**Crop Classification Using Sentinel-2 and USDA CDL Data with Machine Learning in Google Earth Engine**

**Abstract**

Accurate, large-scale crop classification is crucial for agricultural monitoring and resource management. This study presents an interactive tool in Google Earth Engine (GEE) that combines **Sentinel-2 multispectral imagery** with the **USDA Cropland Data Layer (CDL)** to create detailed crop type maps for U.S. counties. The system uses a dynamic user interface (UI) and supports three machine learning classifiers: **Random Forest, CART, and Support Vector Machines (SVM)**. Users can manually tune each classifier’s hyperparameters through an advanced settings panel or opt for an automated grid search to identify the optimal model. To enhance usability, the UI features a dynamic legend that updates based on the available training data and a results panel that displays performance metrics like **overall accuracy and Cohen's kappa**. This paper details the data sources, preprocessing steps, classification methodology, and UI design, demonstrating a robust, scalable, and user-friendly solution for regional crop mapping. The tool achieves high classification accuracy and highlights the efficacy of combining spectral bands with multiple vegetation indices for improved crop separability.

**1. Introduction**

Monitoring crop types and agricultural land use is essential for optimizing food production, managing water resources, and ensuring environmental sustainability. Satellite remote sensing, when combined with machine learning, offers a powerful method for classifying crops over large areas and across different growing seasons.

This work introduces a comprehensive crop classification tool developed in Google Earth Engine (GEE). It integrates high spatial and temporal resolution multispectral data from the Sentinel-2 satellite with the USDA CDL, a reliable source of annual crop labels for supervised learning. The tool supports three popular classifiers—Random Forest, CART, and SVM—and includes both manual hyperparameter tuning and an automated grid search to find optimal model parameters.

A dynamic and intuitive user interface facilitates state and county selection, year input, and classifier configuration. The interface also provides dynamic legends that adapt to the classes present in the training data and a results panel that displays accuracy metrics and allows the export of classified maps. This combination of robust data sources, advanced classification techniques, and a responsive UI provides an accessible and scalable tool for county-level crop mapping.

**2. Data and Preprocessing**

**2.1 USDA Cropland Data Layer (CDL)**

The USDA Cropland Data Layer (CDL) is an annual raster dataset providing high-resolution land cover classification across the continental United States, with a strong focus on crop types. The CDL contains detailed crop-specific classes and non-cultivated land categories, with data available annually since 2008.

For this tool, we use CDL data from 2008 through 2024, focusing on a filtered set of 29 crop classes, including major commodities like corn, soybeans, and wheat, as well as specialty crops. These CDL rasters serve as the ground truth labels for our supervised classification models.

**2.2 Sentinel-2 Imagery and Preprocessing**

Sentinel-2, a satellite mission from the European Space Agency, provides multispectral imagery at a spatial resolution of 10 to 20 meters, which is ideal for agricultural monitoring due to its frequent revisit time.

* **Data Filtering**: We filter Sentinel-2 Surface Reflectance (COPERNICUS/S2\_SR\_HARMONIZED) imagery for the period from April 1 to October 31 of the selected year, covering the typical growing season in the U.S. The imagery is spatially clipped to the selected county and only scenes with less than 20% cloud cover are retained.
* **Cloud Masking**: To ensure data quality, we apply a cloud mask using the Sentinel-2 QA60 band, which identifies and removes pixels flagged as clouds or cirrus.
* **Composite Creation**: The filtered and masked images are combined using a median composite, which helps to minimize temporal variability and noise, resulting in a single representative image for the area of interest (AOI) for the entire growing season.

**2.3 Spectral Bands and Vegetation Indices**

Our classification approach leverages seven raw Sentinel-2 spectral bands: **B2 (Blue), B3 (Green), B4 (Red), B5 (Red Edge 1), B6 (Red Edge 2), B8 (Near Infrared), and B11 (Short-Wave Infrared)**. These bands are chosen for their sensitivity to vegetation characteristics and soil properties.

To improve the separability of different crop types, we also compute and add five vegetation indices as additional bands to the feature space:

* **NDVI (Normalized Difference Vegetation Index)**: A classic index for capturing vegetation greenness and biomass.
* **EVI (Enhanced Vegetation Index)**: Designed to improve sensitivity in high-biomass regions while reducing atmospheric and soil background influences.
* **GNDVI (Green Normalized Difference Vegetation Index)**: An index that is particularly sensitive to chlorophyll concentration.
* **NDWI (Normalized Difference Water Index)**: Used to indicate vegetation water content.
* **SAVI (Soil Adjusted Vegetation Index)**: An index that mitigates soil brightness effects in areas with sparse vegetation.

**3. Methodology**

**3.1 Area of Interest and Sampling**

The tool allows users to select an AOI by first choosing a U.S. state, which then dynamically populates a list of available counties. The AOI is defined by the geometry of the chosen county.

To create a balanced and representative training dataset, we perform **stratified sampling** on the CDL crop labels within the AOI. This method ensures that a proportional number of samples (a minimum of 30 and a maximum of 500 per class) are drawn from each crop type present in the AOI, even for less common crops. The samples are then randomly split into a training subset (70%) and a testing subset (30%) for model evaluation.

**3.2 Machine Learning Classifiers**

We implemented three widely used supervised machine learning classifiers from the GEE API:

* **Random Forest (RF)**: An ensemble of decision trees known for its robustness to overfitting. Its key hyperparameters are the number of trees and the maximum number of nodes per tree.
* **Classification and Regression Tree (CART)**: A single decision tree classifier. Its complexity can be controlled by setting the minimum leaf population, which prunes the tree and prevents overfitting.
* **Support Vector Machine (SVM)**: A powerful classifier that uses a Radial Basis Function (RBF) kernel to create a decision boundary. Its performance is tuned using the **gamma** parameter (kernel width) and the **cost** parameter (regularization penalty for misclassifications).

**3.3 Hyperparameter Tuning and Grid Search**

The UI’s advanced settings panel exposes hyperparameter input fields that change dynamically based on the selected classifier. This allows users to manually fine-tune model parameters.

For automated optimization, the "Compare All Classifiers" checkbox initiates a **grid search**. This process systematically tests a predefined range of hyperparameter combinations for all three classifiers, calculates the accuracy and Cohen's kappa for each, and identifies the best-performing model. The grid search runs sequentially, with progress updates provided in the UI. Upon completion, the best model is automatically retrained and used to classify the entire AOI.

**4. User Interface and Dynamic Features**

Our tool's UI is designed to be intuitive and responsive.

* **Interactive Region and Year Selection**: The cascading dropdowns for state and county and the year input textbox (2008–2024) streamline the AOI and temporal selection process.
* **Dynamic Settings Panel**: The advanced settings panel adapts in real time to show only the hyperparameters relevant to the chosen classifier. A key feature is the **"?" toggle button** next to each parameter, which displays a helpful hint to explain its purpose and influence on the model. When the grid search is selected, this panel and the individual classifier dropdown are disabled to prevent conflicting inputs.
* **Dynamic Loading Label**: A **loading label** dynamically provides real-time feedback to the user, displaying status messages for each step of the process, such as "Loading CDL," "Training Random Forest," and "Classification completed," with color-coded success or error indicators. This ensures the user is informed about the progress of computationally intensive tasks.
* **Dynamic Legend Panel**: The legend automatically updates to display only the crop classes found in the training data for the selected county, reducing visual clutter and improving interpretation.
* **Results Panel and Metrics Display**: A dedicated panel presents key performance metrics, including sample counts, the training/testing split sizes, overall accuracy, and Cohen's kappa, along with the specific model parameters used.
* **Export Functionality**: A "Download Classified Map" button allows users to export the final classified image as a GeoTIFF file, enabling offline use and integration with other GIS software.

**5. Results and Discussion**

The tool enables flexible crop classification workflows, from manual parameter tuning to automated grid searches. The inclusion of five vegetation indices alongside the raw spectral bands proved highly effective at distinguishing various crop types, significantly enhancing model performance compared to using spectral bands alone.

The use of CDL as a ground truth dataset, while an excellent resource, presents a known limitation: the accuracy of the CDL itself places an upper bound on the achievable accuracy of our model. However, the use of stratified sampling and the robust validation with a separate test set provide a reliable estimate of the tool’s performance. The grid search feature proved valuable for objectively identifying optimal classifier configurations without requiring manual experimentation.

Future work will focus on integrating additional data sources, such as Sentinel-1 SAR, to improve performance in cloudy regions. We also plan to explore temporal change analysis for monitoring crop health and rotation over multiple years and to expand our export options to include accuracy metrics.

**6. Conclusion**

This study demonstrates a robust and user-friendly crop classification tool built on the Google Earth Engine platform. By combining high-quality Sentinel-2 and CDL data with advanced machine learning, dynamic UI elements, and a powerful grid search feature, the tool provides a highly effective solution for accurate crop mapping at a county scale, supporting critical agricultural monitoring and management efforts.