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

We present the Cassava Image data sheet 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?	The dataset was created to provide an open and accessible *cassava dataset* with well-labeled, sufficiently curated, and prepared cassava crop imagery. Data scientists, researchers, and the broader machine learning community can use the dataset for various machine learning experiments to build cassava crop disease diagnosis and spatial analysis solutions.
Was there a specific task in mind?	Previously published cassava image datasets have had issues with class mislabeling and real world application for infield usage (Saleem et. al. 2020) due to the limitation with laboratory settings where the images were captured. There was a need to come up with a high quality and well labeled real-world cassava image dataset that depicts different image crop conditions including: Healthy and diseased and these images had to be taken in actual farmer gardens. 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 scientists from the Makerere Artificial Intelligence Lab and the National Crops Resources Research Institute (NaCRRI) in Namulonge, Uganda. NaCRRI is an institute of the National Agricultural Research Organisation (NARO) in charge of crop research.

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?	Each instance in the dataset includes crop image (JPEG), the crop image status (Healthy, Cassava Mosaic Disease and Cassava Brown Streak Disease), crop variety, crop age, location (district, sub-county).
How many instances are there in total (of each type, if appropriate)?	The dataset contains 15000 images in total. Each class that is: Healthy, Cassava Mosaic Disease and Cassava Brown Streak Disease has 5000 images.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	A Sample. The dataset contains different instances that were captured across the Central, Eastern, Northern and Western regions of Uganda. A few samples were collected from specific cassava prominent growing districts that were selected by the agricultural experts within these regions to create a representative dataset. The initial plan was to have an equal representation from each region however this was not the case because for some regions we visited fewer districts. The reason for this is that there was less data because some of the districts are mainly urban centers whilst others there was minimal diseased data because of the change in the season.
What data does each instance consist of? "Raw" data or features?	The data consists of raw image data. Each image data point is accompanied with attributes; the crop variety, plant age, district, sub-county, the GPS location, GPS accuracy, and the date of recording.
Is there a label or target associated with each instance? If so, please provide a description.	The image has an associated annotation file showing the sub-county where the image was captured, disease class, age of plant, variety and date when the data was collected. The given labels per image are: Healthy, Cassava Mosaic Disease and Cassava Brown Streak Disease as shown in Figure 1.
Is any information missing from individual instances?	For the data collected in May 2021, some of the instances do not have the specific name of the local variety in which case this was represented as local. This was because the agricultural experts could not easily tell what the exact variety of the cassava was. All the data collected earlier than May 2021 does not include age and variety data as this was not requested during that period of data collection.
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, development/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 Cassava image data was collected using mobile phones from farmer gardens. The gardens were identified within the different cassava growing prominent districts across the four regions in Uganda.For some of the data instances, the location attribute was generated by analyzing the GPS coordinate to identify the sub county.
What mechanisms or procedures were used to collect the data?	The images were captured using android smartphones running a form-based application supported by ODK[1]. The data collection forms were designed, validated and tested in coordination with domain agricultural experts to ensure that the right attributes were well represented and captured in the application. Sample data collection activities were conducted to validate and improve the data collection forms. Some inbuilt validations were developed within the data collection forms to improve the quality of the data. These include type validation for some attributes like numbers, setting required fields and using predefined values for selection of the diseases. A different account was also created for each data collector to ensure easy tracking and containment of errors. The data was collected using the Adsurv mobile application, a software program that uses a module that enables the crowdsourcing of crop disease surveillance data from farmers' gardens. Adsurv application was installed on mobile phones/tablets used during the data collection process.

[1]https://opendatakit.org

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?	A team of researchers from the Makerere Artificial Intelligence Lab, an agricultural expert from the cereals program at National Crops Resources Research Institute (NaCRRI), and a district agricultural officer. The agricultural officer 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?	The data was collected over 7 days in May 2021. Part of the data was also collected in 2020.
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)?	Cleaning of the data was done after data collection to rectify some errors that were created during collection. This included visually checking for images to ensure that they all showed the objects of interest which are the cassava leaves and all other unrelated objects were discarded. Verification of the variety names was also done and some of the variety names were corrected to reflect the correct and consistent spellings through consultation with the agricultural experts at the Namulonge Agricultural Research Institute. 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? 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/T4RB0B`.
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. 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: Image samples from the cassava dataset showing healthy image on the left, Cassava Brown Streak Image in the middle and Cassava Mosaic image on the right.

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.

Saleem, Muhammad Potgieter, J. Arif, Khalid. (2020). Plant Disease Classification: A Comparative Evaluation of Convolutional Neural Networks and Deep Learning Optimizers. Plants. 9. 1319. 10.3390/plants9101319.

Badu-Apraku, B., Bankole, F. A., Ajayo, B. S., Fakorede, M. A. B., Akinwale, R. O., Talabi, A. O., ... Ortega-Beltran, A. (2021). Identification of early and extra-early maturing tropical maize inbred lines resistant to Exserohilum turcicum in

sub-Saharan Africa. Crop Protection, 139, 105386.

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.