

# Multimodal networks for Species Distribution Modeling

Notebook for the LifeCLEF Lab at CLEF 2024

Aman R. Syayfedinov<sup>1</sup>

<sup>1</sup>Moscow Institute of Physics and Technology (MIPT), Dolgoprudny, Russian Federation

## Abstract

Understanding the spatial and temporal distribution of plant species is important for many biodiversity management and conservation scenarios. This paper presents solution to the GeoLifeCLEF challenge, which involves prediction of the presence of plant species using satellite images and time series, climate time series and other rasterized environmental data. Multimodal model leveraged satellite images, bioclimatic cubes and feature vectors of satellite time series and environmental scalar values. With the selected presence probability threshold for inference this method allowed to reach  $F_1$ -score of 0.347 on public and 0.345 on private leaderboard, placing us 9th on the leaderboard.

## Keywords

Species distribution modeling, Biodiversity, LifeCLEF

This is a technical report for final contribution to the GeoLifeCLEF 2024 challenge, submitted under pseudonym “Lonan Syayf”, with which the ninth place was obtained (out of 51 competitors) on the private leaderboard.

## 1. Introduction

The GeoLifeCLEF 2024 competition [1] is held jointly as part of the LifeCLEF 2024 lab [2] and the FGVC11 workshop. Just like in the GeoLifeCLEF 2023 competition [3] the goal is to predict a list of species most likely to be observed at a given location using various geographical and environmental data such as satellite images and time series, climatic time series, and other rasterized data: land cover, human footprint, bioclimatic, and soil variables. Typically, the task of species distribution modelling [4] has challenges associated with imbalances in species presence and absence in the data, large-scale multimodal learning, and plant species diversity. Its results could be useful for predicting biodiversity change and mitigating environmental pressures from human activities.

The GeoLifeCLEF 2024 training data includes a collection of observations of plants in Europe. Each survey consists of a list of plant species with the GPS coordinates and a set of variables characterizing the landscape and environment around them. There are around 90K surveys with around 5K unique plant species in the dataset.

---


CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

✉ asyayfedinov@gmail.com (A. R. Syayfedinov)

🆔 0009-0005-5170-0829 (A. R. Syayfedinov)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This technical report presents selected approach to the competition, which is a multimodal network based on bioclimatic cubes, sentinel image patches (RGB-patch and NIR-patch) and vector of climate, elevation, human footprint, land cover, soilgrid and landsat time series data. Traing code can be found here<sup>1</sup>.

## 2. Data and Evaluation Metric

Data plays an important role in prediction plant species distribution in a given location and time. In this section, we briefly present the data and the evaluation metric used for the competition.

### 2.1. Data

This paragraph is simply a description of the standard GeoLifeCLEF 2024 dataset. The training dataset contains presence-absence (PA) surveys and presence-only (PO) surveys. PO data includes about 5 million observations and reports only presence and not absence of certain plant species in specific areas. On the other hand, PA data combines around 90K surveys with about 5K unique species of the European flora and reports presence and absence of plant species. In solution only presence-absence surveys were used and everywhere below the report will only be about this type of data. The total number of surveys in the test set was 5K.

Training dataset distribution of the number of observations of each plant species is shown in Figure 1. Almost 50% of plant species in training data have a number of occurrences less than 16 and only 20% have more than 110 occurrences. Almost all observations were made in Western Europe, a map of locations can be seen in Figure 2. More detailed descriptions can be found at competitions's homepage<sup>2</sup>.

Each survey is paired with the following covariates:

- Satellite image patches: 128m×128m RGB-NIR patches centered at each observation, at a resolution of 1 meter per pixel;
- Satellite time series: Up to 20 years of values for six satellite bands (R, G, B, NIR, SWIR1, and SWIR2);
- Environmental rasters Various climatic, pedologic, land use, and human footprint variables at the European scale. It was provided as scalar values, time-series, and original rasters;

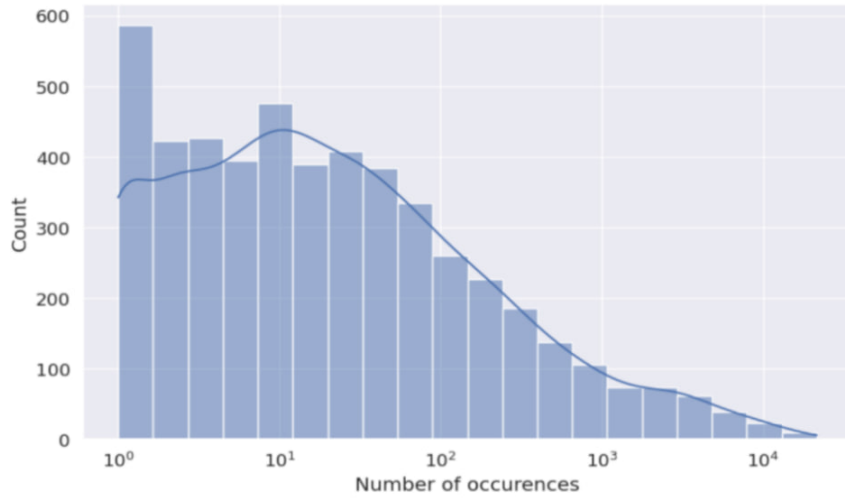
### 2.2. Evaluation Metric

The evaluation metric for the GeoLifeCLEF 2024 competition is the samples-averaged  $F_1$ -score computed on a set made of species presence-absence samples. The  $F_1$ -score is an average measure of overlap between the predicted and actual set of species present at a given location and time. Each observation  $i$  is associated with a list of ground-truth labels  $Y_i$  corresponding

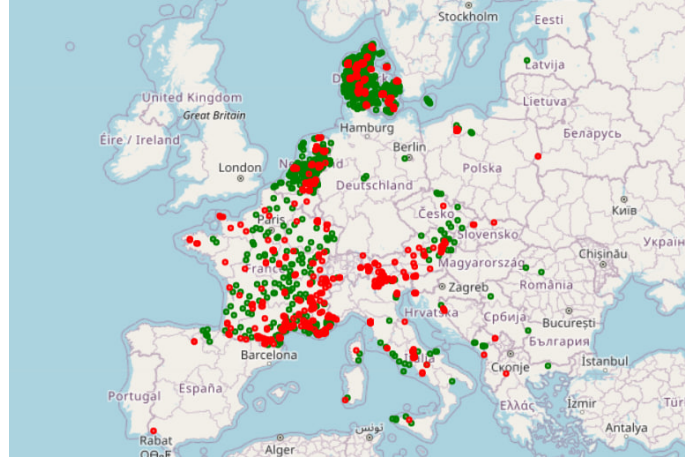
---

<sup>1</sup><https://www.kaggle.com/code/lonansyayf/baseline-with-modifications/notebook>

<sup>2</sup><https://www.kaggle.com/competitions/geolifeclef-2024/data>



**Figure 1:** Histogram for distribution of the occurrences of plant species in the training dataset. Horizontal axis on a logarithmic scale for better understanding.



**Figure 2:** Map of Europe with observation distribution. The train data location is green point, the test data is red points.

to the observed plant species. For each observation, the submissions provide a set of species predicted presence  $P_{i,1}, P_{i,2}, \dots, P_{i,R_i}$ . The micro  $F_1$ -score is then computed using:

$$F_1 = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + (FP_i + FN_i)/2}$$

where  $TP_j$ ,  $FP_j$  and  $FN_j$  are the true positive, the false positive and the false negative of the  $j$ -th input sample, respectively.  $N$  is the number of samples for evaluation.

### 3. Methodology

This section describes the methods that were tried during the competition. Strategy was centered around the baseline model<sup>3</sup> provided by the competition organizers. The baseline  $F_1$ -score is 0.31 on the public set. This model leveraged all environmental data and utilized a multimodal neural network with separated features extractors to return a single prediction set in order to take advantage of every modality (satellite images, bioclimatic cubes, landsat cubes). The main change was to replace landsat cubes with a vector of satellite time series and environmental scalar values, everywhere below it is called feature vector. In addition, plant species with an occurrence number greater than 10 was used to train the model.

#### 3.1. Feature vector

Feature vector consists of climate, elevation, human footprint, land cover, soilgrid and landsat time series data. Methods for compiling this data are taken from the public notebook<sup>4</sup>. Climatic time series data was merged within a 10-year time window. Some positions had missing values, which were filled with spatial interpolation. It appeared that there were densely populated measurements near the missing regions, so missing values were filled with values from the nearest neighbors. Finally, each survey had 1198 values of feature vector. The train and test versions can be found here. Before going to model feature vectors are normalized with standard scaler.

#### 3.2. Model architecture

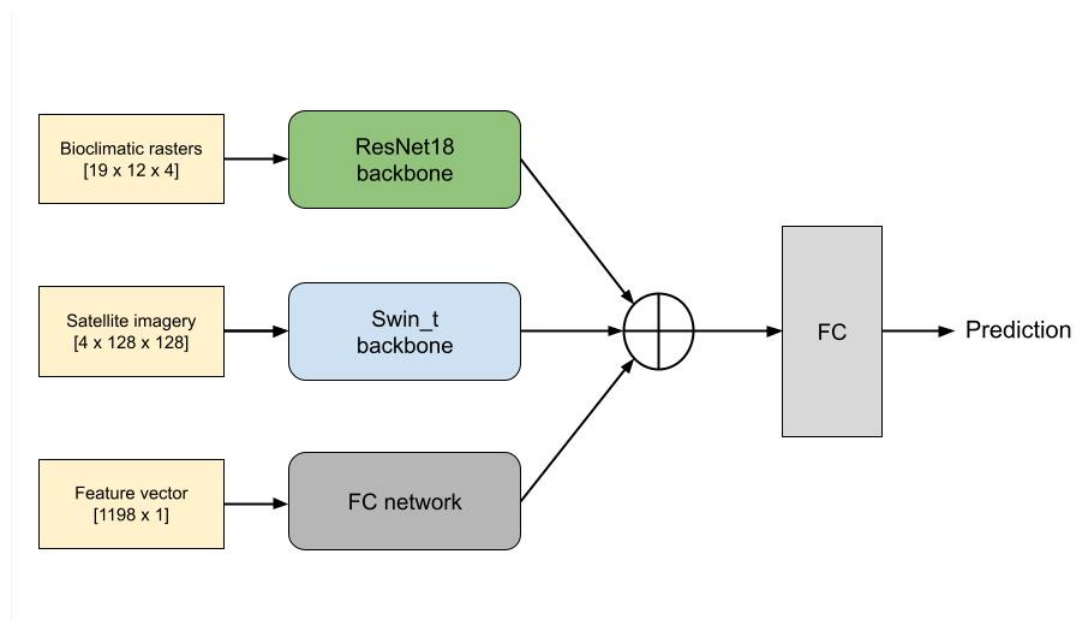
The architecture closely follows the baseline model, incorporating a multimodal neural network that utilizes three distinct feature extractors for bioclimatic rasters (19 channels), satellite images (4-channel RGB with NIR), and feature vectors (1198 channels). These outputs are combined and processed through fully connected layers to generate predictions. The first bioclimatic head involves layer normalization, ResNet18 [5] without pretrained weights, and a dropout [6] with a 0.1 probability. The second image head employs a , swin transformer [7] model with ImageNet [8] weights and a dropout layer with a 0.1 probability. Prior to this stage, image data undergo augmentation techniques like random rotation, random brightness contrast, and normalization. The third head comprises a sequence of layer normalization and three linear layers with GELU [9] activation function, along with dropout set at a 0.1 probability (the first layer mapping from 1198 to 1198, the second and the third layers map to 1000 outputs). Subsequently, the bioclimatic and feature outputs are normalized and combined with the image output. The final classifier is constructed with three linear layers utilizing GELU activation function and dropout at a 0.1 probability.

#### 3.3. Training and inference

The model was trained on PA data for 12 epochs using the Adam optimizer with a learning rate of  $8e-5$  and binary cross entropy (BCE) loss and batch size equal to 128. During training, we

<sup>3</sup><https://www.kaggle.com/code/picekl/sentinel-landsat-bioclim-baseline-0-31626>

<sup>4</sup><https://www.kaggle.com/code/gobyonggeon/preprocess-visualize-spatial-data-eda-xgb>



**Figure 3:** Selected multi-model architecture. Bioclimatic, image and feature heads mapping to 1000, 768, 1000 outputs, respectively. Then stacked outputs pass through linear layers mapping to the 2857 species (species with occurrence number > 10)

focused on plant species with an occurrence number greater than 10, resulting in 2857 unique species out of a total of 5015. It's important to highlight that the occurrence threshold value was determined through experimentation.

In final approach to inference, the strategy used in the baseline notebook was changed. Rather than forecasting the 25 most probable species for every observation in the test dataset, selected threshold of 0.18 was used. This threshold determined that species with probabilities surpassing this value were classified as present. Additionally, test observations featuring fewer than 4 represented species was assigned with the 4 most likely species.

## 4. Experimental results

### 4.1. Experimental settings

Experiments were conducted with the multimodal network described in Section 3.2. The detailed settings of training are shown in Table 1. For comparing different versions of models we used 25 most probable species to remove bias with probability threshold described in Section 3.3.

### 4.2. Usage of feature vector

In order to investigate the impact of using the feature vector head we conducted ablation study. Table 2 represents the detailed results. It seems that with selected hyperparameters combination

**Table 1**  
Frequency of Special Characters

Hyper-parameters	
Batch size	128
Optimizer	Adam
Learning rate	8e-5
Lr scheduler	CosineAnnelingLR
Number of epochs	12

**Table 2**  
Ablation study of usage the feature vector head

Bioclimatic head	Image head	Feature head	Landsat head	$F_1$ -score	
				Public	Private
✓	✓	-	✓	0.315	0.316
✓	✓	✓	✓	0.317	0.317
-	✓	✓	✓	0.306	0.311
✓	✓	✓	-	<b>0.322</b>	<b>0.323</b>

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

Species with number of occurrences	$F_1$ -score	
	Public	Private
>0 (5096 in total)	0.322	0.323
>5 (3425 in total)	0.322	0.326
>10 (2857 in total)	<b>0.326</b>	<b>0.329</b>
>15 (2511 in total)	0.324	0.328

of bioclimatic, image and feature heads gives the best performance of around 0.32 on both public and private scores. The performances of other configurations are about 0.31 or less.

### 4.3. Imbalanced data

As was mentioned before, the dataset is strongly unbalanced, which means that for almost all species the number of observations detecting their presence is much less than the number of observations detecting their absence. we tried to solve this problem in different ways, for example, adding pos\_weight to bce loss, adding different data augmentation. The final option was to limit the number of species on which the model is trained, taking only those with occurrence number greater than 10. Table 2 shows how the score depends on the threshold for the occurrence number. Another thing was lowering the threshold for a species having a probability higher than which it was considered present. For those observations that had fewer than 4 species present we assigned the 4 most likely plant species. Results of different probability thresholds are presented in Table 3.

**Table 4**  
Score depending on the presence probability threshold

Probability threshold	$F_1$ -score	
	Public	Private
0.4	0.309	0.303
0.3	0.334	0.332
0.2	0.346	0.345
0.15	<b>0.345</b>	<b>0.342</b>
0.1	0.329	0.327

## 5. Conclusion

We presented the working principles of submission to the GeoLifeCLEF 2024 challenge and discussed some of the key findings of the results. We have not conducted an expansive, let alone exhaustive hyperparameter search and believe that doing so could raise performance a bit. The main achievement was to use proper model architecture, choosing training data and changing the inference strategy. In final solution, we did not use PO data and training strategies used in previous years [10, 11]. Obviously, using more data would help for better generalization and it is certainly high on the list of improvements that need to be made. Also, possible improvements can be achieved by additionally searching for better backbone models, like Inception-v4 [12] or Vision Transformer, ViT B / 16 [13] for different modalities and using an ensemble of various models.

## References

- [1] 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 presence prediction based on occurrence data and high-resolution remote sensing images, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.
- [2] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Šulc, M. Hruží, M. Servajean, J. Matas, et al., Overview of lifeclef 2024: Challenges on species distribution prediction and identification, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
  - [7] 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.
  - [8] 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.
  - [9] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs) (2016). arXiv:1606.08415.
  - [10] H. Ung, R. Kojima, S. Wada, Leverage samples with single positive labels to train CNN-based models for multi-label plant species prediction, 2023.
  - [11] B. Kellengerger, D. Tuia, Block label swap for species distribution modelling, 2022.
  - [12] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-ResNet and the impact of residual connections on learning, *AAAI Conference on Artificial Intelligence* 31 (2016). doi:10.1609/aaai.v31i1.11231.
  - [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, 2021. arXiv:2010.11929.