

Team no1 – Methodology

Ling Xiang Zou

Ling Feng Zhou

Fan Ye

This project encompasses data collection, data cleaning (EDA), feature engineering, and modeling to analyze satellite data and generate accurate predictions.

During the data collection phase, we obtained Sentinel-1 and Sentinel-2 data from 2021-11-01 to 2022-12-01, using coordinates with 3x3, 5x5, and 7x7 bounding boxes as well as coordinates without bounding boxes.

In the data cleaning and EDA stage, we identified outliers in the xarray-formatted satellite data, particularly extreme values between 2021-11 and 2022-01. These outliers were replaced with the mean value of the coordinates throughout the year.

For feature engineering, we developed several indices to enhance accuracy, including NDVI (Normalized Difference Vegetation Index), SAVI (Soil-Adjusted Vegetation Index), NDWI (Normalized Difference Water Index), EVI (Enhanced Vegetation Index), GCI (Green Chlorophyll Index), LAI (Leaf Area Index), and RVI (Radar Vegetation Index). We calculated various statistical measures for each index across all coordinates throughout the year.

In the modeling phase, we split the data into training and testing sets, performed anomaly detection, and oversampled the data. We evaluated several models with default hyperparameters, such as random forest, support vector machine, neural network, k-nearest neighbor, decision tree, and gradient boosting. Based on their performance, we selected random forest, tuned the model using a parameter grid, and performed feature selection using ANOVA F-value. After validating the results on the testing set, we trained the optimized model with the full dataset to maximize accuracy. Finally, we predicted the given coordinates and submitted the results for evaluation.