Mapping Agricultural Plastics in California with End-to-End CV Pipeline

April Chang [1] **Amelia Li** [1] **Rebecca Qiu** [1] **Peter Wu** [1] **Evan Arnold** [1]

Abstract

Agricultural practices have increasingly incorporated single-use plastics, resulting in significant environmental and economic repercussions. This research addresses the urgent need to monitor and manage plastic usage through the development of an end-to-end computer vision (CV) pipeline. Existing solutions fail to solve this problem because it poses various challenges including class imbalances, data sparsity, and regional variations, which necessitates a robust and generalizable model. Our proposed solution innovates in the application of machine learning models augmented by geographical feature engineering to accurately detect and categorize types of agricultural plastics. Preliminary results demonstrate the system's efficacy, particularly in differentiating between plastic mulch and other plastic types in California's diverse agricultural landscapes.

1. Introduction

The problem we are addressing is the rampant use of single-use plastics such as mulch films and hoop houses in California's agricultural sector. The problem is of importance because California's extensive agricultural activities have both localized and global repercussions. The state's unmonitored use of agricultural plastics contributes to environmental degradation, impacts human health, and raises social justice concerns. As a major player in the U.S. and worldwide agricultural markets, California's practices can set a precedent for other regions. The unchecked plastic usage is not only a direct threat to the state's ecosystems but also exacerbates global issues like climate change. Moreover, the lack of data on plastic use limits our ability to understand its full impact, hindering effective policy-making and conservation efforts. Therefore, addressing this problem can have far-reaching benefits for environmental sustainability, social equity, and global climate goals.

Similar attempts have been made on other regions in the world such as Mexico and Shandong Province of China. They leverage remote sensing data from SENTINAL-2 satellite and various supervised-learning models. However, binary classification approach (plastic/non-plastic) is adopted in these previous work. And they did not address how the model trained on these regions perform when it's generalized to other area. We try to improve these issues by categorizing plastic usage with multi-class labels, particularly differentiating the usage of hoop houses and mulch from others. Moreover, we introduced county-level geographical features including distance to coastline and elevation to encode regional variations in climate pattern and soil characteristics into our data. With the incorporation of additional geographical features, our model gained higher accuracy at county-level plastic classifications.

This study introduces a novel approach that leverages an end-to-end CV pipeline that integrates machine learning with geographical feature engineering to enhance detection accuracy. Additionally, we developed a user-friendly interface for model training and classification visualization. Users can simply upload the labeled data to train a new model, then choose the Californian country they want the model to classify on. This way, a clear picture of the type of plastics and the amount used for each county can be obtained with ease. The intention is to provide a robust and simple-to-use tool for environmental protection agencies and policymakers, aiding in the implementation of more sustainable agricultural practices by enabling them to have a better understanding of the plastic usage in California.

2. Related Works

Various methodologies have been explored for the accurate mapping and monitoring of agricultural plastic structures and mulch. For instance, the use of one-dimensional Convolutional Neural Networks (1D-CNN) in the study *Mapping Plastic Greenhouses with Two-Temporal Sentinel-2 Images and 1D-CNN Deep Learning* (Sun et al., 2021) demonstrated superior classification accuracy over traditional machine learning algorithms like Support Vector Machines (SVM) and Random Forests (RF). While the paper doesn't

[*]Equal contribution [1]Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA. Correspondence to: April Chang <aprilchang@g.harvard.edu>, Amelia Li <weixili@g.harvard.edu>, Rebecca Qiu <zqiu1@fas.harvard.edu>, Peter Wu <pwu@g.harvard.edu>, Evan Arnold <evanarnold@g.harvard.edu>.

explicitly mention it, one might question how the changing reflectance of greenhouses during crop growth could impact the model's long-term reliability.

Another approach involves incorporating ancillary data and manual visual interpretations. The study *High-resolution mapping of protected agriculture in Mexico, through remote sensing data cloud geoprocessing* (Perilla & Mas, 2019) utilized additional data to mask areas where plasticulture is unlikely to occur. It relied on training polygons created through visual interpretation of Sentinel color composites and Google Earth imagery. Although scalability or reliability without manual intervention isn't addressed, it raises questions for further research.

Hoop House Plastic Mulch

Figure 1. **Common Agricultural Plastic Usages in California.** Hoop houses and mulches are the most commonly used plastics for agriculture in California.

Our approach is different from the existing methods as they adopt a binary classification approach, which only informs users whether plastic is used for a specific region without telling them the specific type of plastic used. They also fail to address how environmental factors impact the plastic usage in agricultural practices. We improve the granularity of classification by introducing multiple plastic labels into model training, which are hoop house, mulch, and others. We also recognize that that geographical and environmental factors play a pivotal role in shaping the distribution and utilization of agricultural plastics. Since our goal is to apply the model throughout agricultural lands in California, there can be a lot of variation in agricultural plastic usage due to diverse climate patterns and soil characteristics, which we try to capture with feature engineering. Hence our model has the capability to understand the regional nuances that influences plastic use across different Californian counties.

3. Background

The background for our method involves a comprehensive understanding of CV techniques and feature engineering processes essential for processing satellite imagery. This includes the utilization of spectral features from multi-band satellite images and the calculation of various indices like NDVI (Normalized Difference Vegetation Index) to enhance the model's input data.

Our method also involves an extensive suite of supervised classification models including Support Vector Machine (SVM), Classification and Regression Tree (CART), Random Forest (RF) and Gradient Boosting Tree (GBT), which requires knowledge of each model's architecture, strengths, and limitations.

4. Methods

Our methodology consists of systematic data collection from high-resolution satellite imagery, followed by labeling and processing to create a training dataset representative of the diverse California landscape.

Next, four machine learning models (SVM, CART, RF, GBT) were trained to perform the classification task and evaluated on the processed data. We then selected the best-performing model based on validation accuracy.

Then, we further explored feature engineering for data augmentation, which played a crucial role in model performance. The inclusion of additional geographical features such as distance to coastline and elevation contributed significantly to the model's accuracy.

Finally, we developed a web application that streamlines the process of satellite image processing, model training, and visualization of the classification result.

5. Experiments

5.1. Set Up

Data Collection and Preprocessing Our training data includes two major components: labeled data from diverse Californian locations, which will be used to train multiple supervised classification models, and remote sensing data obtained from SENTINEL-2 satellite, which provides a broad and detailed view of the agricultural landscapes.

The data labeling process is labor-intensive and time-consuming. Fortunately, the labeled data is kindly provided to us by our partner organization, The Nature Conservancy of California (TNC), and it contains data from Oxnard, Santa Maria, Watsonville, and Mendocino counties.

To collect and preprocess the remote sensing data, we leveraged Google Earth Engine Python API, which provided us with a wide variety of satellite images. We process the image collection from Sentinel-2 within a specified date range for each unique date present where the data are recorded before filtering the collection by cloud coverage.

Next, we select spectral bands of the image collection and add in new index features, enhancing them with additional data that can be used for further analysis. These new index features include:

- NDVI (Normalized Difference Vegetation Index): an indicator of live green vegetation.
- NDTI (Normalized Difference Tillage Index): used to indicate soil preparation and tillage.
- PGI (Plastic Greenhouse Index): a custom index for detecting plastic greenhouses.
- PMLI (Plastic-Mulched Landcover Index): another custom index for detecting plastic-mulched land.

Finally, we take the median of the processed image collection across each date range, which can help to reduce noise and variability, and clip it to the region of interest (ROI). We sample the resulting image at the locations of the labeled data, extracting the values of the selected bands and the class labels for each point.

Initial Model Selection The collected data were used to train the Support Vector Machine (SVM), Classification and Regression Tree (CART), Random Forest (RF), and Gradient Tree Boosting (GTB) models via Google Earth Engine. Rigorous testing was conducted to evaluate model performance and optimize parameters. We then chose the best model based on the validation accuracy scores.

Feature Engineering To further improve model performance, we included geographical features into our dataset to retrain the selected model. In addition to the Spectrum Features and Index Features encoding relevant attributes of satellite imageries, we experimented with data augmentation by calculating the distance to coastline and the elevation as additional geographical features. The rationale is grounded in the recognition that geographical and environmental factors play a pivotal role in shaping the distribution and utilization of agricultural plastics. Since our goal is to apply the model throughout agricultural lands in California, there can be a lot of variation in agricultural plastic usage due to diverse climate patterns and soil characteristics, which we try to capture with the added features.

The newly engineered geographical features are:

- The distance to the nearest water body, which is calculated by transforming the water occurrence data from the JRC Global Surface Water dataset into a distance measurement.
- Elevation data from the Shuttle Radar Topography Mission (SRTM).

Below is a full picture of the features engineered for the remote sensing data.

Feature Engineering for Agricultural Plastic Detection		
Spectrum Features	B2	(Blue)
	B3	(Green)
	B4	(Red)
	B6	(Vegetation Red Edge-2)
	B8	(Near infrared, NIR)
	B11	(Shortwave Infrared-1, SWIR1)
	B12	(Shortwave Infrared-2, SWIR2)
Index Features	NDVI	(Normalized Difference Vegetation Index)
	NDTI	(Normalized Difference Tillage Index)
	PGI	(Plastic Greenhouse Index)
	PMLI	(Plastic-Mulched Landcover Index)
Additional Geographical Features	Distance to coastline	
	Elevation	

Figure 2. **Feature Engineering for Agriculture Plastic Detection Task.** Additional to satellite imagery related features, we computed distance-to-coastline and elevation as features to capture the geographical variation across different Californian counties.

Web Application for Model Training and Visualization Finally, We leveraged Streamlit's rapid prototyping and interactivity capability to implement a user-friendly interface for model training and classification visualization.

5.2. Results

Random forest outperforms other models. The bar chart below displays the train and validation accuracies for four different machine learning models: Support Vector Machine (SVM), Classification and Regression Tree (CART), Random Forest (RF), and Gradient Tree Boosting (GTB). Notably, the Random Forest model outperforms the others in terms of both training and validation accuracies, making it our preferred choice for the task. Additionally, the Random Forest's decision trees can be serialized and retrieved with the Google Earth Engine API, offering a valuable advantage for model deployment and future use.

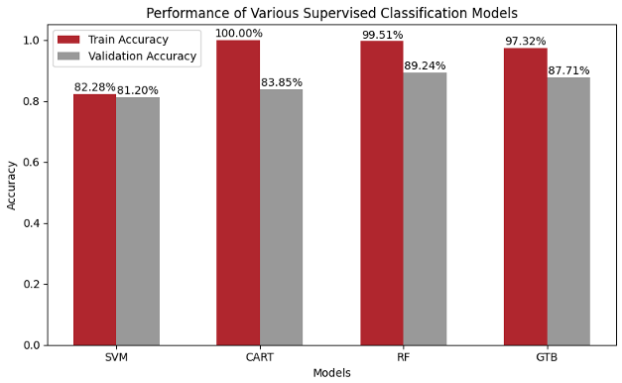


Figure 3. **Training and Validation Accuracies of Various Supervised Classification Model.** We see that Random Forest achieves a higher validation accuracy than other models.

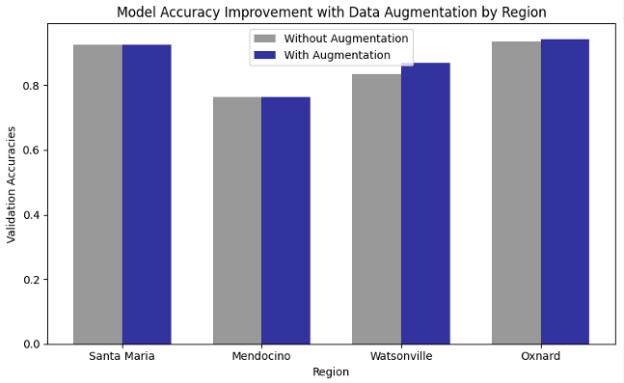


Figure 5. **Validation Accuracies Before and After Data Augmentation.** We see a significant better model performance for Watsonville and Oxnard counties, indicating that geographical features such as distance to coastline and elevation do play an important role in plastic usage across different counties in California.

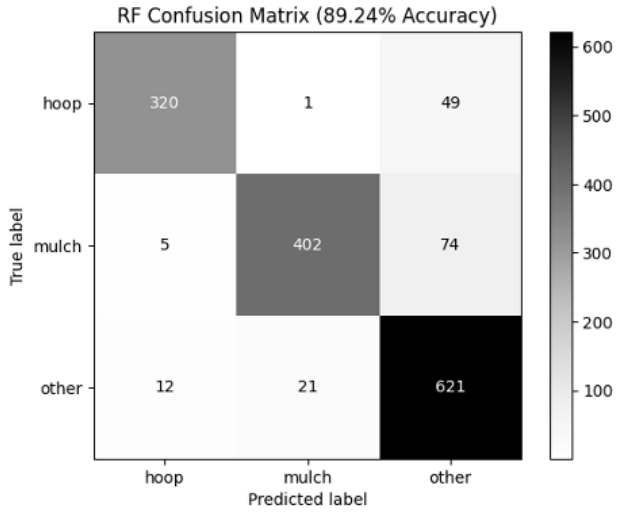


Figure 4. **Confusion Matrix of the Random Forest Model.** We see that Random Forest achieves an overall 89% validation accuracy and it performs well on all 3 classes.

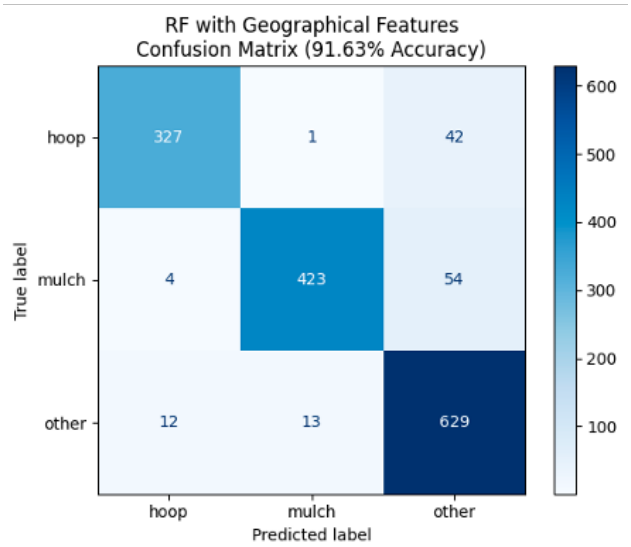


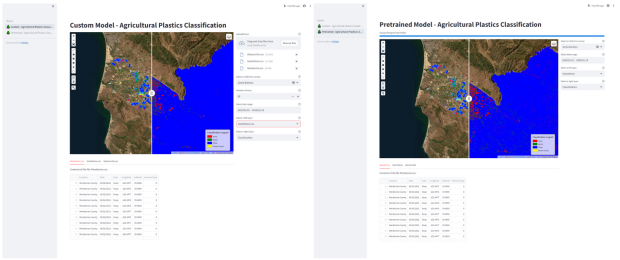
Figure 6. **Confusion Matrix of the Random Forest Model After Data Augmentation.** We see that the new Random Forest model achieves an overall 92% validation accuracy and it performs better on all 3 classes.

Data augmentation with geographical features improve model accuracy on the county-level. By incorporating the additional geographical features, the overall validation accuracy of our random forest model increases from 89.24% to 91.63%. The significant improvement with the county level accuracies in Watsonville and Oxnard also suggests a more comprehensive understanding of the regional nuances that influence plastic use and demonstrates improved accuracy in classifying agricultural plastics across diverse landscapes across California.

Web Application significantly simplified the process of model training, validation, and visualization. Our Streamlit-based web application offers an intuitive platform for both novice and experienced users in data analysis. The app presents a simple yet powerful interface where users can either upload their own CSV data files for model training or utilize our pre-trained model with preloaded data. This dual functionality caters to user needs, from custom analysis to

quick insights.

1. **Data Upload and Preloading:** Users start by choosing to upload their CSV files or utilize our preloaded datasets. This flexibility allows for both personalized and general analyses.
2. **County Selection:** Through a drop-down menu, users can select the Californian county of interest for their classification task, making the analysis region-specific.
3. **Data Presentation:** Post-selection, the app displays the data in a well-organized table format.
4. **Interactive Map Slider:** The centerpiece of our application is an interactive slider embedded in the satellite map. This feature enables users to visually compare the area before and after the classification task, enhancing the analytical experience.
5. **Dynamic Image Display:** To the left and right of the slider, users can toggle between different images. These visuals effectively illustrate the geographic coordinates of the uploaded or preloaded data, with color-coded points representing various classes.



Custom model Streamlit UI Pretrained model Streamlit UI

Figure 7. **The Web Application for Plastic Classification Task.** The web app presents a map of Santa Barbara and an additional layer of the class result. Plastics types identified by the model are labeled in different colors.

Our web application significantly simplifies the computer vision process, offering an easy-to-navigate interface tailored for efficiency and user convenience. It is particularly beneficial when users have new labeled data ready for analysis.

6. Discussion

While our chosen Random Forest model with data augmentation performs achieves high accuracy, it's worth noting that for the Mendocino county, the model's performance is less than ideal, often misclassifying hoop houses as other. The main reason is that regions like Mendocino have a much different landscape of rural forested hills, which greatly impacts how the plastic is utilized in these areas.

To mitigate this issue, the model calls for a more enriched training dataset that covers varying environmental conditions and plastic use cases. A more sophisticated feature engineering to augment the data with more geographical factors is another possible solution.

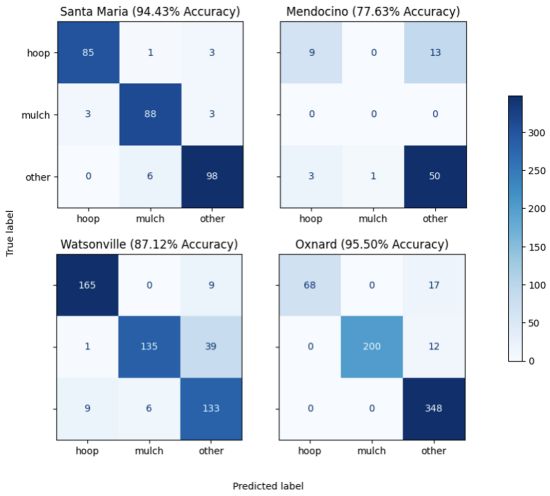


Figure 8. **Misclassification Trend for Mendocino County.** Compared to other counties, the model performed significantly worse on Mendocino, often mistaking hoop houses for other classes.

7. Results

The implementation of a Random Forest model trained on satellite imagery augmented with geographical features has demonstrated success in the classification accuracy of agricultural plastics across different Californian regions. By integrating environmental factors such as distance to coastline and elevation into the model, we've acknowledged the impact of geographical diversity on agricultural practices and, consequently, plastic usage. The overall high accuracy of the model, particularly in regions like Santa Maria and Oxnard, underscores its effectiveness in understanding and adapting to regional nuances.

The model excels in distinguishing between various types of plastics, with a high overall validation accuracy of 91.63%. This is indicative of the model's robustness and the potential for its application in real-world scenarios to aid in the management and reduction of agricultural plastic waste. However, areas for improvement have been identified, such as the misclassification trends observed in 'hoop' and 'other' categories in Mendocino county, suggesting a need for enriching the dataset with a broader spectrum of environmental

conditions and agricultural practices. Other potential strategies could include further enhancing feature engineering or exploring more sophisticated model architectures.

In addition, the web application that we developed facilitates the model's practical use by allowing users to upload data, train the model, and visualize classification results on a user-friendly interface, showcasing the project in an accessible and interactive manner.

Using Google Earth Engine with machine learning, we developed a scalable, cost-effective, and technically robust solution for monitoring agricultural plastic use. Our intention is to contribute to the ongoing efforts in understanding this complex issue, which may help inform policy decisions and lay a foundation for future endeavors aimed at agricultural sustainability.

8. Future Work

As our project progresses, delving deeper into the workings of our machine learning models will be a crucial aspect of future research. This exploration is particularly important to understand and address the failure cases of our models. One key area of focus will be investigating the feature importance in our machine learning models like Random Forest. By examining which features the model deems most significant in making predictions, we can gain valuable insights into its decision-making process. This understanding will help in identifying any potential biases or weaknesses in the model.

Furthermore, aligning the model's prediction patterns with the actual images being analyzed will be a significant step forward. This alignment will allow us to see precisely what aspects of the images the model is focusing on when making predictions. For instance, it may be detecting specific textures, colors, or shapes associated with plastics in agricultural settings that we had not fully considered. Understanding these nuances will enable us to fine-tune the model for greater accuracy and reliability.

Ultimately, the goal of this future work is to refine our machine learning models to a point where they not only predict with high accuracy but also do so in a way that is transparent and understandable. This will not only enhance the models' performance but also bolster the confidence of stakeholders who rely on these models for making informed decisions regarding agricultural practices and environmental policies. By achieving a more profound understanding of our models, we can ensure that they are robust, fair, and aligned with the real-world complexities they are designed to address.

Acknowledgements

We extend our thanks to Darcy and Kirk for their expert guidance as project partners, and to our mentors, Chase and Usha, for their invaluable support. Our appreciation also goes to Yuanyuan for her advisory role and, to TNC's volunteer Brandee, whose data contribution was essential to our project's success.

References

Aguilar, M. , Jiménez-Lao, R., Nemmaoui, A., Aguilar, F. J., Koc-San, D., Tarantino, E., and Chourak, M. Evaluation of the consistency of simultaneously acquired sentinel-2 and landsat 8 imagery on plastic covered greenhouses. *Remote Sensing*, 12(12), 2020. ISSN 2072-4292. doi: 10.3390/rs12122015. URL https://www.mdpi.com/2072-4292/12/12/2015.

Jiménez-Lao, R., Aguilar, F. J., Nemmaoui, A., and Aguilar, M. A. Remote sensing of agricultural greenhouses and plastic-mulched farmland: An analysis of worldwide research. *Remote Sensing*, 12(16), 2020. ISSN 2072-4292. doi: 10.3390/rs12162649. URL https://www.mdpi.com/2072-4292/12/16/2649.

Perilla, G. and Mas, J. High-resolution mapping of protected agriculture in mexico, through remote sensing data cloud geoprocessing. *European Journal of Remote Sensing*, 52:532–541, 01 2019. doi: 10.1080/22797254.2019.1686430.

Sun, H., Wang, L., Lin, R., Zhang, Z., and Zhang, B. Mapping plastic greenhouses with two-temporal sentinel-2 images and 1d-cnn deep learning. *Remote Sensing*, 13 (14), 2021. ISSN 2072-4292. doi: 10.3390/rs13142820. URL https://www.mdpi.com/2072-4292/13/14/2820.

Xiong, Y., Zhang, Q., Chen, X., Bao, A., Zhang, J., and Wang, Y. Large scale agricultural plastic mulch detecting and monitoring with multi-source remote sensing data: A case study in xinjiang, china. *Remote Sensing*, 11:2088, 09 2019. doi: 10.3390/rs11182088.