



**HAL**  
open science

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

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

### ► To cite this version:

César Leblanc, Pierre Bonnet, Maximilien Servajean, Milan Chytrý, Svetlana Ačić, et al.. A deep-learning framework for enhancing habitat identification based on species composition. *Applied Vegetation Science*, 2024, 27 (3), pp.e12802. 10.1111/avsc.12802 . hal-04700157

**HAL Id: hal-04700157**

**<https://hal.science/hal-04700157v1>**

Submitted on 19 Nov 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

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

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

<sup>1</sup> Inria, LIRMM, Université de Montpellier, CNRS, Montpellier, FR

<sup>2</sup> AMAP, Université de Montpellier, CIRAD, CNRS, INRA, IRD, Montpellier, France

<sup>3</sup> LIRMM, AMIS, Université Paul-Valéry - Montpellier 3, CNRS, Montpellier, FR

<sup>4</sup> Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, CZ

<sup>5</sup> Department of Botany, Faculty of Agriculture, University of Belgrade, Belgrade, RS

<sup>6</sup> Antenne Languedoc-Roussillon, Conservatoire botanique national méditerranéen, Hyères, FR

<sup>7</sup> Swiss Federal Research Institute WSL, Birmensdorf, CH

<sup>8</sup> Department of Plant Biology and Ecology, University of the Basque Country UPV/EHU, Bilbao, ES

<sup>9</sup> Department of Life Sciences, University of Siena, Siena, IT

<sup>10</sup> Institute of Biology, Research Center of the Slovenian Academy of Sciences and Art, Ljubljana, SI

<sup>11</sup> School for Viticulture and Enology, University of Nova Gorica, Nova Gorica, SI

<sup>12</sup> Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Skopje, MK

<sup>13</sup> Department of Environmental Biology, Sapienza University of Rome, IT

<sup>14</sup> Vegetation Ecology Research Group, Institute of Natural Resource Sciences (IUNR), Zurich University of Applied Sciences (ZHAW), Wädenswil, CH

<sup>15</sup> Plant Ecology, Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Bayreuth, DE

<sup>16</sup> Department of Geobotany and Ecology, M.G. Kholodny Institute of Botany, National Academy of Sciences of Ukraine, Kyiv, UA

<sup>17</sup> MINES Paris PSL, ISIGE, Fontainebleau, FR

<sup>18</sup> 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

<sup>19</sup> Geobotany & Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, DE

<sup>20</sup> German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,

Leipzig, DE

<sup>21</sup> Faculty of Agricultural and Environmental Sciences, University of Rostock,  
Rostock, DE

<sup>22</sup> South-Ural Botanical garden-institute of Ufa Federal Research Centre of the  
Russian Academy of Sciences, Ufa, RU

<sup>23</sup> UMR CNRS 7058 “Ecologie et Dynamique des Systèmes Anthropisés” (EDYSAN),  
Université de Picardie Jules Verne, Amiens, FR

<sup>24</sup> Department of Ecoscience, Aarhus University, Aarhus, DK

<sup>25</sup> Department of Evolutionary Biology, Ecology and Environmental Sciences and  
Biodiversity Research Institute (IRBio), University of Barcelona, Barcelona, ES

<sup>26</sup> Institute of Botany, Faculty of Biology, Jagiellonian University, Kraków, PL

<sup>27</sup> Department of Biodiversity and Ecology, Plant Science and Biodiversity Center  
Slovak Academy of Sciences, Bratislava, SK

<sup>28</sup> Faculty of Geotechnical Engineering University of Zagreb, Varaždin, HR

<sup>29</sup> Department of Bioscience and Territory, University of Molise, Termoli, IT

<sup>30</sup> Botanical Garden, University of Wrocław, PL

<sup>31</sup> Nature Research Centre, Vilnius, LT

<sup>32</sup> Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Science,  
Sofia, BG

<sup>33</sup> 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.

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

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

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

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

109 **Keywords** — Artificial intelligence, Biodiversity monitoring, Deep learning, European flora, Ex-  
110 pert system, Habitat type, Phytosociology, Vascular plant species, Vegetation classification

## 111 1 Introduction

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

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

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

136 combining observations of species co-occurrence or abundance with environmental estimates to map  
137 habitat distributions across landscapes. Several tools for vegetation classification with different logic  
138 and strategy are available. In particular machine learning algorithms (Hastie et al., 2009) and expert  
139 systems (Noble, 1987). The former are tools for induction of the independent knowledge base, whereas  
140 the latter emulate the process of expert classification done by humans by using explicitly defined  
141 logical formulas. These (numerical) tools can also play a vital role for nature conservation, landscape  
142 mapping and land-use planning and can facilitate biodiversity management (Estopinan et al., 2024).  
143 They make monitoring of species and habitats easier and more accurate, provide decision-support for  
144 nature conservation and guidance for nature restoration and development. Thus, it can be particularly  
145 valuable in the current context where a significant portion of habitats are at risk of collapsing (at least  
146 32% of terrestrial habitats and 18% of marine habitats are threatened (Janssen et al., 2016)). Therefore,  
147 habitat type classification has a crucial role in ecology, and using the EUNIS habitat classification can  
148 serve as a key instrument for assessing progress towards the European Union’s biodiversity targets.

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

- 164 • **strict**, e.g., to be correctly assigned to its habitat, a vegetation plot should contain at least  $n$   
165 species of a given functional species group, or the total cover of a discriminating species group  
166 in a vegetation plot should be greater than the total cover of other discriminating species groups  
167 in the plot,
- 168 • **complex**, e.g., to be correctly assigned to its habitat, the total cover of a functional species  
169 group in a vegetation plot should be greater than that of another functional group, excluding  
170 the species of the former group from the latter group, or the sum of square-rooted percentage  
171 covers of the species belonging to a discriminating species group in a vegetation plot should be  
172 greater than the sum of square-rooted percentage covers of the species of another discriminating  
173 species group,
- 174 • and **idiosyncratic**, e.g., to be correctly assigned to its habitat, a vegetation plot should belong  
175 to a dataset, or a vegetation plot shouldn’t be located in a country.

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

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

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

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

## 210 2 Methods

### 211 2.1 Data

#### 212 2.1.1 EVA: a comprehensive dataset for habitat classification

213 Our data source for training the deep learning framework was drawn from a subset of a data repository  
214 of vegetation-plot observations (i.e., records of plant taxon co-occurrence and cover-abundance at  
215 particular sites in plots ranging from 1 m<sup>2</sup> to a few hundred m<sup>2</sup> which have been collected by vegetation  
216 scientists (Zhongming et al., 2015)) from Europe and adjacent areas. This EVA database (Chytrý et al.,  
217 2016), which was accessed on 22 May 2023, is an initiative of the Working Group European Vegetation  
218 Survey (EVS). Each of the vegetation plots typically contained estimates of cover-abundance of each  
219 species (vascular plant in every vegetation plot, bryophytes and/or lichens in some vegetation plots)  
220 alongside various supplementary details and additional sources of information on vegetation structure,  
221 location and environmental features. Although the EVA database represents a valuable resource for  
222 studying vegetation patterns and dynamics, we were mindful and acknowledged potential limitations  
223 stemming from the representativeness of the data and the possibility of sampling bias (inherent to  
224 sets of data assembled from multiple sources and originally collected for various purposes) (Michalcová  
225 et al., 2011). The final dataset contained a total of 886 260 georeferenced plots (with an average  
226 of approximately around 20 species per vegetation plot), 228 different habitats and 10 481 different  
227 species. Refer to Appendix S4 for a detailed overview of all the preprocessing steps and to Figure 1  
228 for different visualizations.

#### 229 2.1.2 NPMS: an independent dataset to evaluate models

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

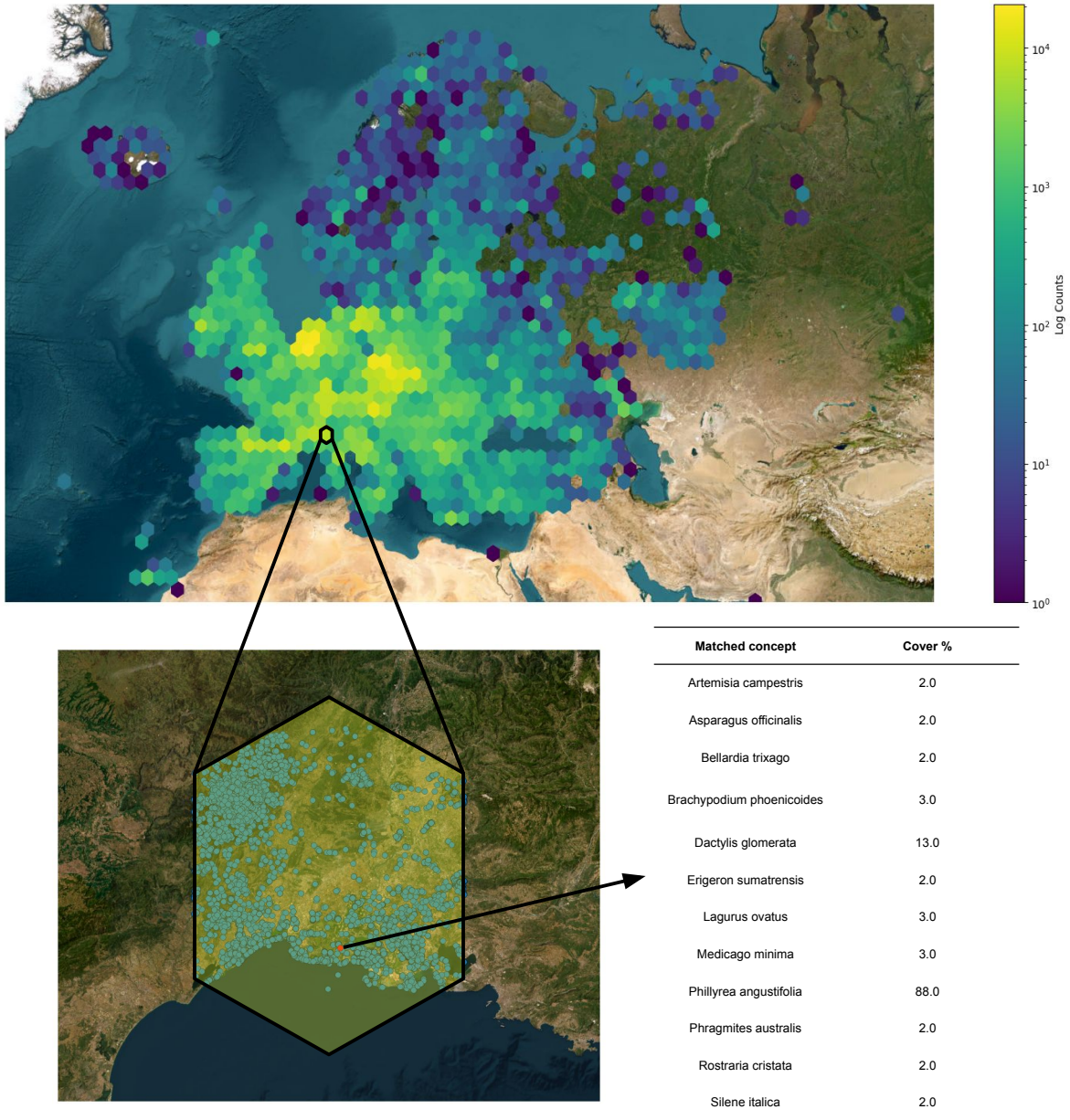


Figure 1: Hexagonal binning showing the distribution of vegetation plots from the training dataset. Zoom in on a specific bin with the raw spatial distribution of the vegetation plots. Further breakdown on a vegetation plot (assigned to the habitat type S51, i.e., Mediterranean maquis and arborescent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

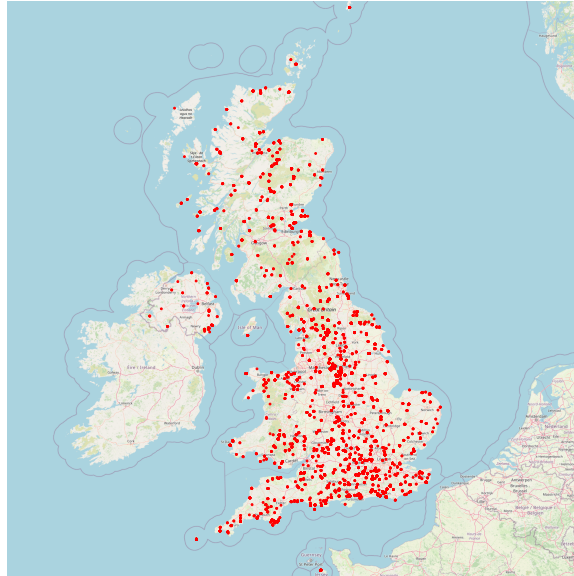


Figure 2: Distribution of vegetation plots in the NPMS test set

238 by citizens (i.e., expert volunteers who carried out surveys of wildflowers and their associated habitats).  
 239 This scheme was designed and developed collaboratively by the Botanical Society of Britain & Ireland  
 240 (BSBI), UK Centre for Ecology & Hydrology (UKCEH), Plantlife and the Joint Nature Conservation  
 241 Committee (JNCC). We specifically chose this dataset because it offered an intriguing opportunity to  
 242 validate the work of numerous European vegetation scientists across generations with a recent citizen  
 243 science project (Bonnet et al., 2023) that employed a systematic protocol and methodology (e.g., the  
 244 participants were allocated a 1km square in which they had to visit five plots in semi-natural habi-  
 245 tats twice a year) and encompassed a wide range of vegetation types, providing valuable insights into  
 246 the potential transferability of our models in a real-world context, beyond expert-driven datasets. It  
 247 offered an interesting contrast by incorporating data collected through citizen science (Bonnet et al.,  
 248 2020), thus expanding our understanding of the generalization of the framework beyond traditional  
 249 scientific datasets. However, this dataset is by nature very different from EVA, and that there is a  
 250 significant distribution shift between the two due to the different collection protocols. So we cannot  
 251 expect the same level of performance. We detail the preprocessing steps to create the test dataset in  
 252 Appendix S4. Refer to Figure 2 for a visual representation of the distribution of the testing dataset.

## 253 2.2 Modeling

### 254 2.2.1 Validation: accounting for the spatial structure of ecological data

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



### 268 2.2.2 Models: using deep neural networks on tabular data for classification

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

- 281 1. A MultiLayer Perceptron classifier (MLP) (Haykin, 1998), i.e., a fully connected class of feed-  
282 forward artificial neural network. It works by taking input data, passing it through multiple  
283 layers of interconnected nodes with weighted connections and activation functions (Bircanoğlu  
284 and Arica, 2018), and producing output predictions based on the learned patterns in the data.
- 285 2. A Random Forest Classifier (RFC) (Ho, 1995), i.e., a meta estimator that fits a number of decision  
286 tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive  
287 accuracy and control over-fitting. A single decision tree works by recursively partitioning the  
288 input data based on the values of its features to create a tree-like structure, where each internal  
289 node represents a feature and each leaf node represents a decision or prediction based on the  
290 input data’s characteristics.
- 291 3. An eXtreme Gradient Boosting classifier (XGB) (Chen and Guestrin, 2016), i.e., an optimized  
292 distributed gradient boosting algorithm designed to be highly efficient, flexible and portable. It  
293 works by iteratively training and adding decision trees to the ensemble model, each focusing on  
294 reducing the residual errors of the previous trees, using a combination of gradient descent opti-  
295 mization (Ruder, 2016), regularization techniques, and hardware-aware optimization to achieve  
296 high accuracy and scalability.
- 297 4. A TabNet Classifier (TNC) (Arik and Pfister, 2019), i.e., a novel high-performance and inter-  
298 pretable canonical deep tabular data learning architecture. It works by selectively attending to  
299 the most informative features of the input data and using a sparse masking technique to allow  
300 for efficient and interpretable feature selection, while employing a multi-step decision-making  
301 process and auxiliary loss functions to enhance its performance and generalization.
- 302 5. A Feature Tokenizer + Transformer classifier (FTT) (Gorishniy et al., 2021), i.e., a model that  
303 transforms all features (categorical and numerical) to embeddings and applies a stack of trans-  
304 former layers to the embeddings. It works by transforming all features to tokens and running a  
305 stack of transformer layers over the tokens, so every transformer layer operates on the feature  
306 level of one object.

### 307 2.2.3 Encodings: mapping current habitat distributions under different constraints

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

- 320 1. The cover-abundance of each species, i.e., the natural logarithm of the raw data from EVA, which  
 321 was recorded using a cover-abundance scale (Westhoff and Van Der Maarel, 1978) and then  
 322 transformed to the arithmetic mid-point percent cover value corresponding to the individual  
 323 cover-abundance class following a comprehensive database management system following the  
 324 default values in the Turboveg database management program (Hennekens and Schaminée, 2001).
- 325 2. The presence-absence of each species, i.e., the binarization of the raw data from EVA. Each  
 326 non-zero entry from the original data is converted to the value one, and every explicit zero are  
 327 preserved (Scherrer et al., 2020).
- 328 3. The reciprocal rank of each species, i.e., the inverse of the ordinal ranking of the raw data from  
 329 EVA. Each species is ranked in descending order of its original value (Brun et al., 2023) and is  
 330 then associated with the value of the inverse of its position in the ranking.

## 331 2.3 Evaluation

### 332 2.3.1 Fitting: evaluating modelling algorithms on selected covariates

333 All details about the models and their optimization are provided in Appendix S1. We evaluated the  
 334 performance of the expert system on the training set we created. We were fully aware that EVA  
 335 was classified using EUNIS-ESy (using its definitions of individual EUNIS habitats based on their  
 336 species composition and geographic location) but we wanted to see if the vegetation plots would  
 337 remain classified to the same habitat after interpreting the taxon names with the GBIF. We thus kept  
 338 the same 886 260 vegetation plots, we took the names from the original database and proceeded to  
 339 normalize them. Furthermore, unlike our experiments for which we kept only vascular plant species  
 340 and species that were observed at least ten times, we also kept in this case species belonging to other  
 341 phyla (especially bryophytes and lichens since they were used by the expert system in the definition  
 342 of some habitats such as *S12*, i.e., moss and lichen tundra) and rare species (as rare species with  
 343 occurrences concentrated in a particular habitat could be used as positive indicators of the habitat by  
 344 the expert system). This process increased the number of observations to 18 867 936 (instead of the  
 345 17 718 306 used to evaluate our models) and the number of different species to 17 885 (instead of the  
 346 10 481 used to evaluate our models). Out of the 886 260 vegetation plots, two of them had no species  
 347 left after the species name matching, and as the expert system (unlike our framework) can't classify  
 348 vegetation plots solely based on external criteria, we added for both vegetation plots a fake species  
 349 named "Unknown species" having a percentage cover of 10%.

### 350 2.3.2 Metrics: computing accuracy to evaluate how well the models are performing

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

- 355 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  
 356 values and  $\hat{y}$  is the predictions. It is the conventional accuracy: the model's prediction must be  
 357 exactly the expected habitat type. This was the most important metric and played a pivotal role  
 358 in our evaluation, as it provided crucial insights into the performance of our approaches when  
 359 we were predicting which habitat was the most likely to be observed at a given location.
- 360 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  
 361 otherwise and where  $y_i$  is a single ground-truth label and  $\hat{y}_{i,k}$  are candidate labels, both associated  
 362 to a sample  $i$ . It means that any of the model's five highest probability predictions must match  
 363 the expected answer. This metric was useful to assess the performances of our methods on similar  
 364 habitats (i.e., habitats that have almost identical species composition and environmental features  
 365 and are thus hard to distinguish from one another) and on scenarios where a vegetation plot was  
 366 associated with several different habitat labels.

367 **2.3.3 Noise: assessing the robustness and generalization of models**

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

383 **3 Results**

384 **3.1 Selection: finding the best performing model**

385 Table 1 contains a comprehensive overview of all the results we obtained (with the models already  
 386 tuned), showcasing the performance of each model-encoding combination. Among the various config-  
 387 urations tested, the model-encoding combination with the best results is a MLP coupled with features  
 388 encoded using the reciprocal rank method. It outperformed other models both with and without noise  
 389 addition to the data and when measuring the performance with the top-one micro average multiclass  
 390 accuracy (since it is the best suited metric in our case, as we want to prioritize the most likely habitat  
 391 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     |

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

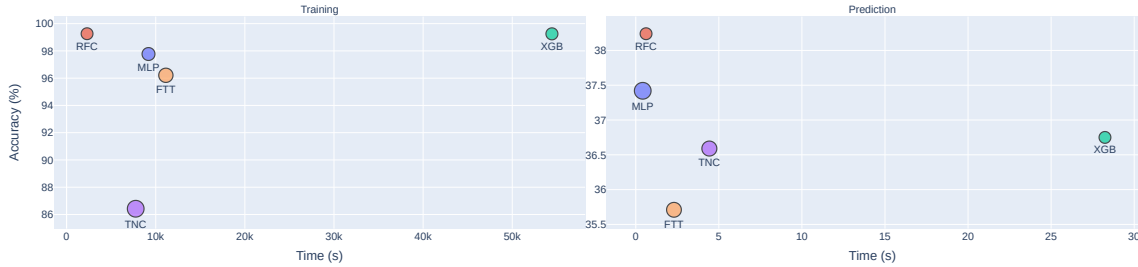


Figure 3: Training with ten different random states on the entire EVA training dataset of 886 260 samples (left) and prediction on the NPMS testing dataset of 7 521 samples (right) time–performance characteristics for selected models, with features encoded with the reciprocal rank method (without noise 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).

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

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

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

### 431 3.2 Evaluation: diving into the performance of the best model

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

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

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

### 462 3.3 Comparison: evaluating the performance of hdm-framework and EUNIS- 463 ESy

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

## 474 4 Discussion

### 475 4.1 Main advantages of hdm-framework

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

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

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

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

## 506 4.2 Potential improvements for practical applications

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

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

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

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

569 Currently, our framework operates by selecting an integer  $K$  (by default set to one) and returning  
570 the top- $K$  habitats with the highest score, a method known as top- $K$  classification. Given the com-  
571 plexity of classifying vegetation plots into a substantial number of habitats (a total of 228), relying on a  
572 single value for  $K$  can lead to challenges in precision. To address this issue, we conducted experiments  
573 with  $K = 3$ . However, our observations revealed that in cases of high certainty, such as T3B (i.e., Pinus  
574 canariensis forest, where our MLP model, trained using the reciprocal rank feature encoding method  
575 without noise addition, achieved an impressive average top-one micro average multiclass accuracy of  
576 98.95% across all ten folds), employing  $K > 1$  resulted in an excessive number of predictions. Con-  
577 versely, for instances characterized by significant ambiguity, like R1L (i.e., Madeiran oromediterranean  
578 siliceous dry grassland, where the same model, trained using the same method, achieved an average  
579 accuracy of 0.00% with the same metric and evaluation procedure, although it should be noted that  
580 only ten occurrences of this habitat are present in EVA), employing  $K \leq 3$  (for example) proved to  
581 be overly restrictive. An alternative and promising strategy to address this challenge is the imple-  
582 mentation of conformal prediction (Gammerman et al., 2013). This approach dynamically adjusts the  
583 number of predicted habitats based on the computed ambiguity for each sample, while still aiming to  
584 maintain an average of  $K$  predictions across all samples, a technique referred to as average- $K$  classifi-  
585 cation (Lorieu et al., 2021). While this approach presents a potential solution for handling ambiguity  
586 more effectively, it is important to note that it has not yet been integrated into our framework but  
587 represents a promising avenue for future development.

## 588 5 Conclusions

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

## 601 Acknowledgements

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

## 622 Author contributions

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



## 636 Data availability statement

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

## 649 ORCID

650 *César Leblanc*  <https://orcid.org/0000-0002-5682-8179>  
651 *Pierre Bonnet*  <https://orcid.org/0000-0002-2828-4389>  
652 *Maximilien Servajean*  <https://orcid.org/0000-0002-9426-2583>  
653 *Milan Chytrý*  <https://orcid.org/0000-0002-8122-3075>  
654 *Svetlana Ačić*  <https://orcid.org/0000-0001-6553-3797>  
655 *Olivier Argagnon*  <https://orcid.org/0000-0003-2069-7231>  
656 *Ariel Bergamini*  <https://orcid.org/0000-0001-8816-1420>  
657 *Idoia Biurrun*  <https://orcid.org/0000-0002-1454-0433>  
658 *Gianmaria Bonari*  <https://orcid.org/0000-0002-5574-6067>  
659 *Juan A. Campos*  <https://orcid.org/0000-0001-5992-2753>  
660 *Andraž Čarni*  <https://orcid.org/0000-0002-8909-4298>  
661 *Renata Čušterevska*  <https://orcid.org/0000-0002-3849-6983>  
662 *Michele De Sanctis*  <https://orcid.org/0000-0002-7280-6199>  
663 *Jürgen Dengler*  <https://orcid.org/0000-0003-3221-660X>  
664 *Tetiana Dziuba*  <https://orcid.org/0000-0001-8621-0890>  
665 *Emmanuel Garbolino*  <https://orcid.org/0000-0002-4954-6069>  
666 *Ute Jandt*  <https://orcid.org/0000-0002-3177-3669>  
667 *Florian Jansen*  <https://orcid.org/0000-0002-0331-5185>  
668 *Maria Lebedeva*  <https://orcid.org/0000-0002-5020-527X>  
669 *Jonathan Lenoir*  <https://orcid.org/0000-0003-0638-9582>  
670 *Jesper Erenskjold Moeslund*  <https://orcid.org/0000-0001-8591-7149>  
671 *Aaron Pérez-Haase*  <https://orcid.org/0000-0002-5974-7374>  
672 *Remigiusz Pielech*  <https://orcid.org/0000-0001-8879-3305>  
673 *Jozef Šibík*  <https://orcid.org/0000-0002-5949-862X>  
674 *Zvezdana Stančić*  <https://orcid.org/0000-0002-6124-811X>  
675 *Angela Stanisci*  <https://orcid.org/0000-0002-5302-0932>  
676 *Grzegorz Swacha*  <https://orcid.org/0000-0002-6380-2954>  
677 *Domas Uogintas*  <https://orcid.org/0000-0002-3937-1218>  
678 *Kiril Vassilev*  <https://orcid.org/0000-0003-4376-5575>  
679 *Thomas Wohlgemuth*  <https://orcid.org/0000-0002-4623-0894>  
680 *Valentin Golub*  <https://orcid.org/0000-0003-3973-6608>  
681 *Alexis Joly*  <https://orcid.org/0000-0002-2161-9940>

## 682 References

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

- 685 Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align  
686 and translate. *arXiv preprint arXiv:1409.0473*.
- 687 Bánki, O., Hobern, D., Döring, M., Ower, G., Roskov, Y., Hernandez-Robles, D., Plata, C., Schalk,  
688 P., and Orrell, T. (2023). Towards a quality assurance and quality control mechanism for species  
689 list building. *Biodiversity Information Science and Standards*, 7:e111665.
- 690 Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new per-  
691 spectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- 692 Bircanoğlu, C. and Arıca, N. (2018). A comparison of activation functions in artificial neural networks.  
693 In *2018 26th signal processing and communications applications conference (SIU)*, pages 1–4. IEEE.
- 694 Bonnet, P., Affouard, A., Lombardo, J.-C., Chouet, M., Gresse, H., Hequet, V., Palard, R., Fromholtz,  
695 M., Espitalier, V., Goëau, H., et al. (2023). Synergizing digital, biological, and participatory sci-  
696 ences for global plant species identification: Enabling access to a worldwide identification service.  
697 *Biodiversity Information Science and Standards*, 7:e112545.
- 698 Bonnet, P., Joly, A., Faton, J.-M., Brown, S., Kimiti, D., Deneu, B., Servajean, M., Affouard, A.,  
699 Lombardo, J.-C., Mary, L., et al. (2020). How citizen scientists contribute to monitor protected  
700 areas thanks to automatic plant identification tools. *Ecological Solutions and Evidence*, 1(2):e12023.
- 701 Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural  
702 networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.
- 703 Botella, C., Deneu, B., Gonzalez, D. M., Servajean, M., Larcher, T., Leblanc, C., Estopinan, J.,  
704 Bonnet, P., and Joly, A. (2023a). Overview of geolifeclef 2023: Species composition prediction with  
705 high spatial resolution at continental scale using remote sensing. *Working Notes of CLEF*.
- 706 Botella, C., Deneu, B., Marcos, D., Servajean, M., Estopinan, J., Larcher, T., Leblanc, C., Bonnet,  
707 P., and Joly, A. (2023b). The geolifeclef 2023 dataset to evaluate plant species distribution models  
708 at high spatial resolution across europe. *arXiv preprint arXiv:2308.05121*.
- 709 Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. (2018). A deep learning approach to  
710 species distribution modelling. *Multimedia Tools and Applications for Environmental & Biodiversity  
711 Informatics*, pages 169–199.
- 712 Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S. M., Chytrý, M., Pillar,  
713 V. D., Jansen, F., Kattge, J., Sandel, B., et al. (2019). *splot*—a new tool for global vegetation  
714 analyses. *Journal of Vegetation Science*, 30(2):161–186.
- 715 Brun, P., Karger, D. N., Zurell, D., Descombes, P., de Witte, L., de Lutio, R., Wegner, J. D., and  
716 Zimmermann, N. E. (2023). Rank-based deep learning from citizen-science data to model plant  
717 communities. *bioRxiv*, pages 2023–05.
- 718 Černá, L. and Chytrý, M. (2005). Supervised classification of plant communities with artificial neural  
719 networks. *Journal of vegetation Science*, 16(4):407–414.
- 720 Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the  
721 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.
- 722 Chytrý, M. et al. (2012). Vegetation of the czech republic: diversity, ecology, history and dynamics.  
723 *Preslia*, 84(3):427–504.
- 724 Chytrý, M., Hennekens, S. M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F., Landucci,  
725 F., Schaminée, J. H., Aćić, S., Agrillo, E., et al. (2016). European vegetation archive (eva): an  
726 integrated database of european vegetation plots. *Applied vegetation science*, 19(1):173–180.
- 727 Chytry, M., Tichy, L., Hennekens, S., Knollova, I., Janssen, J., Rodwell, J., Peterka, T., Marceno, C.,  
728 Landucci, F., Danihelka, J., et al. (2021). Eunis-esy: Expert system for automatic classification of  
729 european vegetation plots to eunis habitats.

- 730 Chytrý, M., Tichý, L., Hennekens, S. M., Knollová, I., Janssen, J. A., Rodwell, J. S., Peterka, T.,  
731 Marcenò, C., Landucci, F., Danihelka, J., et al. (2020). Eunis habitat classification: Expert system,  
732 characteristic species combinations and distribution maps of european habitats. *Applied Vegetation*  
733 *Science*, 23(4):648–675.
- 734 Davies, C. and Moss, D. (1999). Eunis habitat classification. final report to the european topic centre  
735 on nature conservation. *European Environment Agency*, 256.
- 736 De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., Czúcz, B., Dengler,  
737 J., Ewald, J., Faber-Langendoen, D., et al. (2015). A comparative framework for broad-scale plot-  
738 based vegetation classification. *Applied Vegetation Science*, 18(4):543–560.
- 739 Deneu, B., Joly, A., Bonnet, P., Servajean, M., and Munoz, F. (2022). Very high resolution species  
740 distribution modeling based on remote sensing imagery: how to capture fine-grained and large-scale  
741 vegetation ecology with convolutional neural networks? *Frontiers in plant science*, 13:839279.
- 742 Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., and Joly, A. (2021). Convolutional  
743 neural networks improve species distribution modelling by capturing the spatial structure of the  
744 environment. *PLoS computational biology*, 17(4):e1008856.
- 745 Dengler, J., Jansen, F., Glöckler, F., Peet, R. K., De Cáceres, M., Chytrý, M., Ewald, J., Oldeland, J.,  
746 Lopez-Gonzalez, G., Finckh, M., et al. (2011). The global index of vegetation-plot databases (givd):  
747 a new resource for vegetation science. *Journal of Vegetation Science*, 22(4):582–597.
- 748 Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM computing surveys*  
749 *(CSUR)*, 27(3):326–327.
- 750 Estopinan, J., Bonnet, P., Servajean, M., Munoz, F., and Joly, A. (2024). Modelling species distribu-  
751 tions with deep learning to predict plant extinction risk and assess climate change impacts. *arXiv*  
752 *preprint arXiv:2401.05470*.
- 753 Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., and Joly, A. (2022). Deep species distribution  
754 modeling from sentinel-2 image time-series: A global scale analysis on the orchid family. *Frontiers*  
755 *in Plant Science*, 13.
- 756 Evans, D. (2012). The eunis habitats classification—past, present & future. *Revista de investigación*  
757 *marina*, 19(2):28–29.
- 758 Feurer, M. and Hutter, F. (2019). Hyperparameter optimization. *Automated machine learning: Meth-*  
759 *ods, systems, challenges*, pages 3–33.
- 760 Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of*  
761 *statistics*, pages 1189–1232.
- 762 Gammnerman, A., Vovk, V., and Vapnik, V. (2013). Learning by transduction. *arXiv preprint*  
763 *arXiv:1301.7375*.
- 764 Garcin, C., Servajean, M., Joly, A., and Salmon, J. (2022). Stochastic smoothing of the top-k calibrated  
765 hinge loss for deep imbalanced classification. In *International Conference on Machine Learning*,  
766 pages 7208–7222. PMLR.
- 767 Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodolog-*  
768 *ical)*, 14(1):107–114.
- 769 Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. (2021). Revisiting deep learning models  
770 for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943.
- 771 Hall, L. S., Krausman, P. R., and Morrison, M. L. (1997). The habitat concept and a plea for standard  
772 terminology. *Wildlife society bulletin*, pages 173–182.
- 773 Hancock, J. T. and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks.  
774 *Journal of Big Data*, 7(1):1–41.

- 775 Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical*  
776 *learning: data mining, inference, and prediction*, volume 2. Springer.
- 777 Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.
- 778 Hennekens, S. M. and Schaminée, J. H. (2001). Turboveg, a comprehensive data base management  
779 system for vegetation data. *Journal of vegetation science*, 12(4):589–591.
- 780 Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document*  
781 *analysis and recognition*, volume 1, pages 278–282. IEEE.
- 782 Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*,  
783 9(03):90–95.
- 784 Inc, P. T. (2015). Collaborative data science, montreal, qc: Plotly technologies inc.
- 785 Janssen, J., Rodwell, J., Criado, M. G., Gubbay, S., Haynes, T., Nieto, A., Sanders, N., and Calix, M.  
786 (2016). *European red list of habitats*. Publications Office of the European Union Luxembourg.
- 787 Joly, A., Botella, C., Pícek, L., Kahl, S., Goëau, H., Deneu, B., Marcos, D., Estopinan, J., Leblanc,  
788 C., Larcher, T., et al. (2023). Overview of lifeclef 2023: evaluation of ai models for the identification  
789 and prediction of birds, plants, snakes and fungi. In *International Conference of the Cross-Language*  
790 *Evaluation Forum for European Languages*, Springer.
- 791 Joseph, L. N., Field, S. A., Wilcox, C., and Possingham, H. P. (2006). Presence–absence versus  
792 abundance data for monitoring threatened species. *Conservation biology*, 20(6):1679–1687.
- 793 Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. (2021). Well-tuned simple nets excel on tabular  
794 datasets. *Advances in neural information processing systems*, 34:23928–23941.
- 795 Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.
- 796 Leblanc, C., Joly, A., Lorieul, T., Servajean, M., and Bonnet, P. (2022). Species distribution modeling  
797 based on aerial images and environmental features with convolutional neural networks. In *Working*  
798 *Notes of CLEF 2022-Conference and Labs of the Evaluation Forum*, pages 2123–2150.
- 799 Lorieul, T., Joly, A., and Shasha, D. (2021). Classification under ambiguity: When is average-k better  
800 than top-k? *arXiv preprint arXiv:2112.08851*.
- 801 Marcenò, C., Guarino, R., Loidi, J., Herrera, M., Isermann, M., Knollová, I., Tichý, L., Tzonev, R. T.,  
802 Acosta, A. T. R., FitzPatrick, Ú., et al. (2018). Classification of european and mediterranean coastal  
803 dune vegetation. *Applied Vegetation Science*, 21(3):533–559.
- 804 Med, E. (2006). Euro+ med plantbase—the information resource for euro-mediterranean plant diversity.  
805 *October 9 2014*.
- 806 Michalcová, D., Lvončík, S., Chytrý, M., and Hájek, O. (2011). Bias in vegetation databases? a  
807 comparison of stratified-random and preferential sampling. *Journal of Vegetation Science*, 22(2):281–  
808 291.
- 809 Milan, C., Irena, A., Dana, H., Petr, N., Marcela, Ř., Idoia, B., Gianmaria, B., Natálie, Č., Jiří, D.,  
810 Pavel, D., et al. (2022). Floraveg. eu—a new online database of european vegetation and flora. In  
811 *Plant communities in changing environment*. Richard Hrivnák & Michal Slezák.
- 812 Morrison, L. W. (2021). Nonsampling error in vegetation surveys: understanding error types and  
813 recommendations for reducing their occurrence. *Plant Ecology*, 222(5):577–586.
- 814 Moss, D. (2008). Eunis habitat classification—a guide for users. *European Topic Centre on Biological*  
815 *Diversity*.
- 816 Mucina, L., Bültmann, H., Dierßen, K., Theurillat, J.-P., Raus, T., Čarni, A., Šumberová, K., Willner,  
817 W., Dengler, J., García, R. G., et al. (2016). Vegetation of europe: hierarchical floristic classification  
818 system of vascular plant, bryophyte, lichen, and algal communities. *Applied vegetation science*, 19:3–  
819 264.

- 820 Noble, I. (1987). The role of expert systems in vegetation science. *Vegetatio*, 69:115–121.
- 821 Novák, P., Willner, W., Biurrun, I., Gholizadeh, H., Heinken, T., Jandt, U., Kollár, J., Kozhevnikova,  
822 M., Naqinezhad, A., Onyshchenko, V., et al. (2023). Classification of european oak–hornbeam forests  
823 and related vegetation types. *Applied Vegetation Science*, 26(1):e12712.
- 824 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S.,  
825 Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data  
826 with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.
- 827 Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint*  
828 *arXiv:1609.04747*.
- 829 Scherrer, D., Mod, H. K., and Guisan, A. (2020). How to evaluate community predictions without  
830 thresholding? *Methods in Ecology and Evolution*, 11(1):51–63.
- 831 Schmidt, M., Janßen, T., Dressler, S., Hahn, K., Hien, M., Konaté, S., Lykke, A. M., Mahamane,  
832 A., Sambou, B., Sinsin, B., et al. (2012). The west african vegetation database. *Biodiversity and*  
833 *Ecology*, 4:105–110.
- 834 Sietsma, J. and Dow, R. J. (1991). Creating artificial neural networks that generalize. *Neural networks*,  
835 4(1):67–79.
- 836 Stone, M. (1974). Cross-validators choice and assessment of statistical predictions. *Journal of the*  
837 *royal statistical society: Series B (Methodological)*, 36(2):111–133.
- 838 Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep  
839 learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages  
840 13693–13696.
- 841 Szymura, T. H., Kassa, H., Swacha, G., Szymura, M., Zajac, A., and Kacki, Z. (2023). Vegetation  
842 databases augment but do not replace species distribution atlases in species richness assessment.  
843 *Ecological Indicators*, 154:110876.
- 844 Tichý, L., Chytrý, M., and Landucci, F. (2019). Grimp: A machine-learning method for improving  
845 groups of discriminating species in expert systems for vegetation classification. *Journal of Vegetation*  
846 *Science*, 30(1):5–17.
- 847 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polo-  
848 sukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*,  
849 30.
- 850 Walker, K., Pescott, O., Harris, F., Cheffings, C., New, H., Bunch, N., and Roy, D. (2015). Making  
851 plants count. *British Wildlife*, 26(4):243–250.
- 852 Westhoff, V. and Van Der Maarel, E. (1978). *The braun-blanquet approach*. Springer.
- 853 Wisser, S. K., Cáceres, M. d., et al. (2018). New zealand’s plot-based classification of vegetation.  
854 *Phytocoenologia*, 48(2):153–161.
- 855 Yapp, R. H. (1922). The concept of habitat. *Journal of Ecology*, 10(1):1–17.
- 856 Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z.,  
857 et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- 858 Zhongming, Z., Linong, L., Xiaona, Y., Wangqiang, Z., Wei, L., et al. (2015). Linking in situ vegetation  
859 data to the eunis habitat classification: results for forest habitats.

860 **List of Tables**

861 1 Comparison of the top-one (in black) and top-three (in grey) micro average multiclass  
862 accuracy averaged over the ten CV folds for every model and encoding, with and without  
863 noise addition (best top-one result overall with and without noise addition in green  
864 background shading) . . . . . 10

865 **List of Figures**

866 1 Hexagonal binning showing the distribution of vegetation plots from the training dataset.  
867 Zoom in on a specific bin with the raw spatial distribution of the vegetation plots. Fur-  
868 ther breakdown on a vegetation plot (assigned to the habitat type S51, i.e., Mediter-  
869 ranean maquis and arborescent matorral) with the list of co-occurring species. . . . . 6  
870 2 Distribution of vegetation plots in the NPMS test set . . . . . 7  
871 3 Training with ten different random states on the entire EVA training dataset of 886  
872 260 samples (left) and prediction on the NPMS testing dataset of 7 521 samples (right)  
873 time-performance characteristics for selected models, with features encoded with the  
874 reciprocal rank method (without noise addition). The circle size reflects the top-one  
875 micro average multiclass accuracy standard deviation (left) and the size of the model,  
876 i.e., the number of trainable parameters for deep learning algorithms and the number of  
877 estimators (i.e., respectively the number of trees in the forest for RFC and the number  
878 of gradient boosted trees for the XGB) for machine learning algorithms (right). . . . . 11

879 **SUPPORTING INFORMATION**

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

881 **Appendix S1.** Visual overview of the architectures, explanations of the parameters and model  
882 evaluation tables, containing the list of tuned hyperparameters with the search spaces and optimal  
883 values, the list of fixed hyperparameters with the selected values, and the hardware used, time spent,  
884 and result obtained for tuning and optimizing each combination

885 **Appendix S2.** List of all plants species contained in the training dataset from EVA

886 **Appendix S3.** Table listing all habitats from the level three of the EUNIS hierarchy that are present  
887 in the EVA training set or NPMS test set, with their codes, their names, their conservation statuses  
888 based on the European Red List of Habitats and the number of training and testing vegetation plots  
889 assigned to each of them

890 **Appendix S4.** Preprocessing steps to create the two datasets used to evaluate habitat distribution  
891 models and their limitations

892 **Appendix S5.** Comparison and evaluation of the performance of hdm-framework and EUNIS-ESy

893 **Appendix S6.** Understanding how our models reason using interpretability

894 **Appendix S7.** Guide on the classification framework to enhance habitat distribution models