Satellite Settlement Classification

Vincent Zhang

DS 3001: Foundations of Machine Learning

University of Virginia - School of Data Science

December 3, 2025

Background

- Satellite imagery is prevalent in large quantities.
  - Countless artificial satellites orbiting Earth and taking photographs at any given time
- Extracting meaningful data is **hard** and *time-consuming.*
  - Used to be done manually, now done by machines
- Due to the extensive amounts of data embedded in high-resolution imagery, lots of hidden patterns are yet to be discovered:
  - geological activity, geographical trends, land development patterns, nation-state surveillance, and for this project: settlement types

Problem Definition

- The goal of this research project is to:
  - *classify* any **true-color** image taken by a satellite of Earth into one of three categories:
    - an urban settlement
    - a rural settlement
    - uninhabited land
  - The definitions of each of these categories are derived from the publication by Liu et al. (2024).
  - As land-cover heterogeneity cannot be discretized without excessive information loss, the final predictions are soft probabilities that sum to 1.

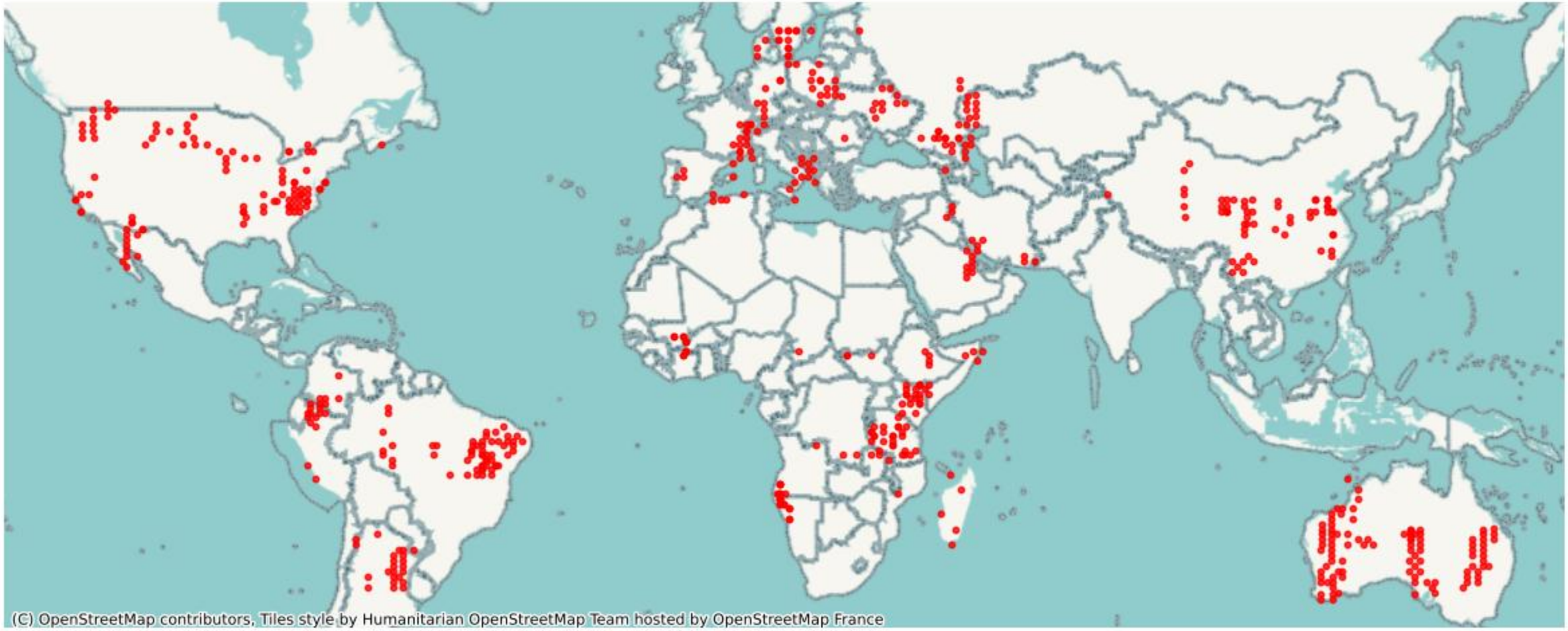
Data Collection

- We want to obtain true-color satellite imagery from a reputable source.
  - That means any generated simulations of satellite data (such as fake layers) are prohibited.
- Microsoft's Planetary Computer gives us datasets from *Harmonized Landsat and Sentinel-2 (HLS) v2.0* at 30-meter resolution.
  - The coverage of the world is extensive, and this will be the dataset of choice.
- The ground truths will be determined by the *2020 Global Rural and Urban Settlement (GURS)* dataset (Liu et al., 2024).
  - A value of “1” represents **urban.**
  - A value of “2” represents **rural.**
  - A value of “127” represents **uninhabited.**

Region Selection

- To ensure equal coverage of most human settlements in the world, I will select six 6 bounding boxes, each representing a geographical area of human activity for every inhabited continent.
  - *Europe and Asia will be considered separate continents since analysis is done on human and not natural geography.*
- For North America, much of northern Canada and all of Alaska will be cut off due to low population density.
- For Asia, much of the Siberian wilderness and the Eurasian Steppe will be omitted due to low population density.
- For South America, most of Patagonia will be cut off due to sparse population density.
- For Africa, most imagery will be selected from a bounding box that focuses on sub-Sahara due to similar reasons as above.
- Australia's outback will be kept to see how the model performs.
- Edge cases, such as ocean data and Antarctica, may appear sporadically but are not the main focus of the model.

Satellite Imagery Coverage



Ground Truth Metadata

Property	Value
Driver	GTiff
Width (pixels)	358509
Height (pixels)	133171
Number of bands	1
Data type	uint8
CRS	EPSG:4326
Transform (Affine)	| 0.00, 0.00,-180.00| | 0.00,-0.00, 77.91| | 0.00, 0.00, 1.00|
Bounds	BoundingBox(left=-179.99999840575052, bottom=-55.81961353348433, right=179.99948973589213, top=77.90541306372046)
Nodata value	127.0
Is tiled	True
Block shapes	[(128, 128)]
Unique pixel values	[np.uint8(1), np.uint8(2)]

Satellite Imagery Preprocessing

- The raw GeoTIFFs all have different coordinate reference systems, but GURS uses WGS 84.
  - Therefore, we convert all the imagery to EPSG:4326.
- Furthermore, to help the model digest the imagery more easily, we need to avoid loading whole GeoTIFFs into RAM at a time.
  - For any raw GeoTIFF, we can slice the raster into 64 quadrants to enhance surface detail.
- Some tiles will be mostly black.
  - If more than 20% of the tile is black pixels, we discard it. Otherwise, we save the output into a new directory on disk.
- This process generates 30,444 distinct satellite tiles.

Ground Truth Preprocessing

- The ground truths of each tile will be extracted from the GURS 2020 raster in two main steps:
  - Isolate the portion of GURS 2020 such that it represents the geographical area covered by any specific tile.
  - Assign a score to each land-cover label in the GURS dataset (urban, rural, uninhabited), reflecting the fractional representation of the respective class within each region.
- Finally, the frequencies are saved as a .csv file representing the ground truths of each tile.

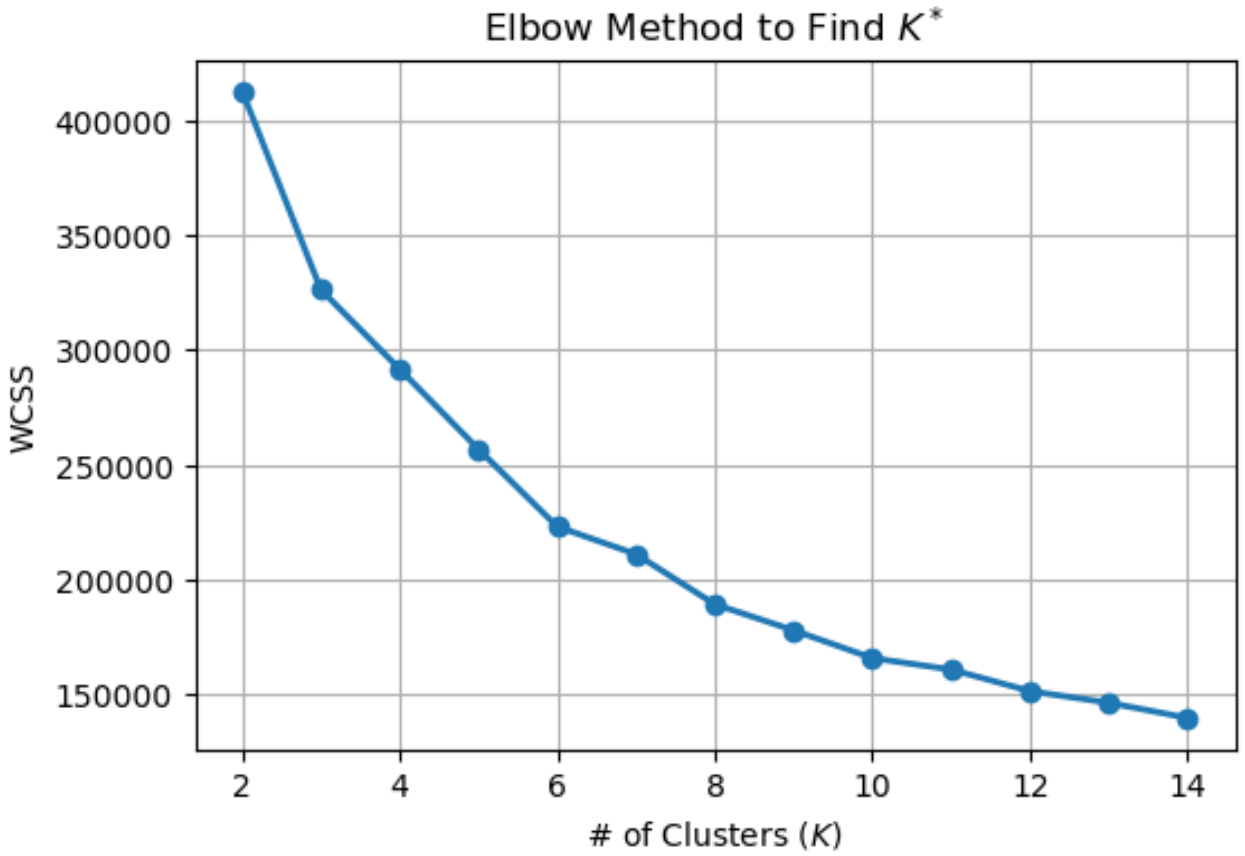
	urban	rural	uninhabited
1	0.0	0.6775210777005185	0.3224789222994815
2	0.0	0.6242062046490563	0.3757937953509436
3	0.4274177980824915	0.36323273920597127	0.20934946271153723
4	0.20468441350766625	0.533722133383501	0.26159345310883264
5	0.0	0.6904357909635832	0.3095642090364168
6	0.03281541780734993	0.7186757800460504	0.2485088021465997
7	0.3583752159635786	0.43313029682258825	0.20849448721383324
8	0.0	0.691072727580315	0.308927272419685
9	0.11511182025447335	0.6274431042626707	0.25744507548285595
10	0.027989022841773268	0.634872957142662	0.337138020015564 6
11	0.30279727001826395	0.45499695600628026	0.24220577397545578
12	0.2645235893684258	0.42328065030324746	0.3121957603283268
13	0.0	0.6250880884618332	0.3749119115381667
14	0.2230179481141095	0.4877008016971907	0.28928125018869977

Unsupervised Learning

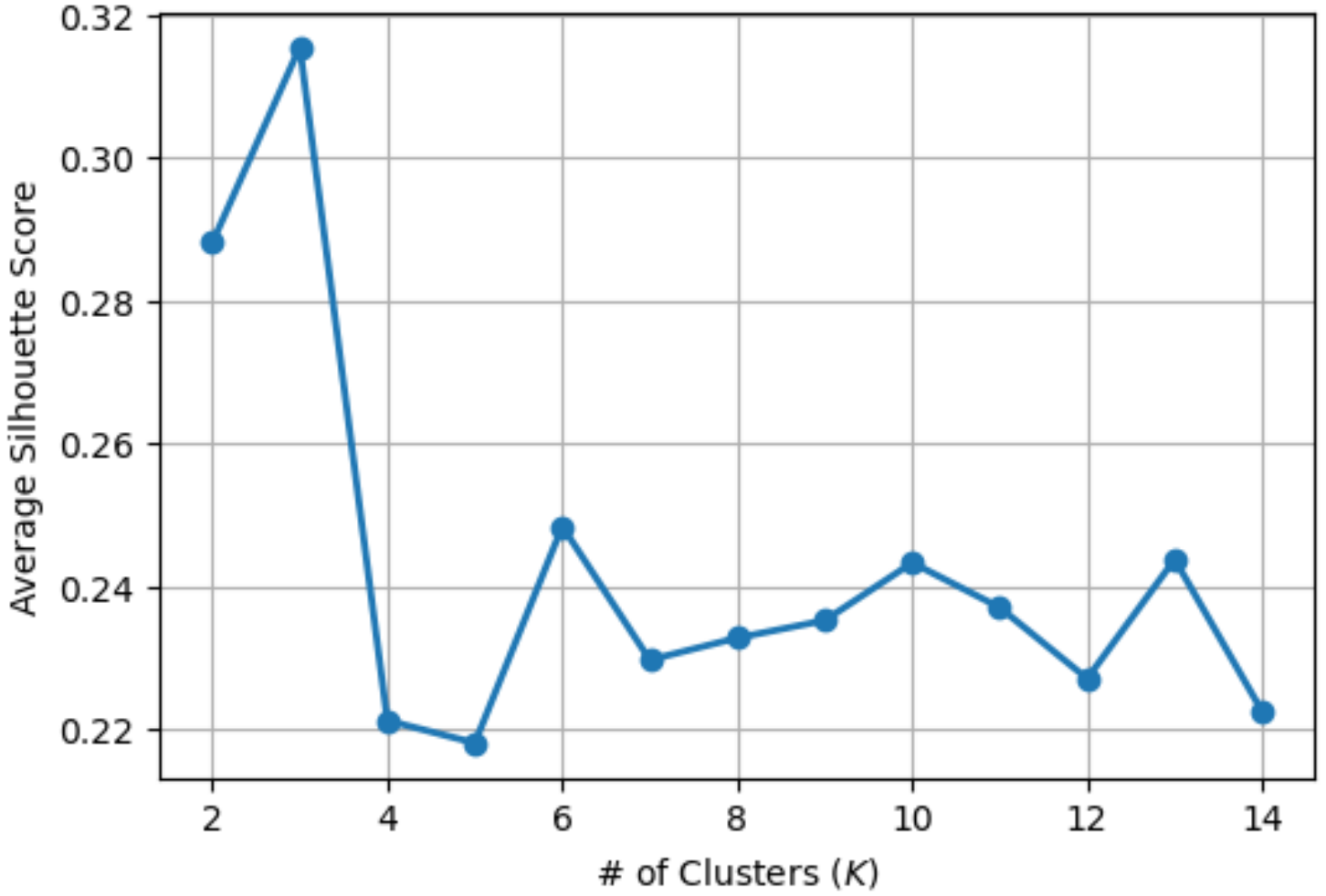
- *k*-means clustering was chosen as the unsupervised learning technique for this project.
- We select 19 distinct features to easily cluster the data based on certain visual aspects of the tile.
  - For example, urban tiles are noisier due to the number of structures. Uninhabited deserts should look smooth and homogeneous.
  - The features are the mean, standard deviation, 10^{th}, 50^{th}, 90^{th} percentiles of each RGB band, as well as some other statistics related to noise and variance of each patch.

Finding the Optimal *k*

- Qualitatively analyzing the elbow plot to the right, we see that at no point does the curvature reach a significant value.
  - In short, a smooth exponential-looking elbow plot tells us the data lacks sharp natural clusters, which is to be expected of homogeneous satellite imagery.

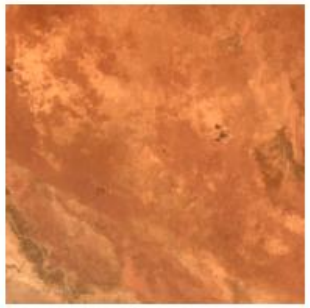


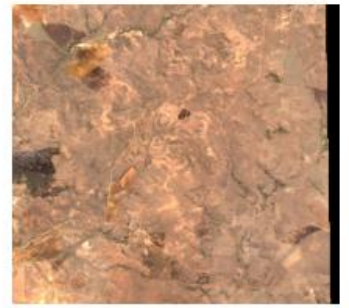
Mean Silhouette Score vs. *k*



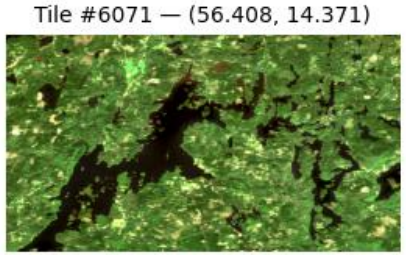
Selected Tiles for $k = 3$ Clusters

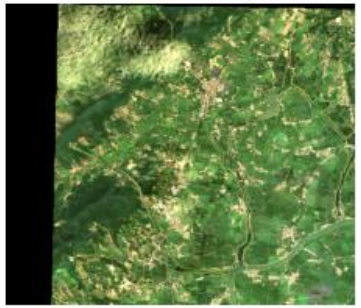
Cluster 0

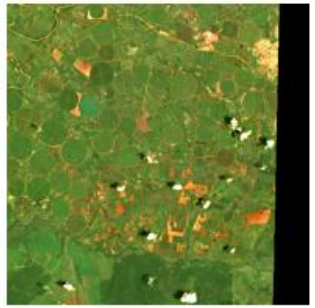




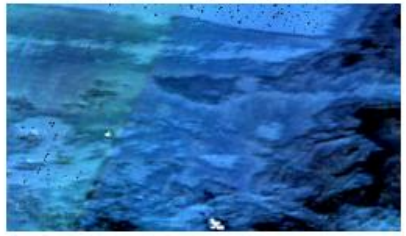
Cluster 1

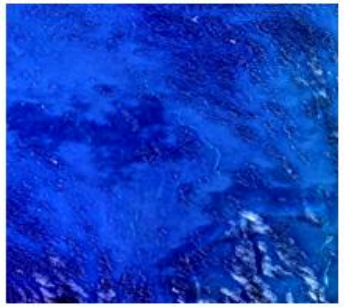


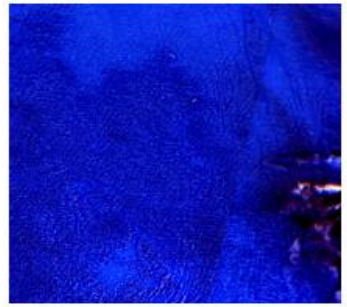




Cluster 2





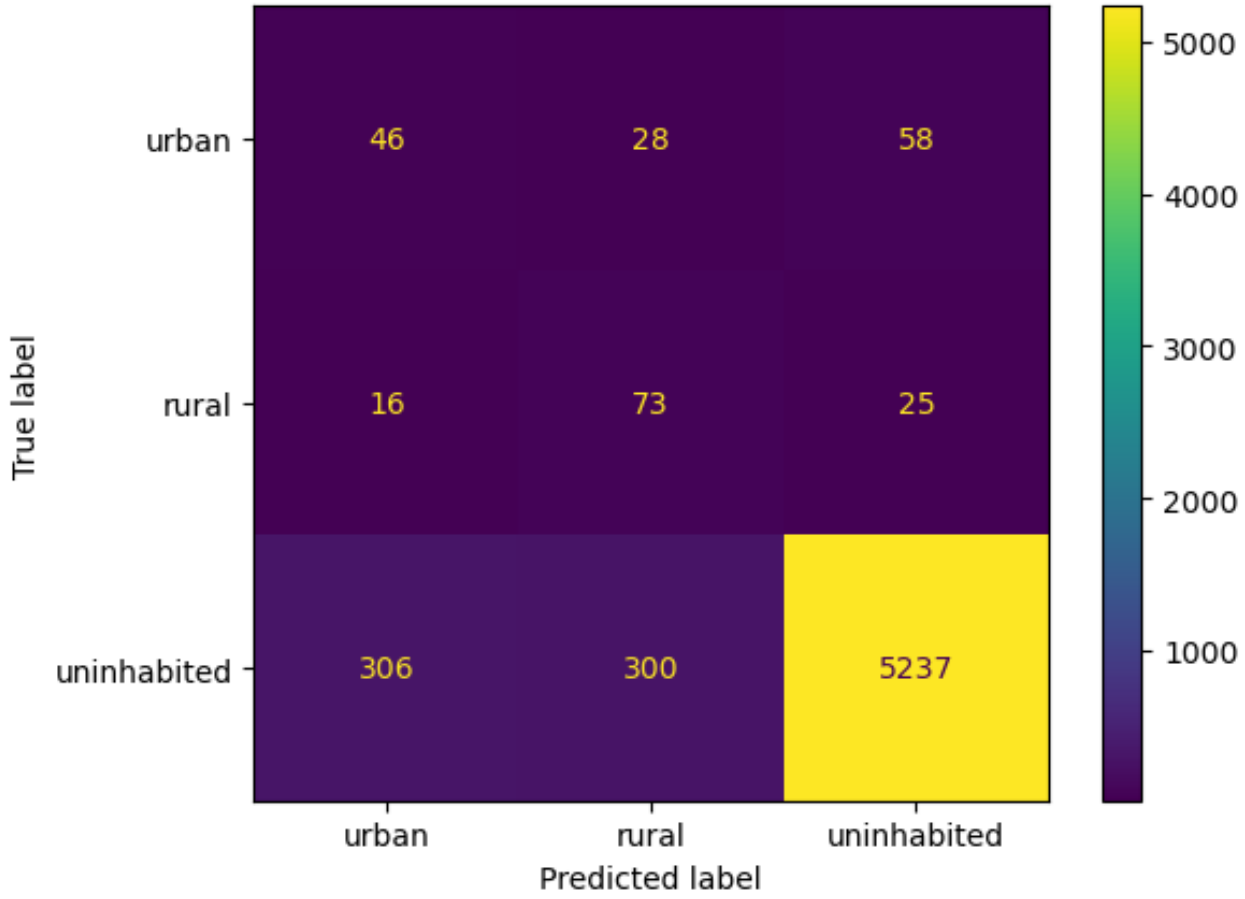


Shallow Supervised Learning

- The first supervised learning technique chosen was *gradient boosting*.
  - For each tile, we extract 12 features containing the mean, standard deviation, minimum, and maximum of each RGB band.
  - Due to extreme class imbalance, we also want to oversample minority classes during the training phase. (This will be true for all supervised learning models.)
- After performing a 7-fold cross validation on a grid of 432 candidates, the best hyperparameters are:
  - $\eta = 0.01$, $\alpha = 0$, $\lambda = 1$, *max*(depth) = 8
  - n_estimators = 800, colsample_bytree = 0.8, subsample = 0.8
- Best cross-validated log loss: 0.142

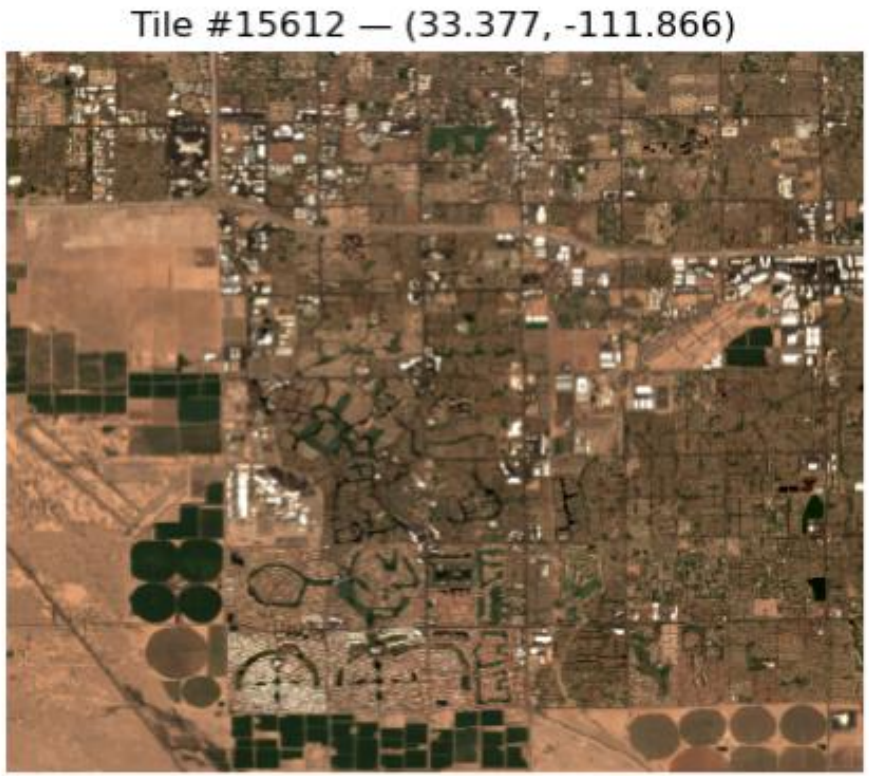
Gradient Boosting Evaluation

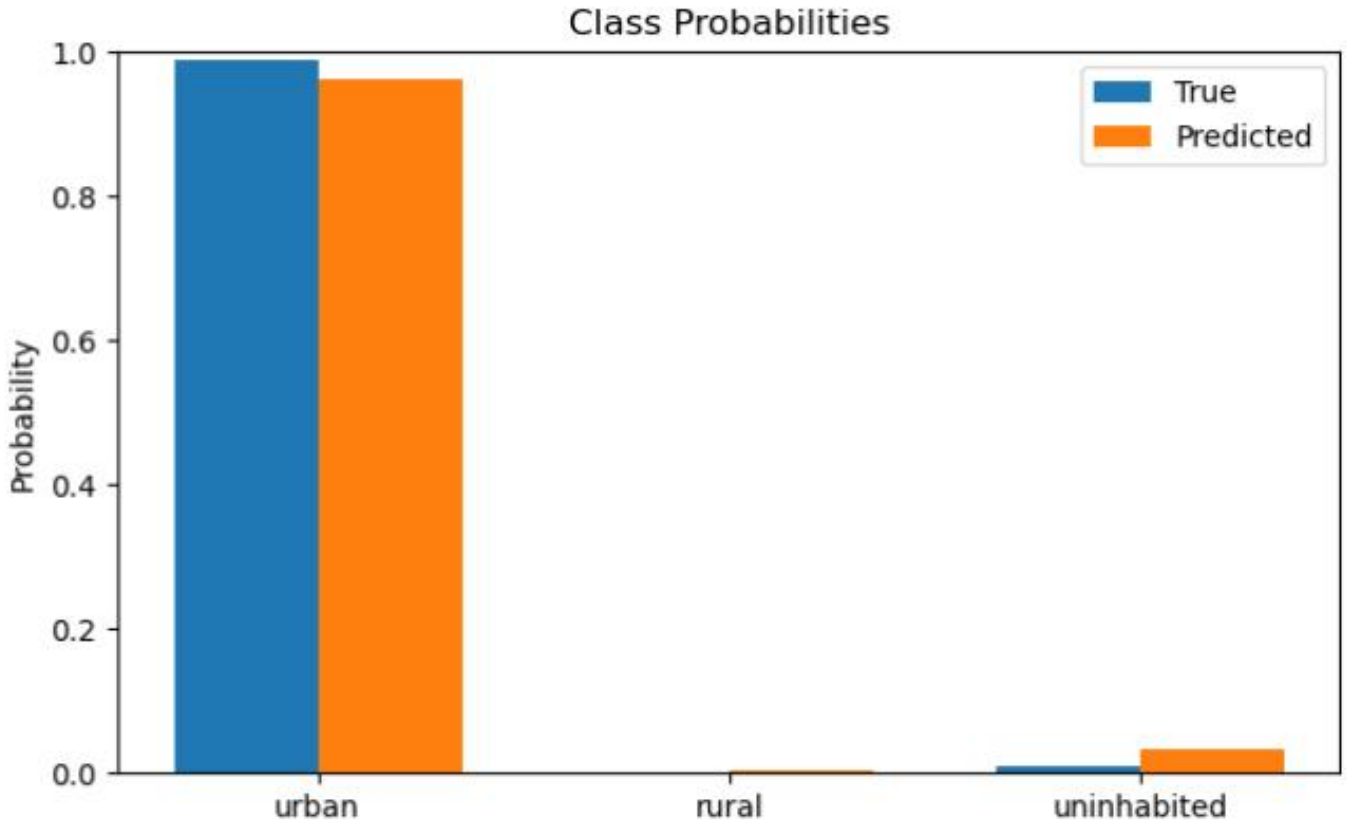
- Evaluation is done on an 80% to 20% train-test split.
- Selected metrics:
  - Cross Entropy = 0.2932
  - Accuracy = 0.8796
  - Macro Precision = 0.4305
  - Macro Recall = 0.6284
  - Macro F1 Score = 0.4686



Gradient Boosting Top Predictions - Urban

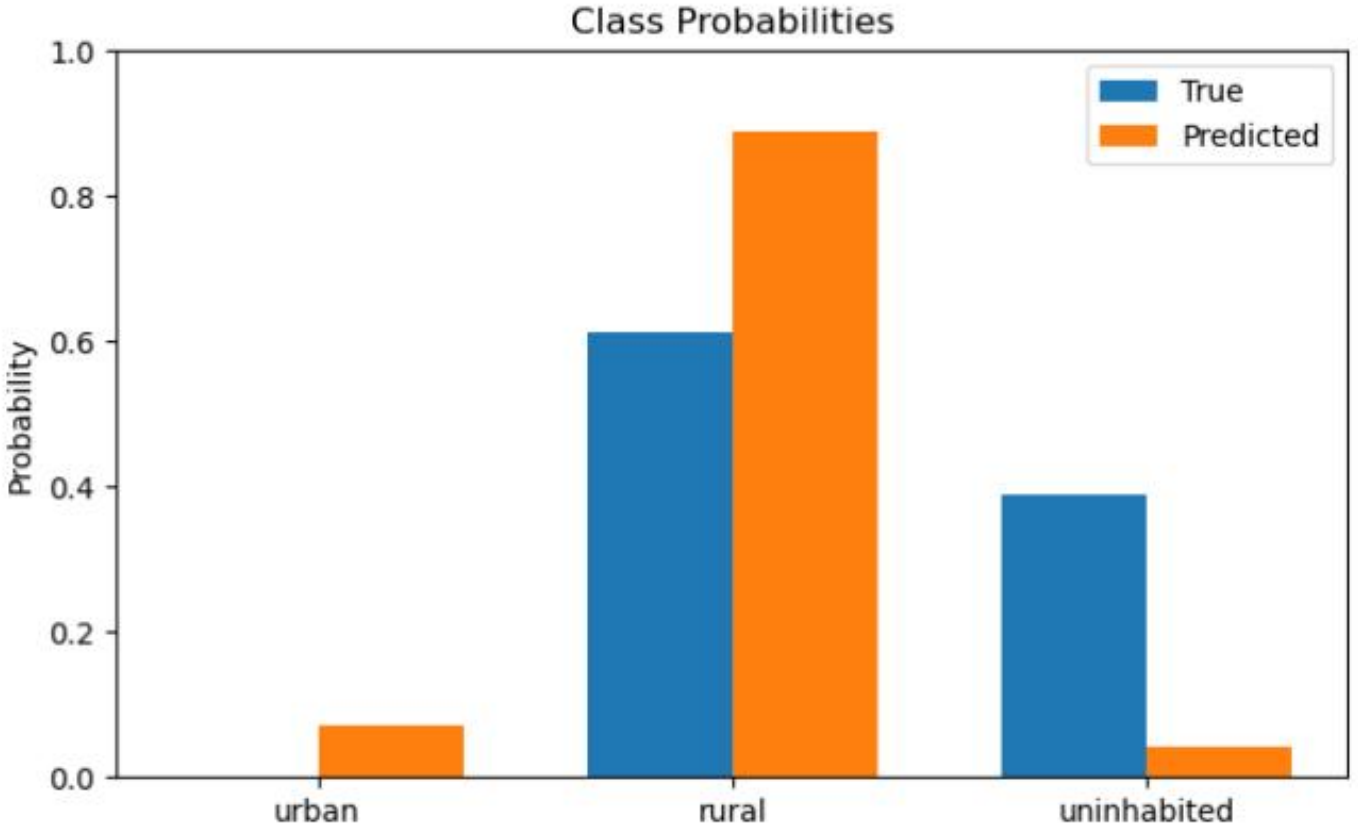
- Phoenix, Arizona, U.S. (https://maps.app.goo.gl/bt2JWSW8afXVU9fX7)





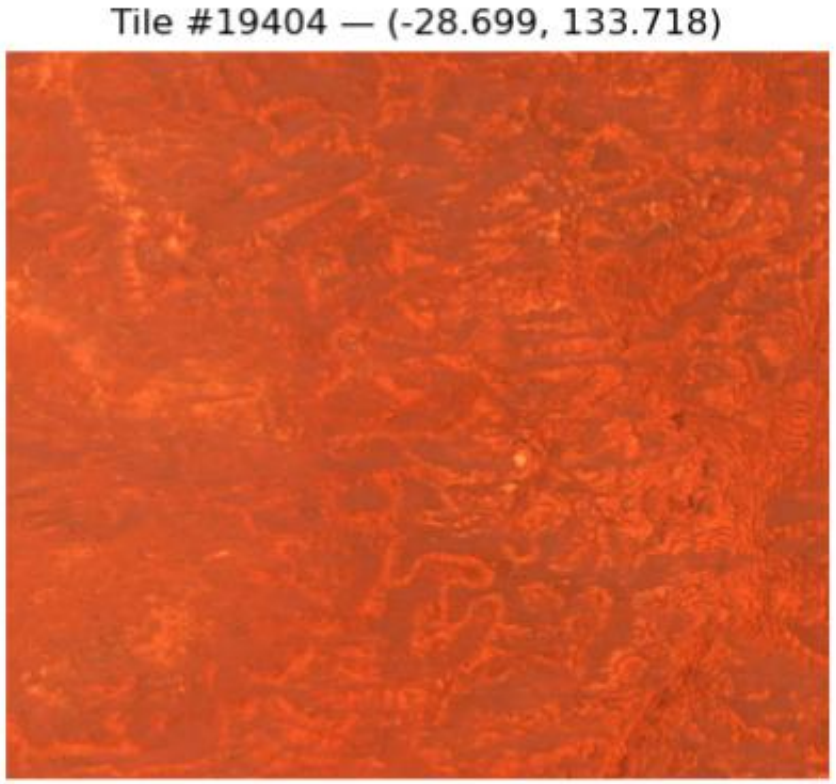
Gradient Boosting Top Predictions - Rural

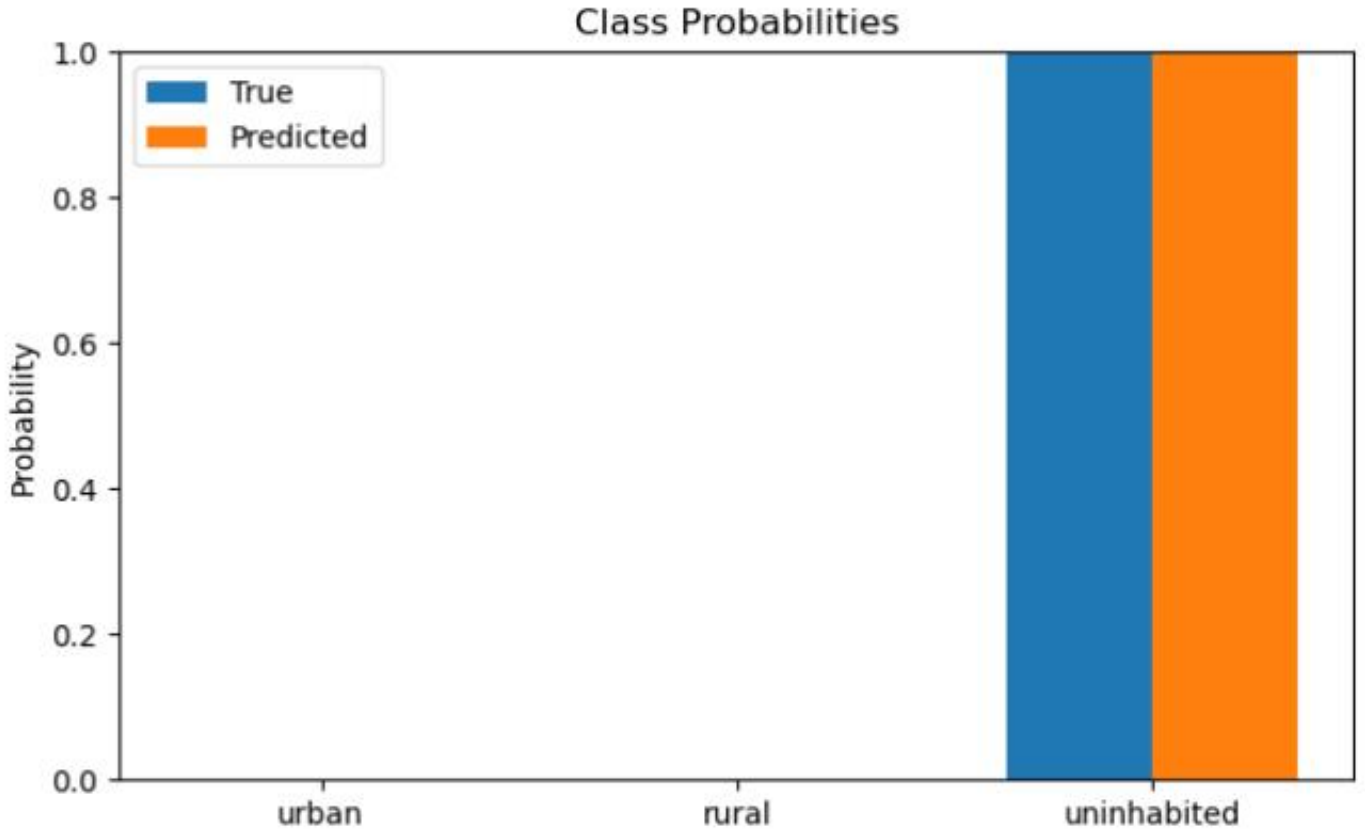
- Linying County, Luohe, Henan, China (https://maps.app.goo.gl/wnEjTqggmo27DeJr5)



Gradient Boosting Top Predictions - Uninhabited

- Tallaringa Conservation Park, South Australia (https://maps.app.goo.gl/ZJA3veYzmqNFn3Vs7)



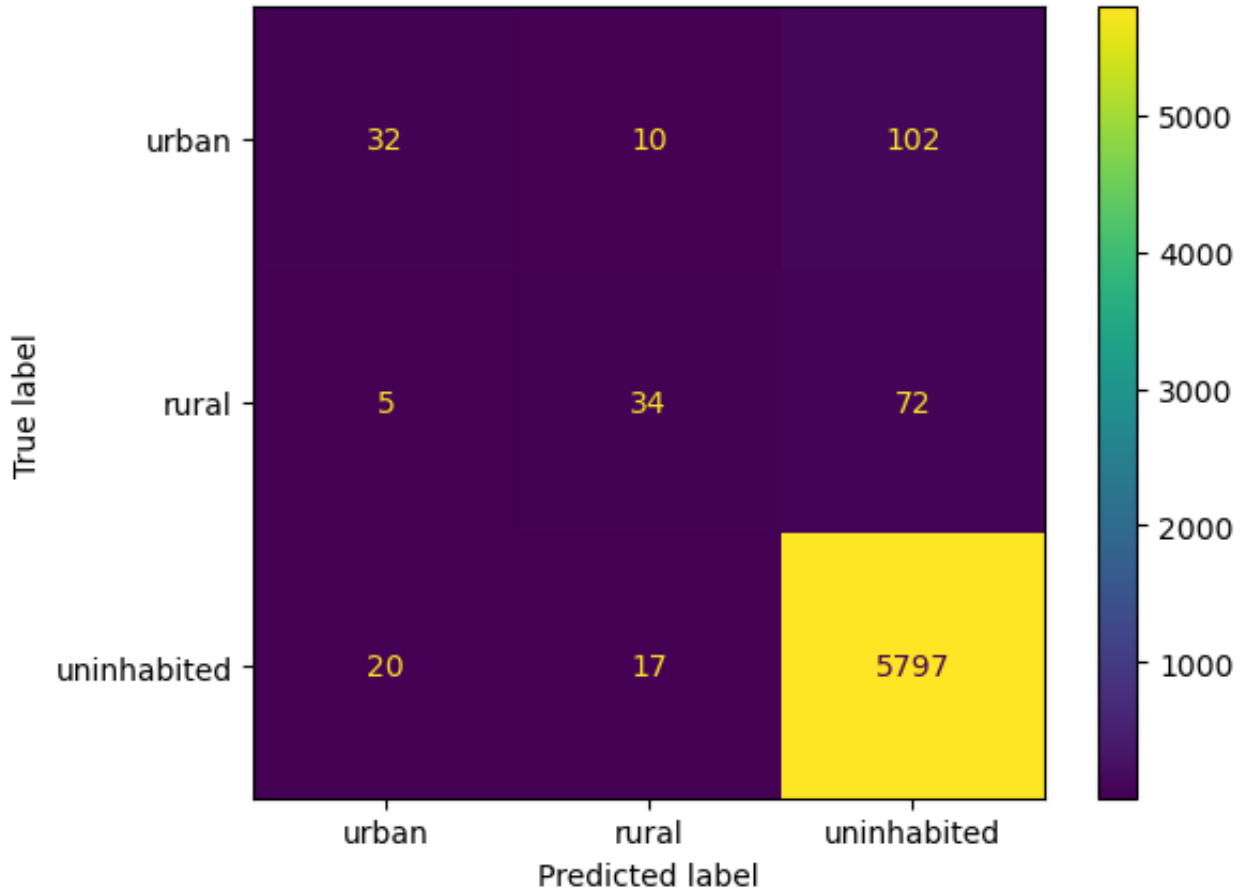


Deep Supervised Learning

- The second supervised learning technique chosen was using *convolutional neural networks*.
  - Due to hardware constraints and the nature of the problem, **ResNet-34** was chosen to be the CNN we will tune.
- He et al. (2015) stated that this type of residual network with 34 parameter layers has around 3.6 billion FLOPs. Given the theoretical peak of our GPU (~4.4 TFLOPS), we can do well over a thousand forward passes per second.
  - Even though real training will be much slower, this still gives us enough headroom to not suffer too much of a slowdown given our GPU.
- After performing a 5-fold cross validation on a grid of 18 candidates with 3 epochs each, the best hyperparameters are: $\eta = 0.0001$, $\lambda = 1e\text{-}05$, $p = 0.2$
  - Lowest validation loss = 0.3195

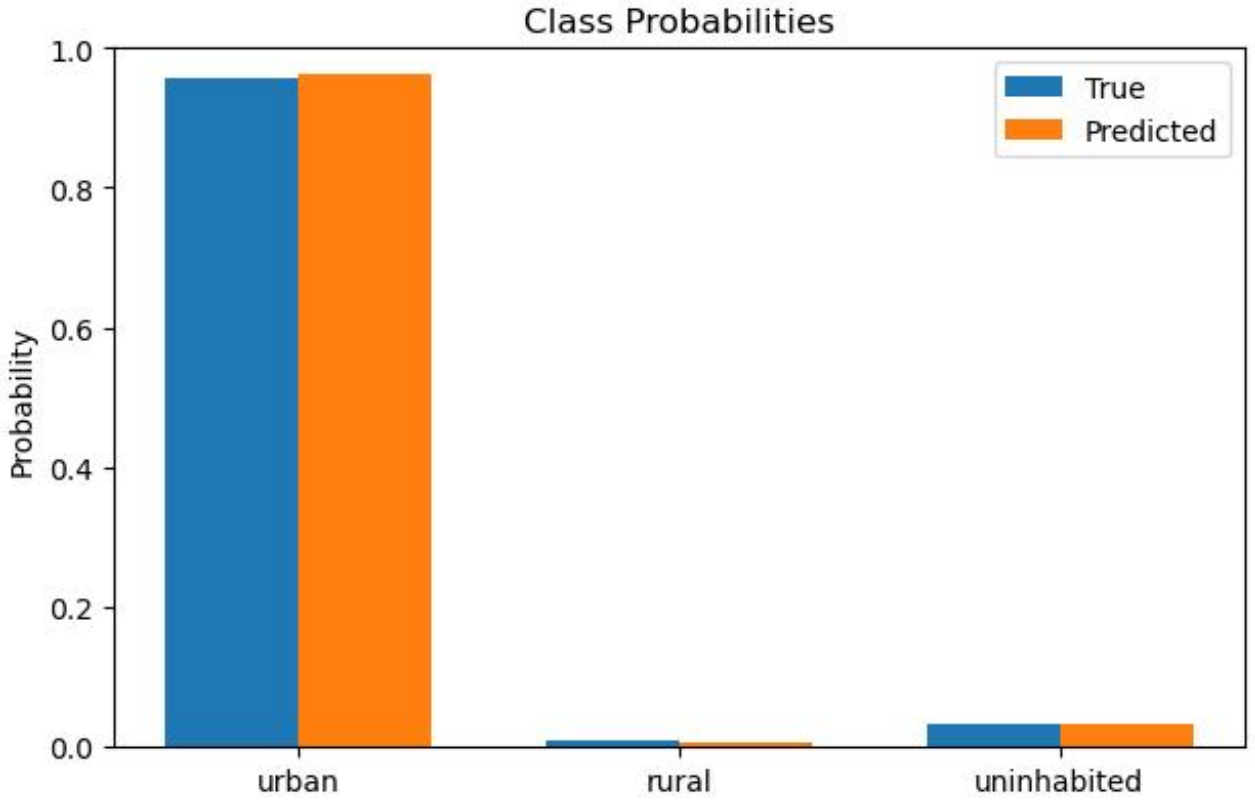
ResNet-34 CNN Evaluation

- Once again, evaluation is done on an 80%/20% train-test split.
  - The final model was trained over 30 epochs. (10x epochs of grid search)
- Selected metrics:
  - Final average validation loss = 0.3172
  - Accuracy = 0.9629
  - Macro Precision = 0.6965
  - Macro Recall = 0.5074
  - Macro F1 = 0.5653



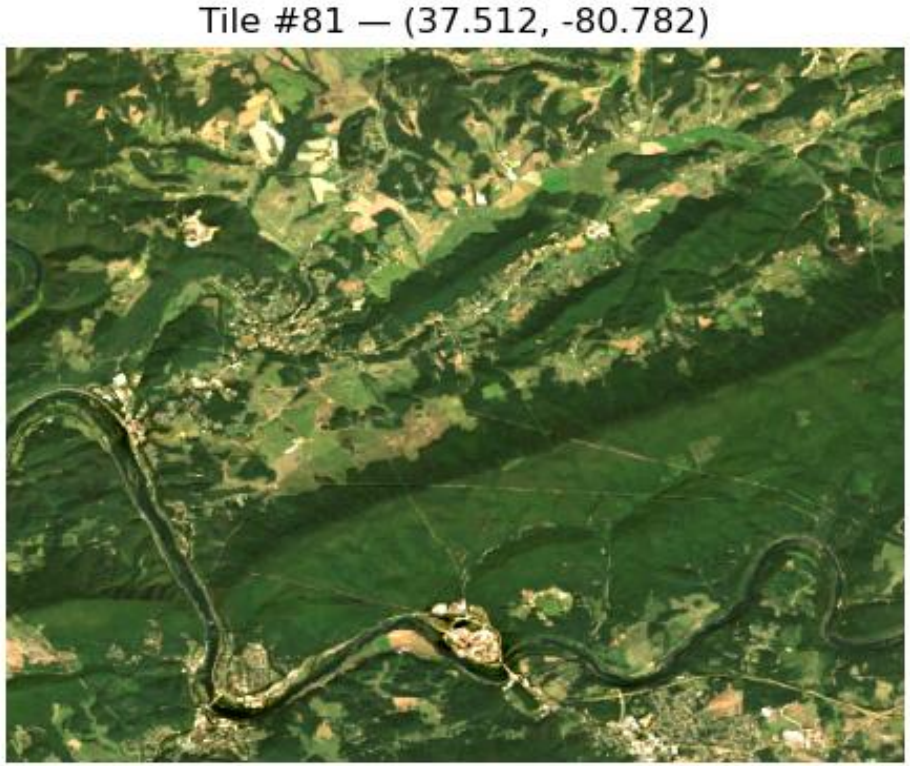
ResNet-34 CNN Top Predictions - Urban

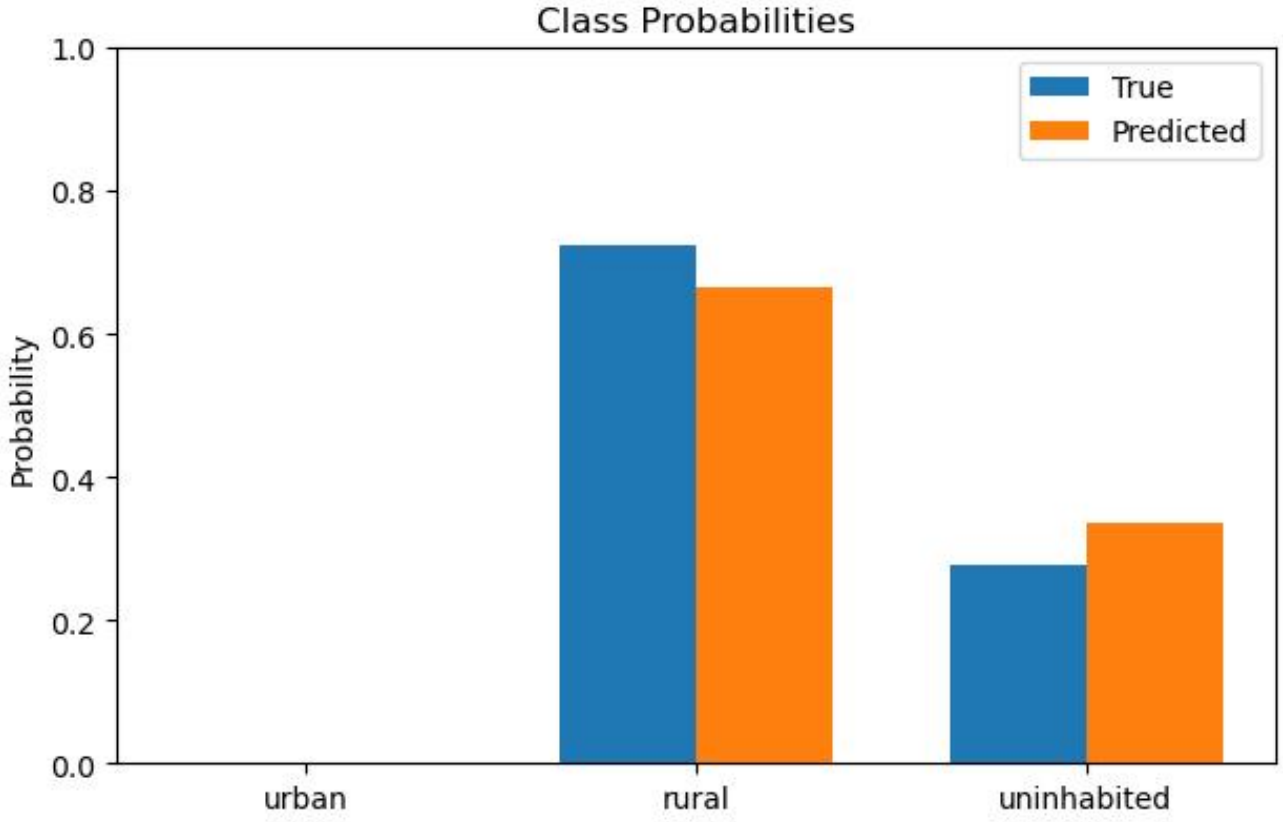
- Western Berlin, Germany (https://maps.app.goo.gl/pisL3rcDGvqDbPZm8)



ResNet-34 CNN Top Predictions - Rural

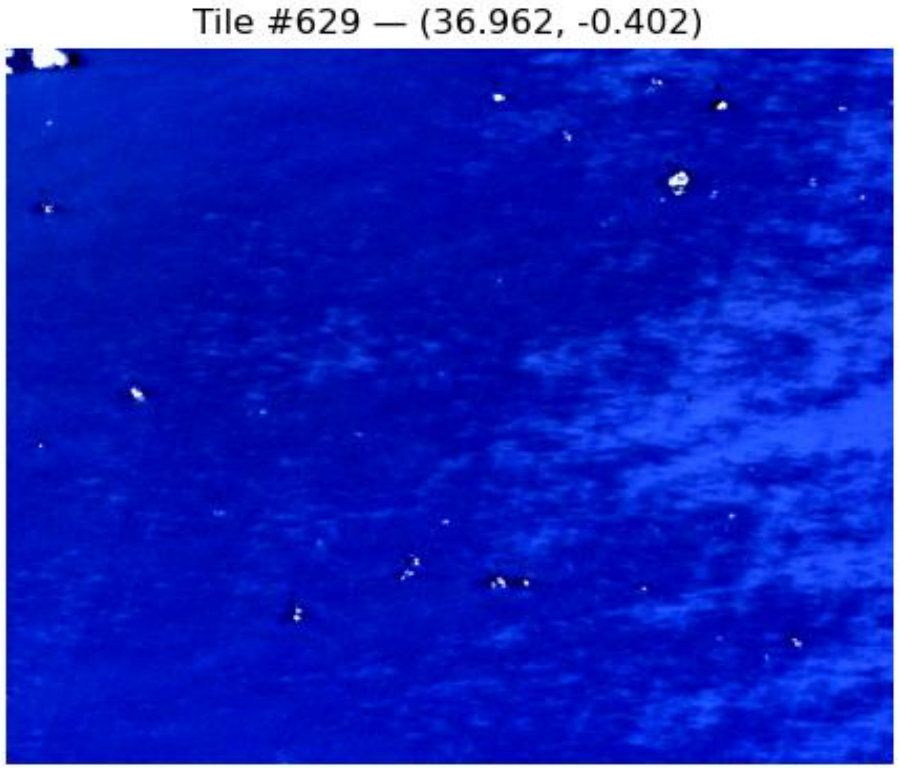
- Peterstown, WV, U.S. (https://maps.app.goo.gl/wA58TmNsWdNdjtct9)

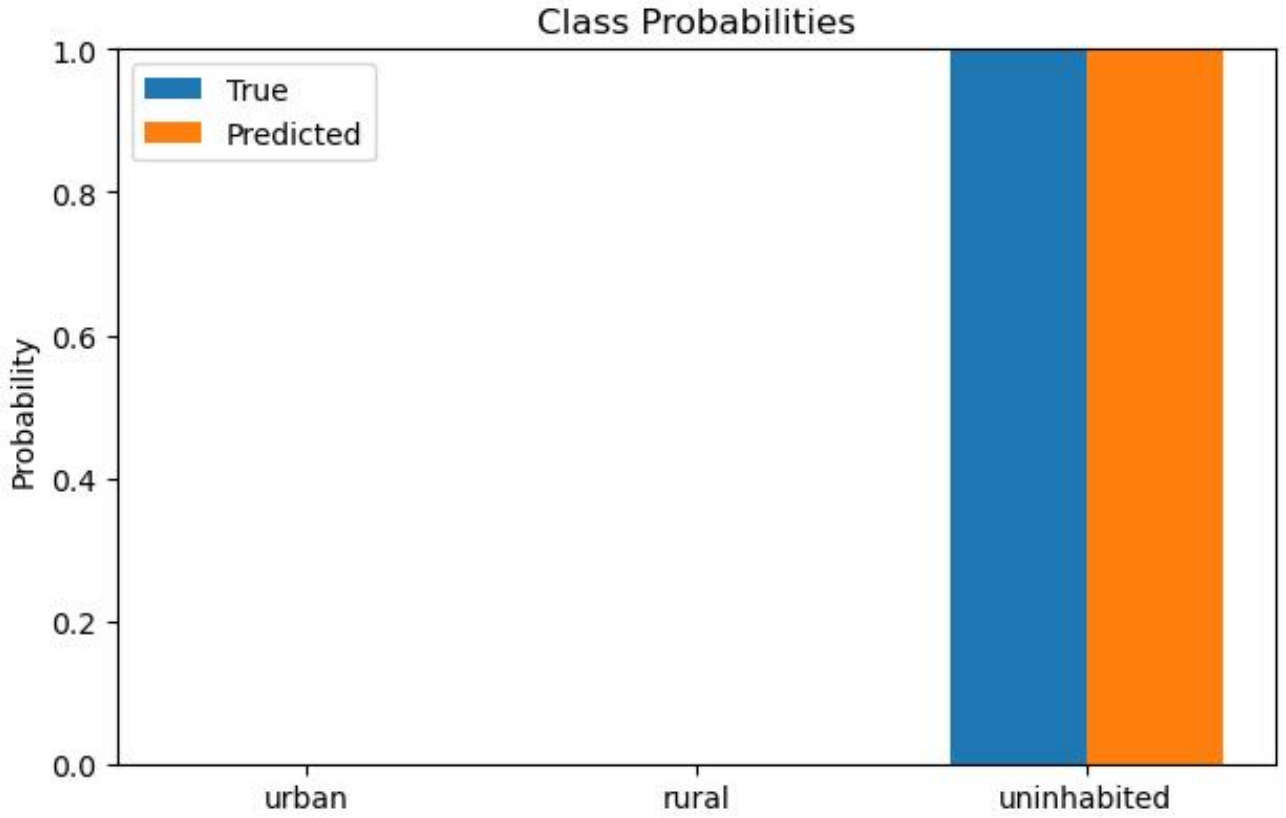




ResNet-34 CNN Top Predictions - Uninhabited

- Alboran Sea (https://maps.app.goo.gl/akkkrvEkJTL13TVT9)





Conclusion

- For unsupervised learning, *k*-means clustering surprisingly clustered more on color than noise.
  - This produced three distinct clusters representing green forestry/grassland, deserts, and bodies of water.
- For supervised learning methods, *gradient boosting* and *CNNs* performed much closer than expectation.
  - Looking at the qualitative top predictions of satellite photographs, gradient boosting made much more intuitive predictions (due to higher recall) than the ResNet-34 model.
  - **However,** the metrics indicate that ResNet-34 outperformed gradient boosting in both accuracy, macro precision and F1 score.
- Although ResNet-34 has more higher metrics than gradient boosting, the extreme class imbalance of the dataset makes recall critical.
  - Misclassifying a minority class (urban/rural) as uninhabited leads to loss of important human-settlement information.
  - In many remote sensing domains, missing urban/rural tiles is worse than misclassifying a few uninhabited tiles.
- Therefore, my final recommendation is to use **gradient boosting** considering the problem domain, metrics, and training efficiency.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. arXiv. https://doi.org/10.48550/arXiv.1512.03385

Liu, Z., Huang, S., Fang, C., Guan, L. & Liu, M. (2024). *Global urban and rural settlement dataset from 2000 to 2020*. *Scientific Data, 11*, Article 1359. https://doi.org/10.1038/s41597-024-04195-y

Microsoft. (n.d.). *Microsoft Planetary Computer* [Platform]. Retrieved November 23, 2025, from https://planetarycomputer.microsoft.com/explore