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

We present the Lacuna Maize data sheet created by by a team of data scientists, agricultural scientists and agricultural officers from the KaraAgro AI Foundation, with support from University of Ghana - Crop Science Department and the Ministry of Food and Agriculture, the Republic of Ghana. 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?	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 maize. Fall armyworm affects 44 countries in Africa. In new household surveys in Ghana and Zambia, 98% of farmers reported maize to be affected. The average maize loss reported by farmers in Ghana was 26.6% and in Zambia 35%. Extrapolating these losses nationally gives an estimate of USD 177m lost value of the annual maize crop in Ghana by (Rwomushana et al. 2018). According to research, Maize Streak Disease which is caused by the Maize Streak Virus is regarded as the third most serious disease affecting maize in sub-saharan Africa (Martin 2009). The prominence of these diseases has greatly affected the yields of Africa's most important food crop (Martin 2009). 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. The image data previously collected is biased and not reproducible It has also not been sufficiently curated, prepared, and shared with the wider community.

Who created the dataset?	The dataset was created by a team of data scientists, agricultural scientists and agricultural officers from the KaraAgro AI Foundation, with support from University of Ghana - Crop Science Department and the Ministry of Food and Agriculture, the Republic of Ghana.
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, image status (Healthy, Fall armyworm, and Maize Streak Virus), crop variety, location (GPS coordinates).
How many instances are there in total (of each type, if appropriate)?	Healthy - 5,392, Fall Armyworm - 5,110 Maize Streak Virus - 5,063
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	A sample. The data set contains different instances that were captured across all the 16 regions of Ghana.
What data does each instance consist of? "Raw" data or features?	Each instance includes: the crop image, image status(Healthy, Maize Streak Virus, Fall Armyworm), crop variety, location (gps coordinates).
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 i.e. diseased or healthy. The labels are Healthy, Maize Streak Virus and Fall Armyworm respectively 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 data associated with each instance was acquired from different maize farms across Ghana. These were different farms that were identified within the maize growing prominent districts of the 16 regions across the country.
What mechanisms or procedures were used to collect the data?	The data was collected using the KaraAgro AI - Collect app, a software program that uses a module that enables crowdsourcing of crop disease surveillance data from farms. This application was installed on Android devices. Data collectors used these Android devices to collect the maize data.
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 team involved in the data collection process is specified below: All sixteen regional agricultural officers from the Ministry of Food and Agriculture in Ghana.District agricultural officers and extension officers supporting regional agricultural officers in each district visited. On-ground farmers in the different districts whose farms we visited for data collection.
Over what timeframe was the data collected?	This data was collected in a period of 4 months from May 2021 to August 2021.
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)?	None
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 (consisting of labeled images) has been saved.
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://www.makesense.ai`
Uses	
Has the dataset been used for any tasks already? If so, please provide a description.	No, the dataset has not yet been used for any tasks.
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?	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? 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/CXUMDS`.
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 KaraAgro AI. 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)?	The image data will be stored initially on Google Drive because it provides the ability to scale when the size of the data increases. The final datasets will be stored both locally on data storage servers within the institutions and remote data storage buckets on Google cloud. These will act as intermediate storage solutions for the data during collection and curation and will provide redundancy to reduce risks of data loss.
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 KaraAgro AI and on remote data storage buckets on 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.

Rwomushana, I., Bateman, M., Beale, T., Beseh, P., Cameron, K., Chiluba, M., Clottey, V., Davis, T., Day, R., Early, R., Godwin, J., Gonzalez-Moreno, P., Kansiime, M., Kenis, M., Makale, F., Mugambi, I., Murphy, S., Nunda. W., Phiri, N., Pratt, C., Tambo, J. Fall armyworm: impacts and implications for Africa. Evidence Note Update, October 2018.

Martin, D. P., & Shepherd, D. N. (2009). The epidemiology, economic impact and control of maize streak disease. Food Security, 1(3), 305-315.