AI for Earth: Evaluating Land Cover Classification from LUCAS Image Chips Using Geospatial Foundation Models

Dong Ha Shin
PiA Course - Politecnico di Milano
Coordinator: Prof. Maria Antonia Brovelli

November 15, 2025

Abstract

This report presents an experimental study on land cover classification using satellite image chips derived from the LUCAS (Land Use/Cover Area Survey) dataset. The work was developed in the *Artificial Intelligence for the Earth: Geospatial Foundation Models* Passion-in-Action course at Politecnico di Milano, coordinated by Prof. Maria Antonia Brovelli.

After extracting multispectral HLS image chips for Lithuania using Google Earth Engine, several datasets with varying sampling strategies were built and evaluated using a controlled CNN baseline implemented with PyTorch Lightning and TerraTorch. Experiments include proportional sampling, equal sampling, and a merged-class strategy designed to address extreme class imbalance. The findings highlight how dataset structure and class aggregation influence classification performance and provide guidance for future fine-tuning of Geospatial Foundation Models (GFMs).

1 Introduction

Geospatial Foundation Models (GFMs)—such as Prithvi and TerraMind—are trained on multi-sensor, multi-temporal Earth observation data to support a wide range of downstream tasks. Before adapting these models, this project explores how dataset composition alone affects land cover classification performance using a simple CNN baseline.

This approach provides a controlled environment to study issues such as class imbalance, sample scarcity, spectral variability and model generalization.

2 Data Preparation

LUCAS points for Lithuania were filtered and processed through Google Earth Engine to generate surface reflectance HLS (HLSS30) chips of size 224 × 224 px. Each chip contains six spectral bands: B2, B3, B4, B8, B11, B12.

Four datasets were produced:

- **Dataset 1:** 100 samples, proportional.
- **Dataset 2:** 500 samples, proportional.
- **Dataset 3:** 10 samples per class
- **Dataset 4:** 50 samples per class

The original LUCAS nomenclature includes ten categories:

1. Arable Land
2. Permanent Crops
3. Grass
4. Wooded Areas
5. Shrubs
6. Bare surface
7. Artificial constructions
8. Inland Water
9. Transitional/Coastal Water
10. Impossible to PI

3 Class Distribution Analysis

Figure 1 shows the histogram of Dataset 1 (100 proportional samples).

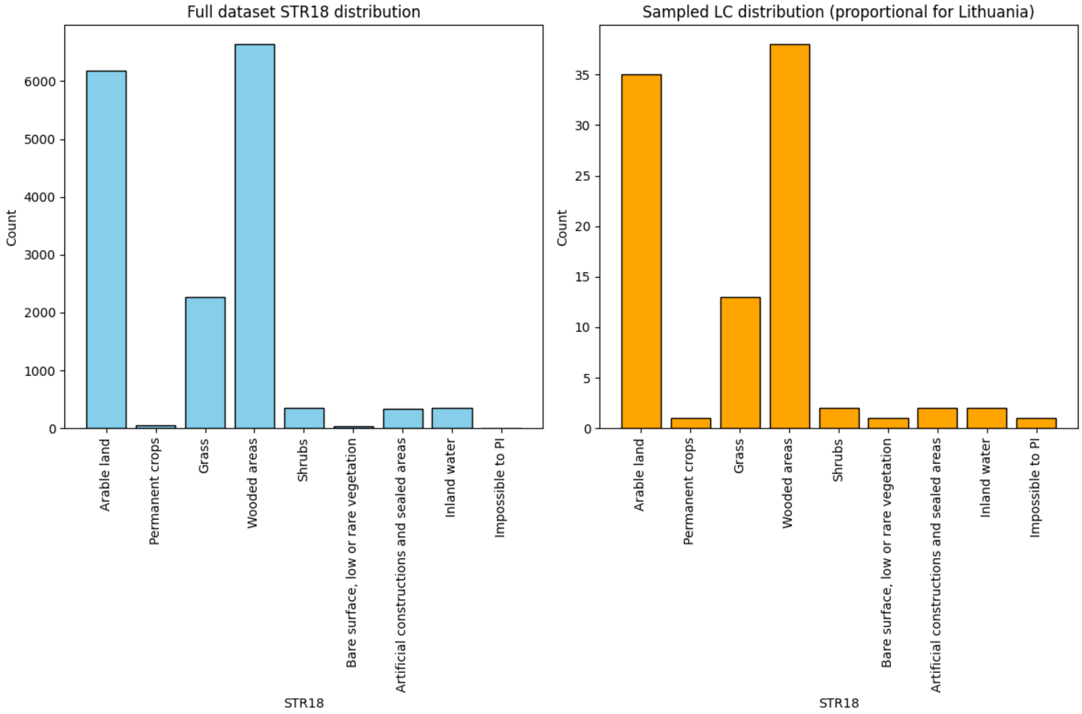

Figure 1: Class distribution in the initial proportional dataset of 100 samples. Several categories appear with extremely low or zero representation.

The distribution reveals severe imbalance:

- some classes appear **only once** (e.g. Class 7),
- others are **entirely absent** (Classes 2, 6, 9, 10),
- Class 1 and Class 4 dominate the dataset.

This imbalance critically affects the learning process, pushing the model to overfit the majority classes while ignoring minority classes.

4 Experiments

All experiments use the same CNN architecture, optimizer and training schedule (17–21 minutes per training session).

4.1 Dataset 1: 100 Chips (Proportional)

- Test Loss: 1.47
- Accuracy: 0.167
- Micro Accuracy: 0.361
- F1 (macro): 0.088

Only Class 1 reached meaningful accuracy and F1 due to dominating representation.

4.2 Dataset 2: 500 Chips (Proportional)

- Test Loss: 1.05
- Accuracy: 0.296
- Micro Accuracy: 0.676
- F1 (macro): 0.257

Performance improves substantially; however, the long tail of rare classes still prevents complete generalization.

4.3 Dataset 3: 10 Chips per Class (Balanced)

Balanced but too small:

- Test Loss: 2.18
- Accuracy: 0.063
- Micro Accuracy: 0.045
- F1 (macro): 0.019

Only Class 7 produced a non-zero score, indicating insufficient sample size.

4.4 Dataset 4: 50 Chips per Class (Equal Sampling)

- Test Loss: 1.92
- Accuracy: 0.226
- Micro Accuracy: 0.230
- F1 (macro): 0.152

Classes 4 and 8 performed relatively well, but several classes (2, 3, 5, 6) remained extremely challenging. Even with 50 examples per class, the CNN baseline struggles to learn complex multispectral patterns.

4.5 Data Augmentation Strategy

To mitigate overfitting caused by the limited size of the datasets, a data augmentation pipeline was applied during training. The augmentation strategy included spatial transformations such as random horizontal flips, random rotations, affine transformations, and random crops. For multispectral data, augmentations were designed to preserve radiometric consistency: no spectral shifting or color jitter beyond small brightness variations was introduced.

These transformations increase intra-class variability and encourage the model to learn more generalizable spectral–spatial features. Data augmentation proved beneficial, particularly for the balanced datasets (10 and 50 samples per class), where it reduced early overfitting and improved training stability. However, augmentation alone was insufficient to compensate for extremely small class sizes or missing categories.

5 Merged-Class Strategy

Given the persistent imbalance and spectral overlap among certain classes, an additional experiment was performed by **merging similar LUCAS categories**. The motivations are:

- reduce fragmentation of classes with extremely low sample count;
- simplify the classification problem when fine-grained spectral boundaries are unclear at 30 m resolution.

The merged scheme grouped the 10 classes into fewer macro-classes (e.g. vegetation, artificial surfaces, water bodies).

Results of the merged model

The merged-class classifier showed a substantial increase in stability:

- minority classes no longer suffered from sample scarcity,
- confusion among similar land cover types was reduced,
- overall accuracy and F1 improved compared to the 10-class setting.

Although still limited by dataset size, the merged-class approach demonstrated that a coarser taxonomy better aligns with the model capacity and the resolution of HLS data.

6 Discussion

Experimental findings highlight several key points:

1. **Dataset size matters**: scaling from 100 to 500 samples improves performance across all metrics.
2. **Severe imbalance prevents learning**: classes with 0–2 samples cannot be learned by a CNN.
3. **Balance is not enough**: 10 samples per class are insufficient for multispectral classification.
4. **Class aggregation mitigates fragmentation**: the merged-class model significantly improves robustness.

Across all experiments, the CNN baseline shows limited ability to generalize, confirming the need for:

- larger and more informative datasets (200–500 samples per class), and/or
- more capable architectures such as Prithvi or TerraMind.

7 Conclusion

This study demonstrates the impact of dataset composition, sampling strategy and class definition on land cover classification from HLS image chips. Although based on a simple CNN, the experiments provide valuable insights for designing datasets and preparing downstream tasks for Geospatial Foundation Models. Future work will involve fine-tuning Prithvi or TerraMind using the best-performing dataset configurations.

Acknowledgements

This work was carried out within the *Artificial Intelligence for the Earth: Geospatial Foundation Models* course at Politecnico di Milano. Special thanks to Prof. Maria Antonia Brovelli and the teaching team for their support.