

Abstract: As used in renewable energy storage, lithium is vital to the global green energy transition. However, traditional exploration methods to find lithium deposits are slow and inefficient. This project introduces a machine-learning approach to accelerate lithium discovery. The work utilizes multi-band satellite imagery taken from four different datasets (**Landsat 8, Sentinel-1, Sentinel-2, NASA SRTM**) to extract and combine various features. The model uses 20 different spectral and geological markers to predict lithium deposits across varied terrains with 89% accuracy. To make it reliable, the algorithm was optimized to reduce classification biases associated with training and validation data in machine learning. This method not only makes lithium exploration faster but can also provide a common framework for discovering other critical or rare-earth minerals.

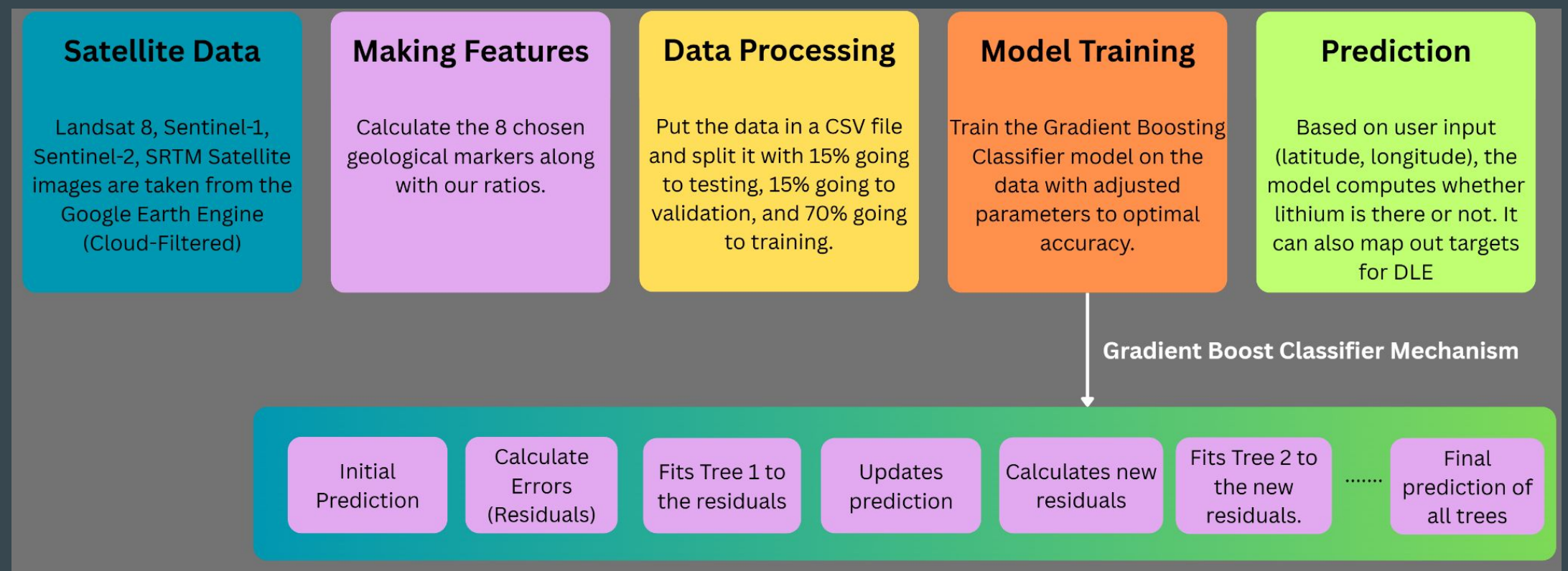
Lithium & importance for the Green Energy Transition

- ## Lithium Exploration Traditional Vs Remote Sensing Method

**AI Generated Image, Gemini 2026*

- ## Proposed Solution & Design Steps

A Machine Learning model trained to find lithium deposits (Brine and Pegmatites) based off 20 different satellite imaging based markers



Domain Research, Problem, solution space study for Remote sensing

- Research about lithium, challenges with exploration, and satellite imaging based markers that can help detect lithium.

Data source for Model training, and Feature Extractions

- Using USGS and MRDS datasets.
- Using Google Earth Engine APIs to get satellite data-sets.
- Extract features from those data-sets and combining them.

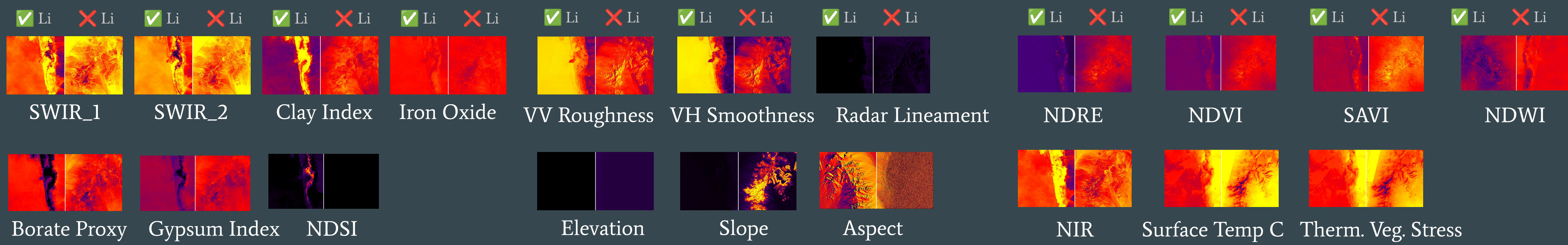
Model Development & Optimization

- Built initial machine learning model for lithium detection and classification.
- Applied advanced algorithms (XGBoost, Random Forest, CatBoost, etc) to boost predictive power.
- Analyzed data to reduce data biases and training inaccuracy, that helped final model to achieve ~89% accuracy.

Satellite Imaging Markers: Spectral Mineralogy

- 1) **SWIR_1 (Short-Wave Infrared 1):** This measures silica content, Granites & Pegmatites are high in silica.
- 2) **SWIR_2 (Short-Wave Infrared 2):** This detects Aluminum rich rocks where lithium resides.
- 3) **Clay Index:** Lithium Pegmatites are often surrounded by clay minerals like Kaolinite and Illite.
- 4) **Iron Oxide:** Lithium Pegmatites barely have any iron so this marker helps ignore iron-rich volcanic rocks
- 5) **Borate Proxy:** This targets minerals containing borate such as Tourmaline which are there with lithium-containing pegmatite dike.
- 6) **Gypsum Index:** This identifies sulfate-rich alteration zones that occur at the edges of mineralized granites.
- 7) **NDSI (Normalized Difference Snow Index):** In rock models, this identifies highly reflective, bright mineral surfaces that lack vegetation

**[Intensity Images created by the researcher with markers data extracted using Google Earth Engine API from various satellites]*



Satellite Imaging Marker: Physical Shape

- 8.) **Radar VV Roughness:** finds jagged pegmatite ridges.
- 9.) **Radar VH Smoothness:** This helps see through light vegetation to feel the actual shape of the rock dikes.
- 10.) **Radar Lineament Ratio:** This marker highlights fault lines and tectonic fractures where the lithium magma was originally injected to find Pegmatites.
- 11.) **Elevation:** Pegmatites are “Erosion Resistant”, they do not wear down as fast as softer rocks do, due to this they often form the backbone of ridges or small hills.
- 12.) **Slope:** Dikes are easily found on steep slopes where the dirt is washed away, leaving the “outcrop” exposed.
- 13.) **Aspect:** This is a correction marker, it’s used to find which way of the mountain faces to avoid confusion by the shadows on the dark sides of mountains or hills.

ected using Google Earth Engine API from various satellites]

Bio-geochemistry / Thermal infrared Markers

- 14) **NDRE (Normalized Difference Red Edge):** When plants are stressed by lithium in their roots, their chlorophyll leaks and the red edge shifts.
- 15) **NDVI (Normalized Difference Vegetation Index):** This maps the plants. Pegmatites are usually barren as nothing can grow on a slab of spodumene granite.
- 16) **SAVI (Soil Adjusted Vegetation Index):** A specialized version of NDVI which filters out desert glare.
- 17) **NDWI (Normalized Difference Water Index):** Pegmatites are non porous and dry. This marker helps distinguish between wet clays and Pegmatites
- 18) **NIR (Near Infrared):** This detects the landscape, provides contrast to separate forests from rocks
- 19) **Surface Temp C:** Measures thermal inertia, solid pegmatite has high thermal inertia relative to topsoil.
- 20.) **Thermal Vegetation Stress:** It identifies areas where the ground is too hot or mineralized for healthy plant growth and is indicative of shallow mineral deposit.

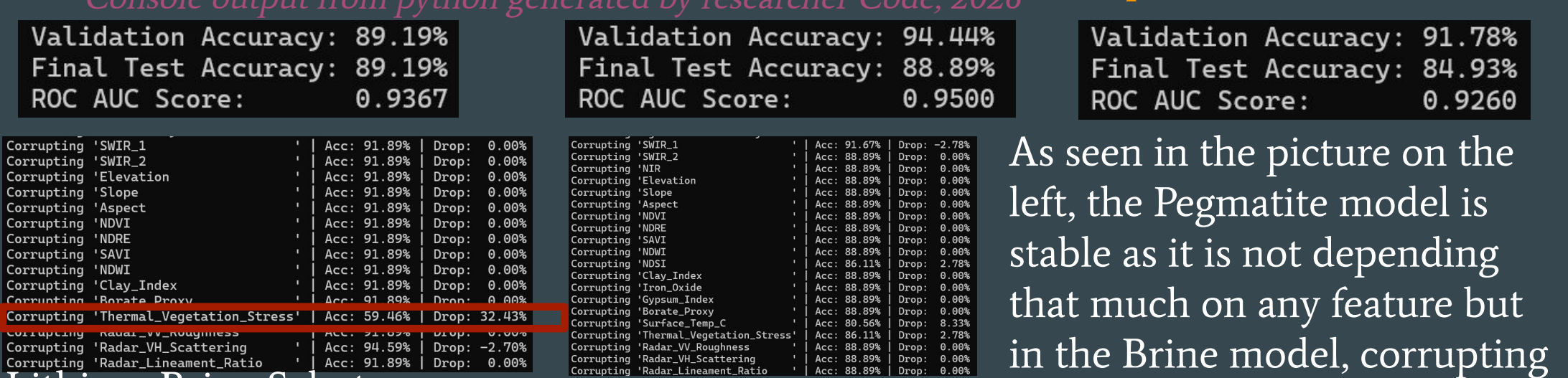
- **Baseline Modeling** ⇒ I started with a simple **Random Forest** model and 30 data points to prove the concept.
- **Changing Model Type** ⇒ I switched to **XGBoost** model (to better correct errors in model) with 100 data points to better filter out background noise.
- **Expanding Markers and Data Points** ⇒ I re-defined **20 geological markers** and increased the dataset to 500 points (mix of Li deposits and no Li) for higher accuracy.
- **Categorization into two different model types** ⇒ I split into two models (Brine vs. Pegmatite for Lithium) to hit **~89% accuracy** for each model as opposed to 85% combined.
- **LOFO Testing** ⇒ Used "**Leave One Feature Out**", a test to identify noise. Removed 6 noisy features from the brine model.
- **Sabotage Test for Model Bias** ⇒ Intentionally sabotaged the model to ensure it wasn't taking shortcuts instead of analysing the data for signatures of Li.

Bibliography/References

- 1.) USGS, MRDS for various data sources, <https://mrdata.usgs.gov/mrds/>,
- 2.) Mindat, for various mineral information <https://www.mindat.org/photo-665586.html>,
- 3.) Google Colab for running codes, colab.research.google.com
- 4.) Google Earth Engine for satellite Images, earthengine.google.com,
- 5.) Python for various programming, python.org,
- 6.) JavaScript for coding,
- 7.) Canva and Google Slides for image depictions,
- 8.) Visual Studio Code, <https://code.visualstudio.com/>,
- 9.) Lithium Harvest , <https://lithiumharvest.com/knowledge/lithium/the-lithium-mining-market/#textimage-6>,
- 10.) EM spectrum, <https://seos-project.eu/remotesensing/remotesensing-c01-p01.html>,
- 11.) Google Gemini for remote sensing and traditional drilling images

I decided to try the *sabotage method*, where it will train the model on all the markers but then on the testing data, it will shuffle up the data for one marker and it will measure the accuracy of that. Identifying “*data bias*” issue and fixing it.

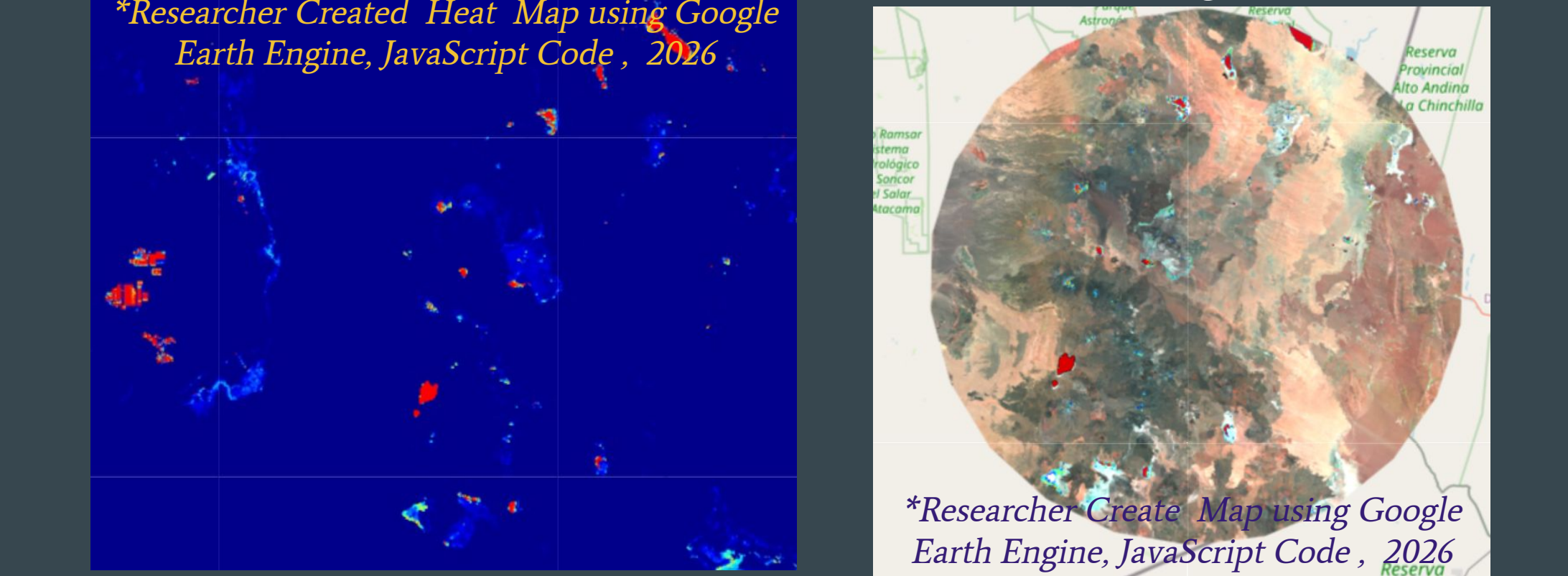
I decided to just train one model on the full 500 data dataset to see that if I did not split the data what the accuracy would have been, it came out to be 85%, proving that splitting was the better idea for higher accuracy.



Lithium Brine Sabotage Results & Evaluation Scores. **Model Acting as a Thermometer!** Which was then fixed

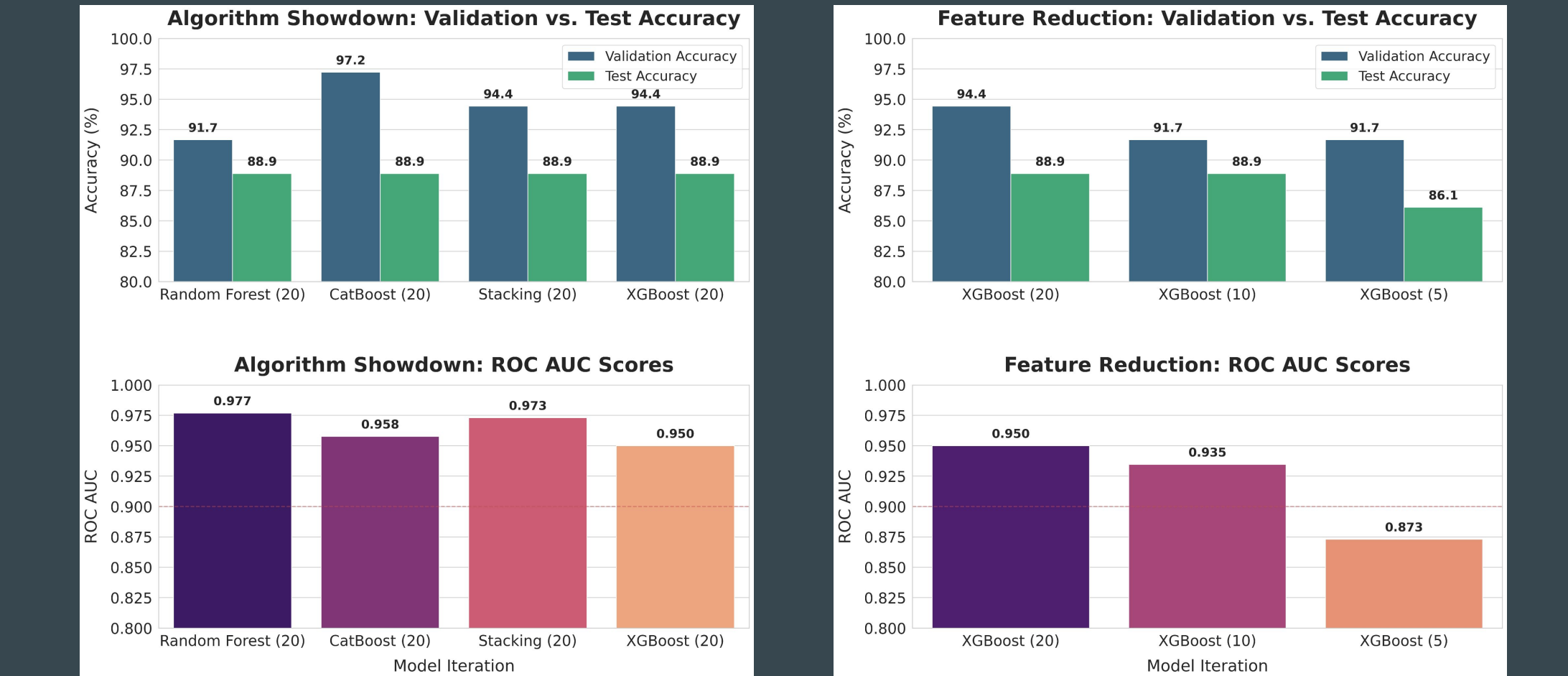
Lithium Pegmatite Sabotage Results & Evaluation Scores. Model Features working in harmony

in the Brine model, corrupting the Thermal Vegetation Stress made the model drop 32% so I decided to remove the feature to remove the “data-bias” and validated against it.

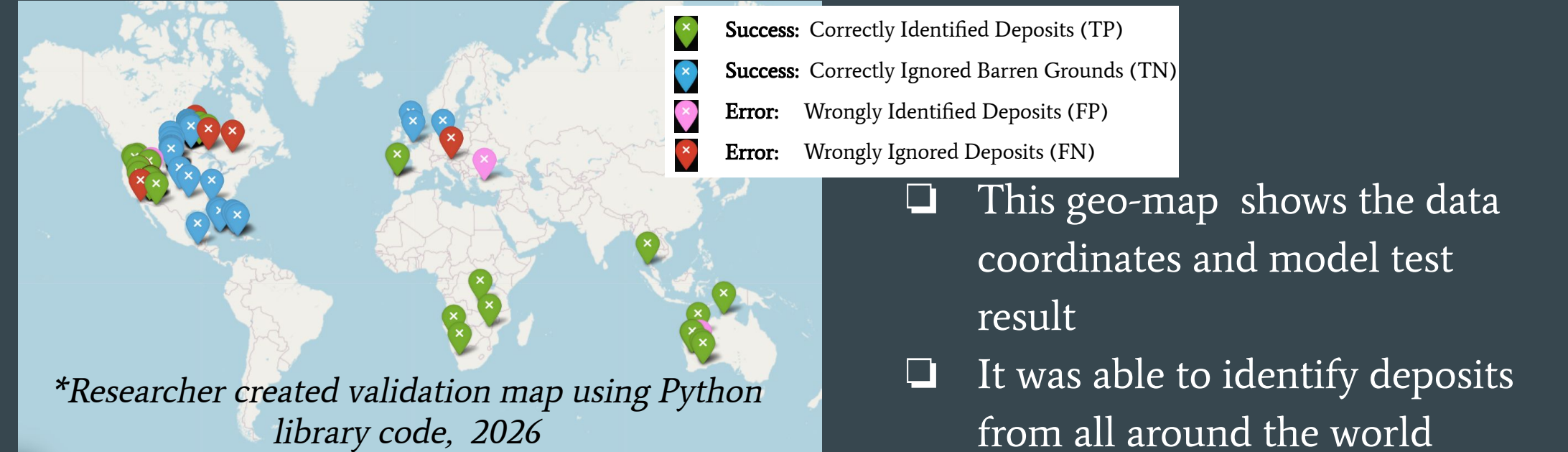


Real-World Validation: Salar de Atacama (Probability Heatmap)

Result proof: The heatmap successfully identified known lithium-rich zones, confirming the model works in the real world. (red >80%, blue: unlikely)



- ❖ Evaluated models using **Accuracy** & **ROC AUC**.
- ❖ **CatBoost** got best validation accuracy & **Random Forest** best ROC AUC score.
- ❖ Tested **XGBoost** with different marker counts to evaluate sensitivity to accuracy.
- ❖ Selected **markers** to maximize **ROC AUC Score**, for most reliable detection.



Analysis and Conclusions / “Takeaways”

- 1.) **Categorization to boost accuracy:** Splitting the model into two separate categories of brine and pegmatite improved detection accuracy from 85% combined to ~89% separately, this helped as the two lithium type has unique signatures.
- 2.) **Preventing Data Leakage:** I separated the data into Training, Validation, and Test sets as opposed to an earlier where during training the validation and test sets were visible to the model. This prevented it from having biases due to data leakages.
- 3.) **Identifying Model's Data Bias:** I discovered a sabotage signature where the model used temperature as a shortcut for accuracy. Removing that marker from the analysis ensured the model was a true lithium detector, not just acting as a thermometer.
- 4.) **Reducing "Noise":** By identifying and removing data markers that negatively affected accuracy, I cleared out "noise" that was confusing the model and lowering accuracy.
- 5.) **Data Stratification to avoid Bias:** I used stratified sampling to ensure both Brine and Pegmatite samples were equally distributed in my training and testing sets. This prevented the model from becoming biased toward one specific type of lithium.
- 6.) **Overcoming Map Overlap:** I solved data overlap issue while combining different datasets different satellites by enlarging the area within each dataset.
- 7.) **Diversity of Markers helping with Accuracy:** Using a diverse set of 20 geo-markers produced higher accuracy than relying on just one or two indicators.

Future Improvements

- 1) **Future Expansion:** This framework can be expanded to find other critical minerals such as Nickel, Copper, and Cobalt.
- 2) **De-noising Data further:** Taking more time-averaged markers to de-noise data.
- 3) **Geographical diversity:** Adding other regions of the world (Asia/China, Russia)
- 4) **Deep Learning Model:** Can help find new insights relative to known feature extraction based analysis with Machine Learning.
- 5) **Time-progression:** Detection of mining Activity on Known sites.
- 6) **Field Verification:** For unknown locations and iterative improvement.