A deep-learning framework for enhancing habitat identification based on species composition

César Leblanc, Pierre Bonnet, Maximilien Servajean, Milan Chytrý, Svetlana Aćić, Olivier Argagnon, Ariel Bergamini, Idoia Biurrun, Gianmaria Bonari, Juan Antonio Campos, et al.

A deep learning framework for enhancing habitat identification based on species composition

César LEBLANC[1,2], Pierre BONNET[2], Maximilien SERVAJEAN[3], Milan CHYTRÝ[4], Svetlana AĆIĆ[5], Olivier ARGAGNON[6], Ariel BERGAMINI[7], Idoia BIURRUN[8], Gianmaria BONARI[9], Juan A. CAMPOS[8], Andraž ČARNI[10,11], Renata ĆUŠTEREVSKA[12], Michele DE SANCTIS[13], Jürgen DENGLER[14,15], Tetiana DZIUBA[16], Emmanuel GARBOLINO[17], Valentin GOLUB[18], Ute JANDT[19,20], Florian JANSEN[21], Maria LEBEDEVA[22], Jonathan LENOIR[23], Jesper Erenskjold MOESLUND[24], Aaron PÉREZ-HAASE[25], Remigiusz PIELECH[26], Jozef ŠIBÍK[27], Zvjezdana STANČIĆ[28], Angela STANISCI[29], Grzegorz SWACHA[30], Domas UOGINTAS[31], Kiril VASSILEV[32], Thomas WOHLGEMUTH[33], and Alexis JOLY[1]

[1] Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, FR
[2] AMAP, Université de Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France
[3] LIRMM, AMIS, Université Paul-Valéry - Montpellier 3, CNRS, Montpellier, FR
[4] Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, CZ
[5] Department of Botany, Faculty of Agriculture, University of Belgrade, Belgrade, RS
[6] Antenne Languedoc-Roussillon, Conservatoire botanique national méditerranéen, Hyères, FR
[7] Swiss Federal Research Institute WSL, Birmensdorf, CH
[8] Department of Plant Biology and Ecology, University of the Basque Country UPV/EHU, Bilbao, ES
[9] Department of Life Sciences, University of Siena, Siena, IT
[10] Institute of Biology, Research Center of the Slovenian Academy of Sciences and Art, Ljubljana, SI
[11] School for Viticulture and Enology, University of Nova Gorica, Nova Gorica, SI
[12] Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Skopje, MK
[13] Department of Environmental Biology, Sapienza University of Rome, IT
[14] Vegetation Ecology Research Group, Institute of Natural Resource Sciences (IUNR), Zurich University of Applied Sciences (ZHAW), Wädenswil, CH
[15] Plant Ecology, Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Bayreuth, DE
[16] Department of Geobotany and Ecology, M.G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, Kyiv, UA
[17] MINES Paris PSL, ISIGE, Fontainebleau, FR
[18] Ecosystem Research Laboratory, Institute of Ecology of the Volga river basin of the Russian Academy of Science, Samara Federal research center of the Russian Academy of Sciences, Togliatti, RU
[19] Geobotany & Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, DE
[20] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,

Leipzig, DE
[21] Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, DE
[22] South-Ural Botanical garden-institute of Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, RU
[23] UMR CNRS 7058 "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN), Université de Picardie Jules Verne, Amiens, FR
[24] Department of Ecoscience, Aarhus University, Aarhus, DK
[25] Department of Evolutionary Biology, Ecology and Environmental Sciences and Biodiversity Research Institute (IRBio), University of Barcelona, Barcelona, ES
[26] Institute of Botany, Faculty of Biology, Jagiellonian University, Kraków, PL
[27] Department of Biodiversity and Ecology, Plant Science and Biodiversity Center Slovak Academy of Sciences, Bratislava, SK
[28] Faculty of Geotechnical Engineering University of Zagreb, Varaždin, HR
[29] Department of Bioscience and Territory, University of Molise, Termoli, IT
[30] Botanical Garden, University of Wrocław, PL
[31] Nature Research Centre, Vilnius, LT
[32] Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Science, Sofia, BG
[33] Research Unit Forest Dynamics, Swiss Federal Research Institute for Forest, Snow and Landscape Research WSL, Birmensdorf, CH

February 6, 2024

Correspondence

César Leblanc, Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, FR & CIRAD, UMR AMAP, Montpellier, FR
Email: cesar.leblanc@inria.fr

Funding information

The research described in this paper was funded by the European Commission through the GUARDEN (safeGUARDing biodivErsity aNd critical ecosystem services across sectors and scales) project and the MAMBO (Modern Approaches to the Monitoring of BiOdiversity) project. These projects received funding from the European Union's Horizon Europe research and innovation program under grant agreements 101060693 (start date: 01/11/2022; end date: 31/10/2025) and 101060639 (start date: 01/09/2022; end date: 31/08/2026), respectively. This work had granted access to the High-Performance Computing (HPC) resources of IDRIS (Institut du Développement et des Ressources en Informatique Scientifique) under the allocation 2023-AD011014219 made by GENCI (Grand Equipement National de Calcul Intensif). The authors declare no financial or personal conflicts of interest related to this work. Any ethical considerations were taken into account during the research process. The views and opinions expressed in this work are those of the authors only and do not necessarily reflect those of the GUARDEN or MAMBO partners, the European Commission, or GENCI.

Abstract

Aims The accurate classification of habitats is essential for effective biodiversity conservation. The goal of this study was to harness the potential of deep learning to advance habitat classification in the European Union (EU). We aimed to develop and evaluate models capable of assigning vegetation-plot records to the habitats of the European Nature Information System (EUNIS), a widely used reference framework for European habitat types.

Location The framework was designed for use in Europe and adjacent areas.

Methods We leveraged deep learning techniques, such as transformers (i.e., models with attention components able to learn contextual relations between categorical and numerical features), that we trained using k-fold cross-validation (CV) on vegetation plots sourced from the European Vegetation Archive (EVA), to show that they have great potential for classifying vegetation-plot records. We experimented different network architectures, feature encodings, hyperparameter tuning and noise addition strategies to identify the optimal model. We used an independent test set from the National Plant Monitoring Scheme (NPMS) to evaluate its performance and compare its results against the traditional expert systems.

Results We explored the use of deep learning applied to species composition and plot-location criteria and we developed a framework for habitat classification containing a wide range of models. Our selected algorithm, applied to European habitat types, significantly improved habitat classification accuracy, achieving an improvement of over twofold compared to the previous state-of-the-art (SOTA) method on an external dataset. The framework is shared and maintained through a GitHub repository.

Conclusions Our results demonstrate the potential benefits of the adoption of deep learning for improving the accuracy of vegetation classification. They highlight the importance of incorporating advanced technologies into habitat monitoring. Indeed, these algorithms have shown to be best suited for habitat type prediction than expert systems. The framework we developed can be used by researchers and practitioners to accurately classify habitats.

Keywords — Artificial intelligence, Biodiversity monitoring, Deep learning, European flora, Expert system, Habitat type, Phytosociology, Vascular plant species, Vegetation classification

1 Introduction

The term habitat (Hall et al., 1997) encompasses a broad range of definitions (Yapp, 1922). In this study, we adopt the following: "plant and animal communities as the characterising elements of the biotic environment, together with abiotic factors (soil, climate, water availability and quality, and others), operating together at a particular scale" (Davies and Moss, 1999). The EUNIS Habitat Classification (Moss, 2008) uses this definition and serves as a comprehensive and hierarchical pan-European system for habitat identification that covers all types of habitats, which are identified by specific codes, names and descriptions. The EUNIS classification system stands nowadays as a widely recognized framework for European habitat types (as it has already played a pivotal role in numerous applications, both research and applied applications (Evans, 2012), and provides a common language for communication among scientists, policy-makers, and other stakeholders). The European Environment Agency (EEA) initiated a (still on-going) process of the revision of the EUNIS habitat classification at level three, i.e., habitat complexes, (and sometimes level four, i.e., biotope complexes) of its classification hierarchy. This revision led to a more consistent and less ambiguous typology. On this work, we focused on eight habitat groups (level one habitats):

1. Littoral biogenic habitats (MA2)
2. Inland habitats with no or little soil and mostly with sparse vegetation (U)
3. Coastal habitats (N)
4. Wetlands (Q)
5. Grasslands and lands dominated by forbs, mosses or lichens (R)
6. Heathlands, scrub and tundra (S)
7. Forests and other wooded land (T)
8. Vegetated man-made habitats (V)

Habitat type classification is a fundamental process integral to ecology, involving automatically classifying an area based on its environmental characteristics and species composition. It is done by

combining observations of species co-occurrence or abundance with environmental estimates to map habitat distributions across landscapes. Several tools for vegetation classification with different logic and strategy are available. In particular machine learning algorithms (Hastie et al., 2009) and expert systems (Noble, 1987). The former are tools for induction of the independent knowledge base, whereas the latter emulate the process of expert classification done by humans by using explicitly defined logical formulas. These (numerical) tools can also play a vital role for nature conservation, landscape mapping and land-use planning and can facilitate biodiversity management (Estopinan et al., 2024). They make monitoring of species and habitats easier and more accurate, provide decision-support for nature conservation and guidance for nature restoration and development. Thus, it can be particularly valuable in the current context where a significant portion of habitats are at risk of collapsing (at least 32% of terrestrial habitats and 18% of marine habitats are threatened (Janssen et al., 2016)). Therefore, habitat type classification has a crucial role in ecology, and using the EUNIS habitat classification can serve as a key instrument for assessing progress towards the European Union's biodiversity targets.

On the one hand, many expert systems that have been published by the global community (Tichý et al., 2019) to protect nature and have long played a crucial role in restoring habitats and species worldwide. Whether they classify the vegetation of precisely-defined phytosociological units (Marcenò et al., 2018) (Novák et al., 2023), the vegetation of entire countries (Chytrý et al., 2012) (Wiser et al., 2018) or even the vegetation of larger areas (Chytrý et al., 2020) (Mucina et al., 2016), these expert systems follow all human decisions. They are usually designed by experts in the field who have extensive knowledge of the characteristics of different habitats and their species composition. These systems thus employ assignment rules (species-based and/or location-based membership conditions) to classify vegetation plots into vegetation or habitat types with formal definitions. However, it's important to note that these definitions can evolve over time, meaning that the structure of the expert systems might need to be modified in order to replace current provisional definitions with improved ones or to use new vegetation-plot records to characterize habitat types. Moreover, the current version of the expert system for automatic classification of European vegetation plots to habitat types of the EUNIS habitat classification (i.e., EUNIS-ESy (Chytry et al., 2021)) contains some definitions that are:

- **strict**, e.g., to be correctly assigned to its habitat, a vegetation plot should contain at least n species of a given functional species group, or the total cover of a discriminating species group in a vegetation plot should be greater than the total cover of other discriminating species groups in the plot,
- **complex**, e.g., to be correctly assigned to its habitat, the total cover of a functional species group in a vegetation plot should be greater than that of another functional group, excluding the species of the former group from the latter group, or the sum of square-rooted percentage covers of the species belonging to a discriminating species group in a vegetation plot should be greater than the sum of square-rooted percentage covers of the species of another discriminating species group,
- and **idiosyncratic**, e.g., to be correctly assigned to its habitat, a vegetation plot should belong to a dataset, or a vegetation plot shouldn't be located in a country.

These intricacies motivate the exploration of alternative approaches, such as the application of deep learning algorithms, which we delve into in this study.

On the other hand, even if they have shown great potential for modeling species distributions (SDM) (Botella et al., 2018), modern deep learning techniques have never been applied to classify EUNIS habitats, and their application (Černá and Chytrý, 2005) to the classification of habitats at a global scale is a relatively unexplored territory (Joly et al., 2023). Deep learning techniques are types of machine learning models that can automatically learn patterns and features from large amounts of data (Botella et al., 2023a) and that are typically designed and trained by data scientists, who have expertise in artificial intelligence (AI) and data analysis. As it had already been done for species (Deneu et al., 2021), we sought to establish that it was feasible to map EU habitats extent at (very) high spatial resolution (Deneu et al., 2022). Thus, we used *in-situ* plant species composition data, information on the location and some environmental features (Leblanc et al., 2022) in a framework

with a diverse range of deep learning models that could be trained for different types of habitats in order to reach an optimal compromise between accuracy and generalization. Habitat type identification has traditionally relied on expert knowledge, a process that is not only time-consuming and costly but also susceptible to subjectivity. Advances in machine learning have opened up new opportunities for automating this process using large datasets of environmental and and other auxiliary data. We built upon these techniques to enable automation and scalability in habitat classification, which forms the cornerstone of our study. AI-powered Habitat Distribution Models (HDMs) should thus be suited to represent how complex ecological niches and spatial dynamics determine the distribution of many habitats in a region. Machine learning could improve predictive performance in HDMs compared to expert systems by better mapping the actual realized distribution of habitat types.

We trained different models on very large volumes of data (by coupling EUNIS types with plant species composition recorded in vegetation plots) to develop, share and maintain a generic, free and open-source deep learning framework capable of accurately classifying vegetation plots to their habitat types. Several crucial features were introduced into the software package to make it generic and reusable in a wide variety of contexts. We focused our work on five key areas for (i) high modularity (for enhanced flexibility), (ii) new data loaders (to handle both internal and external classification criteria (De Cáceres et al., 2015), i.e., respectively species-based and location-based criteria), (iii) new model's architectures (in particular models based on transformers (Vaswani et al., 2017)), (iv) new loss functions (i.e., the penalty for an incorrect classification of a vegetation plot, in particular for species assemblage prediction with an imbalanced top-k loss (Garcin et al., 2022)) and, (v) a new inference module allowing to compute the top-k classification for any user-specified area and plant species composition.

2 Methods

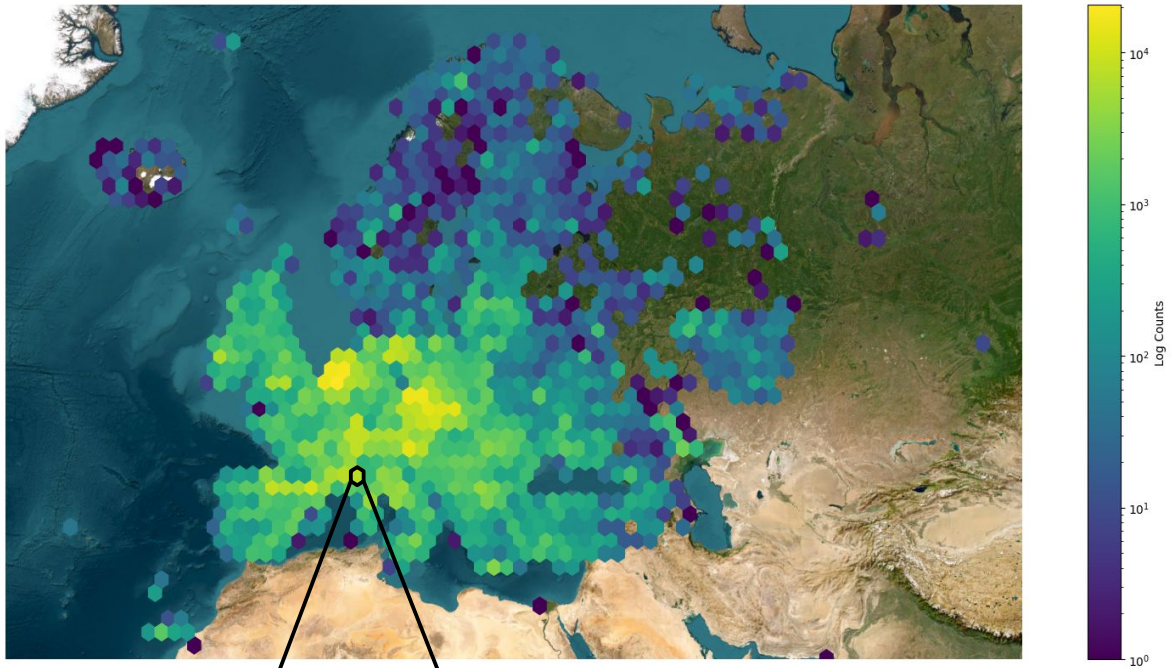
2.1 Data

2.1.1 EVA: a comprehensive dataset for habitat classification

Our data source for training the deep learning framework was drawn from a subset of a data repository of vegetation-plot observations (i.e., records of plant taxon co-occurrence and cover-abundance at particular sites in plots ranging from 1 m^2 to a few hundred m^2 which have been collected by vegetation scientists (Zhongming et al., 2015)) from Europe and adjacent areas. This EVA database (Chytrý et al., 2016), which was accessed on 22 May 2023, is an initiative of the Working Group European Vegetation Survey (EVS). Each of the vegetation plots typically contained estimates of cover-abundance of each species (vascular plant in every vegetation plot, bryophytes and/or lichens in some vegetation plots) alongside various supplementary details and additional sources of information on vegetation structure, location and environmental features. Although the EVA database represents a valuable resource for studying vegetation patterns and dynamics, we were mindful and acknowledged potential limitations stemming from the representativeness of the data and the possibility of sampling bias (inherent to sets of data assembled from multiple sources and originally collected for various purposes) (Michalcová et al., 2011). The final dataset contained a total of 886 260 georeferenced plots (with an average of approximately around 20 species per vegetation plot), 228 different habitats and 10 481 different species. Refer to Appendix S4 for a detailed overview of all the preprocessing steps and to Figure 1 for different visualizations.

2.1.2 NPMS: an independent dataset to evaluate models

To comprehensively assess and compare the transferability of our models and the EUNIS-ESy, we also established an independent and separate test dataset (whose labels weren't generated by the EUNIS expert system nor by our algorithms but relied on human annotations). As most of the existing vegetation-plot databases indexed in the Global Index of Vegetation-Plot Databases (Dengler et al., 2011) (GIVD) and the Global Vegetation Database (Bruelheide et al., 2019) (sPlot) were already included inside EVA, obtaining a representative and high-quality independent dataset for model validation was challenging. To address this, we selected the NPMS (Walker et al., 2015). It aims to survey plant species across different habitats in the United Kingdom (UK) by utilizing data collected



Matched concept	Cover %
Artemisia campestris	2.0
Asparagus officinalis	2.0
Bellardia trixago	2.0
Brachypodium phoenicoides	3.0
Dactylis glomerata	13.0
Erigeron sumatrensis	2.0
Lagurus ovatus	3.0
Medicago minima	3.0
Phillyrea angustifolia	88.0
Phragmites australis	2.0
Rostraria cristata	2.0
Silene italica	2.0

Figure 1: Hexagonal binning showing the distribution of vegetation plots from the training dataset. Zoom in on a specific bin with the raw spatial distribution of the vegetation plots. Further breakdown on a vegetation plot (assigned to the habitat type S51, i.e., Mediterranean maquis and arborescent matorral) with the list of co-occurring species.

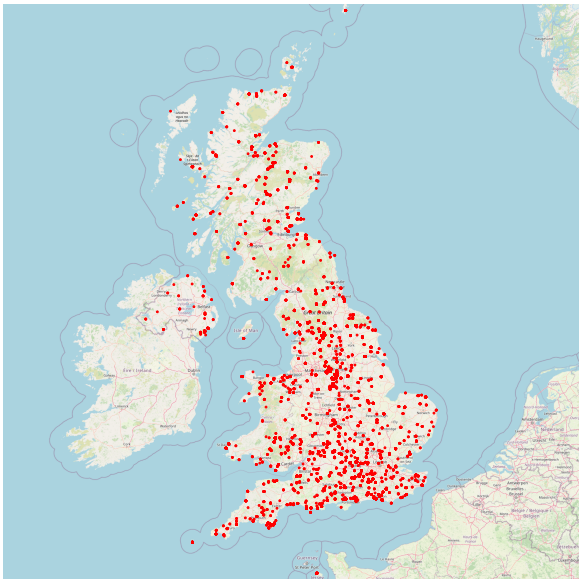


Figure 2: Distribution of vegetation plots in the NPMS test set

by citizens (i.e., expert volunteers who carried out surveys of wildflowers and their associated habitats). This scheme was designed and developed collaboratively by the Botanical Society of Britain & Ireland (BSBI), UK Centre for Ecology & Hydrology (UKCEH), Plantlife and the Joint Nature Conservation Committee (JNCC). We specifically chose this dataset because it offered an intriguing opportunity to validate the work of numerous European vegetation scientists across generations with a recent citizen science project (Bonnet et al., 2023) that employed a systematic protocol and methodology (e.g., the participants were allocated a 1km square in which they had to visit five plots in semi-natural habitats twice a year) and encompassed a wide range of vegetation types, providing valuable insights into the potential transferability of our models in a real-world context, beyond expert-driven datasets. It offered an interesting contrast by incorporating data collected through citizen science (Bonnet et al., 2020), thus expanding our understanding of the generalization of the framework beyond traditional scientific datasets. However, this dataset is by nature very different from EVA, and that there is a significant distribution shift between the two due to the different collection protocols. So we cannot expect the same level of performance. We detail the preprocessing steps to create the test dataset in Appendix S4. Refer to Figure 2 for a visual representation of the distribution of the testing dataset.

2.2 Modeling

2.2.1 Validation: accounting for the spatial structure of ecological data

The goal of this paper is to use the floristic and environmental information in several locations to train a deep learning tabular model that can predict the habitat type of given points. To mitigate the influence of spatial autocorrelation and to ensure that our models generalize well beyond the spatial structure of the training data, we split our dataset into ten folds according to a spatial block holdout procedure (Roberts et al., 2017). All the vegetation plots were assigned into a grid of 10 km $\times$ 10 km cells, all of these cells were then randomly sampled for one of the folds and each fold was used once as an internal validation set while the nine remaining folds formed the training set, allowing us to perform ten-fold CV (Stone, 1974). The performance measure reported by the ten-fold CV was then the average of the values computed in the loop. This method allowed us to evaluate our approaches in a way that limits the effect of the spatial bias in the data without wasting much of it (which can occur when arbitrarily setting aside a validation set). Importantly, it is worth noting that, regardless of the fold designated for validation in each iteration, every habitat category remained present in the training set formed by the remaining nine folds.

2.2.2 Models: using deep neural networks on tabular data for classification

We used the ten-fold CV procedure described above to conduct a rigorous comparative analysis of several machine and deep learning models. Since there was not an established benchmark for tabular data, we had to work with some of the most used and well-established machine and deep learning algorithms in competitions, from ensembles of decision trees (Friedman, 2001) to attention-based models (Bahdanau et al., 2014). To ensure fairness and optimize their performances, we meticulously tuned each model's main hyperparameters (for the rest, we kept the default configurations recommended by the corresponding papers) (Feurer and Hutter, 2019). The existing literature described a wide range of diverse machine and deep learning models for tabular data (Borisov et al., 2022), but none of them could consistently outperform all the others. To comprehensively assess model performance, we adopted a variety of approaches and selected neuron-based, tree-based and transformer-based models. We illustrate each model and the associated training procedure in Appendix S1. Five common algorithms were retained for evaluation:

1. A MultiLayer Perceptron classifier (MLP) (Haykin, 1998), i.e., a fully connected class of feed-forward artificial neural network. It works by taking input data, passing it through multiple layers of interconnected nodes with weighted connections and activation functions (Bircanoğlu and Arıca, 2018), and producing output predictions based on the learned patterns in the data.

2. A Random Forest Classifier (RFC) (Ho, 1995), i.e., a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A single decision tree works by recursively partitioning the input data based on the values of its features to create a tree-like structure, where each internal node represents a feature and each leaf node represents a decision or prediction based on the input data's characteristics.

3. An eXtreme Gradient Boosting classifier (XGB) (Chen and Guestrin, 2016), i.e., an optimized distributed gradient boosting algorithm designed to be highly efficient, flexible and portable. It works by iteratively training and adding decision trees to the ensemble model, each focusing on reducing the residual errors of the previous trees, using a combination of gradient descent optimization (Ruder, 2016), regularization techniques, and hardware-aware optimization to achieve high accuracy and scalability.

4. A TabNet Classifier (TNC) (Arik and Pfister, 2019), i.e., a novel high-performance and interpretable canonical deep tabular data learning architecture. It works by selectively attending to the most informative features of the input data and using a sparse masking technique to allow for efficient and interpretable feature selection, while employing a multi-step decision-making process and auxiliary loss functions to enhance its performance and generalization.

5. A Feature Tokenizer + Transformer classifier (FTT) (Gorishniy et al., 2021), i.e., a model that transforms all features (categorical and numerical) to embeddings and applies a stack of transformer layers to the embeddings. It works by transforming all features to tokens and running a stack of transformer layers over the tokens, so every transformer layer operates on the feature level of one object.

2.2.3 Encodings: mapping current habitat distributions under different constraints

The vegetation plots found within EVA contain comprehensive records of plant species co-occurrence and abundance. All categorical variables (i.e., the country name, the terrestrial ecoregion, the coastline and the location on a coastal dune) are transformed using the simple and widely-used one-hot encoding technique (Hancock and Khoshgoftaar, 2020). It is an encoding method where a particular value of a categorical variable having n possible categories would be encoded with a 1-dimensional feature vector of length n where every component is zero except for the ith component, corresponding to the index of the particular category in the set of possible values, which has the value one. All numerical features (i.e., the degrees of latitude and longitude and the altitude in meters above sea level of the vegetation plot) were left untouched. We proposed different data representations (as it is known that it can be vital for the success or failure of models (Bengio et al., 2013)) to ensure the framework's applicability to both abundance and presence–absence surveys (Joseph et al., 2006). Three distinct techniques for plant species encoding were employed:

1. The cover-abundance of each species, i.e., the natural logarithm of the raw data from EVA, which was recorded using a cover-abundance scale (Westhoff and Van Der Maarel, 1978) and then transformed to the arithmetic mid-point percent cover value corresponding to the individual cover-abundance class following a comprehensive database management system following the default values in the Turboveg database management program (Hennekens and Schaminée, 2001).

2. The presence-absence of each species, i.e., the binarization of the raw data from EVA. Each non-zero entry from the original data is converted to the value one, and every explicit zero are preserved (Scherrer et al., 2020).

3. The reciprocal rank of each species, i.e., the inverse of the ordinal ranking of the raw data from EVA. Each species is ranked in descending order of its original value (Brun et al., 2023) and is then associated with the value of the inverse of its position in the ranking.

2.3 Evaluation

2.3.1 Fitting: evaluating modelling algorithms on selected covariates

All details about the models and their optimization are provided in Appendix S1. We evaluated the performance of the expert system on the training set we created. We were fully aware that EVA was classified using EUNIS-ESy (using its definitions of individual EUNIS habitats based on their species composition and geographic location) but we wanted to see if the vegetation plots would remain classified to the same habitat after interpreting the taxon names with the GBIF. We thus kept the same 886 260 vegetation plots, we took the names from the original database and proceeded to normalize them. Furthermore, unlike our experiments for which we kept only vascular plant species and species that were observed at least ten times, we also kept in this case species belonging to other phyla (especially bryophytes and lichens since they were used by the expert system in the definition of some habitats such as *S12*, i.e., moss and lichen tundra) and rare species (as rare species with occurrences concentrated in a particular habitat could be used as positive indicators of the habitat by the expert system). This process increased the number of observations to 18 867 936 (instead of the 17 718 306 used to evaluate our models) and the number of different species to 17 885 (instead of the 10 481 used to evaluate our models). Out of the 886 260 vegetation plots, two of them had no species left after the species name matching, and as the expert system (unlike our framework) can't classify vegetation plots solely based on external criteria, we added for both vegetation plots a fake species named "Unknown species" having a percentage cover of 10%.

2.3.2 Metrics: computing accuracy to evaluate how well the models are performing

Some of the vegetation plots that were automatically classified by EUNIS-ESy were assigned to several level three EUNIS habitats. In order to deal with that and to evaluate the effectiveness of our classification framework considering the complexity of the habitat classification task, two key metrics were selected:

1. The top-one micro average multiclass accuracy, i.e., $\frac{1}{N}\sum_{i=1}^{N} 1(y_i = \hat{y}_i)$ where y is the target values and $\hat{y}$ is the predictions. It is the conventional accuracy: the model's prediction must be exactly the expected habitat type. This was the most important metric and played a pivotal role in our evaluation, as it provided crucial insights into the performance of our approaches when we were predicting which habitat was the most likely to be observed at a given location.

2. The top-three accuracy, i.e., $\frac{1}{N}\sum_{i=1}^{N} e_i$ where e_i equals 1 if $\forall k \in \{1,2,3\}, \hat{y}_{i,k} = y_i$ and equals 0 otherwise and where y_i is a single ground-truth label and $\hat{y}_{i,k}$ are candidate labels, both associated to a sample i. It means that any of the model's five highest probability predictions must match the expected answer. This metric was useful to assess the performances of our methods on similar habitats (i.e., habitats that have almost identical species composition and environmental features and are thus hard to distinguish from one another) and on scenarios where a vegetation plot was associated with several different habitat labels.

2.3.3 Noise: assessing the robustness and generalization of models

To enhance the robustness (Sietsma and Dow, 1991) of our approaches (to mitigate the risk of the phenomenon of overfitting (Dietterich, 1995)), we experimented with the incorporation of controlled noise to the input data. We introduced 30% of dropout, i.e., when evaluating the performance of the models we gave each present species a 30% chance of being randomly considered absent in the input data. This deliberate introduction of noise served the vital purpose of reducing the risk that our models will overfit the noise in the data by memorizing various peculiarities of some vegetation plots. Instead, it encouraged the models to identify more general and transferable patterns, thus bolstering their ability to make accurate predictions across diverse ecological contexts. It also helped to imitate the omission of plant species during vegetation sampling (e.g., if some species were small and not easily visible) (Morrison, 2021). After encoding the data and adding (or not) noise to it, standardization of the features (i.e., by removing the mean and scaling to unit variance in order to have a mean of observed values of zero and a standard deviation of one) was always initiated (these values were estimated from the training data, and then the transformation was consistently applied across all datasets), as it has been shown that such manipulation can benefit to some models by improving the numerical stability of the calculations (Kuhn et al., 2013).

3 Results

3.1 Selection: finding the best performing model

Table 1 contains a comprehensive overview of all the results we obtained (with the models already tuned), showcasing the performance of each model-encoding combination. Among the various configurations tested, the model-encoding combination with the best results is a MLP coupled with features encoded using the reciprocal rank method. It outperformed other models both with and without noise addition to the data and when measuring the performance with the top-one micro average multiclass accuracy (since it is the best suited metric in our case, as we want to prioritize the most likely habitat for each vegetation plot).

Table 1: Comparison of the top-one (in black) and top-three (in grey) micro average multiclass accuracy averaged over the ten CV folds for every model and encoding, with and without noise addition (best top-one result overall with and without noise addition in green background shading)

Models	Ten-fold CV			Ten-fold CV with 30% dropout		
	Cover abundance	Presence-absence	Reciprocal rank	Cover abundance	Presence-absence	Reciprocal rank
MLP	88.33/97.99	76.69/95.78	88.74/98.55	72.12/86.46	65.83/88.22	73.20/89.19
RFC	80.31/95.72	73.44/93.74	79.39/95.41	72.56/91.88	66.32/89.90	72.62/92.20
XGB	88.33/98.84	76.52/96.23	86.80/98.56	73.18/88.15	64.74/86.08	72.49/88.58
TNC	79.02/91.55	68.73/87.99	80.22/92.24	65.75/81.17	60.37/82.04	67.20/82.95
FTT	86.62/96.88	75.09/93.78	86.98/97.18	71.18/84.83	64.76/86.50	71.68/86.21

Moreover, to gain insights into the run time (since all the experiments were conducted under the same conditions and some people may have to use the models in the regime of a low tuning time budget), we conducted an in-depth analysis and plotted the time-performance characteristic for the models in Figure 3. For each meticulously tuned configuration, we conducted ten experiments, each with different random seeds (all integers ranging from zero to nine) and reported both the (averaged) training performance (denoting how well the models can fit the data it has seen during the training process) and the results obtained on the test set (using the default seed). As the encoding and the noise addition did not significantly affect the training time nor the inference time, we only show the time of the models used with the reciprocal rank and without noise addition. We can see that while the RFC and TNC have the lowest training time, their inference time are significantly higher compared to the

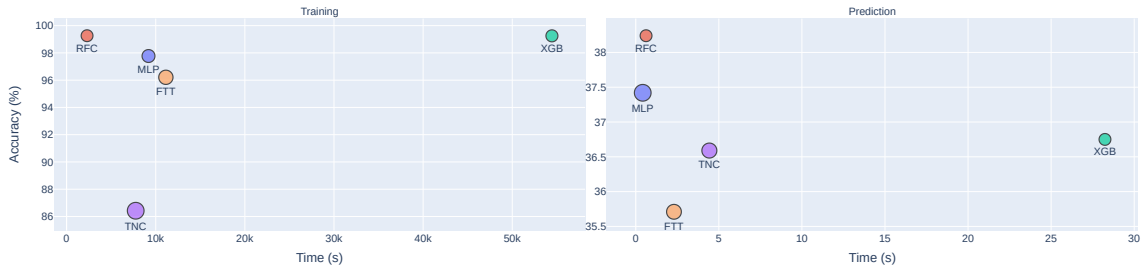


Figure 3: Training with ten different random states on the entire EVA training dataset of 886 260 samples (left) and prediction on the NPMS testing dataset of 7 521 samples (right) time–performance characteristics for selected models, with features encoded with the reciprocal rank method (without noise addition). The circle size reflects the top-one micro average multiclass accuracy standard deviation (left) and the size of the model, i.e., the number of trainable parameters for deep learning algorithms and the number of estimators (i.e., respectively the number of trees in the forest for RFC and the number of gradient boosted trees for the XGB) for machine learning algorithms (right).

MLP, so there is no universally superior solution in terms of time resources. These two comparisons (i.e., Table 1 and Figure 3) allowed us to make some interesting findings, highlighting the nuanced trade-offs between various models and encodings, and emphasizing the importance of selecting the most appropriate approach based on both performance and runtime considerations:

- Models based on decision tree ensembles, such as RFC or XGB, can still outperform some of the deep learning models we kept in our experiments, while requiring either a significantly shorter (RFC) or a significantly higher (XGB) amount of time to train.
- Although there has been a clear trend towards transformer-based solutions in recent years, these models, such as TNC and FTT, do not consistently outperform standard neural network architectures, such as MLP.
- The reciprocal rank encoding usually leads to a better performance than the cover-abundance (except for tree-based models), despite providing less information about the plant species composition in a given vegetation plot.
- Recent state-of-the-art specialized neural network architectures (e.g., TNC and FTT) and strong traditional ML methods (e.g., RFC and XGB) do not provide any benefit over a tuned MLP, which is still more than a simple baseline or a good sanity check (Kadra et al., 2021).

Based on these promising findings, we opted to proceed with the configuration that emerged as the standout performer (i.e., using a MLP classifier with features encoded using the reciprocal rank method and no noise addition) for the subsequent experiments. Indeed, this option was the best trade-off between predictive performance and computational complexity. This strategic choice will be useful for the next phases of our research (i.e., evaluation and interpretability of this configuration and rigorous comparison with the expert system). We dive into the explainability of our models and the ecological interpretability of the results in Appendix S6 (for example, we show that around 85% of the information about the habitat classification of a vegetation plot is brought by vascular plant species only). Having concluded the rigorous process of model selection, which included hyperparameter tuning and the identification of the most effective encoding technique, we proceeded to re-train the chosen model on the entire training dataset. This approach allowed us to evaluate the model's performance in a holistic manner (i.e., without partitioning the available data into sets and holding out one of them for evaluation) to compare it to the EUNIS-ESy.

3.2 Evaluation: diving into the performance of the best model

Up until now, we employed the micro average multiclass accuracy to measure the performance of our models. Due to significant class imbalance within the dataset (e.g., we had almost 10 000 times

more samples of the R22 habitat than samples of the R1L habitat in the training set), we aggregated the contributions of all habitats to compute the average metric. However, in some cases, the micro average may not be the most appropriate metric to evaluate the overall performance of the models. For example, what if we were interested in measuring the performance of the model on each habitat separately, rather than considering the overall performance of the model across all habitats? In such cases, we turn to the macro average multiclass accuracy metric instead (still with $k = 1$ and $k = 3$), which is obtained by computing micro average multiclass accuracy for each class separately and then taking the average over classes. This approach ensures that the habitats with only a few vegetation plots contribute the same as the habitats with thousands of vegetation plots to the assessment of the model's performance. The use of the macro average multiclass accuracy mitigates the potential issue of smaller classes being overshadowed by larger classes in the overall evaluation of the model's performance.

Before delving into the habitat-specific performance of our model, we conducted further experimentation by training two new MLPs with the reciprocal rank encoding using the exact same hyperparameters as before, except for one crucial alteration: the reduction that is applied over labels which we replaced by the macro average (the statistics were calculated for each label and then averaged, but we still used one and three as the numbers of highest probability or logit score predictions considered to find the correct labels). There are much more variations between the different folds and a reduction in overall accuracy compared to our previous micro-average results (across all ten CV folds, the model achieved an average multiclass macro-average accuracy of respectively 73.97% and 90.80% for the top-one and top-three metrics, against an average of 88.74% and 98.55% in micro-average accuracy). While our goal was to maintain consistency by employing the same model throughout our experiments, it is important to acknowledge that for habitat-wise performance assessments it is possible to enhance the results of the MLP model. One promising avenue for improvement is to explore alternative loss functions, for example by switching the currently employed loss function (i.e., the cross-entropy loss (Good, 1952)) for the imbalanced top-one and top-three losses, which, after fine-tuning using a grid of parameter values recommended by the authors of the function, outperformed the model's performance under the existing setup.

3.3 Comparison: evaluating the performance of hdm-framework and EUNIS-ESy

Of all 886 260 vegetation plots from the dataset we used for the expert system, 742 498 were classified to exactly one habitat of the level three of one of the eight habitat groups we considered in this study (i.e., MA2, N, Q, R, S, T, U or V). Among the 143 762 other vegetation plots, 11% (i.e., 15 558 vegetation plots) remained unclassified and 4% (i.e., 5748) were classified to more than one habitat. The rest of the vegetation plots (i.e., 122 456 vegetation plots) were classified as one of: habitat groups (i.e., level one habitats), broad habitat types (i.e., level two habitats) or unrevised habitas (i.e., habitats not part of the current EUNIS list). The expert system achieved an accuracy of 85.20%. As the expert system itself was the tool that was used to classify the vegetation plots from EVA, this study shows the lack of robustness to species names normalization of the expert system which clearly overfits the original data. We dive deeper into this evaluation exercise in Appendix S5.

4 Discussion

4.1 Main advantages of hdm-framework

We explain in detail the methodology and use of hdm-framework in Appendix S7. Our different experiments have highlighted the remarkable efficacy of AI in classifying vegetation-plot records into their respective EUNIS habitats, marking a significant milestone as the first tool to automate this process across Europe using deep learning techniques. Notably, our framework not only surpasses the performance of traditional expert systems but also achieves over double the classification accuracy, all while processing data more than 50 times faster than a recently developed electronic expert system. This efficiency carries profound academic and practical implications, benefiting phytosociologists and related fields by potentially expediting research processes and enabling timely conservation initiatives.

Furthermore, our work not only underscores the potential of AI within this domain but also points toward a broader paradigm shift in favor of advanced AI solutions. While we acknowledge the need for continued exploration and potential challenges on the horizon, our framework lays a robust foundation for future research and applications in habitat classification. It represents a significant leap forward in the practical utility of the EUNIS habitat classification system.

EUNIS-ESy, relying on species cover information, encounters limitations when attempting to classify vegetation plots that only record the presence of species without specifying their covers. In contrast, our hdm-framework seamlessly accommodates presence-only data, extending the applicability of such data. Furthermore, traditional expert systems typically assess every vegetation plot within a database, scrutinizing each one to determine if it aligns with one or more predefined habitat definitions specified in their scripts. This process can sometimes lead to vegetation plots remaining unclassified by the expert system. In contrast, the deep learning models we present in this study were meticulously trained to assign each vegetation plot to (at least) one habitat.

hdm-framework is an HDM platform facilitating the use of species occurrence data and environmental features retrieved from multiple sources. Inspired from the existing literature, we proposed several methods that are fast enough to deliver results for thousands of vegetation plots in less than a second. Provided with a set of 195 tunable parameters, hdm-framework has been designed for high customization flexibility, so it can be adapted to anyone's objectives and computing environment. In contrast to the expert system which doesn't extract itself environmental features, the framework will derive them from the vegetation plot coordinates using the relevant shapefiles already provided and store the calculated values (e.g., location on coastal dunes or in a certain ecoregion) to the header data of the vegetation plots.

4.2 Potential improvements for practical applications

We discuss the inherent limitations of the training and testing dataset in Appendix S4. An essential aspect of our methodology revolves around the normalization of species names using the GBIF Backbone Taxonomy. This step plays a pivotal role in ensuring consistency and facilitating cross-dataset comparisons, making it a necessary component of our approach. However, it is important to acknowledge that this process comes with inherent trade-offs, including the loss of valuable information pertaining to species variations and local taxonomic nuances. The harmonization of species names, while promoting uniformity, can inadvertently lead to the amalgamation of distinct taxa or the division of a single taxon into multiple names. Such outcomes have the potential to influence the accuracy of our classification results. Notably, in some instances, phytosociology experts conducting vegetation surveys may have recorded species at a higher taxonomic level, such as specifying the genus (e.g., *Quercus*), without providing precise species designations. This practice presents a challenge during the normalization process, particularly when the GBIF Backbone Taxonomy relies on explicit species information. Consequently, the normalization of higher-level taxonomic names may not always be feasible, potentially impacting the precision of species classification within our framework. It is imperative to recognize and navigate this inherent trade-off between achieving consistency and comparability through species name normalization and the potential loss of finer taxonomic details. This trade-off significantly influences the interpretation and reliability of our classification results, warranting careful consideration in our biodiversity monitoring efforts. Furthermore, the GBIF API works against data kept in the GBIF Checklist Bank (in partnership with the Catalogue of Life (Bánki et al., 2023)) which taxonomically indexes all registered checklist datasets in the GBIF network. It is important to note that this taxonomy store is constantly evolving through updates and takes taxonomic and nomenclatural information from different and new sources, thus potentially resulting in unreproducible results. However, the widespread public deployment of large language models in recent months (Zhao et al., 2023) might offer new opportunities. For example, it could soon be possible to train AI tools on data that have non-standardized nomenclature.

Moreover, the efficacy of our model is intrinsically linked to the taxonomic diversity of vascular plant species present in the training dataset (EVA). As our models are trained on this dataset, their ability to recognize and classify species is contingent on exposure during training. While in Europe

there are more than 20 000 species of vascular plants (Med, 2006), our framework was trained on a subset comprising 10 481 distinct vascular plants. Consequently, when tasked with classifying plots that contain species not represented in the training set, certain limitations come to the forefront. In instances where our trained models encounter species absent from the training data, we confront a challenge. To address this issue, it becomes necessary to exclude species not encompassed in the training set, as our models may lack familiarity with these unrepresented species. Consequently, this constraint introduces the potential for classification errors, especially in scenarios where a substantial proportion of species within a plot diverge from those within the training set. This limitation is a crucial consideration when applying our framework to novel datasets (Schmidt et al., 2012) or datasets characterized by high species diversity (Botella et al., 2023b). To enhance the framework's utility and robustness, future endeavors could concentrate on broadening the training set to encompass a more extensive spectrum of species. This expansion could be achieved through various means, including the acquisition of supplementary data sources (Estopinan et al., 2022) or collaboration with domain experts to identify and incorporate missing species (Szymura et al., 2023). Exploring strategies to mitigate the impact of species mismatch between training and testing data would be pivotal, further augmenting the framework's versatility and applicability in diverse vegetation classification scenarios.

An essential limitation of our framework pertains to its reliance on predefined habitats for classification. The predictions generated by our models are grounded in the established definitions of EUNIS habitats at the time of model training. In this paper, we focus on eight distinct habitat groups, reflecting the updated EUNIS classification: littoral biogenic habitats, coastal habitats, wetlands, grasslands and lands dominated by forbs, mosses or lichens, heathlands, scrub and tundra, forests and other wooded land, inland habitats with no or little soil and mostly with sparse vegetation and vegetated man-made habitats. However, it's paramount to recognize that the dynamism of environmental classifications can result in evolving habitat definitions or the emergence of entirely new habitats, driven by agencies such as the EEA. In such cases, our models would necessitate retraining with vegetation plots categorized according to these revised or newly established habitat types. This process can be resource-intensive and potentially environmentally taxing, given the associated energy consumption (Strubell et al., 2020). Therefore, we must acknowledge this limitation and emphasize the importance of periodic model updates to align with any changes in habitat definitions. Furthermore, it underscores the need to consider the ecological footprint of these retraining procedures and explore strategies to optimize their efficiency and sustainability. This may encompass efforts to minimize energy consumption, employ renewable energy sources during the training phase, or investigate eco-friendly training methodologies. By doing so, we can ensure that our framework remains adaptable and environmentally responsible in the face of evolving habitat classifications.

Currently, our framework operates by selecting an integer K (by default set to one) and returning the top-K habitats with the highest score, a method known as top-K classification. Given the complexity of classifying vegetation plots into a substantial number of habitats (a total of 228), relying on a single value for K can lead to challenges in precision. To address this issue, we conducted experiments with $K = 3$. However, our observations revealed that in cases of high certainty, such as T3B (i.e., Pinus canariensis forest, where our MLP model, trained using the reciprocal rank feature encoding method without noise addition, achieved an impressive average top-one micro average multiclass accuracy of 98.95% across all ten folds), employing $K > 1$ resulted in an excessive number of predictions. Conversely, for instances characterized by significant ambiguity, like R1L (i.e., Madeiran oromediterranean siliceous dry grassland, where the same model, trained using the same method, achieved an average accuracy of 0.00% with the same metric and evaluation procedure, although it should be noted that only ten occurrences of this habitat are present in EVA), employing $K \leq 3$ (for example) proved to be overly restrictive. An alternative and promising strategy to address this challenge is the implementation of conformal prediction (Gammerman et al., 2013). This approach dynamically adjusts the number of predicted habitats based on the computed ambiguity for each sample, while still aiming to maintain an average of K predictions across all samples, a technique referred to as average-K classification (Lorieul et al., 2021). While this approach presents a potential solution for handling ambiguity more effectively, it is important to note that it has not yet been integrated into our framework but represents a promising avenue for future development.

5 Conclusions

In summary, the deep learning framework presented in this paper has demonstrated its remarkable capability to accurately assign vegetation-plot records to their respective EUNIS habitats, as confirmed through rigorous expert evaluation. This framework not only achieves high accuracy but also ushers in a new era of possibilities. It helps big vegetation data classification and management. The results produced, that are understandable to experts in vegetation classification, highlight the importance of dominant species and the species composition of sites as a whole. The fusion of data sources offers unprecedented flexibility, making it suitable for a wide spectrum of applications across diverse habitat types. For instance, as we consistently assign a substantial number of vegetation plots from various European regions to EUNIS habitat classifications using our framework, it paves the way for precise characterizations of species composition, distribution patterns, and their intricate environmental associations within these habitats. The development of this comprehensive framework represents a significant step towards more efficient, accurate and cost-effective classification of habitat types.

Acknowledgements

The authors extend their gratitude to several individuals and organizations whose contributions were instrumental in the completion of this research. We acknowledge the Observatoire Pluridisciplinaire des Alpes-Maritimes (OPAL) infrastructure at Université Côte d'Azur for providing essential resources, including high-performance computing facilities, AI computing resources, data storage, and computational support. We also express our appreciation to the dedicated scientists who collected the original vegetation-plot observations within the EVA and the generous database owners and representatives who facilitated access to this valuable data. Special thanks to database administrator Ilona Knollová for her assistance in obtaining the data. The NPMS played a pivotal role in organizing and funding this initiative, and we gratefully acknowledge their support. Our deepest thanks go to the volunteers whose data contributions were indispensable for this study's success. All graphical representations in this paper were created using the Plotly (Inc, 2015) and Matplotlib (Hunter, 2007) plotting libraries. We sincerely appreciate the efforts of all contributors to these open-source libraries, which greatly enhanced the quality of our visualizations. Our analysis also benefited from the FloraVeg.EU website (Milan et al., 2022), a valuable resource for European vegetation types, habitats, and plant species information. Furthermore, we acknowledge OpenStreetMap (OSM), an open data platform licensed under the Open Data Commons Open Database License (ODbL) by the OpenStreetMap Foundation (OSMF) that we used to generate all maps presented in this work. We recognize the OSMF community for their dedicated work in maintaining and curating this invaluable geographic resource, which was instrumental in our study. The collective support and resources provided by these organizations and individuals significantly enriched our research endeavors.

Author contributions

C.L., P.B., M.S. and A.J. were involved in the initial idea for the project, helped to define the research questions and objectives and contributed to the overall design of the study; C.L., with contributions from P.B., M.S. and A.J. who conducted analyses to compare the performance of the models to the expert system, was responsible for developing the deep learning framework and ensuring that it was robust, accurate, efficient and well-documented; C.L., with contributions from P.B., M.S. and A.J. who ensured that it was of high quality and suitable for the research questions, was responsible for gathering and organizing the data used in the study; C.L. was responsible for writing the first draft of the paper and ensuring that it met the standards of the journal; all the other authors were responsible for curating the data delivered from EVA for this study and reviewing and editing the paper; P.B., M.S. and A.J. were responsible for overseeing the project as a whole, providing guidance and support to C.L. and ensuring that the research was conducted ethically and rigorously; P.B. was responsible for securing funding for the project, ensuring that the necessary resources were available to conduct the research. The contributions of each author were integral to the successful execution of this research.

Data availability statement

This article utilizes data from the European Vegetation Archive (EVA), a comprehensive multi-contributor database. The EVA data selection used for this project is stored in the EVA archive at https://doi.org/10.58060/QR4B-G979. While we are unable to publicly share the specific dataset used due to third-party restrictions, the vegetation plots we utilized are accessible for research purposes. To replicate our results or conduct further analysis, researchers can submit a proposal to the EVA Coordinating Board to download the data from the archive stored under the above-mentioned Digital Object Identifier (DOI). In contrast, the dataset from the National Plant Monitoring Scheme (NPMS) is available under the terms of the Open Government Licence v3 (OGL), which permits unrestricted use and reuse. Interested parties can freely access and utilize the NPMS dataset, with conditions as specified by the license. For transparency and reproducibility, the scripts used to generate the analyses presented in this paper, along with the corresponding command lines, are publicly available and can be accessed at https://github.com/cesar-leblanc/hdm-framework/tree/main/Experiments.

ORCID

César Leblanc https://orcid.org/0000-0002-5682-8179
Pierre Bonnet https://orcid.org/0000-0002-2828-4389
Maximilien Servajean https://orcid.org/0000-0002-9426-2583
Milan Chytrý https://orcid.org/0000-0002-8122-3075
Svetlana Aćić https://orcid.org/0000-0001-6553-3797
Olivier Argagnon https://orcid.org/0000-0003-2069-7231
Ariel Bergamini https://orcid.org/0000-0001-8816-1420
Idoia Biurrun https://orcid.org/0000-0002-1454-0433
Gianmaria Bonari https://orcid.org/0000-0002-5574-6067
Juan A. Campos https://orcid.org/0000-0001-5992-2753
Andraž Čarni https://orcid.org/0000-0002-8909-4298
Renata Ćušterevska https://orcid.org/0000-0002-3849-6983
Michele De Sanctis https://orcid.org/0000-0002-7280-6199
Jürgen Dengler https://orcid.org/0000-0003-3221-660X
Tetiana Dziuba https://orcid.org/0000-0001-8621-0890
Emmanuel Garbolino https://orcid.org/0000-0002-4954-6069
Ute Jandt https://orcid.org/0000-0002-3177-3669
Florian Jansen https://orcid.org/0000-0002-0331-5185
Maria Lebedeva https://orcid.org/0000-0002-5020-527X
Jonathan Lenoir https://orcid.org/0000-0003-0638-9582
Jesper Erenskjold Moeslund https://orcid.org/0000-0001-8591-7149
Aaron Pérez-Haase https://orcid.org/0000-0002-5974-7374
Remigiusz Pielech https://orcid.org/0000-0001-8879-3305
Jozef Šibík https://orcid.org/0000-0002-5949-862X
Zvjezdana Stančić https://orcid.org/0000-0002-6124-811X
Angela Stanisci https://orcid.org/0000-0002-5302-0932
Grzegorz Swacha https://orcid.org/0000-0002-6380-2954
Domas Uogintas https://orcid.org/0000-0002-3937-1218
Kiril Vassilev https://orcid.org/0000-0003-4376-5575
Thomas Wohlgemuth https://orcid.org/0000-0002-4623-0894
Valentin Golub https://orcid.org/0000-0003-3973-6608
Alexis Joly https://orcid.org/0000-0002-2161-9940

References

Arik, S. O. and Pfister, T. (2019). Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bánki, O., Hobern, D., Döring, M., Ower, G., Roskov, Y., Hernandez-Robles, D., Plata, C., Schalk, P., and Orrell, T. (2023). Towards a quality assurance and quality control mechanism for species list building. *Biodiversity Information Science and Standards*, 7:e111665.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bircanoğlu, C. and Arıca, N. (2018). A comparison of activation functions in artificial neural networks. In *2018 26th signal processing and communications applications conference (SIU)*, pages 1–4. IEEE.

Bonnet, P., Affouard, A., Lombardo, J.-C., Chouet, M., Gresse, H., Hequet, V., Palard, R., Fromholtz, M., Espitalier, V., Goëau, H., et al. (2023). Synergizing digital, biological, and participatory sciences for global plant species identification: Enabling access to a worldwide identification service. *Biodiversity Information Science and Standards*, 7:e112545.

Bonnet, P., Joly, A., Faton, J.-M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A., Lombardo, J.-C., Mary, L., et al. (2020). How citizen scientists contribute to monitor protected areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, 1(2):e12023.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Botella, C., Deneu, B., Gonzalez, D. M., Servajean, M., Larcher, T., Leblanc, C., Estopinan, J., Bonnet, P., and Joly, A. (2023a). Overview of geolifeclef 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing. *Working Notes of CLEF*.

Botella, C., Deneu, B., Marcos, D., Servajean, M., Estopinan, J., Larcher, T., Leblanc, C., Bonnet, P., and Joly, A. (2023b). The geolifeclef 2023 dataset to evaluate plant species distribution models at high spatial resolution across europe. *arXiv preprint arXiv:2308.05121*.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (2018). A deep learning approach to species distribution modelling. *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, pages 169–199.

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S. M., Chytrỳ, M., Pillar, V. D., Jansen, F., Kattge, J., Sandel, B., et al. (2019). splot–a new tool for global vegetation analyses. *Journal of Vegetation Science*, 30(2):161–186.

Brun, P., Karger, D. N., Zurell, D., Descombes, P., de Witte, L., de Lutio, R., Wegner, J. D., and Zimmermann, N. E. (2023). Rank-based deep learning from citizen-science data to model plant communities. *bioRxiv*, pages 2023–05.

Černá, L. and Chytrỳ, M. (2005). Supervised classification of plant communities with artificial neural networks. *Journal of vegetation Science*, 16(4):407–414.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chytrỳ, M. et al. (2012). Vegetation of the czech republic: diversity, ecology, history and dynamics. *Preslia*, 84(3):427–504.

Chytrỳ, M., Hennekens, S. M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F., Landucci, F., Schaminée, J. H., Aćić, S., Agrillo, E., et al. (2016). European vegetation archive (eva): an integrated database of european vegetation plots. *Applied vegetation science*, 19(1):173–180.

Chytry, M., Tichy, L., Hennekens, S., Knollova, I., Janssen, J., Rodwell, J., Peterka, T., Marceno, C., Landucci, F., Danihelka, J., et al. (2021). Eunis-esy: Expert system for automatic classification of european vegetation plots to eunis habitats.

Chytrỳ, M., Tichỳ, L., Hennekens, S. M., Knollová, I., Janssen, J. A., Rodwell, J. S., Peterka, T., Marcenò, C., Landucci, F., Danihelka, J., et al. (2020). Eunis habitat classification: Expert system, characteristic species combinations and distribution maps of european habitats. *Applied Vegetation Science*, 23(4):648–675.

Davies, C. and Moss, D. (1999). Eunis habitat classification. final report to the european topic centre on nature conservation. *European Environment Agency*, 256.

De Cáceres, M., Chytrỳ, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., Czúcz, B., Dengler, J., Ewald, J., Faber-Langendoen, D., et al. (2015). A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science*, 18(4):543–560.

Deneu, B., Joly, A., Bonnet, P., Servajean, M., and Munoz, F. (2022). Very high resolution species distribution modeling based on remote sensing imagery: how to capture fine-grained and large-scale vegetation ecology with convolutional neural networks? *Frontiers in plant science*, 13:839279.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., and Joly, A. (2021). Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS computational biology*, 17(4):e1008856.

Dengler, J., Jansen, F., Glöckler, F., Peet, R. K., De Cáceres, M., Chytrỳ, M., Ewald, J., Oldeland, J., Lopez-Gonzalez, G., Finckh, M., et al. (2011). The global index of vegetation-plot databases (givd): a new resource for vegetation science. *Journal of Vegetation Science*, 22(4):582–597.

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327.

Estopinan, J., Bonnet, P., Servajean, M., Munoz, F., and Joly, A. (2024). Modelling species distributions with deep learning to predict plant extinction risk and assess climate change impacts. *arXiv preprint arXiv:2401.05470*.

Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., and Joly, A. (2022). Deep species distribution modeling from sentinel-2 image time-series: A global scale analysis on the orchid family. *Frontiers in Plant Science*, 13.

Evans, D. (2012). The eunis habitats classification–past, present & future. *Revista de investigación marina*, 19(2):28–29.

Feurer, M. and Hutter, F. (2019). Hyperparameter optimization. *Automated machine learning: Methods, systems, challenges*, pages 3–33.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gammerman, A., Vovk, V., and Vapnik, V. (2013). Learning by transduction. *arXiv preprint arXiv:1301.7375*.

Garcin, C., Servajean, M., Joly, A., and Salmon, J. (2022). Stochastic smoothing of the top-k calibrated hinge loss for deep imbalanced classification. In *International Conference on Machine Learning*, pages 7208–7222. PMLR.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114.

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943.

Hall, L. S., Krausman, P. R., and Morrison, M. L. (1997). The habitat concept and a plea for standard terminology. *Wildlife society bulletin*, pages 173–182.

Hancock, J. T. and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):1–41.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

Hennekens, S. M. and Schaminée, J. H. (2001). Turboveg, a comprehensive data base management system for vegetation data. *Journal of vegetation science*, 12(4):589–591.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Inc, P. T. (2015). Collaborative data science, montreal, qc: Plotly technologies inc.

Janssen, J., Rodwell, J., Criado, M. G., Gubbay, S., Haynes, T., Nieto, A., Sanders, N., and Calix, M. (2016). *European red list of habitats*. Publications Office of the European Union Luxembourg.

Joly, A., Botella, C., Picek, L., Kahl, S., Goëau, H., Deneu, B., Marcos, D., Estopinan, J., Leblanc, C., Larcher, T., et al. (2023). Overview of lifeclef 2023: evaluation of ai models for the identification and prediction of birds, plants, snakes and fungi. In *International Conference of the Cross-Language Evaluation Forum for European Languages, Springer*.

Joseph, L. N., Field, S. A., Wilcox, C., and Possingham, H. P. (2006). Presence–absence versus abundance data for monitoring threatened species. *Conservation biology*, 20(6):1679–1687.

Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. (2021). Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Leblanc, C., Joly, A., Lorieul, T., Servajean, M., and Bonnet, P. (2022). 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*, pages 2123–2150.

Lorieul, T., Joly, A., and Shasha, D. (2021). Classification under ambiguity: When is average-k better than top-k? *arXiv preprint arXiv:2112.08851*.

Marcenò, C., Guarino, R., Loidi, J., Herrera, M., Isermann, M., Knollová, I., Tichỳ, L., Tzonev, R. T., Acosta, A. T. R., FitzPatrick, Ú., et al. (2018). Classification of european and mediterranean coastal dune vegetation. *Applied Vegetation Science*, 21(3):533–559.

Med, E. (2006). Euro+ med plantbase–the information resource for euro-mediterranean plant diversity. *October 9 2014*.

Michalcová, D., Lvončík, S., Chytrỳ, M., and Hájek, O. (2011). Bias in vegetation databases? a comparison of stratified-random and preferential sampling. *Journal of Vegetation Science*, 22(2):281–291.

Milan, C., Irena, A., Dana, H., Petr, N., Marcela, Ř., Idoia, B., Gianmaria, B., Natálie, Č., Jiří, D., Pavel, D., et al. (2022). Floraveg. eu–a new online database of european vegetation and flora. In *Plant communities in changing environment*. Richard Hrivnák & Michal Slezák.

Morrison, L. W. (2021). Nonsampling error in vegetation surveys: understanding error types and recommendations for reducing their occurrence. *Plant Ecology*, 222(5):577–586.

Moss, D. (2008). Eunis habitat classification–a guide for users. *European Topic Centre on Biological Diversity*.

Mucina, L., Bültmann, H., Dierßen, K., Theurillat, J.-P., Raus, T., Čarni, A., Šumberová, K., Willner, W., Dengler, J., García, R. G., et al. (2016). Vegetation of europe: hierarchical floristic classification system of vascular plant, bryophyte, lichen, and algal communities. *Applied vegetation science*, 19:3–264.

Noble, I. (1987). The role of expert systems in vegetation science. *Vegetatio*, 69:115–121.

Novák, P., Willner, W., Biurrun, I., Gholizadeh, H., Heinken, T., Jandt, U., Kollár, J., Kozhevnikova, M., Naqinezhad, A., Onyshchenko, V., et al. (2023). Classification of european oak–hornbeam forests and related vegetation types. *Applied Vegetation Science*, 26(1):e12712.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Scherrer, D., Mod, H. K., and Guisan, A. (2020). How to evaluate community predictions without thresholding? *Methods in Ecology and Evolution*, 11(1):51–63.

Schmidt, M., Janßen, T., Dressler, S., Hahn, K., Hien, M., Konaté, S., Lykke, A. M., Mahamane, A., Sambou, B., Sinsin, B., et al. (2012). The west african vegetation database. *Biodiversity and Ecology*, 4:105–110.

Sietsma, J. and Dow, R. J. (1991). Creating artificial neural networks that generalize. *Neural networks*, 4(1):67–79.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.

Szymura, T. H., Kassa, H., Swacha, G., Szymura, M., Zając, A., and Kącki, Z. (2023). Vegetation databases augment but do not replace species distribution atlases in species richness assessment. *Ecological Indicators*, 154:110876.

Tichỳ, L., Chytrỳ, M., and Landucci, F. (2019). Grimp: A machine-learning method for improving groups of discriminating species in expert systems for vegetation classification. *Journal of Vegetation Science*, 30(1):5–17.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Walker, K., Pescott, O., Harris, F., Cheffings, C., New, H., Bunch, N., and Roy, D. (2015). Making plants count. *British Wildlife*, 26(4):243–250.

Westhoff, V. and Van Der Maarel, E. (1978). *The braun-blanquet approach*. Springer.

Wiser, S. K., Cáceres, M. d., et al. (2018). New zealand's plot-based classification of vegetation. *Phytocoenologia*, 48(2):153–161.

Yapp, R. H. (1922). The concept of habitat. *Journal of Ecology*, 10(1):1–17.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., Wei, L., et al. (2015). Linking in situ vegetation data to the eunis habitat classification: results for forest habitats.

List of Tables

List of Figures

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

Appendix S1. Visual overview of the architectures, explanations of the parameters and model evaluation tables, containing the list of tuned hyperparameters with the search spaces and optimal values, the list of fixed hyperparameters with the selected values, and the hardware used, time spent, and result obtained for tuning and optimizing each combination
Appendix S2. List of all plants species contained in the training dataset from EVA
Appendix S3. Table listing all habitats from the level three of the EUNIS hierarchy that are present in the EVA training set or NPMS test set, with their codes, their names, their conservation statuses based on the European Red List of Habitats and the number of training and testing vegetation plots assigned to each of them
Appendix S4. Preprocessing steps to create the two datasets used to evaluate habitat distribution models and their limitations
Appendix S5. Comparison and evaluation of the performance of hdm-framework and EUNIS-ESy
Appendix S6. Understanding how our models reason using interpretability
Appendix S7. Guide on the classification framework to enhance habitat distribution models