



OPEN

DATA DESCRIPTOR

# Helmet Labeling Crops: Kenya Crop Type Dataset Created via Helmet-Mounted Cameras and Deep Learning

Catherine Nakalembe<sup>1,6</sup>✉, Ivan Zvonkov<sup>1,6</sup>, Hannah Kerner<sup>2</sup>, Diana Botchway Frimpong<sup>1</sup>, Kenneth Mwangi<sup>3</sup>, Jane Kioko<sup>4</sup>, Bhanu Tokas<sup>2</sup>, Kartik Jawanjal<sup>2</sup>, Iman Akhtar Smith<sup>1</sup>, Anjali Paliyam<sup>1</sup>, Christopher Atsianzale Wakhana<sup>1</sup>, Ana María Tárano<sup>2</sup>, Shreya Jha<sup>1</sup> & Priscilla Mawuena Loh<sup>5</sup>

Accurate, up-to-date agricultural monitoring is essential for assessing food production, particularly in countries like Kenya, where recurring climate extremes, including floods and droughts, exacerbate food insecurity challenges. In regions dominated by smallholder farmers, a significant obstacle to effective agricultural monitoring is the limited availability of current, detailed crop-type maps. Creating crop-type maps requires extensive field data. However, the high costs associated with field data collection campaigns often make them impractical, resulting in significant data gaps in regions where crop production information is most needed. This paper presents our inaugural dataset comprising 4,925 validated crop-type data points from Kenya's 2021 and 2022 long-rain seasons. Collaborating with institutional partners and an extensive citizen science network, we collected georeferenced images across Kenya using GoPro cameras. We developed and implemented a deep learning pipeline to process images into crop-type datasets. Our methodology incorporates rigorous quality control measures to ensure the integrity and reliability of the data. The resulting dataset represents a significant contribution to open science and a valuable resource for evidence-based agricultural decision-making.

## Background & Summary

Understanding food production patterns is fundamental to Sustainable Development Goal 2 (Zero Hunger), especially in food insecure regions. While Earth Observation (EO) data are increasingly the most cost-effective method of quantifying agricultural production, creating accurate crop-type maps requires significant investment in field surveys for collecting essential in-situ data for training and validating EO-derived maps<sup>1</sup>. Historically, this cost barrier has limited consistent crop mapping primarily to high-income countries with established agricultural monitoring systems, for example, Canada's Annual Crop Inventory<sup>2</sup>, USDA's Cropland Data Layer<sup>3,4</sup>, and France's "Registre Parcellaire Graphique"<sup>5</sup>. However, consistent and accurate crop data layers are needed in all countries to help realize Sustainable Development Goal 2: Zero-Hunger. Technological and institutional innovations, including approaches to data collection and modeling frameworks, are needed to address this data gap<sup>6</sup>.

Cropland maps provide spatial information on where crops are growing, while crop-type maps go a step further by specifying the type of crop cultivated in each spatial unit (e.g., maize, wheat, or rice)<sup>7</sup>. These maps are critical for EO-based applications, such as crop yield predictions and crop condition assessments, as they enable analysts to focus on pixels that represent cropland or specific crop types. Given that farmers may change the crops grown in a particular field from season to season, crop-type maps must be updated regularly to maintain

<sup>1</sup>Department of Geographical Sciences, University of Maryland, College Park, MD, 20742, USA. <sup>2</sup>School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, 85287, USA. <sup>3</sup>World Resources Institute (WRI) Africa, Nairobi, Kenya. <sup>4</sup>Agriculture Statistics Unit, State Department for Agriculture, Ministry of Agriculture and Livestock Development, Nairobi, Kenya. <sup>5</sup>Department of Urban Forestry, Environment and Natural Resources, Southern University and A&M College, 110E Fisher Hall, Baton Rouge, LA, 70813, USA. <sup>6</sup>These authors contributed equally: Catherine Nakalembe, Ivan Zvonkov. ✉e-mail: [cnakalem@umd.edu](mailto:cnakalem@umd.edu)

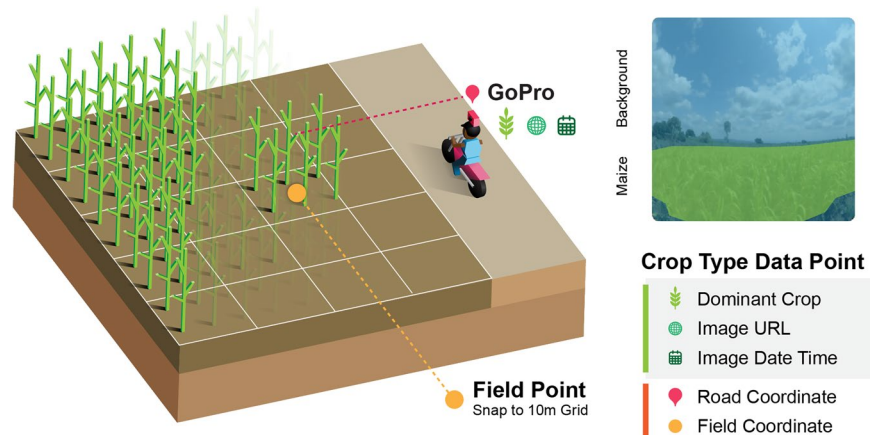


**Fig. 1** Illustration of crop type data generation pipeline. Steps: 1) Field agent with helmet-mounted GoPro drives along agricultural fields, 2) GoPro captures photo of adjacent field (red dotted line) and road coordinate (red point), 3) field coordinate (orange point) is calculated by moving road coordinate 20 meters into field and snapping to 10 meter grid, 4) GoPro photo is used to predict dominant crop.

accuracy<sup>8</sup>. A fundamental input for creating both cropland extent and crop-type maps is labeled data, which consists of georeferenced points indicating cropland and crop types.

While cropland labels are increasingly derived from satellite image interpretation, making cropland maps more accessible<sup>9–11</sup>, crop-type mapping remains a significant challenge. Unlike cropland extent, crop types cannot be determined through image interpretation alone and require extensive field surveys for accurate identification. Traditional methods for collecting crop-type data rely on labor-intensive field visits using GPS devices or smartphone apps, which often result in sparse and unevenly distributed datasets. This is particularly true in low-income regions, where resources for data collection are limited. Consequently, comprehensive crop-type maps with full national coverage or seasonal updates remain scarce, especially in low-income countries where the need for such data is most critical.

Emerging innovations in deep learning and computer vision are transforming in-situ crop type data collection, offering scalable alternatives to traditional methods. Recent studies highlight promising approaches: Paliyam *et al.* developed Street2Sat, leveraging vehicle-mounted cameras and deep learning to generate georeferenced crop type points in Kenya<sup>12</sup>. Wu *et al.* introduced the iCrop dataset, based on smartphone-captured roadside imagery in China<sup>13</sup>. Yan and Ryu demonstrated the use of Google Street View and deep learning to produce crop type points and maps in Illinois and California<sup>14,15</sup>. Similarly, Laguarta *et al.* applied Google Street View and deep learning to map crop types in smallholder agricultural landscapes in Thailand<sup>16</sup>. d'Andrimont *et al.* utilized vehicle-mounted cameras to capture multi-temporal imagery, enabling crop phenology tracking for delineated parcels in the Netherlands<sup>17</sup>. Together, these advancements pave the way for overcoming the limitations of traditional data collection methods.

In this paper, we introduce a scalable, low-cost, and effective method for collecting crop-type data in smallholder agricultural landscapes through the Helmets Labeling Crops project. Our approach, summarized in Fig. 1, combines GoPro cameras and deep learning techniques based on the Street2Sat framework<sup>12</sup> to enable rapid and cost-effective field surveys of crop type. We release an open, georeferenced crop-type dataset with associated roadside imagery from 17 counties in Kenya, providing an alternative to proprietary platforms like Google Street View<sup>18</sup> while adhering to FAIR data principles<sup>19</sup>. This initial release will be followed by datasets from additional countries, including Uganda, Tanzania, Zambia, Nigeria, Uruguay, Senegal, India, South Korea, and Bhutan.

## Methods

**Data Collection.** We used 6th Grain's Maize Density Layer<sup>20</sup> and ESA's Global Crop Mask<sup>21</sup> to select counties for targeted maize-focused data collection, ensuring coverage of high, medium, and low-density maize production areas across Kenya from 2021–2022. We treated maize as “low-hanging fruit” for demonstrating the scalability of our street-level data collection methodology in Kenya's most critical food security crop, which accounts about 65% of total staple food caloric intake and 36% of total food caloric intake<sup>22</sup>. We covered 17 counties Bomet, Bungoma, Busia, Homa Bay, Kakamega, Kericho, Kisii, Machakos, Migori, Nakuru, Nandi, Narok, Nyamira, Trans Nzoia, Uasin Gishu and Vihiga (Fig. 5). These counties were selected to capture diverse agro-ecological zones and maize production intensities rather than comprehensive crop diversity.

This strategic sampling design allowed us to validate our data collection protocols and quality control procedures in a controlled context focused on Kenya's dominant staple crop. Achieving comprehensive crop diversity would require multiple surveys throughout the agricultural calendar to capture different crops during their optimal identification windows—for example, bean/maize rotation systems or cowpea intercropping would necessitate separate data collection campaigns timed to each crop's development cycle. Our single-season approach prioritized methodological validation over crop diversity, providing a foundation for future multi-temporal collection strategies.



**Fig. 2** Field agent being trained on operating helmet-mounted GoPro for crop type data collection.



**Fig. 3** Data Collection Toolkit Content.

In collaboration with the Regional Centre For Mapping Of Resources For Development (RCMRD) and Kenya Ministry of Agriculture, we recruited and trained 25 local agricultural officers as data collectors. These officers received comprehensive training on mounting and operating GoPro cameras to ensure consistent data collection protocols (Fig. 2). As part of their regular crop monitoring duties, these officers provided valuable insights into current season crop performance and distribution patterns, which were essential for optimizing data collection routes and timing.

The data collection strategy utilized Kenya's hierarchical road network classification<sup>23</sup> to maximize coverage efficiency. Primary routes (classes A, B, and C) facilitated movement between counties and major towns, while secondary and tertiary networks (classes D and E) enabled access to local farming communities. Field agents reported that GoPro cameras mounted on motorcycles were particularly effective for village-level data collection, offering superior maneuverability and access to remote agricultural areas compared to vehicle-mounted systems. This mobility advantage resulted in more comprehensive coverage of smallholder farms and diverse agricultural landscapes within each target region.

In Kenya, this method was implemented through a collaboration between the Kenyan Ministry of Agriculture and the Regional Centre for Mapping of Resources for Development (RCMRD), following an initial pilot phase in 2020 with LocateIT. After training 59 agricultural officers and engaging local stakeholders, the team captured 397,190 georeferenced images across all campaigns in Kenya.

To support data collection, we developed a comprehensive Data Collection Toolkit (Fig. 3) for each agricultural officer. The toolkit included a manual detailing the contents of the kit, GoPro camera settings for data collection, GoPro mounting instructions, step-by-step guidelines for capturing data, and instructions for uploading data. The toolkit (Fig. 3) included a GoPro Hero 8 Black, a magnetic car mount (for capturing data in a car), a motorbike helmet mount (for capturing data on a motorcycle), a 512 GB SD, a USB-C to USB-A cable (to charge the GoPro while collecting data), backup GoPro batteries, and a carrying case to keep all items together. We provided a daily GoPro settings checklist for all data collectors to ensure consistent data collection. The checklist

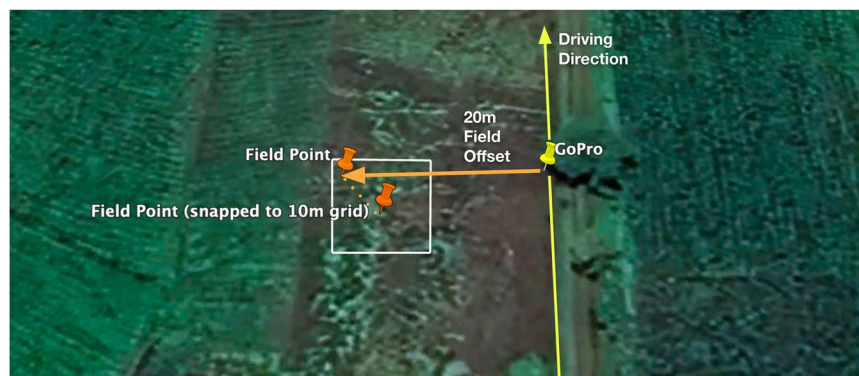


**Fig. 4** Field Offset Algorithm Example.

ensured that the GPS was enabled, the battery was charged or plugged in, and sufficient storage space was available on the provided SD card.

**GoPro Camera Configuration.** Our Data Collection Toolkit included comprehensive instructions for standardized image capture to ensure consistent data quality across all field agents. We specified precise camera settings for the GoPro Hero 8 Black: **Lens:** Narrow field of view, **Format:** Photo, **Mode:** Time-lapse photo (not video), **Interval:** 0.5 seconds (2 Frames Per Second), and **Resolution:** 4000 × 3000 pixels (12 MP). These settings were specifically optimized for the challenges of smallholder agricultural data collection. The narrow lens setting was critical for coordinate estimation accuracy, as it reduces edge distortion that could affect crop segmentation and minimizes GPS coordinate uncertainty relative to image content. Unlike wide-angle modes that capture more area but introduce significant distortion, the narrow field of view provides more focused and geometrically consistent imagery for both crop identification and spatial referencing.

The 0.5-second interval was designed to address the spatial variability of smallholder landscapes. At typical rural driving speeds (20–40 km/h), this interval captures 3–4 photos every 10 meters, providing substantial redundancy that proves essential given the irregular field boundaries and mixed land use patterns characteristic of these agricultural systems. This redundancy serves multiple purposes: it provides backup images in cases of obstruction or poor exposure, ensures coverage of narrow fields that might otherwise be missed with longer intervals, and increases the likelihood of capturing at least one image where each field is centered and fully visible.

The photo mode rather than video was selected to optimize both storage efficiency and image quality for computer vision applications. Individual high-resolution photos provide better detail for crop segmentation algorithms while generating more manageable file sizes for upload and processing in bandwidth-limited rural environments. The time-lapse approach also automatically embeds GPS coordinates and timestamps in EXIF metadata, which is essential for our coordinate estimation pipeline.

These standardized settings were enforced through daily checklists provided to all field agents to ensure data consistency across the diverse geographic and operational conditions encountered during our multi-county data collection campaigns.

We also provided a guide for mounting the camera for both cars and motorcycles, as an improperly mounted GoPro meant unusable images. For data collection in Kenya, it was crucial always to mount the camera to point left to optimize the visibility of roadside fields, since Kenya is a left-side drive country. GoPros do not record camera heading, which is critical to determining where the crop is located. By mounting the GoPros to the face left, we determine the camera direction by rotating the driving direction 90 degrees to the left (counter-clockwise, see Fig. 4).

We created a Google Cloud Storage bucket and a file structure to store all collected images. We provided detailed instructions to all data collectors to upload collected GoPro images using either the Google Cloud user interface or command line interface (to allow for parallel uploads). We stored the images on Google Cloud Storage to allow for a central repository with managed permissions. Each leaf folder contains between 1000 and 20,000 images, depending on the length of time the specific GoPro was on. The approach generates large volumes of data, often tens of thousands of images per collection campaign. This scale makes manual annotation infeasible and necessitates the development of automated processing pipelines to extract meaningful agricultural information efficiently.

**Crop Prediction using Deep Learning.** Smallholder fields in Kenya present two challenges to determining the crop from a GoPro image. First, some fields contain multiple crops (inter-cropping); second, the small size of the fields means multiple fields can be visible in a single image. Unlike prior work, which formulated the crop recognition task as image classification (predicting one crop label for the entire image)<sup>12</sup>, here, we formulated the crop recognition task as image segmentation. By asking the model to segment the image into crops and background, we measure crop proportion through the number of pixels segmented as each crop type. Additionally, we obtain a fine-grained model output, which can be useful for debugging model failure modes. We experimented with formulating the crop recognition task as object detection (as in Paliyam *et al.*<sup>12</sup>) but found that labeling and identifying each crop with a bounding box was labor-intensive and difficult to predict with object detection methods.

**Two-Stage Processing Pipeline.** To optimize computational efficiency for large-scale deployment, we implemented a two-stage processing pipeline. The first stage employs a lightweight binary classification model to filter images containing crops, while the second stage applies semantic segmentation for precise crop identification. This approach significantly reduces computational overhead by eliminating non-agricultural images before applying the more resource-intensive segmentation model.

**Architecture Selection and Comparison.** We conducted comprehensive evaluation of three deep learning architectures: Feature Pyramid Network (FPN)<sup>24</sup>, UNet++<sup>25</sup>, and LinkNet<sup>26</sup>, each tested with multiple encoder backbones including Xception<sup>27</sup>, DenseNet-161<sup>28</sup>, ResNet-18<sup>29</sup>, and EfficientNet-B5<sup>30</sup>. Our evaluation focused on mean average precision (mAP) as the primary performance metric due to its robustness for multi-class segmentation tasks.

We initialized each model with ImageNet pre-trained weights. We used the default Weighted Dice Loss from PyTorch and the Adam optimizer with a learning rate of 0.0001 to train each model. We trained each model for 20 epochs and saved the model weights with the highest iou score.

FPN demonstrated superior stability and performance across all crop types compared to alternative architectures. UNet++ showed notable performance variations across different crops, particularly for rice, maize, sugarcane, and tobacco, while LinkNet exhibited the least stable performance with significant variations across crop classes. We selected FPN as our primary architecture due to its consistent performance across all agricultural scenarios.

For encoder backbone selection within the FPN architecture, Xception achieved the highest mAP across most crop classes while using fewer parameters (22M) compared to DenseNet-161 (26M), making it optimal for both performance and computational efficiency Table 1. This parameter efficiency is crucial for enabling deployment of larger models in subsequent processing stages of our pipeline.

**Crop Segmentation.** Our Feature Pyramid Network (FPN) with an Xception backbone model achieved a 92.5% mean average precision on our held-out validation dataset, representing a significant improvement over alternative architectures tested. The crop-specific performance metrics (Table 2) demonstrate robust detection capabilities across all eight crop types, with precision and recall values ranging from 0.85-0.99 and 0.93-0.99, respectively.

Performance varied by crop type, with soybean, sunflower, tobacco, and wheat achieving the highest precision scores (0.94-0.99), while maize showed slightly lower but still acceptable performance (0.85 precision, 0.93 recall). This variation reflects the inherent visual complexity differences between crop types, with structurally distinct crops like tobacco and sunflower being more readily identifiable than visually similar crops like maize in mixed agricultural landscapes.

To train the image segmentation model we labeled a dataset of 2302 images with 8 different classes (Table 2): maize, banana, rice, soybean, sugarcane, sunflower, tobacco, and wheat. These crops (particularly maize, banana, rice, and wheat) were selected for their importance to food security in East Africa and their high likelihood for cultivation in the 17 counties surveyed. Maize, for instance, is a vital food security crop, contributing 36% of caloric intake in Kenya alone<sup>31</sup>. Similarly, rice, wheat, and bananas are integral to regional diets and have significant nutritional and economic impacts<sup>32</sup>. Collectively, these crops provide a solid foundation for evaluating agricultural patterns and enhancing food security across the region.

We gathered training images from previous field campaigns and existing datasets of crop images (Table 3). The multi-source training data approach ensures model robustness across diverse geographical and climatic conditions, with substantial contributions from Tanzania (3,027 training images), Uganda (486), and supplementary data from the USA (369) and existing agricultural datasets. We manually labeled a segmentation mask in each image using the Roboflow annotation tool. The Smart Polygon Tool with intelligent assistant enables efficient annotation through single-click addition and removal of regions, significantly reducing the time required for manual labeling while maintaining annotation quality. We allocated 30% of the crop images for validation to ensure robust performance evaluation.

**Advanced Preprocessing and Data Augmentation.** We applied several augmentation techniques (rotation:  $\pm 5$  degrees, contrast:  $\pm 5\%$ , horizontal flip, and exposure:  $\pm 3\%$ ) to increase the size of the training set to 4832 images. We resized all images to  $800 \times 800$  pixels to balance computational efficiency with spatial resolution requirements for accurate crop segmentation.

Varying outdoor lighting conditions present a significant challenge for agricultural image analysis, as images are captured under diverse environmental conditions throughout different times of day and weather patterns. A critical preprocessing innovation was the implementation of adaptive gamma correction with maximum entropy optimization<sup>33-35</sup>. This technique was adopted after observing that varying outdoor lighting conditions significantly affected model performance. The gamma correction preprocessing improved our mean average precision from 85.16% to 92.75%, representing a substantial 7.6 percentage point improvement that validates the importance of addressing illumination variability in outdoor agricultural imaging scenarios.

The full segmentation dataset is publicly available on Roboflow: <https://app.roboflow.com/ivan-zvonkov/street2sat-segmentation/overview>.

**Binary Classification for Computational Efficiency.** We found that a significant number of collected images contained no crops. Rather than running the relatively computationally heavy segmentation model on all images, we developed a lightweight binary crop classification model to filter out images that did not contain any crops before feeding crop images to the segmentation model. We call this the CropNop (crop or not crop) model.

| Model           | Background | Banana | Maize | Rice | Sugarcane |
|-----------------|------------|--------|-------|------|-----------|
| EfficientNet-B5 | 0.90       | 0.65   | 0.83  | 0.86 | 0.93      |
| Xception        | 0.93       | 0.65   | 0.87  | 0.90 | 0.93      |
| DenseNet-161    | 0.94       | 0.70   | 0.87  | 0.89 | 0.93      |
| ResNet-18       | 0.93       | 0.65   | 0.83  | 0.90 | 0.93      |

**Table 1.** Crop-wise mean average precision by Encoder Backbone (Selected Results).

| Crop Type | Training | Validation | Precision | Recall |
|-----------|----------|------------|-----------|--------|
| Banana    | 407      | 55         | 0.88      | 0.98   |
| Maize     | 1272     | 184        | 0.85      | 0.93   |
| Rice      | 1194     | 181        | 0.86      | 0.96   |
| Soybean   | 351      | 48         | 0.97      | 0.99   |
| Sugarcane | 174      | 31         | 0.87      | 0.98   |
| Sunflower | 384      | 51         | 0.97      | 0.98   |
| Tobacco   | 66       | 9          | 0.99      | 0.99   |
| Wheat     | 480      | 72         | 0.94      | 0.99   |
| Null*     | 504      | 60         | 0.86      | 0.93   |

**Table 2.** Segmentation dataset distribution and metrics per crop type. \*Prevent False Positives

| Source                    | Training    | Validation | Total       |
|---------------------------|-------------|------------|-------------|
| Tanzania                  | 3027        | 422        | 3449        |
| Kenya                     | 87          | 16         | 103         |
| Uganda                    | 486         | 76         | 562         |
| USA                       | 369         | 50         | 419         |
| Kaggle <sup>42</sup>      | 369         | 52         | 421         |
| Kaggle <sup>43,44</sup>   | 314         | 47         | 361         |
| FlevoVision <sup>45</sup> | 180         | 28         | 208         |
| <b>Total</b>              | <b>4832</b> | <b>691</b> | <b>5523</b> |

**Table 3.** Data distribution by source.

Three lightweight architectures were systematically evaluated for filtering non-crop images, with the goal of optimizing computational efficiency while maintaining high accuracy. The evaluation focused on balancing model performance with parameter count to enable resource allocation toward more computationally intensive downstream segmentation tasks.

As with the segmentation models, we initialized each of the binary classification models with ImageNet pre-trained weights. We used a the default Binary Cross Entropy loss and Adam optimizer from PyTorch to train the models. We trained each classification model for 25 epochs and saved the model weights with the lowest loss value.

We selected a variant of the SqueezeNet model<sup>36</sup> for this task due to its exceptional computational efficiency, using only 1.2M parameters compared to 11M for ResNet18 and 2.5M for MobileNetV3. Despite having 50% fewer parameters than MobileNetV3, SqueezeNet achieved superior performance (99% accuracy) while maintaining the computational efficiency necessary for large-scale deployment. The simpler binary classification task allowed us to use smaller images ( $300 \times 300$  pixels) which further reduced computation requirements.

We created a training dataset using pseudolabels from the segmentation model predictions: if an image was segmented as more than 75% background, the image was labeled non-crop. If the image was segmented as less than 25% background, the image was labeled crop. We gathered crop and non-crop images from Kenya, Uganda, the USA, and Tanzania for a total of 8439 images (Train: 6957 [crop: 3711, non-crop: 3246], and Test: 1482 [crop: 400, non-crop: 1082]). We used the same image preprocessing as for the crop segmentation model. Our CropNop model achieved 99% accuracy on the test dataset, effectively eliminating 73.9% of non-agricultural images from subsequent processing.

This two-stage approach enables significant computational savings by allowing resource conservation for deployment of larger, more sophisticated models in downstream segmentation tasks, while maintaining high accuracy in agricultural content detection.

**Road to Field Coordinate.** For each image, we extracted the GPS coordinates, date, and time from the image's EXIF-formatted metadata. We then projected the coordinates from WGS84 into the Universal Transverse Mercator (UTM) system to enable calculations in meters. We then used the image's coordinates and metadata on the road to compute the coordinates of the crop field captured in each image.

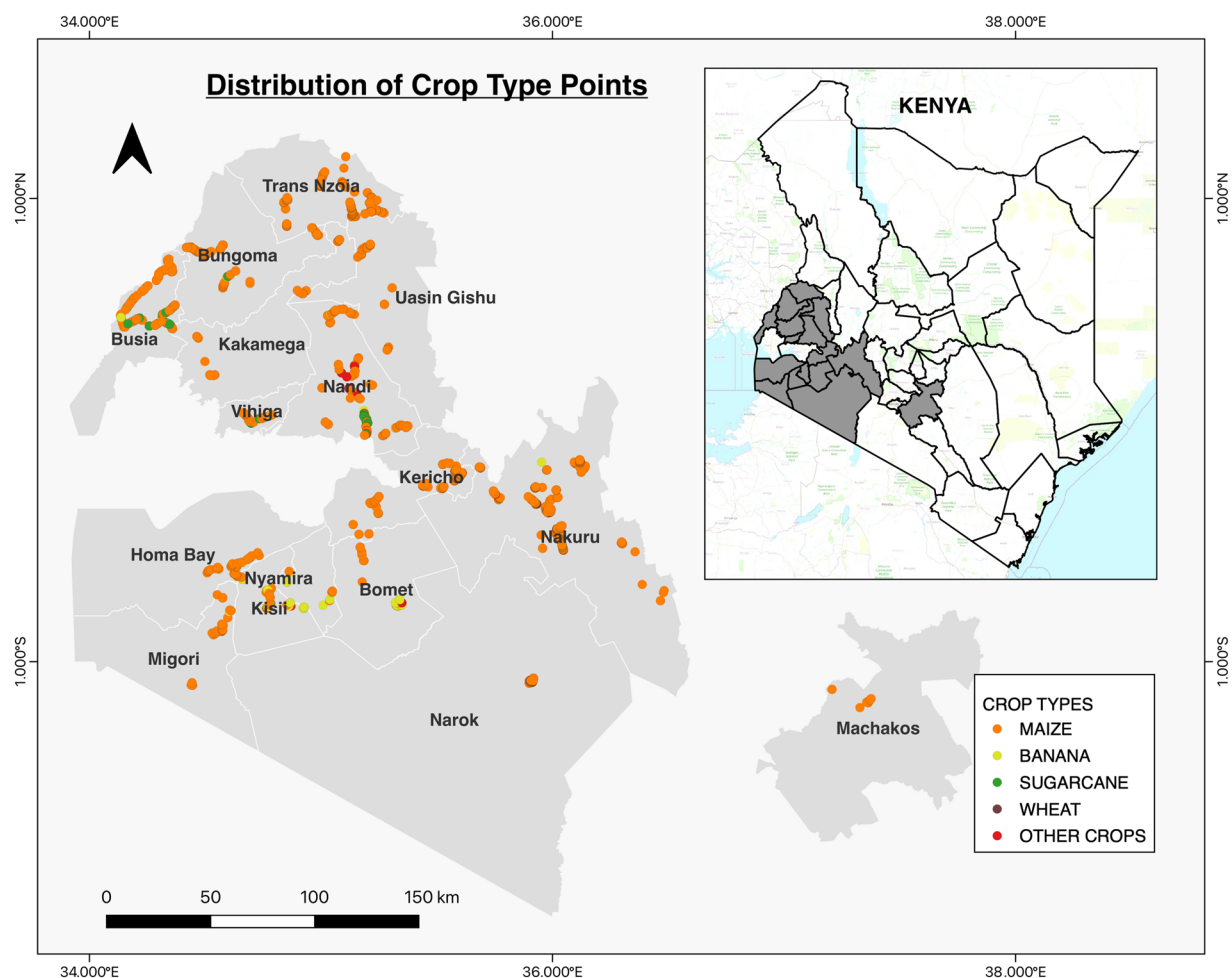


**Fig. 5** Distribution of crop type points.

Next, we computed the vehicle's driving direction to determine the GoPro camera's field-facing direction. For each coordinate, we selected the prior coordinate and calculated the difference between the two coordinates to determine the Northing and Easting components of the driving direction. Since Kenya is a left-hand-drive country and the GoPro cameras are mounted on the passenger side, the field-facing direction is obtained by rotating the driving direction 90 degrees to the left.

We applied a 20-meter offset along the field-facing direction from the GoPro's position to estimate the field coordinates. We experimented with various distance estimation methods but ultimately found that adaptive offsets had a fundamental drawback within our workflow. Our points are intended to be used with Earth Observation data such as Sentinel-2, which has a resolution of 10 meters. A Sentinel-2 pixel has a diagonal of 14.14 meters. Therefore an offset from road point to field point of less than 14.14 meters risked staying within the same Sentinel-2 pixel. A 20 meter offset guarantees that the field point does not land in the same Sentinel-2 pixel as the road point. It also accounts for buffers that are usually present between roads and fields. Additionally, smallholder fields can have widths of just 30 meters (15 meters to the center of the field). Using a 20 meter offset allowed us to not overshoot the center of these small fields and ensure the field point represents a field pixel. At this stage, crop-type data points may form clusters, with consecutive field points spaced only a few meters apart. To ensure compatibility with moderate-resolution Earth observation data, such as Sentinel-2, we snapped all points to the centroid of a 10-meter pixel grid (Fig. 4). If multiple points fell within the same 10-meter pixel, we retained the photo with the highest percentage of crops and discarded the rest.

**End-to-End Automated Pipeline.** In total, we ran our automated pipeline on 32,804 GoPro images in the focus counties. Of these, we progressively eliminated 28,580 images that were not fit to be transformed into crop type points. We eliminated 5.7% of the photos captured in Kenya due to invalid EXIF data (coordinate metadata). Next, we ran the CropNop model on the photos with valid EXIF data, resulting in the removal of 73.9% images that did not contain crops according to the classifier. We ran the segmentation model on the remaining photos. We used a segmentation proportion threshold for the dominant crop to eliminate photos where the segmented area was very small. We set the default threshold to 5%, meaning only photos with a dominant crop covering at least 5% of the image are kept. We adjusted this threshold for some dataset subsets. Finally, we snapped our GoPro field coordinates to a grid and deduplicated points in the same grid. The deduplication step eliminated 53% of

the remaining photos. We packaged the remaining 4,224 points into Google Earth Pro KMZ files, on which we performed further manual verification (see Technical Validation).

**Planned Improvements.** Future development will focus on incorporating active learning capabilities into the segmentation model to enable seamless integration of new crop classes as they appear in collected data. This adaptive approach will allow the system to expand its crop recognition capabilities without requiring complete model retraining, making it more suitable for diverse agricultural environments and changing cropping patterns.

## Data Records

The resulting dataset reflects both our strategic sampling approach and the agricultural landscape realities during our collection periods. While maize dominates the final dataset (88% of points), our model successfully detected multiple crop types during processing, demonstrating the broader applicability of our data collection methodology for diverse agricultural systems. Our final dataset consists of a CSV file representing the crop type points and a zipped folder of roadside images associated with each crop type point. We also host the roadside images on a public Google Cloud Storage bucket for accessibility without the need to download all images. We make the dataset public on Zenodo<sup>37</sup> with the CC BY 4.0 license: <https://zenodo.org/records/15133324>.

In total, we present 4,925 validated crop type points collected across 17 counties in Kenya during 2021 and 2022. The crop distribution accurately reflects Kenya's agricultural priorities, with maize as the dominant staple crop<sup>22</sup> represented by 4,351 points, followed by 301 sugarcane, 140 banana, 106 wheat, 10 tea, 5 beans, 4 sunflower, 2 rice, 2 soybean, 2 cassava, and 1 kale point. Geographically, the 2022 dataset contains points from Nakuru (653) and Machakos (14), while the 2021 dataset spans a broader geographic range including Nakuru (764), Trans Nzoia (550), Nandi (528), Kericho (516), Busia (493), Bungoma (370), Narok (236), Homa Bay (204), Uasin Gishu (170), Kisii (124), Migori (117), Vihiga (84), Bomet (40), Kakamega (34), and Nyamira (28).

We chose to store the crop type points in a CSV file because it is human-readable, easy to open on any computer, and can be easily used in common GIS software (e.g., Google Earth Engine, Google Earth Pro, QGIS) when **latitude** and **longitude** columns are included. We include the road coordinates as a separate row in our CSV so future researchers can use them as non-crop points when training a cropland classifier. Each row of our CSV has the following attributes:

- **latitude** [number]: Field latitude snapped to 10m UTM grid.
- **longitude** [number]: Field longitude snapped to 10m UTM grid.
- **is\_crop** [number]: 1 if field coordinate, 0 if road coordinate.
- **crop\_type** [string]: The dominant crop in the image.
- **capture\_info** [string]: ID for each image.
- **capture\_time** [string]: Date and time of image capture.
- **adm1** [string]: Admin Level 1 from GADM.
- **adm2** [string]: Admin Level 2 from GADM.
- **image\_path** [string]: The path of the roadside image inside the provided image folder.
- **image\_url** [string]: The URL of the roadside image hosted on our Google Cloud.

This dataset provides a valuable resource for maize mapping applications and food security monitoring in Kenya, while the underlying methodology demonstrates scalability for diverse crop types across different agricultural landscapes.

## Technical Validation

**Quality Assessment.** We performed quality assessment on all classified crop type data points to verify the dominant crop prediction and whether the point fell inside a field and spatial coordinate accuracy. This manual verification process is essential given the complexity of Kenyan smallholder agricultural landscapes, where highly variable field boundaries, irregular plot shapes, mixed land use patterns, and inconsistent distances from roads to field edges make purely algorithmic coordinate estimation insufficient for reliable data quality.

We packaged the crop type points into a Google Earth Pro KMZ files which allow viewing the points overlaid on high-resolution satellite imagery. We used Google Earth Pro to select the temporally closest available satellite imagery to the given point, ensuring the most accurate assessment of field conditions at the time of data collection. The quality assessment was conducted as a process of elimination designed to address coordinate accuracy concerns. We eliminated any point that did not fall within an actual cultivated field, thereby directly addressing potential errors from our 20-meter offset approximation that could place points in buffer zones, fallow areas, or beyond field boundaries. We also eliminated any point where the photo contained no crops, ensuring that our coordinate estimation successfully captured agricultural areas rather than roads, settlements, or natural vegetation. Additionally, we removed any point where a human was present in the photo, as this could interfere with crop classification accuracy.

Beyond spatial validation, we corrected misclassified dominant crops at this stage of the assessment, ensuring that the crop type prediction accurately reflected the field conditions visible in both the street-level imagery and corresponding satellite data. This dual verification process (spatial accuracy through satellite imagery and crop accuracy through street-level photos) provides robust quality control that accounts for the inherent uncertainties in coordinate estimation across diverse landscape conditions.

The interface for point assessment is shown in Fig. 7. All KMZ files were checked by a minimum of two independent reviewers to ensure consistency and reduce subjective bias in the assessment process. We reviewed difficult or ambiguous points with representatives from Kenya's Ministry of Agriculture and Livestock Development,



**Fig. 6** Examples of misclassified images: (a) an intercropped banana and beans farm, and (b) a buffer plant occasionally misclassified as a crop.

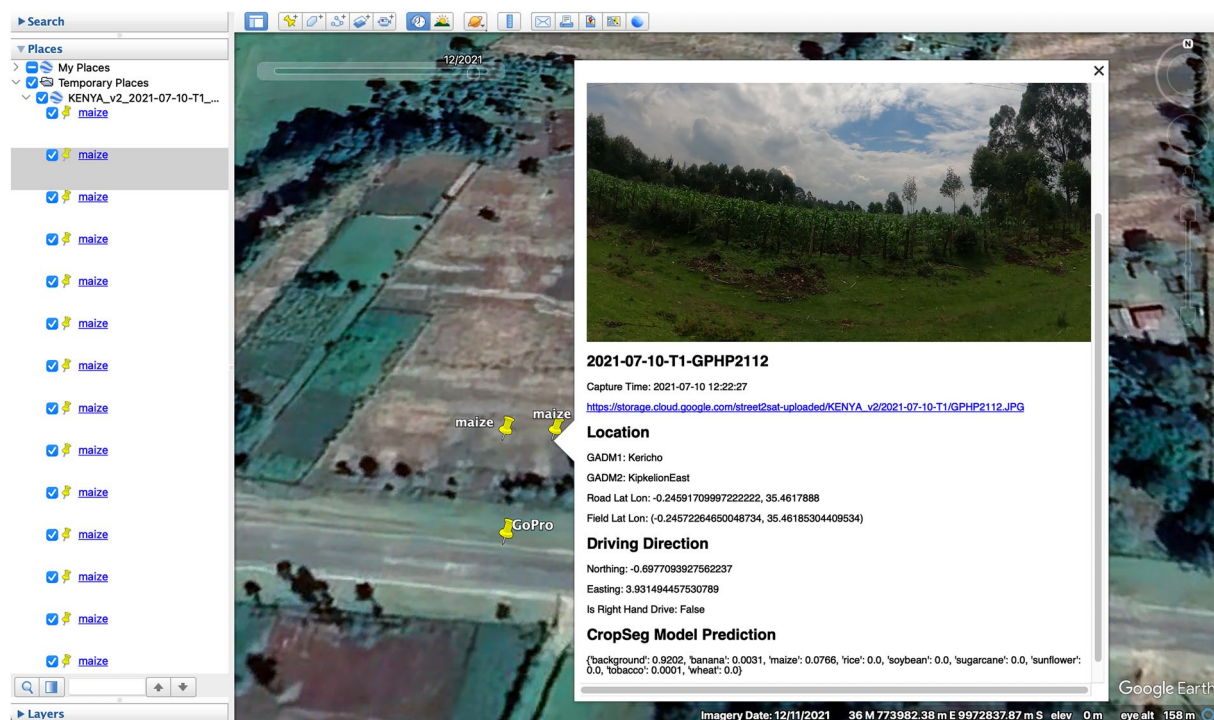


**Fig. 7** Interface for quality assessment.

leveraging their local expertise to resolve cases where landscape complexity made assessment challenging. This collaborative approach was particularly valuable for interpreting intercropped fields, buffer crop arrangements, and seasonal variations in crop appearance that could affect both coordinate placement and crop identification accuracy.

The manual verification process effectively addresses the coordinate estimation challenges that would be difficult to resolve through algorithmic approaches alone. While we previously tested more precise distance measurement methods using tape measures in controlled environments, developing alternative algorithmic approaches to improve coordinate estimation accuracy, including Structure from Motion (SfM) techniques that leverage feature matching between consecutive images to estimate distances without requiring assumptions about crop height, the scalability limitations and landscape heterogeneity made such approaches impractical for large-scale deployment<sup>38</sup>. Our quality assessment protocol provides a scalable solution that maintains data quality while accommodating the spatial variability inherent in smallholder agricultural systems. We then converted all approved KMZ files into a single CSV file, ensuring that only validated, high-quality crop type points with verified coordinate accuracy were included in the final dataset.

**Point Assessment Challenges.** Our quality assessment process revealed several systematic challenges that informed our validation protocols. Spatial accuracy challenges were among the most frequent obstacles, occurring when capture points fell outside field boundaries despite our 20-meter offset approximation. While this offset represented the average distance from road to field edge based on initial field observations, the heterogeneous nature of smallholder landscapes meant that farmers establish field boundaries at highly variable distances from

| Model       | Parameters | Precision | Recall | F1 Score | Accuracy |
|-------------|------------|-----------|--------|----------|----------|
| ResNet18    | 11M        | 0.955     | 0.925  | 0.93     | 0.943    |
| MobileNetv3 | 2.5M       | 0.97      | 0.95   | 0.96     | 0.97     |
| SqueezeNet  | 1.2M       | 0.98      | 0.985  | 0.98     | 0.99     |

**Table 4.** Data Filtration Model Performance Comparison.

roads (5–50+ meters). Points that landed beyond field boundaries or on buffer zones were systematically eliminated during quality assessment. Crop classification challenges arose from both landscape complexity and model behavior. Buffer crops created particular difficulties when they obstructed the view of main crops—for example, roadside maize plants blocking sugarcane fields behind them, leading to misclassification and subsequent point elimination (see Fig. 6a). Our segmentation model also exhibited differential sensitivity to crop types, detecting banana plants more readily than other crops due to their distinct structural features. This frequently resulted in fields with sparse banana trees (planted for shade in predominantly maize or bean fields) being classified entirely as “banana” crops, despite banana representing a minor component of the agricultural system. Intercropping prevalence added another layer of complexity, as many smallholder farmers practice various intercropping arrangements that challenge discrete classification systems. During quality assessment, we prioritized the dominant agricultural practice in each field, correcting cases where visually prominent but spatially minor crops (particularly single banana plants) caused incorrect field-level classification (Fig. 6b). These challenges directly shaped our quality control methodology, emphasizing the necessity of manual verification for ensuring both spatial accuracy and appropriate crop-type representation. Our final dataset provides discrete crop-type labels that accurately reflect the dominant agricultural practice at each location, while our segmentation model’s successful identification of eight crop types during processing (Table 3) demonstrates broader detection capabilities across Kenya’s diverse agricultural landscape.

**Maize Mapping.** We further validated the points by using them to create a maize map over Kenya’s Nakuru county. Nakuru covers an area of 7496.5 square kilometres and is an important agricultural center in Kenya<sup>39</sup>. We chose to map maize in Nakuru county because that is by far the most dominant crop grown in the region Table 4.

**Satellite Data.** We used Sentinel-2 L2A data as the input data source for our maize mapping workflow. We queried images in the area of interest from April 1, 2021 to December 1, 2021 to cover the entire growing season. We used the S2 Cloudless collection to mask out pixels with over 30% cloud cover. We then composited the images into 2 month median composites (April–May, June–July, August–September, October–November). For each median composite, we computed the Normalized Difference Vegetation Index (NDVI). We selected the following bands from each composite: B2, B3, B4, B8, B8A, B9, B11, NDVI.

**Crop Type Data.** We used all crop type points in Nakuru county generated through our pipeline (a total of 1528 points consisting of 764 field points, and 764 corresponding road points). Of the 764 field points, 760 were maize. We increased the amount of non-maize crop points by sampling additional points within the fields of our non-maize crop points. Specifically, we drew field polygons for a sample of non-maize crop points, then applied a 5 meter buffer to the field polygons to remove border effects and sampled exhaustively every 10 meters within the field polygon. This process resulted in 1102 non-maize crop points. Additionally, after an initial round of classification we added 41 maize points. For a total of 801 maize points and 1870 non-maize points to be used for classification.

**Classification.** We used a two step approach to maize classification, first masking out obvious non-crop pixels and then training a random forest classifier to classify remaining pixels as maize or non-maize. To mask out obvious non-crop pixels such as water bodies, we used the WorldCover 2021 land cover map<sup>40</sup>. We masked out all classes except crops, tree cover, and grassland, as we found that the latter two classes in the WorldCover map sometimes contained crop fields. We trained a random forest classifier (with 50 trees) on Google Earth Engine using the satellite data and crop type data described above.

**Map Validation.** We evaluated the created maize map using Copernicus4GEOGLAM<sup>41</sup> field polygons available in the Nakuru region. These polygons represent high-quality ground truth data collected through in-person field visits, providing reliable validation for our classification results. To avoid spatial auto-correlation effects that could inflate accuracy estimates, we extracted only the centroid of each polygon for validation rather than using entire field areas. Our maize classification achieved an overall accuracy of 71.1%, with a user accuracy of 45.7% and producer accuracy of 58.2%. While the overall accuracy demonstrates reasonable performance for distinguishing maize from non-maize areas, the lower user and producer accuracies reflect the inherent challenges of crop-type mapping in heterogeneous smallholder landscapes where field boundaries are irregular, crop mixtures are common, and spectral signatures can be ambiguous.

## Usage Notes

**Code Repository and Pipeline.** All code for the crop-type mapping pipeline is accessible through the helmets-kenya repository: <https://github.com/nasaharvest/helmets-kenya/tree/main>. The most relevant files are:

1. For processing the GoPro images into crop type points; <https://github.com/nasaharvest/helmets-kenya/blob/main/notebooks/GoPro2CropKMZ.ipynb> and,
2. For converting analyzed and approved KMZ files into a single CSV file <https://github.com/nasaharvest/helmets-kenya/blob/main/notebooks/CropKMZtoCSV.ipynb>

**Data Format and Access.** The dataset is provided as a CSV file with georeferenced crop-type points and an associated image folder. Users can access the roadside images either by downloading the complete dataset or accessing individual images via the provided Google Cloud Storage URLs without downloading the entire collection.

**Recommended Software and Analysis Platforms.** Our collected crop-type points can be used to train machine learning models and create crop-type maps. We recommend using Google Earth Engine with our data points because Earth Engine makes it straightforward to analyze crop-type coordinates in conjunction with remote sensing data. We share code for an example maize mapping use case in Nakuru county on Google Earth Engine: [https://code.earthengine.google.com/?accept\\_repo=users/izvonkov/helmets-kenya-public](https://code.earthengine.google.com/?accept_repo=users/izvonkov/helmets-kenya-public). The CSV format also enables direct import into common GIS software including QGIS, ArcGIS, and Google Earth Pro for spatial visualization and analysis. The dataset is compatible with standard Python geospatial libraries (geopandas, rasterio) and R spatial packages (sf, raster) for custom analysis workflows.

**Suggested Processing Steps.** For training crop classification models, users should split the dataset geographically rather than randomly to avoid spatial autocorrelation in validation results. The road coordinates (is\_crop = 0) can serve as non-crop training points for cropland extent mapping. When combining with satellite imagery, we recommend using the 10-meter grid snapping inherent in our coordinates for compatibility with Sentinel-2 data. Users should account for the temporal information (capture\_time) when matching with satellite imagery to ensure appropriate temporal alignment.

**Dataset Limitations and Considerations.** The heavy maize representation (88% of points) makes the dataset particularly suitable for maize mapping applications and food security monitoring in similar smallholder contexts. Users requiring balanced multi-crop training data should consider combining our dataset with other sources or implementing class balancing techniques. The dataset reflects agricultural conditions during specific growing season windows and may not represent year-round crop distributions. Users working in different geographic regions should validate coordinate offset assumptions against local field boundary patterns.

Received: 22 April 2025; Accepted: 1 August 2025;

Published online: 27 August 2025

## References

1. Davidson, A.M., Fiset, T., McNairn, H. & Daneshfar, B. Detailed crop mapping using remote sensing data (Crop Data Layers). In Delincé, J. (Ed.), *Handbook on Remote Sensing for Agricultural Statistics* (Chapter 4). Rome. (2017).
2. Fiset, T. et al. AAFC annual crop inventory. In *2013 Second International Conference on Agro-Geoinformatics (Agro-Geoinformatics)* (pp. 270–274). IEEE. (2013).
3. Han, W., Yang, Z., Di, L. & Mueller, R. CropScape: A Web service based application for exploring and disseminating US continuous geospatial cropland data products for decision support. *Computers and Electronics in Agriculture* **84**, 111–123, <https://doi.org/10.1016/j.compag.2012.03.005> (2012).
4. North, M. A., Hastie, W. W. & Hoyer, L. Out of Africa: The underrepresentation of African authors in high-impact geoscience literature. *Earth-Science Reviews* **208**, 103262, <https://doi.org/10.1016/j.earscirev.2020.103262> (2020).
5. Cantelaube, P. & Carles, M. Le registre parcellaire graphique : des données géographiques pour décrire la couverture du sol agricole. *Cahier des Techniques de l'INRA*, Méthodes et techniques GPS et SIG pour la conduite de dispositifs expérimentaux, 58–64, <https://hal.inrae.fr/hal-02639660> (2014).
6. Kebede, E. et al. Assessing and addressing the global state of food production data scarcity. *Nature Reviews Earth & Environment* **5**, 295–311, <https://doi.org/10.1038/s43017-024-00516-2> (2024).
7. Nakalembe, C. & Kerner, H. Considerations for AI-EO for agriculture in Sub-Saharan Africa. *Environmental Research Letters*. (2023).
8. Nakalembe, C. et al. A review of satellite-based global agricultural monitoring systems available for Africa. *Global Food Security* **29**, 100543 (2021).
9. Potapov, P. et al. Global cropland expansion in the 21st century [Dataset]. Global Land Analysis and Discovery. <https://glad.umd.edu/dataset/croplands> (2021).
10. Digital Earth Africa Digital Earth Africa Cropland Extent Map (2019) [Dataset]. Registry of Open Data on AWS. <https://registry.opendata.aws/deafrica-crop-extent/> (2019).
11. Xiong, J. et al. NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Global Food Security-support Analysis Data (GFSAD) Cropland Extent 2015 Africa 30 m V001 [Dataset]. NASA EOSDIS Land Processes DAAC. <https://doi.org/10.5067/MEaSUREs/GFSAD/GFSAD30AFCE.001> (2017).
12. Paliyam, M., Nakalembe, C. L., Liu, K., Nyirawung, R. & Kerner, H. R. Street2Sat: A Machine Learning Pipeline for Generating Ground-truth Geo-referenced Labeled Datasets from Street-Level Images. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*. <https://www.climatechange.ai/papers/icml2021/74> (2021).
13. Wu, F., Wu, B., Zhang, M., Zeng, H. & Tian, F. Identification of Crop Type in Crowdsourced Road View Photos with Deep Convolutional Neural Network. *Sensors* **4**, 1165, <https://doi.org/10.3390/s21041165> (2021).
14. Yan, Y. & Ryu, Y. Exploring Google Street View with deep learning for crop type mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* **171**, 278–296, <https://doi.org/10.1016/j.isprsjprs.2020.11.022> (2021).
15. Waldner, F. et al. Roadside collection of training data for cropland mapping is viable when environmental and management gradients are surveyed. *International Journal of Applied Earth Observation and Geoinformation* **80**, 82–93, <https://doi.org/10.1016/j.jag.2019.01.002> (2019).

16. Laguarda, J., Friedel, T. & Wang, S. Combining deep learning and street view imagery to map smallholder crop types. *arXiv:2309.05930*. (2023).
17. d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L. & Velde, M. Monitoring crop phenology with street-level imagery using computer vision. *Computers and Electronics in Agriculture* **196**, 106866, <https://doi.org/10.1016/j.compag.2022.106866> (2022).
18. Helbich, M., Danish, M., Labib, S. M. & Ricker, B. To use or not to use proprietary street view images in (health and place) research? That is the question. *Health & Place* **87**, 103244, <https://doi.org/10.1016/j.healthplace.2024.103244> (2024).
19. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**. (2016).
20. 6th Grain Annual Cropped Area Maps: Maize Density Layer for 2016 and 2019. <https://cropmapping.6grain.com> (2019).
21. Pérez-Hoyos, A. Global crop and rangeland masks. European Commission, Joint Research Centre (JRC). <http://data.europa.eu/89h/jrc-10112-10005> (2018).
22. Mohajan, H. Food and Nutrition Scenario of Kenya. MPRA Paper. <https://mpra.ub.uni-muenchen.de/56218/> (2014).
23. Kenya Roads Board Road Network Classification. Kenya Gazette Supplement No. 4 of 2016. <https://maps.krb.go.ke/kenya-roads-board12769/maps/109276/1-road-network-classification-> (2016).
24. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125). (2017).
25. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). (Springer International Publishing 2015).
26. Chaurasia, A. & Culurciello, E. LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE Visual Communications and Image Processing (VCIP)* (pp. 1–4) <https://doi.org/10.1109/VCIP.2017.8305148> (IEEE, 2017).
27. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *CVPR* (pp. 1800–1807). IEEE Computer Society. (2017).
28. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4700–4708) (2017).
29. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE. (2016).
30. Tan, M. & Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Chaudhuri, K. and Salakhutdinov, R. (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 (pp. 6105–6114) (PMLR, 2019).
31. Ngeno, V. Profit efficiency among kenyan maize farmers. *Heliyon*, **10**(2) (2024).
32. McCullough, E. B., Lu, M., Nogue, Y., Arsenault, J. & Zhen, C. Nutrient adequacy for poor households in Africa would improve with higher income but not necessarily with lower food prices. *Nature Food* **5**(2), 171–181 (2024).
33. Lee, Y., Zhang, S., Li, M. & He, X. Blind inverse gamma correction with maximized differential entropy. *Signal Processing* **193**, 108427, <https://doi.org/10.1016/j.sigpro.2021.108427> (2022).
34. Kenya Roads Board (KRB) Map Portal (2021).
35. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3–11) (Springer International Publishing, 2018).
36. Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5MB model size. *arXiv:1602.07360* (2016).
37. Zvonkov, I. *et al.* Helmets Kenya Crop Type Dataset. Available at: <https://zenodo.org/records/15133324> (Zenodo, April 2025).
38. Manimurugan, S., Singaram, R., Nakalembe, C. & Kerner, H. Geo-referencing crop labels from street-level images using Structure from Motion. In *Proceedings of the International Astronautical Congress, IAC*, vol. 2022 (International Astronautical Federation IAF, 2022).
39. Nakuru. *Encyclopedia Britannica*. <https://www.britannica.com/place/Nakuru>
40. Zanaga, D. *et al.* ESA WorldCover 10 m 2021 v200. *Zenodo*. <https://doi.org/10.5281/zenodo.7254221> (2022).
41. European Commission, Joint Research Centre (JRC) Kenya AOI. European Commission, Joint Research Centre (JRC) [Dataset]. <http://data.europa.eu/89h/5b6245d3-e561-4f6c-8c09-627888063d11> (2021).
42. Jaiswal, A. Agriculture crop images. *Kaggle*. <https://www.kaggle.com/datasets/aman2000jaiswal/agriculture-crop-images> (2021).
43. Azam, M. W. Agricultural crops image classification. *Kaggle*. <https://www.kaggle.com/datasets/mdwaquarazam/agricultural-crops-image-classification/data> (2022).
44. Van Tricht, K. *et al.* WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data* **15**(12), 5491–5515, <https://doi.org/10.5194/essd-15-5491-2023> (2023).
45. D'Andrimont, R., Yordanov, M., Martinez Sanchez, L. & Van Der Velde, M. FlevoVision Dataset: Street Level Imagery with BBCH phenology label. *European Commission*. <https://publications.jrc.ec.europa.eu/repository/handle/JRC128955> (2022).

## Acknowledgements

This work was funded by Lacuna Fund Project NO. 2020-01-202 - "Helmets Labeling Crops". Additional Support for data collection in Kneya from the NASA Harvest program No. 80NSSC23M0032, NASA SERVIR Grant No. 80NSSC20K0264, and SwissRe Foundation Grant EO-NAM. Partner organizations in implementation testing included NAACRI (Uganda), RCMRD (Kenya, Zambia), Makerere (Uganda), Milsat (Nigeria), NARSDA (Nigeria). We gratefully acknowledge the invaluable contributions of the many individuals who supported the creation of the Kenya dataset. Special thanks go to the team at LocateIT for initial testing. In particular, we would like to thank Vivianne Meta, Stephen Sande, Majambo Gamoyo, and Julius Buyengo for their instrumental roles in this project. We also thank Chad Smith for the design and creation of the helmets pipeline diagram (Fig. 1). We also extend our appreciation to the following contributors for their efforts in data collection: Anne Koimburi, David Letuati, Josphat Kioko, Fredrick Oscar Owino, Mary Njine, Lucy Wairimu, Macharia Dorothy Lemein, Consela Mwaura, Hussein Misango, Joseph Mwangi Macharia, Rachael Mwangi, Joseph Kering, Benard Makala, Charles Githiri, Moses Gichohi, Cecilia Wachera Kioria, Peris Towet, Annastasia Mbula, Kenneth Muiruri, Lucy Muguchia, Jane Gitonga, Annet Kopito, Francis Ruingu, Julius Mwangi, Joyce Mugwe, Jane Waihenya, Kamakei Benson, Eunice Kavata, Daniel Mutuku, Caroline Mwenze, Peterson Ngai, Ezekiel Letoya, Daniel Nyalita, Martin Ndegwa Nganga, George Muraguri Mwangi, Francis Gathongo, Benard Samoei, Cosmus Muli, Evans Kimuge, Florence Mwai, Jacob Kibet Kirui, K Richard, Samuel Mburu, Peter Langat, Caroline Warui, Stephene Romanos, Martin Turere, Kennedy Kivuva, Simon Kariuki, Charles Mbau Kabia, Rosemary Kimani, Wilson Kirui, Antony Kioko Waitiki Kariuki, John Thumbi Maina, David Kabangi, Simon Wanjohi, and Mary Mwaura. Your collective efforts have been critical in ensuring the quality and comprehensiveness of this dataset, which will contribute significantly to agricultural research and food security in the region.

### Author contributions

C.N., H.K. and I.Z. conceived the study and experiment(s), C.N., K.M., J.K. led the data collection effort, I.Z., K.M., J.K., B.T., K.J., conducted the experiments, I.Z., C.N., D.B.F., J.K., A.P., I.S., C.A.W., A.M.T., S.J., P.M.L. analyzed the results. The manuscript was written by C.N. and I.Z. All authors reviewed the manuscript and provided input on results and experiments.

### Competing interests

The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to C.N.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025