# EARTHSYS 142 Final Project Report

Weston Kirk

March 2025

## 1 Introduction

The Northern Blackland Prairie is a Level IV ecoregion classified by the US EPA, henceforth referred to as Region 32a, consistent with EPA terminology [1], representing the southern extreme of the Great Plains tallgrass prairie. Currently, only 4% of the original tallgrass prairie ecosystem spanning the Great Plains remains intact [2], with roughly 0.0004% of the Northern Blackland Prairie remaining, in other words–only 5000 of the original 12 million acres remain [3].

Figure 1: Blackland Prairie Remnant within Clymer Meadow Preserve, Hunt County, Texas. Wilafa, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

Aside from the many inherent reasons to protect these remaining remnants, preserving extant remnants of a once-larger prairie as blueprints for future restoration acts as a driving motivator. I propose using a pixel-based supervised learning technique to classify prairie remnants using Google Earth Engine (GEE). To do so, I evaluated four built-in machine algorithms (SVM, CART, Random Forest, and Naive Bayes) in Google Earth Engine (GEE) to determine

which is the most effective in supervised pixel-based classification of prairie remnants in the Northern Blackland Prairie of Texas, an endangered eco-region where less than 0.0004% of the original land area remains [3]. Based on previous literature, I predicted the following relative ranking of each algorithm,

1. **SVM** - Proven effectiveness in classifying high-dimensional data [4], and on crop-classification [5].
2. **Naive Bayes** - Given that we are using protected areas as a proxy for prairie remnants, being able to achieve high accuracy despite noise will prove effective.
3. **Random Forest** - Established success on a diverse array of data, though may struggle with the noise in this application compared to Naive Bayes and SVM.
4. **CART** - Similar concerns as Random Forest, along with misclassification tendencies [6].

For a more complete background, please see Final Project Proposal.

## 2 Methods + Data

To define the training area, I used 3 publicly available datasets, **(A)** USGS Land Use Land Cover (LULC) Data for the Continental United States for 2023 (Raster), **(B)** EPA Level IV Ecoregion Boundaries (Vector), **(C)** USGS Protected Lands of the U.S. dataset, (PAD-US) 4.0 (Vector). The pre-processed state of the data can be seen in Figure 2.

- **(A)** USGS Land Use Land Cover (LULC) Data for the Continental United States for 2023 (Raster)
- **(B)** EPA Level IV Ecoregion Boundaries (Vector)
- **(C)** USGS Protected Lands of the U.S. dataset (Vector)

I created two raster layers with 30 x 30 pixel resolution to define two training data classes, being **Remnant** and **Non-Remnant** prairie. To simplify the computation, I first filtered **(A)** to include only grassland [1] pixels, or Class Label `71` [7]. Formally, we defined each as:

- **Remnant:** **(A)** Class Label `71` + **(B)** Region 32a + **(C)** Inside of Protected Areas
- **Non-Remnant:** **(A)** Class Label `71` + **(B)** Region 32a + **(C)** Outside of Protected Areas

[1] Grassland/Herbaceous is defined as: "Grassland/Herbaceous- areas dominated by gramanoid or herbaceous vegetation, generally greater than 80% of total vegetation. These areas are not subject to intensive management such as tilling, but can be utilized for grazing." [7]

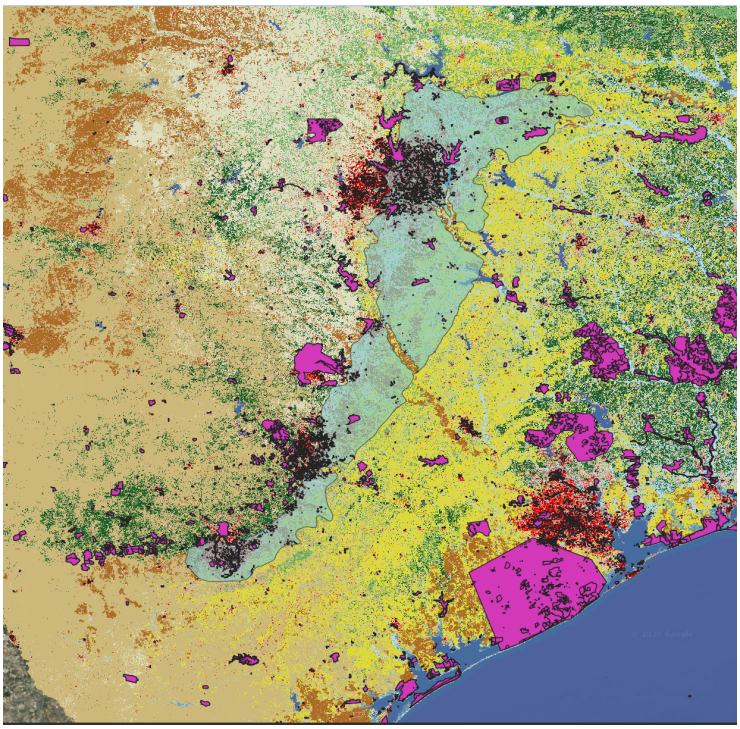

Figure 2: Layers **(A)**, **(B)**, and **(C)** before filtering methods.

Each class was created in QGIS using tools such as "Clip Raster By Map Layer" then exported as a .tif file to GEE so that pixels could be sampled. A high resolution delineation of the two classes can be seen in Figure 9 (Page 13).

It is worth noting that I opted to train the models within GEE, as opposed to exporting data to an external training environment using TensorFlow or PyTorch [8]. This is because I wanted to determine how each model behaves with a low amount of training data in GEE, as many users, especially those without the knowledge of external ML workflows, will be working within GEE. This, along with the fact that the focus of EARTHSYS 142 is getting experience working with GEE, provided reason to explore the limits of computation and machine learning within this framework. Inside the GEE environment, I sampled 500 remnant pixels and 500 non-remnant. For each pixel, the input training data consisted of Landsat 9 Spectral Bands 1-7, where per-pixel information can be

thought of as a 7-vector input to the algorithm:

$$\text{Band information for pixel } i = \begin{bmatrix} \text{Band 1} \\ \text{Band 2} \\ \text{Band 3} \\ \text{Band 4} \\ \text{Band 5} \\ \text{Band 6} \\ \text{Band 7} \end{bmatrix}$$

This selection was made because is encompasses data which is used to calculate known vegetation indices such as NDVI and EVI, removing the need for direct calculation. Furthermore, high resolution spectral data has been shown to be effective for similar classification tasks [9]. Once the training data set was created, the models were trained and then evaluated on a set of 500 pixels from each class, for a total of 1000, where all of the training pixels were removed as choices for the test pixels. Because the total number of pixels for the Non-Remnant class was on the magnitude of millions, and on the magnitude of hundred thousands for the Remnant class. The area for each class, along with a zoomed-in view for clarity, can be seen in Figure 9 (Page 13). Because of this, I down-sampled the data in GEE to ensure random spatial distribution across region 32a. I used built in GEE algorithms for each of the four, keeping all settings default, with the exception of Random Forest, which I set to have 100 trees.

# 3 Results

Using the test data, a confusion matrix was generated for each algorithm, as seen in Figures 3-6.
From these, we computed Overall Accuracy (OA), Cohens Kappa ($\kappa$) [10], and F1 Score [11], to provide more context beyond initial classification. These metrics can be seen in Figure 7.

Thus, based on OA, the final relative ranking is:

1. **Random Forest**
2. **SVM**
3. **Naive Bayes**
4. **CART**

Overall, Random Forest had the highest OA, followed closely by SVM, though SVM achieved a slightly higher F1 score, balancing recall and precision more effectively [11]. All algorithms achieved an OA of over 60%, matching my initial hypothesis.

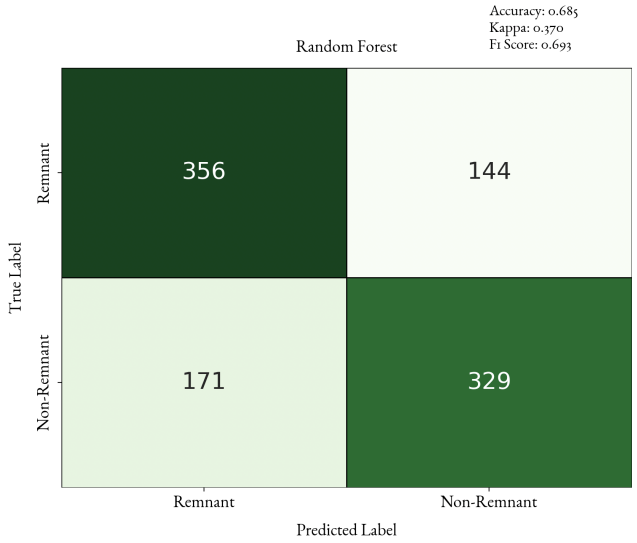


Figure 3: Confusion Matrix for Random Forest

One key motivation for this classification task was to identify areas with potential prairie remnants based on known locations. To do so, I looked at the results across each of the four algorithms, and selected points with True Label of Non-Remnant and Predicted Label of Remnant, as these represent the areas currently not known as prairie remnants, which may be unrealized remnants of ecotones unique to Region 32a. Since each algorithm evaluated the same pixels, I implemented this logic to see which pixels were classified in this way by multiple algorithms, which was done using Python:

This is due to the idea that if a pixel's predicted label is consistently remnant and its true label is non-remnant, there is a higher likelihood that it contains high-quality ecosystem. Each pixel is represented by a point (not to scale) to visualize the spatial extent of these potentially unrecognized remnant prairies, as seen in Figure 8 (Page 12):

# 4 Discussion

The results align with current conclusions in the literature. The original prediction relative ranking underestimated the accuracy of Random Forest compared to other algorithms. RF is used for a myriad of tasks, and performs well with high-dimensional data [12]. Furthermore, the notion that RF is more accurate than SVM in classification tasks, but not as robust [13] is supported by RF achieving a higher OA and $\kappa$ than SVM, while SVM achieved a higher F1 score (balance of precision and recall). SVM performing well aligned with its known performance on similar tasks [5] [14]. Naive Bayes achieved a lower overall ac-

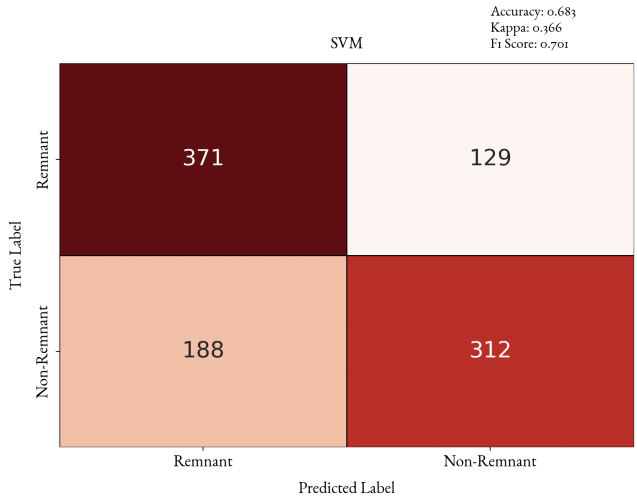


Figure 4: Confusion Matrix for SVM

curacy than SVM or Random Forest, perhaps because input data was limited to only 7 spectral bands, so the noise was not a major factor in this dataset [15]. GEE serves as an accessible tool for geospatial analysis [16]. Most large data projects take place outside the GEE environment, but it is important to understand how known algorithms perform with limited training data and computing power, as many users will rely on these algorithms for similar tasks. The nearest neighbor paper for this work [17], used 10000 training pixels for their model, and though they used different methods to measure accuracy, achieved accuracies of $58 - 66\%$, compared to our model, which used only 1000 training pixels, and achieved classification accuracies of $61 - 69\%$. Thus, with only one-tenth of the training data, I achieved similar OA with built in algorithms in GEE set to their respective defaults, which is a notable result. Furthermore, the results met my initial hypothesis that all 4 algorithms would achieve greater than 60% OA rate, consistent with current literature [18]. As for the potential unidentified remnants, connecting local experts, regional and local governments, and state agencies with the results represents a natural next step. This way, additional surveys and measures to potentially conserve the land could be implemented in the case it were to be developed. Also, pixels currently on public land, or land where the landowner is willing to establish a conservation easement are targets for immediate conservation action. Looking forward, combining approaches such as using ensemble algorithms [19], represents a future direction of work. With regard to applications, working with tools such as PyTorch or TensorFlow, with much larger datasets, could allow for more accurate classification, alongside measures such as parameter tuning. Furthermore, once refined, these methods could be generalized to adapt to input data from any ecoregion or ecotone, depending on the spectral data characteristics and signatures, as a way to locate unknown habitat fragments or remnants. I am considering looking

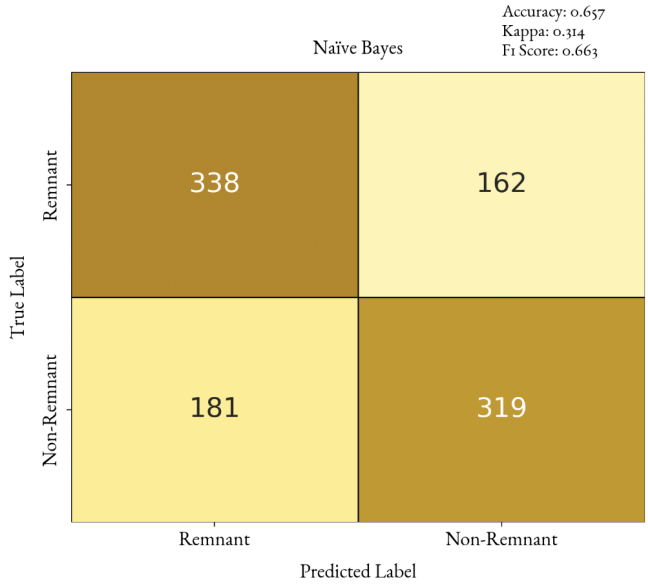


Figure 5: Confusion Matrix for Naive Bayes

into a generalized model as part of a senior honors thesis. Potential limitations to this study include the lack of a "ground-truth" for comparison, as we used protected areas as a proxy for prairie remnants.

# 5 Conclusion

I determined that Random Forest and SVM are the most effective built-in GEE algorithms of the four tested for pixel-based supervised learning methods. All models demonstrated over 60% OA on a relatively small training dataset (1000 pixels), and the results were used to identify potential remnant prairies currently classed as non-remnant prairies. Future work includes expanding to larger datasets, ensemble learning implementations, and generalizing this workflow to other ecoregions.

# 6 Data Sources

**(A)** U.S. Geological Survey, 2023, Annual National Land Cover Database (NLCD) Collection 1 Land Cover Conterminous United States - 2023(published 20241016), accessed at March 7, 2025 at `https://www.mrlc.gov/data?f%5B0%5D=project_tax_term_term_parents_tax_term_name%3AAnnual%20NLCD`.
**(B)** U.S. EPA, 2013, US Level IV Ecoregions shapefile without state boundaries (65 mb) - 2013 (published 20130416), accessed at March 7, 2025 at `https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states`

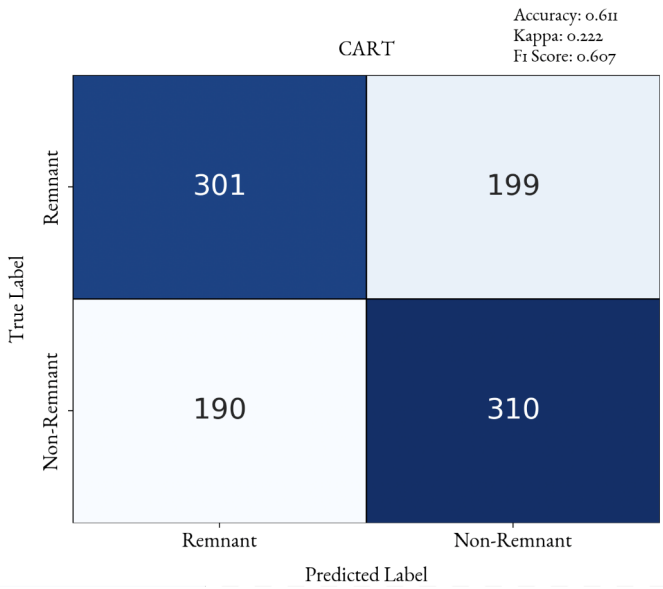


Figure 6: Confusion Matrix for CART

| | SVM | Random Forest | CART | Naïve Bayes |
|---|---|---|---|---|
| **OA** | 0.683 | 0.685 | 0.611 | 0.657 |
| **Kappa** | 0.366 | 0.370 | 0.222 | 0.314 |
| **F1** | 0.701 | 0.691 | 0.607 | 0.663 |

Figure 7: Performance metrics for different classification models.

**(C)** U.S. Geological Survey (USGS) Gap Analysis Project (GAP), 2024, Protected Areas Database of the United States (PAD-US) 4.0: U.S. Geological Survey data release, `https://doi.org/10.5066/P96WBCHS`
**(D)** Earth Resources Observation and Science (EROS) Center. (2020). Landsat 8-9 Operational Land Imager / Thermal Infrared Sensor Level-2, Collection 2 [dataset]. U.S. Geological Survey. `https://doi.org/10.5066/P9OGBGM6`

# References

[1] United States Environmental Protection Agency (EPA). Level iii and iv ecoregions of the continental united states, 2025. Accessed: 2025-01-28.

[2] National Park Service. Tallgrass prairie national preserve, 2025. Accessed: 2025-01-27.

**Algorithm 1** Counting Misclassified Pixels

**Input:** Four classification results $\mathcal{D} = \{D_1, D_2, D_3, D_4\}$, where each $D_i$ is one of the four algorithms (RF, SVM, CART, Naive Bayes) which contains pixel data with true label $y$ and predicted label $\hat{y}$
**Output:** A dataset of Non-Remnant pixels misclassified Remnant pixels along with the number of models misclassifying the pixel in this way

Initialize an empty dictionary $\mathcal{M}$ *Stores misclassified pixels + counts*
**for** each dataset $D_i$ in $\mathcal{D}$ **do**
  **for** each pixel $p \in D_i$ **do**
    **if** $y_p$ = Non-Remnant and $\hat{y}_p$ = Remnant **then**
      **if** $p \notin \mathcal{M}$ **then**
        $\mathcal{M}[p] \leftarrow 1$ *First time misclassified*
      **else**
        $\mathcal{M}[p] \leftarrow \mathcal{M}[p] + 1$ *Increment count*
  **Return** $\mathcal{M}$ *Final misclassified pixel dataset*


[3] Texas Parks and Wildlife Department. *Texas Conservation Action Plan 2012–2016: Texas Blackland Prairies Handbook*. Texas Parks and Wildlife Department, Austin, Texas, 2012. Texas Conservation Action Plan Coordinator.

[4] Reza Khatami, Giorgos Mountrakis, and Stephen V Stehman. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote sensing of environment*, 177:89–100, 2016.

[5] Mohamad Awad. Google earth engine (gee) cloud computing based crop classification using radar, optical images and support vector machine algorithm (svm). In *2021 IEEE 3rd International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 71–76. IEEE, 2021.

[6] Liliane Bel, Denis Allard, Jean-Marie Laurent, Rachid Cheddadi, and Avner Bar-Hen. Cart algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics & Data Analysis*, 53(8):3082–3093, 2009.

[7] Multi-Resolution Land Characteristics (MRLC) Consortium. National Land Cover Database (NLCD) Class Legend and Description, 2023. [Accessed: YYYY-MM-DD].

[8] Google Earth Engine Developers. Classification in google earth engine, 2025. Accessed: 2025-01-28.

[9] Xiaozhi Yu, Dengsheng Lu, Xiandie Jiang, Guiying Li, Yaoliang Chen, Dengqiu Li, and Erxue Chen. Examining the roles of spectral, spatial, and

topographic features in improving land-cover and forest classifications in a subtropical region. *Remote Sensing*, 12(18):2907, 2020.

[10] Giles M Foody. Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification. *Remote sensing of environment*, 239:111630, 2020.

[11] Aaron E Maxwell, Timothy A Warner, and Luis Andrés Guillén. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 1: Literature review. *Remote Sensing*, 13(13):2450, 2021.

[12] Thanh Noi Phan, Verena Kuch, and Lukas W Lehnert. Land cover classification using google earth engine and random forest classifier—the role of image composition. *Remote Sensing*, 12(15):2411, 2020.

[13] Victor F Rodriguez-Galiano and Mario Chica-Rivas. Evaluation of different machine learning methods for land cover mapping of a mediterranean area using multi-seasonal landsat images and digital terrain models. *International Journal of Digital Earth*, 7(6):492–509, 2014.

[14] Dee Shi and Xiaojun Yang. Support vector machines for land cover mapping from remote sensor imagery. *Monitoring and Modeling of Global Changes: A Geomatics Perspective*, pages 265–279, 2015.

[15] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714, 2010.

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

[17] Jane M Kunberger, Brian S Early, Csanyi EL Matusicky, and Ashley M Long. Using remotely sensed data to identify coastal prairie remnants in louisiana. *Natural Areas Journal*, 44(3):190–195, 2024.

[18] Timothy J Assal and Jeffrey A Lockwood. Utilizing remote sensing and gis to detect prairie dog colonies. *Rangeland Ecology & Management*, 60(1):45–53, 2007.

[19] Yuzhen Zhang, Jingjing Liu, and Wenjuan Shen. A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, 12(17):8654, 2022.

# 7 Bibliography Notes, Initial Source Code, Use of LLMs

The initial framework for the GEE Scripts for the model training and evaluation was obtained from this guide (`https://developers.google.com/earth-engine/`

guides/machine-learning). We used LLMs as a translation tool to handle some complex syntax in Google Earth Engine, which was specified as ok via announcements on the first day of class. LLMs were used only for translation/syntax purposes, not for conceptual reasoning/ideation.

# 8 Supplemental Materials

**GEE Scripts:** https://code.earthengine.google.com/dd5379a22cadb9c5e6e72bddfd718af5?accept_repo=users%2Fstacemaples%2FSGC-EE101
https://code.earthengine.google.com/d9ae33f7b3741b028e49445b19977289?accept_repo=users%2Fstacemaples%2FSGC-EE101
**Presentation Slides from 3/11:** https://docs.google.com/presentation/d/1MbwwnmqcnWXpMFMw-5vWPlSGlzHVyiBLrReUlTq4_Fg/edit?usp=sharing

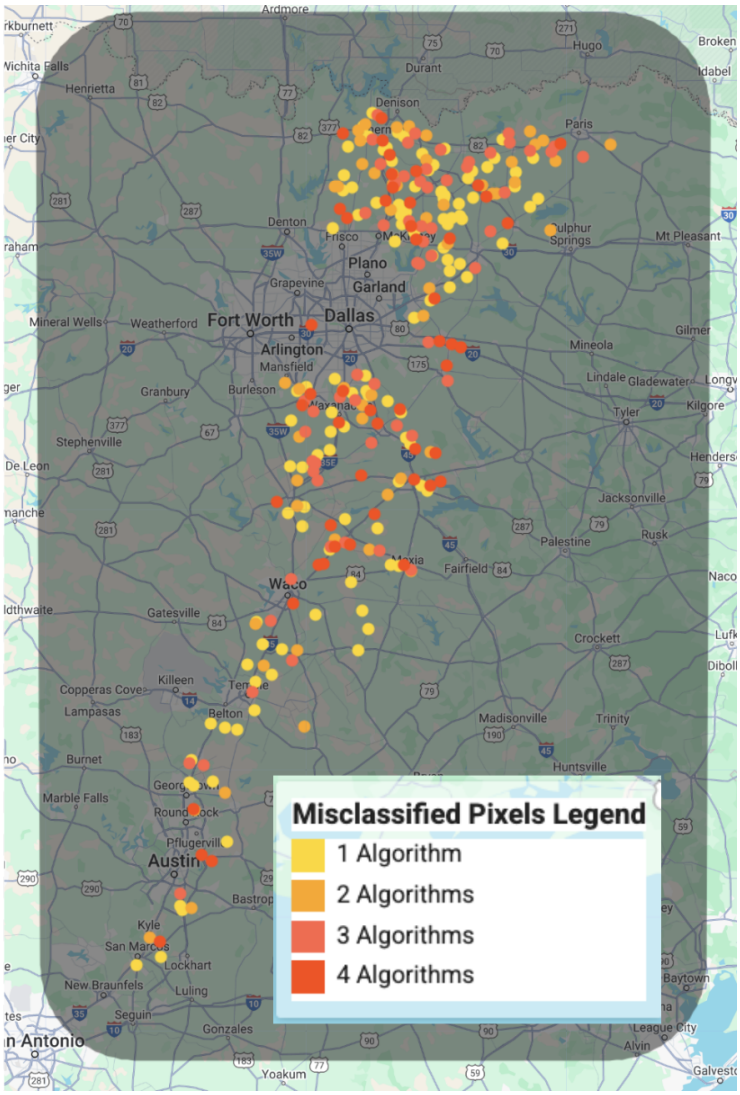


Figure 8: Spatial Distribution of Potential Unidentified Remnants in Region 32a


Figure 9: Remnant Areas in green and Non-Remnant Areas in dark tan.