Predicting Plant Species Distribution with a Multimodal Swin Transformer Network: A GeoLifeCLEF 2025 Report

Aman R. Syayfetdinov[1]

[1]Higher School of Economics (HSE), Moscow, Russian Federation

Abstract

Predicting the spatial and temporal distribution of plant species is a key challenge for biodiversity monitoring and conservation planning. In this report, we present the solution to the GeoLifeCLEF 2025 challenge, which requires predicting the presence of plant species using satellite images and time series, climate time series and other rasterized environmental data. Our approach utilizes multimodal network with encoders for images, climate and satellite time series. During inference, we apply a fixed probability threshold to produce multi-label predictions. Without any pseudo-labeling or ensembling, our model achieves a macro-F1 score of 0.218 on the public leaderboard and 0.192 on the private leaderboard, placing us 6th place. We analyze the impact of each modality and discuss ways for further improvement.

Keywords

Multimodal Deep Learning, Species distribution modeling, Biodiversity, LifeCLEF

1. Introduction

Species distribution modeling [1] plays a crucial role in biodiversity conservation by predicting species occurrence probabilities across spatial-temporal contexts. The recent spread of geolocated species observations, which cover thousands of species, has created opportunities for data-driven approaches. The GeoLifeCLEF 2025 competition [2], part of the LifeCLEF 2025 lab [3] and FGVC12 workshop organized in conjunction with the CVPR 2025 conference, leverages this potential through a large-scale multimodal prediction task: identifying plant species likely observed in different locations using various environmental data [4].

The GeoLifeCLEF 2025 competition presents two main complexities: extreme data heterogeneity and challenging labeling constraints. It is similar to previous editions [5, 6] with a training dataset comprising 90,000 surveys across Europe, each documenting observed plant species (from 5,000 unique taxa) alongside multimodal environmental descriptors. These include satellite imagery (RGB/NIR patches), bioclimatic data cubes, time series climate measurements, land cover classifications, updated human footprint indices, soil properties and elevation data. Furthermore, only a small subset provides complete species labels (Presence-Absence data), while the majority (about 5M records) offer only single positive annotations (Presence-Only data). This creates a strong partial-label scenario for multi-species prediction. In this year, the

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

asysyfetdinov@gmail.com (A. R. Syayfetdinov)

0009-0005-5170-0829 (A. R. Syayfetdinov)

test set includes 10,000 additional plots, resulting in almost 14K overall, with significant location shifts, inducing domain adaptation challenges.

In this paper we introduce an adaptive prediction method that combines a tuned probability threshold with a top-k fallback mechanism. This approach provides more contextually relevant predictions than a fixed-k method while preventing overly sparse outputs. We demonstrate that a single, non-ensembled model can achieve a top-tier ranking. Our experiments show that strategically focusing on the richest data modalities and curating the training set is more effective than including all available data.

2. Data

The GeoLifeCLEF 2025 dataset contains presence-absence (PA) and presence-only (PO) observations. PO data includes about 5 million observations and reports only the presence and not absence of certain plant species in specific areas. However, PA data combines around 90K surveys with about 5K unique species of the European flora and reports the presence and absence of plant species. The total number of surveys in the test set was approximately 16K.

In PA train data, species distribution exhibits extreme imbalance: 50% of taxa have 16 occurrences, while only 20% exceed 110 observations. Geographically, the data skew towards western Europe, a map of the locations can be seen in Figure 1. More detailed descriptions can be found on the competitions' homepage[1].

Each survey pairs GPS coordinates with some environmental data:

- **Satellite imagery**: 128m×128m Sentinel-2 RGB/NIR patches (10m resolution) + Landsat time-series (6-band quarterly composites);
- **Climatic data**: 19 bioclimatic rasters (1km resolution);
- **Soil variables**: 19 SoilGrids properties (e.g., pH, organic carbon);
- **Human footprint**: 16 time-dynamic pressure indices (1993/2009);
- **Land cover**: (500m resolution) and elevation (30m resolution);

3. Evaluation Metric

The competition utilizes a macro-averaged F_1-score to evaluate species presence predictions. This metric provides equal weight to all species, mitigating bias from class imbalance by independently estimating each taxon's performance. For each test plot i with true species set Y_i, the model predicts ranked presence probabilities $P_{i,1}, P_{i,2}, \ldots, P_{i,R_i}$. After thresholding these probabilities to obtain binary predictions, the metric computes:

$$\text{Macro } F_1 = \frac{1}{C} \sum_{c=1}^{C} F_{1,c}$$

[1] https://www.kaggle.com/competitions/geolifeclef-2025/data

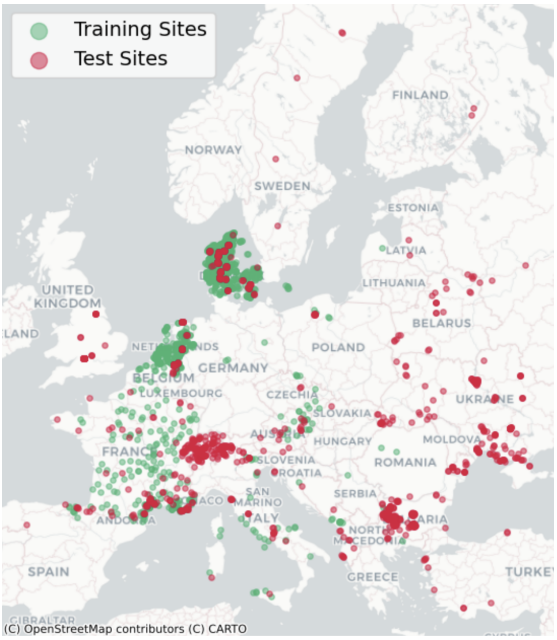

Figure 1: Geographic distribution of training (green) and test (red) presence-absence data for the GeoLifeCLEF 2025 challenge. The map highlights the spatial skew of the training data towards Western Europe and the significant geographical shift of the test set locations.

, where

$$F_{1,c} = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_{i,c}}{TP_{i,c} + (FP_{i,c} + FN_{i,c})/2}$$

Here $TP_{i,c}$, $FP_{i,c}$ and $FN_{i,c}$ are the true positive, the false positive and the false negative of the i-th input sample, respectively, while C represents all 5K species. This formulation emphasizes balanced performance across rare and common taxa.

4. Methodology

This section describes our proposed multimodal network and the methods that were tried during this competition.

4.1. Model architectures

Our solution for the GeoLifeCLEF 2025 challenge was based on the baselines provided by the organizers in current and previous competitions. However, through experimentation, we determined that our approach from the previous year [7], which used vectors of satellite time series and scalar environmental values, did not yield performance improvements this time.

Our final model, illustrated in Figure 2, is a multimodal neural network that exclusively integrates three rich data sources: Sentinel-2 satellite imagery, bioclimatic data cubes, and Landsat data cubes. The core of our architecture consists of three specialized encoders, one

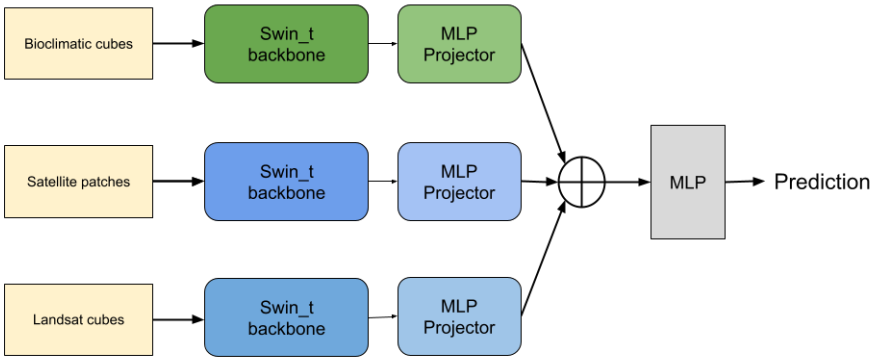

Figure 2: The model consists of three parallel streams, each with a specialized Swin Transformer encoder for Bioclimatic, Sentinel-2, and Landsat data. Each encoder's output is transformed by a projection head into a 1000-dimensional embedding. These embeddings are then fused via concatenation and fed into an MLP classification block, which outputs probabilities for 3,425 species (those with an occurrence count greater than five).

for each data modality, based on a modified Swin Transformer [8] architecture. Each encoder processes its respective data type to generate a feature embedding, which are then fused and passed to a final classification head to predict the species.

The Landsat data encoder utilizes a Swin_t model trained from scratch. To prepare the data, we first apply an initial LayerNorm to stabilize the input distribution. The primary modification involves adapting the model's first convolutional layer (the patch embedding layer) to accept 6 input channels instead of the default 3. This allowed to accommodate the rich multi-spectral and temporal information. To simplify the model structure and focus on feature extraction, we replace its final classification head with an Identity layer, which outputs a 768-dimensional feature vector.

Similarly, the encoder for the 19 bioclimatic variables (structured as a 4-channel input) uses a Swin_t architecture also trained from scratch. It is preceded by a LayerNorm layer, accommodating the diversity of Bioclim data, and its initial patch embedding layer is modified to handle 4 input channels. Like the Landsat encoder, its classification head is removed to output high-level features.

To leverage the high-resolution multispectral Sentinel-2 imagery (Red, Green, Blue, and Near-Infrared bands), we employed a Swin_t model pre-trained on the ImageNet-1K dataset [9]. This supervised pre-training provides a powerful starting point for feature extraction. We adapted the pre-trained model for 4-channel input by modifying the first convolutional layer's weights. The weights for the original three channels (RGB) were preserved, while the weights for the new fourth channel (NIR) were initialized by averaging the RGB channel weights. This strategy retains the valuable features learned during pre-training while accommodating the additional spectral band. The classification head was also replaced with an Identity layer.

The 768-dimensional feature vectors produced by each of the three Swin_t backbones are first passed through separate projection heads. Each projection head consists of a Linear layer, BatchNorm1d, a GELU [10] activation, and Dropout [11], transforming each feature vector into a 1000-dimensional representation.

These projected features are then concatenated into a single 3000-dimension vector. This fused vector is fed into a final classification head—a multi-layer perceptron (MLP) with GELU [10] activations and Dropout [11] —which produces the ultimate species predictions. This multimodal fusion design ensures that the model learns to combine cues from all three data sources effectively. Full implementation can be found in Kaggle[2].

4.2. Training and inference

To manage the highly imbalanced and long-tailed distribution of species occurrences in the training data, we first filtered the dataset. We focused on plant species with more than 5 recorded occurrences, which reduced the number of target classes from the original 5,016 to a more manageable 3,425. This occurrence threshold was empirically determined through experimentation to optimize the trade-off between class coverage and model stability.

The model was trained on the Presence-Absence (PA) data for 12 epochs. We employed the AdamW optimizer with a learning rate of 8e-5 and a Cosine Annealing Learning Rate Scheduler (CosineAnnealingLR) to promote stable convergence. Given the multi-label nature of the problem (multiple species can be present at one location), we used Binary Cross-Entropy (BCE) loss. Training was conducted with a batch size of 300.

For the final predictions on the test set, we developed a hybrid, threshold-based strategy that deviates from the baseline approach of simply predicting a fixed number of the most probable species. Our method is designed to be more adaptive to the local biodiversity of each observation point. After passing the test data through the model, we apply a Sigmoid function to the output logits to obtain a probability score between 0 and 1 for each of the 3,425 species. We then apply a probability threshold of 0.18. Any species with a score exceeding this threshold is classified as present. This threshold was carefully tuned on a validation set to balance precision and recall. To ensure a minimum number of predictions for each observation and avoid generating overly sparse results, we implemented a fallback mechanism. If the number of species predicted using the 0.18 threshold is fewer than 14, we discard those predictions and instead select the 14 species with the highest probability scores for that specific observation. This fallback value was chosen as it approximates the median number of species per plot in the training data, providing a data-driven baseline that prevents overly conservative predictions while being robust to outliers.

5. Experimental results

5.1. Experimental settings

To tune hyperparameters and validate our architectural choices, we partitioned the official training set into a training subset (80%) and a validation subset (20%). The training subset was

[2]https://www.kaggle.com/code/lonansyayf/2025-model-geolifeclef

Table 1
Training configuration and hyperparameters for our final model.

Parameter	Value / Setting
Batch and Epochs	
Batch Size	300
Epochs	12
Optimizer	
Optimizer Algorithm	AdamW
Initial Learning Rate	8×10^{-5}
LR Scheduler	Cosine Annealing

Table 2
Ablation study of input modalities.

Bioclim.	Sentinel	Feature	Landsat	F1-Score	
				Public	Private
✓	✓	-	✓	**0.197**	**0.170**
✓	✓	✓	✓	0.188	0.163
-	✓	✓	✓	0.191	0.165
✓	✓	✓	-	0.194	0.168
✓	-	✓	✓	0.185	0.160

used for model optimization, while the validation subset provided an unbiased estimate of performance for model selection. For comparing different model versions during this development phase, we benchmarked performance using the top-25 most probable species for each observation. This standardized metric allowed for a consistent and direct comparison between models, removing the potential bias introduced by our custom probability thresholding strategy (described in Section 4.2). The detailed settings of training are shown in Table 1.

5.2. Imbalanced data

The dataset exhibits a strong long-tailed distribution, where most species have far fewer presence records than absence records. We explored several mitigation techniques. While methods like applying a pos_weight to the Binary Cross-Entropy (BCE) loss were tested, they did not yield significant improvements. Our most successful strategy was a multi-faceted approach combining data curation and regularization. We constrained the training problem by focusing on species with an occurrence count greater than 5. This reduced the number of classes from 5,016 to 3,425, allowing the model to learn more robust features for species with a reasonable number of examples. Table 3 shows the impact of species filtering on model performance. Our second strategy involved the hybrid inference approach detailed in Section 4.2: we applied a tuned probability threshold and used a top-14 fallback for observations with sparse predictions. The performance of different thresholding strategies is presented in Table 4.

Table 3
Score depending on the number of occurrences of plant species for model training

Minimum Occurrences	No. of Species	F1-Score	
		Public	Private
> 0 (All species)	5096	0.197	0.170
> 5	3425	0.199	0.172
> 10	2857	0.195	0.168
> 15	2511	0.192	0.166

Table 4
Score depending on the presence probability threshold.

Thresholding Strategy	F1-Score	
	Public	Private
Top-25 Predictions	0.198	0.171
0.10	0.197	0.170
0.15	0.200	0.173
0.18	0.204	0.178
0.20	0.201	0.173
0.25	0.198	0.170

5.3. Encoder Architecture

A key finding from our ablation studies was the superior performance of the Swin Transformer [8] architecture (see Table 5). We conducted comparative experiments using a standard ResNet18 [12] model as the backbone for each modality encoder. The Swin_t-based encoders consistently outperformed their ResNet-based counterparts on our validation set. This suggests that the Swin Transformer's hierarchical feature representation and its ability to model long-range dependencies are particularly well-suited for extracting discriminative information from the complex spatial patterns found in satellite, bioclimatic, and Landsat data cubes.

A key strategy for improving overall score was different model regularization. To prevent overfitting and improve generalization, we integrated standard image augmentations (such as rotation and random brightness/contrast adjustments). We utilized Dropout [11] along whole network and Batch Normalization in projector layers after modality encoders.

6. Conclusion

In this paper, we presented the multimodal deep learning framework that secured 6th place in the GeoLifeCLEF 2025 competition. Our approach integrated Sentinel-2 imagery, Landsat time-series, and bioclimatic data cubes using a network of specialized Swin Transformer encoders. We demonstrated that this architecture is particularly effective for processing complex, rasterized environmental data. The success of our method hinged on three key contributions: the tailored Swin Transformer backbones, a pragmatic data curation strategy that prioritized signal quality

Table 5
Ablation study results.

Model / Component	Private Score	Δ Score
Baseline	0.170	—
+ Subset of train	0.172	0.002
+ Positive threshold (0.18)	0.178	0.006
+ Model regularizations	0.184	0.006
+ **Swin-T encoders**	0.192	0.08

over data quantity, and a novel hybrid inference method combining a tuned probability threshold with a top-18 fallback.

To support reproducibility and encourage further research, the complete source code for our solution is publicly available on Kaggle[3]. Looking ahead, several avenues for enhancement exist. While our model achieves strong performance, a more exhaustive hyperparameter optimization could yield further gains. The most significant opportunity for improvement, however, lies in data enrichment. A key future direction would be to revisit the integration of the Presence-Only (PO) data, potentially using advanced techniques to correct for sampling bias, which could substantially improve the model's generalization across rare species. Finally, incorporating additional environmental data layers and exploring more sophisticated data augmentation methods remain promising areas for future research.

7. Declaration on Generative AI

During the preparation of this work, the author used DeepSeek to check grammar and spelling, as well as to paraphrase and reword the text. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] N. E. Zimmermann, T. C. Edwards Jr., C. H. Graham, P. B. Pearman, J.-C. Svenning, New trends in species distribution modelling, Ecography 33 (2010) 985–989. doi:10.1111/j.1600-0587.2010.06953.x.

[2] L. Picek, C. Leblanc, T. Larcher, M. Servajean, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2025: Plant species presence prediction with environmental and high-resolution remote sensing data, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[3] L. Picek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification,

[3] https://www.kaggle.com/code/lonansyayf/2025-model-geolifeclef

in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2025.
[4] L. Picek, C. Botella, M. Servajean, R. Palard, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Geoplant: Spatial plant species prediction dataset, NEURIPS, 2024.
[5] L. Picek, C. Botella, M. Servajean, B. Deneu, D. Marcos Gonzalez, R. Palard, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of geolifeclef 2024: Species composition prediction with high spatial resolution at continental scale using remote sensing, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
[6] C. Botella, B. Deneu, J. Estopinan, M. Servajean, D. Marcos Gonzalez, A. Joly, Overview of GeoLifeCLEF 2023: Species presence prediction based on occurrence data and high-resolution remote sensing images, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[7] A. Syayfetdinov, Multimodal networks for species distribution modeling, 2024. URL: https://ceur-ws.org/Vol-3740/paper-208.pdf.
[8] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021. doi:10.1109/ICCV48922.2021.00986.
[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, ImageNet: a Large-scale hierarchical image database, in: Conference: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[10] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs) (2016). arXiv:1606.08415.
[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
[12] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Conference: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.