Land Cover Classification over Delhi-NCR using Satellite Imagery

This repository contains a complete pipeline for performing supervised land cover classification using RGB satellite imagery and ESA for the Delhi-NCR region.

1 Project Tasks Overview

This project addresses the following core tasks:

- **Q1**: Grid Construction and Image Filtering using shapefiles and geospatial mapping.
- **Q2**: Patch Extraction and Land Cover Labeling using ESA WorldCover raster.
- **Q3**: Supervised Evaluation using a deep learning classifier and multiple evaluation metrics.

2 Objectives

- Learn to manipulate vector and raster geospatial data using GeoPandas and RasterIO.
- Generate structured image-label pairs for training from raw satellite and land cover data.
- Train a convolutional neural network (ResNet18) for multiclass classification.
- Evaluate model performance using F1 scores, classification reports, and confusion matrices.

3 Workflow Summary

1. **Geo-filtering and Grid Overlay**:
   - Reprojected Delhi-NCR boundary to UTM (EPSG:32643) for accurate distance-based grid creation.
   - Created a uniform 60×60 km grid and **stored all cell corners properly** for downstream spatial filtering.
   - The issue in earlier versions where grid corners were not fully captured (due to a misplaced extend() call) has been corrected.
2. **Dataset Creation and Labeling (Q2)**:
   - 128×128 pixel patches were extracted from ESA WorldCover 2021 raster, centered at the image coordinates.
   - Each patch was assigned a label based on the dominant land cover class ($\geq$60% occurrence).
   - Patches without a clear dominant class or with no-data pixels were discarded.
   - Class distribution was visualized using bar plots to highlight imbalance.
   - A 60/40 train-test split was performed with stratification.
3. **Model Training and Evaluation (Q3)**:

- A pre-trained ResNet18 model was fine-tuned on the RGB image dataset.
- The model was trained for 5 epochs using CrossEntropyLoss and Adam optimizer.
- Evaluation metrics included accuracy, macro F1 score, and a confusion matrix.
- Torchmetrics was used alongside Scikit-learn to validate the F1 score.

4. **Labeling**:
   - Extracted land cover class codes from ESA WorldCover `.tif` raster patches.
   - Mapped ESA class codes to 11 standardized land cover labels (e.g., 10 → "Tree Cover", 50 → "Built-up") using a predefined dictionary.
   - Only assigned labels where the dominant class (>60%) was present in the patch.

For example: ESA Class Mapping

ESA Code	Class Name
10	Tree Cover
20	Shrubland
30	Grassland
40	Cropland
50	Built-up
60	Bare/Sparse
70	Snow/Ice
80	Wetlands
90	Water
95	Tundra
100	Mangroves

4 Results

Metric	Value
Training Accuracy	98.8%
Test Accuracy	98.0%
Macro F1 Score	0.7265

- Best performance was observed for **Cropland** and **Built-up** classes. - Some misclassifications occurred in **Shrubland** and **Tree Cover** due to class imbalance. - Visualizations included heatmaps and class-wise confusion matrices.

5 Tools and Libraries

- PyTorch, torchvision, torchmetrics
- GeoPandas, RasterIO, Shapely
- Leafmap (for satellite basemap visualization)
- Scikit-learn, Seaborn, Matplotlib

6 Data Sources

- **RGB Satellite Images**: Downloaded patches with geospatial filenames (e.g., 28.6129_77.2295.png)
- **ESA WorldCover 2021**: Used for labeling based on land cover classes
- **Delhi-NCR GeoJSON**: Provided shapefile for spatial filtering and overlay

7 How to Run (Google Colab)

```
# Mount Google Drive
from google.colab import drive
drive.mount('/content/drive')
And run the python file.
```

8 Class Distribution and Imbalance Analysis

After filtering and labeling, we visualized the distribution of land cover classes across the dataset using a bar plot.

- A total of 6,563 patches were retained after applying a dominance threshold ($\geq$60%).
- Classes like **Cropland** and **Built-up** dominated the dataset, with thousands of examples.
- Other classes such as **Shrubland**, **Tree Cover**, and especially **Grassland** and **Wetlands** were underrepresented or nearly absent.

This imbalance can affect model generalization, especially for rare classes. Although we used stratified sampling for the train-test split, future work should include:

- Data augmentation for minority classes
- Class weighting during training
- Sampling techniques like SMOTE for balancing

A sample visualization of the class distribution (generated with Seaborn) is in the folder.

- Evaluation metrics include accuracy, macro F1-score, and confusion matrix. - The macro F1-score was computed using both a custom implementation and the torchmetrics library to ensure correctness and consistency.

The confusion matrix compares actual versus predicted labels for five major ESA WorldCover land classes: Built-up, Cropland, Grassland, Shrubland, and Tree Cover. The results show:

1. High accuracy for dominant classes:
2. Cropland and Built-up classes are classified with strong accuracy. Out of 1933 Cropland samples, 1896 were correctly predicted.
3. Misclassifications in less represented classes:
4. Grassland had no correct predictions — all five samples were confused with either Built-up or Shrubland.
5. Tree Cover was sometimes confused with Shrubland (12 out of 57 cases).
6. Moderate confusion occurred between urban (Built-up) and peri-urban (Cropland or Shrubland) areas, suggesting texture-based overlaps in RGB imagery.

The confusion matrix validates that:

- The model has learned dominant land classes well, supported by high diagonal values.
- However, minority class performance remains limited, especially for Grassland, which lacks enough examples to generalize.
- The misclassification trends also indicate that some ESA classes share visual features when represented as 128×128 RGB patches.
- To improve performance on underrepresented classes, further steps may include:
  - Data augmentation for low-sample classes
  - Applying class reweighting or focal loss
  - Incorporating additional spectral bands or vegetation indices