Drone-based Agricultural Dataset for Crop Yield Estimation

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 creation from KaraAgro and Makerere AI Lab serves multifaceted purposes. The dataset was conceived to advance research and development in the realm of cashew and cocoa yield estimation. It pioneers the incorporation of drone data into cashew and cocoa yield estimation research, providing a diverse and well-annotated assortment of cocoa and cashew crop images from Ghana's Boro-Region and Kade. This compilation aids the development of accurate, efficient, and scalable crop yield estimation methods. The dataset can also be described as having dual objectives. It first revolves around crafting a comprehensive crop drone dataset, meticulously labelled and accessible for users. Its primary focus is on crop yield estimation, with broader applications in variety identification, crop classification, and more, particularly impactful for Africa's coffee, cashew and cocoa-dependent households. Additionally, the dataset emphasizes crop fruits, catering to researchers and machine learning experts with its labelled compilation, fostering model construction for various objectives, including yield estimation and quality assurance. The unifying thread is empowering farmers with data-driven decisions for enhanced sales and income prospects.

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.

Coffee is also a primary source of income for more than 12 million households in Africa, particularly for rural-based populations. Yield estimation allows farmers to make good business decisions and appropriately plan ahead for their equipment, fuel, and labor needs, ensure they have enough storage, cash-flow budgeting, and make early marketing decisions. This vital work can provide farmers the opportunity to ensure healthy and fresh produce, and therefore better sales and income The creation of an open and accessible crop 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, organisation)?

The dataset was created by two teams of data scientists from Ghana and Uganda. The Ghana team is made up of a team of data scientists from the KaraAgro AI Foundation, with support from agricultural scientists and officers from Ghana. The Uganda set was created by scientists from the Makerere Artificial Intelligence Lab - Uganda, Marconi Lab -Uganda 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 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.: 19497.51.

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 depicting maturity and development stage (flower, immature, mature, ripe, spoilt, tree), and file type (images and bounding box annotations). For coffee data, each instance includes an image status, which can be either the side view or the aerial view of the image. In particular, the side images are required to exhibit visible coffee beans that are clean and clear. Therefore, the instances in the dataset depict different perspectives of coffee imagery, emphasising the presence of discernible and pristine coffee beans in the side-view images. Each instance also contains the image details, i.e., crop age, and location including the district and sub-county.

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

There are 4,715 instances of cashew images, 4,069 instances of cocoa images, 3098 instances of cashew images from Uganda and 3200 coffee images. Alongside these images, valuable annotations are provided, indicating the positions of fruits/beans within blusters (for coffee) and maturity stages of fruits for all datasets. For coffee, the clusters exhibit beans in three primary states: raw, ripe and spoilt. For cashew the annotated stages were, the cashew tree, flowers, immature fruits, mature fruits, ripe fruits and spoilt fruits. Lasty, immature fruits, mature fruits, ripe fruits and spoilt fruits were annotated for cocoa crops.

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. From Uganda, the dataset encompasses data from the NaCORI field, Kyotera district, Masaka district, Bukomansimbi district, Ngora district, Kumi district, Soroti district, Abim district, Napak district, Lira district and Nakasongola district. The dataset has image samples collected from significant cashew nut growing districts selected with the guidance of agricultural experts to obtain a representative dataset. Coffee specialists also expertly chose these fields to ensure comprehensive data collection.

What data does each instance consist of? “Raw” data or features?

Each instance includes: the crop image, image status(Flower, Immature, Mature, Ripe, Spoilt, Cashew Tree) for both cocoa and cashew and location (gps coordinates). The top-view images reveal distinctive features helpful in identifying different coffee varieties, while the side-view images display characteristics indicating the maturity stage of coffee beans.

Is there a label or target associated with each instance? If so, please provide a description.

Cashew

Each instance is associated with a class label based on the maturity stage of the crop i.e. flower, immature, mature, ripe, spoilt or cashew tree. Flower; This is displayed as miniature flowers that often appear at the end of the branch Premature; The nut and/or fruit displays a reddish-maroon colour. Unripe; The nut and/or fruit displays a light green colour. Ripe; The nut displays a gray colour.Fruit displays a yellow/ red colour Spoilt; The nut and/or fruit displays a black colour. There may be oozing of a gray sap-like liquid from the fruit and/or the nut. Tree: This label encompasses the tree that is in centre focus in the image.

Coffee

Though not explicitly specified, we ensured that the image selection process prioritized trees with well-captured coffee beans. Consequently, the side-view images prominently showcase these beans, and subsequently, annotations were applied to highlight and identify them accurately

Is any information missing from individual instances?

None

Are relationships between individual instances made explicit?

Yes, there are three sets of data, the cocoa dataset, cashew dataset and coffee dataset. However 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?

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 for cashew was acquired from different cashew farms in the Bono Region and cocoa farms in the Eastern Region in Ghana and the cashew nut growing prominent districts across three regions in Uganda.

What mechanisms or procedures were used to collect the data?

The images were taken with a 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. The data collection protocol/mechanism/procedures were collaboratively developed by a team comprising researchers from Makerere Artificial Intelligence Lab - Uganda, The Marconi Lab-Uganda, and coffee experts from NaCORI - Uganda. We ensured that each image was captured with optimal visibility of the nuts and fruits. We ensured maximum illumination and appropriate exposure such that the nuts and fruits are easily identifiable in the images. We also ensured that all images have location information (GPS coordinates) and timestamps.

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?

KaraAgro

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

Makerere

The team consisted of researchers from Makerere AI Lab and the Marconi Lab, coffee experts from NaCORI, and a skilled drone pilot. Furthermore, an agricultural expert from National Crops Resources Research Institute (NaCRRI), and lastly a district agricultural officer

Over what timeframe was the data collected?

KaraAgro AI Labs

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 2022.

Makerere AI Labs

The coffee data was collected from October 2022 to June 2023 while the cashew data was collected from March 2023.

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)?

Coffee

We conducted thorough data cleaning, eliminating blurry and overexposed images while resolving any inconsistencies. The coffee specialist from NaCORI expertly handled the annotation process, and the Makerere team gave the training on using an offline tool called VGG Image Annotator (VIA) to annotate the images.(Abhishek 2021).The annotated data is provided in YOLO format with 5 class ids representing the coffee labels; 0:Unripe, 1:Ripening, 2:Ripe, 3:Spoilt, 4:Coffeetree

Cashew

We carried out data cleaning to remove blurry images and over exposed images, and resolution of inconsistencies. The data was labeled using an online annotation tool called Makesense AI. The annotated data is provided in YOLO format with 6 class ids representing the cashew labels; (0)Tree (1)Flower, (2)Premature, (3)Unripe, (4)Ripe, (5)Spoilt

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*
The link to the annotation tool for coffee is available: https://www .robots.ox.ac.uk/~vgg/software/via/

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**.** The dataset and the associated metadata are stored on Hugging Face which is a platform that allows users to share machine learning models and datasets. The dataset was assigned a Digital Object Identifier: http://doi.org/10.57967/hf/0959

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

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

When will the dataset be distributed?

The dataset is available under the specified DOI as of August 2023

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 4.0 BY license, allowing users to share and adapt the dataset as 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, Makererer AI Lab and the Marconi 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 - 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

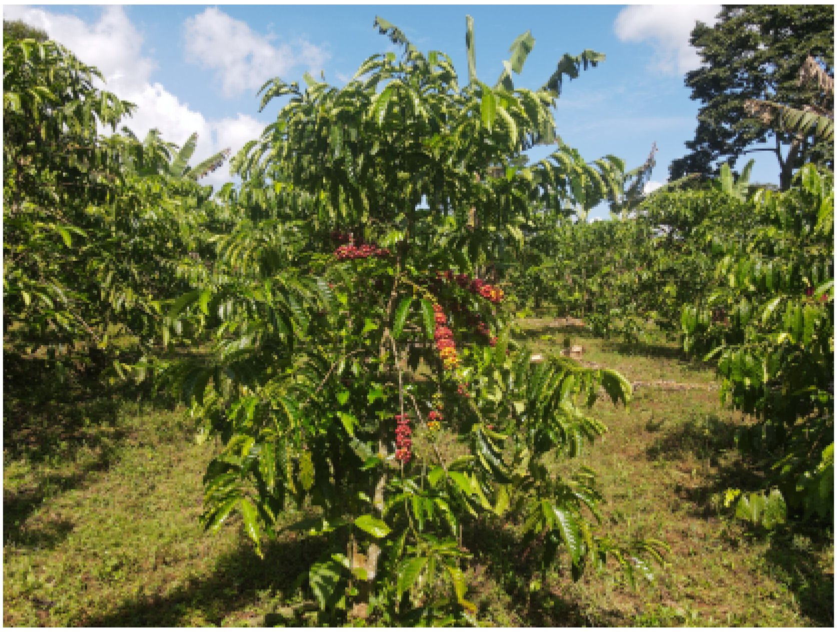
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

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

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.

Figure 1: Sample of Coffee Data



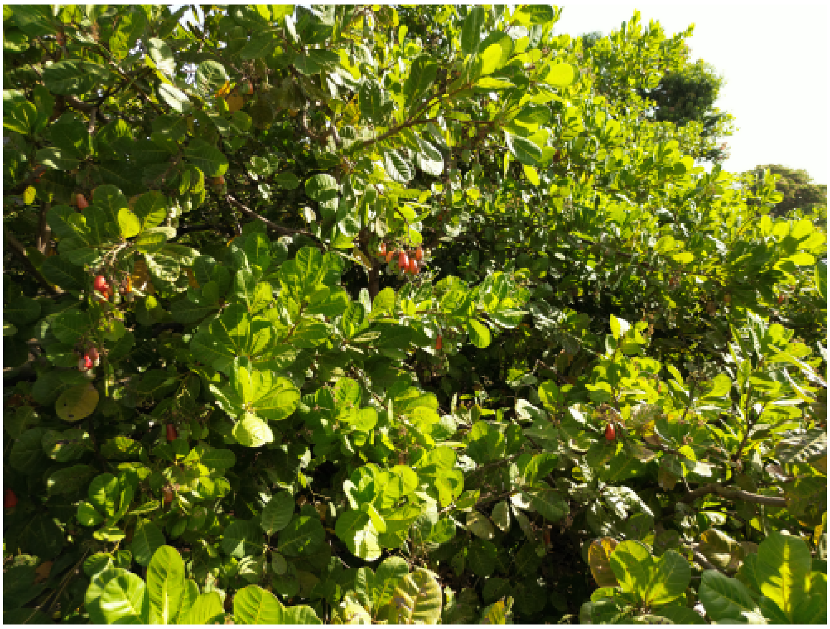


Figure 2: Cashew nut tree

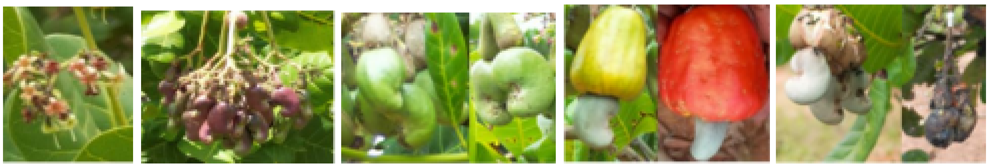


Figure 3: Cashew labels (L-R) Flower, Immature (Premature), Unripe (Mature), Ripe, Spoilt

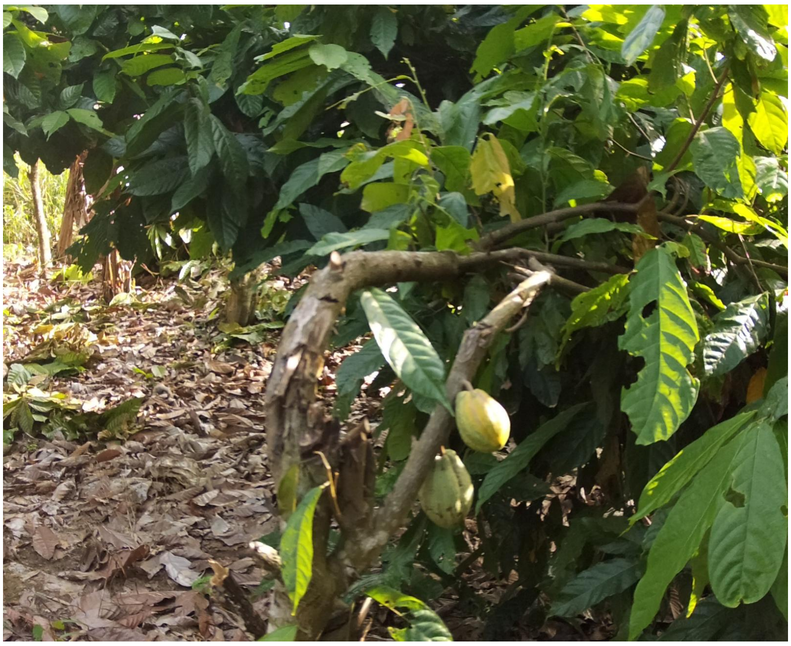


Figure 4: Cocoa tree