Localized Data Work as a Precondition for Data-Centric ML: A Case Study of Full Lifecycle Crop Disease Identification in Ghana

Darlington Akogo [1] Issah Samori [1] Cyril Akafia [1] Harriet Fiagbor [1] Andrews Kangah [1]
Donald Kwame Asiedu [1] Kwabena Fuachie [1] Luis Oala [2]

Abstract

The Ghana Cashew Disease Identification with Artificial Intelligence (CADI AI) project demonstrates the importance of sound data work as a precondition for the delivery of useful, localized data-centric solutions for public good tasks such as agricultural productivity and food security. Drone-collected data and machine learning are utilized to determine crop stressors. Data, model and the final app are developed jointly and made available to local farmers via a desktop application.

Cashew is a significant cash crop in Ghana (Rabany et al., 2015), with small and medium farmers relying on it for income. Cashew cultivation is concentrated in specific regions of Ghana. However, farmers face challenges including insect, plant disease and abiotic stress factors that reduce their yields (ICAR; Jayaprakash et al., 2023; Mensah et al., 2023; Timothy et al., 2021). To address these issues, the Cashew Disease Identification With Artificial Intelligence (CADI AI) project was launched to provide a data-centric solution. The project encompasses three stages. Comprehensive data work encompassed stakeholder consultation, data collection, data annotation and labelling. The collected drone data is open for researchers and data scientists to develop innovative machine learning applications to improve food security. Model work involved the training of an object detection model to diagnose and detect stress factors in cashew crop images. Finally, the model was integrated into a desktop application for farmers, allowing them to input their own data and receive diagnoses. The application also displays the precise location of the image, enabling farmers to identify affected areas on their farms.

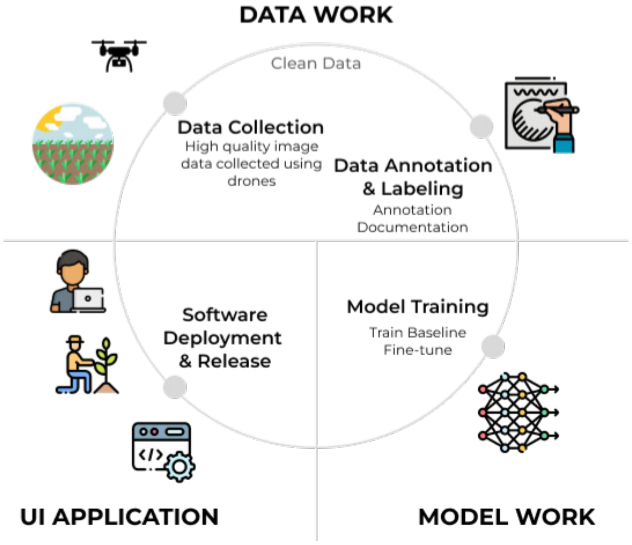


Figure 1: A visual summary of the application lifecycle: data work (data collection with farmers, data annotation and labelling), model work (model training and fine-tuning), and UI application (software deployment and release to farmers).

1. Data work

The data was collected from cashew farms in the Bono Region of Ghana, necessitating two separate trips to the farms to accommodate seasonal variations and diversity of data. In total, the data collection process spanned six days. The dataset is diverse in maturity stages, camera angles, time of capture, and various types of stress morphology. All images were captured with the P4 Multi-spectral drone (Dji) at image resolution of 1600 x 1300 pixels. The images comprise close up shots and shots from distance of the cashew plant abnormalities. The total number of images collected is 4,736. Full details and datasheet in the appendices. Further improvements to the dataset could be made by capturing across more regions during blooming cycles or varying devices for robustness testing (Oala et al., 2023).

*Equal contribution [1]KaraAgro AI, Accra, Ghana [2]Dotphoton AG, Zug, Switzerland. Correspondence to: Harriet Fiagbor <harrietfiagbor@gmail.com>, Cyril Akafia <kwakucyril@gmail.com>.

Published at the Data-Centric Machine Learning Research (DMLR) Workshop of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. July 2023. Copyright 2023 by the author(s).

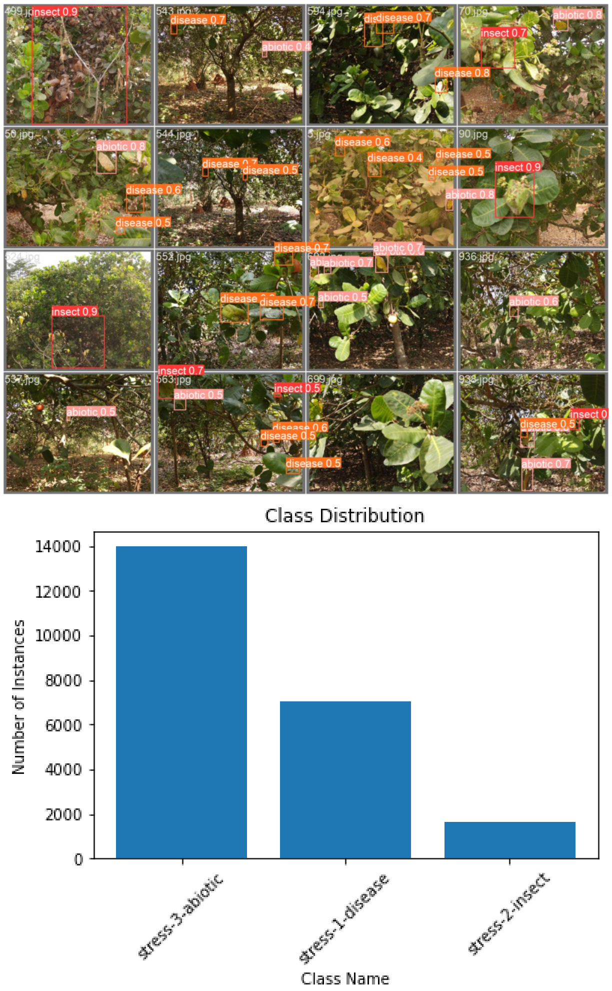


Figure 2: Top: Sample instances from the annotated dataset. For a higher resolution sample see the appendices. Bottom: Distribution of labels in the annotated data.

1.1. Data Annotation/Labelling

The data was annotated by the project team with labelling tools makesense (Makesense) and roboflow (Roboflow). Refer to appendix A.1 for annotation guidelines developed by an agricultural scientist with expertise in crop health and disease management from a local Ghanaian university. Each stress instance is associated with a class label based on the status of the crop. The labels are **"insect"**, **"disease"** and **"abiotic"** respectively as depicted in Figure 2. The data was split into train, validation and test sets i.e., 3788, 710 and 238 images respectively. During the training, it was found that the dataset is significantly skewed towards the abiotic class.

2. Model work

We utilized the YOLO v5X (Jocher, 2020) architecture, known for its strong performance on object detection benchmarks, as the foundation for this study. The experiments were conducted on the high-performance DFKI GIZ cluster.The dataset has a significant skew towards the abiotic class (see Figure 2), and measures were taken to balance the data by augmenting other classes. However, preserving the skewness was important to reflect the higher occurrence of abiotic factors on farms. The best model achieved a mean average precision (mAP) of 0.648. See Table 1 and Figure 4 in the Appendices for detailed experimental evaluations and baselines. The model has a few limitations that affect its performance in distinguishing between the disease class and the abiotic class. The primary challenge lies in the similarity between these two classes within a typical farm setting. The model may encounter difficulties in accurately differentiating between them due to their overlapping characteristics. This limitation is an inherent challenge in the dataset and can impact the model's accuracy when classifying these instances. However, it is worth noting that the model exhibits strong performance when it comes to the insect class. This is attributed to the distinct characteristics of insect class, which make them easier to identify and classify accurately.

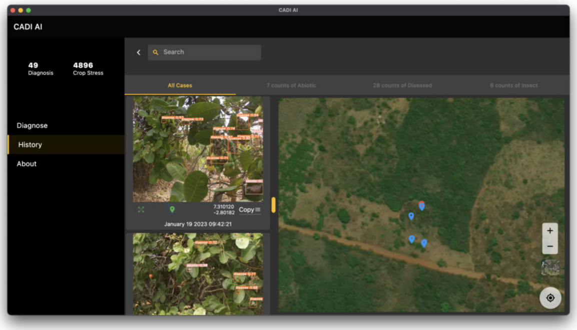


Figure 3: Screenshot of the final UI application. For more details see appendices.

3. Closing the loop: UI application

A software application built with Flutter (Google, 2019) was infused with the CNN model trained on the collected data, allowing farmers to use the model on new data for crop disease management. Our objective is to make the CADI AI project's data, model, and software widely accessible to maximize their impact within the agriculture community. The data is available through prominent platforms such as Kaggle and Hugging Face dataset hub. These platforms provide user-friendly interfaces, allowing researchers, developers, and enthusiasts to access the data for their specific needs. Furthermore, the model itself can be accessed through the Hugging Face platform, enabling users to leverage its capabilities in their own ML applications. A summary of all resources is in the appendices.

References

AGPL, G. Gnu affero general public license, Feb 2023. URL https://en.wikipedia.org/wiki/GNU_Affero_General_Public_License.

Dji. P4 multispectral - dji. URL https://www.dji.com/p4-multispectral.

Exif. Exif exchangeable image file format, Feb 2014. URL https://www.loc.gov/preservation/digital/formats/fdd/fdd000146.shtml.

Google. Flutter - beautiful native apps in record time, 2019. URL https://flutter.dev/.

ICAR. Cashew pest database. https://cashew.icar.gov.in/pestsite/. (Accessed on 02/02/2023).

Jayaprakash, V., Rajagopal, M. K., et al. Cashew dataset generation using augmentation and ralsgan and a transfer learning based tinyml approach towards disease detection. *arXiv preprint arXiv:2304.08766*, 2023.

Jocher, G. Yolov5 by ultralytics, 2020. URL https://github.com/ultralytics/yolov5.

KaraAgroAI. Cadi ai - cashew disease identification with artificial intelligence, May 2023. URL https://github.com/karaagro/cadi-ai.

Makesense. Makesense ai tool for annotation. URL https://www.makesense.ai.

Mensah, P. K., Akoto-Adjepong, V., Adu, K., Ayidzoe, M. A., Bediako, E. A., Nyarko-Boateng, O., Boateng, S., Donkor, E. F., Bawah, F. U., Awarayi, N. S., et al. Ccmt: Dataset for crop pest and disease detection. *Data in Brief*, pp. 109306, 2023.

Oala, L., Aversa, M., Nobis, G., Willis, K., Neuenschwander, Y., Buck, M., Matek, C., Extermann, J., Pomarico, E., Samek, W., Murray-Smith, R., Clausen, C., and Sanguinetti, B. Data models for dataset drift controls in machine learning with optical images. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=I4IkGmgFJz.

Rabany, C., Rullier, N., and Ricau, P. The african cashew sector in 2015. *General Trends and Country Profiles. African Cashew Initiative (ACI) Report*, 2015.

Roboflow. Roboflow: go from raw images to a trained computer vision model in minutes. URL https://roboflow.com/.

Timothy, M., John, O., Aibinu, A., and Adebisi, B. Detection and classification system for cashew plant diseases using convolutional neural network. In *The 5th International Conference on Future Networks & Distributed Systems*, pp. 225–232, 2021.

A. Data Annotation

A.1. Guidelines for Annotation

Below are some guidelines for labeling or annotating the data set by team.

- Provide enough space to capture the affected area without cutting any part off, but avoid introducing too much space.
- If possible, zoom into the affected area to ensure accurate annotation.
- Avoid including too much gap between two affected areas. Treat a wide gap as another annotation.
- To ensure completeness, try to annotate every possible instance as long as it is visible.
- Even for images that appear to have no affected areas, zoom in and carefully examine the image to ensure nothing is missed. It is better to have an annotation than to remove the image altogether. However, if truly nothing exists, discard the image from the database.
- Avoid using images with human faces whenever possible.
- If an image with a human face is necessary, but the face is not centered, consider cropping it out and replacing the original image with the cropped version

A.2. Deciding the Labels

- **Insect/ pest stress factors** represent the damage to crops by insects or pests
- **Diseased factors** represent attacks on crops by microorganisms.
- **Abiotic stress factors** represent stress factors caused by non-living factors, e.g. environmental factors like weather or soil conditions or the lack of mineral nutrients to the crop.

The decision to use the labels ”abiotic”, ”disease”, and ”insect” for our object detection task was recommended by an agricultural scientist with expertise in crop health and disease management, Dr. Torkpor Stephen from University of Ghana.

B. Evaluation and Results

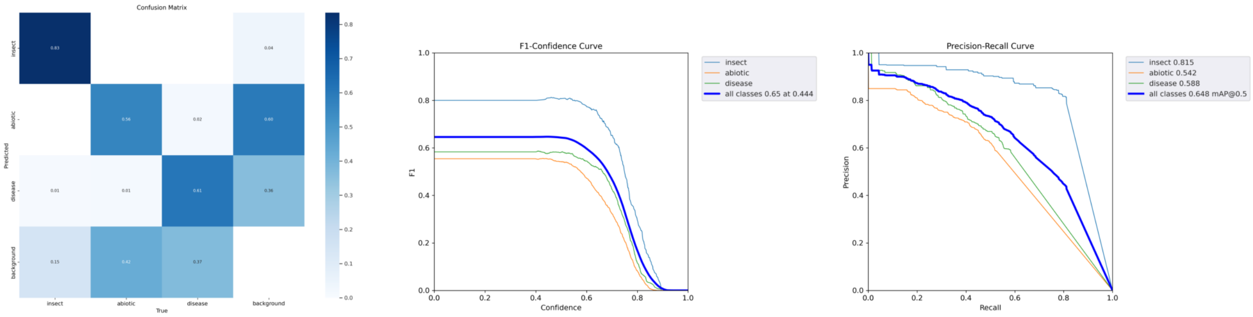


Figure 4: Left: Confusion Matrix of Evaluation results. Center: F1-Curve of Evaluation results. Right: PR-Curve of Evaluation results

Table 1: Table showing Experimental Procedures using YOLO architecture

Training Data	Pretrained Model Used	Epoch	Image size	Batch Size	Comments	mAP
Original and background images	yolov8x	30	640	16	Little overfitting	0.6
	yolov8x	30	640	16	More background images were included. This lead to overfiting.	0.56
	yolo8m	30	640	16	Tried the above experiment with a smaller architecture	0.57
Only original images	yolov8m	50	640	16	Early overfitting	0.59
	yolov8n	50	640	16	Did not converge	0.53
	yolov8l	30	640	16	Overfitting	0.58
	yolov8x	30	640	16	Overfitting	0.61
	yolov8n	60	640	16	Did not converge	0.55
	yolov8n	100	640	16	Overfitting	0.57
	yolov8m	50	960	16	Overfitting	0.59
	yolov8n	80	1280	16	Overfitting	0.58
Original, background and augmented images	yolov5x	30	640	16	Converged	0.6
	yolov5t	45	640	16	Overfitting	0.58
	yolov5m6	50	640	16	Overfitting	0.572
	yolov5x	35	640	32	Increased batch size lead to increased mAP	0.615
	yolov5x	35	640	64	Overfitting	0.601
	yolov5x	35	640	48	Little Overfitting	0.603
	yolov5x	**25**	**640**	**56**	**Batch size lead to a significant increase in performance. The model converged.**	**0.648**
	yolov5x	40	640	56	Overfitting	0.606
	yolov5x	40	640	72	Overfitting	0.616
	yolov5x	40	640	72	Overfitting	0.616
	yolov5x	35	640	48	Overfitting	0.58
	yolov5x	35	640	48	Overfitting	0.57
	yolov8x	40	640	32	Overfitting	0.55
	yolov8n	30	640	32	Overfitting	0.55
	yolov8n	30	640	32	Overfitting	0.48
	yolov8n	50	640	32	Overfitting	0.53

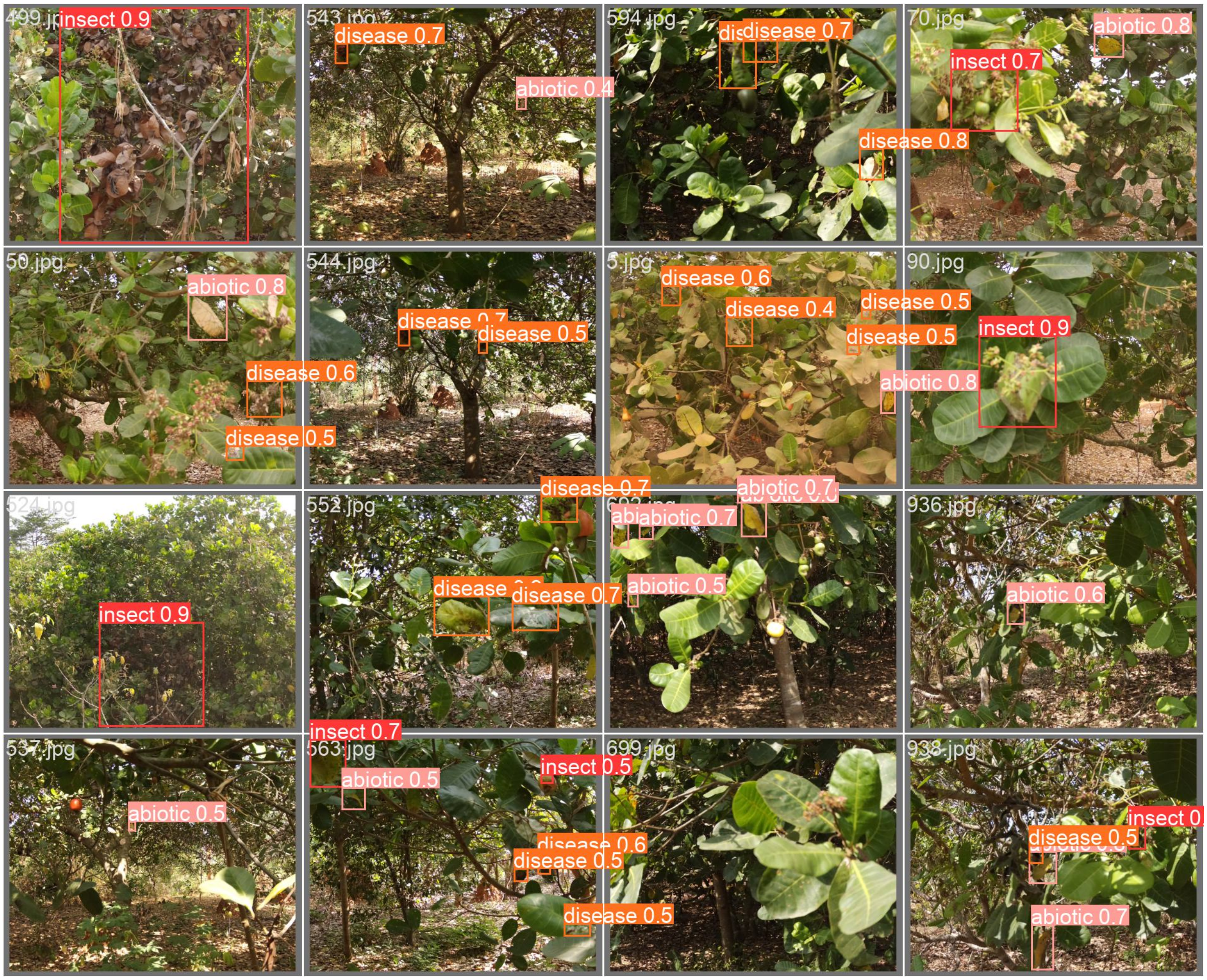
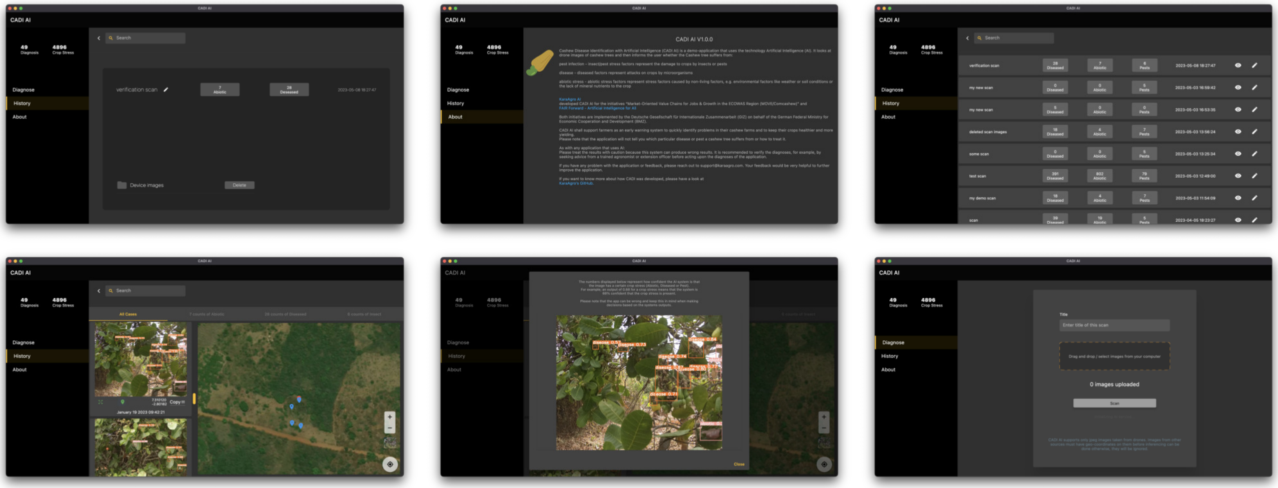


Figure 5: Sample instances from the annotated dataset.



Cashew Disease Identification with Artificial Intelligence (CADI AI) is a demo-application that uses the technology Artificial Intelligence (AI). It looks at drone images of cashew trees and then informs the user whether the Cashew tree suffers from:

pest infection - insect/pest stress factors represent the damage to crops by insects or pests

disease - diseased factors represent attacks on crops by microorganisms

abiotic stress - abiotic stress factors represent stress factors caused by non-living factors, e.g. environmental factors like weather or soil conditions or the lack of mineral nutrients to the crop

KaraAgro AI
developed CADI AI for the initiatives "Market-Oriented Value Chains for Jobs & Growth in the ECOWAS Region (MOVE/Comcashew)" and
FAIR Forward - Artificial Intelligence for All

Both initiatives are implemented by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ).

CADI AI shall support farmers as an early warning system to quickly identify problems in their cashew farms and to keep their crops healthier and more yielding.
Please note that the application will not tell you which particular disease or pest a cashew tree suffers from or how to treat it.

As with any application that uses AI:
Please treat the results with caution because this system can produce wrong results. It is recommended to verify the diagnoses, for example, by seeking advice from a trained agronomist or extension officer before acting upon the diagnoses of the application.

If you have any problem with the application or feedback, please reach out to support@karaagro.com. Your feedback would be very helpful to further improve the application.

If you want to know more about how CADI was developed, please have a look at
KaraAgro's GitHub.

verification scan	28 Diseased	7 Abiotic	6 Pests	2023-05-08 18:27:47
my new scan	0 Diseased	0 Abiotic	5 Pests	2023-05-03 16:59:42
my new scan	5 Diseased	0 Abiotic	0 Pests	2023-05-03 16:53:35
deleted scan images	18 Diseased	4 Abiotic	7 Pests	2023-05-03 13:56:24
some scan	0 Diseased	0 Abiotic	5 Pests	2023-05-03 13:25:34
test scan	391 Diseased	802 Abiotic	79 Pests	2023-05-03 12:49:00
my demo scan	18 Diseased	4 Abiotic	7 Pests	2023-05-03 11:54:09
scan	39 Diseased	19 Abiotic	5 Pests	2023-04-05 18:23:27

The numbers displayed below represent how confident the AI system is that the image has a certain crop stress (Abiotic, Diseased or Pest).
For example; an output of 0.68 for a crop stress means that the system is 68% confident that the crop stress is present.

Please note that the app can be wrong and keep this in mind when making decisions based on the systems outputs.

CADI AI supports only jpeg images taken from drones. Images from other sources must have geo-coordinates on them before inferencing can be done otherwise, they will be ignored.

Figure 6: Screenshots of the final app (from left to right, top to bottom): About page, edit window, project history, stress locator, stress analysis, upload screen.

C. CADI AI Software

The software is named CADI AI; an abbreviation for Cashew Disease Identification with Artificial Intelligence. CADI AI has the following features:

- A desktop interface for accessing all other features of the application
- A diagnosis widget where a user could select multiple number of images for diagnoses without limit
- A history widget which shows a history list of all scans made
- A widget for viewing all images uploaded in a particular scan with a satellite map that shows where the images were taken

CADI AI does not rely on an internet connection to work. It is, in an ideal world, a totally offline application. Because the application spins up a local flask server that enables communication with the CNN model, there is sometimes a need to connect to the internet to download metadata which are then cashed. The local cache can expire after some time and create the need for a re-sync using an internet connection.

Note that CADI AI only runs scans on images that are having coordinates in their Exif metadata (Exchangeable Image Format).

The Exif specifications define a pair of file types, mainly intended for recording technical details associated with digital photography (Exif, 2014).

Exif is a metadata standard that defines formats for sharing metadata related to images, sound, and ancillary tags used by digital cameras, scanners and other systems that handle image recorded by digital cameras.

In this case, it is required that images selected for scan in CADI AI must have metadata following the Exif standard, and must have geographical location coordinates in the metadata before these images can be scanned by the CNN model.

This is because the application shows location pins on a satellite map that enables farmers to locate the region of the farm where a disease, abiotic stress or pest is found present. Without GPS coordinates or geographical location coordinates, farmers would be oblivious to where exactly issues are found in their farms.

The development of CADI AI was started in-house at KaraAgro AI and the codebase was later published as an open source repository welcoming contributions from the open source community and serving up the application to the general public on github (KaraAgroAI, 2023)

D. Resource Coordinates

Table 2: Table Showing URLs to Open Source Resources

	Location
Data	https://www.kaggle.com/datasets/karaagroaiprojects/cadi-ai https://huggingface.co/datasets/KaraAgroAI/CADI-AI
Model	https://huggingface.co/KaraAgroAI/CADI-AI
App Code and Executable	https://github.com/karaagro/cadi-ai

E. Contribution and Commercial Use

As an open-source project under the GNU Affero General Public License(GNU AGPL)(AGPL, 2023), the CADI AI project encourages contributions from the wider community. Users are free to modify and build upon the existing resources, tailoring them to their unique needs. Moreover, the resources can be utilized for commercial purposes, providing attribution to the original creators is duly given. This flexibility fosters collaboration and allows the project's benefits to extend to diverse domains and industries.

F. Datasheet

F.1. 3.1 Motivation

For what purpose was the data set created? Was there a specific task in mind?

The creation of this dataset represents a first contribution of drone data to the field of cashew crop research: Providing an open and accessible resource of high-quality, well-labeled drone imagery collected from Ghana Bono-Region, this dataset will offer data scientists, researchers, and social entrepreneurs within Sub-Saharan Africa and beyond, opportunities for innovative machine learning experiments and the development of solutions for **infield cashew crop disease diagnosis and spatial analysis**.

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

Yes. The threat of pests and diseases to the agricultural sector in Ghana is a constant concern, with climate change contributing to the potential for new and more damaging types of outbreaks (Yeboah et al., 2023). 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.

However, the current methods of identifying agricultural pest and disease outbreaks, such as land surveys and on-site observations by individuals, are limited in their effectiveness and efficiency. Thus, there is a need for more innovative and efficient solutions to improve the monitoring and management of crop health. 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-labeled, curated, and prepared imagery can provide a valuable resource for data scientists, researchers, and social entrepreneurs to develop innovative solutions towards infield pest and disease detection and 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.

Who funded the creation of the data set? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The creation of this dataset was made possible through funding of the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) through their projects "Market-Oriented Value Chains for Jobs & Growth in the ECOWAS Region (MOVE)" and "FAIR Forward - Artificial Intelligence for All", which GIZ implements on behalf the German Federal Ministry for

Economic Cooperation and Development (BMZ).
The MOVE initiative aims to support market-oriented and resilient value chains that contribute to the creation of income and employment in the ECOWAS region
The FAIR Forward initiative aims to promote a more open, inclusive, and sustainable approach to AI on an international level, by partnering with countries such as Ghana, India, Indonesia, Kenya, Rwanda, South Africa and Uganda.

F.2. 3.2 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 (Disease, Abiotic, and Insect), file type (images and bounding box annotations) and location (this variable though is without values).

Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

No. There is only one instance type, which represents cashew crops based on drone images with various attributes.

Does the data set contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the data set is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representation was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset contains various instances that were captured in the Bono region, which is renowned for its cashew production. The data was collected in two rounds: The first data collection happened in November 2022, the second in January 2023. The data captured represents cashew data from a year where cashew blooming was particularly late. Given that the data was collected punctually only twice, it might be that not all blooming variations of Cashews have been captured, potentially influencing the variety of the collected data.

Thus, the dataset collected is not representative. While more continuous data collection across various regions during the blooming cycle could have been beneficial, we still consider our dataset to be sufficiently diverse. This is due to the inclusion of different maturity stages, camera angles, time of capture, and various types of stress morphology.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance includes: the crop image 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 status of the crop. The labels are "*insect/pest*", "*disease*" and "*abiotic*" respectively as depicted in Figure 1:

- **Insect/ pest** stress factors represent the damage to crops by insects or pests
- **Diseased** factors represent attacks on crops by microorganisms.
- **Abiotic** stress factors represent stress factors caused by non-living factors, e.g. environmental factors like weather or soil conditions or the lack of mineral nutrients to the crop.

The decision to use the labels ”abiotic”, ”disease”, and ”insect” for our object detection task was recommended by an agricultural scientist with expertise in crop health and disease management, Dr. Torkpor Stephen from University of Ghana.

It is important to note that while these labels provide a general categorization of crop damage, they may not fully capture the complexity of the underlying causes. In addition, the labels may not be exhaustive and other types of damage may not be captured by these categories. As with any dataset, users should be aware of the limitations and context of the labels used and exercise caution when interpreting the results of models trained on this data.

Examples of the limitations and complexities involved includes

- A plant may exhibit symptoms of both insect damage and disease, making it difficult to assign a single label to the damage.

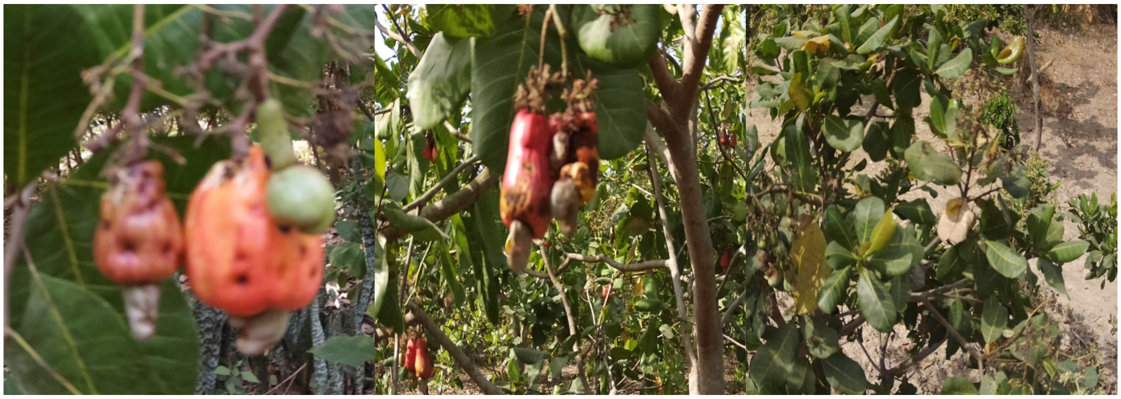


Figure 7: Class labels associated with data (from left to right): Insect, disease, abiotic

- Damage caused by abiotic factors such as drought or nutrient deficiency may be similar to damage caused by disease or insect infestation, leading to confusion when assigning labels.
- Damage caused by multiple factors may not fit neatly into a single label category, requiring more nuanced and complex labeling.
- Different species of insects or diseases may cause similar damage to crops, making it difficult to distinguish between them using only the three labels.
- Other factors such as environmental stress, mechanical damage, or chemical exposure may also cause damage to crops, but may not be captured by the current labels.

After extensive discussions, the project team decided to opt for more general labels, e.g. “pest”, instead of specific labels for certain diseases, e.g. Helopeltis. The reasons for this decision were related to the AI use case for which the data was planned to be further used (a more general early-warning system for farmers) and also a weighting of the available resources, e.g. for collecting and annotating data.

Can you provide a brief description of the number of images and instances in the dataset, as well as any skewness or imbalances in the data?

The dataset is significantly skewed towards the abiotic class, which could potentially introduce bias into machine learning models trained on the data. To address this issue, we took experimental measures during model development, such as augmenting other classes to balance the data. However, we found that preserving the skewness was also important, as it reflected the higher occurrence of abiotic factors in a typical farm. This approach helped the model recognize the prevalence of abiotic factors without overemphasizing their importance in predicting other classes.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, the data set contains all the required information.

Is the data set self-contained?

Yes, the data is self-contained, it does not rely on any other external sources.

Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate. Does the data set contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals’ nonpublic communications)? If so, please provide a description.

No, the datasheet is not confidential and it is self-contained.

Does the data set contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the data set relate to people?

No. Given the slight possibility to use the GPS location of cashew trees to identify individual farms, the project team decided to strip the GPS location data in the published dataset.

F.3. 3.3 Collection process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age, or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance was acquired from Bono region cashew farms in Ghana.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The images for the cashew data collection process were captured using a drone that was flown manually. The drone was flown at different altitudes to ensure that comprehensive information about the cashew crops was gathered. The photos of the cashew crop 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.

What quality management systems were put in place to ensure the validity, accuracy, and reliability of data collected?

1. Careful selection of flight altitude and angle to ensure comprehensive data collection.
2. Use of high-quality drone cameras to capture images of the cashew crops.
3. Careful storage and recording of images, including GPS coordinates and timestamps, to provide a record of location and time.
4. Exclusion of blurry or indistinct images, as well as those with defects such as low lighting, to maintain the quality of the dataset.

What were the challenges faced in using drones for data collection and how did they impact the dataset?

Challenges faced in data collection using drones:

- Limitations of battery life and flight time
- Weather conditions affecting drone performance
- Difficulties in navigating terrain with obstacles
- Drone whirling up leaves and thus making steady pictures difficult

Impacts on the dataset:

- Limited coverage area due to battery life and flight time restrictions
- Incomplete data due to weather conditions and obstacles
- Need for manual inspection and correction of data to ensure accuracy

To overcome these challenges, the team implemented measures such as carefully selecting flight paths, checking weather conditions, regularly calibrating and maintaining drones, and manually inspecting and correcting data (example, if collected data seemed blurry or not visible)

Despite these challenges, the use of drones likely provided a more efficient and cost-effective means of data collection compared to traditional methods (traditional methods may involve manual inspections or measurements of crops.) However, the limitations of drone technology and the need for careful data inspection should also be considered when analyzing and interpreting the dataset.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

The KaraAgro AI team took upon themselves to collect this data.

Over what time frame was the data collected?

The data was collected in two rounds: The first data collection happened in November 2022, the second in January 2023.

Does this time frame match the creation time frame of the data associated with the instances (e.g., a recent crawl of old news articles)? If not, please describe the time frame in which the data associated with the instances was created.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please describe these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, there were no ethical review processes conducted.

Does the data set relate to people? If not, you may skip the remaining questions in this section.

No, the data set does not relate to people.

F.4. 3.4 Preprocessing/ Cleaning/ labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, preprocessing and labeling of the data were done during the data annotation stage using annotation tools (makesense.ai). Preprocessing of the data involved removing crop images in which human figures or faces were accidentally captured. Also blurry images were deleted.

How was it ensured that the annotation of data was performed accurately and efficiently, and what methods were used to validate the data and ensure that the annotations were consistent and of high quality?

The data was labeled by data scientists of KaraAgro AI who worked on this project. To ensure the accurate and efficient annotation of data, the team used advanced annotating tools (makesense.ai, roboflow) that offered various annotation formats (xml, yolo). Before the annotation process began, an expert in Agricultural Science reviewed the cashew images and provided comprehensive training on the annotation process, including appropriate labels (abiotic, insect, and diseased) to assign to each image.

During the annotation process, the team was diligent in checking the images and ensuring that they aligned with the correct labels. Any inconsistent images were sent to the expert for further analysis and suggestions. The team only included clear and high-quality images in the annotated dataset and excluded blurry, indistinct, or defective images to maintain the quality of the dataset. The team also checked and removed any images not related to the desired crop to prevent inaccuracies or confusion in the analysis.

Following the annotation process, the team conducted a thorough peer review of the annotated cashew images to ensure the quality and accuracy of the annotations made by team members. This step was crucial in ensuring that all annotations were consistent and comprehensive, and that the annotated dataset was of high quality.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw unprocessed data (consisting of labeled images) has been saved.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

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

F.5. 3.5 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes, the dataset will be distributed to third parties outside of the entity since it will be made publicly available for different companies, institutions, and organizations.

How will the dataset be distributed (e.g., tarball on the website, API, GitHub)

The dataset can be distributed on Kaggle and Dataverse.

When will the dataset be distributed?

Distribution is planned for May 2023.

What license (if any) is it distributed under? Are there any copyrights on the data?

The data will be licensed under the GNU AGPL license where the credit must be given to the creator, the data can be used for commercial use and adaptations and it will be shared under identical terms.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No, there are no fees or regulatory restrictions that apply to this dataset.

F.6. 3.6 Uses

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

The dataset was utilized for an object detection task, where a model was trained to recognize areas affected by abiotic factors, diseases, and insect infestations of cashew fields.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. What (other) tasks could the dataset be used for?

What were the main challenges faced by the team during the data collection and annotation process, and how were these challenges overcome to ensure the quality of the data collected?

During the data collection and annotation process, the team faced several challenges, such as incomplete or inconsistent labeling, difficulty in distinguishing between similar classes, and limited availability of experts for domain-specific knowledge. To address these challenges, the team implemented several strategies:

- The team established clear guidelines for labeling by inviting Dr.Stephen Torkpor from University of Ghana who is an Agricultural Scientist to provide extensive training to team annotators to ensure consistency and accuracy.
- The team shared ambiguous or uncertain labeling cases with Dr. Stephen Torkpor of University of Ghana for clarification.
- The team conducted regular team-internal evaluations of the annotated data to ensure that the dataset was of high quality and suitable for the intended use.

Below are some guidelines for labeling or annotating the data set by team.

- Provide enough space to capture the affected area without cutting any part off, but avoid introducing too much space.
- If possible, zoom into the affected area to ensure accurate annotation.

Annotation of large affected areas

- Avoid including too much gap between two affected areas. Treat a wide gap as another annotation.
- To ensure completeness, try to annotate every possible instance as long as it is visible.

Careful annotation

- Even for images that appear to have no affected areas, zoom in and carefully examine the image to ensure nothing is missed. It is better to have an annotation than to remove the image altogether. However, if truly nothing exists, discard the image from the database.

Avoid human faces

- Avoid using images with human faces whenever possible.

Cropping out human faces

- If an image with a human face is necessary, but the face is not centered, consider cropping it out and replacing the original image with the cropped version.

Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description? Is there anything a future user could do to mitigate these undesirable harms?

None that the KaraAgro AI team is aware of.

F.7. 3.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

KaraAgro AI will support, host, and maintain the dataset.

How can the dataset owner/curator/manager be contacted (e.g., email address)?

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

Is there an erratum? If so, please provide a link or other access point.

No

Will the dataset be updated (e.g., correcting labeling errors, adding new instances, deleting instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will be open sourced such that people can contribute to it. Any changes to the dataset will be documented in a form of dataset updates

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

The dataset will be hosted on public servers. Older version links will still be available in a form of versioning.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description. Other researchers are allowed to extend this dataset.

The interested researchers should send an email to darlington@gudra-studio.com owned by Darlington Akogo and they will be able to discuss the dataset building, extensions, and contributions.

[Refer to sustainability strategy documentation on how to contribute to this project]

References

1. Patrick Ateah Yeboah, Bismarck Yelfogle Guba, Emmanuel K. Derbile, Smallholder cashew production and household livelihoods in the transition zone of Ghana, Geo: Geography and Environment, 10.1002/geo2.120, 10, 1, (2023).