Time-Series Analysis of Crop Types in Tanzania Using Crowdsourced Data and Sentinel-2 Imagery

Michael L. Mann* Lisa Colson† Rory Nealon‡
Stellamaris Wavamunno Nakacwa

Abstract

This study introduces a robust methodology for crop type classification in Tanzania, utilizing a novel integration of time-series feature from Sentinel-2 satellite data and crowdsourced ground truth data collected by the YouthMappers network. By combining advanced remote sensing techniques with extensive local knowledge, the research addresses significant gaps in agricultural monitoring within resource-limited settings. The application of machine learning algorithms to analyze temporal and spectral data enables the precise identification of crop types, showcasing the enhanced accuracy and utility of combining technological and human resources. We achieve 0.80 kappa accuracy scores, across a diverse multi-class dataset including challenging crops including cassava, millet, sorghum, and cotton amongst other. This methodological innovation not only improves crop classification accuracy but also contributes to sustainable agricultural practices and policy-making in developing countries, making a significant impact on food security and land management.

*The George Washington University, Washington DC 20052, Corresponding author. Email mmann1123@gmail.com
†USDA Foreign Agricultural Service, Washington DC 20250
‡USAID GeoCenter, Washington DC 20523

(**stella?**) add you details above

Introduction

Background and Context

The free access to remotely sensed data, such as imagery from satellites (e.g. Sentinel-2, LandSat) has revolutionized the field of crop type classification in developing countries. By leveraging the power of advanced imaging technologies combined with machine learning algorithms, researchers and practitioners can now identify and map different crop types over large geographic areas at low cost. This has the potential to improve food security, land use planning, and agricultural policy in regions where ground-based data collection is limited or non-existent.

In recent years, machine learning approaches have emerged as powerful tools for crop type classification using remotely sensed data. Specifically, methods based on machine learning algorithms have gained recognition for their effectiveness in matching valuable spectral information from satellite imagery to observations of crop type for particular locations. Machine learning algorithms, including decision trees, random forests, support vector machines (SVM), and k-nearest neighbors (KNN), have been successfully used to classify imagery into unique agricultural types. These algorithms leverage the rich spectral information captured by satellite sensors, allowing them to identify distinctive patterns associated with different crop types. By training on large labeled datasets where ground-validation information on crop types is linked to corresponding image patches, these models can effectively learn the relationships between the spectral characteristics of crops and their respective classes.

The strength of traditional machine learning approaches lies in their ability to exploit both the spectral patterns within the remotely sensed data. For instance, decision tree-based algorithms partition the feature space based on the spectral bands, enabling the identification of different crop types based on their unique spectral signatures. Random forests extend this concept by combining multiple decision trees to improve classification accuracy and handle more complex scenarios.

These traditional machine learning approaches offer advantages in terms of interpretability and computational efficiency compared to deep learning architectures. They provide insight into the decision-making process and can be more readily understood and explained by domain experts. Additionally, these methods are generally less computationally demanding and require less training data, making them suitable for applications with limited computational resources.

The development of salient features on a pixel-by-pixel basis from remotely sensed images remains a challenge. Traditional machine learning algorithms require the extraction of relevant features from the raw data to effectively classify crop types. These features are typically derived from the spectral bands of the satellite imagery, such as the enhanced vegetation index (EVI) and basic time series stastics (e.g. mean, max, minimum, slope) for the growing season. Meanwhile a broader set of time series statistics may be more relevant for a number of applications. For instance the skewedness of EVI might help distinguish crops that greenup early vs later in the season, measures of the numbers of peaks in EVI might help differentiate intercropping or multiple plantings in a season. However, the selection and extraction of these features can be time-consuming and labor-intensive, requiring domain expertise and manual intervention.

Field-collected data provides the necessary validation and calibration for remote sensing-based models. It serves as the benchmark against which the model's predictions are evaluated and refined. Ground truth data collected through field visits, observation, and interactions with local farmers offer essential insights into the specific crop types present in the study area. Validating and training models with accurate ground reference information ensures that the spectral patterns captured by remote sensing data are correctly associated with the corresponding crop classes. By combining the spectral information from satellite imagery with ground truth data, researchers can develop robust models that effectively differentiate between different crop types based on their unique spectral signatures.

The collection of field observations and ground truth data is a critical input to the development of machine learning models for crop type classification. However, obtaining accurate and timely ground truth data

can be challenging in developing countries due to limited resources, infrastructure, and capacity. In many cases, researchers rely on crowdsourced data from volunteers or citizen scientists to supplement or validate ground truth data collected through traditional methods. Projects like (Tseng et al. 2021) point to the near complete lack of multi-class crop type datasets globally. This is a significant gap in the field of crop type classification, as the availability of high-quality training data is essential for the development of accurate and reliable machine learning models.

In this study we aim to address two critical challenges in the field of crop type classification: the lack of multi-class crop type datasets and the need for automated methods of developing salient time-series features for agricultural applications.

We propose a novel approach that combines crowdsourced data with time series features extracted from satellite imagery to classify crop types in Tanzania. By leveraging the power of crowdsourcing and remote sensing technologies, we aim to develop a robust and scalable solution for crop type classification that can be applied in other regions and contexts.

Data & Methods

Data for this study were collected from multiple sources, including satellite imagery, and crowdsourced ground truth observations.

Study Area

(**anyone?**) _____ revise this _____ The study area is located in Tanzania, a country in East Africa known for its diverse agricultural landscape. We focus on the 15 northern most ?states? including Arusha, Dodoma, Geita, Kagera, Katavi, Kigoma, Kilimanjaro, Manyara, Mara, Mwanza, Shinyanga, Simiyu, Singida, Tabora and Tanga. The region is characterized by a mix of smallholder farms, commercial plantations, and natural vegetation, making it an ideal location for studying crop type classification. ???

(**rory?**) or stella - Can you create a *pretty* map of the study area? the bounds file is northern_TZ_states.geojson

Crowd Sourced Crop Data

(**anyone?**) _____describe YM data collection_____

(keep this ->) The land cover classes included in the analysis were rice, maize, cassava, sunflower, sorghum, urban, forest, shrub, tidal, cotton, water, and millet. Classes excluded due to irrelevance or insufficient data included 'Don't know', 'Other (specify later)', 'water body', 'large building', and several others which were either ambiguous or represented a negligible fraction of the data.

Additional training data was collected utilizing high resolution imagery from Google Earth. This data was used to supplement the crowdsourced data and improve the model's ability to distinguish between crops and more common land cover types like forests, urban areas, and water.

Satellite Imagery

Satellite imagery was obtained from the Sentinel-2 satellite constellation, which provides high-resolution multispectral data at 10-meter spatial resolution. The imagery was acquired over the study area during the growing season, capturing the spectral characteristics of different crop types. The Sentinel-2 data were pre-processed to remove noise and atmospheric effects, ensuring that the spectral information was accurate and reliable for classification purposes.

In our study, cloud and cloud shadow contamination was mitigated using the 's2cloudless' machine learning model on the Google Earth Engine platform. Cloudy pixels were identified using a cloud probability mask,

with pixels having a probability above 50% classified as clouds. To detect cloud shadows, we used the Near-Infrared (NIR) spectrum to flag dark pixels not identified as water as potential shadow pixels. The projected shadows from the clouds were identified using a directional distance transform based on the solar azimuth angle from the image metadata. A combined cloud and shadow mask was refined through morphological dilation, creating a buffer zone to ensure comprehensive coverage. This mask was applied to the Sentinel-2 surface reflectance data to exclude all pixels identified as clouds or shadows, enhancing the reliability of the dataset for environmental analysis.

Monthly composites were collected for January through August of 2023 for the the bands B2, B6, B8, B11, and B12. We also calcuate the Enhanced Vegetation Index (EVI) and 'hue' the color spectrum value. This computed hue value provides the basic color as perceived in the color wheel, from red, through green, blue, and back to red. Due to the high prevelence of clouds in the region, we used linear interpolation to fill in missing data in the time series using `xr_fresh` (Mann, Michael L. 2024). These bands were selected based on their relevance to crop type classification and their ability to capture the unique spectral signatures of different crops. The composites were used to generate time series features for each pixel in the study area, providing valuable information on the temporal dynamics of crop growth and development.

Time Series Features

Time series features capture the temporal dynamics of crop growth and development, providing valuable information on the phenological patterns of different crops. We leverage the time series nature of the satellite imagery to extract relevant features for crop type classification.

In this study, we utilized the `xr_fresh` toolkit to compute detailed time-series statistics for various spectral bands, facilitating comprehensive pixel-by-pixel temporal analysis (Mann, Michael L. 2024). The `xr_fresh` framework is specifically designed to extract a wide array of statistical measures from time-series data, which are essential for understanding temporal dynamics in remote sensing datasets.

The metrics computed by `xr_fresh` in this study include basic statistical descriptors, changes over time, and distribution-based metrics, applied to each pixel's time series for selected spectral bands (B12, B11, hue, B6, EVI, and B2). The list of computed time-series statistics encompasses:

- **Energy Measures**: Absolute energy which provides a sum of squares of the values.
- **Change Metrics**: Absolute sum of changes to quantify overall variability, mean absolute change, and mean change.
- **Autocorrelation**: Calculated for three lags (1, 2, and 3) to assess the serial dependence at different time intervals.
- **Count Metrics**: Count above and below mean, capturing the frequency of high and low values relative to the average.
- **Extreme Values**: Day of the year for maximum and minimum values, providing insight into seasonal patterns.
- **Distribution Characteristics**: Kurtosis, skewness, and quantiles (5th and 95th percentiles) to describe the shape and spread of the distribution.
- **Variability Metrics**: Standard deviation, variance, and whether variance is larger than standard deviation to evaluate the dispersion of values.
- **Complexity and Trend Analysis**: Time series complexity and symmetry looking, adding depth to the analysis of temporal patterns.

Notably, certain statistics like `longest_strike_above_mean` and `longest_strike_below_mean` were excluded due to computational constraints related to GPU memory capacity on the JAX platform. Additionally, some planned metrics such as OLS slope, intercept, and R-squared calculations were not implemented.

The integration of `xr_fresh` into our analytical workflow allowed for an automated and robust analysis of temporal patterns across the study area. By leveraging this toolkit, we could efficiently process large datasets, ensuring that each pixel's temporal dynamics were comprehensively characterized, which is critical for accurate environmental monitoring and change detection.

Data Extraction

To partially accound for variation in field size we extract pixels based on a buffer around field point locations. Small fields were buffered by only 5 meters, medium by 10m and large by 30m. This approach allowed us to capture the time series features from the surrounding area, providing a more comprehensive representation of the field's characteristics. The use of larger buffers was explored but found to decrease model performance as fields tended to be heterogenous - for instance containing patches of trees. To account for this in our modeling we treat observations from the same field as a "group" in our cross-validation scheme - as described below.

Machine Learning Models

In our study, we utilized the extracted time-series features from satellite imagery, described above, to analyze crop classifications. Notably, features were centered and scaled from the `scikit-learn` library to normalize the data, followed by the application of a variance threshold method to reduce dimensionality by excluding features with low variance (Pedregosa et al. 2011).

In this study, we employed `Optuna`, an optimization framework, to conduct systematic model selection and hyperparameter tuning (Akiba et al. 2019). Our methodology involved defining a study using Optuna where each trial proposed a set of model parameters aimed at optimizing performance metrics. Specifically, we used stratified group k-fold cross-validation with the number of splits set to three, ensuring that samples from the same field were not split across training and validation sets to prevent data leakage. The scoring metric utilized was the kappa statistic, chosen for its suitability in evaluating models on imbalanced datasets.

This approach allowed us to rigorously evaluate and compare different classifiers, including LightGBM, Support Vector Classification (SVC), and RandomForest, and their configurations under a variety of conditions. The final selection of the model and its parameters was based on the ability to maximize the kappa statistic, ensuring that the chosen model provided the best possible performance for the classification of land cover types in our dataset.

Interpretation and Feature Selection

To interpret the contributions of individual features to the model predictions, we employed SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017). This approach, based on game theory, quantifies the impact of each feature on the prediction outcome, providing insights into which features are most influential in determining land cover types.

In our feature selection process, we incorporate both the mean and maximum SHAP values to comprehensively assess the influence of features on model predictions. The mean of the absolute SHAP values across all samples, provides a measure of the average impact of each feature, highlighting its overall importance across the dataset. This approach emphasizes features that consistently affect the model's output but might underrepresent the significance of features causing substantial impacts under specific conditions. To address this, we also consider the maximum absolute SHAP values. Sorting features by their maximum absolute SHAP values allows us to identify those that have significant, albeit possibly infrequent, effects on individual predictions. This method ensures that features crucial for particular scenarios are not overlooked, thus offering a more nuanced understanding of feature importance that balances general influence with critical, situation-specific impacts.

Feature selection then is the union of the top 30 time series features found with both the mean and maximum SHAP values, resulting in 33 total features. This approach ensures that the selected features are both consistently influential across the dataset and capable of exerting substantial impacts under specific conditions, providing a comprehensive set of features for model training and evaluation.

Results & Discussion

Crowd Sourced Data

In addressing the significant gap in available crop type datasets, particularly in developing regions, this study harnessed the power of crowdsourced data to enhance the robustness and applicability of our machine learning models. Crowdsourced data collection, an innovative approach in the agricultural domain, involves gathering data from a large number of volunteers or citizen scientists, who provide valuable ground truth information. This method has proven especially useful in areas where traditional data collection methods are challenging due to logistical, financial, or infrastructural constraints.

By leveragign the YouthMappers student organization, with over 400 active chapter in _____ countries, we were able to collect a large dataset of crop type observations in Tanzania. Moreover this exercise provided an important opportunity for students to gain practical experience in data collection, analysis, and interpretation, contributing to their professional development and capacity building in the geospatial domain.

@ Someone finish _____

Land Cover and Crop Type

The distribution of primary land cover types within the training dataset used for the model is represented in Figure 1. The dataset consists of a diverse range of land cover types, each contributing differently to the total number of observations. Maize is the most prevalent land cover type, accounting for the highest percentage of the observations, followed by rice and sunflower. This is indicative of the agricultural dominance in the region being studied. Lesser common land covers such as millet, sorghum, and urban areas represent intermediate percentages, suggesting a varied landscape that includes both agricultural and urbanized zones.

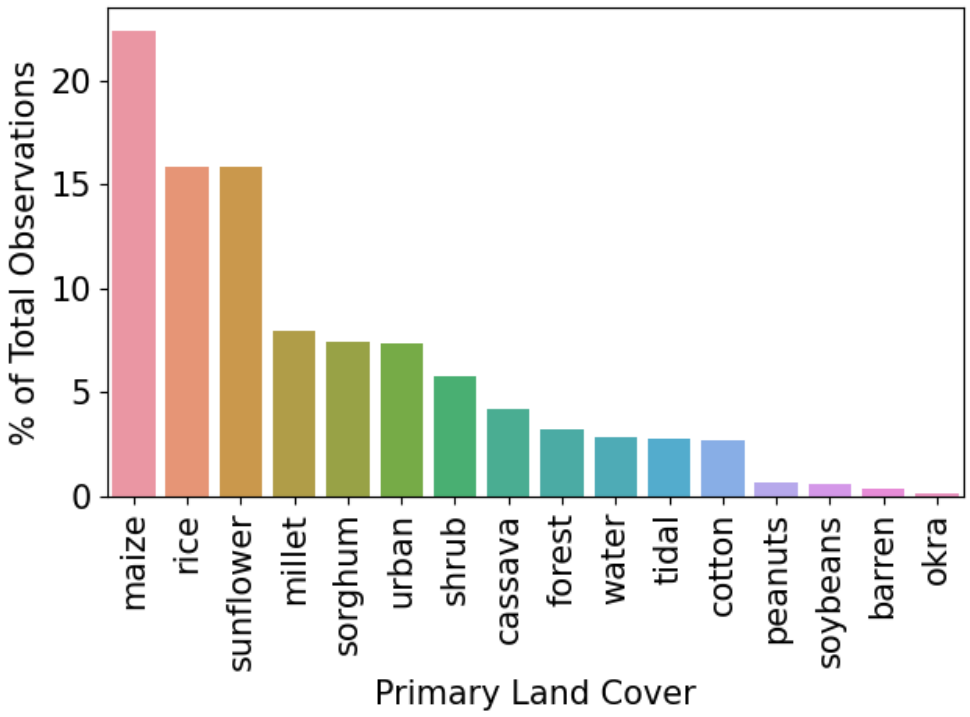

Figure 1: Land Cover Percentages

Feature Importance

The interpretation of model behavior using SHAP values has allowed for a deeper understanding of how different spectral features impact the model's predictions, which is critical for refining the feature selection process. By analyzing both the mean and maximum SHAP values, we were able to prioritize features based on their overall impact as well as their critical contributions to specific model decisions.

In the two summary plots below, SHAP values for each feature to identify how much impact each feature has on the model output for individuals in the validation dataset. Features are sorted by the sum of the SHAP value magnitudes across all samples. The figures bar color (hue) represents the mean contributions to explaining each land class value. This visualization provides a comprehensive overview of the feature importance, highlighting the key predictors that drive the model's predictions. For example, features that are highly influential for "maize" may not be as impactful for "rice" or "sorghum", reflecting the unique spectral signatures of these crops.

Mean SHAP Values In Figure 2, the mean SHAP values provide insights into the average impact of each feature across all predictions. This analysis highlights the features that consistently influence the model's output across various scenarios. For example, the mean value of B11 (B11.mean) and the 5th percentile of hue (hue.quantile.q.0.05) features were found to have substantial average impacts on model outputs, suggesting their strong relevance in distinguishing between different crop types. Reflecting on the colors of the bars we can see that 'B11.mean' is important in distinguishing sunflow, sorghum, and millet to a roughly equal degree, and has some small impact on distiguishing other classes. While 'hue.quantile.q.0.05' has the strongest effect distiguishing rice, sunflower, and to a lesser degree cotton. Looking down the list we can see that features like "EVI.standard.deviation" are most effective at isolating urban areas, and 'B12.mean.second.derivative.central' substantively differentiates shrub from other classes. Note that the mean second derivative of B12 is a measure of the rate of change of the rate of change of the B12 band, so positive values indicate increasing rate of change (increasingly upward trend), and negative values decreasing rate of change (increasingly downward trend).

Maximum SHAP Values

On the other hand, Figure 3, maximum SHAP values uncover features that, while perhaps not consistently influential, have high impacts under particular conditions. This aspect of the analysis is crucial for identifying features that can cause significant shifts in model output, potentially corresponding to specific agricultural or environmental contexts. Features such as "hue.median" and "B11.maximum" show high maximum SHAP values, indicating their pivotal roles in certain classifications. For instance, "B11.maximum" reflects peak reflectance in the Short-Wavelength Infrared (SWIR), which could be critical in identifying crops at their maximum biomass, like sunflower at full bloom compared to other crops at different stages of growth.

The final selection of features for model training was carefully curated to include all 30 of both the highest mean and maximum SHAP values, ensuring a comprehensive set of predictors for accurate and reliable classification of crop types in Tanzania. This strategic selection process has not only improved model accuracy but also enhanced our understanding of the spectral characteristics most relevant for distinguishing among the diverse agricultural landscape of the region.

Model Performance

The classification model demonstrated robust performance across multiple land cover classes, as evidenced by the out-of-sample mean confusion matrix with a Cohen's Kappa score of 0.7975, indicating substantial agreement between predicted and actual classifications. The confusion matrix (Figure 4) shows high diagonal values for most classes, highlighting the model's ability to accurately identify specific land covers. For instance, 'rice' and 'urban' categories achieved classification accuracies of 90% and 94%, respectively. Other well-classified categories included 'forest' and 'millet', each with over 70% accuracy. However 'forest' is primarily confused with the category 'shrub', which is likely a result of poor training data obtained from high-res imagery.

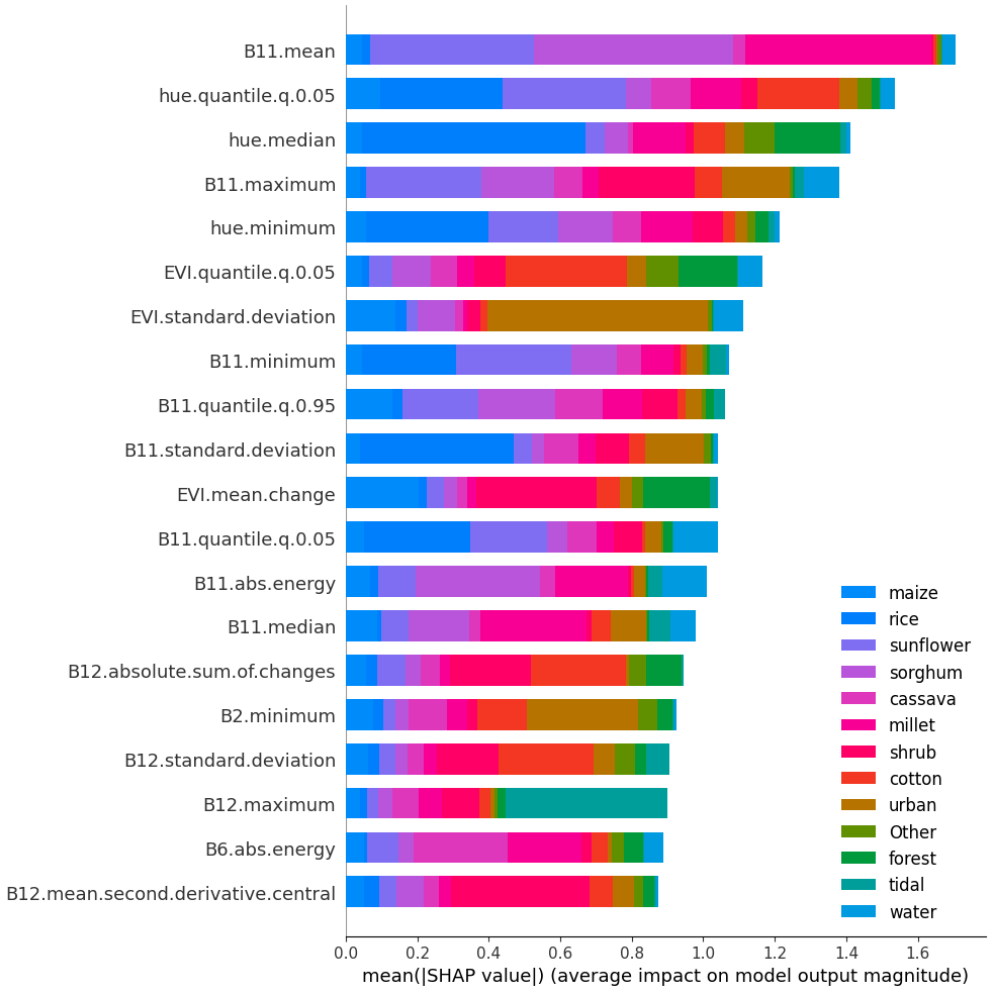

Figure 2: Top 20 Mean SHAP Feature Importance by Land Cover Type

Categories such as 'sorghum' and 'cotton' displayed moderate confusion with other classes, indicating potential areas for model improvement, especially in distinguishing features that are common between similar crop types. Notably, the 'other' category showed a broader distribution of misclassifications, likely due to its encompassing a diverse range of less frequent land covers, achieving a lower accuracy of 40%.

The overall high performance across the majority of categories suggests that the model is effective for practical applications in land cover classification, though further refinement is recommended for categories showing lower accuracy and higher misclassification rates.

Conclusion

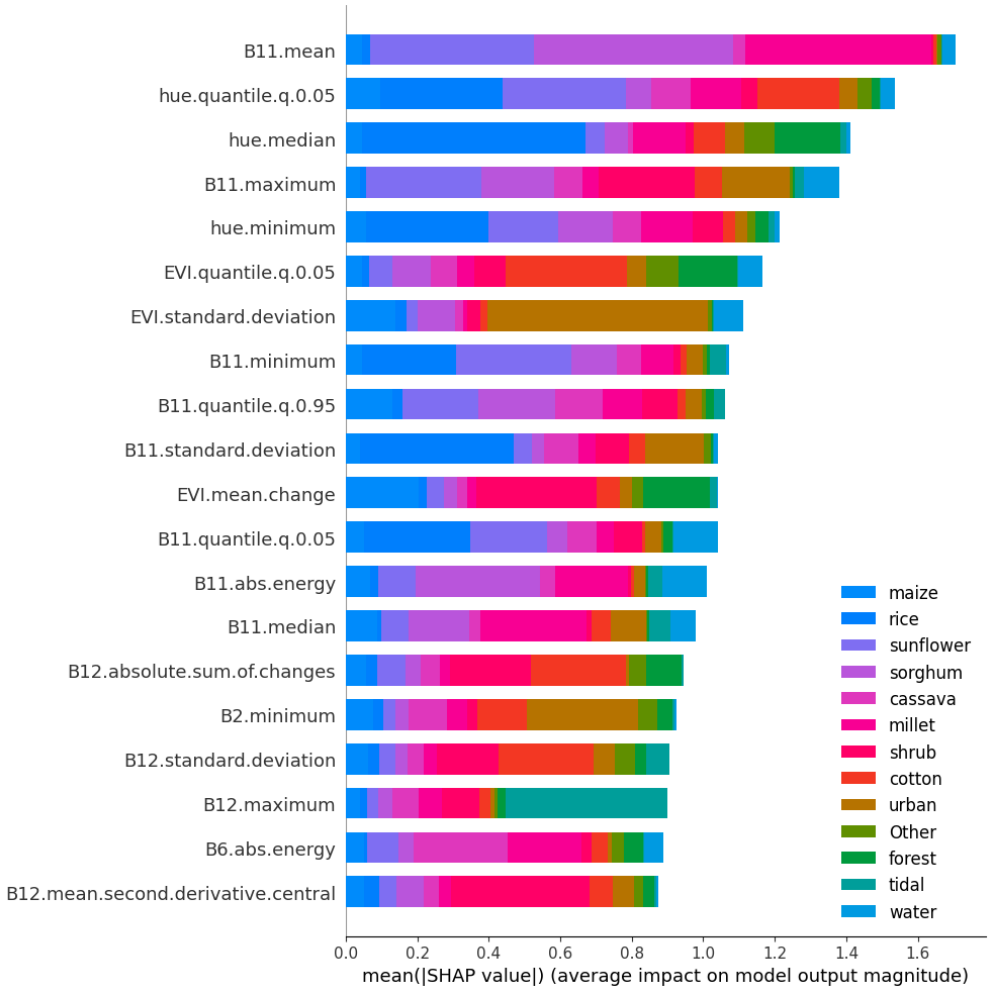

Figure 3: Top 20 Max SHAP Feature Importance by Land Cover Type

References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. "Optuna: A Next-Generation Hyperparameter Optimization Framework." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems 30*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mann, Michael L. 2024. "xr_fresh: Python Package for Feature Extraction from Raster Data Time Series." Zenodo. https://doi.org/10.5281/zenodo.12519007.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Tseng, Gabriel, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. 2021. "CropHarvest: A Global Dataset for Crop-Type Classification." In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=Jtjz UXPEaCu.

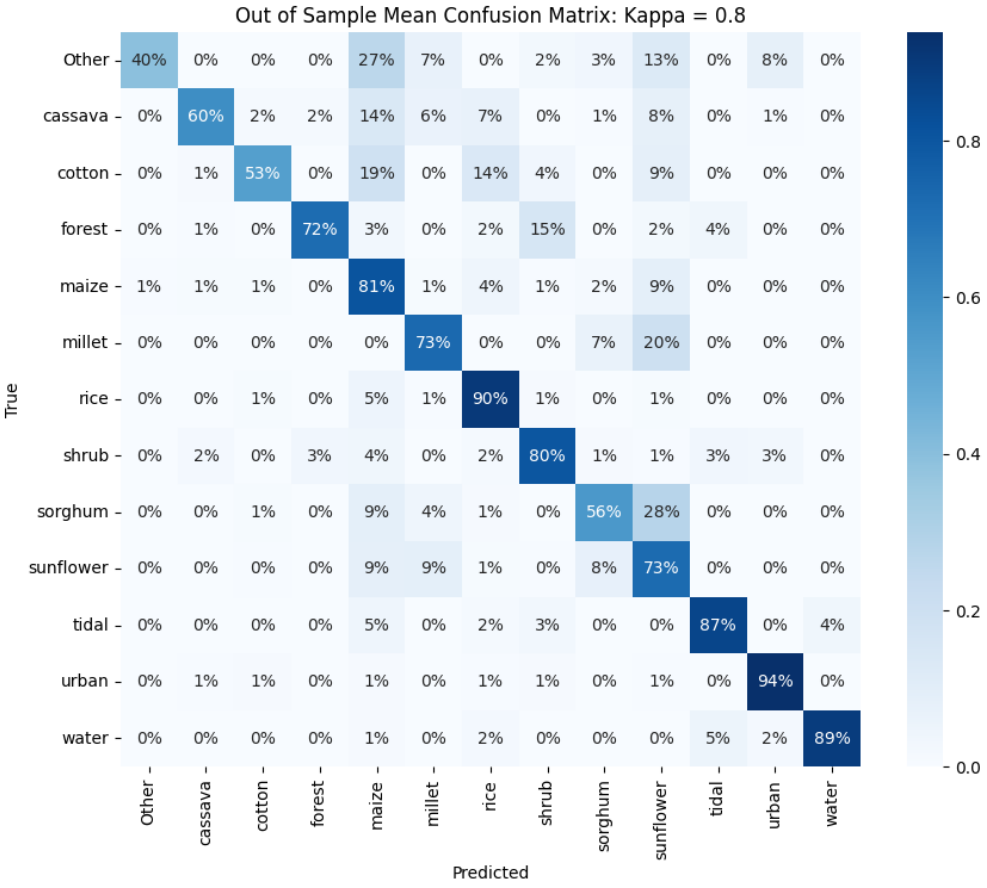

Figure 4: Out of Sample Confusion Matrix