Multispectral Satellite Image Segmentation for Agricultural Land Classification: Mapping Pineapple Plantations in Costa Rica

Sage McGinley-Smith
sagems@stanford.edu

Hannah Norman
hnorman@stanford.edu

Abstract

This study leverages Sentinel-2 multispectral satellite image data to identify pineapple plantations in Costa Rica using a Random Forest (RF) model. Given ground-truth data for the years 2018-2019, we comparatively explore the effectiveness of a Decision Tree classifier, RF classifier, and Gradient-Boosting classifier on the task of pineapple plantation identification across the entire country of Costa Rica. We find that the RF model outperforms the others in terms of overall accuracy and scalability. Then, we apply the optimized RF model to unlabeled satellite data from 2020-2023 to address the deficit in available agricultural land-use mapping data for those years in Costa Rica. We conclude the study with suggestions for further optimizations of our own model, and we consider the merits of more advanced, state-of-the-art deep learning image segmentation models. Project page: https://github.com/sagems/pineapple_classification.

1. Introduction

The task of image segmentation is a central focus of modern computer vision research. While early image segmentation was done using low-level image features, like color and brightness values [19], advances in sensors that can capture light outside of the visible spectrum have changed the landscape of what is possible in pixel classification. Notably, the application of classification tools, like tree-based machine learning models, to satellite imagery has opened up the field of terrestrial remote sensing to further investigations into human-environment interaction [1].

The availability of preprocessed, fine-scale resolution imagery of the planet has increased substantially in recent years [4]. Specifically, missions such as NASA's Landsat and the European Space Agency's Sentinel satellites have increased the amount of multi-spectral image data available for scientific use [7]. Applications of both tree-based and deep learning models to perform segmentation tasks on these images have been successful, particularly in projects involving land-use/land-cover (LULC) classification [15].

This project expands upon applications of segmentation tools to classify pineapple plantations in Costa Rica. Pineapple is a cornerstone agricultural product in the Costa Rican economy [3], but currently no high quality dataset exists mapping pineapple plantations across the country in the past four years. Through collaboration with researchers at Centro Natcional de Alta Tecnología (CeNAT) in Costa Rica and the lab of Erin Mordecai at Stanford University, we obtained labeled datasets showing pineapple plantation distribution across the country for the years 2015-2019. While those datasets are ground-truthed at around 99% accuracy, they are extremely labor intensive to create, and the researchers at CeNAT are seeking less time and resource-intensive methods of classifying the satellite data for the years that do not currently have labeled datasets. This project aims to fill that gap by identifying and deploying models to predict pineapple plantation distributions for 2020-2023 for future use by the CeNAT researchers.

The challenges of creating such datasets lie first in the ability to obtain high quality, cloud-free images of the entire country, a challenge that is particularly pertinent in the subtropical climate of Costa Rica. Second, the selection of indexes to extract as features for the model as well as the choice of model itself presents a myriad of options, of which there are many tenable combinations. The challenge of selecting the correct model and features to most accurately detect pineapple plantations based on their spectral reflectance signatures is a central objective of this project.

Through in-depth analysis of relevant tree-based machine learning models in the literature [15], we ultimately fit and employ a Random Forest (RF) model to classify unlabeled satellite image data of Costa Rica at 30m resolution. Our model performs with about 88% accuracy out-of-sample on the training data and is used to segment multispectral satellite images of the country for the past four years, 2020-2023, which lack labeled datasets from CeNAT.

2. Related Work

Our work builds upon a foundation of computer vision research on LULC classification [15]. The increasing availability of multispectral image data combined with advances in image segmentation has greatly expanded the landscape of satellite-based LULC classification [1]. Satellite imagery provides access to bands outside of the visible spectrum, which allows for the extraction of indices like the Normalized Difference Vegetation Index (NDVI) [2]. These indices alongside the bands provide additional feature extraction capabilities beyond those traditionally seen in color image segmentation. It follows that with increasingly complex LULC data representation, application to machine learning models proves even more fruitful, and examples abound. For instance, in 2017, Kulkarni explored generalized crop classification using K-Means based approaches [10]. In 2023, Tariq *et al.* mapped tobacco, wheat, barley, and gram croplands with decision tree based models [16]. Moreover, in 2024, Masolele *et al.* developed a deep learning framework to investigate land use following deforestation across the continent of Africa [12]. Each of these studies employed remote sensing data in the training of their diverse segmentation models, and they are but a few examples of the myriad explorations into applying multispectral satellite image segmentation to LULC classification.

Surveys have also been conducted to examine the efficacy of various machine learning classification models to multispectral satellite image segmentation problems in an effort to determine which perform best. Talukdar *et al.* applied Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN) algorithms alongside three other more complex models to the same LULC task [15]. They found RF to be the most effective classifier of the six, which aligned with conclusions found in their literature review ranking RF and ANN classifiers as optimal for LULC remote sensing tasks [15]. Our own research supports such a ranking of models.

Across relevant work to multispectral satellite image segmentation, tree based models consistently arise as reliable, effective classifiers for remote sensing tasks [8, 14–18]. Tree based models extend decision tree estimators in their implementation. Decision Trees (DT) themselves are the most basic of such models, and they often prove the most computationally efficient, as they do not require extensive training time like certain deep learning classifiers, like SVMs and ANNs [16]. An RF model, then, is a step up from a DT since it is an ensemble learning method which builds and combines multiple decision trees in its execution. For multispectral satellite image segmentation tasks, RF models consistently achieve overall predictive accuracy ratings in the upper 80th to lower 90th percentiles [8, 14–18].

Despite strong results achieved with tree based models, research has expanded into different solution domains in response to the ever prevalent task of segmenting remote sensing data. Chief among them sits deep learning (DL). With the recent advancements of DL methods within computer vision, efforts have been made to expand their application to segmentation of satellite imagery [20]. Yuan *et al.* conduct a comprehensive review of these DL methods, highlighting noteworthy DL models like the Fully-Convolutional Network (FCN) and the U-Net [20]. These two models have been employed extensively across remote sensing literature. Maggiori *et al.* extend an FCN and apply it to open-source geographic map data for dense classification problems [11]. Masolele *et al.*, alternatively, employ the U-Net architecture to assess LULC factors that contributed to deforestation across 30 African countries [12]. These are only two examples of multitudes, but they represent an increasingly popular new line of exploration into multispectral satellite image segmentation approaches. Although we recognize the relevance of DL models to remote sensing segmentation tasks, this work limits its scope to tree based models: specifically, DT, RF, and Gradient-Boosting (GB) classifiers.

With regards to the data itself, this project builds off of an existing CeNAT dataset [5, 9] that maps pineapple plantations for 2015-2019. Although the data is highly accurate, it was generated using an extremely resource-intensive process of repeated hand-labeling and ground-truthing. Our project aims to find a more reasonably scoped approach for remotely classifying pineapple plantations, which relies on CeNAT work for both data labels and visual checks of model performance in combination with publicly available Sentinel 2 multispectral satellite image data[1].

3. Methodology

Our basic methodology mirrors similar work described in the Related Work section of this report, specifically with regards to training data collection and model development. We select 2018 and 2019 as the years for which to collect labeled data, as those are the earliest years for which corresponding 10m resolution Sentinel 2 data exists for Costa Rica. After initially selecting 7000 pixels labeled as pineapple plantation and 7000 labeled as non-plantation from across both years in the CeNAT data, we extract band and feature data from the corresponding Sentinel images for each pixel based on its year. During the process, certain datapoints are pruned from the dataset due to various reasons. The finalized dataset contains 10,636 datapoints, which we use to train, test, and tune several models.

3.1. Training Dataset

Pixel values for the training dataset and the prediction regions come from the European Space Agency's Sentinel 2 10m resolution dataset, publicly available on the Google Earth Engine data catalogue [6]. This project uses the Level 2A dataset, which offers the highest quality data since Level 2A data has been corrected for atmosphere re-

[1] https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED

flectance and shadow detection. The selected pixels are chosen based on labeled data from the CeNAT pineapple mapping project [5, 9], from which data was provided through the Mordecai Lab at Stanford University, in conjunction with Python support libraries defining the borders of Costa Rica[2]. For each labeled pixel, the B2, B3, B4, B8, and B11 bands are extracted, which represent the Blue, Green, Red, Near-Infrared, and Short Wave Infrared bands, respectively. Python libraries which support geospatial raster data format, namely Rasterio[3] and GDAL[4], aided in this process. Using those bands, the Normalized Difference Moisture Index, Normalized Difference Water Index, Soil-Adjusted Vegetation Index, and NDVI are extracted as additional features for each pixel. Finally, we clean the training dataset using the Pandas library[5] and export to a readable format (CSV) suitable for model development.

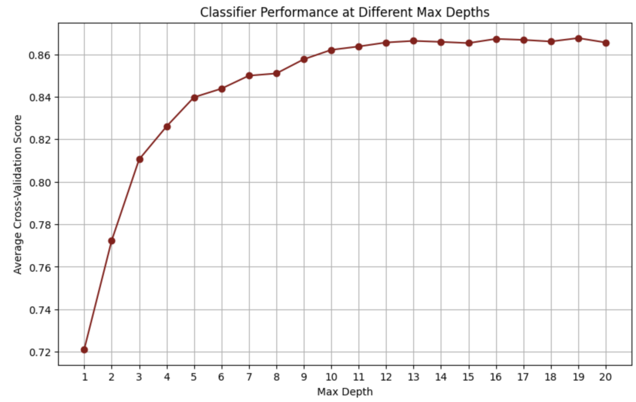

Figure 1. **Random Forest classifier depth tuning.**

3.2. Model Development

Given the training data in CSV format, we load the data points and separate features from class labels, denoting 0 as non-plantation area and 1 as plantation area. The models we will build and run are targeted at a per-pixel task, and we follow the typical design of an 80-20 train-test split across the 10,636 data points (pixels) spanning remote sensing satellite imagery from the years 2018 and 2019. As a result, our training set comprises 8,509 data points and our testing set comprises 2,127 data points. We apply these train-test sets to a Decision Tree (DT), Random Forest (RF), and Gradient-Boosting (GB) model through the use of the Scikit-learn Python library [13].

Training and testing adhere to the same flow for each model. First, hyper-parameters are tuned via one of two methods, to be described next. Then, the classifier is fit to the training data. Next, probabilities for each observation classification are calculated to aid with subsequent model analysis. Finally, the optimized model is fed the testing data, predictions are made, and the results are analyzed through a variety of metrics, including overall predictive accuracy, precision, sensitivity, f1-score, and AUROC.

Model optimization follows two routes. For the first, we explore tuning only the maximum depth parameter for each tree-based model. To do so, each model is run with different max depth values, ranging from 1 to 20, and the resulting overall accuracy is plotted for each value. For instance, Figure 1 shows the plot corresponding to our optimal RF model. The optimal depth is returned from the helper method and subsequently fed into the final classifier definition. The second route entails more comprehensive parameter tuning. As opposed to restricting tuning to the max-depth parameter, we expand considerations to a number of other hyperparameters for each model and utilize the Randomized Search Cross Validation method available through Scikit-learn [13]. Max-depths in the same range are considered for each model. For the DT, we also consider between 2 and 20 samples as the minimum required to split an internal node and between 1 and 20 samples as the minimum required to be at a leaf node. For RF, we also consider between 50 and 500 trees used in the ensemble method; either square-root two, log two, or no limits on the number of features to consider when looking for the best split; and between the same ranges as those for the DT when considering the minimum number of samples required to split an internal node and/or be at a leaf node. Finally, for GB, we also consider between 50 and 500 trees used in the ensemble method and a learning rate of either 0.01, 0.1, 0.2, or 0.3. Due to reasons that will be expanded upon in the Results section of this report, we use the initial, solely max-depth tuned RF model for our final predictive model.

3.3. Prediction Mapping

After an in-depth analysis of model performance on the training data, we select the final depth-tuned RF model to use for prediction on the unlabeled data for the years 2020-2023. We download raster data for the entire country of Costa Rica for each year from the same Sentinel dataset used for the construction of the training set, but we are limited to 30m resolution due to the compute resources available. The same bands and indices as used in the training set are extracted across the region for each raster. The model then makes predictions for each pixel in the rasters, generating a final map of predictions for each year. The final result is a 30m resolution prediction map of pineapple plantations for each of the years 2020-2023.

[2] https://developers.google.com/earth-engine/datasets/catalog/FAO_GAUL_2015_level0
[3] https://rasterio.readthedocs.io/en/stable/
[4] https://gdal.org/index.html
[5] https://zenodo.org/records/10697587

Figure 2. **Predictive plantation mapping from the years 2020-2023.** Given the ground-truth plantation mapping (left) for a region in Costa Rica in 2019, we apply unlabeled data from the years 2020-2023 to our RF model to produce predictive mappings of pineapple plantations in the same region, where the white pixels represent new classifications that were not labeled in the original data.

Model	Runtime	Overall Accuracy
Decision Tree	12s	85.3%
Random Forest	4m18s	87.9%
Gradient-Boosting	21m6s	87.5%

Table 1. **Max-depth hyperparameter tuning.**

Model	Runtime	Overall Accuracy
Decision Tree	7s	86.2%
Random Forest	12m22s	88.1%
Gradient-Boosting	45m44s	87.4%

Table 2. **Generalized hyperparameter tuning.**

Model	Acc.	Precision	Sensitivity	F1-Score	AUC
DT	85.096	85.093	85.096	85.081	0.91
RF	87.917	87.931	87.917	87.900	0.95
GB	87.635	87.645	87.635	87.619	0.95

Table 3. **Detailed depth-tuned model results.**

4. Results

As evaluation metrics for the final models, we consider overall accuracy, precision, sensitivity, f1-score, and AU-ROC. Tables 1 and 2 compare tuning approach runtimes and their resultant optimal model accuracies.

4.1. Quantitative Analysis

Table 3 presents several evaluation metrics for DT, RF, and GB models evaluated at their optimal maximum depth. RF outperforms DT by a clear margin on all metrics, though its performance in comparison to GB presents a much tighter race. The AUROC stats provided in the rightmost column of Table 3 offer an additional quantitative perspective on each model's performance. Once again, DT boasts the weakest performance with an AUC of 0.91, whereas RF and GB are tied at an AUC of 0.95. Please see the Appendix for plots of the ROC curves for each model.

Where RF clearly outperforms GB is in its tuning time. Across both approaches, RF completes its tuning execution between 4-5 times faster than GB. Although both models achieve comparable results, we note that our experiments are run on a small number of nine features and are limited to a dataset of roughly 10,000 points. Given more complex data and in much larger magnitudes, our RF model would scale more optimally than a GB model in terms of tuning time without sacrificing on performance. For this reason, we selected the RF model as the final model with which to make our predictions on the unlabeled data from Costa Rica for the years 2020, 2021, 2022, and 2023.

4.2. Qualitative Analysis

We observe that the predictions for 2020-2023 match closely on the map with the predictions from 2019, while also showing expansion of the plantations over the span of the subsequent years. Figure 2 offers an example a particular region and demonstrated plantation growth over the past four years. This provides visual confirmation that the models are correctly mapping the plantations. Visually, we also observe that the models clearly misclassify areas of rivers, roads, and clouds as plantations (see Appendix), and also struggle to classify urban areas as non-plantation areas.

5. Conclusion

This project succeeds in its goals of building a training dataset from the 2018 and 2019 CeNAT labels, and in using that training data to build and test multiple tree based models for satellite image segmentation for pineapple plantations in Costa Rica. We select a final RF model, demonstrate its efficacy on the training data, tune its hyperparmeters, and successfully deploy it on the unclassified satellite images of Costa Rica for 2020-2023. Ultimately, the predictions are visually confirmed to mirror the predictions from 2019, and they also show the plantations growing over time. On a country-wide scale, the model struggles with misclassification of roads, rivers, urban areas, and clouds, but in localized regions it appears to very accurately map those areas, although this conclusion is based off of observing the 2019 maps and not based on ground truthing of the

2020-2023 predictions. We conclude that our random forest model was able to construct datasets for 2020-2023 at a 30m resolution that are useful to the CeNAT researchers on a local scale and that further filtering of the data to reduce noise and misclassification of specific features would increase the usability of the dataset on a country-wide scale.

5.1. Limitations and Future Work

From the data perspective, this project is limited by the availability of cloudless images of the entire country of Costa Rica. Available compute resources is also a limiting factor which restricted the prediction maps to only 30m resolution, which corresponded to roughly 53 million pixels per year to classify. Future work includes reducing noise noise in the dataset by filtering out pixels that are isolated and not a part of a larger plantation area, as those are most likely misclassified as well as increasing the resolution of the predictions to 10m resolution. The training dataset can also be augmented with more labeled roads, urban areas, and rivers to prevent misclassification of those areas.

6. Contributions

Work on this project was shared equally across team members. Sage took the lead with data retrieval, preprocessing, conversion to a readable format, application to final predictive model, and mapping visualizations. Hannah was responsible for model development–including model training, testing, optimization, and analysis–across all three classifier types. Both team members supported each other in their respective tasks and wrote the project proposal, midterm status progress report, and final report jointly. We feel the work was shared fairly.

7. Acknowledgements

We thank our mentors in the Mordecai Lab at Stanford, in particular Caroline Glidden, for providing the inspiration for this project. We also thank the researchers at CeNAT for the 2015-2019 data labels that were used in the building of the training dataset. Finally, we also wish to thank Alina Chou, our CS 131 TA mentor, for her feedback and support during this project.

References

[1] Mariana Belgiu and Lucian Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogrammetry Remote Sens.*, 114:24–31, 2016. 1, 2

[2] Sebastian Candiago, Fabio Remondino, Michaela De Giglio, Marco Dubbini, and Mario Gattelli. Evaluating multispectral images and vegetation indices for precision farming applications from uav images. *Remote Sens.*, 7(4):4026–4047, 2015. 2

[3] Z Castro. The agro-industrial pineapple system in costa rica. *Acta Horticulturae*, pages 41–49, 2019. 1

[4] Francisco Eugenio and Javier Marcello. Very high resolution (vhr) satellite imagery: Processing and applications. *Remote Sens.*, 2019. 1

[5] Catalina Arguedas González, Christian Vargas Bolaños, and Cornelia Miller Granados. Conservando la biodiversidad por medio del manejo de paisajes productivos: Monitoreo del estado de la piña en costa rica para el año 2019, asociado con la pérdida y ganancia entre la cobertura forestal. Technical report, Centro Nacional de Alta Tecnología (CeNAT), 2021. 2, 3

[6] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.*, 202:18–27, 2017. 2

[7] Mohammad Ali Hemati, Mahdi Hasanlou, Masoud Mahdianpari, and Fariba Mohammadimanesh. A systematic review of landsat data for change detection applications: 50 years of monitoring the earth. *Remote Sens.*, 13(15):2869, 2021. 1

[8] Konrad Hentze, Frank Thonfeld, and Gunter Menz. Evaluating crop area mapping from modis time-series as an assessment tool for zimbabwe's "fast track land reform programme". *PLOS One*, 11(6), 2016. 2

[9] Sofía Hernández Hernández, Heileen Aguilar Arias, Iván Ávila Pérez, Sofía Acuña López, Yerlin Vargas Solano, Armando Vargas Céspedes, María Fernanda Obando Picado, Yorleny Calvo Elizondo, Milagro Jiménez Rodríguez, Jennifer Fernández Garro, Ezequiel Fallas Montero, David Romero Badilla, Marilyn Manrow Villalobos, and Cornelia Miller Granados. Paisaje productivo de pastos año 2019: Informe final. Technical report, Centro Nacional de Alta Tecnología (CeNAT), 2022. 2, 3

[10] Namita M Kulkarni. Crop identification using unsuperviesd isodata and k-means from multispectral remote sensing imagery. *Int. J. Eng. Res. Appl.*, 7(4):45–49, 2017. 2

[11] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Tran. Geosci. Remote Sens.*, 55(2):645–656, 2016. 2

[12] Robert N Masolele, Diego Marcos, Veronique De Sy, Itohan-Osa Abu, Jan Verbesselt, Johannes Reiche, and Martin Herold. Mapping the diversity of land uses following deforestation across africa. *Sci. Rep.*, 14(1681), 2024. 2

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(85):28252830, 2011. 3

[14] Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, and Gérard Dedieu. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. *Remote Sens. Environ.*, 187:156–168, 2016. 2

[15] Swapan Talukdar, Pankaj Singha, Susanta Mahato, Shahfahad, Swades Pal, Yuei-An Liou, and Atiqur Rahman. Land-

use land-cover classification by machine learning classifiers for satellite observations—a review. *Remote Sens.*, 12(1135), 2020. 1, 2

[16] Aqil Tariq, Jianguo Yan, Alexandre S Gagnon, Mobushir Riaz Khan, and Faisal Mumtaz. Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geospat. Inf. Sci.*, 26(3):302–320, 2023. 2

[17] Francesco Vuolo, Martin Neuwirth, Markus Immitzer, Clement Atzberger, and Wai-Tim Ng. How much does multi-temporal sentinel-2 data improve crop type t classification? *Int. J. Appl. Earth Obs. Geoinformation*, 72:122–130, 2018. 2

[18] Ahmed Mohamed Youssef, Hamid Reza Pourghasemi, Zohre Sadat Pourtaghi, and Mohamed M Al-Katheeri. Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at wadi tayyah basin, asir region, saudi arabia. *Landslides*, 13:839–856, 2016. 2

[19] Ying Yu, Chunping Wang, Qiang Fu, Renke Kou, Fuyu Huang, Boxiong Yang, Tingting Yang, and Mingliang Gao. Techniques and challenges of image segmentation: A review. *Electronics*, 12:1199, 2023. 1

[20] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.*, 169, 2021. 2

Appendix

A. Github Repository

Project page: `https://github.com/sagems/pineapple_classification`.

B. Additional Figures

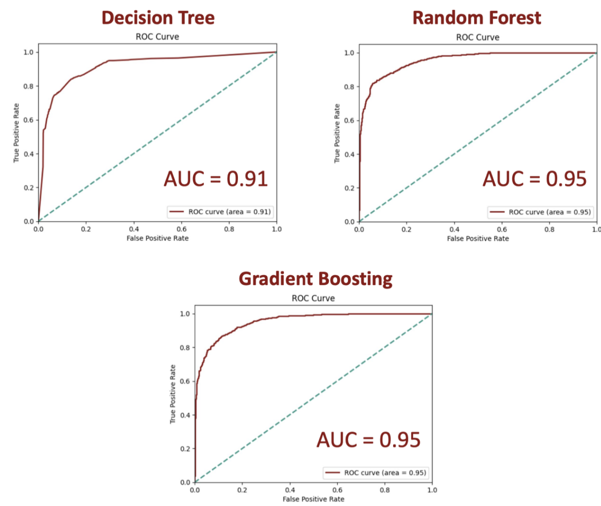

Figure 3. **ROC curves for each model.**

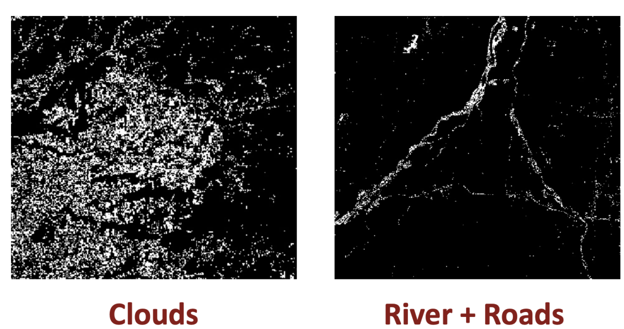

Figure 4. **Clouds, river, and roads misclassified by final model.**