



Data Article

A labelled dataset of healthy and diseased common bean (*Phaseolus vulgaris*) from Tanzania



Neema Mduma^{a,*}, Hudson Laizer^b

^aSchool of Computational and Communication Sciences and Engineering, The Nelson Mandela African Institution of Science and Technology, P.O Box 447, Tengeru, Arusha, Tanzania

^bSchool of Life Sciences and Bioengineering, The Nelson Mandela African Institution of Science and Technology, P.O. Box 447, Tengeru, Arusha, Tanzania

ARTICLE INFO

Article history:

Received 13 April 2026

Revised 10 May 2026

Accepted 29 May 2026

Available online 31 May 2026

Dataset link: [A Labeled Dataset of Healthy and Diseased Common Beans from Tanzania \(Original data\)](#)

Keywords:

Phaseolus vulgaris

Image dataset

Machine learning

Deep learning

Crop disease detection

Bean anthracnose

Bean rust

ABSTRACT

Common bean (*Phaseolus vulgaris*) is an important food and cash crop in Tanzania, where it contributes to household nutrition and income among smallholder farmers. Its production is, however, constrained by diseases such as bean rust and bean anthracnose, which can cause substantial yield losses and reduce crop quality. This article presents a labelled image dataset of common bean leaves collected under field conditions in Northern Tanzania to support research and application development in computer vision, machine learning and digital crop health. The dataset comprises 155,842 labelled multi-season images belonging to three classes: healthy leaves, bean rust and bean anthracnose. Images were acquired from farms in Arusha, Kilimanjaro and Manyara regions using smartphone cameras and were subsequently reviewed and validated with support from agricultural extension officers and plant pathologists to improve annotation reliability. The dataset is organized into class-specific compressed files and is publicly available through the Zenodo repository. By providing a large field-based image resource captured under variable real-world conditions, this dataset can support the development, training and evaluation

* Corresponding author.

E-mail address: neema.mduma@nm-aist.ac.tz (N. Mduma).

Social media: [@nakadori](#) (N. Mduma)

Specifications Table

Subject	Computer Sciences
Specific subject area	Computer Vision for Detection of Common Bean Diseases (Bean Rust and Bean Anthracnose)
Type of data	Image Raw
Data collection	Data were collected using Samsung Galaxy M13 smartphones equipped with 13-megapixel cameras. The Open Data Kit (ODK) application was installed on the devices to capture common bean leaf images directly from the field. The raw images were categorized into three classes: healthy, bean rust and bean anthracnose. The data collection process involved researchers and farmers, while agricultural extension officers and plant pathologists conducted quality assurance and validation.
Data source location	• Institution: The Nelson Mandela African Institution of Science and Technology, Tanzania Agricultural Research Institute • City/Town/Region: Arusha • Country: Tanzania
Data accessibility	Repository name: Zenodo Data identification number: doi: 10.5281/zenodo.15685315 Direct URL to data: https://zenodo.org/records/15685315

1. Value of the Data

- This dataset provides a large, field-based collection of labelled common bean leaf images representing healthy leaves, bean rust and bean anthracnose from major bean-growing areas of Northern Tanzania. Because the images were captured under real farm conditions rather than in controlled laboratory settings, the dataset reflects practical variation in background, lighting, leaf orientation, symptom expression and image quality. This makes it practically useful for developing models intended for real-world agricultural applications.
- The dataset is valuable to researchers working in computer vision, machine learning, digital agriculture and plant pathology. It offers a locally relevant resource for training, validating and benchmarking image-based disease recognition models for common bean. Since many publicly available plant disease datasets are developed under controlled conditions or from regions outside East Africa, this dataset helps address the need for context-specific data that better represent the production environments in which diagnostic tools are expected to operate.
- The dataset can support a wide range of analytical and applied tasks such as image classification, comparative testing of deep learning architectures, transfer learning, model generalization studies and the development of mobile or web-based decision-support tools for disease diagnosis. It may also be useful for studies examining the effect of field variability on model performance, especially where robust disease detection under non-standardized conditions is required.
- This dataset can serve as a baseline resource for integrating additional disease classes, broader geographical coverage, symptom severity levels or complementary metadata in future studies. As an openly accessible dataset with a persistent repository link, it supports transparency, reproducibility and wider reuse in both methodological research and practical agricultural innovation.

2. Background

Common bean (*Phaseolus vulgaris*) is among the most widely cultivated grain legumes in Tanzania and across Sub-Saharan Africa, playing an important role in the livelihoods of smallholder farmers [1,2]. It is widely consumed in both rural and urban households and serves as an accessible source of dietary protein, minerals and income [3]. In many farming systems, common bean is cultivated either as a sole crop or in association with other staple crops, making it an important component of household nutrition and local market systems [4]. Its contribution to food security is especially important in regions where animal protein is less accessible or less affordable [5]. Despite these, common bean productivity is frequently constrained by pests and diseases, which continue to reduce yields and affect crop quality [6,7]. Among the major foliar diseases, bean rust and bean anthracnose are of particular concern because of their widespread occurrence and their potential to cause substantial production losses under favourable conditions [8]. These diseases affect leaf health and plant vigor and, when not managed in time, may reduce both yield and market value [9]. Their effects are often more serious in smallholder production systems where routine disease monitoring and timely technical support may be limited [10,11].

Accurate recognition of plant diseases at an early stage is essential for effective crop management. In practice, diagnosis in many farming areas relies mainly on visual inspection by farmers, extension officers or local experts [12,13]. This approach can be constrained by limited access to specialist knowledge, variation in field conditions and overlap in visible symptoms across diseases or stress factors [14]. These limitations have increased interest in digital approaches that can assist with crop health diagnosis in a faster and more standardized way [15]. Image-based diagnosis supported by artificial intelligence has emerged as a promising area in agricultural research, with computer vision methods showing strong potential for identifying visible disease symptoms from leaf images [16,17]. The usefulness of these methods, however, depends greatly on the quality, quantity and representativeness of the training data used to develop them [18]. Datasets collected in controlled settings may not adequately reflect the variability encountered in real farm environments, where differences in lighting, background, camera angle, symptom severity and field conditions can affect model performance [19]. For this reason, there is a growing need for datasets collected under natural production conditions and from locations where these tools are intended to be applied.

This work presents a labelled dataset of common bean leaf images collected from farms in Arusha, Kilimanjaro and Manyara regions of Tanzania. The dataset includes healthy leaves together with leaves showing symptoms of bean rust and bean anthracnose. It was developed to provide an openly accessible image resource that reflects field-based conditions and that can support future work in machine learning, computer vision and digital crop health applications. By describing the dataset and its organization, this article contributes to the availability of reusable data for common bean disease research and related technological development.

3. Data Description

This dataset consists of common bean leaf images collected from farms in Tanzania, including both healthy samples and those affected by bean rust and bean anthracnose. It was developed to support research on disease diagnosis and to enhance crop productivity through data-driven approaches. The dataset is suitable for a range of computer vision tasks, including image classification and object detection. All images are organized in compressed (.zip) files based on class categories. The bean anthracnose class is provided in “anthra.zip” containing 13,542 images. Bean rust images are distributed across two files, “rust1.zip” with 10,300 images and “rust2.zip” with 10,307 images. Healthy common bean leaf images are divided into eleven files, where “healthy1.zip” to “healthy10.zip” each contain 11,000 images, and “healthy11.zip” includes 11,693 images. In total, the dataset comprises 155,842 labeled leaf images.

Table 1
Number of common bean leaf images per categories.

Class name	Number of images
Healthy	121,693
Bean rust	20,607
Bean anthracnose	13,542

As shown in Table 1, the dataset exhibits class imbalance, with a substantially larger number of healthy images compared to diseased classes, particularly bean anthracnose. While this reflects real-world field conditions, it may influence model training and evaluation. Users of the dataset may therefore need to consider appropriate strategies such as resampling, class weighting or the use of balanced evaluation metrics when developing machine learning models.

Each ZIP archive contains images in JPEG (.jpg) format. In addition, image-level metadata is embedded within the files in EXIF format. This includes information such as image dimensions, device make and model, capture date and time and camera acquisition parameters (e.g., exposure time and focal length). Where available, location-related metadata may also be present. This embedded metadata enables users to access both visual and contextual information directly from the images, supporting reproducibility and extended analyses.

This dataset is distinct from existing image datasets in both domain and application. Unlike general-purpose or non-agricultural datasets (e.g., those used for tasks such as floor count estimation), it focuses specifically on common bean (*Phaseolus vulgaris*) leaves under real field conditions. The images capture natural variability in lighting, background and disease presentation, which enhances their relevance for practical agricultural applications. Furthermore, the dataset provides labeled examples of key diseases affecting common bean production, making it suitable for developing and evaluating machine learning models for plant disease detection and decision support in digital agriculture.

Although this article does not include model development or performance evaluation, the dataset has been structured to support machine learning applications such as crop disease detection. The availability of labeled images across multiple classes, along with real field variability, makes it suitable for training and benchmarking computer vision models. Users may apply standard machine learning and deep learning approaches to evaluate classification performance using this dataset.

4. Experimental Design, Materials and Methods

4.1. Field data collection

The dataset comprises common bean leaf images collected through a joint effort involving researchers and collaborators working in Northern Tanzania. Image acquisition was conducted using Samsung Galaxy M13 smartphones equipped with 13-megapixel cameras, with the Open Data Kit (ODK) application deployed to enable structured data capture in field settings (Fig. 1).

Data were collected over a one-year period, from January to December 2024, across three regions i.e., Arusha, Kilimanjaro and Manyara. These regions were selected based on their importance in common bean production and the observed prevalence of key diseases affecting the crop. The data collection process involved research assistants and smallholder farmers to ensure representation of real farming conditions. Images were primarily captured in situ under natural field conditions to reflect realistic disease presentation. Image resolution was determined by the device settings and maintained consistently across the data collection process. In some cases, multiple images of the same plant or leaf may be present due to repeated sampling under different angles or conditions.

Common Bean

* Type of Picture

☐ Healthy

☐ Bean rust

☐ Bean anthracnose

☐ Other

* Leaf Photo

Take Leaf Picture

Take Picture

Choose Image

Comment

Specify any observations (e.g. variety)

* Location

Please wait until phone gets reading

Start GeoPoint

NEXT >

Fig. 1. ODK form for collecting image data.

To ensure data quality and consistency, all collected images were reviewed and validated by agricultural extension officers and plant pathologists. The annotation process followed a structured workflow in which each image was independently assessed by at least two experts based on visible disease symptoms. A standardized symptom checklist, derived from established field diagnostic guidelines for common bean diseases was used to guide labeling decisions. In cases of disagreement, images were jointly reviewed and discussed until a consensus was reached. Where necessary, uncertain samples were excluded from the final dataset to maintain label reliability. This multi-step validation process helped ensure the accuracy and consistency of the assigned labels.

4.2. Data preprocessing

Following field collection, the common bean leaf images were subjected to a series of preprocessing steps prior to publication in an open-access repository (Fig. 2). The raw data captured were initially stored in cloud-based spreadsheets, where labels provided by field officers and domain experts were reviewed, verified and confirmed.

Preprocessing involved transferring the images from the cloud to a local server environment, where automated scripts developed in Python (within a Conda environment) were used to organize and prepare the dataset. Libraries such as Pandas were utilized to manage metadata and ensure correct mapping between images and their labels as part of the preprocessing and organization stage.

Several data quality challenges were addressed during preprocessing. These included handling incomplete or corrupted image files, inconsistencies in file naming and mismatches between image files and their corresponding metadata. Automated checks were applied to detect and remove unusable images, while standardized naming conventions were introduced

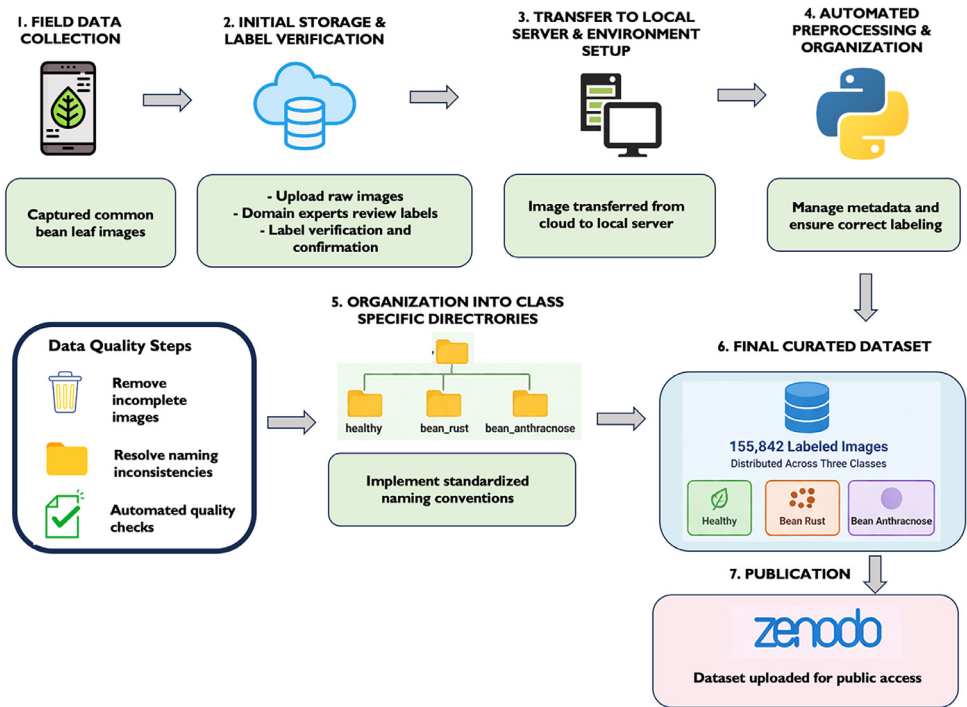


Fig. 2. Data preprocessing pipeline of the common bean imagery leaves.

to ensure consistency and traceability. A structured renaming process was implemented, linking each image to its metadata (e.g., GPS information) through generated reference files.

The processed images were then organized into class-specific directories namely healthy, bean rust and bean anthracnose, based on the validated annotations. The final curated dataset comprises 155,842 labeled images distributed across the three classes. The dataset was uploaded to a publicly accessible repository for wider use [20].

The preprocessing steps were designed to ensure data quality, consistency and usability for downstream applications such as crop disease detection. These steps focused on data cleaning, validation and structured organization rather than model development, in line with the objective of preparing a reliable dataset for reuse. Particular care was taken throughout preprocessing to maintain data integrity and ensure reliability for downstream computer vision applications.

Limitations

The dataset consists of common bean leaf images collected from selected regions in Northern Tanzania specifically Arusha, Kilimanjaro and Manyara. While these areas are important for bean production, the geographical coverage is limited and may not fully represent conditions in other parts of the country. In addition, the dataset focuses on only two major diseases i.e., bean rust and bean anthracnose which are known to significantly affect productivity in the selected regions. Other diseases that may impact common bean production are not included.

Although expert validation and quality control procedures were applied during annotation and preprocessing, the possibility of residual labeling errors cannot be completely eliminated. Furthermore, the dataset may contain variability associated with environmental conditions during image acquisition, differences in disease progression and changes in leaf appearance such as

partial drying after collection. Potential sampling biases related to specific farming practices or locations may also exist.

To ensure high data quality, only images captured under suitable lighting conditions were retained, while those taken in poor lighting environments were excluded. As a result, the dataset may have limited representation of challenging real-world conditions. Future work could extend the dataset by including additional disease categories and broader geographical coverage to improve dataset diversity and applicability.

Ethics Statement

Data were gathered from farms following a formal consent procedure. Participating farmers signed consent forms granting permission for data collection to take place on their fields.

CRedit Author Statement

Neema Mduma: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition; **Hudson Laizer:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, Project administration.

Data Availability

[A Labeled Dataset of Healthy and Diseased Common Beans from Tanzania \(Original data\)](#) (Zenodo)

Acknowledgements

The authors would like to thank the Grow Further for funding this project. Also, we wish to extend our gratitude to Mr. Deodatus Stanley Kiriba and Mr. Shida Mahenge Nestory from Tanzania Agricultural Research Institute (TARI), farmers, extension officers and researchers involved in this project for their assistance and technical advice during field data collection and annotation.

Declaration of Competing Interest

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

References

- [1] M.K. Muchomba, E.M. Muindi, N. Kandie, J. Ngeno, Overview of common bean (*Phaseolus vulgaris* L.), economic importance, ecological requirements, and production constraints in Kenya, *Int. J. Plant Soil Sci.* 38 (2026) 193–214, doi:[10.9734/ijpss/2026/v38i15939](https://doi.org/10.9734/ijpss/2026/v38i15939).
- [2] E. Mkuna, Determinants of horticultural export and welfare impact of smallholder farmers: evidence from common beans (*Phaseolus Vulgaris* L) farming in Arusha Tanzania, in: 2022: pp. 267–292. https://doi.org/10.1007/978-981-19-4211-2_12.
- [3] S. Lisciani, S. Marconi, C. Le Donne, E. Camilli, A. Aguzzi, P. Gabrielli, L. Gambelli, K. Kunert, D. Marais, B.J. Vorster, K. Alvarado-Ramos, E. Reboul, E. Cominelli, C. Preite, F. Sparvoli, A. Losa, T. Sala, A.-M. Botha, M. Ferrari, Legumes and common beans in sustainable diets: nutritional quality, environmental benefits, spread and use in food preparations, *Front. Nutr.* 11 (2024), doi:[10.3389/fnut.2024.1385232](https://doi.org/10.3389/fnut.2024.1385232).

- [4] M.A. Uebersax, K.A. Cichy, F.E. Gomez, T.G. Porch, J. Heitholt, J.M. Osorno, K. Kamfwa, S.S. Snapp, S. Bales, Dry beans (*Phaseolus vulgaris* L.) as a vital component of sustainable agriculture and food security—A review, *Legume Sci.* 5 (2023), doi:[10.1002/leg3.155](https://doi.org/10.1002/leg3.155).
- [5] C. Ojiewo, D.J.D.H. Keatinge, J. Hughes, A. Tenkouano, R. Nair, R. Varshney, M. Siambi, E. Monyo, N. Ganga-Rao, S. Silim, The role of vegetables and legumes in assuring food, nutrition, and income security for vulnerable groups in Sub-Saharan Africa, *World Med. Health Policy* 7 (2015) 187–210, doi:[10.1002/wmh3.148](https://doi.org/10.1002/wmh3.148).
- [6] I. Karavidas, G. Ntatsi, V. Vougeleka, A. Karkanis, T. Ntanasi, C. Saitanis, E. Agathokleous, A. Ropokis, L. Sabatino, F. Tran, P.P.M. Iannetta, D. Savvas, Agronomic practices to increase the yield and quality of common bean (*Phaseolus vulgaris* L.): a systematic review, *Agronomy* 12 (2022) 271, doi:[10.3390/agronomy12020271](https://doi.org/10.3390/agronomy12020271).
- [7] S. Wani, Q. Nisa, T. Fayaz, Naziya Nabi, Aasiya Nabi, I. Lateef, A. Bashir, R.J. Rashid, Z. Rashid, G. Gulzar, U. Shafi, Z.A. Dar, A.A. Lone, U.C. Jha, B.A. Padder, in: *An Overview of Major Bean Diseases and Current Scenario of Common Bean Resistance, Diseases in Legume Crops*, Springer Nature Singapore, Singapore, 2023, pp. 99–123, doi:[10.1007/978-981-99-3358-7_5](https://doi.org/10.1007/978-981-99-3358-7_5).
- [8] A. Mtonga, M.N. Maruthi, Diseases of common bean, in: 2024: pp. 1–52. https://doi.org/10.1007/978-3-030-35512-8_18-1.
- [9] R.K. Mishra, S. Pandey, Major fungal diseases of pulse crops and their management to mitigate its losses, in: *Diseases of Field Crops: Diagnostics and Management*, Springer Nature Singapore, Singapore, 2024, pp. 279–294, doi:[10.1007/978-981-97-6160-9_13](https://doi.org/10.1007/978-981-97-6160-9_13).
- [10] H.C. Laizer, M.N. Chacha, P.A. Ndakidemi, Farmers' knowledge, perceptions and practices in managing weeds and insect pests of common bean in northern Tanzania, *Sustainability (Switzerland)* 11 (2019) 4076, doi:[10.3390/su11154076](https://doi.org/10.3390/su11154076).
- [11] Z.J. Ngoya, A.G. Mkindi, S.J. Vanek, P.A. Ndakidemi, P.C. Stevenson, S.R. Belmain, Understanding farmer knowledge and site factors in relation to soil-borne pests and pathogens to support agroecological intensification of small-holder bean production systems, *Front. Sustain. Food Syst.* 7 (2023), doi:[10.3389/fsufs.2023.1094739](https://doi.org/10.3389/fsufs.2023.1094739).
- [12] U. Mwaibale, N. Mduma, H. Laizer, B. Mgawe, Enhancing detection of common bean diseases using Fast Gradient sign Method-trained Vision transformers, *Front. Artif. Intell.* 8 (2025), doi:[10.3389/frai.2025.1643582](https://doi.org/10.3389/frai.2025.1643582).
- [13] H. Laizer, N. Mduma, D. Machuve, R. Maganga, Common beans imagery dataset for early detection of bean rust and bean anthracnose diseases, *Data Brief.* 54 (2024) 110508, doi:[10.1016/j.dib.2024.110508](https://doi.org/10.1016/j.dib.2024.110508).
- [14] I. Buja, E. Sabella, A.G. Monteduro, M.S. Chiriaco, L. De Bellis, A. Luvisi, G. Maruccio, Advances in plant disease detection and monitoring: from traditional assays to In-field diagnostics, *Sensors* 21 (2021) 2129, doi:[10.3390/s21062129](https://doi.org/10.3390/s21062129).
- [15] M.E. Brown, S. Mugo, S. Petersen, D. Klausner, Designing a pest and disease outbreak warning system for farmers, agronomists and agricultural input distributors in East Africa, *Insects.* 13 (2022) 232, doi:[10.3390/insects13030232](https://doi.org/10.3390/insects13030232).
- [16] M. Shafay, T. Hassan, M. Owais, I. Hussain, S.G. Khawaja, L. Seneviratne, N. Werghi, Recent advances in plant disease detection: challenges and opportunities, *Plant Methods* 21 (2025) 140, doi:[10.1186/s13007-025-01450-0](https://doi.org/10.1186/s13007-025-01450-0).
- [17] M. Naveed, M. Majeed, K. Jabeen, N. Hanif, R. Naveed, S. Saleem, N. Khan, Plant disease diagnosis with artificial intelligence (AI), in: 2024: pp. 217–234. https://doi.org/10.1007/978-981-99-9621-6_15.
- [18] C.A. Elinisa, N. Mduma, Smart Agricultural Technology mobile-based convolutional neural network model for the early identification of banana diseases, *Smart Agric. Technol.* 7 (2024) 100423, doi:[10.1016/j.atech.2024.100423](https://doi.org/10.1016/j.atech.2024.100423).
- [19] K.G.F. da Silva, P.V.M. Rozatto, K.O. e Sousa, L.D. Hudson, A.W.S. Meyer, A.F. Silva, I.T. Mendes, A.R. Borges, L.E. Morais, L.M.S. Maciel, S.M. Villela, H. Pedrini, M.B. Vieira, Bean leaf image dataset annotated with leaf dimensions, segmentation masks, and camera calibration, *Data Brief.* 59 (2025) 111328, doi:[10.1016/j.dib.2025.111328](https://doi.org/10.1016/j.dib.2025.111328).
- [20] N. Mduma, H. Laizer, D. Kiriba, A Labeled Dataset of Healthy and Diseased Common Beans from Tanzania, *Zenodo*, 2025, doi:[10.5281/zenodo.15685315](https://doi.org/10.5281/zenodo.15685315).