AI Model Development for Urban Fabric Segmentation

Barbara Metzler and Dani Arribas-Bel

Table of contents

This report includes the preliminary research and experiments run to design an AI model for making predictions on the urban fabric of England, as of November 2024.

1 Executive Summary

This technical report details the development and evaluation of artificial intelligence (AI) models for urban fabric classification and segmentation using satellite imagery. The project spans from June 2024 to March 2025, encompassing model design, development, training, and application to European urban strategy analysis. Our research compares various approaches including classification and segmentation tasks, using different foundation models as backbones. The primary goal is to develop accurate and efficient methods for predicting urban fabric from satellite imagery.

2 Classification vs segmentation

In satellite image analysis, classification and segmentation address spatial labeling at different levels of granularity, with classification assigning a single label to an image tile or cell, while segmentation provides pixel-level detail. In our study, the label dataset does not always correspond directly to identifiable features in the imagery, making classification a potentially more suitable approach as it generalizes each tile's dominant land cover type without requiring exact pixel alignment. However, we explore both approaches: classification for broader pattern recognition and segmentation for finer, boundary-specific mapping. This dual approach enables us to evaluate how each method performs given the scale and feature alignment limitations within the dataset.

3 Methods

3.1 Data

3.1.1 Satellite imagery (input)

The satellite image data used is accessed from the GHS-composite-S2 R2020A dataset [1]. The dataset is a global, cloud-free image composite derived from Sentinel-2 L1C data, covering the period from January 2017 to December 2018. We use three bands (RGB) and the images have a resolution of 10 meters per pixel.

3.1.2 Urban fabric classes (outcome)

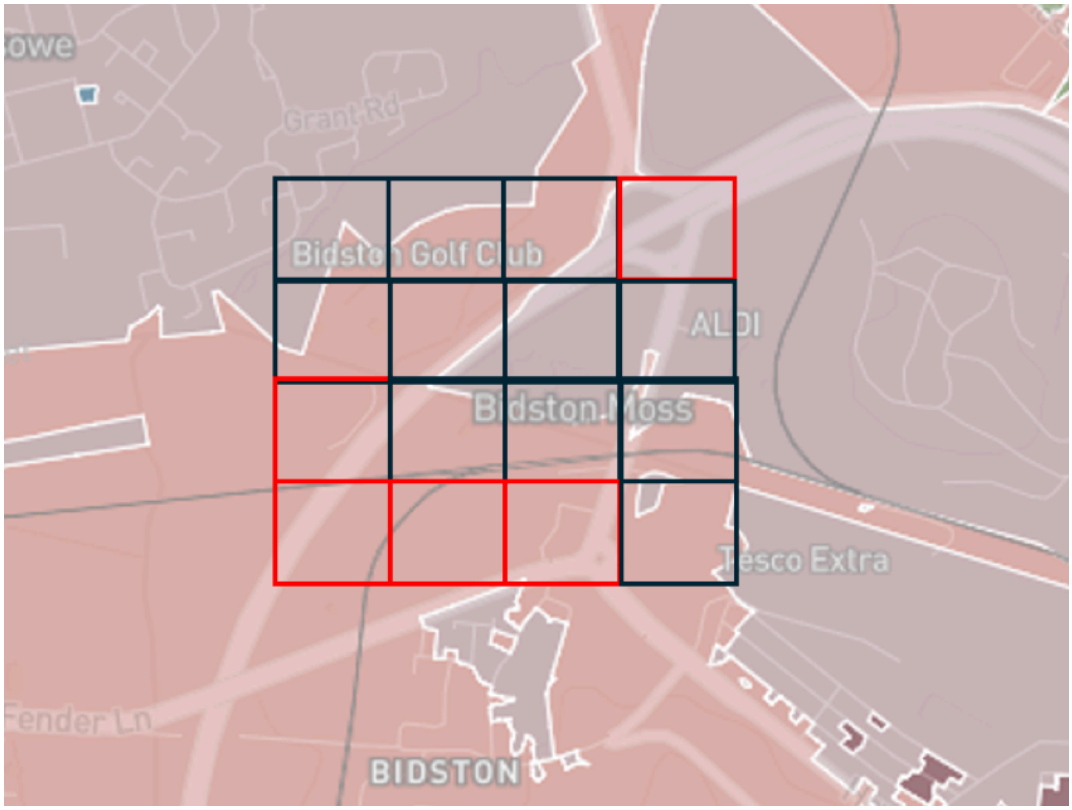
We use labels from the Spatial Signatures Framework [2], which provides a typology of British urban environments based on both form (physical appearance) and function (usage). Spatial signatures capture complex urban characterizations, offering insights into how different areas look and operate. While our primary interest is in urban fabric classification focused on form—an approach we expect may be simpler since form is visible in imagery—this classification scheme is still under development. Therefore, we currently use the more comprehensive Spatial Signatures Framework as a proxy, as it aligns with the goals of urban characterization in our project.

3.2 Data preprocessing

For our analysis, we use two datasets of image tiles at different scales: larger tiles (224 x 224 pixels, covering 2240 x 2240 meters) for segmentation tasks, and smaller tiles (56 x 56 pixels, covering 560 x 560 meters) for classification tasks. The segmentation dataset includes 26,753 tiles (21,402 for training and 5,351 for testing), while the classification dataset consists of 403,722 tiles (342,648 for training and 61,074 for testing). For consistent sampling and comparison, we only use tiles that fully overlap with the spatial signatures, ensuring that each tile aligns with the urban form and function typology in our labeling framework. This alignment supports robust comparison of classification and segmentation outcomes on a pixel-level.

[1] Corbane, C. et al., 2020. A global cloud-free pixel-based image composite from Sentinel-2 data. Data in brief, 31, p.105737.

[2] Fleischmann, M. & Arribas-Bel, D., 2022. Geographical characterisation of British urban form using the spatial signatures framework. Scientific Data, 9(1), p.546.

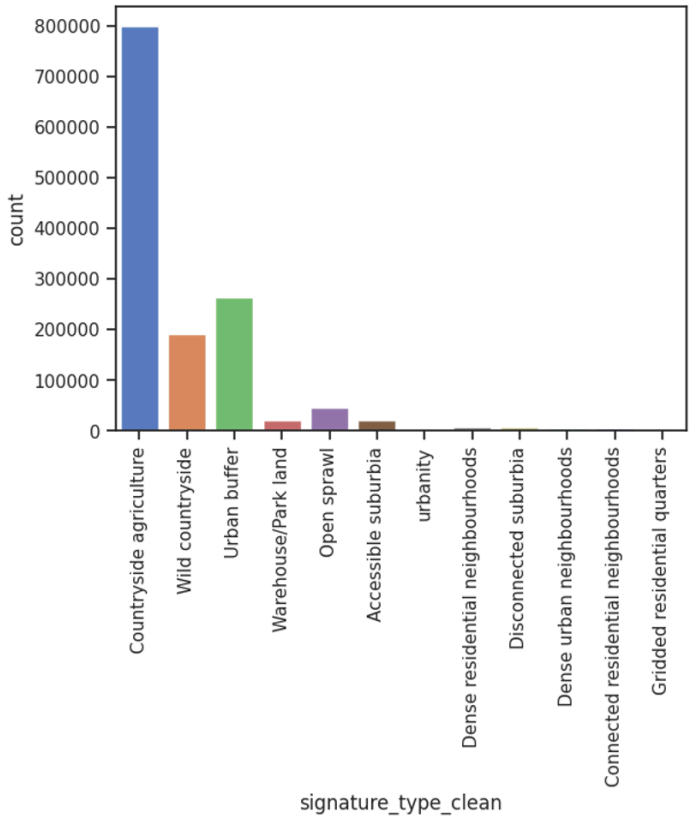


A significant challenge in our dataset is class imbalance, where certain urban fabric types are substantially more represented than others. This imbalance influenced our decisions regarding model architecture and loss function selection, leading us to explore specialized approaches for handling uneven class distributions.

3.2.1 Train/test split

We split the dataset into 80% train and 20% test data. The test datasets for segmentation and classification overlap.

3.2.2 Unbalanced dataset



4 Model architectures

Our study involves three main experiments to analyze urban fabric classification and segmentation. First, we conduct a baseline experiment using image embeddings from the SatlasPretrain model [3], which we fit to an XGBoost classifier to predict urban fabric classes (Approach A). Second, we fine-tune three different geospatial foundation models—SatlasPretrain, Clay, and IBM/NASA's Prithvi model—to perform segmentation tasks (Approach B). Third, we take the best-performing geospatial foundation model from the segmentation experiments (Clay) and fine-tune it specifically for a classification task (Approach C). To evaluate and compare the results, we report weighted pixel-level accuracy, F1 score, and Intersection over Union (IoU) metrics across the experiments.

- Approach A: Image embeddings + XGBoost model

[3]Bastani, F. et al., 2023. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16772-16782.

- Approach B: Fine-tuned geospatial foundation model (segmentation)
- Approach C: Fine-tuned geospatial foundation model (classification)

4.1 Baseline approach (Approach A)

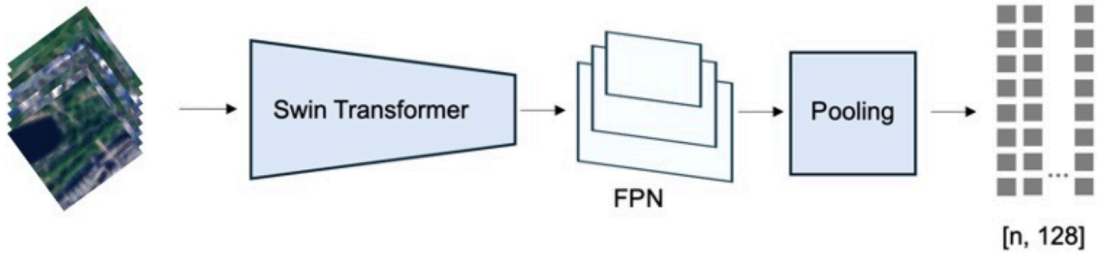
The tiles are fed into the geospatial foundation model SatlasPretrain [4] that has been pretrained with more than 302 million labels on a range of remote sensing and computer vision tasks.

The model operates in two main steps:

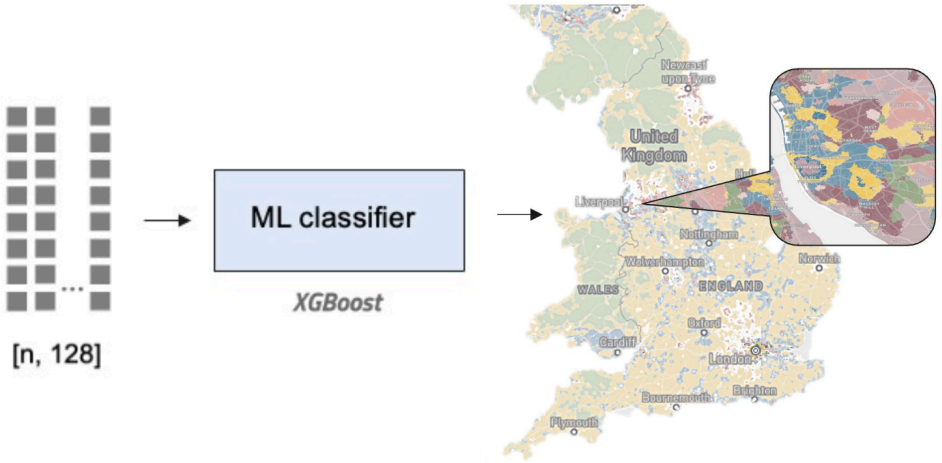
1. Foundation Model: A vision transformer model with a feature pyramid network (FPN) and a pooling layer is used to derive image embeddings—lower-dimensional representations of the images (Fig. 2).
2. Machine Learning Classifier: The image embeddings are then input into an XGBoost classifier to predict urban fabric classes across England.

Our baseline model achieved a moderate prediction accuracy of approximately 61% with varying accuracy across the various spatial signature classes as seen in the figure below.

[4] Bastani, F. et al., 2023. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 16772-16782.

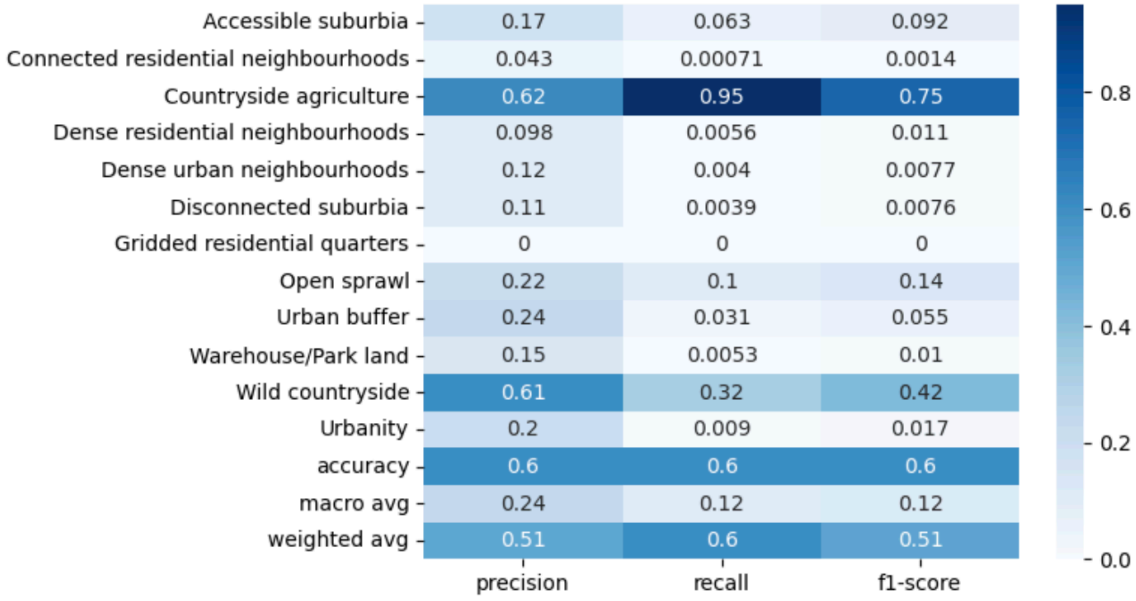

Swin Transformer
FPN
Pooling
[n, 128]

Step 1


ML classifier
XGBoost
[n, 128]
United Kingdom
Liverpool
Nottingham
Wolverhampton
Norwich
WALES
ENGLAND
Oxford
Cardiff
London
Bournemouth
Brighton
Plymouth

Step 2

4.1.1 Results across spatial signatures

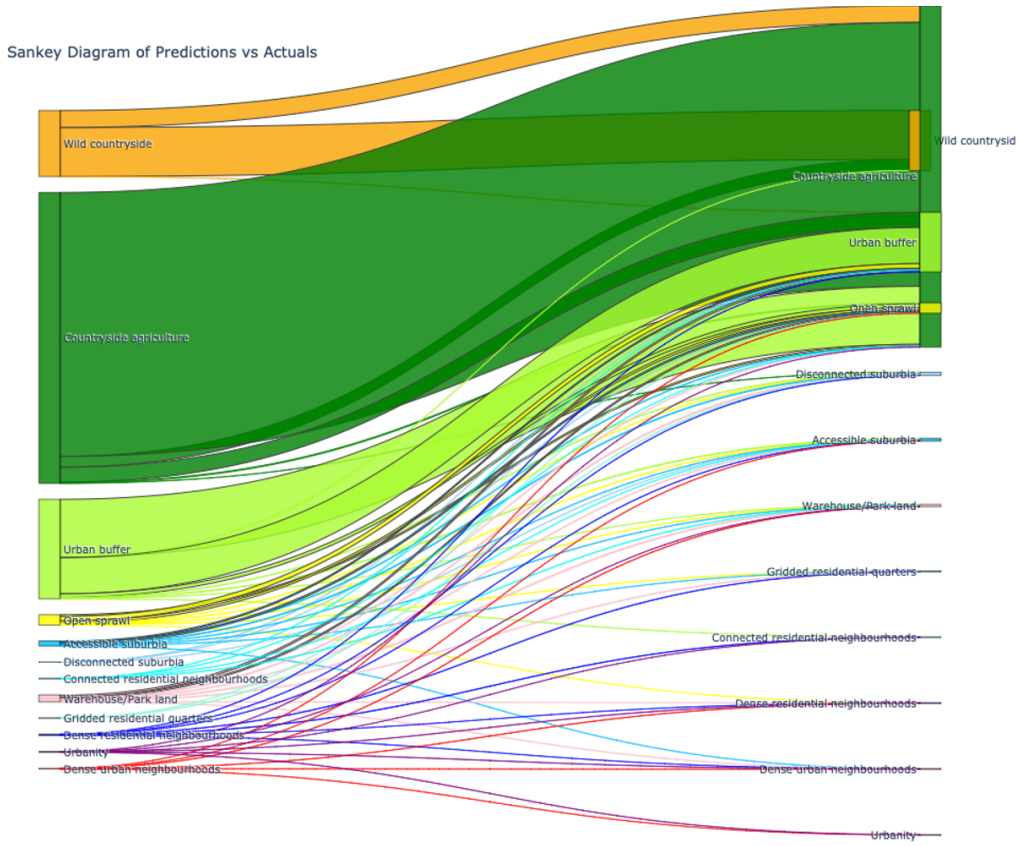


Baseline approach (ordinal)

In addition to the general classification task, we explored an ordinal regression task to account for the continuous nature of the spatial signatures, which are not strictly categorical. We applied the following ordinal mapping:

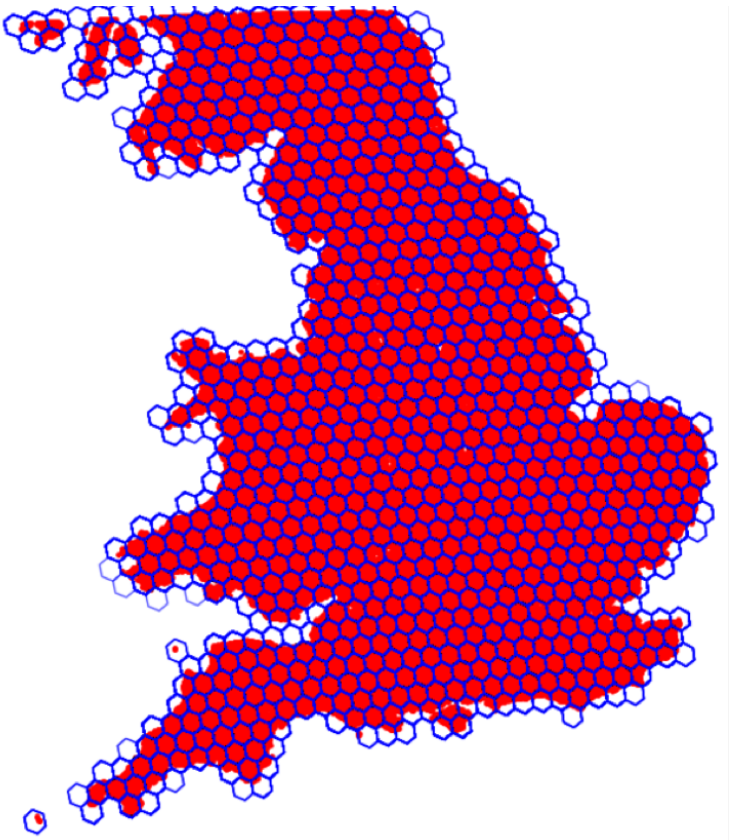
```
ordinal_mapping = {     'Wild countryside': 0,     'Countryside agriculture':
1,     'Urban buffer': 2,     'Open sprawl': 3,     'Disconnected suburbia':
4,     'Accessible suburbia': 5,     'Warehouse/Park land': 6,     'Gridded
residential quarters': 7,     'Connected residential neighbourhoods': 8,
'Dense residential neighbourhoods': 9,     'Dense urban neighbourhoods': 10,
'Urbanity': 11, }
```

This ordinal approach produced a Mean Absolute Error (MAE) and Mean Squared Error (MSE) of 0.28, along with an R^2 score of 0.62. The Sankey diagram below highlights the primary misclassifications, which tend to occur between similar classes.



Baseline approach + spatial context

To enhance model performance, we incorporated spatial context, enabling the model to account for regional location. We included H3 Level 5 hexagon identifiers as a categorical variable, where each hexagon (approximately 560x560 meters) encompasses around 80 tiles.



4.2 Segmentation (Approach B)

In Approach B, we fine-tuned a state-of-the-art geospatial foundation model for a segmentation task. We used the 224x224x3 image tiles as input. We evaluated three state-of-the-art foundation models, each with unique characteristics as listed below:

Model	Architecture	Dataset Size	Image Sources
Satlas [5]	SwinT	302M labels	Sentinel-2
Clay [6]	MAE/ViT	70M labels	Multiple+
Prithvi [7]	MAE/ViT	250 PB	Sentinel-2/Landsat

+Multiple sources include Sentinel-2, Landsat, NAIP, and LINZ

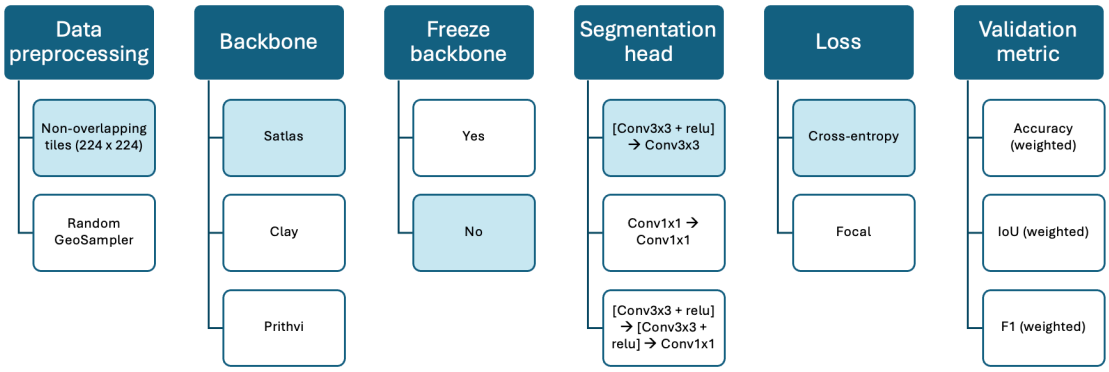
[5] Corbane, C. et al., 2020. A global cloud-free pixel-based image composite from Sentinel-2 data. Data in brief, 31, p.105737.

[6] https://huggingface.co/made-with-clay/Clay

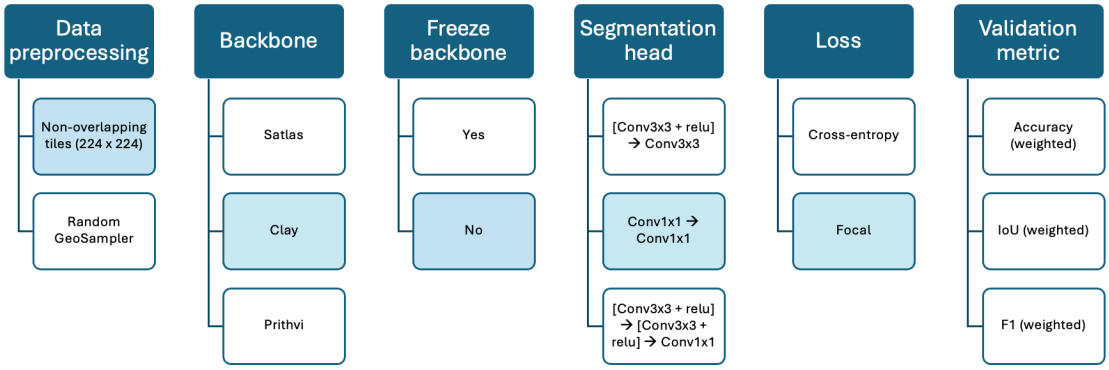
[7] https://huggingface.co/ibm-nasa-geospatial/Prithvi-100M

The following visualisations show the varying model configurations for the three different approaches tested for the segmentation task. The main difference is the varying backbone.

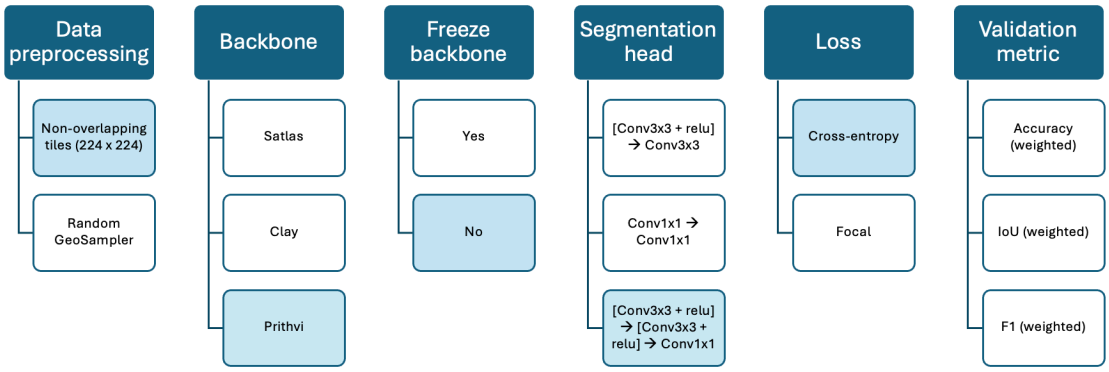
4.2.1 Model A: Satlas



4.2.2 Model B: Clay



4.2.3 Model C: Prithvi



After fine-tuning each foundation model for 10 epochs, we observed the following performance metrics:

Metric	Satlas	Clay	Prithvi
Weighted Accuracy	0.57	**0.72**	0.62
Weighted IoU	0.33	**0.58**	0.41
Weighted F1	0.41	**0.69**	0.58
Training Time/Epoch	9 mins	8 mins	20 mins
Parameters	90M	86M	120M
Implementation Score	5/10	6/10	7/10

The Clay model consistently outperformed other foundation models across all metrics, while also maintaining reasonable training times and computational requirements.

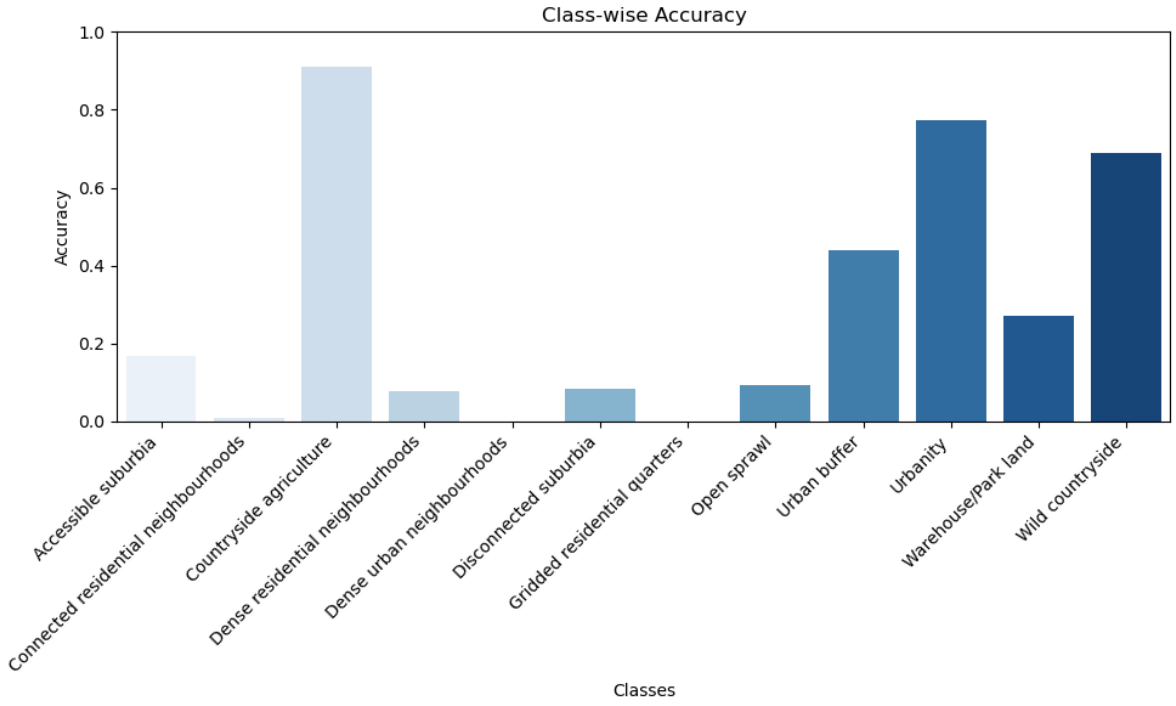
Loss Function Impact: The choice of loss function significantly influenced model performance. Focal loss proved particularly effective in handling class imbalance, especially when combined with the Clay model architecture.

4.3 Classification (Approach C)

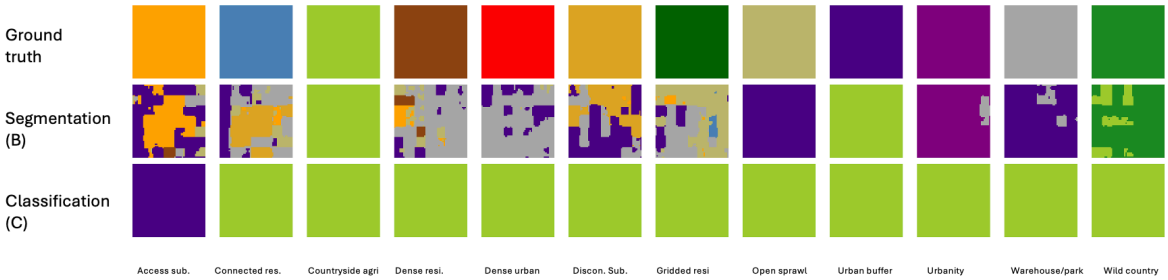
Finally, in Approach C, we fine-tuned a geospatial foundation model for a classification task. We used the 56x56x3 image tiles as input.

For this approach we only used the Clay model as backbone, since it performed the best in the previous experiments.

The accuracy varied across the different classes as seen in the figure below:



This figure shows a comparison between the predicted classes for the fine-tuned geospatial foundation model with segmentation approach (B) and the classification (C).



4.3.1 Evaluation Metrics

We employed multiple complementary metrics to evaluate model performance, as follows:

1. **Intersection over Union (IoU):** This metric measures the overlap between predicted and ground truth segmentations, ranging from 0 (no overlap) to 1 (perfect overlap). IoU

is calculated as the area of intersection divided by the area of union between the predicted and actual segmentation masks.

2. **Weighted F1 Score:** This metric provides a balanced measure of precision and recall, particularly important for imbalanced datasets. It is calculated as the harmonic mean of precision (how many of the predicted positives are correct) and recall (how many of the actual positives were identified), weighted by class frequencies.
3. **Weighted Accuracy:** This metric calculates the proportion of correct predictions, weighted by class frequencies to account for class imbalance. It provides a more representative measure of model performance across all classes, regardless of their frequency in the dataset.

5 Results

Comparing the results is a non-trivial task because the image tiles do not correspond to each other and do not perfectly overlap (42px vs 224 px). To make a fair comparison we thus calculate the pixel-level accuracy scores across the approaches. For this purpose, we predict the full map of the test set and compare the overlapping tiles (as described in sampling). We then calculate the following metrics on a per-pixel level.

5.1 Overall model performance comparison (Pixel-level)

Our comprehensive evaluation revealed varying levels of performance across the three different approaches:

Approach	Global Accuracy	Macro Accuracy	F1 Score	IoU
Classification (embeddings)	0.76 (0.66)	0.22 (0.13)	0.23	0.63
Classification + H3 level 5	**0.87** (0.82)	**0.42** (0.35)	**0.45**	**0.79**
Classification + H3 ordinal	0.80 (0.80)	0.26 (0.26)	0.26	0.69
Classification (Clay)	0.59 (0.68)	0.09	0.12	0.38
Segmentation (Clay)	0.73	0.31	0.30	0.58

The baseline classification approaches demonstrated varying levels of success: - Basic embedding classification achieved 76% global accuracy (66% balanced) - Integration with H3 level 5

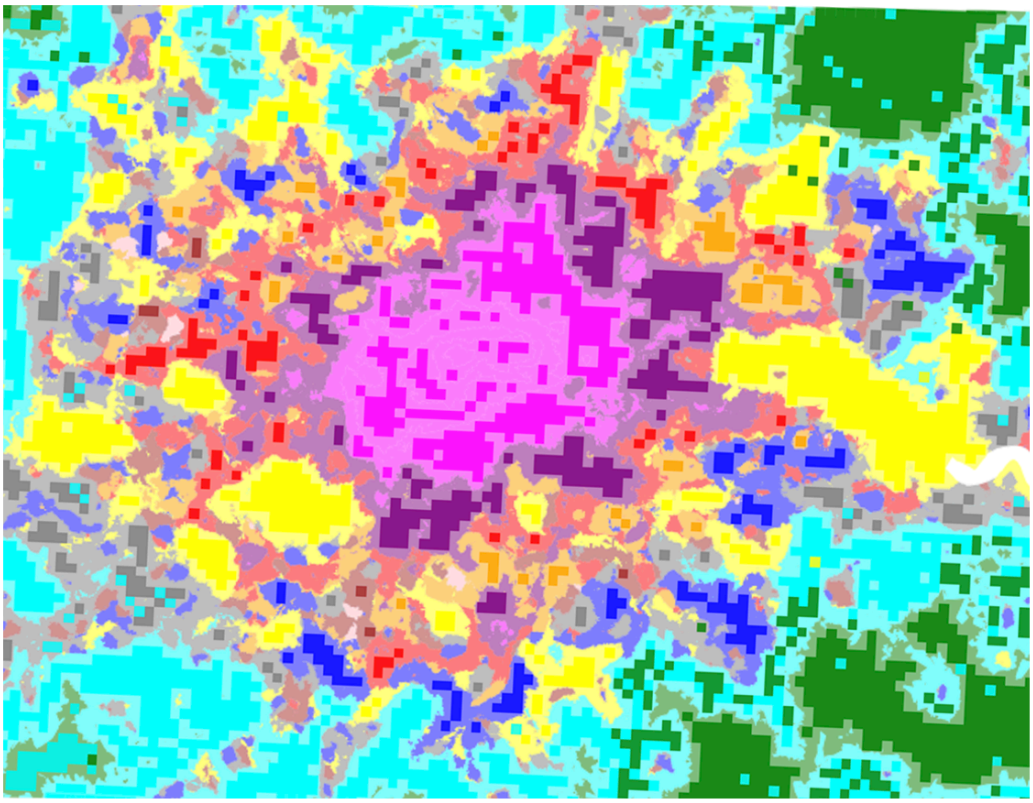
spatial indexing significantly improved performance to 87% global accuracy (42% balanced) - H3 level 5 ordinal classification reached 80% accuracy (26% balanced)

The fine-tuned geospatial foundation model performed better than the fine-tuned classification, with an accuracy score of 0.56 and 0.73 respectively.

Overall, the baseline approach with regional information performed best. This approach is not only the best performing but also relatively efficient to implement. Once the image embeddings are created, the downstream classification can be done in minutes.

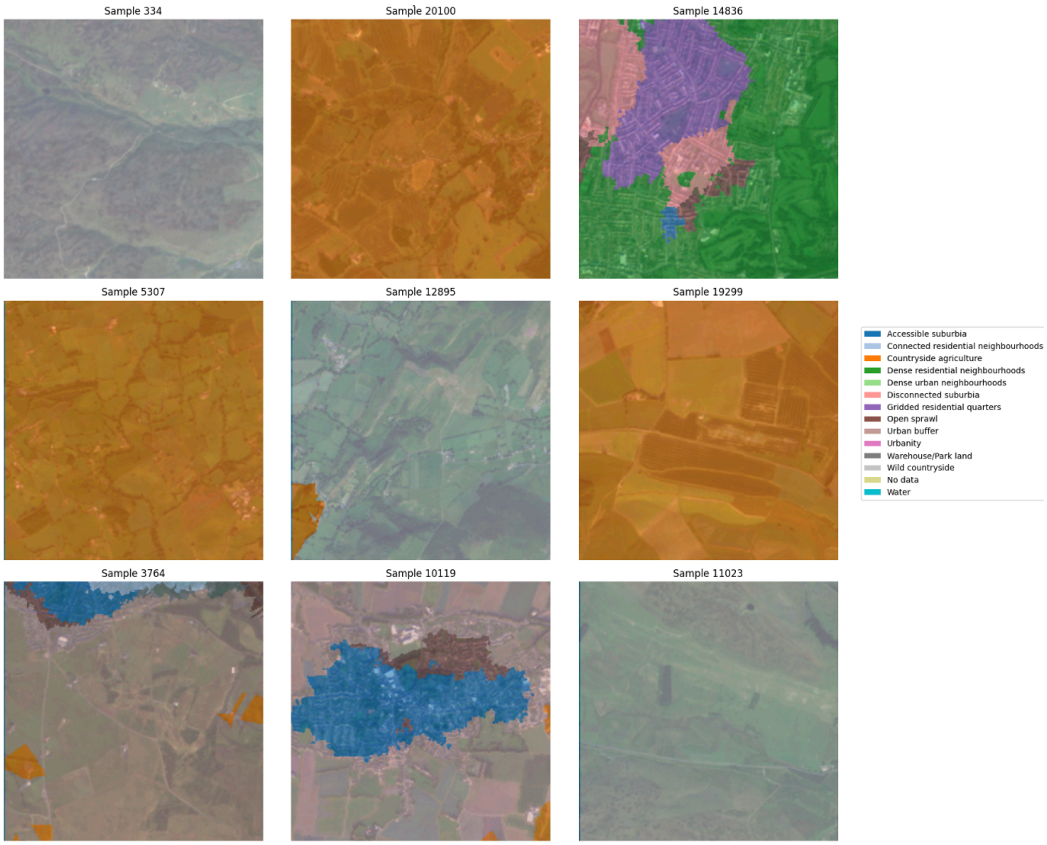
5.1.1 Prediction example: London

The following figure shows an example of a prediction for the London area using the whole dataset. Each colour represents a different signature. The background colour represents the ground truth.



6 Discussion

6.0.1 Spatial signatures: Example tiles



6.1 Key Findings

1. **Regional trends:** The integration of regional information into the model significantly improving model performance across all metrics.
2. **Model performance:**:

6.2 Challenges and Limitations

Our research encountered several significant challenges:

1. **Class Imbalance:** The dataset exhibited substantial variations in class representation, which impacted model performance. While focal loss helped address this issue, some classes remained challenging to predict accurately.
2. **Computational Resources:** Training times varied significantly between models, with Prithvi requiring substantially more computational resources. This presents practical considerations for model deployment and scaling.
3. **Comparison Complexity:** The different tile sizes between classification (42px) and segmentation (224px) approaches made direct comparisons challenging, requiring careful consideration of evaluation metrics and methodology.

7 Conclusions