Data Sheet for Cocoa and Cashew Data Collected in Ghana Under the Lacuna Project

We present the Lacuna Cashew and Cocoa data sheet created by KaraAgro AI Lab.
We follow the datasheet for the 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 with the specific purpose of advancing research and development in the field of cashew and cocoa yield estimation. This dataset may represent a first contribution of drone data to the field of cashew and cocoa yield estimation research. By collecting the dataset from Ghana, a diverse and well-annotated collection of cocoa and cashew crop images collected from Ghana Boro-Region and Kade will be made available to support the development of accurate, efficient, and scalable methods for crop yield estimation.

Was there a specific gap that needed to be filled? Please provide a description.

Based on multi-stakeholder engagements conducted by KaraAgro AI, also with women smallholder cashew farmers, stakeholders have identified pest and disease detection and yield estimation as critical concerns. Thus, there is a need for more innovative and efficient solutions to improve the monitoring and estimation of crop yield. This highlights a gap in the available tools and resources, which can be addressed through the use of advanced technologies such as machine learning and image analysis.

The creation of an open and accessible cashew dataset with well-labelled, curated, and prepared imagery can provide a valuable resource for data scientists, researchers, and social entrepreneurs to develop innovative solutions towards yield estimation.

Who created this data set (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

The dataset was created by a team of data scientists from the KaraAgro AI Foundation, with support from agricultural scientists and officers.

Composition

What do the instances that comprise the data set represent (e.g., documents, photos, people,

countries)?

Each instance in the dataset includes crop image (JPEG), image status (flower, immature, mature, ripped, spoilt, tree), and file type (images and bounding box annotations).

How many instances are there in total (of each type, if appropriate)?

There are 4,715 instances of cashew images and 4,069 instances of cocoa images.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of the cases from a larger set?

The dataset contains various instances that were captured from the Bono Region and Eastern Region for Cashew and Cocoa data respectively..

What data does each instance consist of? "Raw" data or features?

Each instance includes: the crop image, image status(Flower, Immature, Mature, Ripped, Spoilt, Cashew Tree) for both cocoa and cashew and 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 maturity stage of the crop i.e. flower, immature, mature, ripped, spoilt or cashew tree.

Is any information missing from individual instances?

None

Are relationships between individual instances made explicit?

Yes, there are two sets of data, the cocoa dataset and the cashew 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?

No, the dataset is self-contained, it does not rely on any other external sources

Does the dataset contain data that might be considered confidential?

No, the dataset does not contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No, the dataset does not contain data that might be offensive, insulting, threatening or data that may cause anxiety.

Collection Process

How was the data associated with each instance acquired?

The data associated with each instance was acquired from different cashew farms in the Bono Region and cocoa farms in the Eastern Region in Ghana

What mechanisms or procedures were used to collect the data?

The images were taken with a d DJI P4 multispectral drone. The images were captured using a drone that was flown manually. The drone was flown at different altitudes to ensure that comprehensive information about the crops was gathered. The photos of the cashew and cocoa crops were taken at different angles with altitudes ranging from 2 to 10 meters. This altitude range provides a good balance between capturing a close-up view of the fruits and

their growth stages and a wider perspective that allows for variation.

If the dataset is a sample from a larger set, what was the sampling strategy?

The final dataset is the complete dataset and not a sample of any other dataset

Who was involved in the data collection process?

The karaAgro team, district agricultural officers and extension officers and farmers.

Over what timeframe was the data collected?

The cashew dataset was collected in two rounds: The first data collection happened in November 2022, the second in January 2023. The cocoa data was collected in one round in December 2023.

Were any ethical review processes conducted (for example, by an institutional review board)?
No

Preprocessing, cleaning, and labeling

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)?

No

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 labelled 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.

Yes, the annotation tool *makesense.ai* can be accessed *here*

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

During the annotation, the cashew dataset was used to develop models to train object-detection models to speed up annotation in a semi-supervised approach.

Is there a repository that links to any or all papers or systems that use the dataset?

None at the moment

What (other) tasks could the dataset be used for?

1. 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?

Nothing about the composition of the dataset would affect future use for the use case/task the dataset was curated for.

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. How will the dataset be distributed (for example, tarball on website, API, GitHub)?

Yes, the dataset will be made publicly available

Does the dataset have a digital object identifier (DOI)?

https://doi.org/10.57967/hf/0959

When will the dataset be distributed?

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

License: cc-by-4.0

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)?

Darlington Akogo can be be contacted on his email address - darlington@gudra-studio.com

Is there an erratum?

No

Will the dataset be updated (for example, to correct labelling errors, add new instances, or delete instances)?

Updates to the dataset will be communicated to the public through the datasheet or data cards on data hosting websites.

Will older versions of the data- set continue to be supported/hosted/ maintained? If so, please describe how.

The data which is publicly available will be maintained by karaAgro AI and Makerere University. Information regarding dataset version will be communicated through datasheets and data cards on online hosting platforms

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

The dataset and the datasheet will be made publicly available. Any contribution can be directed to the authors, KaraAgro AI and Makerere University.

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.