Rice Crop Classification

EY Open Science Data Challenge 2023

BUSN 41204 Winter 2023

Hongwon Jang, Ruisi Liu, Yuqi Niu, Matthew Yang

March 10, 2023

1 Introduction

Food insecurity is one of the most pressing issues the world faces today. Nearly 350 million people around the world experience acute food insecurity, which is ten times more than it was five years ago. [1] Climate change has played a large role in this increase in food insecurity, with changing climates decreasing crop yields and making certain agricultural land incapable of growing crops. Thus, improving agricultural efficiency is necessary to combat the world's worsening food insecurity problem. A critical step in this effort is to identify land that is capable of producing high crop yields. Satellite data has proven instrumental in classifying land types and monitoring land degradation at a large scale.

As part of the EY Open Science Data Challenge, [2] our team built a classification model to identify plots of land where rice is grown. Though this classification task is somewhat narrow in scope, modeling techniques discussed in this report can be extended to answer important and practical questions. For example, we can apply the machine learning techniques used in this paper to satellite data to predict/forecast crop yields, so that policy-makers and farmers can ensure that sufficient rice is being grown to feed the population. In this way, machine learning techniques will be critical to reduce food insecurity as farmers adapt to climate change effects.

This goal of our analysis (as part of the EY challenge) was to build a classification model that accurately classifies land (given it's satellite imagery) as a rice field or not. Our model was scored by EY based on a test data set for which we did not have access to labels. Our best classification model achieved a test set accuracy of 0.99, which ranked tied for 12th among all challenge participants as of the date of this report. The remainder of this report details the analysis that we performed.

2 Data

2.1 Background on Satellite Data

The emergence and accessibility of satellite data has played a crucial role in enabling researchers to gain a deeper understanding of agricultural trends, including fluctuations in land productivity, soil degradation, and water consumption. By using remote sensing techniques, researchers can gather information on crop health, soil moisture levels, and land-use patterns, which can be used to identify areas that are at risk of experiencing declines in productivity. This information can be used to develop strategies to improve agricultural practices, reduce water usage, and prevent soil degradation, ultimately leading to more sustainable and productive agricultural systems. By leveraging satellite data, researchers and farmers alike can make more informed decisions that help support the long-term health of our agricultural systems and the planet as a whole.

[1] https://www.wfp.org/global-hunger-crisis

[2] https://challenge.ey.com/challenges/level-1-crop-identification-global/overview

The two main types of data used in agricultural research are radar and optical data. Radar satellite data uses microwave radiation to detect and measure objects on the Earth's surface, while optical satellite data captures images of the Earth's surface using visible and infrared light. One key difference between radar and optical data is that radar technology can penetrate clouds and thus can collect useful data at all times of the day, whereas optical data is only useful when land can be seen from space (i.e., when there is limited cloud cover). Nonetheless, both radar data can be useful in land type classification.[3]

In the next subsection, we describe the types of satellite data used in our analysis.

2.2 Data Collection and Preprocessing

EY provided challenge participants with two datasets. The first dataset included 600 locations (described with longitude/latitude coordinates) and a label denoting whether the location was a rice field or not. Of the 600 locations in the training dataset, 300 were rice fields and 300 were non-rice fields. Therefore, we would expect a random guess to be accurate 50 percent of the time. The second dataset was a test dataset that contained 250 locations, but no labels. After submitting a file to the challenge website with classification predictions for each of the 250 test locations, EY provided the achieved accuracy on the test set. All of the locations in the training and testing datasets were observed in 2020 from a certain region in Vietnam.

Before building a classification model, our team had to collect satellite time series data on each of the locations in the training and testing datasets. That is, for each of the locations, we collected all of the available satellite data for the calendar year 2020. For our analysis, we relied on the following three satellite data sources, all of which were accessed through the publicly available Microsoft Planetary Computer API. [4]

1. **Sentinel-1**
   Sentinel-1 is a satellite mission that collects radar data. It was developed by the European Space Agency (ESA) as part of the European Union's Copernicus program. [5] Initially launched in 2014, the Sentinel-1 mission consists of two satellites (Sentinel-1A and Sentinel-1B) that collect radar signals every few days at discrete time points.[6] Furthermore, the Sentinel-1 satellites transmit and receives both vertical and horizontal signals ("bands"), which are referred to as the "vv" and "vh" bands. [7] Analyzing the "vv" and "vh" bands in conjunction provides a more complete understanding of the land surface than just analyzing a single band alone.

   The Sentinel-1 mission collected the vv and vh signals for approximately 91 time points throughout 2020 for each of the locations in the data.[8] As we describe in the next section on feature engineering, we utilized the entire 2020 time series for each location to compute informative features to be used in our models.

2. **Sentinel-2**
   Sentinel-2 was launched as part of the European Commission's Copernicus program on June 23, 2015. Sentinel-2 collects optical data. [9] It comprises two satellites placed in the same orbit but phased at 180°. [10] Moreover, Sentinel- 2 carries an optical instrument that collects image data on 13 spectral bands (e.g, blue, green, red, NIR, etc.). [11] NIR refers to Near Infra-Red Band, and blue, green, red are just the natural color bands. They are important visible light

[3] For example, Moumni and Lahrouni (2021) used both radar and optical data for land classification models. See: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8081626/

[4] https://planetarycomputer.microsoft.com

[5] https://sentinel.esa.int/web/sentinel/missions/sentinel-1; https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar/product-overview/polarimetry

[6] https://sentinel.esa.int/web/sentinel/missions/sentinel-1/overview/mission-summary

[7] https://ceos.org/document_management/SEO/DataCube/Laymans_SAR_Interpretation_Guide_2.0.pdf. "VV" refers to "vertical transmission vertical reception" and "VH" refers to "vertical transmission horizontal reception."

[8] Some locations had a few observations more than 91.

[9] https://www.usgs.gov/centers/eros/science/usgs-eros-archive-sentinel-2

[10] https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview

[11] https://hatarilabs.com/ih-en/how-many-spectral-bands-have-the-sentinel-2-images

to support vegetation, land cover, and environmental monitoring.

The Sentinel-2 mission collects blue, green, red, and NIR bands for about 73 time points in 2020 for each of the locations in the data. One issue of optical data is that it cannot penetrate clouds. So if the region is covered with clouds at a certain time, the data is unusable. Hence, we need to filter the data to remove clouds. After applying such cloud filtering, which we describe further below, we used the entire 2020 data for each location to generate features related to the presence of rice crops.

Figure 1 below shows an example of an image taken by Sentinel-2 for a rice-field and non-rice field in the data.

(a) Rice Field (b) Non-Rice Field

Figure 1: Sentinel-2 Optical Data

3. **Landsat**
   The Landsat Program, is a series of Earth-observing satellite missions that collect optical data. The program is managed by NASA and the U.S. Geological Survey. Landsat 1 was launched on July 23, 1972. Since then, the mission has collected more than 8 million images of Earth and additional Landsat satellites have launched to the space. The currently operational satellites are Landsat 8 and 9,[12] which are also the data source of our project.

   Landsat 8 was launched in 2013 and carries two sensors, the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS) instruments. OLI captures signals with higher precision, allowing improved identification of land cover conditions. [13]Landsat 9 was later launched in 2021. With higher radiometric resolution, it allows to detect more subtle differences. With more accurate surface temperature measurements, it plays a greater role in tracking land use and change. [14] They record NIR, blue, green, and red light for approximately 45 time slices in 2020. We use entire optical data in 2020 to identify the distribution of major crops after removing cloud effect.

The following two subsections describe data preprocessing steps.

2.2.1 Bounding Box

Instead of simply extracting the radar and optical data at the exact latitude/longitude coordinates provided in the data, we created a rectangular bounding box centered at the provided coordinates and extracted the average value of the radar and optical data within the box.

[12] https://www.usgs.gov/faqs/what-landsat-satellite-program-and-why-it-important
[13] https://www.usgs.gov/landsat-missions/landsat-8
[14] https://www.usgs.gov/landsat-missions/landsat-9

In the case of the Sentinel-1 data, creating a bounding box of 5x5 pixels and averaging the "vv" and "vh" data within that area is better than taking the "vv" and "vh" values at a single point because it provides a more representative measure of the properties of the land cover within that area. In essence, data obtained from a single point/pixel may be noisy, so averaging the data over a nearby region will reduce variance. [15] This is a popular approach in image processing. [16]

We also apply the bounding box approach when pulling data from Sentinel-2 and Landsat.

2.2.2 Cloud Filtering

As mentioned above, one of the major drawbacks of optical data is that it relies on the visible and near-infrared spectrum of light, which cannot penetrate clouds or other forms of atmospheric interference. This means that if there are clouds in the area being observed, the optical data collected may be incomplete or inaccurate. This limitation can hinder the ability of optical data to provide a complete picture of a given area, particularly in regions with frequent cloud cover or other forms of atmospheric interference.

It is important to be able to distinguish cloud cover in optical data from the ground. For instance, if we did not attempt to remove cloudy observations from the data, we might think that both a rice and non-rice field look white, because both locations will have some cloudy observations. Therefore, it is critical to filter out the cloudy observations before using the optical data in our classification model. We use "SCL" band and "QA" band to set the threshold and assess each pixel for the Sentinel-2 and Landsat respectively. Any pixel with a brightness value above the threshold is classified as a cloud and will be removed, while any pixel below the threshold is classified as clear sky. Fir some observations, after performing this adjustment, all pixels in the bounding box were classified as a cloud (and changed to NA), rendering the observation useless. Figure 2 below shows an example of a rice field with cloud cover.[17]

Figure 2: Sentinel-2 Optical Data with Cloud

3 Feature Engineering

There are many different vegetation indices that can be computed from radar and optical data. Such indices can be extremely informative and serve as important features in land classification models. Below, we describe the indices we computed from the data that we use in our models. Additionally, we describe various feature engineering techniques that we applied to the indices in order to better separate the classes (rice and non-rice fields).

[15] https://www.mdpi.com/2072-4292/11/10/1184

[16] https://nanonets.com/blog/image-processing-and-bounding-boxes-for-ocr/

[17] Note that the image shown is pulled using a much larger bounding box than we used to extract the data used in our model for purposes of visualization.

3.1 Radar Vegetation Index

The Radar Vegetation Index (RVI) is commonly used to measure the health and density of vegetation in a given area using radar data. RVI can be used to distinguish between types of land cover because radar waves interact differently with vegetation than with other types of land cover, such as soil or water. Specifically, radar waves are absorbed more strongly by dense vegetation than by bare soil or open water. [18] Therefore, higher RVI values indicate denser and healthier vegetation, while lower values indicate less dense or less healthy vegetation.

The RVI is a function of the VV and VH bands that can be extracted from the Sentinel-1 radar data. Below is one common formula for RVI, though there are many variations of the formula that researchers use. In the formula, the subscript it denotes the observed value for location i at time t.

$$\text{RVI}_{it} = 4 * \sqrt{\frac{VV_{it}}{VV_{it} + VH_{it}}} * \frac{VH_{it}}{VV_{it} + VH_{it}}$$

We analyze the RVI at all available time slices across 2020 for each of the locations in the training data to understand the differences between RVI fluctuations in rice and non-rice fields. The plot on the left of Figure 3 below shows the RVI for each of the 300 rice fields in the training data. As we can see, the data is fairly noisy, so we apply a moving average transformation to each time series (using a 5 time point rolling window) to obtain a smoother series. The moving average RVI values are shown on the right in Figure 3.

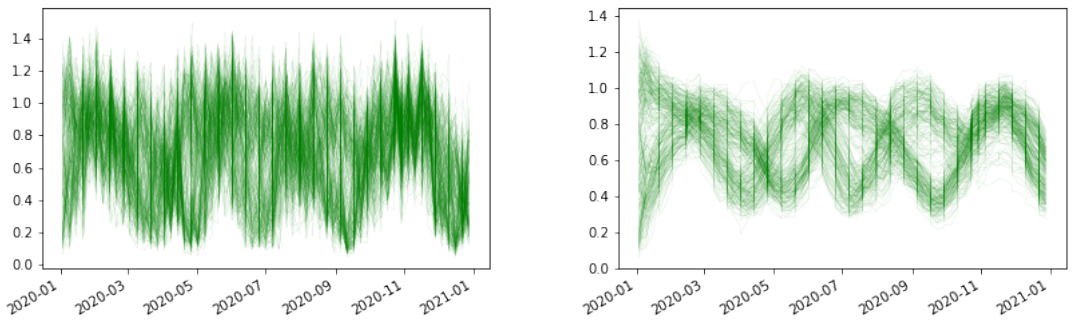


Figure 3: RVI for Rice Crops 2020, with and without Moving Average

We can now compare the RVI time series for the rice and non-rice locations in the training data. Figure 4 below shows the RVI (with and without moving average transformation) for all 600 locations in the training data.

There are several takeaways from Figure 4. First, the RVI time series for non-rice fields is extremely noisy. After applying the rolling average to the non-rice locations, we see that RVI values are relatively flat throughout the year compared to rice fields. Second, after applying the rolling average smoothing adjustment, rice crops have a much more clear and consistent seasonal pattern compared to non-rice crops. In fact, Figure 4 exhibits two distinct seasonal patterns for rice crops. These two patterns correspond to double and triple cropping schedules.[19] Third, the rice fields systematically have lower minimum RVI values compared to non-rice fields. Additionally, the rice crop RVI time series' tend to have larger variance compared to non-rice crops.

Therefore, we create features for simple statistics describing each RVI time series, including minimum, maximum and variance. We also create features that capture elements (such as periodicity) of the time series patterns, as we describe in the next subsection on Fast Fourier Transforms.

[18] `https://www.mdpi.com/2072-4292/10/11/1776`

[19] Double cropping involves planting and harvesting rice twice in a year, whereas triple cropping involves planting and harvesting three rice crops in a single year.

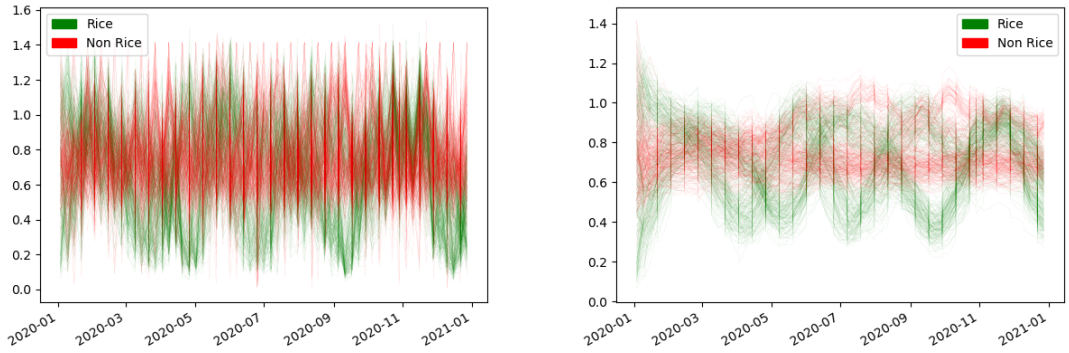


Figure 4: RVI for Rice and Non Rice Fields 2020, with and without Moving Average

3.1.1 Fast Fourier Transform

Fast Fourier Transform (FFT) is a mathematical algorithm that can transform a time series into its corresponding "frequency-domain" representation.[20] That is, FFT decomposes a time series from a waveform that varies over time to a set of frequencies and amplitudes that describe the same signal but in a different way.

We can use FFT to decompose the RVI time series for each location in the data and extract relevant features. For example, we can calculate the power spectral density (PSD) to see how much power is present at different frequencies, which can help us identify periodic patterns or oscillations in the data.[21] We can also identify the dominant frequency, which can tell us the frequency that occurs most frequently or is most prominent in the time series.[22]

Overall, FFT is a powerful tool that can summarize elements of a time series into simple representations. Figure 5 shows the dominant frequency and PSD values for each of the RVI time series in the training data. As shown, these two features, can help distinguish the two classes. We see that the dominant frequency is relatively similar across all rice fields, but varies for non-rice fields. This is consistent with our observation in Figure 4 that rice fields exhibit a consistent seasonal pattern whereas non-rice fields do not. Additionally, the RVI time series for rice fields tend to have higher PSD values because of the stronger seasonal pattern.

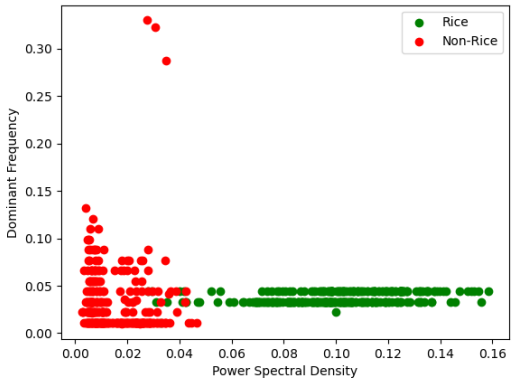
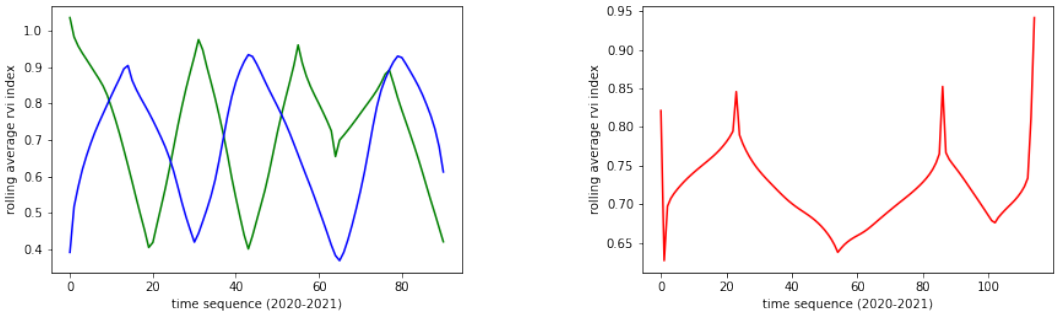


Figure 5: FFT Features for RVI Time Series

[20] https://www.princeton.edu/~cuff/ele201/kulkarni_text/frequency.pdf
[21] https://scicoding.com/calculating-power-spectral-density-in-python/
[22] https://eplabworks.com/dominant-frequency/

3.1.2 Time Series K Means Clustering

K means clustering is an unsupervised learning technique that clusters data points based on their similarities, which can be measured by Euclidean distance or other quantitative methods. It can also be used in comparing the similarities between different time series using dynamic time warping metrics.[23] As we show in Figure 3 and Figure 4, we find that there are two slightly shifted time series patterns in RVI values for rice fields, while there is no clear trend in RVI values for non-rice fields. In general, the RVI values for non-rice fields are relatively flat throughout the year (compared to rice fields). Since there are some variations in their quantities, the clustering method can provide a more stable measure of the similarities between different classes of lands. We apply K means clustering to the RVI values for rice and non-rice fields to obtain 2 clusters for rice fields and one cluster for non-rice fields. These clusters are shown in Figure 6. After calculating these clusters, we calculate a similarity score to each of the three clusters for each location and use these scores as features in our models. Specifically, there are two features: similarity to rice cluster[24], and similarity to non-rice cluster. In the next subsection, we describe the soft dynamic time warping method for calculating similarity.



(a) Rice Field (b) Non-Rice Field

Figure 6: RVI Time Series Clustering

It is worth noting that we used K Means here rather than a supervised method since we cluster based on the classes of lands. We note that this could potentially bias our feature selection process because the clusters were computed on the same training data that the similarity features are computed on. Ideally, we could perform the clustering on data that is independent of the training data, however, we do not have enough labeled data to implement this.

3.1.3 Soft Dynamic Time Warping

Unlike the Euclidean distance, Dynamic Time Warping (DTW) can compare time series of variable size and is robust to shifts or dilatations across the time dimension. Thus, it's suitable to use for this context since the RVI variations could shift based on different growing schedules (i.e., double vs triple cropping). It can compute the best possible alignment between two time series, even with different length. As shown in Figure 7 below, we see that the soft dynamic time warping scores can be a strong indicator of the classes of the lands, meaning the RVI time series' for rice and non-rice fields have distinct behaviors.

[23] `https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3`

[24] Since there are two clusters for rice-fields, we define the feature as the maximum similarity score to the two clusters

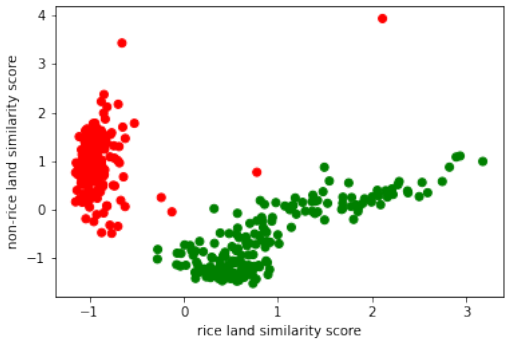


Figure 7: Time Series Clustering Classification

3.2 Optical Band Indices

Optical data (e.g., Landsat or Sentinel-2) contains many spectral bands (e.g., Red, Green, Blue, Red-Edge, NIR, SWIR, etc.) that can be related to the presence or growth of rice crops. Researchers often use statistical combinations of these bands, called indices. Some of the common indices for agriculture are the Normalized Difference Vegetation Index (NDVI), including Enhanced Vegetation Index (EVI), and Soil Adjusted Vegetation Index (SAVI).

As we noticed through data exploration, despite both collecting optical data, the Landsat and Sentinel-2 data have some slight discrepancies. Previous researchers have also observed such differences.[25]. Therefore, in order to be comprehensive, we compute the vegetation indices using both Sentinel-2 and Landsat.

3.2.1 Normalized Difference Vegetation Index

Normalized Difference Vegetation Index (NDVI) is a widely used vegetation index to measure the plant growth and vegetation cover captured in a satellite image (i.e., optical data). [26] It ranges from -1 to 1. Higher NDVI value corresponds to dense vegetation such as tropical forests. Negative NDVI almost always corresponds to water, cloud and snow. Different crops tend to have different NDVI values, so NDVI can be useful in crop classification.

The NDVI uses NIR (which plants intensively reflect) and Red (which plants absorb) channels which can be extracted from sentinel 2 data. The formula is as follows:

$$\text{NDVI} = \frac{NIR - Red}{NIR + Red}$$

3.2.2 Enhanced Vegetation Index

The Enhanced Vegetation Index (EVI) is similar to NDVI used to quantify vegetation greenness. However, EVI incorporates some atmospheric conditions and canopy background noise, and it is more sensitive in high biomass areas, so it will improve upon quality of NDVI production. [27]

The EVI is a function of NIR, Red and Blue bands. Below is one common formula for the EVI.[28] C1 and C2 are coefficients of the aerosol resistance term. G is gain factor. L is the canopy

[25] https://www.sciencedirect.com/science/article/pii/S0034425718305212
[26] https://eos.com/make-an-analysis/ndvi/
[27] https://www.usgs.gov/landsat-missions/landsat-enhanced-vegetation-index
[28] For Landsat data, the equation becomes

$$2.5 \times \frac{NIR - Red}{NIR + 6 \times Red - 7.5 \times Blue + 1}$$

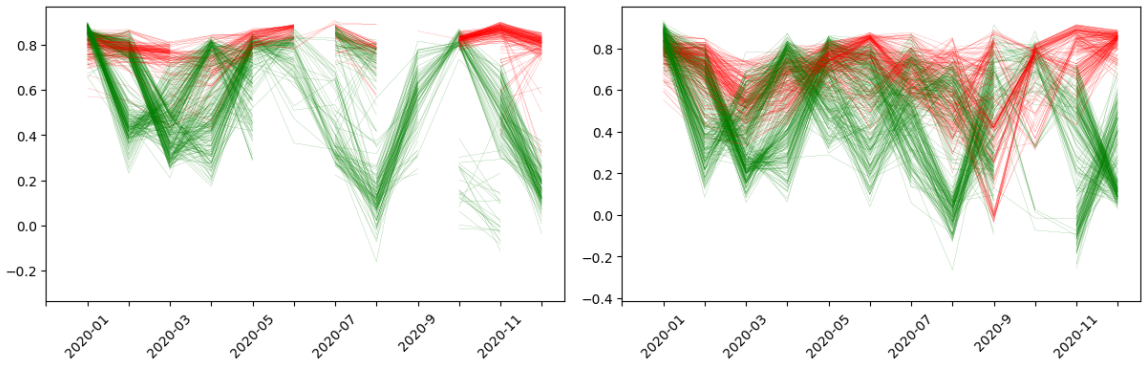


Figure 8: NDVI for Rice and Non Rice Fields 2020, from Landsat (Left) and Sentinel-2 (Right)

background adjustment.

$$\text{EVI} = G \times \frac{NIR - Red}{NIR + C1 \times Red - C2 \times Blue + L}$$

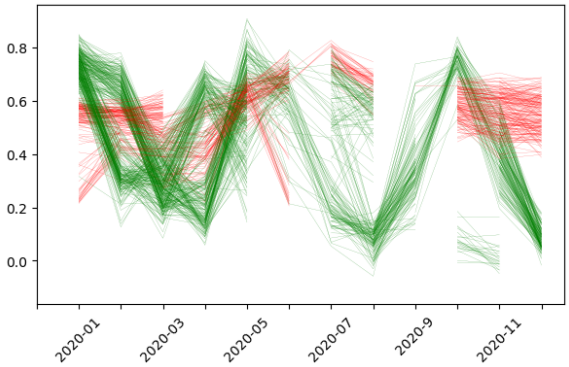


Figure 9: EVI for Rice and Non Rice Fields 2020, from Landsat

Due to a scaling issue, we have decided to use Landsat data for the EVI index. We analyze the EVI for each location from January to May in 2020 to analyze crop growth pattern and detect land use.

3.2.3 Soil Adjusted Vegetation Index

The Landsat Soil Adjusted Vegetation Index (SAVI) is used to correct NDVI to minimize the soil brightness influence when vegetation cover is low. It is a reliable indicator at the early stage of crop. When crop cover is poor, the soil effect affects the accuracy of NDVI values. By using correction factor in the computation of the SAVI, it allows to ignore the soil reflectivity and focus on plant growth only.[29]

SAVI is also calculated as a ratio between NIR and Red values. L refers to a soil brightness correction factor which ranges from 0 for very high vegetation cover to 1 for very low vegetation cover.[30]

$$\text{SAVI} = \frac{NIR - Red}{NIR + Red + L} \times (1 + L)$$

[29] `https://support.geoagro.com/en/kb/savi-index/`

[30] $L = 0.5$ for Landsat data, and $L = 0.428$ for Sentinel-2 data

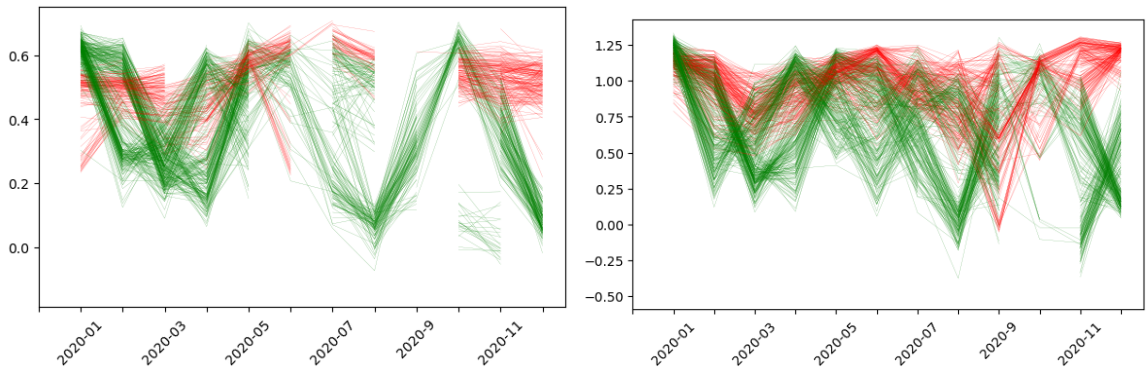


Figure 10: SAVI for Rice and Non Rice Fields 2020, from Landsat (Left) and Sentinel-2 (Right)

3.2.4 Features from Indices

Unlike radar data, optical data can be affected and distorted by cloud cover. In order to obtain clear optical data for land observation, cloud filtering is necessary. This is achieved using specific spectral bands that can identify areas of clouds or water. For Landsat data, the "QA" band is used for cloud filtering, while for Sentinel-2 data, the "SCL" band is used.

Filtering out clouds from optical data makes data sparse. Landsat data from June to December, in particular, is so sparse and it seems inappropriate to use the entire dataset. Therefore, we would limit the analysis of Landsat data to the first cropping cycle of rice, which usually takes place between January and May when data availability is higher.

Due to the sparsity of optical datasets, obtaining a continuous series of data is impossible. Therefore, rather than using time-series analytical methods, a more appropriate approach is to engineer features from the available optical data that can account for missing values. In this study, we have created four variables for each monthly index: mean, variance, minimum, and maximum. These variables could be useful for classification, as depicted in Figure 11.[31]

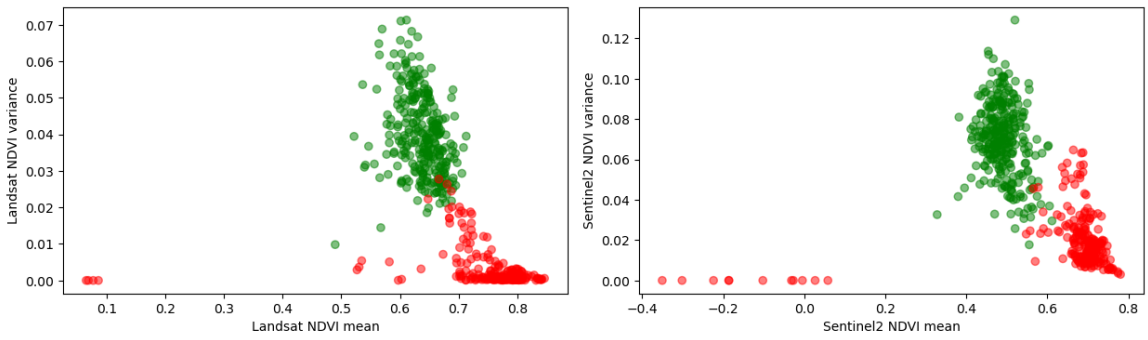


Figure 11: Monthly NDVI Mean and Variance, from Landsat (Left) and Sentinel-2 (Right)

3.3 Variables Obtained

As a result of feature engineering, we have obtained 27 variables to use in the models. To be specific, seven features are engineered from the radar data, and twenty features are engineered from

[31]There are several other potential ways to address this issue, such as extending the temporal scope to gather more data or using other datasets to impute the missing values. However, these approaches often require significant amounts of additional data and computational resources. Given our limited resources, we have endeavored to create the best possible outcome by engineering features from the available optical data. While this approach may have its own limitations, we believe that it provides a reasonable balance between the available data and computational constraints.

the optical data. The features are listed in Table 1.

	Features Engineered
Radar Data	power spectral density, dominant frequency, RVI maximum & minimum & variance, maximum similarity to rice cluster & non rice cluster
Optical Data	monthly NDVI mean & variance & minimum & maximum (Landsat and Sentinel-2), monthly SAVI mean & variance & minimum & maximum (Landsat and Sentinel-2), monthly EVI mean & variance & minimum & maximum (Landsat)

Table 1: Engineered Features from Radar and Optical Data

4 Modeling Methodology

We explored several different classification algorithms for predicting whether a plot of land is a rice field or not. Specifically, we tried logistic regression, XGBoost, and neural networks. We chose these models because they are popular tools in machine learning for solving classification problems.

4.1 Hyper-parameter Tuning

We used 10-fold cross validation on the entire training set to select the best hyper-parameters. Since EY ultimately evaluated our submission based on accuracy, we used the accuracy score as our evaluation metric in the hyper-parameter tuning process. We also used feature selection techniques to select the most important features for each model.

4.2 Variable Selection

Variable selection serves several purposes in data analysis. In some cases, variable selection is important to make a model more interpretable. In other cases, variable selection can help decrease variance in predictions. A further objective of variable selection may be to simplify computational efficiency. For our project, the main goal is prediction accuracy, therefore, we perform variable selection in order to stabilize predictions. Using fewer predictors in a model can result in higher bias but lower variance, ultimately leading to more accurate predictions.

5 Results

In Table 2, we show the test set accuracy achieved by each of the three models. Each of the three models are able to achieve near-perfect accuracy on the test set, as scored by EY. Below, we provide more details on the optimal features and hyperparameters for each model.

Algorithm	Test Set Accuracy
Logistic Regression	0.98
XGBoost	0.98
Neural Network	0.99

Table 2: Model Comparison

5.1 Logistic Regression

We used cross-validation to select the optimal λ value, the Lasso penalization parameter. The lowest cross-validation error is achieved by including all of the radar data features, but none of the optical data features. Furthermore, by using the 1-standard error approach to selecting λ, we obtain a smaller model with only 3 of the radar data features: 1) maximum similarity to rice cluster, 2) similarity to non-rice cluster, and 3) dominant frequency. Therefore, even though each individual feature appears to be predictive on the training set (as shown in the exploratory plots above), including all of the features may be redundant and not improve prediction accuracy.

5.2 XGBoost

We ran the model on all the features and tried various variable selection methods, including Boruta, RFE, and RFA. It is important to perform variable selection in conjunction with other hyperparameter tuning in order to not bias the variable selection and hyperparameter selection. We used the shap-hypetune library[32] to simultaneously conduct hyperparameter tuning and feature selection. When using all the data, our accuracy score on the test set was 0.98. Boruta selected 12 features, RFE selected 5 features, and RFA only selected 2 features. Boruta and RFE yielded the best performance, achieving an accuracy score of 0.98, while RFA achieved 0.93.

5.3 Neural Network

We used all available features on the neural network, with 7 variables from radar data and 20 variables from optical data and implemented maximum absolute scaling to divide every observation by the maximum value of the feature. It seems that the performance of neural networks is quite stable. That's because we add L2 regularization to the neural network and also tune its magnitude to prevent overfitting. Neural network also reached a test accuracy of 0.99.

6 Conclusion

Based on the results of our analysis, we can conclude that logistic regression, XGBoost, and neural networks are all effective approaches for rice crop classification. By leveraging a wide range of feature engineering techniques based on radar and optical data, we were able to achieve near-perfect test set accuracy for all of the models. Our prediction scored the top 15 among about 150 teams that participated.

Our findings suggest that the incorporation of diverse data sources and feature engineering methods can significantly enhance the performance of classification models in crop classification applications. Furthermore, our findings highlight the potential for machine learning models to provide insights into agricultural practices and support decision-making in the field of agriculture. Our hope is that state-of-the-art machine learning techniques continue to be applied to agricultural applications, including crop classification, in order to improve food security and land management.

[32] `https://github.com/cerlymarco/shap-hypetune`