

## Assessing cropland disagreement in Tanzania using machine learning methods with Sentinel-2 and Planet Scope imagery

Vincent Wayuga Ogweno <sup>a,b</sup>, Nathan Moore  <sup>a,c</sup> and Dan Wanyama  <sup>d</sup>

<sup>a</sup>Department of Geography Environment and Spatial Sciences, Michigan State University, East Lansing, MI, USA; <sup>b</sup>Department of Mathematical Sciences (Climate Option), AIMS Rwanda, Kigali, Rwanda; <sup>c</sup>Center for Global Change and Earth Observation, Michigan State University, East Lansing, MI, USA; <sup>d</sup>Department of Geography, University of Connecticut, Storrs, CT, USA

### ABSTRACT

East Africa faces major land use pressures arising from the harsh effects of climate change and fast-increasing human populations. Recent efforts to increase food production in this region have focused on cropland mapping and approaches to improve yield. Generating accurate cropland maps provides critical support to inform policy, investment, and logistical decisions that address food security, but cropland disagreement assessment studies are lacking mostly due to the unavailability of fine-scale data. Understanding the inconsistent nature of pre-existing global cropland mapping products necessitates disagreement assessment as the most valuable urgent tool to align agricultural resource allocation and policy realignment to tackle food security challenges. This study evaluates cropland disagreement in Kilombero and Ulanga districts in Morogoro, Tanzania based on land use land cover (LULC) maps generated using machine learning methods and high-resolution PlanetScope and Sentinel-2 data sets acquired from 1 January 2017 to 31 December 2018. We created time-variant median composites and performed a three-step assessment; (i) generating multi-class LULC maps in Google Earth Engine, (ii) generating croplands and non-cropland LULC classes from the multi-class images, and (iii) assessing cropland disagreement based on GIS conditional statements. Results show minimal disagreement in mapping cropland based on higher-resolution PlanetScope and Sentinel-2. Results based on the Random Forests and Support Vector Machines relatively outperform those from Classification and Regression Trees achieving 93%, 89%, and 83% with PlanetScope and 91%, 86%, and 76% with Sentinel-2, respectively. On average, 73% of the pixels were consistently classified while 27% were misclassified between cropland and non-cropland classes. Prior unsupervised land cover cluster analysis may substantiate the quality of sampling for accurate land cover classification. Our results indicate that in data-sparse regions where confirmatory ground truth data are lacking, remotely sensed cropland identification may be more successful using Support Vector Machines on newer high-resolution imagery, thereby improving crop forecasts.

### ARTICLE HISTORY

Received 28 June 2023

Accepted 9 October 2023

### KEYWORDS

PlanetScope; Google Earth Engine; machine learning; binary classification; Sentinel-2; East Africa



## 1. Introduction

Accurate cropland mapping is an integral part of making agricultural resource allocations and subsequent productivity projections. These projections are critical to aid world food production goals as well as for making accurate estimates about food security (Buchhorn et al. 2020; Li et al. 2021; Löw et al. 2015; Nkwasa et al. 2022; Teluguntla et al. 2018; Xiong et al. 2017; Yadav and Congalton 2018). In the developing world, cropland area estimates are difficult to calculate due to insufficient field data collection, smallholder dominant agricultural systems, constantly shifting land use, and weather variability (Waldner et al. 2015; Xiong et al. 2017) which are critical to authenticate possible disagreements in existing cropland masks. Yet, in these countries, cropland extent is a crucial issue in achieving important Millennium Development Goals such as food security, poverty reduction, and famine early warning (Li et al. 2021; Liu et al. 2020). In addition, accurate cropland mapping provides information vital to environmental assessments, carbon accounting, crop-type distribution maps, and natural resource management. Poor cropland information quality can be managed to some degree through hybridized maps (See et al. 2015) and through more intensive stakeholder involvement to resolve large disagreements in the cropped areas (Fritz et al. 2012). Efforts to include citizen science like Geo – Wiki (<https://www.geo-wiki.org>) also show some promise, but participation and training remain low in many parts of the world where land cover observations would be most helpful (See et al. 2015).

In the past, cropland mapping in East Africa has largely been a global and/or continental mapping effort seeking to understand land cover change dynamics and the subsequent level of cropland expansion in the sub-Saharan region, using a variety of approaches including Bayesian algorithms and shapelets (Li et al. 2021; Xiong et al. 2022). Unfortunately, the high complexity of land cover types, farm size, and local landscapes has consistently made it difficult to classify smallholder cropping systems carefully and accurately to aid production-level forecasts (Kirimi et al. 2018; Thonfeld et al. 2020; Verburg et al. 2009), hence the need for more robust and locally accurate cropland maps.

The wetlands of the Kilombero floodplains are prospective food production zones due to their capacity to consistently deliver clean water and, as a result, favourable conditions for agricultural production (Thonfeld et al. 2020). National and international efforts to make Kilombero a major producer for Tanzania are showing some progress, but a lack of abundant observations on crop yield and crop extent makes it challenging for System of Rice Intensification (SRI) and project yields. This makes Kilombero a fruitful area to test. Twisa's (Twisa and Buchroithner 2019) long-term assessment to understand changes in land use and land cover (LULC) patterns of the upstream and downstream Wami River basin is a further motivation to focus on Kilombero, as the region experiences almost similar environmental factors as the Kilombero River basin.

More broadly, land cover classification has recently been used to investigate environmental change and its drivers and consequences in the Kilombero Region of Tanzania, a region undergoing fast societal and environmental turbulence, to derive inferences about LULC change in the region. Work by (Johansson and Abdi 2020) demonstrates that land cover classification and local perceptions are inconsistently inaccurate within this region. Additionally, LULC change detection algorithms do not conform to the local perceptions regarding changes in forest cover which is believed to have been encroached rapidly over

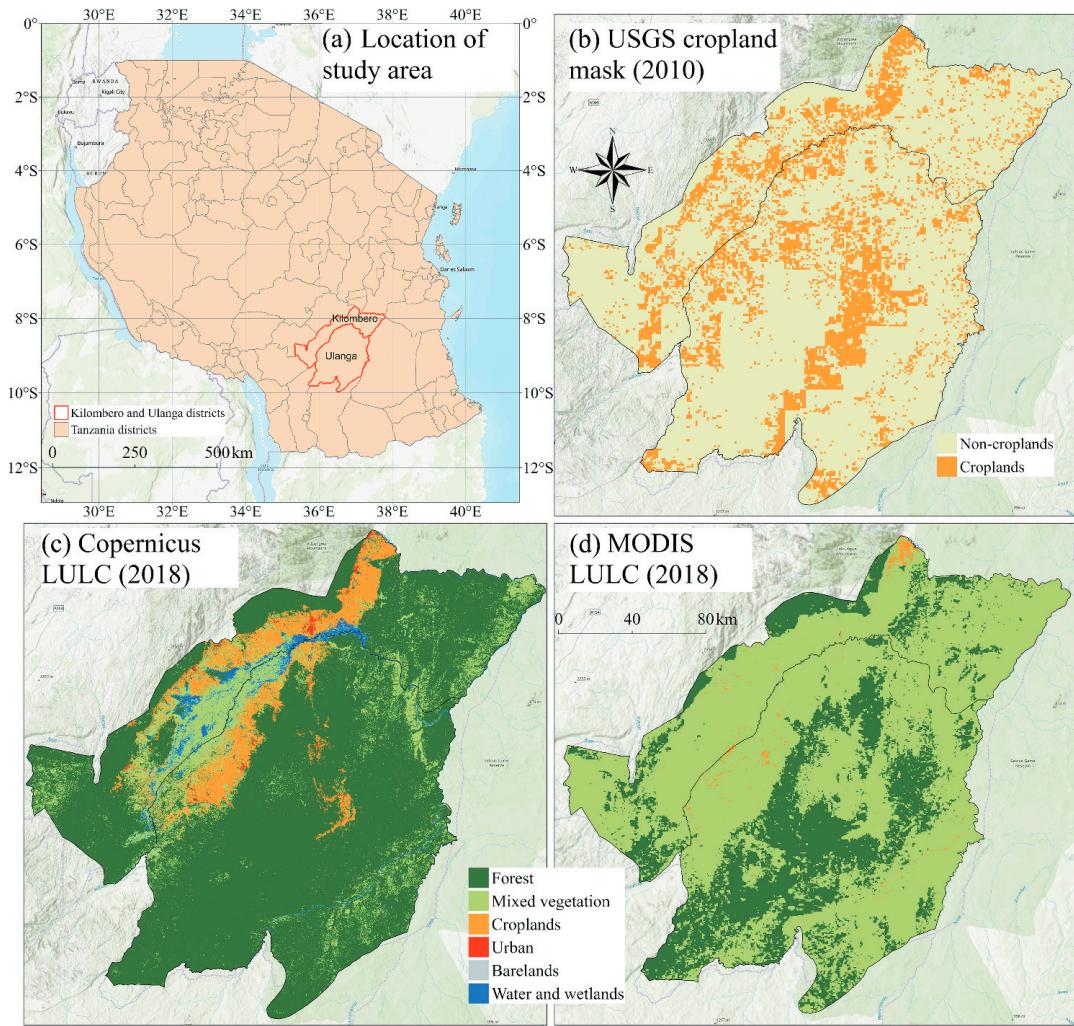
time. Consequently, the Stiegler's Gorge Dam is a promising initiative set to level out water table fluctuations hence giving Tanzania a comparative advantage within the region in the context of the East African Power Pool with an estimated cost of energy production of \$US2 c/though with associated exposure to flood risk (Kichonge 2018).

Figure 1 illustrates the existing disagreement among global land cover map products and subsequently the inaccuracy in determining the cropland masks from these maps. Coordination with farmers and other national and regional stakeholders like AgriSense-STARS (Dempewolf et al. 2015), RCMRD (Ottichilo 2006), GEOGLAM (Becker-Reshef et al. 2019), and SAGCOT (Cisneros-Araujo et al. 2021) can substantially improve such mapping products and their use in agricultural resource allocation management efforts. Past and ongoing work points out that seasonal variability complicates cropland accuracy around the sub-Saharan region among others, which calls for better information on cropping systems distribution (Becker-Reshef et al. 2019). SAGCOT recognizes that prior LULC studies demonstrate a decline in wetland and forest area due to farmland expansion from 1999 to 2016 (Johansson and Abdi 2020). Careful LULC mapping done based on Sentinel-2 and PlanetScope's high-resolution data proceeded by discovering existing disagreements could give room for necessary adjustments while developing reliable resource allocation methodologies. A critical research gap, therefore, is determining where LULC classifications disagree, especially in locations where ground validation is inadequate. This could guide future ground validation efforts.

### 1.1. Research gap

Globally available satellite data have been shown to enable national and global agricultural mapping based on expert analysis and change detection techniques. The challenge however is that there has never existed globally consistent yet locally relevant, multidecadal, cropland time-series data at spatial resolutions of approximately 30 m per pixel (Li et al. 2021). This calls for more analytical spatiotemporal techniques for performing accurate land-use mapping to identify hotspots for sustainable agricultural intensification (Potapov et al. 2020, 2022). Nevertheless, all of the attempts discussed above have been constrained by the lack of high-resolution reference data which is crucial for developing an accurate and precise knowledge base for machine learning algorithms, class identification and labelling, and class validation (Gumma et al. 2011). The success of these attempts has been further hindered by coarse-resolution, satellite-derived global land cover maps originally designed for climate change studies, such as GlobCover, and maps made using MODIS data, which are some of the main sources of current cropland information (Fritz et al. 2012).

Thonfeld on (Thonfeld et al. 2020) presented the first land cover classification study based on 30 m Landsat-8 data for the Kilombero area. With the emergence of the Sentinel-2 Multispectral Instrument and PlanetScope platforms, we now have significantly higher-resolution data available to improve cropland mapping by using data that can inform more accurate land cover categorization for robust training samples collection aimed at enhancing pre-existing classification mechanisms. The current study overcomes the limitations of prior work by collecting more high-resolution reference training and validation data from Planet Labs' Norway's International Climate and Forest Initiative (NICFI) (Gumma et al. 2011; Xiong et al. 2017). To further minimize these challenges, we have considered a regional focus with multiple machine learning evaluation (Thonfeld et al. 2020; Xiong et al. 2017).



**Figure 1.** Location of the Kilombero and Ulanga districts and comparison of the spatial distribution of croplands and other land uses and land covers from three different sources. (a) location of Kilombero and Ulanga districts relative to other districts in Tanzania. (b) spatial patterns in the distribution of croplands in 2010 based on the USGS cropland mask. (c) land use and land cover in Kilombero and Ulanga districts based on the Copernicus global land cover data for 2018. (d) land use and land cover based on the MODIS MCD12Q1 land cover data for 2018.

Therefore, this study proposes a spatial disagreement assessment methodology by integrating the consistency and inconsistency of different LULC products to evaluate the LULC classification accuracy assessment (Dong et al. 2022). The performance of the proposed method is demonstrated using the LULC classification products of Kilombero and Ulanga Districts of Tanzania using PlanetScope and Sentinel-2 high-resolution data with machine learning models, as a case study. In the case study, we first standardized and reclassified multi-resolution remote sensing data into seven classes, then applied supervised classification methods to reveal the regions with profound multiple classifications according to different machine learning models with the two data sets used, and finally carried out the disagreement assessment method to assess the extent of the disagreements to make generalizations about pre-existing LULC products.

## 2. Materials and methods

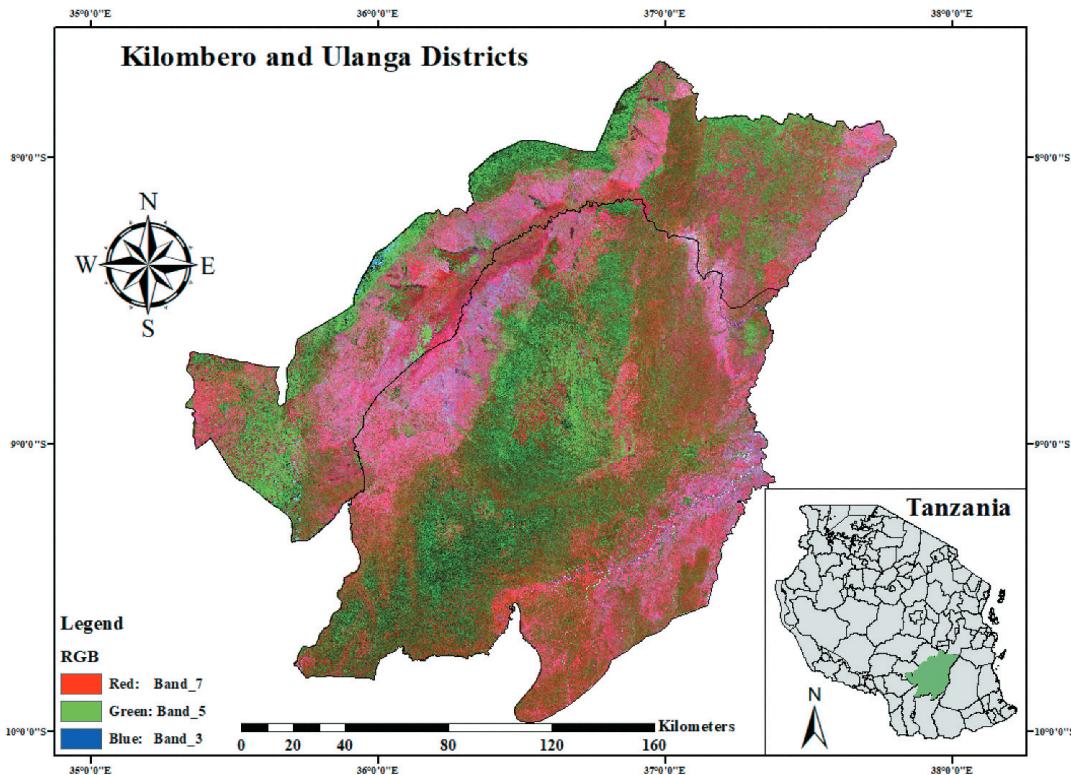
### 2.1. Kilombero and ulanga districts

Kilombero and Ulanga districts are two of the six districts in the Morogoro region of southwestern Tanzania. These districts are situated in the vast floodplain between the Kilombero River in the southeast and the Udzungwa Mountains in the northwest. The study area is located at the foot of the Great Escarpment of East Africa in the southern half of Tanzania, about 300 km from the Indian Ocean coast, lies between longitudes 34.563°E and 37.797°E and latitudes 7.654°S and 10.023°S. The seasonal hydrological variation is substantial – the floodplain becomes regularly inundated during the wet season, while it dries up during the dry season, except for areas with permanent wetlands and water bodies. This region is marked by a rainy season from November to May and a dry season from June to October, with annual rainfall and temperature ranging between 1200 and 1800 mm and 26°C and 32°C, respectively (Balama et al. 2013).

Within the floodplain, primary LULC types are agriculture, forestry, urbanization, and protected lands (Thonfeld et al. 2020). This area is predominantly rural, with the only major settlement being the semi-urban district headquarters in Ifakara. The majority of the citizens are subsistence farmers of maize and rice. The Kilombero region is a good place to study agricultural expansion for many reasons. The region has developed in the last century into a major rice producer in Tanzania and is hence a critical focus for rice mapping. Additionally, this area has been critical for the government's SRI efforts since 2009. However, this region has also been known for data poverty and a lack of efficient monitoring (Symeonakis and Drake 2004).

The Kilombero region serves as an excellent site to expand and intensify food production. The Tanzanian government has engaged extension officers to provide technical assistance to farmers in order to boost productivity in this region. Accurate and reliable data on cropland and the locations of major crop varieties are essential for making future policy, economic, and logistical decisions that address food security. Adequate knowledge of the current situation is a necessity for making a thorough assessment of the causes and consequences of future changes in land use. Through this research study, we hope to raise awareness of the outdated and subpar regional cropland maps and solicit high-resolution data for immediate improvements. The improved, more accurate cropland maps are necessary for the government to plan adequately for their ever-increasing population. This also informs strategic decision-making and policy alignment for a secure future for the population.

Figure 2 illustrates the spatial heterogeneity upon which different land use land cover map products have been generated. The classifications are shown in Figure 7 have inconsistencies in accurately mapping cropland extent in part due to the small differences in reflectance between grasslands, wetlands, and croplands. This figure gives a priori in understanding this work ahead to our classification attempts and hence the best approach to analysing LULC patterns in a region with such distinct natural characteristics.



**Figure 2.** RGB median composite image (2017–18) from Sentinel-2 for Kilombero and Ulanga districts in Tanzania.

## 2.2. Data and processing

Table 1 illustrates the specific datasets obtained from PlanetScope and Sentinel-2 publicly available data from Google Earth Engine (GEE) with a specific focus on the earliest dates for both datasets. The PlanetScope data used were atmospherically corrected bi-annual median composites for tropical Africa with 4 m spatial resolution from Planet Lab's NICFI plug-in on GEE. On the other hand, we used Level-2 atmospherically corrected Surface Reflectance (SR) imagery from Sentinel. These data were resampled from 20 m to obtain the 10 m spatial resolution imagery. At this scale, cropland can be identified by a trained user and used to select training data for the machine learning classification algorithms.

We selected all PlanetScope scenes and Sentinel-2 Multispectral Instrument imagery available on the GEE platform from NaN Invalid Date to NaN Invalid Date. We integrated imagery from the Sentinel-2 satellite containing spectral bands 2, 3, 4, 8, a True Colour Image (TCI), Aerosol Optical Thickness (AOT) and Water Vapour

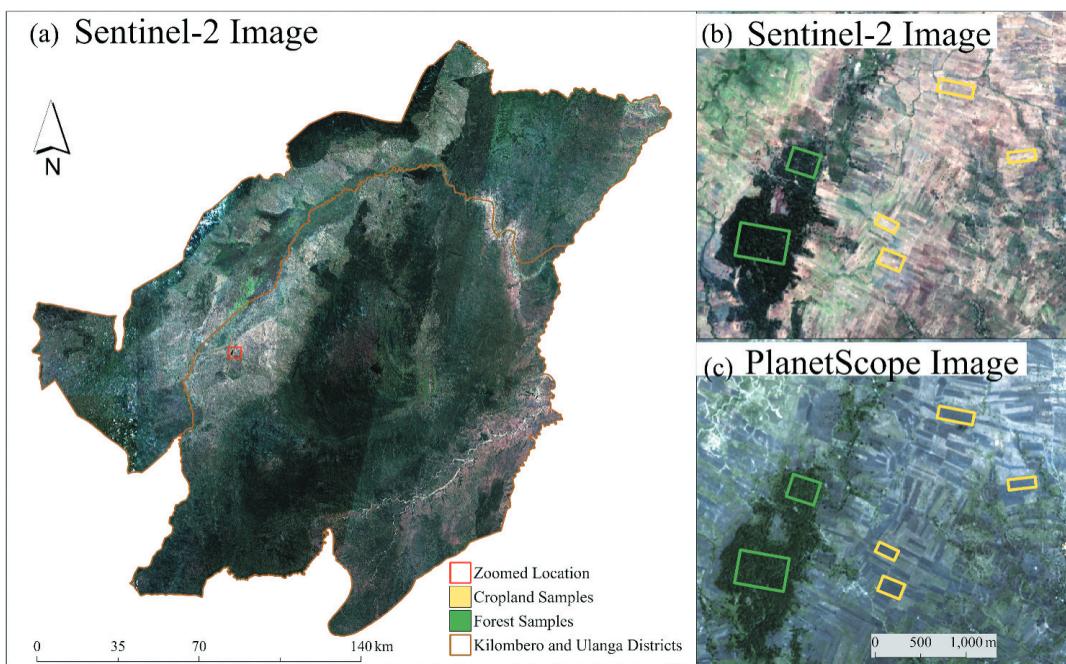
**Table 1.** Description of data used in this study.

Categories	Variables	Time	Spatial scale	Data source
Overall features	Spectral bands	2017/2018	10 m	Sentinel-2
	Spectral index	2017/2018	4 m	Planet Scope
Spatial feature	Elevation, slope	2017/2018	90 m	STRM
	Textural features (GLCM)	2017/2018	-	GLCM

(WV)). Five vegetation indices were also calculated from these imagery, namely the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), Bare soil Index (BSI), normalized difference water index (NDWI), and normalized difference built-up index (NDBI) (Saadi and Wijayanto 2021). Sentinel-2 data sets used in this study can be accessed through USGS Earth Explorer: <https://earthexplorer.usgs.gov/>. We applied a 20% cloud filter to the imagery to obtain cloud-free scenes. To reduce signal noise, fill gaps due to masked cloudy pixels, and ultimately improve the quality of the imagery to better distinguish between LULC classes, we generated two-year median composites for the period 2017–2018. To aid in visualization, we displayed the shortwave infrared, near-infrared, and red bands in the Red, Green, and Blue colour guns to display distinct land cover characteristics. We additionally included spatial reference features including elevation, slope, and textural features from Shuttle Radar Topography Mission (SRTM) and the Gray Level Co-occurrence Matrix (GLCM) respectively to aid accurate training and testing data collection.

With the help of PlanetScope high-resolution imagery, we zoomed into the study area and obtained representative data samples for training our machine learning classification models as shown in Figure 3. Notably, the clarity of these images differs from one dataset to the other depending on their respective spatial resolutions. Our study reveals that PlanetScope data (with the highest spatial resolution) was best for collecting training samples hence very influential in the classification results.

We generated samples for the target LULC classes by creating feature collections in the Kilombero region. The distribution of these features was proportional to the relative extent of the LULC classes. Our main land cover labels were croplands and non-croplands (forest, water, woodlands, wetlands, built-up area, and bare soil). We zoomed into specific



**Figure 3.** Comparison between Sentinel-2 (10 m spatial resolution) and PlanetScope (4 m spatial resolution) images. Examples of training samples for cropland and forests are shown.



locations in the study area and obtained representative data samples for training our machine learning classification models as shown in [Figure 3](#). We defined the class labels of these training data while considering the spectral properties of the surfaces as observed from high-resolution imagery from Sentinel-2 and PlanetScope. The PlanetScope base maps we used were atmospherically corrected bi-annual median composites for tropical Africa with 4 m spatial resolution from Planet Lab's NICFI plug-in on GEE.

Based on a manual assessment of the original Sentinel-2 data and the Planet NICFI high-resolution images accessible on GEE, the training and testing samples were gathered using a simple random sampling criterion ([Thanh Noi and Kappas 2017](#)). We generated 5504 samples over the entire study area and used these to extract from input bands the values of pixels at those locations. We split and used 70% of these data for training the models and the remaining 30% for independent validation of model results.

### **2.3. Classification algorithms and hyperparameter tuning**

When employing classification and regression trees (CART), random forests (RF), and support vector machines (SVM), tuned parameters are crucial for obtaining high-accuracy outcomes. Different tuning steps and tuned parameters are used with each classifier ([Thanh Noi and Kappas 2017](#)). We tested a range of tuning parameters for each classifier, with the ideal values chosen based on the highest overall classification accuracy. In this study, the performance of classifiers was compared using the categorized results under each classifier's optimal parameters.

### **2.4. Classification and Regression Trees (CART)**

CART is a binary decision tree classifier that uses logical if-then statements that evaluate the input variables and select the variable with the most information gain as the basis for the node splitting at each level ([Breiman 2001](#)). With this method, the input data is randomly split into a set number of groups, and trees are created utilizing all but one of the groups. The pruned tree that shows the least deviation is chosen after the left-out group is utilized to validate the tree. In this study, we used the 'Classifier.smileCart()' function available in GEE to perform the LULC classification. We selected 10 as the best cross-validation factor for pruning. [Table 2](#) shows the best set of tuned parameters used in this study.

### **2.5. Random Forests (RF)**

RF is an aggregation of k CART classifiers that overcomes the overfitting challenges with CART and is the most extensively used classifier in LULC classifications. The 'bagging' strategy is used by RF, which randomly chooses a subset of characteristics from the input observations for each tree. The number of trees and the

**Table 2.** Input parameter values for CART classifier.

Description	Value
Minimum training points to allow the creation of a node	1
Minimum training points at a node to allow splitting	1
Maximum depth of the initial tree	10
Cross-validation factor for pruning	10

variables at each split is the primary input parameters for RF. A very large number of trees does not always mean that the classification is more accurate, as, after a certain number of trees, more trees become redundant and do not contribute as much to the label prediction. The optimal tree count number is 100 or 500 with the square of the number of variables being the ideal choice to determine the best split (Belgiu and Drăguț 2016). We used the 'Classifier.smileRandomForest()' function available in the GEE library to perform the RF classification. Table 3 shows the best set of input variables selected for the RF classifier.

## 2.6. Support vector machines (SVM)

SVM is a popular classification algorithm that works by seeking to find an ideal hyperplane that separates the decision boundary across different classes. The choice of the cost parameter C, Gamma, and kernel functions heavily influences the selection of the support vectors. The cost parameter determines the severity of the penalty for incorrectly categorized data i.e. the higher the value of C, the less the amount of incorrectly classified data. Therefore, a grid search method is deployed to achieve the best values for C and Gamma. Studies like those by Hsu (Hsu, Chang, and Lin 2003) claim that because C is a scale parameter, an exponentially expanding sequence of C provides a better approximation to appropriate parameter selection. Additionally, linear kernel training is favoured for large datasets. The best set of tuned parameters for the support vector classifier is shown in Table 4.

## 2.7. Cropland agreement/disagreement assessment

Using ArcGIS and QGIS software we performed cropland disagreement and agreement assessment. Using GIS conditional statements to perform a binary analysis of classification results, we extracted cropland area from the classified maps and obtained the disagreement and agreement using the difference and sum functionalities, respectively.

**Table 3.** Input parameter values for random forests classification algorithm.

Description	Value
Number of variables per split	The square root of input variables
Number of trees	200

**Table 4.** Input parameter values for Support Vector Machines classifier.

Description	Value
SVM type	C_SVC
Kernel type	radial basis function(RBF)
Cost parameter	3510



## 2.8. Accuracy assessment

An effective way to assess machine learning models' performance is to train them with known data and evaluate their classification performance using ground truth data or a different dataset (Vabalas et al. 2019). An equally reliable approach, commonly called Train/Test Split, is to separate a portion of data before developing a machine learning model and use that data only for validation (Kohavi 1995). Unknown data is used to test machine learning models, which provides an objective assessment of how well the model will perform when it is used to make predictions in the real world (Vabalas et al. 2019). Based on our 70%, 30% splits, we constructed a confusion matrix that we used for the computation of the Producer's accuracy, User's Accuracy, Overall Accuracy, and Kappa Statistics using the predefined formulas in the GEE library (Adam et al. 2014). Based on the accuracy results obtained from different ML algorithms on our different data sets, we constructed and evaluated a 95% confidence level significant test by obtaining the Z statistics associated with the accuracy results obtained from each pair of the accuracy results.

The evaluated hypothesis states;

$$H_0 : \alpha \geq 0.05; \quad Vs \quad H_1 : \alpha < 0.05 \quad (1)$$

**Figure 4** Shows precisely the steps taken towards obtaining reliable disagreement assessment which is the profound objective of this work.

## 3. Results

### 3.1. Classification results

To clarify the disagreements highlighted while evaluating global land cover products, we used Sentinel-2 and PlanetScope data as discriminant thresholds because of their relatively high spatial and temporal resolution. This study produced LULC maps for each of the three ML algorithms on the two datasets.

## 4. Discussions

**Figure 5** shows the results from the three different machine learning classification algorithms trained on the same samples from Sentinel-2 and PlanetScope. The LULC based on Sentinel-2 identifies vegetation more effectively, albeit with inferior sensitivity to built-up areas and bare soil classes. Alternatively, LULC based on PlanetScope data has shown higher sensitivity to the built-up area and bare soil classes though with slightly lower sensitivity to vegetation cover especially forests and croplands using the SVM algorithm.

Trained on the same datasets, LULC maps obtained from Sentinel-2 data consistently show cropland areas towards the eastern part of the study area as seen from the cropland masks in **Figure 6a**. These LULC patterns are inconsistent with pre-existing global and regional LULC products (Buchhorn et al. 2020) and hence an indication of profound disagreement in LULC patterns at local scales.

**Figure 5** shows cropland masks generated from Sentinel-2 and PlanetScope data classification using the three ML classification algorithms in yellow. More generally,