TETRA TECH

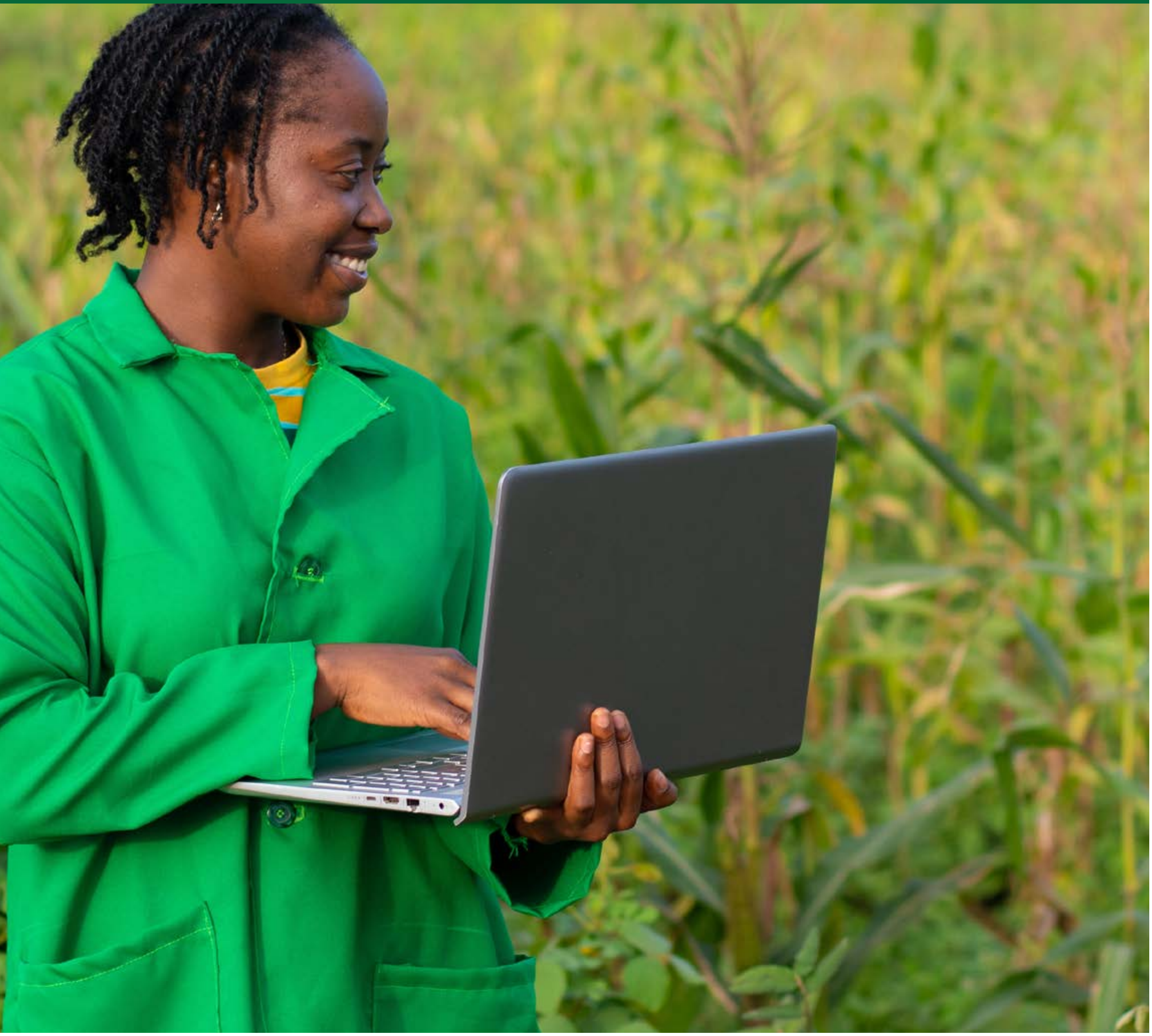
Enabling Crop Analytics At Scale

ENABLING CROP ANALYTICS AT SCALE (ECAAS)

# An Open-Source Framework For Crop Type Mapping In Africa

# Contents

# 1
# Overview



Digital Earth Africa (DE Africa) is a continental-scale, not-for-profit initiative focused on improving access to Earth observation (EO) across sectors in Africa. DE Africa is built on partnerships with regional geospatial organizations, African governments and in-country expertise to create sustained capacity development in Africa. DE Africa is working with the African and international community to ensure that EO data is analysis ready, rapidly available, and readily accessible to meet the needs of our users.

DE Africa is guided by a Governing Board co-chaired and represented by African ministers, advised by a Technical Advisory Committee (TAC) with the majority of its members based in Africa and works closely with the AfriGEO community responding to the information needs, challenges and priorities of the African continent. In 2022, DE Africa transitioned out of the establishment phase led by Geosciences Australia (GA) into a distributed network of implementing partners across Africa with a program management office hosted by the South African National Space Agency (SANSA).

This report provides a summary of the deliverables for the ECAAS project "Open-Source workflow for crop-type mapping in Africa". Our workflow provides an end-to-end framework that allows users to develop their own field sampling approach, collect ground truth data using the newly developed Enabling Crop Analytics at Scale (ECAAS) ODK toolkit, upload ground truth data directly into the DE Africa platform, explore features of importance for crop type separation, and run and assess the accuracy of machine learning algorithms to create crop type maps for their area of interest.

We leveraged the DE Africa platform driven by the African GEO community and stakeholders across the continent to develop our workflow through an established process of co-production within the DE Africa Product Development Task Team (PDTT). We developed our workflow in partnership with the Regional Centre for Mapping of Resources for Development (RCMRD) for a priority use case focused on crop type mapping in Zambia. In partnership with AGRHYMET, one of the DE Africa Implementing Partners, we are currently testing out extension of the workflow into Niger.

Our open-source framework allows for future innovation in the development of EO solutions for increasing food security, which can be extended and scaled to other regions in Africa through the PDTT and our user network. Specific objectives of the framework include the demonstration of an end-to-end workflow comprising the following steps:

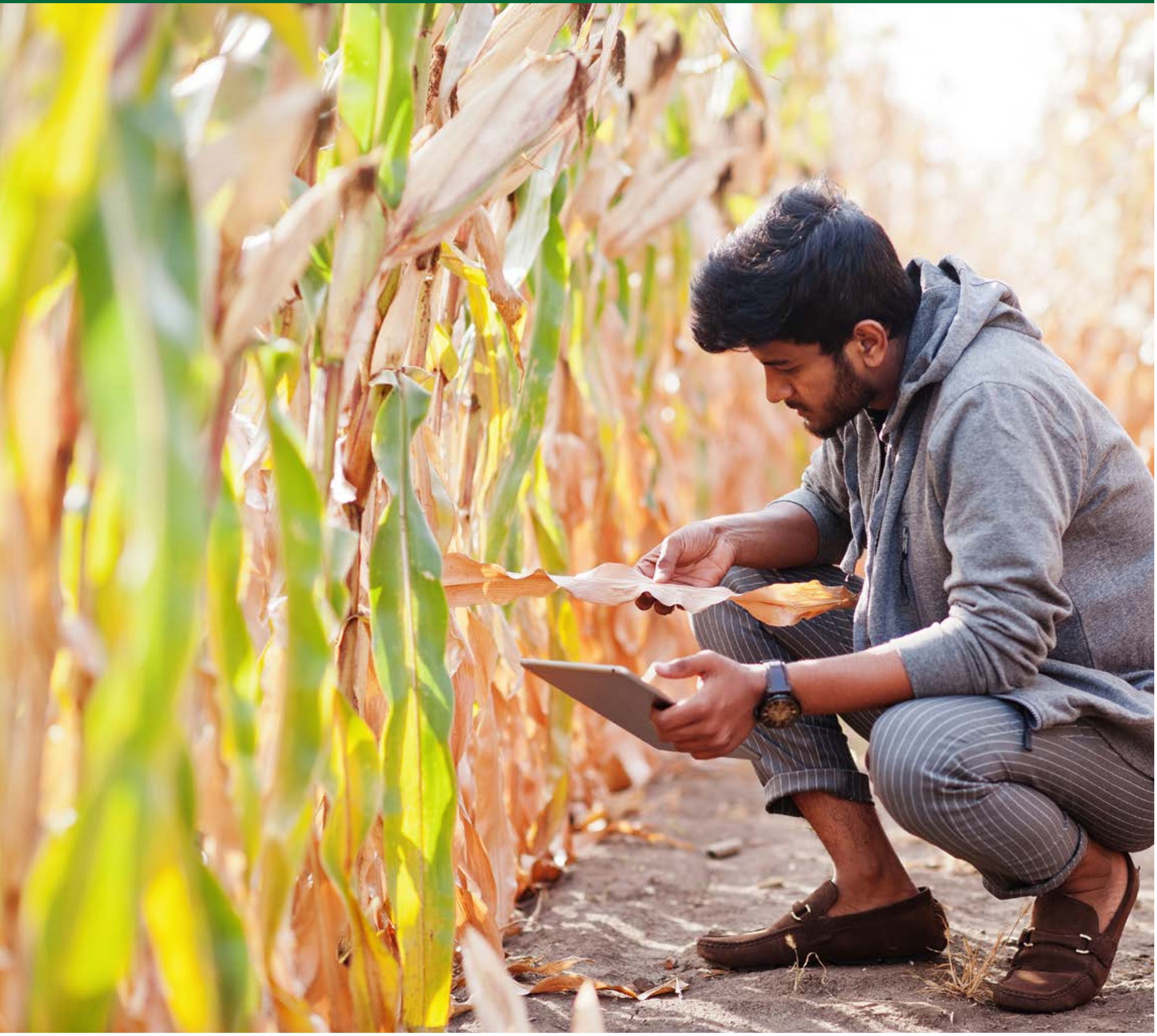
1. **Sampling:** Creation of representative sampling strategies through the combination of the newly created continental DE Africa crop mask dataset and unsupervised machine learning.
2. **Data Collection:** Collection of ground truth data on crop types using the newly created ECAAS ODK toolkit, and tools to pre-process ground truth data for machine learning on the DE Africa sandbox.
3. **Feature Exploration:** Tools to identify important features and understand correlation between them, which helps to select key features for machine learning models.
4. **Model Development:** Training of machine learning models and performance estimation.
5. **Map Production:** Development of crop type map for areas of interest using trained machine learning models.

This framework can be easily adapted, built upon, and is open and freely available to all users. Our workflow is well documented through a series of Jupyter notebooks that will be used as part of our capacity development efforts. We have identified universal data inputs layers useful for crop type mapping that DE Africa can provide as a continental-scale service for scaling and extending workflow to other areas (e.g., statistical EO products). We have brought together key partners in EO and the food security sector to inform our framework and will promote continued innovation and collaboration. This is the first step towards operational crop monitoring and yield prediction services in Africa.

Our project increases accessibility to EO solutions for food security issues through development of an open-source crop-type mapping workflow that is built on Digital Earth Africa's continental EO products, allowing it to be applied anywhere in Africa.

# 2

# Workflow



# 2.1 Summary

The primary goal of this project was to develop an end-to-end workflow that enabled users to collect representative crop type data, train machine learning models, and use the trained models to create crop type maps over their area of interest. The workflow is designed to be generic enough that a user with little machine learning experience can run a robust workflow and expect a reasonable outcome, and modular enough that a user with more expertise can customise different steps of the workflow. The workflow uses best practice approaches like nested cross-validation for model comparison (appropriate for the small sample sizes often seen in labelled crop type datasets), and data transformation pipelines to allow for customised selection of features during the training stage without data leakage from the hold-out set.

The workflow is designed to be used in the Digital Earth Africa sandbox, which has access to several continent-wide analysis-ready satellite datasets and derivative products. The workflow is provided as a collection of documented Jupyter notebooks in an open GitHub repository, along with a copy of the data and supplementary files required to run the workflow for Central province in Zambia. This ensures that a new user can run the workflow “as is” to learn how it works, before applying it to their own data

### The workflow steps are:

1. Sampling design – notebooks
2. Field data collection — ECAAS ODK toolkit
3. Data preparation – notebook
4. Feature extraction – notebook
5. Feature exploration – notebook
6. Machine learning training and performance estimation – notebook
7. Review of trained model on test areas – notebook
8. Production of crop type map for area of interest — notebook

Details on methods for all the workflow steps are included in the Detailed Methodology Annex associated with this Final Report.

## Clean Boundary Data for Machine Learning-Based Data Analysis

Private companies operating in African countries do not know which smallholders are using their products, since seeds and fertilizer travel through multiple distributors, retailers, and agents before being used on small (sub-hectare) plots. Lack of information and field records also keeps farmers from knowing if the inputs they used improved their yields or if a specific crop was profitable. Analysts can use machine learning-based algorithms to create new datasets that can support decision making.

The ability to visualize crop type, variety, yield and individual field boundary, visualize crop mapping/classification, and linking these locations to simple indices that provide evidence of vegetation on each field is the main selling point for commercial users. Being able to find, download and share field data, validate its accuracy and completeness in an easy, online platform brings great value to users.

Machine learning (ML) is a form of artificial intelligence that imitates how humans learn to make predictions from data. 6th Grain (6G) and partners work in data scarce environments with very few reliable datasets that enable and accelerate business decisions; as such, ML algorithms are central to producing timely and affordable datasets.

### All Ml Systems Have Three Basic Steps:

1. A problem statement: ML algorithms are used to make a prediction or classification which results in improved insights or better decisions. The ML algorithm uses input data, which can be labeled or unlabeled, to produce an estimate about a pattern in the data.
2. An error estimation: This function serves to evaluate the accuracy of the model and to improve it as more data becomes available.
3. A model optimization process: By understanding the accuracy requirements of the business decision, the model can be modified to better fit the data points in the training data set via weighting functions to reduce the discrepancy between the known example and the model estimate. By using accuracy thresholds, the model can be optimized to reduce cost while increasing the usability of the output (UC Berkeley, 2020).

All ML processes require high quality training data through which algorithms can be built. The problems that can be solved by ML models will increase as more data become available. A key objective of the DSP project is to expose more data for ML experts across public, provide and academic sectors. In this section we describe use cases for the data provided in the DSP and how it can be used to improve decision making in the agriculture sector.

## 2.2 Application of workflow to Central province, Zambia

This section provides a summary of how we leveraged our workflow to classify crop types across Central province in Zambia.

### 2.2.1 Sampling design

The first step was to generate a sampling strategy to guide field collection of ground truth points. It is important to get a representative sample of crop types in the region to reliably assess the performance of any developed machine learning methods. We used an unsupervised classification method (k-means clustering) to identify groups according to spectral variability, then used this to stratify our suggested sample. From this, we produced a list of suggested sampling sites for the in-country team to visit.

### 2.2.2 Field data collection

The field campaign for this project was conducted between the 10th and the 29th of April 2022 by RCMRD and the Zambia Ministry of Agriculture, using the ECAAS toolkit. Three field teams covered the 12 districts of Central province, meeting with District Agricultural Coordination Officers (DACO) who are knowledgeable of their respective administrative units. The DACO guided field movements and advised on the route plans and target areas. Figure 1 shows snapshots of the field data collection exercise by the staff of the Ministry of Agriculture. Over 1,000 Crop Labels were captured across all districts in the Central province.

**Figure 1:**
Crop labels collection using the ECAAS Toolkit.

### 2.2.3 Data preparation

After field collection, the data were reviewed for quality assurance, and then cleaned to be used for machine learning. We focussed on cleaning crop type points rather than crop type polygons, as they comprised the majority of our sample (1081 points, 188 polygons). We discuss this further in Section 5.2. After cleaning, our sample contained 901 labelled crop points.

### 2.2.4 Feature extraction and exploration

For each cleaned crop type point, we calculated 113 features to use for machine learning. Our feature space included the following sources:

- Sentinel-2 GeoMAD and key indices for three-, six-, and twelve-month periods
- Sentinel-1 geomedians for three-month periods
- Landsat Fractional Cover medians for three-month periods
- CHIRPS rainfall averages for three-month periods
- SRTM Digital Elevation Model slope

After calculating all features, we used a Random Forest model to investigate feature importance and identified that rainfall was ranked as the most important feature in distinguishing between crop types. The scores for the top 20 features are shown in **Figure 2.**

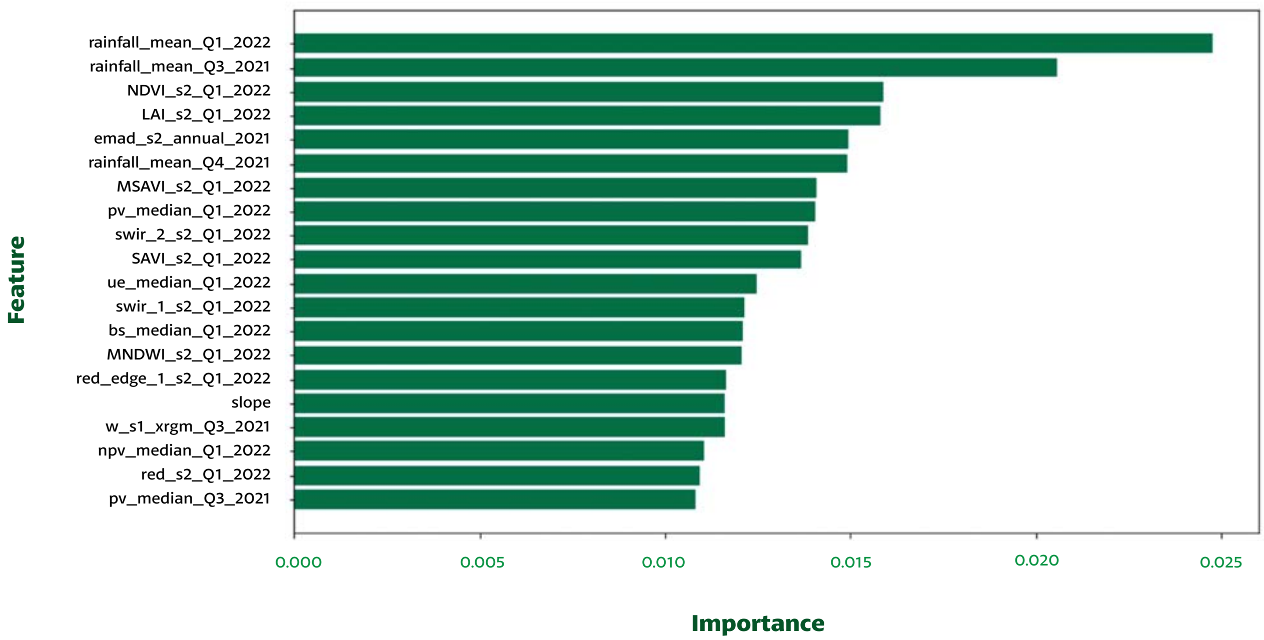



**Figure 2:**
Top 20 (out of 113) features when fitting an untuned Random Forest model to all available data. In feature names, "s2" indicates the feature was derived from Sentinel-2; "s1" indicates the feature was derived from Sentinel-1; The "bs", "pv", "npv", and "ue" correspond to Landsat Fractional Cover products, and the "rainfall" prefix corresponds to CHIRPS average rainfall. "Q3_2021" corresponds to Aug-Oct 2021, "Q4_2021" corresponds to Nov 2021-Jan 2022, and "Q1_2022" corresponds Feb-Apr 2022.

### 2.2.5 Machine learning training and performance estimation

After calculating all features for our crop type points, we fitted two supervised machine learning models (Random Forest and AdaBoost). Given the small size of our sample data, we implemented nested cross-validation to allow us to compare the performance of both models after hyperparameter tuning. To reduce redundancy in the features, we also identified and removed highly correlated variables.

After training both models through nested cross-validation, we found that the Random Forest algorithm outperformed the AdaBoost algorithm. To assess model performance, we used the macro F1 score, which considers both the precision and recall of each class and weights all classes equally. Over the three outer cross-validation folds, the Random Forest algorithm achieved an average macro F1 score of 0.84 (standard deviation of 0.01); the AdaBoost algorithm achieved an average macro F1 score of 0.80 (standard deviation of 0.01).

### 2.2.6 Creation of crop type maps

After training the final model, we used the Digital Earth Africa Sandbox platform to produce the crop type map for Central Province in Zambia, along with a map of model confidence.

To make effective use of the Sandbox's available computing power, the region was processed on a tile-by-tile basis, which ensured all data needed for prediction could be loaded into computer memory.

After running the machine learning prediction step, we masked the outputs using the Digital Earth Africa continental crop mask product, meaning we only kept regions that had been previously identified as crop in 2019.

# 3
# Results

## 3.1 Feature importance

**Figure 3** shows the top 20 features for the final model, with the ranking of all features shown in the Appendix of the Detailed Methodology Annex. The feature importance for the final model revealed that the top three features were average rainfall values, which suggested that the crop types in our study were challenging to distinguish from multispectral satellite data alone. The importance of rainfall may indicate that there are key geographical distributions of cropping that are correlated with the expected rainfall for a given region.

Following rainfall, Fractional Cover indices, such as the proportion of photosynthetic vegetation (pv), non-photosynthetic vegetation (npv) and bare soil (bs) in different time periods featured highly in the model importance. This highlights the value of derivative satellite products for crop type classification.

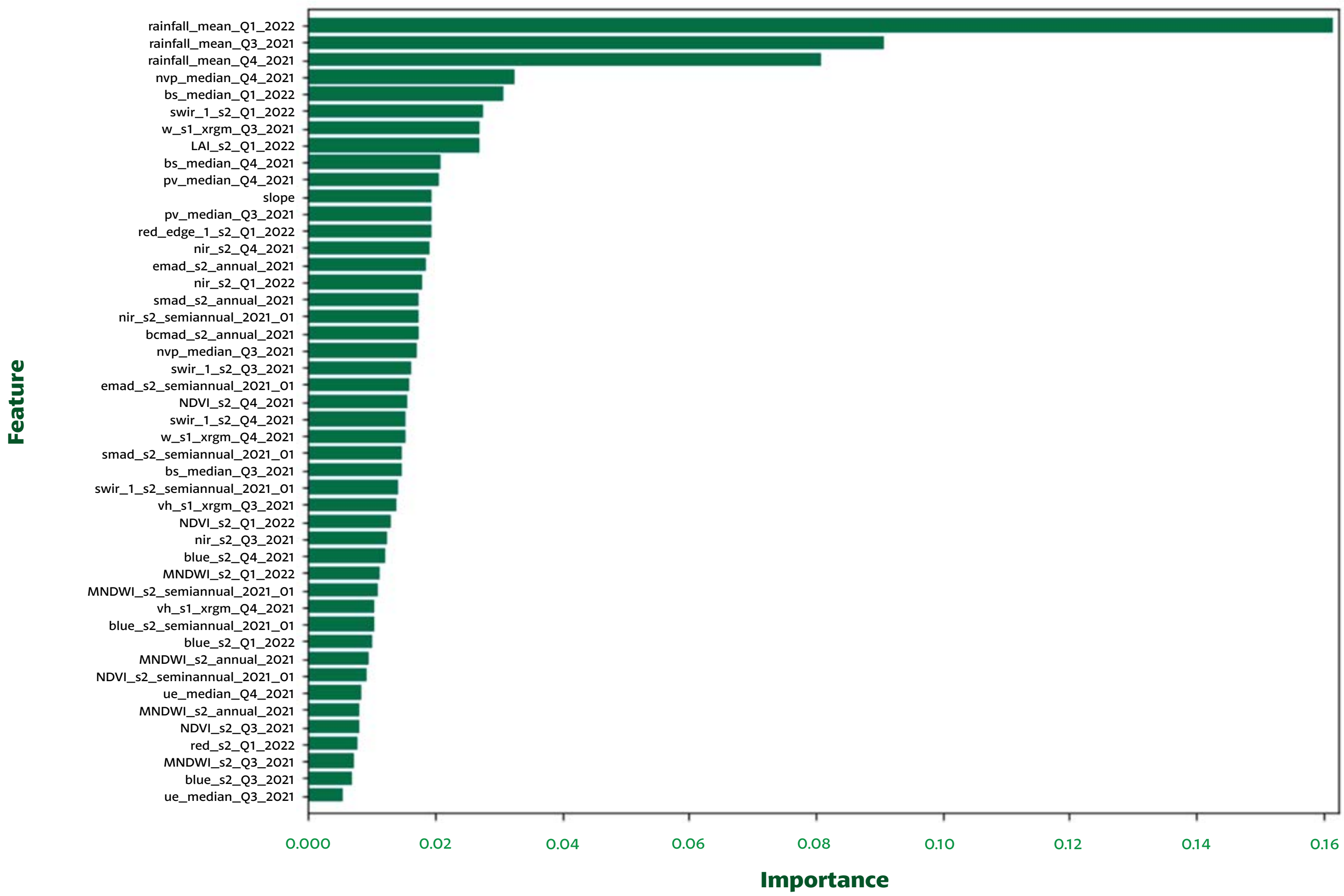



**Figure 3:**

Relative importance of the top 20 features used in the final Random Forest model. In feature names, "s2" indicates the feature was derived from Sentinel-2; "s1" indicates the feature was derived from Sentinel-1; The "bs", "pv", "npv", and "ue" correspond to Landsat Fractional Cover products, and the "rainfall" prefix corresponds to CHIRPS average rainfall. "Q3_2021" corresponds to Aug-Oct 2021, "Q4_2021" corresponds to Nov 2021-Jan 2022, and "Q1_2022" corresponds Feb-Apr 2022.

## 3.2 Accuracy assessment

As discussed in **Section 2.2.5**, we used nested cross-validation to estimate the performance of our model when exposed to unseen data. Our overall estimate of performance comes from taking the average macro F1 score for all three outer folds of our cross-validation approach. For our Random Forest model, we achieved an average macro F1 score of 0.84, with standard deviation of 0.01.

It is also useful to look directly at the confusion matrices for each outer fold, as they can give insight into class-level performance. In this section, we investigate the first fold only, and provide the results from the other two folds in the **Appendix section** of the Detailed Methodology Annex.

**Table 1** shows the confusion matrix values, along with the User's, Producer's, and overall accuracies. **Figure 4** shows the row-normalised confusion matrix, indicating the proportion of samples that were predicted as a given class, relative to the total number of true samples for that class.

**Table 1:**

The confusion matrix for the first outer fold of the Random Forest model, showing the total number of test samples for the fold (N), as well as overall (O), User's (U) and Producer's (P) accuracies.

| | beans | cassava | cotton | ground-nut | maize | millet | sorghum | soybean | sunflower | sweet potatoe | Total | p |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| beans | 47 | 3 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 57 | 0.82 |
| cassava | 0 | 83 | 0 | 2 | 15 | 0 | 0 | 5 | 0 | 0 | 105 | 0.79 |
| cotton | 0 | 0 | 53 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 60 | 0.88 |
| groundnut | 0 | 0 | 1 | 37 | 9 | 0 | 0 | 1 | 3 | 0 | 51 | 0.73 |
| maize | 2 | 1 | 0 | 2 | 1416 | 7 | 3 | 33 | 5 | 10 | 1479 | 0.96 |
| millet | 0 | 0 | 0 | 3 | 2 | 16 | 0 | 8 | 1 | 0 | 30 | 0.53 |
| sorghum | 0 | 2 | 0 | 0 | 6 | 0 | 31 | 0 | 0 | 0 | 39 | 0.79 |
| soybean | 4 | 7 | 0 | 6 | 74 | 4 | 0 | 445 | 0 | 0 | 540 | 0.82 |
| sunflower | 0 | 0 | 0 | 0 | 22 | 0 | 0 | 0 | 179 | 0 | 201 | 0.89 |
| sweet potato | 0 | 1 | 0 | 0 | 21 | 0 | 0 | 1 | 0 | 109 | 132 | 0.83 |
| **Total** | | | 54 | 50 | 1579 | 27 | 34 | 493 | 188 | 119 | **N**2694 | |
| **U** | | | 0.98 | 0.74 | 0.90 | 0.59 | 0.91 | 0.90 | 0.95 | 0.92 | | **O**0.90 |

| True Label | beans | cassava | cotton | groundnut | maize | millet | sorghum | soybean | sunflower | sweet potatoe |
|---|---|---|---|---|---|---|---|---|---|---|
| beans | 0.82 | 0.53 | 0 | 0 | 0.12 | 0 | 0 | 0 | 0 | 0 |
| cassava | 0 | 0.79 | 0 | 0.019 | 0.14 | 0 | 0 | 0.048 | 0 | 0 |
| cotton | 0 | 0 | 0.88 | 0 | 0.12 | 0 | 0 | 0 | 0 | 0 |
| groundnut | 0 | 0 | 0.02 | 0.73 | 0.18 | 0 | 0 | 0.02 | 0.059 | 0 |
| maize | 0.00140 | 0.0068 | 0 | 0.0014 | 0.96 | 0.0047 | 0.002 | 0.022 | 0.0034 | 0.0068 |
| millet | 0 | 0 | 0 | 0.1 | 0.067 | 0.53 | 0 | 0.27 | 0.033 | 0 |
| sorghum | 0 | 0.051 | 0 | 0 | 0.15 | 0 | 0.79 | 0 | 0 | 0 |
| soybean | 0.0074 | 0.013 | 0 | 0.011 | 0.14 | 0.0074 | 0 | 0.82 | 0 | 0 |
| sunflower | 0 | 0 | 0 | 0 | 0.11 | 0 | 0 | 0 | 0.89 | 0 |
| sweet potato | 0 | 0.0076 | 0 | 0 | 0.16 | 0 | 0 | 0.0076 | 0 | 0.83 |

**Predicted Label**

**Figure 4:**
The row-normalised confusion matrix for the first outer fold of the Random Forest model. The Producer's accuracy appears along the diagonal of the matrix. Each cell is coloured according to its value, with yellow representing high values, and blue representing low values.

While the overall accuracy score of 0.90 and macro F1 score of 0.84 indicate strong performance, the confusion matrices reveal challenges with the model. Displayed most obviously in the Detailed Methodology Annex, each non-maize class has a reasonable proportion of its samples predicted as maize (between 6% and 18%), which is likely due to the imbalanced nature of our sample, with maize representing around 50% of all samples. Millet was the hardest class to predict (Producer's accuracy of 53%), likely because it had the smallest presence in our sample. Interestingly, millet is more often confused with soybean than maize, with 27% of millet samples predicted as soybean and 6.7% of millet samples predicted as maize.

# 3.3 Map results

While the accuracy assessment provides a quantitative assessment of map quality, a qualitative visual assessment is also important and can identify areas where the model is failing for future improvement. The RCMRD team identified areas where the model is working **(Figure 5)** and where the model is failing (Figure 6). The RCMRD team will address issues of model failing in future iterations through additional data collection and model refinement.

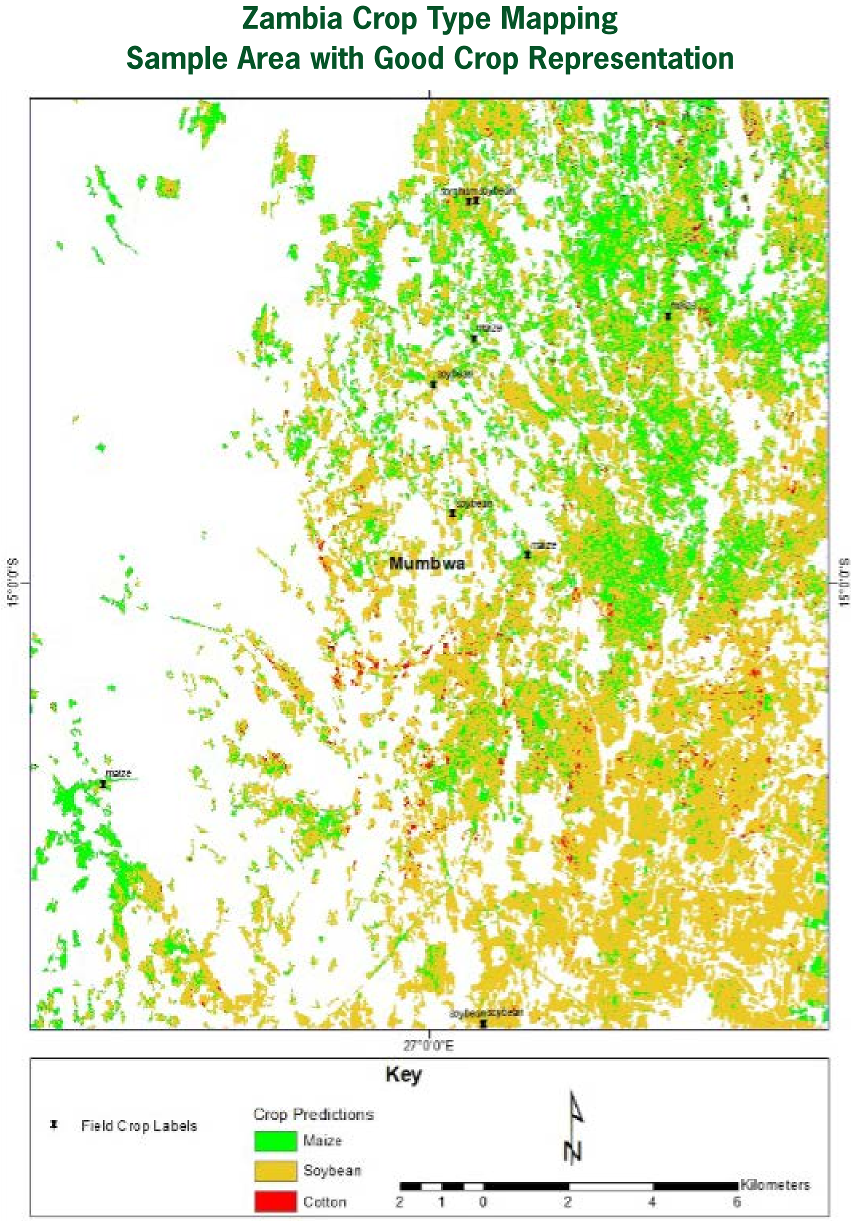



**Figure 5:**
Area with high accuracy in mapping crop types.

**Zambia Crop Type Mapping**
**Sample Area with Incorrect Crop Representation**

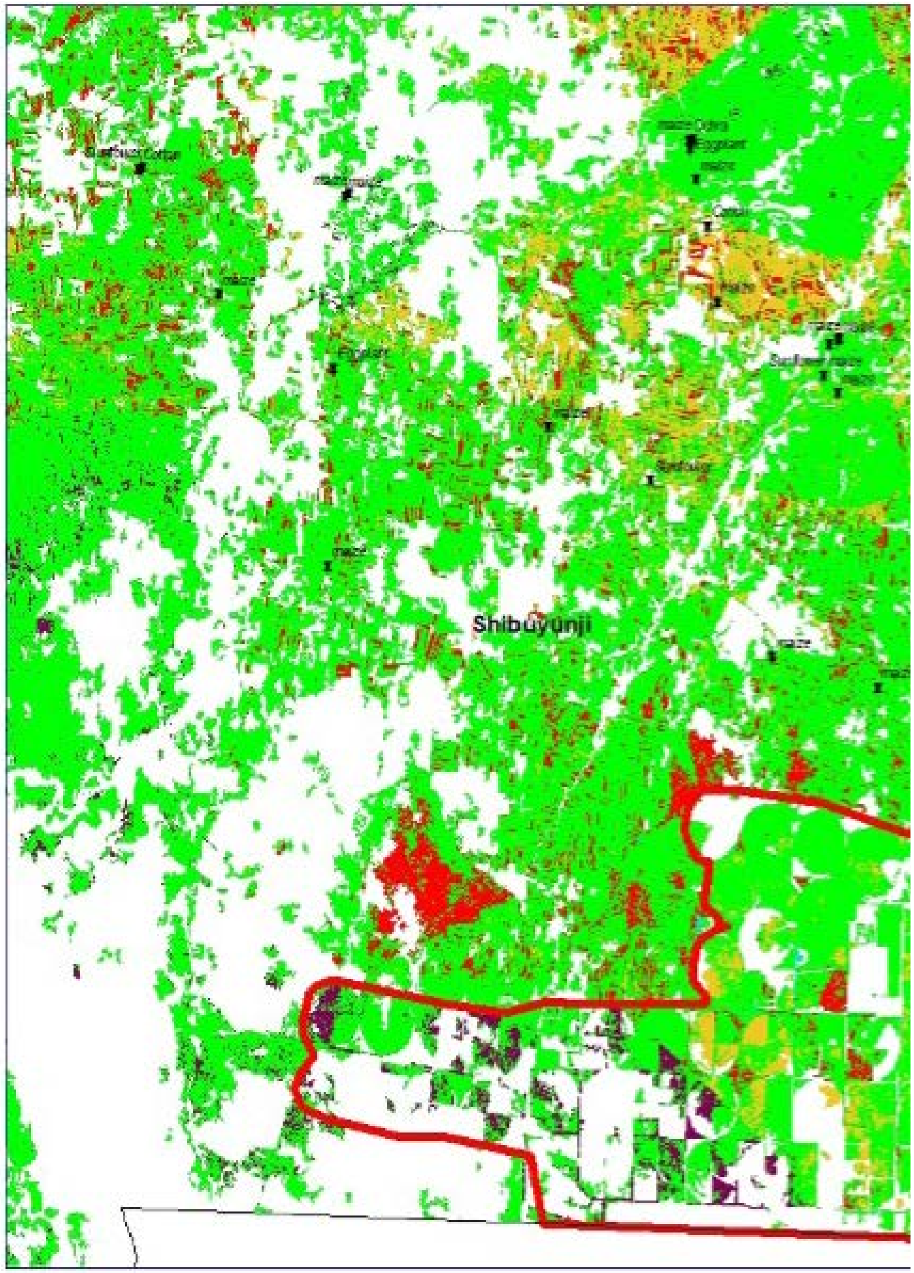



**Figure 6:**
Area where the current workflow is not working correctly and will be corrected as part of the ongoing iterative process. In this example, the circled red area is incorrectly mapped as maize instead of sugarcane. Sugarcane was not included as crop type in our model as we only had 1 crop type sample of sugarcane.

While not visible in the above examples, our produced map had inconsistencies in predicted crop types from the presence of Landsat 7 striping in the Fractional Cover product, as well as tiling artifacts from the large-scale map production. Our project team is currently investigating these issues, to be remediated in future versions of the workflow.

# 4

# Extending and scaling workflow to other African countries

The goal of this project was to develop an open-source flexible workflow that will allow countries to develop EO solutions for crop type mapping. Therefore, throughout the project, we have communicated progress and results with the DE Africa Product Development Task Team. The PDTT allows for peer-to-peer knowledge transfer and extension of the crop type mapping workflow. Currently, the workflow is being implemented with the AGRYHMET team to map crop types in Niger. The first components of the workflow have been completed including development of unsupervised training, sampling design, and field data collection. **Figure 7** shows the results from the unsupervised classification element of the sampling design work, clearly indicating different regions of spectral variability across Niger.

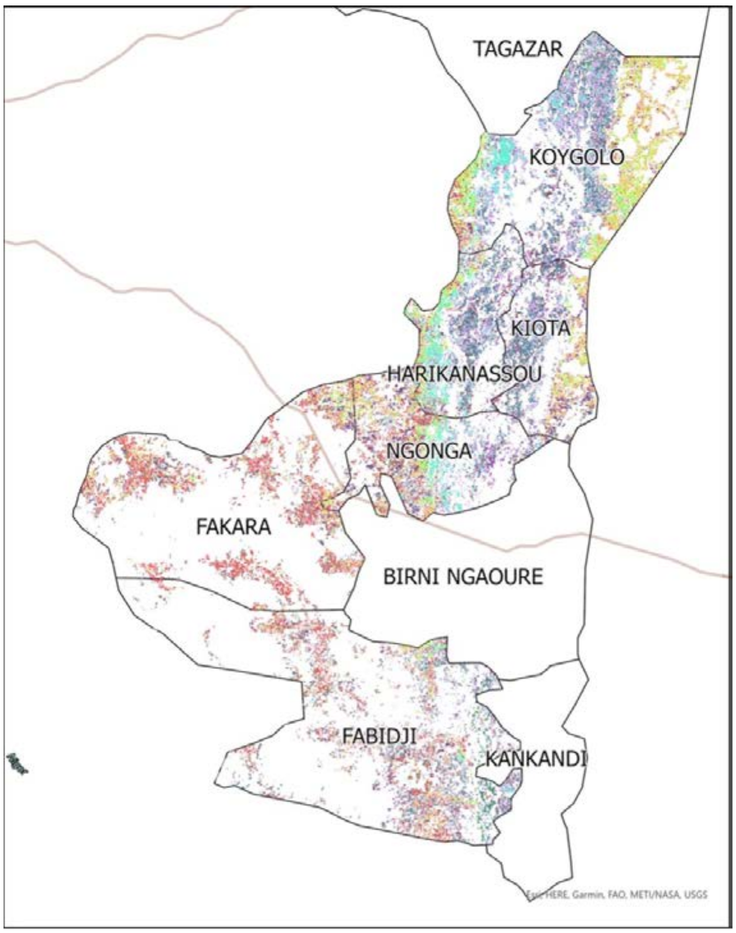



**Figure 7:**
Unsupervised classification using DE Africa crop mask for area of interest in Niger. Different colours represent the 15 classes and demonstrate the spectral variability across the region. Unsupervised classification was used to facilitate a balanced sample design during the field campaign.

DE Africa is working with the AGRYHMET to train them on adapting and running the Jupyter notebooks to create crop type maps for Niger **(Figure 8)**. To facilitate learning and extension of workflow to the DE Africa network, AFRIGIST based in Nigeria, will be integrating components of the workflow into online training modules.

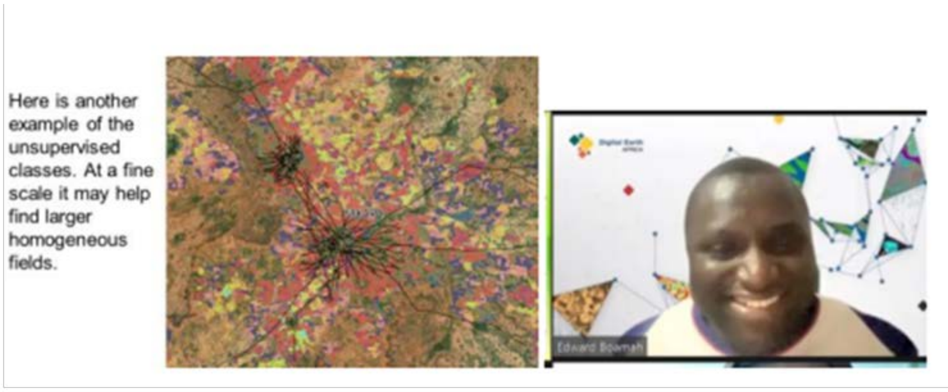



**Figure 8:**
Screenshot of the workflow training for the AGRYHMET team led by Edward Boamah (above) and photo of the AGRYHMET team in the training (below).

# 5

# Key Findings

## 5.1 Sampling strategies must focus on areas around main roads

Many small roads were inaccessible due to poor weather, meaning that the sampling design needed to be adapted on the fly to make the best use of the time. In the future, stratification should still be performed relative to the whole area, but the sampling strategy should focus on collecting representative points from as close to main roads as possible and look to collect multiple points within short distance of each other to reduce the amount of time spent driving between sampling locations. In Zambia, data collection was along a main thoroughfare in districts where the field team received permission. In Niger, field data collection occurred around targeted villages. In both cases, the unsupervised classification aided in collecting a representative sample dataset.

## 5.2 The ECAAS ODK toolkit is useful, but polygon collection is challenging

We tested the ECAAS ODK toolkit in Zambia. This is a useful tool for our program as it helps standardize the data collection effort and allows for easier integration into our workflow.

However, polygons collected through the ECAAS ODK toolkit were not usable in many instances because of multiple errors, such as the capture of small polygons (equivalent to a single Sentinel-2 pixel), invalid topologies, and errors in position. See Appendix 1 for examples.

## 5.3 Crop confusion is common when working with multispectral satellite data

We found that rainfall ranked highest in the Random Forest feature importance score, suggesting we did not have sufficient information from multispectral satellite data to distinguish between some crop types. This may be due to multiple crops having similar planting and harvest times (see Detailed Methodology Annex), meaning the model relied on other geographic factors to distinguish between crop types.

## 5.4 Continental products are the key to having a scalable workflow

At the time of the project, DE Africa provided annual and semi-annual geomedians and median absolute deviation products, which act as scientifically valid cloud-free composites, as well as capturing variation in the landscape. Our workflow also calculates these products over three-monthly periods, but this calculation took significant processing time, meaning map processing took days instead of hours. To have an easily useable and scalable workflow, Digital Earth Africa should continue to focus on producing continental scale products, reducing the processing time for users across Africa.

## 5.5 The PDTT is an important framework for extending workflows and knowledge sharing between African organizations.

The PDTT structure allows not only for dissemination of components of the workflow, but also provides structure for support and knowledge sharing amongst the African geospatial community. The PDTT, with the support of the DE African technical and science leads, can support sustainable co-development processes that have local ownership and build in-country capacity. It's important to highlight the capacity that has been built through this project. For the RCMRD team capacity was built in co-designing the workflow and methodologies, training the team on the workflow components, and extending training of workflow components (i.e., ECAAS ODK toolkit, sampling methods) to the Ministry of Agriculture in Zambia. For the AGRHYMET team capacity was built on implementing the workflow through training with the DE Africa technical lead and from RCMRD. For all the PDTT partners capacity was built on using the ODK toolkit, through a training led by RCMRD. We also expect future capacity building of this workflow through training and collaborative work, and peer-to-peer knowledge exchange through the PDTT.

# 6

# Next steps for crop type classification in Zambia

## 6.1 Increase training data

A shift in the project start date caused field sampling to start later in the season and our field efforts were cut short. While we were still able to collect over 1,000 points and further increase the sample size by using multiple pixels per point, we had challenges with predicting minority classes, including significant confusion with the majority class of maize. This can be addressed both through collecting additional points for minority classes, and class balancing options, such as selecting fewer maize crop samples for model training.

Further model iterations will include additional points from ECAAS partners and the NASA SERVIR program.

Directly related to the work in Zambia, RCMRD through the SERVIR ESA project is seeking to extend crop type mapping activities to other districts in Zambia in the coming cropping season. This initiative will be preceded by the training of Ministry of Agriculture staff on crop labels collection before the real works begin. In this initiative, RCRMD is working to increase the number of crop labels to covers areas beyond Central province.

RCMRD has also obtained more data from the Bird's-Eye project. This is a collaboration between Pula and Tetra Tech under the Enabling Crop Analytics at Scale Project. Just like the Zambia Crop type mapping project, Bird's-Eye project is meant to foster the collection of ground truth georeferenced agriculture labelled datasets that would improve predictive analytics in small-holder agriculture, using artificial intelligence and machine learning (AI/ML) and remote sensing.

RCRMD is also looking at ways to bring on board the LACUNA Helmets Crop Labels initiative which is already ongoing in Uganda, Kenya, Tanzania and Rwanda. Crop labels densification process entails the development of a common GIS database of crop labels for Zambia. At the moment, most of the captured labels are in point form and RCMRD is converting them into polygons to cover more cropland using high resolution Planet and Sentinel imagery. This will provide more pixels for training machine learning models.

# 6.2 Improve the model

While this project has focussed on developing a general workflow for anywhere in Africa, there are specific steps that can be taken to improve the crop type model for Zambia.

### Remove Landsat 7 from provisional Fractional Cover features

While Fractional Cover was a significant input feature, it created striping effects in the crop type map due to the inclusion of Landsat 7 scenes. The model for Zambia will be updated to remove the Landsat 7 scenes when calculating Fractional Cover features. This change will also be applied to the general workflow.

### Investigate class balancing

In conjunction with the collection of additional training data, future model development should include a round of testing to see whether reduction of maize samples in the training set results in less confusion between maize and other crops.

### Propose method for working with low sampling numbers

Key crops, like sugarcane, were not included in the Zambia model due to having insufficient samples for running the training workflow with cross-validation. This resulted in the incorrect prediction identified in Figure 6. Future work could design and investigate methods for increasing the number of points for these classes, either through additional field collection, or statistical sampling.

# 7

## Next steps for workflow

Our workflow was designed for further iteration and improvement. We have identified the following activities that will increase model accuracy, fix visual issues, and increase user uptake across Africa.

## 7.1 Develop a continental-scale rolling GeoMAD service

A big challenge for our team was being able to run multiple iterations of machine learning. This is mainly due to the processing time of calculating several input datasets, such as the 3-month Sentinel-2 geomedian. Nevertheless, the geomedian had high model importance when we tested it as an input variable. Our project team have relayed this information to the DE Africa Lead Scientist and PDTT. Pre-calculating a rolling 3-month GeoMAD product will improve the efficiency of model runs allowing for easier model iteration. The DE Africa Technical Advisory Committee (TAC) recently agreed to produce a 3-month rolling GeoMAD as a continental service. This continental service will be available as an open-source dataset for other including those interested in crop type mapping or other food security issues.

## 7.2 Address tiling artifacts

Artifacts from tiling were evident in the preliminary workflow outputs. We are investigating this issue and will adapt the tiling process to remove this issue from the workflow. Tiling was necessary to produce the product for the entire area of interest due to the computational memory limits of the Digital Earth Africa sandbox.

## 7.3 Promote workflow

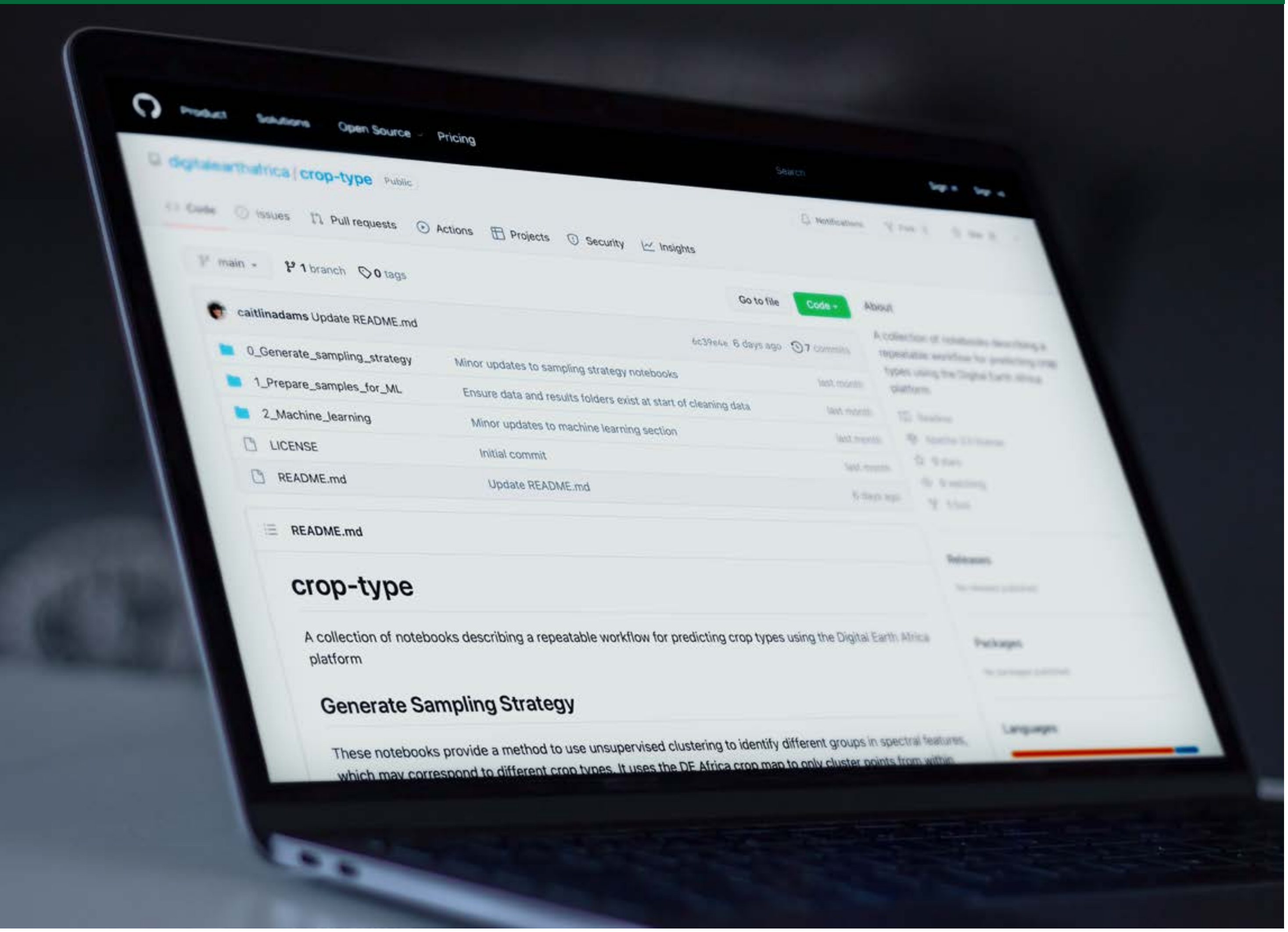
A critical next step for DE Africa upon completion of this project is to promote the workflow for broader use and application. The DE Africa team will do this in several ways:

1. Workflow components will be integrated into our online training modules developed by African Regional Institute for Geospatial Information Science and Technology (AFRIGIST).
2. DE Africa will walk through components of the workflow during the DE Africa weekly live sessions with users across Africa.
3. DE Africa team members will present the workflow at the 2022 AGU scientific conference to share with the broader international scientific community.
4. DE Africa will continue to work with additional partners and countries that are interested in applying the workflow to other parts of Africa.
5. DE Africa will continue to work with the SERVIR program and other ECAAS project partners in building capacity for African-led EO solutions.

# 8

## Data Availability

We provide a csv file containing the cleaned output from the ECAAS toolkit as part of the open-source GitHub repository. The labelled crop type dataset has been submitted to Radiant Earth to increase open access and make it more widely available.



# 9
# Appendix

## 9.1 Examples of polygon challenges

**Appendix 1:** Examples of polygons collected during field sampling shown in red, with Sentinel-2 is used as the reference imagery. Top left demonstrates that polygons captured are much smaller than visible field sizes in Sentinel-2 imagery, with bottom left showing that polygons are smaller than a single pixel. The top right and bottom right images show unusable polygons, due to being narrow or having invalid topologies (self-intersections).

Enabling Crop Analytics At Scale

# An Open-Source Framework For Crop Type Mapping In Africa

**info.ecaas@tetratech.com**

**cropanalytics.net**

**Final report**

October 13, 2022

**Prepared by:**

Dr. Meghan Halabisky, Senior Science Advisor, Digital Earth Africa
meghan.halabisky@digitalearthafrica.org

Dr. Caitlin Adams, Senior Data Scientist, FrontierSI
cadams@frontiersi.com.au

David Ongo Nyang'acha, Senior GIS Analyst, RCMRD
dnyangacha@rcmrd.org