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

We present the Cassava Spectral and Image 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?	The dataset was created to provide an open and accessible spectrometry dataset with well-labeled, sufficiently curated, and prepared cassava crop spectral data that will be used by data scientists, researchers, the wider machine learning community, and social entrepreneurs within Sub-Saharan Africa and worldwide for various machine learning experiments so as to build solutions towards infield for early disease diagnosis.

Was there a specific task in mind?	The current state of disease identification is through visible symptoms. However, for this method, once symptoms have manifested on the aerial part of the plant e.g. diseases in cassava, the root which is the edible part of the plant could have been totally destroyed. This dataset considered the use of visible and near infrared spectral information for early disease detection thus allowing early action measures. The spectral dataset defined in this datasheet came from two experiments: (a) the screenhouse and (b) open field. The screen house setup rules out the influence of other diseases, pests or severe weather conditions while in an open field, crops grow under a natural environment, also exposed to crop pests. Healthy cassava stems were identified from clean cassava gardens by the agricultural experts. The plants were distributed across three varieties (NAROCass, TME14 and a local variety "Kwatamubale"). These varieties were chosen on the basis of being tolerant (NAROCass), susceptible (TME14, Kwatamubale) to Cassava Brown Streak Disease (CBSD). Initially, planting materials were thoroughly cleaned, which included the sterilization of the soil to ensure that no gaps led to disease transmission. At week four of growth, these plants were inoculated with Cassava Brown Streak Disease and Cassava Mosaic Diseases while maintaining a section as a healthy control. The disease inoculation process follows the procedures defined in our initial work in (Owomugisha et al. 2021). The confirmation of a successful transmission of the target viruses can be determined in two ways: molecular-based disease detection and PCR-based disease detection techniques (Mbanzibwa et al. 2021). Spectral data, chemical lab data, leaf image data and expert scoring were captured for a period of 19 and 15 consecutive weeks of disease monitoring for the screenhouse and open field respectively. The data collection process ended when the disease symptoms started showing in some plants visibly seen by the human eye.
Who created the dataset?	The dataset was created by scientists from the Makerere Artificial Intelligence Lab and the National Crops Resources Research Institute (NaCCRI) in Namulonge, Uganda. NaCCRI 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?	The dataset is composed of two experiments: screenhouse and open field experiment. Each experiment contains the following data types. 1. Spectral data. The spectrograms were acquired by a handheld spectrometer. 2. Leaf image data. For each plant, a corresponding image was acquired by a smartphone camera. 3. Biochemical data. Lab chemical data as a ground truth on disease propagation. 4. Expert scoring. At the same, plants were scored every week by the agricultural experts on visual symptoms.
How many instances are there in total (of each type, if appropriate)?	Yes, the dataset involves multiple instances. For a particular plant, a spectrogram was captured. At the same time, a leaf image was captured with a phone camera. Biochemical analysis was provided. Finally, the expert provided a score on whether the plant was diseased or healthy based on visual analysis. The cassava leaf has multiple lobes and in our experiments, we sampled the middle lobe. For each leaf, two readings were taken on the middle leaf lobe: one taken as Reading 1 (not visibly showing symptoms) and the other as Reading 2 (part showing visible disease symptoms) as shown in Figure 1. The number of instances are summarised in Table 2.
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	All instances of the dataset are presented as per designed data collection protocol. E.g spectral data, corresponding leaf image, chemical lab data and expert score to represent a plan.
What data does each instance consist of? "Raw" data or features?	Each instance consists of spectral data points (raw and preprocessed data), plant leaf image, chemical data and the expert score.

Is there a label or target associated with each instance? If so, please provide a description.	The data was collected using the following naming type. E.g **A 1 HLT 1 R a 1g** where, the *variable A* stands for A stands for cassava variety. We had three varieties (A, B and C). Variety A represents NAROCass, B represents TME14 and C represents a local variety "Kwatamubale". The *variable 1* means week 1 of data collection. HLT stands for the disease class. The dataset is represented by three classes (HLT meaning healthy control class, CBSD and CMD cassava diseases). The disease class is followed by *variable 1* stands for plant number, and *variable R* stands for reflectance mode. After the mode, the variable a stands for leaf, data is collected from three leaves per plant, that is from leaf a, b and c. The last variable "1g" is treated as a constant or placeholder. Table 3 gives a summary of the naming format used.
Is any information missing from individual instances?	A few plant leaf images are missing in the dataset. This error was caused at the data collection point during image upload to the server and the error is less than 1%. However, a single datapoint should have a spectral point, its corresponding leaf image, chemical data, and expert score.
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)?	The nature of this dataset involves picking data from the same plants over time. However, training and testing on data from the same plants is not proper for validation of the model. By the naming format given, the user should know the plants that are used in training and those that are used in testing. Thus Leave-One-Plant-Out validation or similar schemes would be appropriate for this dataset.
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 offensive, insulting, threatening, or might otherwise cause anxiety?	No.
Collection Process	

How was the data associated with each instance acquired?	All the instances associated with the data were acquired. That is spectral data, leaf image data, chemical data and expert score. However, with spectral data, several instances/indexes can externally be calculated from the spectral bands (L3Harris).
What mechanisms or procedures were used to collect the data?	The data set was validated as described below: 1. Spectral data: Spectral data was collected using visible and near-infrared handheld spectrometer; CI-710 MINIATURE LEAF SPECTROMETER (CIDBioscience). 2. Leaf Image data: Image data was captured using a smartphone camera of a 12-megapixel. 3. Chemical lab data: The biochemical data was analyzed using a 96-well Real-Time PCR instrument; StepOnePlus™ Real-Time PCR System (Thermo). 4. Expert score: Agricultural experts score plants as healthy or diseased based on visual symptoms that appear on the plants. Validation procedure: For the spectral and chemical lab data, the devices are calibrated using their standard measure as given in their manuals.
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: Research Scientists and graduate students at the Makerere Artificial Intelligence Lab at Makerere University, Kampala-Uganda. Research Scientists at National Crops Resources Research Institute (NaCCRI).
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 dataset holds both raw information and preprocessed data. The data pre-processing, cleaning and annotation process involved: checking missing values, aggregating files and ensuring that all spectral data points have corresponding values. Data continuity was key in this data collection process where each week, all the four data instances were captured from a given plant. Errors pointing to missing data could arise if a plant has been discontinued or dried up. Therefore, this process was able to detect such plants were eliminated from the dataset. The end of this process is a one csv file that holds spectral data concatenated with its corresponding chemical lab data, expert score, plant image name, and a set of labels at the end.
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.	Yes, both raw data and pre-processed data have been provided.
Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.	No
Uses	
Has the dataset been used for any tasks already? If so, please provide a description.	No, this particular dataset has not been used. However, a similar dataset that was collected in previous experiment [1] [2].
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?	No.
Distribution	

[1] https://github.com/godliver/IEEE$_A CCESS_C assava_S pectral_D ata$

[2] doi: 10.1109/ACCESS.2021.3087231

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/R0KL7R`.
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 should send an email to manager one owned by Owomugisha Godliver or to to discuss the dataset extension and contribution.
Funding	This work was carried out with support from: (i) 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. (ii) Data Science Africa Research Award 2020.

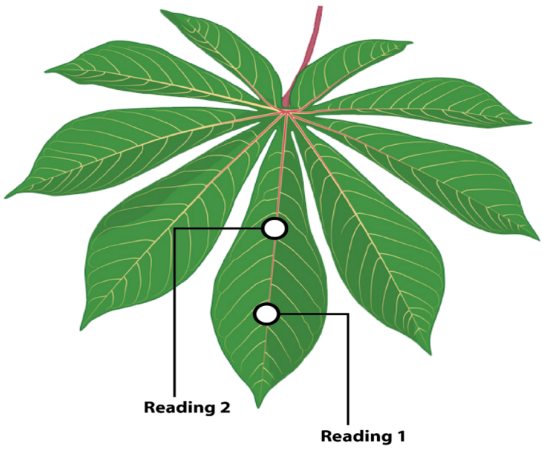

Figure 1: Depiction of spectral data Reading 1 and Reading 2 from the cassava leaf.

Data type	Screen house experiment	Open field experiment
Spectrograms	Reading 1: 4080, Reading 2: 4080	Reading 1: 3780, Reading 2: 3780
Leaf Images	4080 images	3780 images
Chemical Lab data	4080	3780
Expert Score	4080	3780

Table 2: Number of data instances in the data set.

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.

Example of the datapoint name	Variable Meaning
A 1 HLT 1 R a 1g	
A	Cassava Variety
1	Week of data collection
HLT	Disease class
1	Plant number
R	Reflectance mode
a	Plant leaf number
lg	Placeholder/ Constant variable

Table 3: Data Labeling Scheme.

Mbanzibwa, D. R., Tian, Y. P., Tugume, A. K., Mukasa, S. B., Tairo, F., Kyamanywa, S., ... & Valkonen, J. P. T. (2011). Simultaneous virus-specific detection of the two cassava brown streak-associated viruses by RT-PCR reveals wide distribution in East Africa, mixed infections, and infections in Manihot glaziovii. Journal of virological methods, 171(2), 394-400.

Owomugisha, Godliver, Ephraim Nuwamanya, John A. Quinn, Michael Biehl, and Ernest Mwebaze. "Early detection of plant diseases using spectral data." In Proceedings of the 3rd International Conference on Applications of Intelligent Systems, pp. 1-6. 2020.

CID Bio-Science, Inc., "CI-710 MINIATURE LEAF SPECTROMETER," 2010. [Online]. Available: `http://www.cid-inc.com`

Thermo Fisher Scientific, "StepOnePlus™ Real-Time PCR System." [Online]. Available: `https://www.thermofisher.com/order/catalog/product/4376600`

L3Harris Geospatial, "Vegetation analysis: Using vegetation indices in envi," 2022. [Online]. Available: https://www.l3harrisgeospatial.com/Support/Self-Help-Tools/Help-Articles/Help-Articles-Detail/ ArtMID/10220/ArticleID/16162/Vegetation-Analysis-Using-Vegetation-Indices-in-ENVI