Lite Learning: Efficient Crop Classification in Tanzania Using Traditional Machine Learning & Crowd Sourcing

Michael L. Mann, Lisa Colson, Rory Nealon, Ryan Engstrom, Stellamaris Nakacwa

***Abstract*—This study introduces a novel approach to traditional machine learning methodology for crop type classification in Tanzania, by integrating crowdsourced data with time-series features extracted from Sentinel-2 satellite imagery. Leveraging the YouthMappers network, we collected ground validation data on various crops, including challenging types such as cassava, millet, sunflower, sorghum, and cotton across a range of agricultural areas. Traditional machine learning algorithms, augmented with carefully engineered time-series features, were employed to map the different crop classes. Our approach achieved high classification accuracy, evidenced by a Cohen's Kappa score of 0.82 and an F1-micro score of 0.85. The model often match or outperform broadly used land cover models which simply classify 'agriculture' without specifying crop types. By interpreting feature importance using SHAP values, we identified key time-series features driving the model's performance, enhancing both interpretability and reliability. Our findings demonstrate that traditional machine learning techniques, combined with computationally efficient feature extraction methods, offer a practical and effective "lite learning" approach for mapping crop types in data-scarce environments. This methodology facilitates accurate crop type classification using a low-cost, resource-limited approach that contributes valuable insights for sustainable agricultural practices and informed policy-making, ultimately impacting food security and land management in resource-limited contexts, such as sub-Saharan Africa.**

***Index Terms*—Remote sensing, machine learning, crop classification, time-series analysis, crowdsourcing, Sentinel-2, feature engineering.**

I. Introduction

A. Background and Context

The free access to remotely sensed data, such as imagery from satellites (e.g. Sentinel-2, Landsat), has allowed for 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 no or 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 pixels, 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 and time-series patterns within the remotely sensed data. 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 [?], [?], [?], [?], [?].

Traditional machine learning algorithms require the extraction of variables (e.g. max EVI, mean blue band) that can help distinguish different plant or crop types [?]. The development of salient time-series features to capture phenological differences between locations from remotely sensed images remains a challenge. These features are typically derived from the spectral bands (e.g. red edge, NIR) of the satellite imagery or indexes, such as the enhanced vegetation index (EVI), and basic time series statistics (e.g. mean, max, minimum, slope) for the growing season [?]. Meanwhile a broader set of time series statistics from bands or indexes may be more relevant for a number of applications. For instance the skewness of EVI might help distinguish crops that green-up earlier 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.

In contrast, deep learning methods have dominated the most recent literature [?], [?]. These methods include both recurrent neural networks (RNN) and convolutional neural networks (CNN). Recurrent Neural Networks (RNNs) are a class of neural networks that are particularly powerful for modeling sequential data such as time series, speech, text, and audio. The fundamental feature of RNNs is their ability to

Michael L. Mann and Ryan Engstrom are with The George Washington University, Washington DC 20052 (e-mail: mmann1123@gmail.com).

Lisa Colson is with the USDA Foreign Agricultural Service, Washington DC 20250.

Rory Nealon is with USAID GeoCenter, Washington DC 20523.

Stellamaris Nakacwa is with YouthMappers, Texas Tech University, Lubbock TX 79409.

maintain a 'memory' of previous inputs by using their internal state (hidden layers), which allows them to exhibit dynamic temporal behavior. RNNs and its variants allow the integration of time-series imagery, significantly improving crop type classification outcomes especially in data rich environments [?], [?]. Deep learning approaches however typically require much larger sets of training data, may be more prone to overfitting especially with small sample sizes, have significant limitations to interpretability, and require expensive compute [?], [?], [?], [?], [?]. Although recent efforts have closed the gap e.g. [?], the lack of readily available and reliable ground truth data or benchmark datasets for training, as discussed earlier, may limit the applicability of deep learning for a variety of tasks including crop classification and make researchers more reliant of less reliable techniques like transfer learning or zero-shot or low shot methods [?], [?], [?]. Moreover, training data for extreme events, like crop losses, disease, and lodging are largely non-existant. Interpretability is also a salient weakness as interpretation of models allows us to gain scientific insight and assess trustworthiness and fairness in so far as outputs affect policy decisions.

II. Methodology

A. Study Area

B. Data Collection

C. Satellite Imagery and Preprocessing

D. Feature Engineering

Using the `xr_fresh` toolkit, we computed time-series statistics, including absolute energy, autocorrelation, skewness, and variance. These features enhance the interpretability and efficiency of traditional machine learning classifiers.

E. Model Selection and Evaluation

Optuna was employed for hyperparameter tuning across multiple classifiers, including LightGBM and Random Forest. Model performance was assessed using stratified group k-fold cross-validation, ensuring robustness against data leakage. SHAP values provided insights into feature importance.

III. Results and Discussion

The model achieved an overall accuracy of 85%, with particularly strong performance for maize and rice classification. SHAP-based feature analysis revealed that SWIR bands and vegetation indices played a critical role in distinguishing crop types.

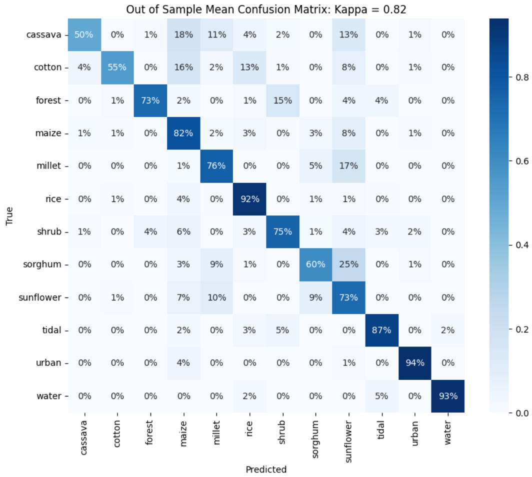

Fig. 1: Confusion Matrix of Crop Classification Model

IV. Conclusion

This study demonstrates that traditional machine learning methods, when combined with engineered time-series features, offer an interpretable and computationally efficient solution for crop classification in Tanzania. Future work should explore integrating additional spectral indices and expanding the dataset for improved model generalizability.