GEO-AI CHALLENGE FOR CROPLAND MAPPING BY ITU

14TH PLACE REPORT

JODI JHOURANDA SIREGAR

jwst.28@gmail.com

ABSTRACT

Timely and accurate crop maps are essential for various applications in agriculture as well as other relevant research fields, such as natural resources, environment, health, and sustainability. Cropland extent maps are the basic products for the practical agricultural applications. Numerous algorithms have been proposed for cropland mapping using satellite observations, and several freely available land cover products provide global cropland extent maps. However, the data have several limitations. With several new earth observation plans implemented and others to be implemented soon, more satellite imagery with increasing spatial and temporal resolution are available. Machine learning and artificial intelligence promise to improve the accuracy and robustness of land cover classification with satellite images (Zindi).In this competition report, we explore several machine learning and satellite data to build crop detection models. To predict crops, we build simple machine-learning pipelines with Sentinel-2 and Sentinel-1 satellite imagery as end-to-end machine-learning frameworks. The pipelines will cover data preprocessing, missing data imputation, and machine learning models. Finally, we compare all pipelines with different scenarios for crop prediction to find the best data and model to accomplish the objective. Sentinel-2 and stacking some classifiers are the best scenarios for field crop detection models.

DATASET

Competition Dataset

1. Satellite images: the participants should use all free-accessible satellite data (e.g. Landsat, Sentienl-1/2) in the test regions, the data pre-processing methodology are not limited, either. We will provide 15-day composited Sentinel-2 time series data in the Iran and Sudan test regions. For the test region in Afghanistan, the participants will have to collect the imagery independently. The participants need to share the availability of the data they used and how the pre-processing is conducted, if they used more data than the official provided data set.
2. Training samples: We will provide a limited number of training samples for cropland mapping, participants can also collect some training samples by themselves as cropland extent can be visually interpreted from Google Earth and Satellite images. Particularly, for the cropland mapping in Afghanistan, training samples were collected in April 2022, and the model is used to identify cropland extent in the same period.

Collected Open Public Dataset

1. Satellite data: We explore Sentinel-1 and Sentinel-2 satellite imagery as the model input. Monthly composite satellite imagery was collected from February 2019 to March 2023. All datasets were downloaded using Google Earth Engine.
2. ITU provides 3000 annotated crop points in Iran, Sudan, and Afghanistan. We added 953 manual annotations using Google Earth Engine to enrich data variability to reduce overfitting probability.

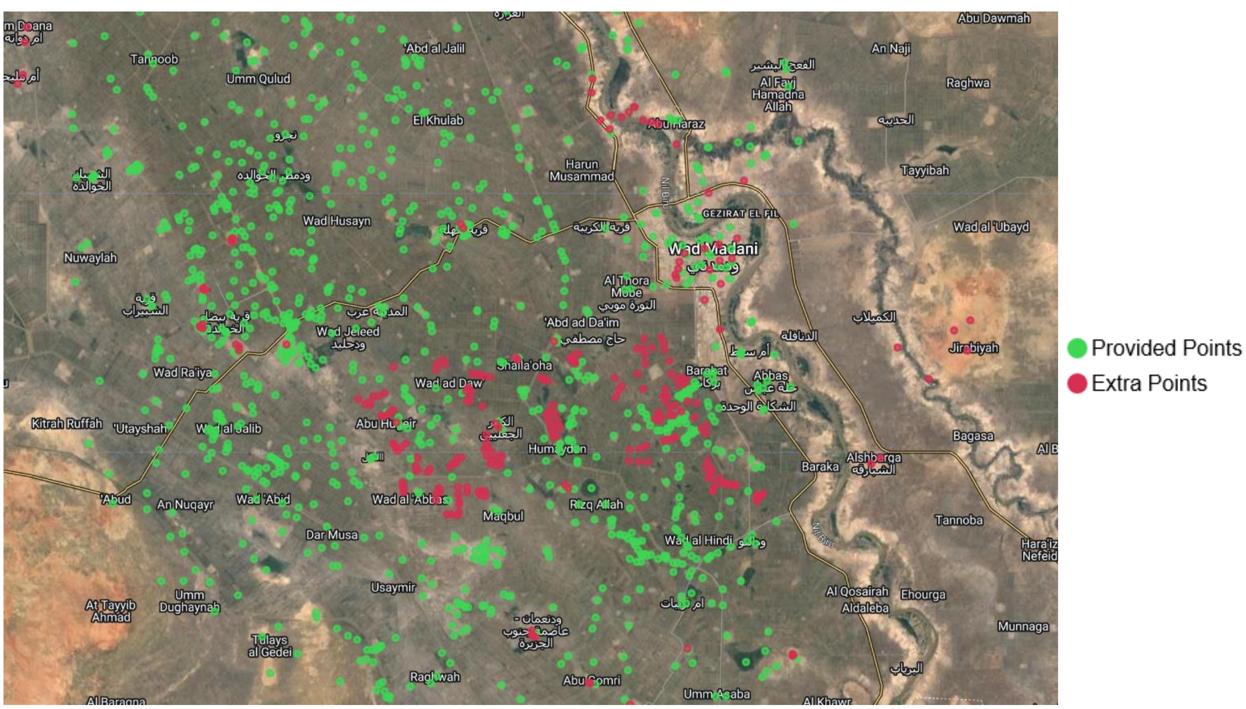

Figure 1 Samples for extra annotated dataset

METHODOLOGY

In this part, we define several machine learning pipelines for crop detection. The pipelines containing data preprocessing and machine learning training process :

1. Sentinel 2 Transformer
   In this pipeline, we use several bands from Sentinel-2 dataset. Several spectral indices are also calculated based on available bands to gather more spatial information. Yet, we join all bands and spectral indices for all monthly composited Sentinel-2 satellite imagery. Lastly, we calculate statistics (minimum, maximum, average, and standard deviation) for every band and spectral indices through all available time.

Table 1 Sentinel-2 Bands and Spectral Indices

Bands	**Description**
B01	Aerosols
B02	Blue
B03	Green
B04	Red
B05	Red Edge 1
B06	Red Edge 2
B07	Red Edge 3
B08	NIR
B09	Water vapor
B11	SWIR 1
NDVI (Normalized Difference MIR/NIR Normalized Difference Vegetation Index)	(B08-B04)/ (B08+B04)
GNDVI (Normalized Difference NIR/Green Green NDVI)	(B08-B03)/ (B08+B03)
EVI (Enhanced Vegetation Index)	2.5 * (B08 - B04) / ((B08 + 6.0 * B04 - 7.5 * B02) + 1.0)
EVI2(Enhanced Vegetation Indeks 2)	2.4 * (B08 - B04) / (B08 + B04 + 1.0)
AVI (Ashburn Vegetation Index)	(B08 * (1 - B04) * (B08 - B04))
BSI (Bare Soil Index)	((B11 + B04) - (B08 + B02)) / ((B11 + B04) + (B08 + B02))
SI (Soil Index)	((1 - B02) * (1 - B03) * (1 - B04))
NDWI (Normalized Difference Water Index)	(B03 - B08) / (B03 + B08)
NDMI (Normalized Difference 820/1600 Normalized Difference Moisture Index)	(B08 - B11) / (B08 + B11)
NPCRI (Normalized Pigment Chlorophyll Ratio Index)	(B04 - B02) / (B04 + B02)
NDBI (Normalized Difference NIR/Green Green NDVI)	(B11 - B08) / (B11 + B08)

2. Sentinel 1 Transformer
   In this pipeline, we use VV and VH bands from Sentinel-2, as shown in Table 2. The radar vegetation index was calculated to gather more spatial information. We also calculate statistics for all features through all monthly available datasets.

Table 2 Sentinel-2 bands and RVI

Bands	**Description**
VV	Single co-polarization, vertical transmit/vertical receive
VH	Dual-band cross-polarization, vertical transmit/horizontal receive
RVI	Radar Vegetation Index

3. Linear Pre-processing: Data Standardization and K-NN Imputer
   In this pipeline, we standardize all features to accelerate model convergences. We also use K-Nearest Neighbor imputation to impute missing data from Sentinel-1 and Sentinel-2

4. Machine learning models : Xgboost Classifier, Random Forest Classifier, and Catboost Classifier
   In this pipeline, we train several machine-learning models, such as random forest, XGBoost, and Catboost. We also explore the best model's parameters with a random search fine-tuning algorithm. Using a stacking Classifier, we build an ensemble machine-learning model to increase model prediction performance. We use logistic regression to combine several classifiers. All possible models that will be used in this mini-research are :

Table 1 All possible data and model

Data	Model
Sentinel-2	Random Forest
	Xgboost
	Catboost
	RF & XGB
	RF,XGB, Catboost
Sentinel-1	Random Forest
	Xgboost
	Catboost
	RF & XGB
	RF,XGB, Catboost
Sentinel-2 & Sentinel-1	Random Forest
	Xgboost
	Catboost
	RF & XGB
	RF,XGB, Catboost

5. Combining pipelines
   After defining all the pipelines we use in this competition, we combine those pipelines to build end-to-end machine learning pipelines to predict crops using a remote sensing dataset, as shown in Figure 2. We use a similar pipeline to Sentinel-1, Sentinel-2 and bombination of both datasets. We evaluate all pipelines with k-fold (k=5) with accuracy as the final metric evaluation.

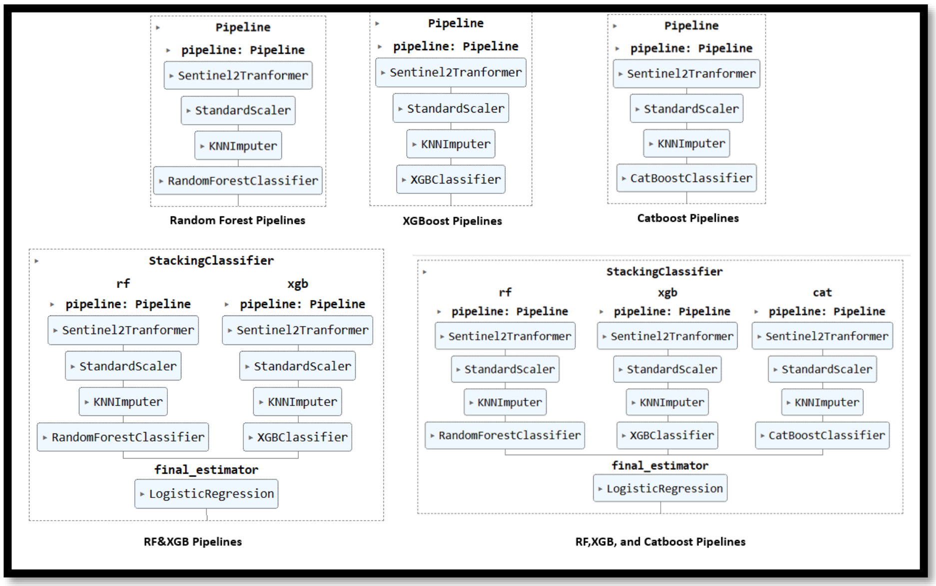

Figure 2 All Senitnel-2 end-to-end machine learning pipeline

6. Model comparison and pseudo-science research founding.
   In this final step, we compare all pipelines to find the best data and the best model for the crop detection task, which includes:
   - Comparing prediction accuracy using provided data points and extra annotated data points
   - Comparing best satellite imagery as machine learning input. In this phase, we also explore the best features from each satellite imagery using random forest feature importance.
   - Comparing the best machine learning models
   - Share some pseudo-science research findings that we learned in the research process. It is all undocumented lessons learned that we find in early model devel

RESULTS

1. Dataset
   In the data preparing process, using our pipelines, we have 1242 features for Sentinel-2 inputs and 112 features for Sentinel-1 inputs, including all bands from each satellite imagery, spectral indices, and statistics of spectral indices. The images span from February 2019 to May 2023. We have 3000 data points provided by Zindi and an extra 953 data points to increase data variability to reduce overfitting.
2. Comparing prediction accuracy using provided data points and extra annotated data points.

Table 2 Provided data points vs extra annotated data performance

Data Points	Data	Model	Mean Accuracy	Std	Private Leaderboard
3000	Sentinel-2	Random Forest	0.955	0.020	0.8838
3953	Sentinel-2	Random Forest	0.937	0.022	0.92

In this comparison, we found that using the provided dataset gives higher accuracy in k-fold evaluation than using an extra 953 manually annotated data points. However, using only 3000 data points in the testing data process caused model overfitting with only 88.38% accuracy in real test data. Adding more data points increases prediction performance by around 4% of accuracy.Comparing the best satellite imagery

In this part, we compare all machine learning pipelines using 3953 data points to learn the best satellite images for crop detection classifications. As shown in Table 3, Sentinel-2 outperformed Sentinel-1 in every trained model. Combining Sentinel-1 and Sentinel-2 results in a small increase in model accuracy and adds extra cost for model training.

Table 3 All posible pipelines comparison

Data	Model	Mean Accuracy (k=5)	Std
Sentinel-2	Random Forest	0.937	0.022
	Xgboost	0.909	0.030
	Caboost	0.938	0.038
	RF & XGB	0.933	0.023
	RF,XGB, Catboost	0.940	0.032
Sentinel-1	Random Forest	0.878	0.028
	Xgboost	0.838	0.037
	Caboost	0.895	0.054
	RF & XGB	0.878	0.029
	RF,XGB, Catboost	0.890	0.890
Sentinel-2 & Sentinel-1	Random Forest	0.936	0.024
	Xgboost	0.911	0.911
	Caboost	0.946	0.032
	RF & XGB	0.934	0.023
	RF,XGB, Catboost	0.943	0.032

Using random forest feature importance (for simplicity), we find that vegetation-related spectral indices statistics are the most important feature to predict crop location, as shown in Figure 3. Other building and water indices statistics also share useful information to predict non-crop areas. B12 (SWIR2) bands in March 2021 and 2022 also significantly impact prediction because the feature is useful for measuring soil and vegetation's moisture content, and they provide good contrast between different vegetation types. Some in-depth data exploration need to be done to learn more about crop condition in March to predict crop using only single-month satellite imagery efficiently.

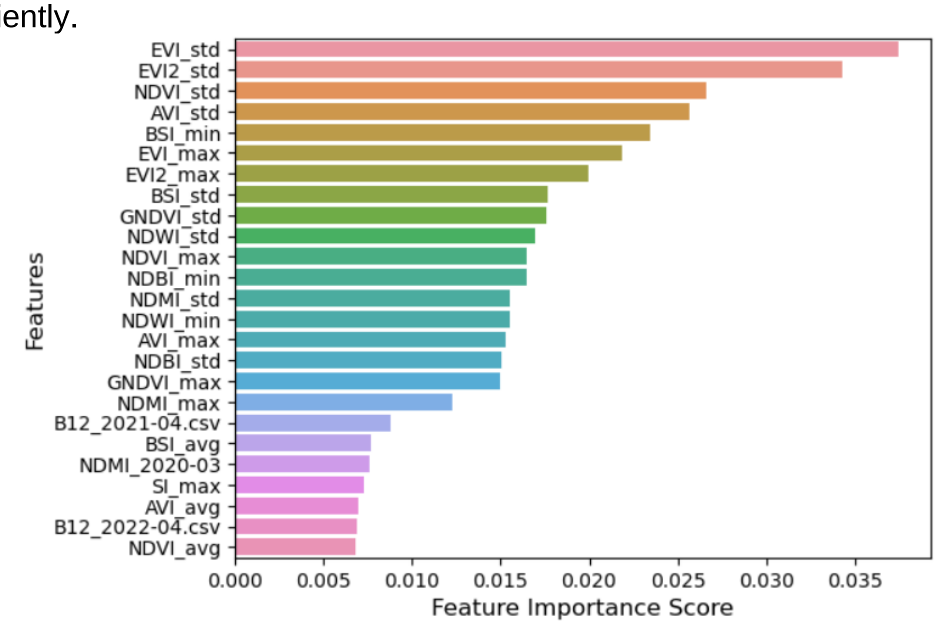

Figure 3 Top 25 Sentinel-2 random forest feature importance

As for Sentinel-1, we found that VH and VV bands in May, June, and July are the most important features for crop detection prediction, as shown in Figure 4. Latitude location also has a moderate impact on model prediction, which is different from the Sentinel-2 dataset in that location information only shares a small impact on model predictions.

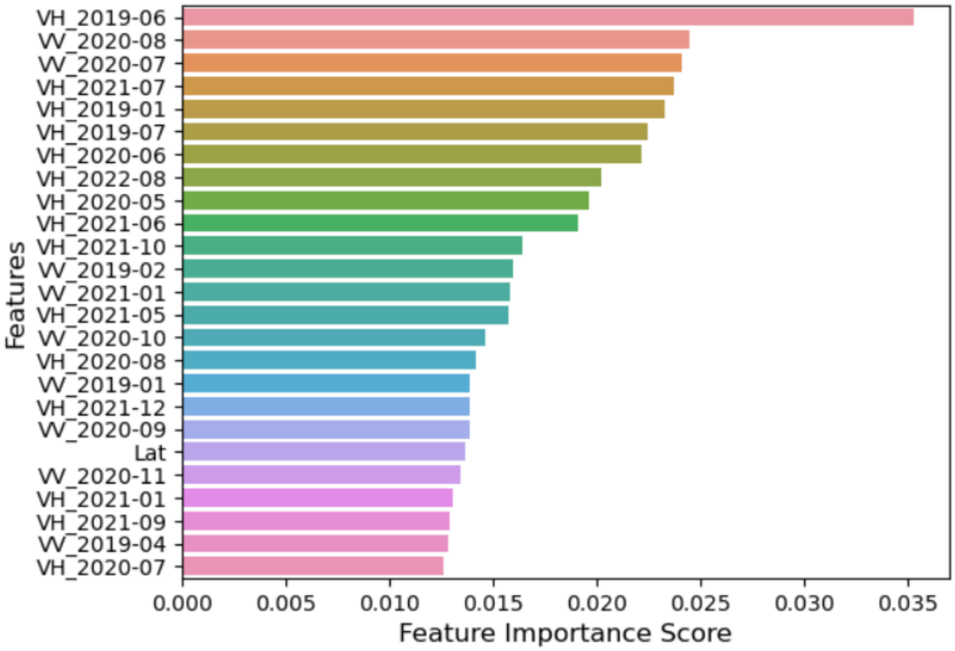

Figure 4 Top 25 Sentinel-1 random forest feature importance

3. Comparing machine learning models
   As shown in Table 3, we also compare the best machine-learning models. We found that stacking RFXGB and RFXGBCatboost is the best crop detection model. We use both models as our final prediction submissions.
4. pseudo-science research founding
   During the research process, we found several undocumented pseudo-sciences that might be useful in future research development, such as :
   - Using a longer timespan for image satellite imagery results in the best performance. We use a short timespan in early development, and the model only performs around 80% accuracy.
   - Using a time series deep learning model (InceptionTime), not increasing performance—the best accuracy is only around 80%. However, changing the data point of view is a potential method to be explored as data engineering gives more performance increases than model fine-tuning.
   - Adding SRTM-DEM slope into the dataset only adds extra noise for model training. Although, in other research, elevation slope might increase model performance.

CONCLUSION

In this mini research, we conclude some points :

1. The sentinel-2 dataset is better satellite imagery (in terms of accuracy) for crop detection with machine learning. Feature engineering, such as data imputation, spectral indices and statistics, is the most important step in whole modelling pipelines. Selecting efficient intervals and more data points also increases model performance.
2. Stacking regressor increases model performance than using an elemental machine learning model.
3. There is more data and model exploration that could increase the model based on our pseudo-science research funding, such as exploring more satellite data, feature engineering and modelling approaches.