Overview of GeoLifeCLEF 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data

Titouan Lorieul[1], Elijah Cole[2], Benjamin Deneu[1], Maximilien Servajean[3], Pierre Bonnet[4] and Alexis Joly[1]

[1]*Inria, LIRMM, Univ Montpellier, CNRS, Montpellier, France*
[2]*Department of Computing and Mathematical Sciences, Caltech, USA*
[3]*LIRMM, AMI, Univ Paul Valéry Montpellier, Univ Montpellier, CNRS, Montpellier, France*
[4]*CIRAD, UMR AMAP, Montpellier, France*

Abstract

Understanding the geographic distribution of species is a key concern in conservation. By pairing species occurrences with environmental features, researchers can model the relationship between an environment and the species which may be found there. To advance research in this area, a large-scale machine learning competition called *GeoLifeCLEF 2022* was organized. It relied on a dataset of 1.6 million observations from 17K species of animals and plants. These observations were paired with high-resolution remote sensing imagery, land cover data, and altitude, in addition to traditional low-resolution climate and soil variables. The main goal of the challenge was to better understand how to leverage remote sensing data to predict the presence of species at a given location. This paper presents an overview of the competition, synthesizes the approaches used by the participating groups, and analyzes the main results. In particular, we highlight the ability of remote sensing imagery and convolutional neural networks to improve predictive performance, complementary to traditional approaches.

Keywords

LifeCLEF, evaluation, benchmark, biodiversity, presence-only data, environmental data, remote sensing imagery, multi-modal data, species distribution, species distribution models

1. Introduction

In order to make informed conservation decisions, it is essential to understand where different species live. Citizen science projects now generate millions of geo-located species observations every year, covering tens of thousands of species. But how can these point observations be used to predict what species might be found at a new location?

A common approach is to build a *species distribution model* (SDM) [1], which uses a location's *environmental covariates* (e.g., temperature, elevation, land cover) to predict whether a species

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 21–24, 2022, Bucharest, Romania
titouan.lorieul@inria.fr (T. Lorieul); ecole@caltech.edu (E. Cole); benjamin.deneu@inria.fr (B. Deneu); servajean@lirmm.fr (M. Servajean); pierre.bonnet@cirad.fr (P. Bonnet); alexis.joly@inria.fr (A. Joly)
0000-0001-5228-9238 (T. Lorieul); 0000-0001-6623-0966 (E. Cole); 0000-0003-0640-5706 (B. Deneu); 0000-0002-9426-2583 (M. Servajean); 0000-0002-2828-4389 (P. Bonnet); 0000-0002-2161-9940 (A. Joly)

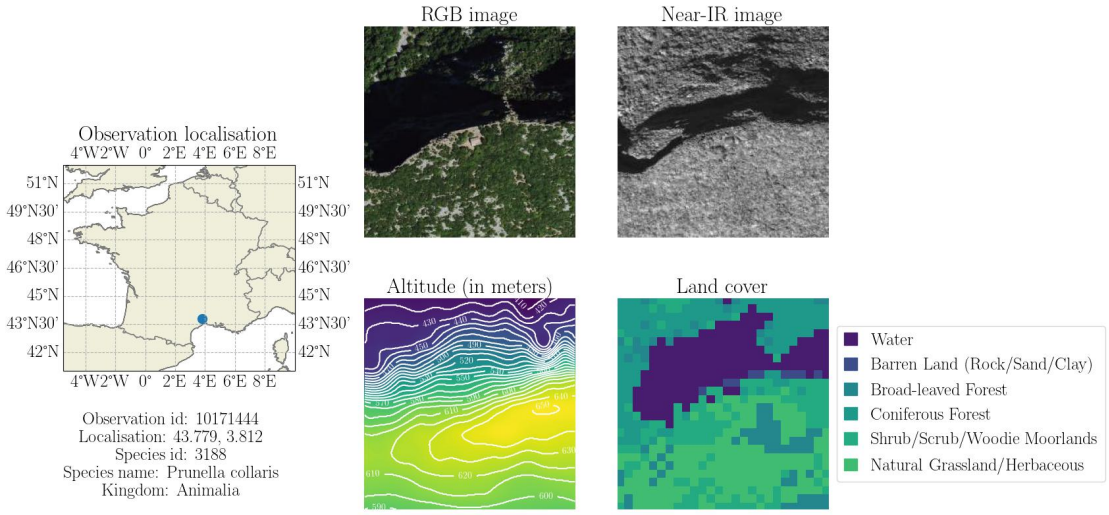


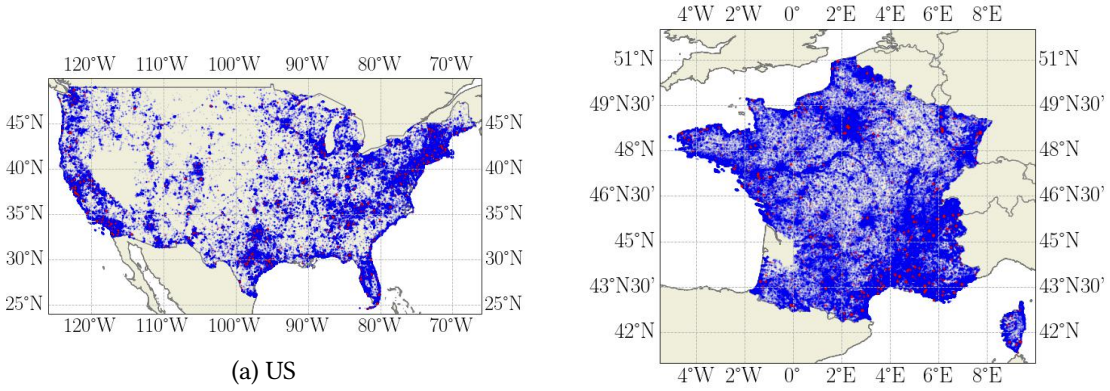
Figure 1: Illustration of the patches corresponding to observation `10171444` on the Pic Saint-Loup mountain, France. Each species observation is paired with high-resolution covariates (clockwise from top left: RGB imagery, NIR imagery, land cover, altitude).

is likely to be found there. Once trained, the model can be used to make predictions for any location where those covariates are available.

Developing an SDM requires a dataset where each species observation is paired with a collection of environmental covariates. However, many existing SDM datasets are both highly specialized and not readily accessible, having been assembled by scientists studying particular species or regions. In addition, the provided environmental covariates are typically coarse, with resolutions ranging from hundreds of meters to kilometers per pixel.

In this work, we present the results of the GeoLifeCLEF 2022 competition which is part of the LifeCLEF evaluation campaign [2] and co-hosted in Ninth Workshop on Fine-Grained Visual Categorization (FGVC9)[1] at CVPR 2022. This competition is the fifth GeoLifeCLEF challenge. In the first two editions, GeoLifeCLEF 2018 [3] and GeoLifeCLEF 2019 [4], each observation was associated only with environmental features given as vectors or patches extracted around the observation. Like the two last year's campaigns (GeoLifeCLEF 2020 [5] and GeoLifeCLEF 2021 [6]), GeoLifeCLEF 2022 is aimed at bridging the previously mentioned gaps by (i) sharing a large-scale dataset of observations paired with high-resolution covariates and (ii) defining a common evaluation methodology to measure the predictive performance of models trained on this dataset. The dataset contains over 1.6 million observations of plant and animal species. Each observation is paired with high-resolution remote sensing imagery—see Figure 1—as well as traditional environmental covariates (i.e., climate and soil variables). To the best of our knowledge, GeoLifeCLEF dataset is the largest publicly available dataset to pair remote sensing imagery with species observations. Our hope is that this analysis-ready dataset and associated evaluation methodology will (i) make SDM and related problems more accessible to machine learning researchers and (ii) facilitate novel research in large-scale, high-resolution,

[1] https://sites.google.com/view/fgvc9



(a) US

(b) France

Figure 2: Observations distribution over the US and France. Training observation data points are shown in blue while test data points are shown in red.

and remote-sensing-based species distribution modeling.

2. Dataset and evaluation protocol presentation

Data collection. The dataset used for the 2022 edition is a cleaned-up version of the data of the two previous years. A detailed description of the original GeoLifeCLEF 2020 dataset is provided in [7]. The following modifications were made for the 2022 version:

- Removed observations of species from kingdom different from *Plantae* and *Animalia* (29,240 observations, 2,072 species).
- Completed species metadata with genus, family, and kingdom information from GBIF.
- Kept only one observation when it is duplicated—same latitude, longitude, and species (110,556 observations removed).
- Removed all observations at exactly the same location—same latitude and longitude, different species (103,887 observations removed).
- Removed observations from species only present in the test set (208 observations, 188 species).
- Removed species with strictly less than 3 observations in the train set (13,336 observations, 9,913 species).
- `species_id` updated (not aligned with GeoLifeCLEF 2020 and 2021), in the end, 17,037 species are retained.
- Updated all altitude patches:
  - Re-extraction of patches using bi-cubic interpolation instead of bi-linear interpolation.
  - An issue which resulted in artifacts on a few altitude patches from past years was also fixed.

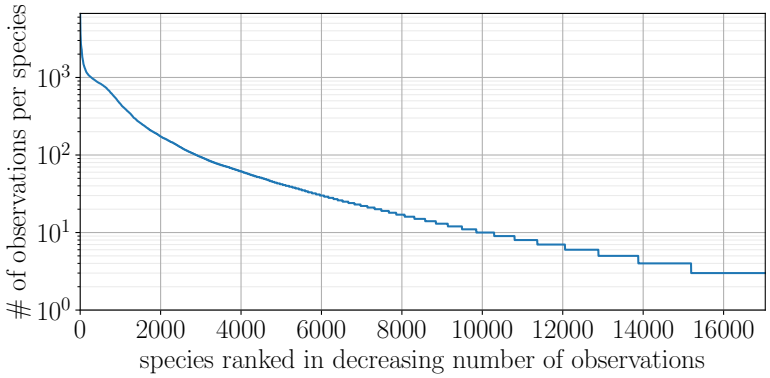


Figure 3: Distribution of observations across species highlighting the long-tail class-imbalance of the dataset.

The final GeoLifeCLEF 2022 dataset consists of 1,663,996 observations covering 17,037 plants and animal species distributed across US (975,357 observations, 14,135 species) and France (688,539 observations, 4,858 species), as shown in Figure 2. The number of observations of each species is not uniform and follows a long-tail distribution shown in Figure 3. Each species observation is paired with high-resolution covariates (RGB-NIR imagery, land cover, and altitude) as illustrated in Figure 1. These high-resolution covariates are re-sampled to a spatial resolution of 1 meter per pixel and provided as 256×256 images covering a 256m $\times$ 256m square centered on each observation. RGB-NIR imagery come from the 2009-2011 cycle of the National Agriculture Imagery Program (NAIP) for the US[2], and from the BD-ORTHO® 2.0 and ORTHO-HR® 1.0 databases from the IGN for France[3]. Land cover data originates from the National Land Cover Database (NLCD) [8] for the US and from CESBIO[4] for France. All elevation data comes from the NASA Shuttle Radar Topography Mission (SRTM)[5]. In addition, the dataset also includes traditional coarser resolution covariates: 19 bio-climatic rasters (30arcsec^2/pixel, i.e., 1km^2/pixel, from WorldClim [9]) and 8 pedologic rasters (250m^2/pixel, from SoilGrids [10]). The details of these rasters are given in Table 1.

Train-test split. The full set of occurrences was split into training and testing sets using a spatial block holdout procedure as illustrated in Figure 4. This limits the effect of *spatial auto-correlation* in the data [11]. Using this splitting procedure, a model cannot perform well by simply interpolating between training samples. The split was based on a global grid of 5km $\times$ 5km quadrats. 2.5% of these quadrats were randomly sampled and the observations falling in those formed the test set. 10% of those observations were used for the public leaderboard on Kaggle while the remaining 90% allowed to compute the private leaderboard providing the final results of the challenge. Similarly, another 2.5% of the quadrats were randomly sampled to provide an official validation set. The remaining quadrats and their associated observations were assigned to the training set.

[2]https://www.fsa.usda.gov
[3]https://geoservices.ign.fr
[4]http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/
[5]https://lpdaac.usgs.gov/products/srtmgl1v003/

Table 1
Summary of the low-resolution environmental variable rasters provided. The first 19 rows correspond to the bio-climatic variables from WorldClim [9]. The last 8 rows correspond to the pedologic variables from SoilGrid [10].

Name	Description	Resolution
bio_1	Annual Mean Temperature	30 arcsec
bio_2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	30 arcsec
bio_3	Isothermality (bio_2/bio_7) (* 100)	30 arcsec
bio_4	Temperature Seasonality (standard deviation *100)	30 arcsec
bio_5	Max Temperature of Warmest Month	30 arcsec
bio_6	Min Temperature of Coldest Month	30 arcsec
bio_7	Temperature Annual Range (bio_5-bio_6)	30 arcsec
bio_8	Mean Temperature of Wettest Quarter	30 arcsec
bio_9	Mean Temperature of Driest Quarter	30 arcsec
bio_10	Mean Temperature of Warmest Quarter	30 arcsec
bio_11	Mean Temperature of Coldest Quarter	30 arcsec
bio_12	Annual Precipitation	30 arcsec
bio_13	Precipitation of Wettest Month	30 arcsec
bio_14	Precipitation of Driest Month	30 arcsec
bio_15	Precipitation Seasonality (Coefficient of Variation)	30 arcsec
bio_16	Precipitation of Wettest Quarter	30 arcsec
bio_17	Precipitation of Driest Quarter	30 arcsec
bio_18	Precipitation of Warmest Quarter	30 arcsec
bio_19	Precipitation of Coldest Quarter	30 arcsec
orcdrc	Soil organic carbon content (g/kg at 15cm depth)	250 m
phihox	Ph x 10 in H20 (at 15cm depth)	250 m
cecsol	cation exchange capacity of soil in cmolc/kg 15cm depth	250 m
bdticm	Absolute depth to bedrock in cm	250 m
clyppt	Clay (0-2 micro meter) mass fraction at 15cm depth	250 m
sltppt	Silt mass fraction at 15cm depth	250 m
sndppt	Sand mass fraction at 15cm depth	250 m
bldfie	Bulk density in kg/m3 at 15cm depth	250 m

Evaluation metric. For each occurrence in the test set, the goal of the task was to return a candidate set of species likely to be present at that location. Due to the *presence-only* [12] nature of the observation data used during the evaluation of the methods, for each location in the test set, we only have the knowledge of the presence of one species—the one observed—among the different ones which can actually be found all together at that point. To measure the precision of the predicted sets while accommodating with this limited knowledge, a simple *set-valued classification* [13] metric was chosen as the main evaluation criterion: top-30 error rate. Each observation i is associated with a single ground-truth label y_i corresponding to the observed species. For each observation, the submissions provided 30 candidate labels $\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,30}$. The top-30 error rate is then computed using

$$\text{Top-30 error rate} = \frac{1}{N}\sum_{i=1}^{N} e_i \quad \text{where} \quad e_i = \begin{cases} 1 & \text{if } \forall k \in \{1,\ldots,30\},\ \hat{y}_{i,k} \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

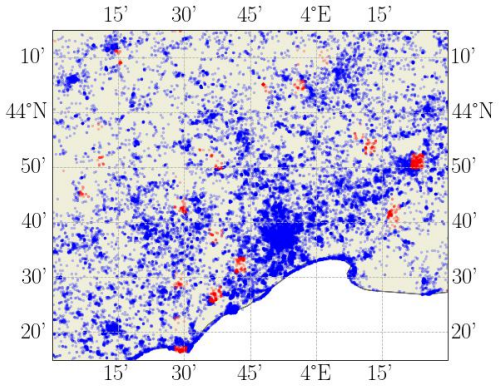


Figure 4: Observations located around Montpellier, France. Training observation data points are shown in blue while test data points are shown in red.

Note that this evaluation metric does not try to correct the sampling bias [14] inherent to present-only observation data (linked to the density of population, etc.). The absolute value of the resulting figures should thus be taken with care. Nevertheless, this metric does allow to compare the different approaches and to determine which type of input data and of models are useful for the species presence detection task.

Course of the challenge. The training and test data were publicly shared on 9th March 2022 through the Kaggle platform[6]. Any research team wishing to participate in the evaluation could register on the platform and download the data. Each team could submit up to 3 submissions per day to compete on the public leaderboard. A submission takes the form of a CSV file containing the top-30 predictions of the method being evaluated for all observations in the test set. For each submission, the top-30 error rate was first computed only on a subset of the test set to produce the public leaderboard which was visible to all the participants while the competition was still running. Once the submission phase was closed (25th May 2022), only 5 submissions per team were retained to compute the private leaderboard using the rest of the test set. These submissions were either hand-picked by the team or automatically chosen as the 5 best performing submissions on the public leaderboard. The participants could then see the final scores of all the other participants on the private leaderboard as well as their final ranking. Each participant was asked to provide a *working note*, i.e., a detailed report containing all technical information required to reproduce the results of the submissions. All LifeCLEF *working notes* were reviewed by at least two members of the LifeCLEF organizing committee to ensure a sufficient level of quality and reproducibility.

3. Baseline methods

Four baselines were provided by the organizers of the challenge to serve as comparison references for the participants while developing their own methods. They consisted in:

- **Top-30 most present species:** a constant predictor returning always the same list of

[6]https://www.kaggle.com/c/geolifeclef-2022-lifeclef-2022-fgvc9/

the most present species, i.e., the ones having the most occurrences in the training set.

- **RF on environmental variables:** a random forest (RF) model [15] trained on environmental feature vectors only, i.e., on the 27 climatic and soil variables extracted at the position of the observation (using scikit-learn [16] implementation with 100 trees of max depth 16).
- **CNN on 3-channels patches:** a ResNet-50 [17] convolution neural network (CNN) trained on the high-resolution 256×256 patches using PyTorch [18]. Two different baselines are provided:
  - **CNN on RGB patches:** standard ResNet-50 pre-trained on ImageNet taken from Pytorch Hub[7] finetuned using stochastic gradient descent (SGD) with a learning rate of 0.01, a Nesterov momentum of 0.9, a batch size of 32, and early stropping on top-30 error rate. Standard data augmentation was used: a random rotation of 45°, a random crop of size 224×224, and random flipping (both horizontal and vertical).
  - **CNN on RG+NIR patches:** same method (and same hyperparameters) than for RGB patches but with input images where the blue channel has been replaced by the near-infrared patch.

These baselines were designed to be simple to leave room for the ideas of the participants while providing some comparison to classical models in the SDM literature using machine learning models on tabular data [19, 20, 21] and to more recent approaches using CNNs on image patches [22].

4. Participants and main results

52 teams participated and submitted at least one prediction file through the Kaggle[8] page of the GeoLifeCLEF 2022 challenge for a total number of submissions in the course of the competition of 261. The final standing is shown in Figure 5.

Out of these 52 teams, 7 managed to beat the weakest non-constant baseline provided and 5 the strongest one. These 7 top participants are *Sensio team*, *Matsushita-san* from *EPFL* (Ecole Polytechnique Fédérale de Lausanne) [23], *New moon* from *CDUT* (Chengdu University of Technology), *Cesar LEBLANC* from *LIRMM / Inria* [24], *Mila_gang* from *UdeM / Mila* (Université de Montréal / Mila, Quebec AI institute) [25], *Sachith Seneviratne* from *UoM* (University of Melbourne), and, *Juntao Jiang* from *ZJU* (Zhejiang University) [26]. In the rest of the paper, we will be referring to the participants using their affiliations. Figure 6 shows the standings of these 7 top participants using their affiliations. The submissions of 4 of those participants are further developed in their individual working notes [23, 24, 25, 26]. As the winning team *Sensio Team* did not submit a working notes paper but did provide some information about their method[9], this input is reported in Appendix A for completeness. However, due to the conciseness of their feedback, they did not supply any details of the performance of their individual models, no ablation study, nor further analysis.

[7] https://pytorch.org/vision/stable/models.html
[8] https://www.kaggle.com/c/geolifeclef-2022-lifeclef-2022-fgvc9
[9] https://www.kaggle.com/competitions/geolifeclef-2022-lifeclef-2022-fgvc9/discussion/327055

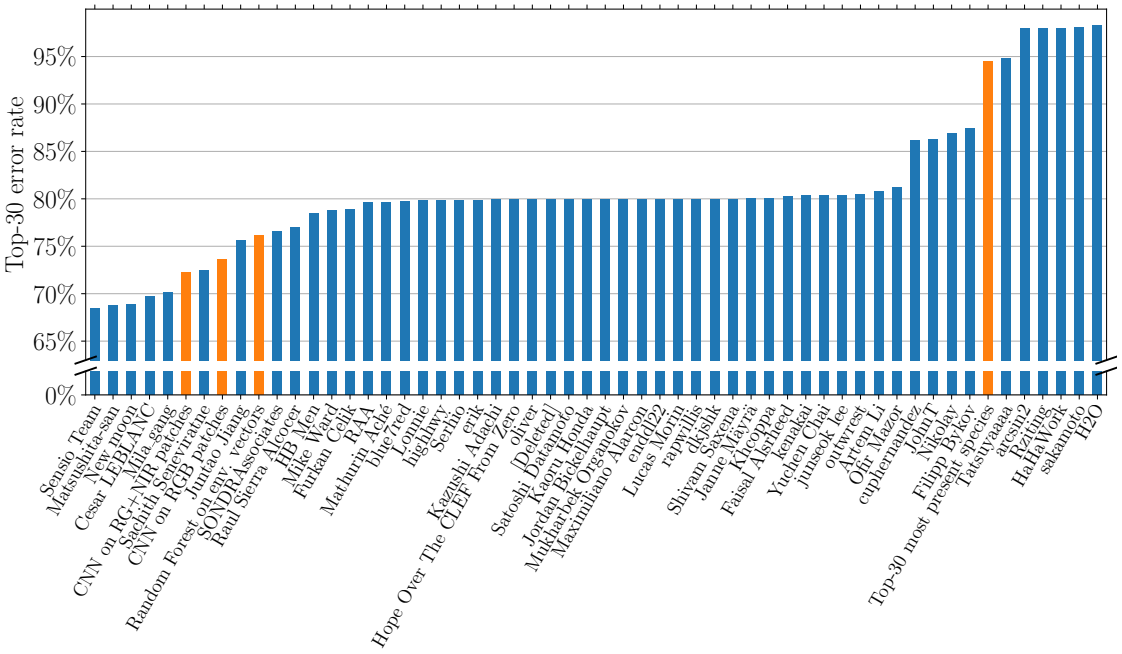


Figure 5: Results of the GeoLifeCLEF 2022 task as provided by the private leaderboard of Kaggle (with Kaggle team names). The top-30 error rates of the best submission of each participant are shown in blue. The provided baselines are shown in orange.

5. Methods

In this section, we highlight the main methods used by the participants.

5.1. Multi-modal models

The main challenge of this competition was to find a proper way to aggregate the heterogeneous sources of data and to deal with their respective characteristics: while RGB and NIR patches are standard images, other data was not directly provided in this format. For instance, altitude can not be cast in `uint8` without loss of information, land cover data is a categorical variable, bioclimatic and pedologic data have a resolution and range of their own, and, localization (GPS coordinates) is a punctual information. Interestingly, the participants did try different means of aggregating this heterogeneous data with more or less success and conflicting results. Figure 7 summarizes the main architectures tested in the course of GeoLifeCLEF 2022 to aggregate the different modalities. Note that they are not mutually exclusive and that some participants actually used several different aggregation methods at the same time.

Early input aggregation. The input patches are aggregated together resulting in a final patch with additional channels. This patch can then be fed to a single CNN whose first layer was adapted to accept more than three channels. In GeoLifeCLEF 2020, [27] used this approach to train a model from scratch using most modalities available. This year, *Sensio Team* and *UdeM / Mila* used this approach to aggregate together the RGB patches with the NIR ones. They

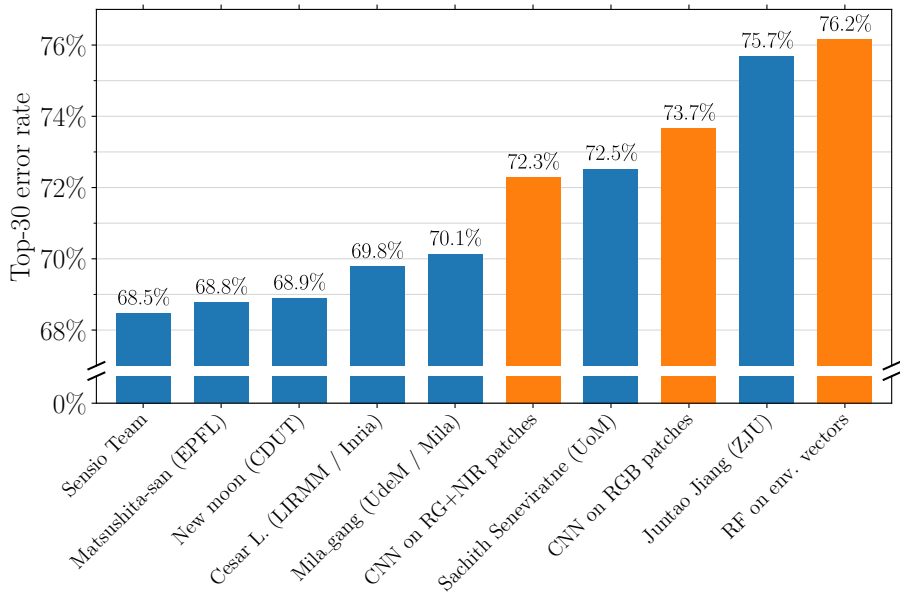


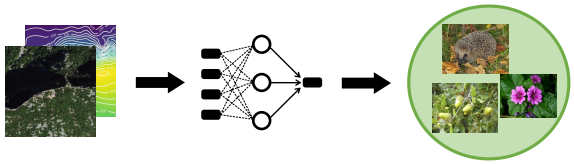
Figure 6: Results of the GeoLifeCLEF 2022 task of the 7 top participants. The top-30 error rates of the best submission of each participant are shown in blue. The provided baselines are shown in orange.

used a pre-trained model and randomly initialized the added filters for the NIR channels (no details were provided on that regard by *Sensio Team*). The advantage of this method is that the resulting model is rather simple and little compared to the other aggregation methods. On the other hand, all the modalities have to be given as patches of the same size[10]. Moreover, as the modalities are aggregated early, the model might struggle to build a relevant feature space if these modalities are too different from one another.

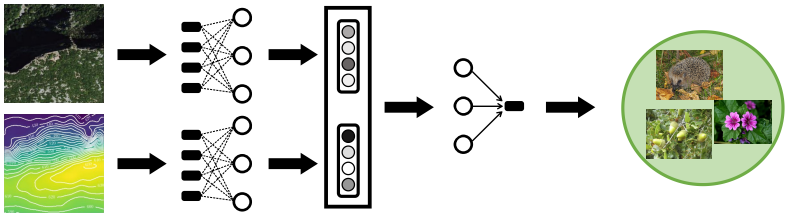
Independent feature extractors. Instead of directly aggregating the modalities at the input of the model, these modalities can go through separate networks to extract different representations adapted for each of them. These representations can then be concatenated to create a global multi-modal feature vector which can then be fed to a classifier—with a single or multiple linear layers. This approach has been successfully applied during GeoLifeCLEF 2021 [28] and was yet again used for GeoLifeCLEF 2022 by top teams *Sensio Team* (winning solution), *EPFL* and *UdeM / Mila*. These different teams have rather contradictory conclusions on the effectiveness of this approach. *EPFL* used two feature extractors based on CNNs, one for RGB and another one for a 3-channel patch containing NIR, altitude, and NDVI[11] data. They report some instability issues during training and a decrease in performance compared to using a single CNN on RGB patches. *Sensio Team* and *UdeM / Mila* used one feature extractor based on a CNN for either RGB+NIR (*Sensio Team* and *UdeM / Mila*) or NIR+GB (*Sensio Team*) patches and another one based on an neural network on tabular data, a multi-layer perceptron (MLP) for *Sensio*

[10]Note that, using this approach, it is technically possible to incorporate vector data by creating constant patches as did *LIRMM / Inria* to create patches from the coordinates of the observations.

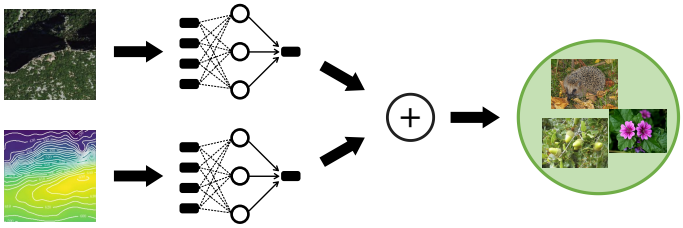
[11]The normalized difference vegetation index (NDVI) is a simple measure of vegetation computed from the red channel from RGB patches and the near-infrared data.



(a) Early input aggregation.



(b) Independent feature extractor for each modality.



(c) Late predictions averaging.

Figure 7: Possible architectures for multi-modal models, ranging from early to late aggregation of the different modalities. Note that it is possible to aggregate modalities given as patches but also as vectors.

Team and a tabular ResNet [29] for *UdeM / Mila*. According to *UdeM / Mila*, this approach also degraded the performance compared to their baseline model. It seems however that *Sensio Team* managed to obtain some gains. The strength of this independent feature extractors approach is that it is more likely to be able to properly extract the relevant information from very different modalities. Nevertheless, it is still not clear how to properly train such a model. Moreover, the resulting model is much bigger and has to be jointly trained. This can trigger some memory issues. [27] carried out post-challenge experiments that provide some insights on those two aspects.

Late predictions averaging. A straightforward and easy-to-implement aggregation approach consists in training separate models and averaging their predictions. *LIRMM / Inria* successfully used this approach as their main submission by learning 8 separate CNNs (thus solely on patches) for RGB, NIR, altitude, land cover, 3 selected temperature variables, 3 selected precipitation variables, 3 selected pedological variables, and the coordinates of the observations. *Sensio Team* also partly used this approach but rather as an ensembling method of good performing models than as an explicit way to aggregate different modalities. The advantage of this approach is its simplicity and the fact that models can be trained independently, it is thus

easy to add or remove one modality. However, the backlash is that there is no joint training of the model and, as there is no learned fusion layer, the modalities are only weakly mixed in the model which might harm its predictive performance.

Finally, besides these three main approaches, *CDUT* [30] used different ways to aggregate the modalities using an architecture based on a Swin transformer [31]. This approach seems promising, further experiments are necessary to measure the exact performance of such methods.

5.2. Species imbalance

Another important trait of the dataset is its imbalance shown in Figure 2: a few species account for most of the observations, while a lot of them have only been observed a handful of times. *EPFL* and *ZJU* tried to use specialized method for this type of data such as focal loss [32], balanced softmax [33] or more advanced methods. These did not help improve their scores, most likely because the test set shares the same imbalance as the training set and the evaluation metric did take it into account (the fixed list of metrics implemented by Kaggle did not allow us to use a class-averaged top-30 error rate).

5.3. Presence-only observation data

One last major characteristic of the dataset is that the observation data provided is presence-only data: at a given location, we only know that one species is present and do not have access to the complete list of species present or the ones absent. The winning team *Sensio Team* and *EPFL* tried to address this by using a grid of squared cells to aggregate the species observed into each cell. They then used this information in a different manner. The winning team tried to map the 30 species closest to each training point falling into its cell and used this list as the new label. Unfortunately, in the given time, this approach only resulted in overfitting. On the other hand, *EPFL* successfully used the aggregated observations as a regularization method by replacing the label assigned to each training observation by another species from its cell 10% of the time.

5.4. Other techniques used

Participants tried out different CNN architectures such as ResNet [17], DenseNet [34], Inception-V4 [35], and EfficientNet [36]. But also transformers such as ViT [37] tested without success by *EPFL* and Swin transformer [31] by *CDUT*. However, the results of these latter models were mixed.

Different approaches for model pre-training were also tested: no pre-training, pre-training on ImageNet, and pre-training on another dataset closer to GeoLifeCLEF 2022 (by *UdeM / Mila*). In the end, using models pre-trained on ImageNet gave consistently better results.

Multi-task learning has been used by two participants, *EPFL* and *UdeM / Mila*. *EPFL* modified the models to predict the different levels of taxonomy, i.e., species, genus, family, and kingdom. On the other hand, *UdeM / Mila* added two additional tasks beyond species prediction: land cover semantic segmentation and country prediction. Unfortunately, both attempts were unsuccessful.

While most participants used a single model for both countries, US and France, *ZJU* used two separate models. *EPFL* tried both approaches and did not notice any difference in predictive performance.

Finally, *EPFL*, instead of solely using the raw input data, computed the NDVI, a classical vegetative index, using the red channel of RGB patches and the NIR patches. However, as their modality aggregation approach was not fully successful, it is unclear whether it is better to compute explicitly this index or to let the model learn to compute it if it manages to do so and finds it relevant for the task.

6. Conclusion and perspectives

The 2022 edition of GeoLifeCLEF has shown a growing interest from the machine learning community towards the challenge. Several dozen people/research groups conducted experiments on the provided dataset and 7 of them managed to obtain better performances than the basic models provided by the organizers. Several participants expressed their satisfaction at having participated and emphasized the fact that this challenge had allowed them to address new issues with respect to their past experience in machine learning. The following two aspects, in particular, were highlighted:

1. the design and use of multi-modal networks requiring to mix structured with unstructured data, and finding effective solutions to capture features specific to each modality as well as interactions across modalities.
2. the design of new methods to tackle the presence-only problem which is rarely discussed in the machine learning community.

The challenge has thus allowed the experimentation of new approaches, some of which will be the subject of subsequent publications by the participants.

One way to improve the challenge would be to include presence/absence data as an additional test set (and possibly validation set). Having only presence-only occurrences in the test set indeed makes the evaluation of methods more difficult, especially to define an appropriate evaluation metric. The top-K error has indeed the defect to depend on the parameter K and not taking into account the variability of the number of species.

Acknowledgement

This project has received funding from the French National Research Agency under the Investments for the Future Program, referred to as ANR-16-CONV-0004, and from the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project). The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

[1] J. Elith, J. R. Leathwick, Species Distribution Models: Ecological Explanation and Prediction Across Space and Time, Annual Review of Ecology, Evolution, and Systematics (2009).

[2] A. Joly, H. Goëau, S. Kahl, L. Picek, T. Lorieul, E. Cole, B. Deneu, M. Servajean, A. Durso, H. Glotin, R. Planqué, W.-P. Vellinga, A. Navine, H. Klinck, T. Denton, I. Eggel, P. Bonnet,

M. Šulc, M. Hruz, Overview of LifeCLEF 2022: an evaluation of machine-learning based species identification and species distribution prediction, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022.
[3] C. Botella, P. Bonnet, F. Munoz, P. Monestiez, A. Joly, Overview of GeoLifeCLEF 2018: location-based species recommendation, CLEF: Conference and Labs of the Evaluation Forum (2018).
[4] C. Botella, M. Servajean, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2019: plant species prediction using environment and animal occurrences, CLEF: Conference and Labs of the Evaluation Forum (2019).
[5] B. Deneu, T. Lorieul, E. Cole, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF), in: CLEF task overview 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.
[6] T. Lorieul, E. Cole, B. Deneu, M. Servajean, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2021: Predicting species distribution from 2 million remote sensing images, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, 2021.
[7] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, A. Joly, The GeoLifeCLEF 2020 dataset, arXiv preprint arXiv:2004.04192 (2020).
[8] C. Homer, J. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. Herold, J. Wickham, K. Megown, Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover change information, Photogrammetric Engineering & Remote Sensing 81 (2015) 345–354.
[9] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, A. Jarvis, Very high resolution interpolated climate surfaces for global land areas, International Journal of Climatology: A Journal of the Royal Meteorological Society 25 (2005) 1965–1978.
[10] T. Hengl, J. M. de Jesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, et al., SoilGrids250m: Global gridded soil information based on machine learning, PLoS one 12 (2017).
[11] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al., Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, Ecography 40 (2017) 913–929.
[12] J. L. Pearce, M. S. Boyce, Modelling distribution and abundance with presence-only data, Journal of applied ecology 43 (2006) 405–412.
[13] E. Chzhen, C. Denis, M. Hebiri, T. Lorieul, Set-valued classification–overview via a unified framework, arXiv preprint arXiv:2102.12318 (2021).
[14] S. J. Phillips, M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, S. Ferrier, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecological applications 19 (2009) 181–197.
[15] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Pro-

ceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, 2019, pp. 8024–8035.
[19] J. Franklin, Mapping species distributions: spatial inference and prediction, Cambridge University Press, 2010.
[20] J. S. Evans, M. A. Murphy, Z. A. Holden, S. A. Cushman, Modeling species distribution and change using random forest, in: Predictive species and habitat modeling in landscape ecology, Springer, 2011, pp. 139–159.
[21] A. Guisan, W. Thuiller, N. E. Zimmermann, Habitat suitability and distribution models: with applications in R, Cambridge University Press, 2017.
[22] C. Botella, A. Joly, P. Bonnet, P. Monestiez, F. Munoz, A deep learning approach to species distribution modelling, in: Multimedia Tools and Applications for Environmental & Biodiversity Informatics, Springer, 2018, pp. 169–199.
[23] B. Kellenberger, T. Devis, Block label swap for species distribution modelling, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
[24] C. Leblanc, T. Lorieul, M. Servajean, P. Bonnet, A. Joly, Species distribution modeling based on aerial images and environmental features with convolutional neural networks, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
[25] M. Teng, S. Elkafrawy, Convolution neural network fine-tuning for plant and animal distribution modelling, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
[26] J. Jiang, Localization of plant and animal species prediction with convolutional neural networks, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.
[27] B. Deneu, M. Servajean, A. Joly, Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge, in: CLEF working notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece., 2020.
[28] S. Seneviratne, Contrastive representation learning for natural world imagery: Habitat prediction for 30,000 species, in: CLEF working notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania., 2021.
[29] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, Advances in Neural Information Processing Systems 34 (2021) 18932–18943.
[30] Y. Zhou, P. Peng, G. Wang, et al., A multimodal species distribution model incorporating remote sensing images and environmental features (2022).
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[33] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, et al., Balanced meta-softmax for long-tailed visual

recognition, Advances in Neural Information Processing Systems 33 (2020) 4175–4186.
[34] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[35] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.
[36] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[38] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., Searching for MobileNetV3, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314–1324.
[39] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics (2001) 1189–1232.

A. Winners solution summary

As the winners did not submit a working notes paper but described their method in the "Discussion" tab on the Kaggle page of the competition[12], we summarize it here for archiving and referencing purposes.

A.1. Solution description

The final solution, illustrated in Figure 8, consisted of an ensemble—averaging of the predictions—of three models:

1. A bi-modal network with NIR+GB on a pre-trained ResNet-34 [17] stacking its final layer to an MLP with 3 layers taking as input the environmental vectors, latitude, longitude, country, altitude mean, max-min altitude, and "dothot" encoding (this is how they called the softmax-onehot encoding) of land covers. These two models were connected to the final classification layer.
2. Another bi-modal network similar to the previous with the same MLP but where the ResNet-34 was replaced by a pre-trained MobileNetV3-large [38] taking RGB+NIR as input. After the concatenation of the outputs of these two models, an extra linear layer of size 2,048 with dropout and ReLU was added before the final classification layer of size 17K.
3. A random forest with 32 trees and a depth of 12 using the same inputs as the previous MLPs with, in addition, the 25th, 50th, and 75th percentiles of each of the R, G, B, and

[12] https://www.kaggle.com/competitions/geolifeclef-2022-lifeclef-2022-fgvc9/discussion/327055

NIR layers. In total, 81 input features were used. The whole training set (training and validation subsets) was moreover used for fitting the final model.

The first two models used data augmentation to train the CNN models: random vertical and horizontal flips, rotations, and 5-10% of brightness and contrast. Test-time augmentation (TTA) was also used by averaging the mean predictions of 5 random image transformations for each sample. Finally, adding validation data to train data was tried. It improved slightly the performance but probably it could have been done better.

A.2. Inconclusive tests

Aggregating close labels to obtain multi-label observations was tried in different ways and using different loss functions but none improved the performance compared to single labels. However, the participants believe that there has to be a way to make it work.

Other architectures for the CNN models and training from scratch were tried. For instance, the participants also tried using 3 different CNNs for (i) RGB+NIR, (ii) altitude, and (iii) land cover patches before aggregating their outputs to the MLP on tabular data. None of these gave better results but the participants think there is room for some of those ideas to improve the performance of their solution. Also, they tried using gradient boosting trees [39] but with 17K classes, it did not fit the amount of RAM available to them.

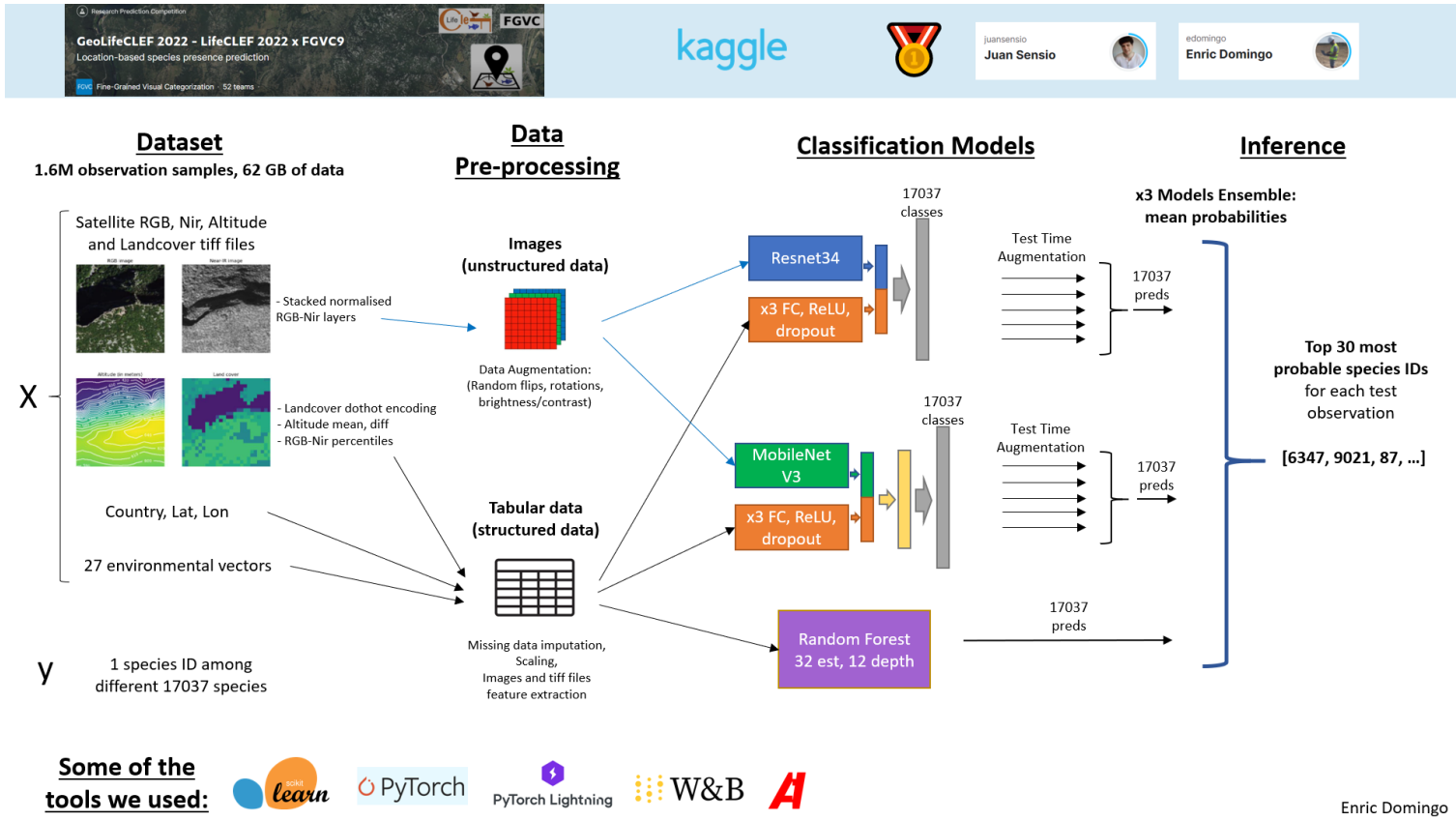


Figure 8: Winning solution summary figure provided by Sensio team. Credits: Enric Domingo.