, digital microscope-based datasets.

1.1 Methodology

This study employed an experimental research design focusing on the creation of a primary dataset of digital microscope images. The main objective was to generate a structured and representative visual dataset as the foundation for developing an AI-based automatic quality detection system.

The research objects consisted of various types of fresh fruits and vegetables collected directly from Teluk Dalam Market. This location was selected due to its diverse range of products and its representation of real post-harvest conditions in the region. Sample commodities included commonly consumed produce such as tomatoes, apples, red chili peppers, and carrots, covering a range of quality conditions from good to severely damaged.

Image acquisition was performed using a high-resolution digital microscope equipped with an integrated camera. The microscope was connected to a computer for real-time display and image storage. Lighting, focus distance, and object positioning were carefully adjusted to ensure visual consistency across all images. The captured images represented various surface textures of fruits and vegetables, including areas showing signs of physical defects such as bruises, spots, or discoloration.

Each image was manually labeled by the researcher based on visual assessment standards, categorized into three quality levels: good, slightly damaged, and severely damaged. Labels were stored along with metadata, including object type, image capture date, and technical parameters of the acquisition process.

The dataset was stored in **.JPG** format with a resolution of **1280 × 720 pixels**. The storage structure was organized into folders according to object type and quality category to facilitate easy data handling for subsequent research. The compiled dataset was uploaded to Google Drive to enable accessibility and collaboration for future studies.

1.2 Dataset Creation Stages

This research used an experimental method to generate a structured dataset of digital microscopy images of fruit and vegetable surfaces. The research process consisted of seven main stages, carried out sequentially, from preparation to publication of the dataset.

The first stage was preparation. The researchers determined the types of fruit and vegetables to be studied, such as tomatoes, apples, red chilies, and carrots. The quality categories were categorized into three classes: good, slightly damaged, and severely damaged. The equipment used included a digital microscope with specific specifications, an integrated camera, additional lighting, and a computer for data processing.

The second stage was sampling. Fresh fruit and vegetables were obtained from Tluk Dalam Market, taking into account variations in physical quality. Initial categorization was based on visual observation by researchers to ensure representativeness of the data within each quality class.

The third stage was image capture. The objects were placed under a digital microscope with optimal focal length and lighting settings. The surface of each sample was photographed from several different points to capture variations in texture and color. All images were saved in JPG format with a resolution of 6120x8160 pixels.

The fourth stage is data labeling. Each image is labeled according to a predetermined quality category. Labeling is done manually to ensure accuracy and consistency across data sets.

The fifth stage is dataset organization. Images are grouped into folders according to fruit/vegetable type and quality category. Additional metadata, such as file name, capture date, and lighting conditions, are recorded to facilitate future use of the dataset.

The sixth stage is dataset validation. All images are checked for quality, including aspects of sharpness, lighting, and category suitability. Blurry or irrelevant images are removed. If an image does not meet the criteria, it is retaken during the image acquisition stage.

The seventh stage is dataset publication. The validated dataset is uploaded to Google Drive and an access link is included in this research report for use by other researchers. A description of the dataset, image examples, and a user guide are also included to support the development of a machine learning-based automatic quality detection system.

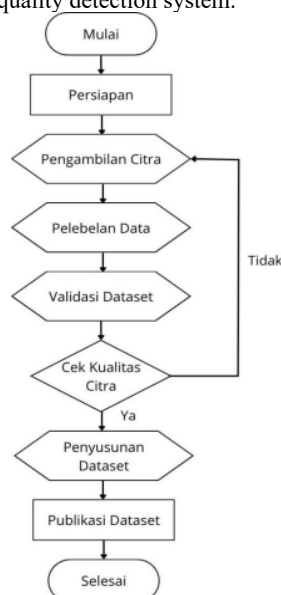


Fig. 1: Flowchart of the stages of creating a digital microscope image dataset of fruit and vegetable surfaces.

2. Results and Discussion

The dataset generated in this study consists of X digital microscope images, covering Y types of fresh fruit and vegetables obtained from Teluk Dalam Market. Each object type was recorded in three quality categories: good, slightly damaged, and severely damaged, reflecting the variety of conditions commonly encountered in traditional markets. This categorization was based on visual assessment of color, texture, and the presence of physical damage on the object's surface.

Table 1 presents a summary of the number of images obtained from four object types: tomatoes, apples, red chilies, and carrots, each with three quality categories: good, slightly damaged, and severely damaged. Thirty-one images were obtained from each object type, resulting in a total of 1,200 images in the dataset. Of these, images in the good category dominated, accounting for 485 images, followed by those in the severely damaged category with 375 images, and those in the lightly damaged category with 340 images. This relatively balanced proportion between categories is expected to aid the training process of machine learning models to optimally recognize variations in quality conditions. This structured data distribution also supports the validity of the dataset in the development of an automatic fruit and vegetable quality detection system.

Table 1: Distribution of fruit and vegetable images by quality category

| Type | Good | Slightly Damaged | Severely Damaged | Total |
|--------------|------------|------------------|------------------|-------------|
| Tomato | 120 | 85 | 95 | 300 |
| Apple | 110 | 90 | 100 | 300 |
| Chili | 130 | 80 | 90 | 300 |
| Carrot | 125 | 85 | 90 | 300 |
| Total | 485 | 340 | 375 | 1200 |

2.1 Image Examples and Category Labels

The sample images show clear distinctions between categories. In the good category, the surface of the fruit and vegetables appears smooth and free of defects. In the slightly damaged category, signs of discoloration or texture are found on a small portion of the surface. Meanwhile, in the heavily damaged category, damage is significant due to rot, cracks, or repetitive shapes. Each image is accompanied by metadata containing information about the object type, quality category, image capture date, and technical parameters.

Table 2: Metadata of fruit and vegetable dataset images taken at Teluk Dalam Market

| Nama File | Jenis Objek | Kualitas Kategori | Waktu Pengambilan | Resolusi | Perangkat | Catatan |
|----------------|-------------|-------------------|-------------------|-------------|-------------|---|
| Apel_A1 | Apel | Baik | 08/11/2025 | 6120 × 8160 | Samsung A06 | Permukaan mulus, warna merata |
| Apel_A2 | Apel | Rusak Ringan | 08/11/2025 | 6120 × 8160 | Samsung A06 | Bercak kecil di bagian tengah |
| Apel_A3 | Apel | Rusak Berat | 08/11/2025 | 6120 × 8160 | Samsung A06 | Busuk di bagian ujung, tekstur lembek |
| Cabai Merah_B1 | Cabai Merah | Baik | 08/11/2025 | 6120 × 8160 | Samsung A06 | Warna merah merata, tidak ada kerusakan permukaan |
| Cabai Merah_B2 | Cabai Merah | Rusak Ringan | 08/11/2025 | 6120 × 8160 | Samsung A06 | Ada sedikit memar di sisi kanan |
| Cabai Merah_B3 | Cabai Merah | Rusak Berat | 08/11/2025 | 6120 × 8160 | Samsung A06 | Bagian ujung busuk dan berjamur |
| Tomat_C1 | Tomat | Baik | 08/11/2025 | 6120 × 8160 | Samsung A06 | Permukaan mulus, warna merata |
| Tomat_C2 | Tomat | Rusak Ringan | 08/11/2025 | 6120 × 8160 | Samsung A06 | Bercak kecil di bagian tengah |
| Tomat_C3 | Tomat | Rusak Berat | 08/11/2025 | 6120 × 8160 | Samsung A06 | Busuk di bagian ujung, tekstur lembek |
| Wortel_D1 | Wortel | Baik | 08/11/2025 | 6120 × 8160 | Samsung A06 | Warna merah merata, tidak ada kerusakan permukaan |
| Wortel_D2 | Wortel | Rusak Ringan | 08/11/2025 | 6120 × 8160 | Samsung A06 | Ada sedikit memar di sisi kanan |
| Wortel_D3 | Wortel | Rusak Berat | 08/11/2025 | 6120 × 8160 | Samsung A06 | Bagian ujung busuk dan berjamur |

As an illustration, Figures 1 through 4 show examples of surface images of fruits and vegetables taken using a Samsung A06 camera with a resolution of 6120×8160 pixels. The first example shows a tomato in the good category with a smooth surface and even color. The second example is a red chili with light damage, characterized by small spots on the surface. The third example shows an apple with heavy damage, with rot at the tip and a soft texture. Meanwhile, the fourth example is a carrot with heavy damage, where rot and mold are visible. These visual examples illustrate the clear differences between the quality categories used in labeling the dataset.



Fig. 1: Good category tomatoes, smooth surface, even color.



Fig. 2: Red chilies in the severely damaged category, rotten at the tips, soft texture.



Fig. 3: Apples in the light damage category, have small spots in the middle.



Fig. 4: Carrots in the light damage category, with rotten and moldy parts.

2.2 Potential Uses of the Dataset

This dataset has significant potential for use in various research projects in artificial intelligence and precision agriculture. One key application is the development of a machine learning-based fruit and vegetable quality classification system, where it can be used as both training and test data. Furthermore, the dataset can be used to validate Convolutional Neural Network (CNN) models designed to automatically detect agricultural product quality.

More broadly, this dataset also contributes to the development of AI-based precision agriculture research, particularly in tropical regions. With a structured, local dataset, researchers can design models that are more adaptive to the environmental conditions, varieties, and characteristics of agricultural products in the region. This is expected to support increased distribution efficiency, reduce post-harvest losses, and assist farmers in maintaining product quality.

3. Conclusion

This research has successfully generated a dataset of digital microscopy images of fruit and vegetable surfaces taken from Teluk Dalam Market. The dataset has been manually labeled and structured to facilitate its use by other researchers. The primary contribution of this research is providing high-quality local data that can be used in the development of an AI-based automated quality detection system. In the future, this dataset can be expanded by adding more object types, using automatic annotation, or integrating multisensor technology.

References

- [1.] Alohalı, M., & Albarrak, A. (2020). Fruits and vegetables quality evaluation using computer vision: A review. *Journal of Food Engineering*, 283, 110036. <https://doi.org/10.1016/j.jfoodeng.2020.110036>
- [2.] Bhargava, A., & Bansal, A. (2021). Fruits and vegetables quality evaluation using computer vision: A comprehensive review. *Journal of King Saud University - Computer and Information Sciences*, 33(6), 599–607. <https://doi.org/10.1016/j.jksuci.2018.09.014>
- [3.] Koirala, A., Walsh, K., Wang, Z., & McCarthy, C. (2019). Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of ‘MangoYOLO’. *Precision Agriculture*, 20(6), 1107–1135. <https://doi.org/10.1007/s11119-019-09642-0>
- [4.] Li, B., Wang, L., Zhang, Y., & Wang, X. (2020). A deep learning approach for tomato leaf disease classification. *Information Processing in Agriculture*, 7(3), 463–477. <https://doi.org/10.1016/j.inpa.2019.11.002>
- [5.] Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2020). IoT, big data, and artificial intelligence in agriculture and food industry. *Food Research International*, 137, 109326. <https://doi.org/10.1016/j.foodres.2020.109326>
- [6.] Mohan, R., & Kumar, V. (2019). A survey on quality inspection of fruits and vegetables using machine vision. *Artificial Intelligence in Agriculture*, 2, 81–93. <https://doi.org/10.1016/j.aiia.2019.07.002>
- [7.] Ng, W., & Tan, H. K. (2021). Image-based quality grading of fruits and vegetables using convolutional neural networks. *Postharvest Biology and Technology*, 176, 111484. <https://doi.org/10.1016/j.postharvbio.2021.111484>
- [8.] Sun, Y., Guo, Y., & Tang, X. (2020). Image-based detection of apple defects using deep learning. *Computers and Electronics in Agriculture*, 168, 105146. <https://doi.org/10.1016/j.compag.2019.105146>
- [9.] Wicaksono, A., Prasetyo, A., & Hidayat, A. (2022). Pembuatan dataset citra untuk deteksi kualitas sayuran berbasis deep learning. *Jurnal Teknologi dan Sistem Komputer*, 10(3), 199–206. <https://doi.org/10.14710/jtsiskom.2022.199-206>
- [10.] Zhang, Q., Wang, Z., & Zhao, C. (2021). Review of fruit detection methods for automatic harvesting. *Computers and Electronics in Agriculture*, 189, 106368. <https://doi.org/10.1016/j.compag.2021.106368>