IDENTIFICATION OF PADDOCKS WITH AGRICULTURE IN AFGHANISTAN, IRAN, AND SUDAN USING ARTIFICIAL INTELLIGENCE AND SATELLITE IMAGERY.

Adrián Cal

INTRODUCTION

This paper presents a model for the identification of agricultural paddocks using artificial intelligence and satellite imagery. This work arises as a result of participation in the GEO-AI Challenge for Cropland Mapping by ITU (zindi.africa/competitions/geo-ai-challenge-for-cropland-mapping-with-satellite-imagery) on the zindi.africa platform held in 2023 and supported by AI for Good (aiforgood.itu.int). The objective of this competition was to use artificial intelligence and freely accessible satellite imagery to map paddocks with annual crops in Afghanistan, Iran, and Sudan.

DATASET

For the competition, a total of 3000 ground truth points were provided for Afghanistan, Iran, and Sudan. Each country had 1000 points, 50% for training and 50% for testing. Within the training set, half corresponded to points with agriculture (1) and points without agriculture (0). For the test set, participants did not know the value of the target variable. For Iran and Sudan, the training and test points correspond to the period from July 2019 to June 2020. In the case of Afghanistan points, they correspond to the period of April 2022. In Table 1 and Figure 1, the quantity and location of the training and test points are presented.

COUNTRY	TRAIN		TEST	TOTAL
	AGRICULTURE	NON-AGRICULTURE		
IRAN	250	250	500	1000
SUDÁN	250	250	500	1000
AFGANISTÁN	250	250	500	1000
TOTAL	750	750	1500	3000

Table 1. Training and testing points by country and type of coverage (agricultural and non-agricultural).

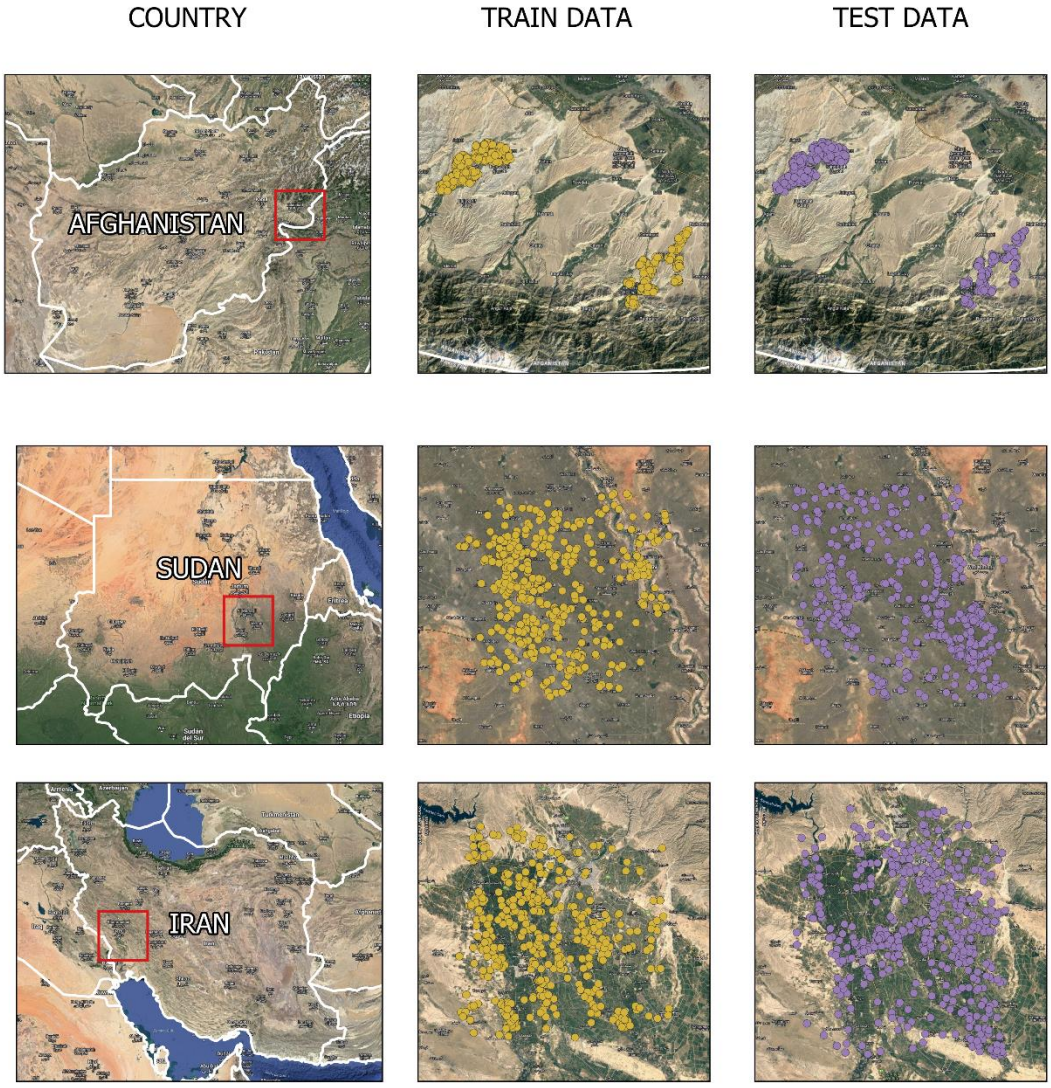

Figure 1. Location of the training and test points within the area of interest (red rectangle) by country.

SATELLITE IMAGES

For this competition, images from Sentinel 2 satellites (A and B) were used, specifically the Level-2A surface reflectance product of the Earth. The images were obtained and processed in the cloud with the planetary-scale platform for Earth science data and analysis Google Earth Engine (earthengine.google.com), which is free to use. Sentinel 2 has a sensor that covers 12 spectral bands, ranging from 444 nm to 2200 nm, with different pixel sizes (Table 2). Additionally, the images have other bands related to the quality of each pixel and other attributes, including the SCL, which classifies each pixel according to the values in Table 3. In Google Earth Engine (GEE), the Sentinel 2 product is specifically called COPERNICUS/S2_S2_HARMONIZED.

NAME	PIXEL SIZE (m)	WAVELENGTH (nm)	DESCRIPTION
B1	60	443.9 (S2A) / 442.3 (S2B)	AEROSOLS
B2	10	496.6 (S2A) / 492.1 (S2B)	BLUE
B3	10	560 (S2A) / 559 (S2B)	GREEN
B4	10	664.5 (S2A) / 665 (S2B)	RED

B5	20	703.9 (S2A) / 703.8 (S2B)	RED EDGE 1
B6	20	740.2 (S2A) / 739.1 (S2B)	RED EDGE 2
B7	20	782.5 (S2A) / 779.7 (S2B)	RED EDGE 3
B8	10	835.1 (S2A) / 833 (S2B)	NIR
B8A	20	864.8 (S2A) / 864 (S2B)	RED EDGE 4
B9	60	945 (S2A) / 943.2 (S2B)	WATER VAPOR
B11	20	1613.7 (S2A) / 1610.4 (S2B)	SWIR 1
B12	20	2202.4 (S2A) / 2185.7 (S2B)	SWIR 2
. . .			
SCL	20		Scene Classification Map (The "No Data" value of 0 is masked out)

Table 2. Sentinel 2 Level-2A image bands.

VALUE	DESCRIPTION
1	Saturated or defective
2	Dark Area Pixels
3	Cloud Shadows
4	Vegetation
5	Bare Soils
6	Water
7	Clouds Low Probability / Unclassified
8	Clouds Medium Probability
9	Clouds High Probability
10	Cirrus
11	Snow / Ice

Table 3. SCL Classes.

EXTRACTION OF TEMPORAL SERIES OF SPECTRAL INDICES

For Afghanistan, images from the period 01/07/2022 to 01/07/2023 were used. For Iran and Sudan, images from 01/07/2019 to 01/07/2020 were used. From the satellite images, it was decided to calculate the following spectral indices: DATT1, IRECI, NBR2, NDREI, NDVI, and NDWI. These indices are obtained by combining two or more bands: GREEN, NIR, RED, RED EDGE 1 to RED EDGE 4, SWIR 1, and SWIR 2. By calculating these indices, information from 560 to 2200 nm of the electromagnetic spectrum is being utilized.

INDEX	FORMULA	DESCRIPTION
DATT1	(RED EDGE 1 - RED EDGE 4) / (RED EDGE 4 - RED)	
IRECI	(RED EDGE 3 - RED) / (RED EDGE 1 / RED EDGE 2)	INVERTED RED EDGE CLOROPHYLL INDEX
NBR2	(SWIR 1 - SWIR 2) / (SWIR 1 + SWIR 2)	NORMALIZED BURN RATIO 2
NDREI	(RED EDGE 3 - RED EDGE 1) / (RED EDGE 3 + RED EDGE 1)	NORMALIZED DIFFERENCE RED EDGE INDEX
NDVI	(NIR - RED) / (NIR + RED)	NORMALIZED DIFFERENCE VEGETATION INDEX
NDWI	(GREEN - NIR) / (GREEN + NIR)	NORMALIZED DIFFERENE WATER INDEX

Table 4. Spectral indices used.

Before calculating the selected indices, each image was filtered to choose suitable-quality pixels and exclude poor-quality ones. Suitable quality pixels are those that, in the SCL band, have values of 4 (vegetation), 5 (bare soils), 6 (water), or 7 (Clouds Low Probability / Unclassified). On the other hand, low-quality pixels include, for example, those with values 1 or 3, which correspond to pixels with the presence of cloud shadows, saturated or defective. Once the images for each index have been obtained and filtered by quality, gaps appear in the places of the eliminated pixels. To get images without missing data, it was decided to generate synthetic images for ten days, according to the calendar in Table 5. Each pixel of the synthetic image takes the highest value of all the images that fall into the corresponding period, according to the table. In this way, there are 36 images for the analyzed period in each country.

SYNTHETIC IMAGE	IRAN - SUDAN		AFGHANISTAN	
	START DATE	END DATE	START DATE	END DATE
001	2019/07/01	2019/07/10	2022/07/01	2022/07/10
002	2019/07/11	2019/07/20	2022/07/11	2022/07/20
003	2019/07/21	2019/07/31	2022/07/21	2022/07/31
004	2019/08/01	2019/08/10	2022/08/01	2022/08/10
005	2019/08/11	2019/08/20	2022/08/11	2022/08/20
006	2019/08/21	2019/08/31	2022/08/21	2022/08/31
007	2019/09/01	2019/09/10	2022/09/01	2022/09/10
008	2019/09/11	2019/09/20	2022/09/11	2022/09/20
009	2019/09/21	2019/09/30	2022/09/21	2022/09/30
010	2019/10/01	2019/10/10	2022/10/01	2022/10/10
011	2019/10/11	2019/10/20	2022/10/11	2022/10/20
012	2019/10/21	2019/10/31	2022/10/21	2022/10/31
013	2019/11/01	2019/11/10	2022/11/01	2022/11/10
014	2019/11/11	2019/11/20	2022/11/11	2022/11/20
015	2019/11/21	2019/11/30	2022/11/21	2022/11/30
016	2019/12/01	2019/12/10	2022/12/01	2022/12/10
017	2019/12/11	2019/12/20	2022/12/11	2022/12/20
018	2019/12/21	2019/12/31	2022/12/21	2022/12/31
019	2020/01/01	2020/01/10	2023/01/01	2023/01/10
020	2020/01/11	2020/01/20	2023/01/11	2023/01/20
021	2020/01/21	2020/01/31	2023/01/21	2023/01/31
022	2020/02/01	2020/02/09	2023/02/01	2023/02/09
023	2020/02/10	2020/02/19	2023/02/10	2023/02/18
024	2020/02/20	2020/02/29	2023/02/19	2023/02/28
025	2020/03/01	2020/03/10	2023/03/01	2023/03/10
026	2020/03/11	2020/03/20	2023/03/11	2023/03/20
027	2020/03/21	2020/03/31	2023/03/21	2023/03/31
028	2020/04/01	2020/04/10	2023/04/01	2023/04/10
029	2020/04/11	2020/04/20	2023/04/11	2023/04/20
030	2020/04/21	2020/04/30	2023/04/21	2023/04/30
031	2020/05/01	2020/05/10	2023/05/01	2023/05/10
032	2020/05/11	2020/05/20	2023/05/11	2023/05/20
033	2020/05/21	2020/05/31	2023/05/21	2023/05/31
034	2020/06/01	2020/06/10	2023/06/01	2023/06/10
035	2020/06/11	2020/06/20	2023/06/11	2023/06/20

036	2020/06/21	2020/06/30	2023/06/21	2023/06/30

Table 5. Start and end dates for synthetic images by country.

Once the synthetic images of the selected spectral indices were obtained, the respective temporal series of these indices were extracted for each training and test point in each country, intersecting the points with the stack of the 36 synthetic images. This processing stage was carried out entirely in GEE, and as a result, tables in CSV format were obtained.

PROPOSED MODEL

For the development of the classification model, different algorithms belonging to the family of ensemble methods were evaluated. Among these, Random Forest (bagging type) and boosting-type algorithms, such as CatBoost, LightGBM, and XGBoost, were examined. After carrying out several tests, LightGBM was chosen. This algorithm was not used with its default parameters (vanilla mode). Still, instead, a strategy was applied to find the optimal values of the most relevant parameters to maximize performance on the test set. To optimize the parameters of LightGBM, the Hyperopt algorithm was chosen, which is based on Bayesian optimization to find the values of the parameters of the algorithm in question, in this case, LightGBM, that maximize an objective function, in this case, the F1-Score.

Before running Hyperopt, it is necessary to define which parameters of LightGBM will be optimized and establish a search space for the optimal values of each. Table 6 presents the optimized parameters and the designated search space for each.

PARAMETER	SEARCH SPACE
num_leaves	5 to 50
learning_rate	exp(-5) to exp(0)
max_depth	2 to 10
min_child_samples	5 to 30
colsample_bytree	0.1 to 0.75
subsample	0.1 to 0.75
n_estimators	10 to 100
reg_alpha	0 to 1
reg_lambda	0 to 1
subsample_for_bin	10 to 50

Table 6. Search space for LightGBM parameters.

TRAINING AND EVALUATION

The training and evaluation stage was done in Google Colab (colab.google.com). During the training phase, the temporal series of the six spectral indices corresponding to Afghanistan, Iran, and Sudan were used as input data. The LightGBM model was trained with five folds, stratifying according to the "Target" variable. The training was carried out by conducting 2000 runs with the Hyperopt algorithm to optimize the parameters of LightGBM to maximize the F1-Score.

When using stratified k-fold with k = 5, in each run of Hyperopt, LightGBM was trained with four folds, leaving one-fold for evaluation. This process was repeated for all possible fold combinations; for example, it was trained with folds 1, 2, 3, and 4, evaluating fold 5, and calculating the F1-Score. Then, it was trained with folds 1, 2, 3, and 5, evaluating on fold 4, and

the F1-Score was recalculated. In each run, the average F1-Score of the five-fold combinations was calculated.

In Table 7, the optimized values of the LightGBM parameters are presented. With these parameters and using the complete training set, LightGBM was trained. This trained model was used to classify the test set, which comprised the temporal series of the six indices for the three countries. As a result of applying this strategy, an accuracy of 0.93777 was obtained in the public score and 0.93333 in the private score of the competition.

PARAMETER	OPTIMAL VALUE
num_leaves	25
learning_rate	0.118
max_depth	10
min_child_samples	23
colsample_bytree	0.101
subsample	0.438
n_estimators	90
reg_alpha	0.323
reg_lambda	0.539
subsample_for_bin	26

Table 7. Optimized values of LightGBM parameters.