Data Sheet for Maize Image Data Collected in Tanzania Under the Lacuna Project

We present the Lacuna Maize data sheet created by a group of researchers from The Nelson Mandela African Institution of Science and Technology and the Tanzania Agricultural Research Institute (TARI) in Tanzania. We follow the datasheet for dataset framework created by (Gebru et al. 2021).

Motivation	
For what purpose was the data set created?	The dataset was created to provide an open, well-labelled, sufficiently curated and accessible *maize image dataset*. Data scientists, researchers, and the broader machine learning community can use the dataset for various machine learning experiments to build maize crop disease diagnosis and spatial analysis solutions.
Was there a specific task in mind?	Although 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 Maize. The images target the diagnostics of Maize Lethal Necrosis (MLN) and Maize Streak Virus (MSV) diseases. We are motivated in developing end to end tools to help farmers diagnose diseases and also help to improve the production of maize in the country.The current state of data collection and crop pest and disease diagnosis is transitioning from disease identification using visible symptoms to using data-driven solutions applying machine learning and computer vision techniques. The image data previously collected has not been sufficiently curated, prepared, and shared with the broader community.
Who created the dataset?	The dataset was created by Scientists and Masters students from the Nelson Mandela African Institution of Science and Technology and the Tanzania Agricultural Research Institute (TARI) in Tanzania.
Who funded the creation of the dataset?	This work was carried out with support from Lacuna Fund, an initiative co-founded 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?	Each instance in the dataset consists of a crop image with an image status, i.e., Healthy, Maize Lethal Necrosis, and Maize Streak Virus, crop variety, crop age, and location (district, sub-county).
How many instances are there in total (of each type, if appropriate)?	The dataset consists of 17,277 labeled images where Healthy are 5542, Maize Lethal Necrosis are 5068, and Maize Streak Virus are 6667.)
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	The dataset has image samples collected from significant maize growing districts selected with the guidance of agricultural experts to obtain a representative dataset.
What data does each instance consist of? "Raw" data or features?	The data consists of raw image data.
Is there a label or target associated with each instance? If so, please provide a description.	Each instance is associated with a class label based on the status of the crop: healthy or diseased. The given labels per image are: Healthy, Maize Streak Virus and Maize Lethal Necrosis 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 (for example, training, devel- opment/validation, testing)?	We do not specify any data splits.
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.	None
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The dataset is self-contained.
Does the dataset contain data that might be considered confidential?	No.
Does the dataset contain data that, if viewed directly, might be of- fensive, insulting, threatening, or might otherwise cause anxiety?	No.
Collection Process	

How was the data associated with each instance acquired?	The maize leaf images for this dataset were collected using cameras on mobile phones from farmer's gardens. The AdSurv mobile application installed on Samsung phones was used to take photos of maize leaves.
What mechanisms or procedures were used to collect the data?	The data was collected using the Adsurv application, which is a mobile application that enables crowdsourcing of crop disease data from farmers' gardens. Adsurv application was installed on mobile phones/tablets used during the data collection process.
If the dataset is a sample from a larger set, what was the sampling strategy?	The dataset is not from a larger set.
Who was involved in the data collection process?	The data was collected by a team of researchers and students from the The Nelson Mandela African Institution of Science and Technology and Tanzania Agricultural Research Institute (TARI) .
Over what timeframe was the data collected?	The data was collected in the range of six months, periodically between February 2021 and July 2021 from farms in Arusha, Kilimanjaro and Manyara regions in Tanzania.
Were any ethical review processes conducted (for example, by an institutional review board)?	No.
Preprocessing, cleaning, and labelling	

Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?	The following steps were taken to process the data: 1. Gathering raw images: First the maize leaves images were collected using the AdSurv application installed on smartphones. 2. Eliminating duplicate images: The identified duplicate images were removed but a very small number (that were not initially identified) might still exist in the dataset. 3. The number of remaining duplicates should be small enough so as not to significantly impact training/testing. The dataset contains distinct images that are not defined to be duplicates but are extremely similar. 4. Labeling: The images were labeled to indicate the belonging class (healthy, Maize Lethal Necrosis, Maize Streak Virus). 5. Curation: The images were annotated for various computer vision tasks such as image classification, image object detection and image segmentation. 6. Renaming: The images in each class were renamed to comprise image number
Was the "raw" data saved in addition to the preprocessed/cleaned/ labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.	The raw unprocessed data 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? If so, please provide a description.	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.

What (other) tasks could the dataset be used for?	The dataset can be used for building object detection, segmentation, and time-series analysis models.
Is there anything about the composition of the dataset or the way it was collected and preprocessed/ cleaned/labeled that might impact future uses?	No.
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? If so, please provide a description.	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 dataset and the associated metadata are stored on the Harvard DataVerse which is an open-source data repository. The dataset is assigned a Digital Object Identifier: `https://doi.org/10.7910/DVN/GDON8Q`.
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)?	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?	No.
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	No.
Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The dataset will be maintained by the research team at the Makerere Artificial Intelligence Lab and The Nelson Mandela African Institution of Science and Technology . 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? If so, please describe how.	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: Maize Data Labels.

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.