Data in Brief 54 (2024) 110508

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Common beans imagery dataset for early detection of bean rust and bean anthracnose diseases

Hudson Laizer[a,*], Neema Mduma[b], Dina Machuve[c], Reinfrid Maganga[d]

[a] Mbeya University of Science and Technology, Department of Natural Sciences, P O Box 131, Mbeya, Tanzania
[b] The Nelson Mandela African Institution of Science and Technology, Department of Information and Communication Sciences and Engineering, P O Box 447, Tengeru, Arusha, Tanzania
[c] DevData Analytics Limited, P O Box 13855, Arusha, Tanzania
[d] Tanzania Agricultural Research Institute - Uyole, Department of Crop Protection, P O Box 400 Mbeya, Tanzania

ARTICLE INFO

Article history:
Received 22 November 2023
Revised 10 March 2024
Accepted 6 May 2024
Available online 11 May 2024

Dataset link: Common Beans Imagery Dataset for Early Detection of Crop Diseases (Original data)

Keywords:
Common beans
Bean anthracnose
Bean rust
Leaves
Image

ABSTRACT

Common bean plays a crucial role in the agricultural sector in Tanzania. To most smallholder farmers, the crop serves as a principal source of protein and an essential source of income. Despite its significance, common bean production is often affected by diseases, particularly bean rust and bean anthracnose, resulting in low yields and diminished economic returns. To address this challenge, a comprehensive dataset of common bean leaf images has been collected by using smartphone cameras to capture the visual characteristics of healthy and diseased leaves. The dataset contains more than 59,072 labeled images, offering a valuable resource for developing machine learning models and user-friendly tools capable of early detection and diagnosis of bean rust and bean anthracnose diseases. The aim of generating this dataset is to facilitate the development of machine learning tools that will empower agricultural extension officers, smallholder farmers, and other stakeholders in agriculture to promptly identify and diagnose affected crops, enabling timely and effective interventions before causing significant economic loss. By equipping farmers with the knowl-

* Corresponding author.
E-mail address: hudson.laizer@must.ac.tz (H. Laizer).
Social media: @hudson_laizer (H. Laizer), @nakadori (N. Mduma), @DMachuve (D. Machuve)

https://doi.org/10.1016/j.dib.2024.110508

edge and tools to combat these diseases, we can safeguard bean production, enhance food security, and strengthen the economic well-being of smallholder farmers in Tanzania and other parts of Africa.

Specifications Table

Subject	Applied Machine Learning
Specific subject area	Computer vision techniques for the detection of diseases affecting common beans specifically Bean Rust and Bean Anthracnose
Type of data	Image
How the data were acquired	Data were collected using Samsung Galaxy A03 Core 8MP camera. Open Data Kit (ODK) application was installed on the smartphones to capture images of common bean leaves in the field. Common bean image leaves were classified as either healthy or affected by bean rust or bean anthracnose. Data collection process involved research assistants and smallholder farmers while agricultural officers and plant pathologists were in charge of the data quality check.
Data format	Raw
Description of data collection	Images were collected in the field for seven months i.e. between October 2022 and April 2023. The target was to generate common bean imagery dataset with consideration of bean rust and bean anthracnose diseases which highly affect the productivity. All data samples were labelled to indicate whether they are healthy or affected by the labelled diseases.
Data source location	• Institution: Mbeya University of Science and Technology (MUST), Tanzania Agricultural Research Institute (TARI) • City/Town/Region: Mbeya • Country: Tanzania
Data accessibility	Repository name: Zenodo Data identification number: DOI: 10.5281/zenodo.8286125 Direct URL to data: https://zenodo.org/records/8286126

1. Value of the Data

- Common bean imagery dataset can be used to train machine learning models for early detection of bean rust and bean anthracnose diseases that cause significant loss to its production.
- Common bean leaf images dataset can be used by researchers in the field of machine learning to diversify options for early disease identification, disease diagnosis, and modeling disease spread.
- The created dataset serves as a valuable resource for facilitating development of machine learning models that can be deployed in mobile applications for example plant diseases finder (for banana disease diagnosis) which can be used by agricultural extension officers, farmers and other stakeholders in the early diagnosis of affected crops for early intervention [1].
- The dataset is among the openly accessible imagery datasets of common bean in Tanzania.

2. Objective

The development of a high-quality open-source dataset for the early detection of bean rust and bean anthracnose diseases affecting common bean is a significant step towards advancing machine learning research and addressing food security challenges in Tanzania. This dataset will

Fig. 1. Imagery sample of common bean leaves (a) healthy (b) bean rust (c) bean anthracnose.

provide researchers and developers with the necessary data to train and refine machine learning models capable of accurately identifying and classifying these diseases, enabling the development of effective real-world tools for crop disease management. The availability of this dataset is expected to accelerate the adoption of machine learning in agriculture, fostering innovation and contributing to the alleviation of food security issues in Africa. The dataset's versatility supports various computer vision tasks, including classification and object detection, further enhancing its potential for addressing crop disease challenges. Therefore, this open-source dataset represents a valuable resource for machine learning research and crop disease management, with the potential to significantly contribute to improving food security in Tanzania and beyond.

3. Data Description

The dataset comprises of common bean imagery leaves of both healthy and those affected by bean rust and bean anthracnose diseases collected from the farms in Tanzania. The dataset has a total of 59,072 labeled images in jpeg format. The images were labeled to indicate the name of the class based on the image number. To ensure accuracy and consistency in the labeling process, a systematic approach was employed which involved multiple rounds of annotation by trained researchers followed by a verification by agricultural extension officers and plant pathologists to address any discrepancies. Each image was cross-checked against established agronomic criteria to confirm the presence of disease symptoms. In cases of annotation disagreements, a panel of experts from TARI was consulted to reach a consensus. On data preparation, preprocessing steps were undertaken to maintain the quality and integrity of the dataset. This included cleaning and curating the data, wherein any irrelevant or low-quality images were removed to ensure a high standard of dataset fidelity. Special attention was given in maintaining uniformity of image characteristics such as resolution and orientation. In the repository, data were uploaded into 3 separate folders; 1 folder of healthy data submitted in zip format named healthy.zip, 1 folder of bean rust data submitted in zip format named rust.zip and 1 folder of bean anthracnose data submitted in zip format named anthra.zip. All folders were named to indicate its corresponding image class. Healthy folder contains all images of healthy common bean leaves, rust folder contains images of common bean leaves affected by bean rust and anthra folder contains images of common bean leaves affected by bean anthracnose. Data were separated in 3 zipped folders to allow easily downloading and uploading of data for model training. Fig. 1 shows imagery samples of common bean leaves available in the dataset.

4. Experimental Design, Materials and Methods

4.1. Field data collection

The dataset consists of images of common bean leaves collected by Mbeya University of Science and Technology (MUST) and Tanzania Agricultural Research Institute (TARI) Uyole cen-

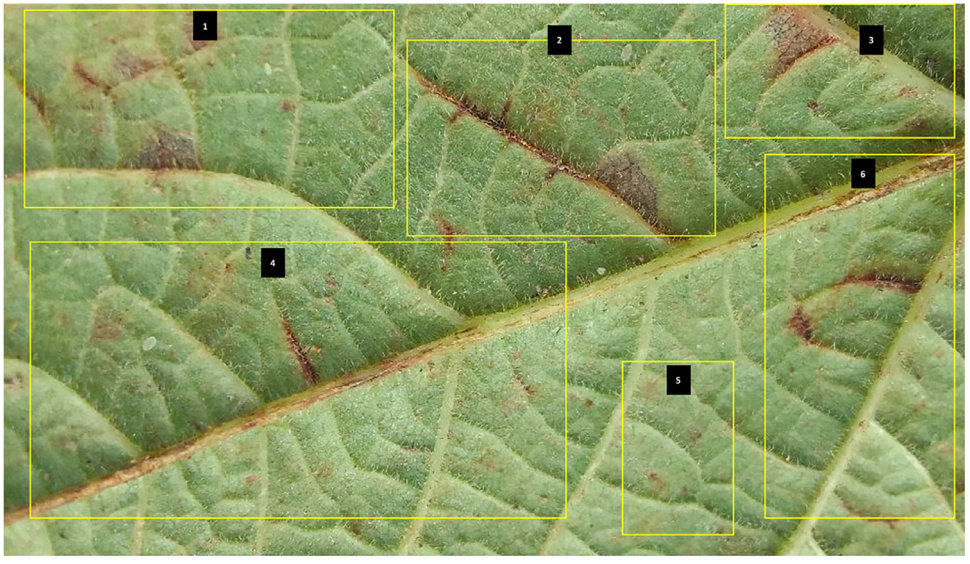

Fig. 2. Sample of the annotated common bean leaf.

tre located in Mbeya region, Southern Highlands of Tanzania. Open Data Kit (ODK) application installed in a Samsung Galaxy A03 Core smartphone was used to collect images of common bean leaves. Image data were collected from three regions i.e. Mbeya, Iringa and Ruvuma and was done by researchers and smallholder farmers, while data quality check was done by plant pathologists and agricultural extension officers. The data were collected for seven months (from 20th October 2022 to 10th April 2023) and encompasses images taken from 47 farms, distributed across three regions: 19 in Mbeya, 17 in Iringa, and 11 in Ruvuma. The farms were carefully chosen by the agricultural extension officers and farmers representatives based on the high incidence of the two diseases affecting common bean production in these regions [2].

4.2. Data preprocessing

Field data was first cleaned, labelled and annotated prior to being made available in a public repository. The data collected using the ODK tool was stored on Google Drive as Google sheets. During the dataset establishment, machine learning experts and agricultural extension officers were engaged in labeling crop leaf images at farm level. The data preprocessing involved transferring the leaf images to a local server, identifying and renaming incomplete images according to their respective classes. Python scripts were utilized in a Conda virtual environment, equipped with Python 3.9, GoogleDriveDownloader 0.4, and Pandas 1.3.5, to execute the preprocessing steps. Three directories were established on the server (Ubuntu 20.04, 1 Tb disk, 8GB RAM) and original sized images were sorted and stored into these folders based on labels from the Google sheet. Then another Python script was executed on the dataset in each folder to detect incomplete files. A total of 57 incomplete files were detected and removed from the folders. The remaining images were then annotated for computer vision image classification (Fig. 2). In the final preprocessing phase, images in each class folder were renamed in a format combining the class name and a number (e.g., “rust567.jpg” or “anthra88.jpg”). The renaming script also generated a csv file listing both original and new filenames, linking the images to additional metadata like GPS coordinates. The preprocessing scripts are available at

Table 1
Common beans dataset after preprocessing.

Class name	Images preprocessed
Healthy	24,973
Bean rust	20,568
Bean anthracnose	13,531

https://github.com/devdatanalytics/ai4afsmust. Table 1 shows the count of preprocessed common bean images.

Images were grouped into three classes; healthy, bean rust and bean anthracnose. Then the curated images were uploaded in the Zenodo open repository [3].

Limitations

This article features a dataset comprising images of common bean leaves, gathered specifically from the three regions (Mbeya, Iringa, and Ruvuma) from Southern Highlands Tanzania. However, the scope of the dataset is confined to two specific diseases: bean rust and bean anthracnose that were reported to affect productivity in the three regions. The primary objective was to maintain high-quality imagery in the dataset, thus only images captured in well-lit conditions have been included in the dataset, while those taken under low-light have been excluded. Future studies could consider expanding to other regions in Tanzania that were not included in this data collection and could also incorporate other common bean diseases that affect productivity.

Ethics Statements

The study does not involve experiments on humans or animals.

Data Availability

Common Beans Imagery Dataset for Early Detection of Crop Diseases (Original data) (Zenodo).

CRediT Author Statement

Hudson Laizer: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition; **Neema Mduma:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration; **Dina Machuve:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration; **Reinfrid Maganga:** Conceptualization, Methodology, Supervision.

Acknowledgments

This research project was financially supported by the International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (SIDA) through the Artificial Intelligence for Agriculture and Food Systems Innovation Research Network (AI4AFS-IRN) administered by the African Technology Policy Studies Network (ATPS) with Grant Award Number: AI4AFS/GA/AFS-2504001568.

The authors acknowledge the project partners; Tanzania Agricultural Research Institute (TARI), Makerere University AI Lab and Southern Corridor Alliance of Agriculture Producers (SCAAP).

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] 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 (2023) 100423.
[2] M. Beatrice, et al., Viruses infecting common bean (Phaseolus vulgaris L.) in Tanzania: a review on molecular characterization, detection and disease management options, Afr. J. Agric. Res. 12 (18) (2017) 1486–1500.
[3] H. Laizer, Common beans imagery dataset for early detection of crop diseases, Zenodo (2023), doi:10.5281/zenodo.8286126.