A Sentinel-2 Image Dataset for Mining Detection Across Mining Proportion Ranges in the Brazilian Legal Amazon

Leonardo Fajardo Grupioni[1], **Thomas Jean Georges Gallois**[2], **Felipe Valencia de Almeida**[1]

[1]Escola Politécnica – Universidade de São Paulo, São Paulo, SP – Brazil

[2]Iepé - Instituto de Pesquisa e Formação Indígena, São Paulo, SP – Brazil

{leogrupioni,fvalencia}@usp.br, thomas@institutoiepe.org.br

Abstract. *The recent expansion of mining in the Amazon is a major driver of environmental degradation. Machine learning and satellite imagery are effective tools to monitor this problem, but most datasets treat mining detection as a strictly binary task. This ignores how the actual mining proportion inside an image affects classification results. We present a Sentinel-2 dataset designed to explore this behavior. It contains 2,600 image patches collected across the Brazilian Legal Amazon in 2024 using Google Earth Engine and MapBiomas Collection 10 data. The dataset includes metadata detailing geographic locations, mining proportion ranges in 10% intervals, and a suggested practical split for experiments.*

1. Summary table

Keywords Amazon, mining detection, remote sensing, dataset.
WCAMA 2026 topic Environmental monitoring and remote sensing.
Data type Multispectral satellite images.
Brief data description The dataset contains 2,600 Sentinel-2 image patches. We divided the data evenly, providing 1,300 images without mining and 1,300 images with mining activity distributed across ten proportion intervals of 10%.
Data format PNG images with accompanying CSV metadata.
Collection site Brazilian Legal Amazon.
Public repository Zenodo (`https://doi.org/10.5281/zenodo.18983646`).

2. Introduction

The rapid expansion of illegal mining in the Amazon drives severe environmental degradation. This activity heavily impacts forests, rivers, and indigenous territories, creating deep ecological and social consequences [MapBiomas 2023, Global Initiative Against Transnational Organized Crime 2023].

Remote sensing provides a practical way to monitor this problem continuously across large areas. Researchers increasingly rely on satellite imagery and deep learning to identify visual patterns and detect artisanal mining zones [Gallwey et al. 2020, Camalan et al. 2022].

The main issue is that current literature usually frames mining detection as a strictly binary task [Camalan et al. 2022]. Most available datasets provide only pixel-level segmentation masks or basic presence and absence labels for the entire scene. In

reality, mining often occupies just a tiny fraction of a satellite image, making detection much harder. We still need to understand how classification models actually perform when the mining proportion varies.

We built this Sentinel-2 dataset precisely to explore that challenge. We organized the image patches based on the actual proportion of mining pixels they contain. This approach gives the community the right data to run controlled experiments and see how classification performance shifts as mining becomes more visually dominant in the scene.

3. Material and Methods

We built the dataset using Sentinel-2 imagery from 2024 at a 10-meter spatial resolution [European Space Agency 2025]. We processed all data and extracted the patches through Google Earth Engine [Gorelick et al. 2017]. To identify actual mining locations, we relied on land cover data from MapBiomas Collection 10. This collection offers annual land use maps for Brazil and features a specific class for mining activity. We used this class as our ground truth to estimate mining presence inside every satellite patch.

We started the pipeline by mapping the spatial distribution of mining pixels across the Legal Amazon using the MapBiomas raster. We then divided this region into candidate sampling locations. For the optical data, we accessed Sentinel-2 imagery and filtered the collection by acquisition year and cloud probability. We applied the s2cloudless method to mask out clouds and ensure a clean final mosaic [Sentinel Hub 2024].

From this clean mosaic, we sampled image patches using geographic bounding boxes with a target size of roughly 128 by 128 pixels. However, the actual dimensions vary slightly across the dataset due to how Google Earth Engine handles pixel grids during spatial exports. Because of these natural variations, anyone training machine learning models will need to include a standard preprocessing step, such as resizing or center cropping, to unify the input shapes.

During extraction, we evaluated the mining coverage for each patch solely to assign it to one of the 10% intervals detailed in Table 1. We intentionally balanced the sample count across these ranges. To ensure data quality and practical relevance, a domain expert visually inspected the final collection to validate the structural integrity of the patches and the real-world representation of the mining scenes. This balanced design is what enables researchers to run controlled experiments and observe model behavior as mining grows visually dominant.

Finally, we exported all images in PNG format and grouped them into folders based on their proportion range. We also compiled a metadata file containing patch identifiers, geographic bounding boxes, the assigned mining proportion range, and a practical suggestion for training and testing splits.

4. Data Availability

We hosted the complete dataset and all related metadata on Zenodo (`https://doi.org/10.5281/zenodo.18983646`). The public repository contains the image patches along with a detailed manifest and clear documentation explaining the folder structure. Inside the metadata file, researchers will find the geographic coordinates, the assigned mining proportion ranges, and our suggested dataset split. We designed this open

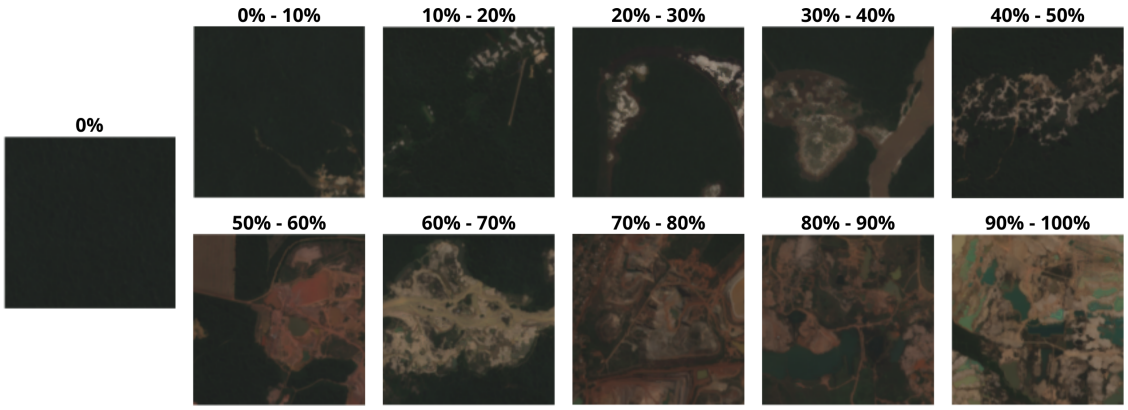

Figure 1. Example patches from the dataset illustrating different mining proportion ranges.

Table 1. Mining proportion ranges and dataset composition.

Mining proportion range	Binary class	Images
0%	No mining	1300
0–10%	Mining	130
10–20%	Mining	130
20–30%	Mining	130
30–40%	Mining	130
40–50%	Mining	130
50–60%	Mining	130
60–70%	Mining	130
70–80%	Mining	130
80–90%	Mining	130
90–100%	Mining	130
Total	–	2600

package to give the community everything they need to reproduce experiments, map the spatial distribution of the samples, and confidently train new machine learning models.

5. Conclusion

We created this Sentinel-2 dataset to push mining detection in the Amazon past the limitations of simple binary classification. By organizing a balanced collection of satellite patches into specific mining proportion ranges, we give the research community a tool to properly investigate how models behave when the target occupies different fractions of a scene.

This structure opens the door for several practical applications in machine learning and remote sensing. Researchers can use these images to evaluate the robustness of convolutional neural networks, especially when the visual signal of mining is extremely weak. In real-world scenarios, mining often appears as tiny clearings or subtle sediment plumes in rivers. Our dataset provides a structured benchmark to train models specifically for these challenging cases. It also serves as a strong foundation for exploring transfer learning techniques and testing new data augmentation strategies tailored for environmental

monitoring.

Ultimately, we built this resource to help bridge the gap between theoretical model performance and the complex reality of environmental tracking. We believe this dataset will directly support the development of more accurate and reliable tools to mitigate the impacts of illegal mining across the Amazon. Future work could expand the dataset by incorporating additional spectral bands and derived indices such as NDVI, NDWI, and SWIR composites, which may improve models sensitivity to subtle mining signatures. Combining Sentinel-2 data with complementary sensors such as Landsat could further increase temporal coverage and spatial diversity of the collection.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. We gratefully acknowledge the the organizations that provided the essential data and tools for this work. We appreciate the European Space Agency for the Sentinel-2 mission and the MapBiomas project for the land use data. Finally, we recognize Google Earth Engine for enabling our geospatial processing and the Sentinel Hub platform for their cloud detection tools.

References

Camalan, S., Cui, K., Pauca, V. P., Alqahtani, S., Silman, M., Chan, R., Plemmons, R. J., Dethier, E. N., Fernandez, L. E., and Lutz, D. A. (2022). Change detection of Amazonian alluvial gold mining using deep learning and Sentinel-2 imagery. *Remote Sensing*, 14(7):1746.

European Space Agency (2025). Sentinel-2 user handbook. Available at: `https://sentinels.copernicus.eu/web/sentinel/copernicus/sentinel-2`. Accessed on: 13/01/2026.

Gallwey, J., Robiati, C., Coggan, J., Vogt, D., and Eyre, M. (2020). A Sentinel-2 based multispectral convolutional neural network for detecting artisanal small-scale mining in Ghana: Applying deep learning to shallow mining. *Remote Sensing of Environment*, 248:111970.

Global Initiative Against Transnational Organized Crime (2023). Amazon underworld: economias criminosas na maior floresta tropical do mundo. Technical report, GI-TOC and Amazon Watch and InfoAmazonia, Geneva, Switzerland. Institutional report.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27.

MapBiomas (2023). Amazônia concentra mais de 90% do garimpo no Brasil. Available at: `https://brasil.mapbiomas.org/2023/09/22/amazonia-concentra-mais-de-90-do-garimpo-no-brasil/`. Accessed on: 13/01/2026.

Sentinel Hub (2024). s2cloudless: Machine learning cloud detector for Sentinel-2 imagery. Available at: `https://docs.sentinel-hub.com`. Accessed on: 13/01/2026.