Data Sheet for Beans Image Data Collected in Uganda Under the Lacuna Project

We present the Lacuna Beans data sheet created by Makerere Artificial Intelligence Lab created by a group of researchers from the Makerere Artificial Intelligence Lab in Makerere University in Uganda. We follow the datasheet for dataset framework created by (Gebru et al. 2021).

Motivation	
For what purpose was the data set created? Was there a specific task in mind?	The dataset was created to provide an open, well-labelled, sufficiently curated and accessible *beans image dataset*. Data scientists, researchers, and the broader machine learning community can use the dataset for various machine learning experiments to build beans crop disease diagnosis and spatial analysis solutions.
Was there a specific gap that needed to be filled?	Despite the fact that the agricultural sector is a national economic development priority in sub-Saharan Africa, crop pests and diseases have been the challenge affecting major food security crops like beans. In 2020, a study conducted on bean rust in Uganda showed that the disease resulted in an estimated loss of 5% to 67% in the six varieties considered for the study (Odogwu and Rubaihayo 2021). In 2017 a study on Angular Leaf Spot (ALS) and its sources in the sub-Saharan African region showed that ALS contributed an extreme yield loss estimated at 384.2 tons per year in the whole region (Kijana et al. 2017). The current state of data collection and crop pest and disease diagnosis is transitioning from disease identification using visible symptoms to the use of data-driven solutions applying machine learning and computer vision techniques. Smallholder farmers and agricultural experts are equipped with mobile phones loaded with software to automatically collect field-level Geo-coded and time-stamped data. However, the image data previously collected has not been well-curated and shared with the wider machine learning community.
Who created the dataset?	The dataset was created by a team of scientists from the Makerere Artificial Intelligence Lab, Marconi Society Machine Learning Laboratory Lab, and the National Crops Resources Research Institute (NaCRRI) in active collaboration. NaCRRI is an institute of the National Agricultural Research Organization (NARO) in charge of crop research in Uganda.

Who funded the creation of the dataset?	This work was carried out with support from Lacuna Fund, an initiative cofounded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. The views expressed herein do not necessarily represent those of Lacuna Fund, its Steering Committee, its funders, or Meridian Institute.: 0328-S-001.
Composition	
What do the instances that comprise the dataset represent?	The dataset includes bean crop trifoliate images, each instance includes the trifoliate image accompanied by the image category, i.e., Healthy, Bean Rust, Angular Leaf spot, and a group of attributes associated with the crops from which the image was taken.
How many instances are there in total?	The dataset contains 5031 Angular Leaf Spot images, 5020 Bean Rust images, and 5284 healthy images. The total number of image instances in the dataset is 15,335.
Does the dataset contain all possible instances or is it a sample of instances from a larger set?	The dataset consists of bean image crops spread collected across the different regions in Uganda. Data were collected by random sampling from the areas where bean crop farming is practiced; these areas were identified by the experts from NaCRRI. A few samples were collected from the identified areas to generate a dataset that represents the overall bean farming in the country.
What data does each instance consist of? "Raw" data or features?	Each instance includes a trifoliate image, accompanied with attributes; the crop variety, plant age, district, sub-county, the GPS location, GPS accuracy, and the date of image capture.
Is there a label or target associated with each instance?	Each instance is associated with a class label to assert whether or not it was taken from a diseased crop. The labels are Healthy, Angular Leaf Spot, and Bean Rust as shown in Figure 1.
Is any information missing from individual instances?	None
Are relationships between individual instances made explicit?	There are no relationships between the different image instances in the dataset.
Are there recommended data splits?	None
Are there any errors, sources of noise, or redundancies in the dataset?	None
Is the dataset self-contained?	Yes. The dataset does not rely on any external sources, it is sufficient on its own.
Does the dataset contain data that might be considered confidential?	No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?	No
Does the dataset relate to people?	No
Collection Process	
How was the data associated with each instance acquired?	The beans image data was collected using mobile phones from bean farmer gardens.
What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated?	The data was collected using the Adsurv application, which is a mobile application that enables crowdsourcing of crop disease data from farmers' gardens. Android devices were available for data collection. Each device was configured with Adsurv for a specific data collector.
Who was involved in the data collection process?	The overall data collection exercise was conducted by a team including researchers from the Makerere Artificial Intelligence research Lab, the Marconi Society Machine Learning Laboratory at Makerere University and agricultural experts from Legumes program at the National Crops Resources Research Institute and an agricultural extension worker. The extension worker enabled us to bridge the language gap between the data collectors and the on-ground farmers in the different regions of the country.
Over what timeframe was the data collected?	April 2021 & May 2021
Does this time frame match the creation time frame of the data associated with the instances?	Yes. The time frame matches the creation time frame of the data associated with the instances.
Were any ethical review processes conducted?	None
Does the dataset relate to people?	None
Preprocessing, cleaning, and labelling	

Was any preprocessing/cleaning/labeling of the data done?	For quality assurance data cleaning was done taking into consideration the missing values, defective images, resolution of inconsistencies, and removal of outliers. During data collection process, the data collectors had to manually input some values for predefined attributes; for example, crop variety, however, some of these attributes changed. Some of the images were taken in the middle of the day which led to an overexposure defect as a result of too much light hitting the camera sensor. This affected the quality of the images making it difficult to see the disease symptoms. These images were preprocessed to minimize the overexposure defect as shown in Figure 2. The data was labeled using a custom web tool built on top of the VIA annotation tool (Abhishek 2021).
Was the "raw" data saved in addition to the preprocessed/cleaned/ labeled data (for example, to support unanticipated future uses)?	The raw unprocessed is stored locally on data storage servers in the Makerere Artificial Intelligence Lab.
Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.	The link to the annotation tool is available: `https://github.com/AI-Lab-Makerere/web-annotation-tool`
Uses	
Has the dataset been used for any tasks already?	Yes, we have used the dataset to build baseline disease classification models.
Is there a repository that links to any or all papers or systems that use the dataset?	No.
Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?	None.
Are there tasks for which the dataset should not be used?	None.
Distribution	
Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created?	Yes, the dataset will be made publicly available.

How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	The beans dataset and the associated metadata are stored on the Harvard DataVerse; an open-source data repository software that accepts a wide range of data types in different formats. This data was assigned a Digital Object Identifier `https://doi.org/10.7910/DVN/TCKVEW`.
When will the dataset be distributed?	The dataset is available under the specified DOI.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? None.	The dataset is licensed under the CC BY license that allows users to share and adapt the dataset so long as they give credit to data set creators Have any third parties imposed IP-based or other restrictions on the data associated with the instances?
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	None.
Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The dataset will be maintained by the research team at the Makerere Artificial Intelligence Lab. The team will support, host, and maintain the dataset.
How can the owner/curator/ manager of the dataset be contacted (for example, email address)?	The dataset manager can be contacted via email.
Is there an erratum?	No.
Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)?	All updates to the dataset will be documented and communicated through the Makerere AI Lab GitHub repository.
Will older versions of the data- set continue to be supported/hosted/ maintained?	Yes, the older versions will be stored locally on data storage servers in the Makerere Artificial Intelligence Lab and on remote data storage buckets on the Google cloud.
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	Interested researchers can send an email to data managers manager one and manager two to discuss the dataset extension and contribution.

Figure 1: Beans Data Class Labels.

Figure 2: On the left is the trifoliate image after preprocessing, on the right is the original trifoliate image with the overexposure defect.

References

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

Odogwu, B. A., Nkalubo, S. T., Rubaihayo, P. (2021). Yield loss of common bean caused by rust disease in Uganda. Journal of Underutilized Legumes, 3(1), 1-12.

Kijana, R., Abang, M., Edema, R., Mukankusi, C., Buruchara, R. (2017). Prevalence of angular leaf spot disease and sources of resistance in common bean in eastern Democratic Republic of Congo. African Crop Science Journal, 25(1), 109-122.
Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 4 pages.