Introduction to deep learning methods for multi-species predictions

Yuqing Hu, Sara Si-Moussi, Wilfried Thuiller

Received: 7 February 2024 | Accepted: 14 October 2024

DOI: 10.1111/2041-210X.14466

Methods in Ecology and Evolution

RESEARCH ARTICLE

Introduction to deep learning methods for multi-species predictions

Yuqing Hu | Sara Si-Moussi | Wilfried Thuiller

Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, LECA, Grenoble, France

Correspondence
Wilfried Thuiller
Email: wilfried.thuiller@univ-grenoble-alpes.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-21-AAFI-0001 and ANR-19-P3IA-0003; HORIZON EUROPE Climate, Energy and Mobility, Grant/Award Number: 101060429 and 101134954; Office Français de la Biodiversité

Handling Editor: Giovanni Strona

Abstract

1. Predicting species distributions and entire communities is crucial for ecologists, to enhance our understanding of the drivers behind species distributions and community assembly and to provide quantitative data for conservation efforts.
2. Popular species distribution models use statistical and machine learning methods but face limitations with multi-species predictions at the community level, hindered by scalability and data imbalance sensitivity. This paper explores the potential of deep learning methods to overcome these challenges and provide more accurate multi-species predictions.
3. Specifically, we introduced four distinct deep learning models that use site × species community data but differ in their internal structure or on the input environmental data structure: (1) a multi-layer perceptron (MLP) model for tabular data (e.g. in-situ/raster climate or soil data), (2) a convolutional neural network (CNN) and (3) a vision transformer (ViT) models tailored for image data (e.g. aerial orthophotographs, satellite imagery), and a multimodal model that integrates both tabular and image data. We also show how adapted loss functions can address imbalance issues.
4. We applied these deep learning models to a plant community dataset comprising 130,582 vegetation surveys encompassing 2522 species located in the French Alps. The tabular environmental data consisted of climate, terrain and soil information, while the images were derived from aerial photographs. All models achieved approximately 70% true skill statistics on hold-out data, demonstrating high predictive capacity for community data, the multimodal model being the best performing one. Additionally, we showcased how interpretability tools can illuminate community structure as seen by deep learning models.
5. Deep learning models offer a broad array of features for predicting entire species communities. They handle imbalance issues and accommodate various data types, from tabular datasets to images, while also being equipped with insightful interpretation tools. The versatility extends to tabular datasets and images, with no clear superiority between the two. The last hidden layers can provide valuable features for modelling other species, and the trained models can be used

to support transfer learning to related tasks. The field of ecology now possesses an additional, potent tool in its arsenal that can foster basic and fundamental research.

KEYWORDS
co-occurrence, deep neural networks, explainable AI, species community, species distribution models

1 | INTRODUCTION

Species distribution models (SDMs), also known as ecological niche models or habitat suitability models (Guisan & Thuiller, 2005), are computational tools used in ecology, biogeography and conservation biology to predict and analyse the spatial distribution of species in a given geographic area. These models are based on the idea that a species' presence or absence (or its abundance) in a particular location is influenced by environmental factors. By analysing these factors and their relationship with species occurrence data, SDMs aim to provide insights into where a species is likely to be found in unobserved conditions (Elith & Leathwick, 2009; Guisan et al., 2017). Beyond single-species predictions, this goal can be extended to encompass communities of multiple species coexisting within the same geographical area. Modelling these communities proves valuable for monitoring regional biodiversity, as these species often share similar or closely related environmental preferences. Furthermore, exploring how these communities respond to specific environmental variables offers insights into their ecological niches, thereby enhancing our understanding of conservation efforts (Pollock et al., 2020).

Traditional SDMs usually rely on established machine learning methods that link species occurrences with environmental factors (Anderson et al., 2011; Araújo & Guisan, 2006; Elith et al., 2006; Franklin, 2010; Guisan & Thuiller, 2005; Guisan & Zimmermann, 2000). Among these methods, popular choices encompass MAXENT (Booth et al., 2014; Kramer-Schadt et al., 2013; Phillips et al., 2006; Phillips & Dudík, 2008), random forest (Bradter et al., 2013; Cutler et al., 2007) and boosted regression trees (De'Ath, 2007; Elith et al., 2008; Moisen et al., 2006). However, despite their effectiveness in single-species predictions, these methods often face challenges when dealing with multiple species. This limitation primarily arises from several reasons. (1) Limited scalability for modelling species-rich communities: Modelling species communities can be immensely complex, involving the simultaneous consideration of thousands of species across hundreds of thousands of sites, particularly across large spatial scales. Few available frameworks are equipped to handle such extensive datasets, and recent implementations, such as joint SDMs, frequently struggle to address this issue effectively. (2) Imbalanced species occurrences: When modelling species communities, a significant challenge lies in the uneven distribution of species occurrences. The well-documented ecological phenomenon of skewed species distributions, where some species are rare while others are very frequent, poses a challenge. Traditional methods tend to prioritize modelling frequent species, neglecting proper representation of rarer ones. (3) The majority of existing methods and frameworks are limited to tabular data formats that encapsulate environmental features, typically in the form of a matrix associating sites with their respective environmental attributes. This constraint overlooks the potential wealth of information held within other data types, such as satellite imagery or airborne data, which could provide valuable insights yet remain unexplored. In essence, while traditional SDMs and joint SDMs have proven effective in various ecological contexts, they encounter significant challenges when applied to species-rich communities, imbalanced species occurrences and the integration of diverse data sources, ultimately highlighting the need for more advanced and adaptable modelling approaches.

Owing to their exceptional performance, deep learning methods (Bengio et al., 2021; LeCun et al., 2015), which include various types of neural networks, have gained widespread recognition across various domains, including computer vision (Krizhevsky et al., 2012a; Redmon et al., 2016), natural language processing (Devlin et al., 2018; Vaswani et al., 2017), gaming (Silver et al., 2016, 2017), audio analysis (Noda et al., 2015; Purwins et al., 2019) and, more recently, even in the field of biology (Jumper et al., 2021; Senior et al., 2020). When coupled with a substantial volume of data, neural networks have shown their ability to efficiently exploit input information. Furthermore, when combined with an appropriately tailored model and a training strategy aligned with the specific task at hand, these methods consistently outperform traditional approaches. Another compelling advantage of deep learning is its adaptability in network design to match the characteristics of the input data. For example, convolutional neural networks (CNNs) (He et al., 2016; LeCun et al., 2010) are particularly well-suited for image data due to their consideration of translation invariance and equivariance, along with parameter efficiency when extracting features from large images. On the contrary, multi-layer perceptron (MLP) models (Amari, 1967; Bengio et al., 2000) excel at processing big tabular data, capable of handling complex non-linear problems effectively. Nevertheless, despite the remarkable achievements of deep learning techniques across diverse domains, their application in ecology, particularly in the context of SDM, remains relatively uncharted territory (Brodrick et al., 2019; Chen et al., 2017). Early endeavours in this domain involved the utilization of neural networks with simplistic architectures featuring a single hidden layer

(Baran et al., 1996; Lek et al., 1996). These models, as implemented in the R package biomod2 (Thuiller et al., 2009), exhibited a susceptibility to underfitting and often yielded suboptimal performance on training data, thereby yielding unreliable predictions (Zhang et al., 2020). More recently, a notable example is the work by (Deneu et al., 2021), which introduced CNNs to predict the likelihood of a specific plant species being present, conditioned on the observation of a plant at a given location. Additionally, (Estopinan et al., 2022) harnessed the temporal dimension of Sentinel-2 data to model the global distribution of orchids using CNNs. These initiatives show a growing interest in the integration of deep learning methodologies within species distribution modelling. Nonetheless, it is imperative to acknowledge that deep learning methods do not offer a universally applicable solution. The choice of the deep learning architecture and the suitability of these methods hinge on the distinctive characteristics of the data and the particular research objectives at hand. Moreover, it is worth noting that deep learning models frequently demand substantial volumes of data, and ecological datasets may be inherently limited in scope, presenting challenges in certain applications.

In this paper, we propose a guided tour into the realm of deep learning methods and investigate their potential for multispecies predictions (Figure 1). We have crafted models tailored to handle two of the most prevalent data structures: tabular data and image data. Tabular data represent the standard format extensively used in species distribution modelling, encompassing various established algorithms such as Maxent, boosted regression trees and generalized linear models. This data format includes environmental information for each sampling point, such as soil pH, mean annual temperature and vegetation type, often collected either in situ or obtained from online databases like CHELSA (Karger et al., 2016, 2020) or SoilGrid (Hengl et al., 2017; Poggio et al., 2021). In contrast, image data typically originate from sources such as satellites or aerial sensors and are characterized by a specific resolution. It can manifest in various forms, including multiple spectral bands (e.g. Sentinel-2 data), LiDAR data, or RGB-infrared images. Given the diverse nature of these two data types, we implemented multispecies models following three well-known deep learning models for our study: A MLP model designed to handle tabular data containing environmental variables; a CNN model and a vision transformer (ViT) model specifically engineered for processing image data. Finally, we also developed a multimodal model to profit out of both the tabular and the image data (Figure 1). To contrast and validate the efficacy of these models, we applied them all on the same case study involving 130582 plant community plots in the French Alps, encompassing a total of 2522 species, while accounting for a significant class imbalance (a 3464: 1 imbalance ratio in our case). Finally, our paper delves into the analysis of the models, employing interpretability tools that enhance our understanding of their underlying characteristics. We specifically conduct a comprehensive comparison between the MLP and the CNN, exploring their respective behaviours, similarities and distinctions. In summary, our paper is guided by the following objectives:

- Introducing fundamental deep learning methods for community-level species distribution modelling, featuring state-of-the-art models tailored to address image and tabular data.
- Demonstrating the performance capabilities of the proposed methods on a relatively large dataset.
- Conducting a thorough comparison among the proposed models, beyond merely assessing their predictive abilities, to discern the underlying learned patterns and identify the key features driving each model's decisions.
- Discussing extensively the challenges and opportunities that deep learning methods are facing in the context of SDM.

Through this work, we aim to shed light on the potential of deep learning methods in the domain of species distribution modelling, providing valuable insights for future research and applications in this field. The subsequent sections provide a structured overview of our approach. In Section 2, we present the multi-species prediction problem in mathematical terms, introduce state-of-the-art deep learning models and review key architectures in (see also Figure 1). In Section 3, we further detail the case study and the corresponding results, while offering insights into our trained models tailored for image and tabular data by applying interpretability tools. This also includes a comparative analysis of the commonalities and distinctions among models, especially between CNN and MLP. Concluding our model exploration, we delve into the prospects for future work and offer our concluding remarks in Sections 4 and 5, respectively.

2 | METHODOLOGY

In this study, we introduce four distinct deep neural network models, each tailored to handle a specific type of input data or having a different internal structure (Figure 1): (1) a MLP model that copes with tabular input data; (2) a CNN model; (3) a ViT model that tackles image input data; and finally (4) a multimodal modal that combines both data structures. In the following, we initially describe the problem setting, then we introduce the general structure and functionality of a neural network, and we present our applied models, while summarizing the characteristics of each model.

2.1 | Problem statement

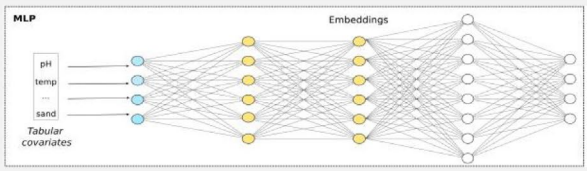
Let's here consider that our data are made of one training dataset and one testing dataset for evaluation. The training and testing datasets can either originate from the same initial dataset, partitioned into two components using a random scheme or more sophisticated strategies (as suggested by Roberts et al., 2017), or they can be independent datasets. In the context of multi-species distribution prediction problems, we are given a training dataset consisting of N species community plots, referred to as $\mathcal{D}_{train} = \{(x_n, y_n)\}_{n=1}^{N}$, and a testing dataset of M species community plots, termed $\mathcal{D}_{test} = \{(x'_m, y'_m)\}_{m=1}^{M}$. Each community plot within the sets $\mathcal{D}_{train}$

Methods in Ecology and Evolution

Mini-review of deep learning architectures

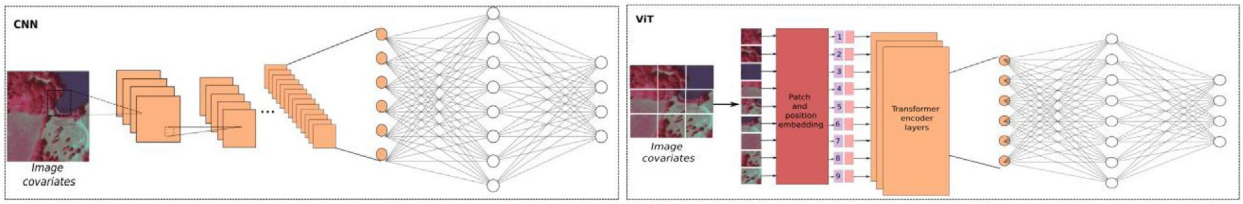
The ability of neural networks to extract complex patterns from large datasets has led to significant advancements in various research and industry domains (Bengio, Courville, & Vincent, 2013). For supervised learning problems (e.g., SDM) a typical deep learning architecture comprises two components: a feature extractor and a classification head. The choice of architecture for the feature extraction depends on the type of input.

Tabular feature extractors



Tabular data are usually processed with multi-layer perceptrons (MLP) but tabular attention models (TabNet (Arik & Pfister, 2021)) and differentiable treebased models (NODE (Popov, Morozov, & Babenko, 2019)) better capture multiple feature interactions.

Image feature extractors



Convolutional neural networks (CNN) have long been the gold standard due to their ability to capture local spatial information efficiently and to learn hierarchical feature representations, while being invariant to rotation and translation. Over the last decade, various developments improved on early CNN architectures (AlexNet (Krizhevsky, Sutskever, & Hinton, 2012b)) by going deeper to learn complex features (VGG (Simonyan & Zisserman, 2014)), learning efficiently by combining features at multiple scales (Inception (Szegedy et al., 2015)) and addressing issues of training stability (e.g. vanishing gradient) in deep networks with residual connections (ResNet (He et al., 2016)). Further developments made use of Neural Architecture Search (NAS (Elsken, Metzen, & Hutter, 2019)) to automate the design of optimal architectures (EfficientNet (Tan & Le, 2019)) with less parameters and a better accuracy that can readily be embedded into mobile devices (MobileNet (Howard et al., 2017)).

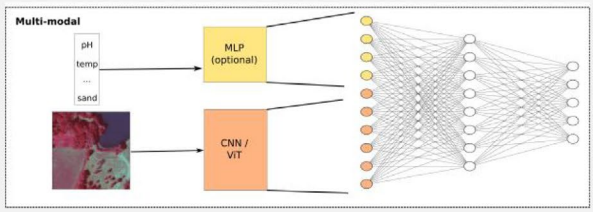
Despite these developments, CNNs are unable to capture long-range dependencies within images (beyond their receptive field size) and can be computationally expensive for very large images.

Vision Transformers (ViT) treat images as sequences of fixed-size patches, each of which is linearly embedded into a vector. By applying self-attention mechanisms to the patch embeddings, ViTs can capture long-range dependencies and relationships between different parts of an image. ViTs (Dosovitskiy et al., 2020) are notoriously hard to train from scratch and are voracious in data. To improve data efficiency while retaining competitive performances, Data-efficient Image Transformers (DeiT (Touvron et al., 2021)) use training tricks (data augmentation, regularization) and knowledge distillation. With the same goal, Shifted Window Transformers (Swin-t (Liu et al., 2021)) use a local rather than a global self-attention mechanism within non-overlapping windows to reduce model complexity. By shifting the local window across layers, Swin-t can learn both local context and global context. Additionally, Swin-t gradually merge image patches to construct a hierarchical feature representation and generate multi-scale embeddings that proved useful in segmentation and object detection tasks (Scheibenreif, Hanna, Mommert, & Borth, 2022).

FIGURE 1 General overview of deep learning model architectures for supervised classification/regression of tabular/image data.

Hybrid architectures such as Deit-C or Convolutional ViTs leverage the complementary strengths of both CNNs (local feature embedding) and ViTs (global dependencies) typically by using CNNs to extract better patch embeddings and ViTs for reasoning about inter-patch dependencies (Yunusa et al., 2024). Innovations in ViTs inspired the development of Next-generation CNNs (ConvNexT, (Liu et al., 2022)) by incorporating ideas from ViTs (e.g. GELU activations), improving training stability with sample-level normalizations (LayerNorm) and using novel components (such as Inverted bottlenecks).

Multi-modal architectures



Multiple feature extractors dedicated to different inputs (i-e multi-modal) can be combined by concatenating their respective output embeddings. These architectures are useful to combine inputs of different types (image, text, tabular, sound, graphs) or different sensors (e.g. sentinel-1 and sentinel-2).

A different approach to multi-modality is adopted in approaches like CLIP (Contrastive Language–Image Pre-training by (Radford et al., 2021)) which learn to map two modalities (e.g. image and text) into a shared embedding space using contrastive training, allowing it to associate and understand the relationship between different modalities across diverse tasks. This self-supervised approach is typically used to pretrain the feature extractors without the need for annotations. Recent literature showed that embeddings produced by such pretrained models, aka foundation models, generalize better to downstream tasks (Klemmer, Rolf, Robinson, Mackey, & Rußwurm, 2023; Vivanco Cepeda, Nayak, & Shah, 2024).

Classification head is a neural network that maps extracted embeddings of the feature extractor(s) and predicts the class probabilities (classification) or the quantities of interest (regression). This component is specific to the modeling problem. Multiple response variables can be modeled/predicted at once (e.g. multiple traits, multiple species or both, etc.) in which case the architecture is referred to as multi-task. A special case is multi-label architectures where multiple binary outputs (e.g. species presence/absence) are jointly modeled and predicted.

FIGURE 1 (Continued)

and $\mathcal{D}_{test}$ is annotated by a list of labels (y_n and y'_m respectively) representing the species observed in the plot amongst the set of K modelled species S. Additionally, each community plot is associated with environmental features in the form of a d-dimensional vector x_n, x'_m and/or an image of dimensions (height h, width w) and a depth of c channels representing the measured signals (for instance RGB images have three channels: red, green, blue).

A deep multi-species distribution model is thus a function $f(x;\theta)$ that predicts the probability of each label (species) given the input x'_m such that f is a neural network with weight parameters θ. We to fit f on the training dataset $\mathcal{D}_{train}$ and evaluate it on the test dataset $\mathcal{D}_{test}$.

2.2 | Neural networks

Neural networks (Goodfellow et al., 2016) serve as the fundamental building blocks for most deep learning methods. Internally, a general L-layered neural network comprises neurons with learnable parameters denoted as $\theta = \mathcal{W}, \mathcal{B}$. These parameters consist of weights $\mathcal{W} = \boldsymbol{W}_{l=1}^{L}$ and biases $\mathcal{B} = \boldsymbol{b}_{l=1}^{L}$ for each layer. Activation functions, such as ReLU (Agarap, 2018) and GeLU (Hendrycks & Gimpel, 2016), introduce non-linearity to neuron outputs, effectively addressing the vanishing gradient issue (Pascanu et al., 2013). The final layer uses activations like softmax (Nwankpa et al., 2018) or sigmoid (Jamel & Khammas, 2012) to map outputs into distributions.

Functionally, a neural network, as previously mentioned, estimates a prediction function $f(\cdot)$ that maps input data (community plots in our case) to desired outputs (observed species distributions). In general, $f(\cdot)$ consists of two essential functions:

$$f(\cdot) = \left(f_e \circ f_c\right)(\cdot), \tag{1}$$

where $f_e(\cdot)$, known as a feature extractor, non-linearly transforms the inputs (images or tabular environmental variables) into new vector representations, termed 'feature vectors' or 'embeddings'. Subsequently, a generalized linear function $f_c(\cdot)$, referred to as a classifier, executes the final prediction, determining species probabilities. Consequently,

a neural network is a synergistic combination of a feature extractor involving multiple transformation layers up to the penultimate layer (layer $L-1$), and a classifier corresponding to the final layer (layer L).

2.2.1 | MLP model

For tabular data inputs in the context of multi-species predictions, we leveraged the MLP, a well-established neural network architecture. MLPs are equipped to handle tabular data efficiently, including the ability to manage large volumes of input data, solve complex nonlinear prediction tasks, facilitate fast learning and maximize the model's capacity to identify feature connections. The feature extractor $f_e(\cdot)$ in MLPs comprises a series of non-linear transformations performed via fully connected layers. In each layer, the transformation is expressed as follows:

$$\boldsymbol{x}_l = \sigma\left(\boldsymbol{W}_{l-1}^T \times \boldsymbol{x}_{l-1} + \boldsymbol{b}_{l-1}\right). \tag{2}$$

This process involves multiplying the input of layer l by the layer's weights, adding the bias to create a linear relationship, and applying a non-linear activation function, denoted as $\sigma(\cdot)$, to obtain the layer's output. This output then becomes the input for the subsequent layer, with this operation repeated for each layer until reaching the penultimate layer.

As for the classifier, denoted as $f_c(\cdot)$, it is characterized as a linear classifier that takes the features extracted from $f_e(\cdot)$ as input and generates scores, often referred to as 'logits', for the targeted labels. Mathematically, the logits, represented as $\boldsymbol{z}$, can be expressed as follows:

$$\boldsymbol{z} = \boldsymbol{x}_L = \boldsymbol{W}_{L-1}^T \times \boldsymbol{x}_{L-1} + \boldsymbol{b}_{L-1}. \tag{3}$$

In our proposed MLP model, we designed a straightforward network architecture comprising two hidden dense layers. The rectified linear unit (ReLU) serves as the activation function in each hidden layer, with batch normalization (Ioffe & Szegedy, 2015) applied before activation to adjust the data distribution. To counteract overfitting, where the model becomes overly tailored to the training data and fails to generalize, a dropout layer (Srivastava et al., 2014) is introduced during training to randomly exclude neurons. Considering the simultaneous modelling of all K species, the final layer of our MLP model comprises K neurons. Consequently, the logits $\boldsymbol{z} = \left[z_1, \ldots, z_k, \ldots z_K\right]$ form a K-dimensional vector, where each component z_k represents the score for species k. These logits are subsequently mapped into species distributions, $p(y|\,\boldsymbol{x};\theta)$, via a sigmoid function:

$$p\left(y = s_k|\,\mathbf{x};\theta\right) = \frac{1}{1+\exp\left(-z_k\right)}, \tag{4}$$

which indicates the probability of the presence of species k given the input $\boldsymbol{x}$. The use of the sigmoid function enables the neural network to predict the presence of each species independently with dedicated neurons, allowing for the prediction of multiple species simultaneously or even no species at all. In contrast, the softmax function, another popular activation function, normalizes the outputs to sum to one, essentially treating the prediction as a 'one-species-vs-all' classification (see Deneu et al., 2021). This makes softmax less flexible for capturing species of all co-occurrence patterns.

2.2.2 | CNN model

While MLP models excel in handling 1D tabular data, they are less suited for 2D image data due to the extensive number of neurons required (typically, one neuron per pixel). Additionally, MLP layers are densely connected, leading to an abundance of parameters that can quickly result in redundancy and overfitting. To address these limitations, we turn to CNNs when dealing with image data. CNNs employ convolution operations, with shared weights applied across the entire image. For vision-related tasks like object classification, this approach offers two significant advantages. First, it enables model equivariance with respect to translation, meaning that detecting a translated object in an image does not necessitate learning entirely new weights. Second, it reduces the number of model parameters that need to be learned. In a CNN, the convolution operation for layer l is expressed as follows:

$$\boldsymbol{x}_l = \sigma\left(\boldsymbol{W}_{l-1}^T * \boldsymbol{x}_{l-1} + \boldsymbol{b}_{l-1}\right), \tag{5}$$

where * signifies a convolution operation. Unlike the multiplication approach outlined in Equation (2), here, the input and weights interact through convolution. For image data, $\boldsymbol{W}_l$ is systematically applied to different positions within the image, resulting in the creation of feature maps after activation. This process is repeated for each convolutional layer.

In our proposed CNN model, we adopt the well-established ResNet50 architecture (He et al., 2016). The feature extractor component includes an initial convolutional layer and a subsequent max-pooling layer (Nagi et al., 2011). This is followed by multiple blocks, each containing three convolutional layers. Finally, another max-pooling layer is used to reduce the feature maps to feature vectors, which are subsequently fed into the classifier component. The classifier portion mirrors the structure used in the MLP model, where the features undergo processing through a linear classifier to produce logits. Similar to Equation (4), we used the sigmoid function for the mapping process to compute probabilities.

2.2.3 | Vision transformer model

Vision transformers (Dosovitskiy et al., 2021) are a novel architecture used in computer vision that makes use of the attention mechanism at the basis of Transformers (Vaswani et al., 2017) that revolutionized the field of Natural Language Processing.

Different from CNNs that apply convolutional operations on image data, a ViT firstly dismounts an image input into several patches

transformed by a ViT encoder into feature vectors, followed by a linear classifier. A ViT encoder usually consists of the following elements: (1) a projector that transforms flattened image patches into new vectors that are more suitable for the model; (2) a positional encoder that adds positional information in each dimension of a vector, providing more contexts for the model to learn; and (3) multiple repeated attention blocks containing self-attention layers (also termed transformer encoder layers) that capture the relationships between vectors. And it is the key component differentiating a ViT with other models. For an image input, such model could potentially better capture the relationships between patches, making predictions based on the common dependency patterns that the model learns.

To test the performance of ViT in our application, we hereby propose to use a basic ViT presented in Figure 1 to predict multiple species. The model reshapes an image into patches of size 16×16, then performs a linear projection and a positional encoder with trainable parameters, followed by 12 repeated attention blocks before entering into the classifier.

2.2.4 | Multimodal model

Multimodal architectures are designed to integrate and interpret information in diverse formats (modalities) to enhance the accuracy and efficiency of deep learning methods.

Here, we propose to construct a multimodal model that takes both tabular and image data as inputs, benefiting from the potential complementarity of environmental features and aerial images in our case study. With regard to the model structure, we concatenate the previously proposed MLP and CNN models in the feature space, followed by a single classifier at the end that performs predictions based on concatenated features, as depicted in Figure 1.

For the case study in this paper, we applied our above models and gave their performance as well as analysis. And in Section 4, we discussed the choice of model architectures in more details.

3 | CASE STUDY

3.1 | Dataset

To test the proposed deep learning models for multi-species predictions, we focused on a plant community dataset collected by the National Alpine Botanical Conservatory (termed CBNA hereafter) over the last 20 years in the French Alps (Poggiato et al., 2023). This plant community data contains a total of 130,582 vegetation relevés (around 10×10 m plots) covering 2522 different plant species with a location uncertainty of at most 100 m. Each species has a minimum of 4 and a maximum of 13,857 observed occurrences. Thus, species prevalences are heavily unbalanced. The dataset includes both image and tabular data, allowing different models to be built and compared with respect to their performances. More details on the CBNA dataset can be found in Appendix S1.

3.2 | Training strategies and evaluation metrics

All of our models are learned following a supervised training scheme (see Appendix S1 for more in-depth details). During training, regularization techniques are applied to prevent the overfitting so that the model can generalize better to new data. In addition, given the heavy imbalance of species occurrences in the dataset, we propose to use the class-balanced (CB) focal loss function (Cui et al., 2019), preventing the model from ignoring rare species. Our own experiments with other lost functions were less positive and we reached out the best performing models with the CB focal loss function (see Sections A2 and A4, Table S1 in Appendix S1). In terms of evaluation, the true skill statistics (TSS) score averaged across species (macro-TSS) or across samples (micro-TSS) (Somodi et al., 2017) was used to evaluate the match between observations and predictions (see Section A3 in Appendix S1 for more details).

Data and code to reproduce the models are available here: https://github.com/yhu01/CBNA/tree/1.0.0.

3.3 | Results

Employing the previously outlined training strategies for all models, we achieved commendable predictive accuracy on the testing dataset: a macro-TSS of around 69% and micro-TSS around 70% (see Table 1). Notably, despite the different types of environmental data used (tabular environmental variables for the MLP and images for the vision models), the macro-TSS scores for these models were similar, with the MLP even outperforming the CNN slightly. However, in the case of micro-TSS, the CNN exhibited slightly better performance at 75.24%, compared with the MLP's 71.41%. The discrepancy between micro- and macro-TSS scores suggests the dominant influence of well-predicted species on the former, whereas the

TABLE 1 True skill statistics (TSS) scores of the proposed deep learning models on the testing plant community dataset.

Model	No. of species	Loss function	Macro-TSS	Micro-TSS
MLP	2522	CB focal loss	69.61%	71.41%
CNN			69.67%	75.24%
ViT			64.26%	66.67%
Multimodal			71.35%	76.87%

Note: All models are trained from scratch using the same loss function, and the threshold in each model was tuned using the validation dataset.

latter is more affected by poorly predicted species. As for the ViT, we obtained results slightly lower than the proposed CNN model. This may be due to the fact there were not enough training data, since a ViT usually requires a large number of inputs to learn well (Touvron, Cord, El-Nouby, et al., 2022). The strong class imbalance may also influence the model's ability to associate semantic features for rare species. Lastly, we can see that by using a multimodal model that fuses MLP and CNN in a situation where both input structures are available brings extra gain on the performance than using only single modalities.

In terms of training time, the proposed vision models required approximately 15 h for training using a single GPU V100. In contrast, the MLP did not necessarily demand a GPU for training and can be learned using a CPU in about 1 h. This makes the MLP a faster and more accessible option, particularly when limited computing resources are available.

We then assessed the models' proficiency (with TSS) in handling species with varying occurrence levels (Figure 2). In general, the models showed commendable results across all species, including rare ones with significantly fewer occurrences, underscoring their effectiveness in addressing data imbalance. Yet, it is noticeable that models for rare species displayed greater variance in their performance. This variance was expected since the limited number of training and testing data can lead to substantial score fluctuations. Additionally, all models exhibit similar performance trends, suggesting that MLP can be as effective as CNN for multi-species predictions and that the multimodal option is the most performing since it somehow takes the best of the two data structures. Yet, this is interesting to note that, in average but also in terms of variance across species, the MLP tends to be slightly better than the CNN for rarer species, a trend that reverses when it comes to more common species (Figure 2). The multimodal model was in average but also in variance better than the MLP and CNN for all categories of occurrence excepted for the rarest one, where the MLP achieved better performances.

3.4 | Interpretability

Explainable AI (xAI) techniques (Molnar, 2022) can help explain black box model predictions at various interpretability levels and with varying degrees of model-specificity.

At the level of the feature extractors, model internals such as feature maps (Qin et al., 2018) and class-activation maps (Chattopadhay et al., 2018; Selvaraju et al., 2017a) for CNNs, or attention masks for ViTs (Chefer et al., 2021) are model-dependent and tailored to specific architectures, thus requiring a deep understanding of the layers and parameters of the architecture.

As depicted in Figure 1, all supervised deep learning architectures share a common intermediate output namely the feature vectors or embeddings which can be visualized to analyse compositional similarities between plots. Given the high dimensionality of feature embeddings, tools such as PCA (Mackiewicz & Ratajczak, 1993), UMAP (McInnes et al., 2018) and t-SNE (Van der Maaten & Hinton, 2008) are designed to reduce the feature dimension effectively, without losing essential information. In addition, further analyses based on clustering is often applied to observe commonalities of the learned features. Techniques applied to the embeddings have the advantage of not being tied to a specific neural model, and do not require a surgical understanding of the architecture.

Finally, at the classifier level, the great wealth of model-agnostic machine learning interpretability tools can be mobilized. This includes tools to measure the relative importance of features: globally (e.g. Permutation importance; Breiman, 2001), their local feature contribution (e.g. LIME; Ribeiro et al., 2016) or the variation

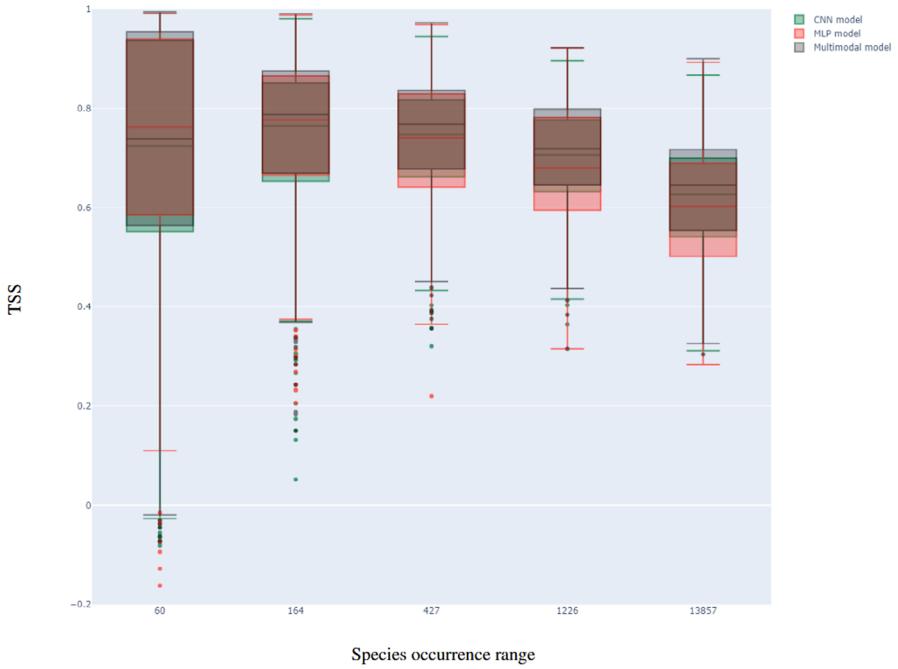


FIGURE 2 Single species true skill statistics (TSS) as a function of species occurrence ranges for multi-layer perceptron (MLP), convolutional neural network (CNN) and the multimodal models (the vision transformer (ViT) is not represented since it presented a lower general performance than the others and for the sake of readability). Two thousand five hundred twenty-two species, divided into five equal parts, each containing 20% of the species, depending on their frequency in the training dataset. The first box illustrates TSS scores for species with occurrences ranging from the minimum (four in our case) to 60. The second box shows performance for species with occurrences ranging from 60 to 164, and so on.

partitioning across all features (e.g. SHAP; Lundberg & Lee, 2017). Response curves that illustrate the output response as a function of input features also fall in this category, including Evaluation Strip (Elith et al., 2005), Partial Dependence Plots (PDP, Friedman, 2001) or Accumulated Local Effects (ALE, Apley & Zhu, 2020).

While these methods are model agnostic, they are often applied to MLP and other traditional machine learning models on tabular data, as they converge more slowly on vision models.

In this section, we used some popular interpretability tools at three different levels (feature extractor, embeddings and classifier), and apply them to our trained models (mainly MLP and CNN, since the ViT showed lower performance in general) to observe their behaviours. Note that these techniques are not restricted to multi-species prediction task but can be broadly deployed to analyse other learning tasks such as object classification, detection and segmentation.

3.4.1 | Inside the feature extractor

For vision models such as CNN or ViT, that are often considered less interpretable due to their use of images, several approaches are available to highlight which features in the feature maps are being utilized, or not. This interpretability can be invaluable for gaining insights into how the landscape or environment influences species distributions or to diagnose potential biases in the data. Here, we provide a demonstration of the CNN model in the context of multi-species predictions. First, we visualized feature maps generated by our CNN. Typically, an input is convoluted with different filters (often referred to as 'kernels' or 'filters') to produce various feature maps. Each feature map corresponds to the filter's response to the input, with each element in the feature map indicating the activation level of a specific neuron within the network. Consequently, activated neurons (those with positive responses) in a feature map correspond to features learned by the filter, while their values signify the presence of these features in the input.

In Figure S3, we provide an example of an input image, which we feed through our trained CNN to extract corresponding feature maps from early, middle and final convolution layers. These feature maps revealed a hierarchical learning process. The feature maps generated in the early layers of the CNN predominantly captured low-level features like edges, boundaries and corners. As we progress through the network's layers, feature maps become capable of learning more complex high-level features, including shapes or even entire regions.

In addition to visualizing feature maps, we also employ the Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2017b) technique to generate class-activation maps for each targeted species. Specifically, an activation map for a species is computed as a weighted sum of feature maps obtained from the final convolution layer in the CNN.

When overlaid on the input image, these activation maps provide insights into the areas in the image that the model identifies as important for predicting the presence of the targeted species. In Figure 3, we provide an example image containing six different species. For each of these species, we display the corresponding activation map (from Figure 3a to f) and indicate the model's prediction for that species in red. These activation maps reveal that different areas of the image are emphasized for each species, demonstrating CNN's ability to predict the presence of multiple species effectively in a multi-species prediction setting.

3.4.2 | Embedding visualization

While our proposed models were able to accurately predict multiple species altogether, understanding the underlying patterns behind these predictions is crucial. This is feasible due to the network's capability to learn generalized features during training, aligning with the principles of Transfer Learning (Pan & Yang, 2009; Torrey & Shavlik, 2010). Additionally, the feature extractor reduces the input dimensionality significantly, streamlining the process while retaining essential input information. To demonstrate this, we computed the image embeddings for the training images D_{train} using CNN's feature extractor, generating a total of N points, each represented as a 2048-dimensional feature vector. We then project these

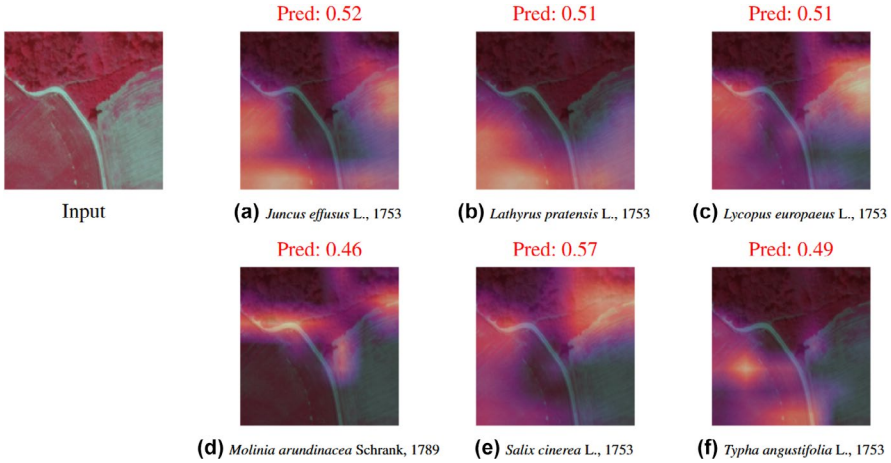


FIGURE 3 Activation maps for targeted species contained in an example input image, and (a–f) the prediction for each targeted species is described on top in red.

high-dimensional feature vectors into a three-dimensional space for visualization purposes using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) (Figure 4a). Unsurprisingly, the 3-D projection unveiled distinct clusters of community plots, each group sharing common image-derived features. This concept may resonate with those experienced in ecological ordinations, somewhat akin to community plot ordinations mapped onto a 3-D space through environmental principal component analysis (PCA). However, the distinguishing factor here is that the primary components emerging from UMAP reflect highly intricate features learned directly from the images, providing a unique perspective on community patterns.

To highlight differences, we employed a *K*-means algorithm (Krishna & Murty, 1999) to cluster the different community plots (Figure 4b), with the number of clusters (8 in our case) chosen using the elbow method (Syakur et al., 2018). Community plots in $\mathcal{D}_{train}$ belonging to the same cluster tend to be explained by the same image features (i.e. landscape characteristics and context) and are likely to share similar species or species with similar functional characteristics.

To further illuminate the insight offered by these clusters, we selected three clusters (designated as Cluster 1, 2, and 3) from Figure 4b. Within these clusters, we present nine representative images, as showcased in Figure 5. It is worth observing that these diverse images captured from the clusters depict typical landscapes found in the European Alps.

Cluster 1, for example, appears closely linked to sparsely vegetated landscapes found in the alpine zone. Meanwhile, Cluster 2 shows densely vegetated forests from the montane zone, and Cluster 3 represents low-elevation and human-dominated sparsely forested regions bordered by roads and cultivated areas. These clusters not only provide insight into distinct landscape types but also help identify the dominant species characterizing these communities. For instance, Table 2 shows the top five dominant species presented in each of the selected cluster, and we note that a species can belong to multiple clusters, a common scenario for closely neighbouring clusters in the 3-D dimension space (Figure 4b), as well as for generalist plant species. In essence, the CNN models communities by grouping images with similar landscapes, recognizing the profound influence of landscape characteristics on species distributions.

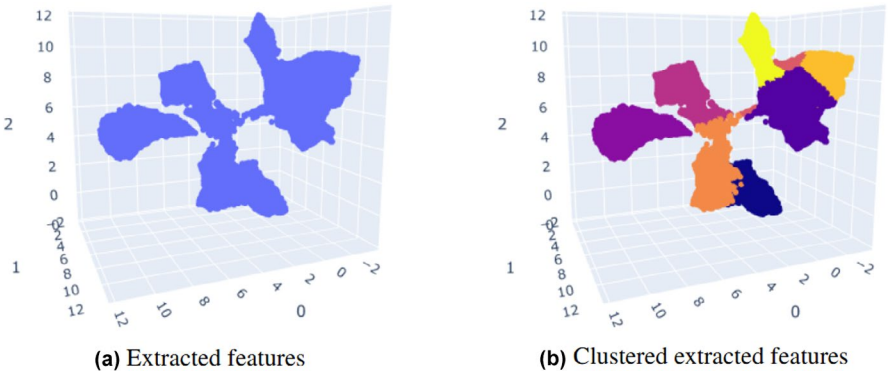
3.4.3 | Classifier-level interpretability

In this section, we delve into understanding the interpretability of a model at the classifier level to gain insights on how it makes the final predictions based on features. The study at this level is often conducted on the MLP model as it gives clearer indications on the influence of input variables leading to the global predictions. Therefore, for our MLP model concerning multispecies distribution predictions based on tabular environmental data, we employed the SHAP (SHapley Additive exPlanations) implementation of Shapley values as introduced by Lundberg et al. (Lundberg & Lee, 2017).

For a given observation and for each variable, the core idea of Shapley values is to compute the change in prediction when including the focal variable(s) compared with when it is excluded i-e randomized, averaged across all the coalitions of the remaining environmental variables. The difference between the model's prediction from the focal observation and its average prediction across the dataset, referred to as payout, is equal to the sum of Shapley values across all variables. This property known as Efficiency coupled with the local to global consistency of Shapley values provides a theoretical foundation for the use of Shapley values to partition the model predictions into the relative contributions of the various environmental predictors for each species (Molnar, 2022).

The SHAP algorithm is an efficient method to estimate Shapley values using a local linear model for each observation. Here, we use the Deep SHAP (Shrikumar et al., 2017) implementation adapted for neural networks.

To match our approach for the CNN model, here we computed the SHAP values for our MLP based on clusters as well. Following the same procedure as for the CNN model, we first used the feature extractor of our trained MLP and obtained 128-dimensional feature vectors for the training data. Then, we applied *K*-means clustering (here we set the cluster number as 8 to compare with the CNN model) to group similar feature vectors based on their Euclidean distances within the feature space. Subsequently, we illustrated the variable importance for two distinct clusters (Figure 6; Figure S4), by aggregating the feature contributions of the five dominant species of each cluster. For the identified cluster in Figure 6, variables such as temperature seasonality (TSeason), precipitation seasonality (PSeason), annual sum of precipitations (PTotY) and the soil



(a) Extracted features **(b)** Clustered extracted features

FIGURE 4 Visualization of features extracted from the feature extractor of the trained convolutional neural network (CNN), projected onto 3-D space using UMAP. Figure (b) is the clustered version of figure (a) using *K*-means algorithm, in which each colour represents a cluster.

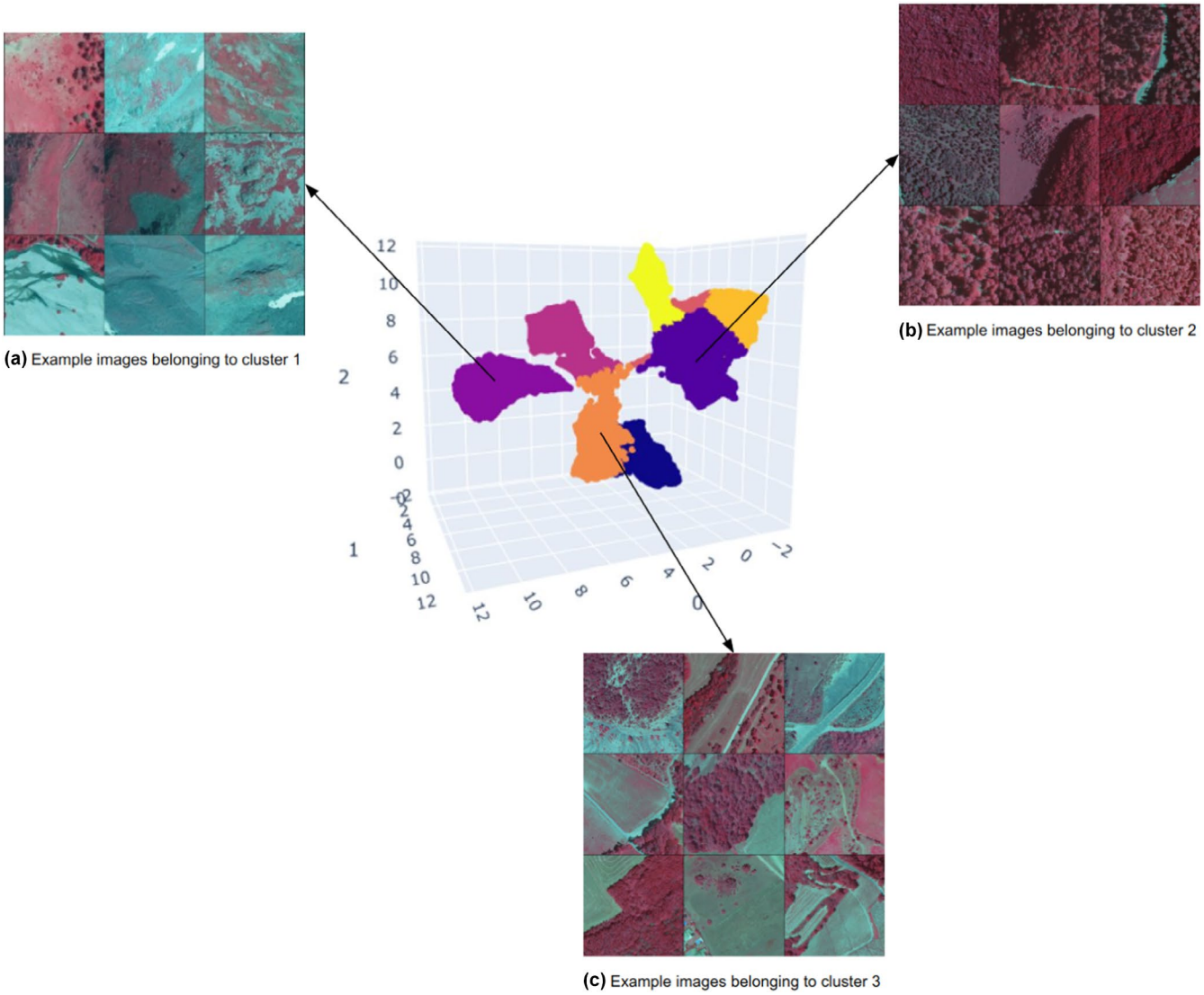


FIGURE 5 (a–c) Example clusters learned by the proposed convolutional neural network (CNN) in the feature space, with some input images selected and shown for each cluster.

TABLE 2 Top five dominant species presented in each of the selected cluster illustrated in Figure 5.

Cluster	Cluster 1	Cluster 2	Cluster 3
Dominant species	Bistorta vivipara (L.) Delarbre, 1800; Festuca violacea Ser. Ex Gaudin, 1808; Plantago alpina L., 1753; Poa alpina L., 1753; Carex sempervirens subsp. sempervirens Vill., 1787	Bromopsis erecta (Huds.) Fourr., 1869; Crataegus monogyna Jacq., 1775; Poterium sanguisorba L., 1753; Quercus pubescens Willd., 1805; Teucrium chamaedrys L., 1753	Aria edulis (Willd.) M.Roem., 1847; Bromopsis erecta (Huds.) Fourr., 1869; Buxus sempervirens L., 1753; Fagus sylvatica L., 1753; Teucrium chamaedrys L., 1753

carbon–nitrogen(CN) ratio significantly influenced the predictions of these species. On the contrary, the cluster depicted in Figure S4 was primarily impacted by mean annual temperature (TMeanY) and temperature seasonality (TSeason). Interestingly, to provide a more visual representation of what these clusters entail, we included sample images alongside the plot. It is important to note that these images are solely for illustrative purposes and were not used in the model, in contrast to the CNN. Yet, we can see that these clusters based on environmental variables only do also represent similar landscape context and structure. This explains why both MLP and CNN, while not based on the same data input and structure, reached similar predictive accuracy. They somehow 'see' the same things. In conclusion, the SHAP analysis based on the cluster illustrates the MLP model's capacity to model communities by giving precedence to various environmental variables, thus enhancing the interpretability of multi-species distribution predictions. As a side note, SHAP is also able to calculate variable importance directly on communities, which might offer very interesting explanatory tools for ecologists.

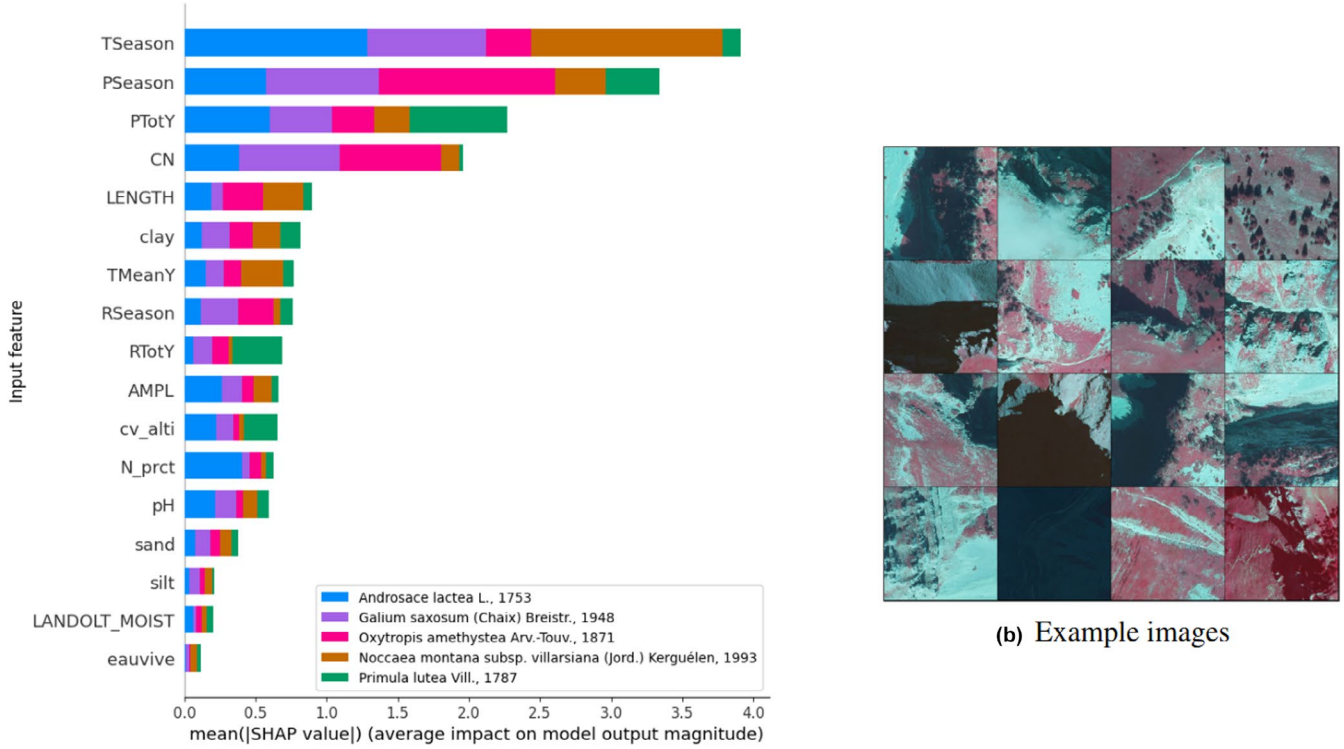


FIGURE 6 Environmental variable contributions of a cluster (a) computed using SHAP for the multi-layer perceptron (MLP) model, with sample images (b) associated with the cluster (which are not the inputs here) illustrated alongside.

In the previous section, we demonstrated how the CNN is adept at grouping inputs that exhibit similarity in terms of visual content in images. This is especially beneficial in our problem, where images depicting similar landscapes are assembled into coherent clusters. However, the approach taken by the MLP differs from this. For instance, in Figure S4b, there are two types of landscapes differentiated by their texture. Images with a cyan texture identify mineral conditions at very high altitudes with few apparent vegetation patches in brown (referred to as high-elevation mineral landscapes), while more brown pictures rather display low productive high-altitude meadows (referred to as high-elevation sparsely vegetated landscapes). Interestingly, our CNN successfully distinguishes between these two types of images and assigns them to separate clusters. In contrast, the MLP groups them into the same cluster. This raises the question of how the MLP assembles species community plots and what criteria it uses to do so. Upon closer inspection, it becomes evident that these images share certain common characteristics, such as barren lands, the absence of water, and sparse vegetation. Importantly, these landscape characteristics are associated with the environmental features in the input data. Thus, in Figure S2, we depict histograms of specific input features (TSeason, CN, and LENGTH in our case) for each type of landscape and compare their similarities using Earth Mover's Distance (EMD, Rubner et al., 2000). The EMD quantifies the optimal distance required to transform one histogram into the shape of the other. Smaller EMD values indicate greater similarity between the two histograms, signifying approximate feature values. From Figure S2, it is evident that the 'cyan' landscape and the 'brown' landscape share similar TSeason, CN and LENGTH value ranges, with standardized EMD values of 0.21, 0.32 and 0.35, respectively. These small EMD values explain why the MLP groups these two landscapes into the same cluster.

4 | DISCUSSION

In this work, we have achieved several significant goals: (1) introducing deep neural network models designed for multi-species predictions and presentation of four distinct models tailored to handle tabular and image inputs, respectively; (2) implementing these models on a real dataset comprising 2522 plant species. Our models achieved commendable performances across both MLP and CNN, the ViT being less performing and the multimodal being the best; (3) demonstrating the models' versatile capabilities, extending mere species predictions to identify and display common landscape and environmental features that explain similar communities; and (4) conducting a thorough comparative analysis mainly between the proposed MLP and CNN on model structure, training time and explainability.

From this, we can highlight in general some similarities and differences among the proposed models, especially between the MLP and the vision models (CNN and ViT) in the context of multi-species predictions.

Similarities:

- Both types of models adhere to the general architecture of a neural network, incorporating a feature extractor responsible for extracting crucial input features and a classifier for executing predictions.
- They effectively tackle data imbalance concerns by implementing the same loss function, leading to great predictive performance. The MLP, handling 1D tabular data, achieves competitive TSS scores compared with the CNN and ViT on 2D image data.
- Both the vision models and MLP exhibit the capacity to learn communities by discerning common features within their respective inputs.
- In terms of performance in our case, both the MLP and vision models exhibit similar trends across species occurrences, with high variances on rare species, and less variant results as species prevalence grow.

Differences:

- Divergent processing of distinct data structures as inputs, with the CNN and ViT adeptly handling image data and the MLP specifically engineered for tabular data.
- Architecturally, feature extractors for the MLP and vision models diverge significantly. While the CNN and ViT both incorporate specifically designed layers for image processing, the MLP integrates multiple fully connected layers.
- The proposed MLP is notably faster to train and can operate without the need for GPUs, whereas the CNN and ViT require GPUs and longer training times.
- Variable approaches in aggregating environmental/image features for modelling species communities. The vision models excel at recognizing distinct landscapes and grouping those with similar textures, while the MLP focuses on identifying similarities in environmental features and grouping inputs with similar values of relevant features.
- Regarding performance, the MLP tends to outperform the vision models on rare species, and the results reverse for more common species. Using direct environmental features known to influence species ecology (e.g. soil chemical properties and temperature) might lead the MLP to better model the complexity of rare species niches.

The presented multimodal model obtained superior results by leveraging the complementarity of the tabular and image modalities. With these findings, we anticipate that ecologists will gain a more nuanced understanding of the intricacies associated with employing deep learning models for modelling multiple species. This, in turn, should encourage a broader embrace of the flexibility offered by these innovative approaches.

There are yet several promising avenues for future research, each capable of enhancing the performance and understanding of these models.

4.1 | Overestimation of species richness

As traditional SDMs and many joint SDMs that tend to produce overly optimistic predictions of species presence, deep learning methods are also challenged with the overestimation of species occurrences within communities. In our case, due to the strong imbalance of our dataset that contains far more species absences than presences, a neural network model, without any weight assignment, tends to bypass the hard training and predict absences all the time. In order to avoid the issue, we assigned present species with a weight (also called an 'importance factor') that is significantly larger than absent species (to a 1000:1 ratio). However, assigning large weights to present species could also over-encourage the model to predict presences, resulting in the overestimation problem. Therefore, solutions are required, for a strongly unbalanced dataset, to find the right trade-off between (1) a model's training ability to not only predict the majority cases (absences), and (2) overestimation of species richness by predicting too many presences. Yet, it should be noted that this overestimation of species richness is likely to arise from other ecological processes that are not explicitly modelled such as priority effects, demography, competitive exclusion and dispersal limitations (Deschamps et al., 2023). Further studies should investigate how to overcome those limitations by for instance constraining overall species predictions by a carrying capacity of a community or modelled species richness or by an explicit consideration of dispersal limitations in the models.

4.2 | Imbalance trade-off

Like many species' community data, our dataset showed a strong imbalance in terms of the species occurrences, with the ratio of common versus rare species up to 3464:1. This is challenging for a neural network model, as it tends to only focus on the common species. Therefore, to address this issue, we adopted a loss function (CB focal loss; Cui et al., 2019) that is imbalance-aware and assigns a weight on each species based on its effective occurrences (Janson, 1986) (see Section A4 in Appendix S1 for comparison with other loss functions). That way, the loss function down-weights common species, making them much less significant compared with rare species. However, besides the above weighting scheme, there are other strategies such as Inverse of Number of Samples (INS) and Inverse of Square Root of Number of Samples (ISNS) (Abdi & Hashemi, 2015; Venkatesan & Er, 2016), or label-distribution aware margin-based losses (Cao et al., 2019) attempting to find the right balance for predictions between common and rare species. Adapting and choosing the ideal loss function is an active field of research, notably to find the optimal weighting scheme to take into account the rare items while keeping a decent performance on the frequent ones (overall performance in other words).

4.3 | Choice of model architectures

We here introduced the four popular neural network architectures that are widely applied in other domains. While they have a similar global structure (feature extractor plus classifier), there are trade-offs on the choice of model architectures. Generally speaking, MLP-based models are preferred for tabular data, as they explore each variable extensively, with fully connected layers, the relationship with other input variables. However, for image data, a MLP model has to counter a large number of pixels, which results in slow computations, and overfitting as the model learns on a pixel level. This would make the model focus too much on the specifics of an image, such that it fails at generalizing to new images. Therefore, CNN and ViT models are often considered for image data, since they both have the ability to learn the spatial relationships between subsets of pixels in an image, reducing computational time and the overfitting, while also capturing essential information. Between CNN and ViT, even though ViT models show competitive, or even superior results on various tasks in other research domains, they have major drawbacks compared to CNNs in that (1) they generally require a large amount of data to be able to stabilize and achieve competitive accuracy; (2) the training of ViT models is difficult and requires sophisticated strategies (more details are discussed in the next sections); and (3) they are slow in terms of inference execution time, as they possess larger model sizes compared to CNNs. Therefore, for applications that require good performance and real-time execution, CNNs are preferred and still much used in real-world applications. In the meantime, multimodal models are gaining momentum since they can take the best from the different data sources. This paper demonstrates the use of two types of data in a single model and has better performance compared with single-modal models. However, challenges of multimodal models exist in terms of fusion, feature alignment, etc. that need to be further studied (Ngiam et al., 2011).

In the domain of species distribution modelling more broadly, there are also other deep learning architectures that deserve experimenting and have started to be tested. For instance, deep learning models were adapted to capture relevant data-generating processes, for example modified to accommodate co-occurrence (Deneu et al., 2019), or to account for species detection probabilities (Joseph, 2020). The choice of models using deep learning can be beyond the above-proposed models, targeting specific goals in the field. We thus hereby encourage more studies on the model structures with a standardized evaluation scheme so that rigorous comparisons between models can be demonstrated.

4.4 | Choice of training schemes

Different strategies can be used to train the feature extraction components of deep learning models. In this paper, given the availability of ground truth annotations (labels), we adopted an end-to-end (1) supervised learning scheme. This scenario is convenient when the amount of data is fairly large such that the model effectively learns representations that best disentangle given labels. However, in real-world scenarios, the acquisition of these ground truths could be quite expensive and time-consuming, and their qualities can be a determinant factor of the model performance.

In cases where only a few data annotations are available, (2) semi-supervised learning (Zhu & Goldberg, 2022) allows training with few-annotated data by making use of existing annotations while exploring the structure of the non-annotated data points. Using data augmentation with generative models (Yang et al., 2022) (e.g. variational autoencoders, generative adversarial networks), label propagation in similarity graph (Wan et al., 2021) or self-training techniques (Amini et al., 2022), and so on semi-supervised methods learn feature embeddings that have better generalization capability to unseen data.

In the extreme case of no annotated samples, (3) self-supervised learning involves training models on unlabeled data by creating surrogate tasks where the model generates its own labels from the input data. Example tasks include rotation prediction, image reconstruction, colorization or contrastive learning. As an example of image reconstruction tasks, Masked AutoEncoders (MAE) (He et al., 2022), involves masking parts of the input image and training the model to reconstruct the missing parts. On the contrary, contrastive techniques like SimCLR (Chen et al., 2020), or MoCo (He et al., 2020) involve generating positive (similar) and negative (dissimilar) pairs by applying deformations on input images, such that the models are trained to maximize agreement between positive pairs and minimize it between negative pairs. Self-distillation, a technique where the model improves itself by learning from its own soft (i-e probabilistic) predictions, showed promising results with ViTs (Caron et al., 2021; Oquab et al., 2023) with applications in remote sensing (Tolan et al., 2024).

At the embedding and classifier level, pretrained models can be used as feature extractors to extract embeddings which are then used as fixed predictors in any classifier/regression model. But one can also leverage (4) transfer learning techniques to transfer the knowledge, learned through large datasets, to new prediction tasks. This is achieved by further training, that is fine-tuning the whole network or a subset of its layers on the new prediction task. Various techniques have been developed for extreme cases with very scarce (few shot) or no data (zero-shot).

In the domain of species distribution modelling, we need to consider difficult scenarios, with no ground truth or few ground truths and develop training methods accordingly in order to learn robust embeddings and facilitate further work.

4.5 | The training of ViT models

As briefly mentioned, ViT models require large datasets as well as well-designed training strategies in order to reach on-par or superior performance compared to CNNs. Therefore, ongoing research has been actively looking for ways to train ViTs more effectively. For instance, in Shiv and Quirk (2019); Wu et al. (2021); Chu et al. (2021),

the authors designed positional encoding techniques to improve awareness of a ViT with image patches. In Touvron, Cord, and Jégou (2022); Touvron, Cord, El-Nouby, et al. (2022), the authors proposed techniques such as layer parallelism and 3-Augment to reduce model redundancy and improve generalization. In addition, (Touvron, Cord, El-Nouby, et al., 2022) also suggested a more effective way using transfer learning, that is to fine-tune only the multi-head attention layers and freeze the feedforward network. This would allow the fixed layers, which dominate the model parameters, to be applied to various tasks. Generally speaking, transfer learning on ViTs is preferred compared to training from scratch. And even more recently, ViTs have been applied with self-supervised training methods (Caron et al., 2021; Chen et al., 2021; Li et al., 2021), with the goal being mainly to learn feature representations using data without labels (i.e. ground truths). These innovative methods would save immense efforts on the data annotation, while achieving state-of-the-art results.

Therefore, while slightly under-performing in the task of multi-species predictions, there are still many potential avenues to explore regarding the ViT models. Importantly in the field of SDMs, ViTs need to be more robust against data imbalance, more efficient in terms of training, and more adaptive to multiple tasks with relatively few data.

4.6 | Data processing

For both classic and deep learning methods in species distribution modelling, data can come from various sources with different qualities and spatial scales. Moreover, data used at large spatial scales often originates from heterogeneous sources with varying geo-location accuracy. Pre-screening input images for data quality (e.g. location uncertainty, low visibility and edge effects) and normalization (centre-scaling or min-max) of both tabular and image features are examples of preprocessing that needs to be tailored to the data type.

Some of these issues can also be addressed within the deep learning training itself. For instance, at the training stage, Data Augmentation can be used to encourage the models to learn representations that are robust to arbitrary noise on the data (e.g. rotation and sensor noise) and at the same time increase the size of the dataset for a cheap price. In addition, some important normalization methods are deeply integrated into the model structure. For example, batch normalization (Ioffe & Szegedy, 2015) handles the scales of an incoming data batch, while layer normalization (Ba et al., 2016) deals with scales in the feature dimension. Oftentimes a neural network would apply such normalizations on a regular basis in each block of its structure, leading to stable training and more persistent performance.

4.7 | Generalization to unobserved species

To broaden the applicability of these models, it is crucial to evaluate their capacity to generalize to new environments and new species. There are several methods to increase the generalization capacity in the field of transfer learning. Recently, methods such as (Hu et al., 2021; Mangla et al., 2020) added self-supervised learning tasks (e.g. predicting the rotations of input images) inside the supervised training scheme to further improve the generalization capacity. In terms of application to unseen data, and by that we mean unseen labels/classes during training, 'zero-shot' strategies rely on the model's generalization capacity without further training. As an example, one case could use our embeddings optimized for vegetation communities to predict bird species in similar environments. However, zero-shot strategies can perform poorly when there is a domain mismatch between the training and inference/prediction dataset. This could happen for instance if we use the CNN trained on alpine landscapes to predict vegetation in urban areas. In this case, it is recommended to fine-tune the models on data for the new tasks. Exploring the application of transfer learning to adapt the trained parameters to new datasets in different ecological domains and spatial data scales, and understanding how domain differences affect model performance is a promising research avenue. As described above, experiments can be conducted in two directions: (1) develop a more sophisticated training procedure to learn better features that are robust to spatial scales, and (2) fine-tune with unseen data using better strategies.

5 | CONCLUSIONS

In this paper, we introduced deep learning methods for multi-species distribution predictions. We proposed neural network models tailored to accommodate two common data structures in species distribution modelling: tabular data and image data. Through experiments conducted on a specific dataset, we showcased the models' capacity to predict multiple species simultaneously, simplifying the task compared with traditional methods that model species individually. Additionally, our models exhibited the ability to handle species imbalance effectively. We also emphasized model interpretability, providing insights into how the models learn discriminative features and identify shared characteristics to predict communities. Notably, we discovered that in our specific case, the MLP on tabular data can be just as effective as the CNN and other models on image data for the prediction task, achieving a similar level of performance.

Deep learning methods have gained popularity across various domains and applications. In the context of species distribution modelling, this work can serve as a foundational reference for researchers interested in applying neural networks to related subjects. Given the perception of deep learning models as 'black boxes', we also encourage further studies on model interpretability, alongside efforts to enhance performance within the field of ecology.

AUTHOR CONTRIBUTIONS

Wilfried Thuiller and Sara Si-Moussi designed the study. Yuqing Hu implemented and ran the models with the help of Sara Si-Moussi.

Sara Si-Moussi prepared the input datasets. Yuqing Hu prepared all the materials and figures for the paper. Wilfried Thuiller and Yuqing Hu wrote the paper, with significant inputs from Sara Si-Moussi.

ACKNOWLEDGEMENTS

We thank Vincent Miele for useful discussions on model results. This work was funded through the European Union's Horizon Europe under grant agreement nos. 101060429 (NaturaConnect) and 101134954 (Obsgession), the ANR FishPredict (ANR21-AAFI-0001) projects and the MIAI@Grenoble Alpes (ANR-19-P3IA-0003) institute. We also acknowledge funding from the French Biodiversity Office through the FloreAlpes and MODIC projects.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14466.

DATA AVAILABILITY STATEMENT

Data are available via the GitHub Repository https://github.com/yhu01/CBNA and the Zenodo Repository https://doi.org/10.5281/zenodo.13983722 (Poggiato et al., 2023).

ORCID

Yuqing Hu https://orcid.org/0000-0002-9093-1356
Sara Si-Moussi https://orcid.org/0000-0002-0519-8699
Wilfried Thuiller https://orcid.org/0000-0002-5388-5274

REFERENCES

Abdi, L., & Hashemi, S. (2015). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering*, *28*(1), 238–251.

Agarap, A. F. (2018). Deep learning using rectified linear units (ReLU). *arXiv preprint*, arXiv:1803.08375. https://doi.org/10.48550/arXiv.1803.08375

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, *16*(3), 299–307.

Amini, M.-R., Feofanov, V., Pauletto, L., Hadjadj, L., Devijver, E., & Maximov, Y. (2022). Self-training: A survey. *arXiv preprint*, arXiv:2202.12040. https://doi.org/10.48550/arXiv.2202.12040

Anderson, R. P., Martínez-Meyer, E., Nakamura, M., Araújo, M. B., Peterson, A. T., Soberón, J., & Pearson, R. G. (2011). *Ecological niches and geographic distributions (MPB-49)*. Princeton University Press.

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *82*(4), 1059–1086.

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688.

Arik, S. Ö., & Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, 6679–6687.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint*, arXiv:1607.06450. https://doi.org/10.48550/arXiv.1607.06450

Baran, P., Lek, S., Delacoste, M., & Belaud, A. (1996). Stochastic models that predict trout population density or biomass on a mesohabitat scale. *Hydrobiologia*, *337*, 1–9.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828.

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, *13*, 932–938.

Bengio, Y., Lecun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, *64*(7), 58–65.

Booth, T. H., Nix, H. A., Busby, J. R., & Hutchinson, M. F. (2014). Bioclim: The first species distribution modelling package, its early applications and relevance to most current maxent studies. *Diversity and Distributions*, *20*(1), 1–9.

Bradter, U., Kunin, W. E., Altringham, J. D., Thom, T. J., & Benton, T. G. (2013). Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution*, *4*(2), 167–174.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brodrick, P. G., Davies, A. B., & Asner, G. P. (2019). Uncovering ecological patterns with convolutional neural networks. *Trends in Ecology & Evolution*, *34*(8), 734–745.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, *32*, 1567–1578.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9650–9660). Computer Vision Foundation.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)* (pp. 839–847).

Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 782–791).

Chen, D., Xue, Y., Fink, D., Chen, S., & Gomes, C. P. (2017). Deep multispecies embedding. In C. Sierra (Ed.), *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI 2017, Melbourne, Australia, August 19–25, 2017* (pp. 3639–3646). IJCAI. https://doi.org/10.24963/IJCAI.2017/509

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607). PMLR.

Chen, X., Xie, S., & He, K. (2021). An empirical study of training self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9640–9649).

Chu, X., Tian, Z., Zhang, B., Wang, X., & Shen, C. (2021). Conditional positional encodings for vision transformers. *arXiv preprint*, arXiv:2102.10882. https://doi.org/10.48550/arXiv.2102.10882

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268–9277).

Cutler, D. R., Edwards, T. C., Jr., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792.

De'Ath, G. (2007). Boosted trees for ecological modeling and prediction. *Ecology*, *88*(1), 243–251.

Deneu, B., Servajean, M., Bonnet, P., Botella, C., Munoz, F., & Joly, A. (2021). Convolutional neural networks improve species distribution

modelling by capturing the spatial structure of the environment. *PLoS Computational Biology*, *17*(4), e1008856.

Deneu, B., Servajean, M., Botella, C., & Joly, A. (2019). Evaluation of deep species distribution models using environment and co-occurrences. *Experimental IR meets multilinguality, multimodality, and interaction: 10th international conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, proceedings 10* (pp. 213–225).

Deschamps, G., Poggiato, G., Brun, P., Galiez, C., & Thuiller, W. (2023). Predict first–assemble later versus assemble first–predict later: Revisiting the dilemma for functional biogeography. *Methods in Ecology and Evolution*, *14*(10), 2680–2696. https://doi.org/10.1111/2041-210X.14203

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *9th international conference on learning representations, ICLR 2021, virtual event, Austria, May 3–7, 2021*. https://openreview.net/forum?id=YicbFdNTTy

Elith, J., Ferrier, S., Huettmann, F., & Leathwick, J. (2005). The evaluation strip: A new and robust method for plotting predicted responses from species distribution models. *Ecological Modelling*, *186*(3), 280–289.

Elith, J., Graham, H. C., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. M., Peterson, A. T., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*, 677–697.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813.

Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., & 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*, 839327.

Franklin, J. (2010). *Mapping species distributions: Spatial inference and prediction*. Cambridge University Press.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: Offering more than simple habitat models. *Ecology Letters*, *8*(9), 993–1009.

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: With applications in r*. Cambridge University Press.

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, *135*(2–3), 147–186.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16000–16009).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729–9738).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint*, arXiv:1606.08415. https://doi.org/10.48550/arXiv.1606.08415

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One*, *12*(2), e0169748.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint*, arXiv:1704.04861. https://doi.org/10.48550/arXiv.1704.04861

Hu, Y., Gripon, V., & Pateux, S. (2021). Leveraging the feature distribution in transfer-based few-shot learning. *International conference on artificial neural networks* (pp. 487–499).

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. Q. (2016). Deep networks with stochastic depth. *Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, part iv 14* (pp. 646–661).

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning* (pp. 448–456).

Jamel, T. M., & Khammas, B. M. (2012). Implementation of a sigmoid activation function for neural network using FPGA. *13th scientific conference of Al-Ma'moon University College* (vol. 13).

Janson, S. (1986). Random coverings in several dimensions. *Acta Mathematica*, *156*, 83–118.

Joseph, M. B. (2020). Neural hierarchical models of ecological populations. *Ecology Letters*, *23*(4), 734–747.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, *596*(7873), 583–589.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2016). *CHELSA climatologies at high resolution for the earth's land surface areas* (version 1.0). World Data Cent.

Karger, D. N., Wilson, A. M., Mahony, C. R., Zimmermann, N. E., & Jetz, W. (2020). Global daily 1 km land surface precipitation based on cloud cover-informed downscaling. *Scientific Data*, *8*, 307.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., & Rußwurm, M. (2023). SatCLIP: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint*, arXiv:2311.17179. https://doi.org/10.48550/arXiv.2311.17179

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013). The importance of correcting for sampling bias in maxent species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379.

Krishna, K., & Murty, M. N. (1999). Genetic *k*-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *29*(3), 433–439.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012a). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012b). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. *25*). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253–256).

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, *90*(1), 39–52.

Li, C., Yang, J., Zhang, P., Gao, M., Xiao, B., Dai, X., Yuan, L., & Gao, J. (2021). Efficient self-supervised vision transformers for representation learning. *arXiv preprint*, arXiv:2106.09785. https://doi.org/10.48550/arXiv.2106.09785

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012–10022).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *arXiv preprint*, arXiv:2201.03545. https://arxiv.org/abs/2201.03545

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Mackiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). *Computers & Geosciences*, *19*(3), 303–342.

Mangla, P., Kumari, N., Sinha, A., Singh, M., Krishnamurthy, B., & Balasubramanian, V. N. (2020). Charting the right manifold: Manifold mixup for few-shot learning. *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2218–2227).

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29), 861. https://doi.org/10.21105/joss.00861

Moisen, G. G., Freeman, E. A., Blackard, J. A., Frescino, T. S., Zimmermann, N. E., & Edwards, T. C., Jr. (2006). Predicting tree species presence and basal area in utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, *199*(2), 176–187.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book

Nagi, J., Ducatelle, F., Di Caro, G. A., Cires¸an, D., Meier, U., Giusti, A., Nagi, F., Schmidhuber, J., & Gambardella, L. M. (2011). Max-pooling convolutional neural networks for vision-based hand gesture recognition. *2011 IEEE international conference on signal and image processing applications (ICSIPA)* (pp. 342–347).

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689–696).

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence*, *42*, 722–737.

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint*, arXiv:1811.03378. https://doi.org/10.48550/arXiv.1811.03378

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint*, arXiv:2304.07193. https://doi.org/10.48550/arXiv.2304.07193

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. *International conference on machine learning* (pp. 1310–1318).

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with maxent: New extensions and a comprehensive evaluation. *Ecography*, *31*(2), 161–175.

Poggiato, G., Gaüzere, P., Martinez-Almoyna, C., Deschamps, G., Renaud, J., Violle, C., Münkemüller, T., & Thuiller, W. (2023). Predicting combinations of community mean traits using joint modelling. *Global Ecology and Biogeography*, *32*(8), 1409–1422.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). Soilgrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *The Soil*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Pollock, L. J., O'Connor, L. M., Mokany, K., Talluto, M. V., & Thuiller, W. (2020). Protecting biodiversity (in all its complexity): New models and method. *Trends in Ecology & Evolution*, *35*(12), 1119–1128.

Popov, S., Morozov, S., & Babenko, A. (2019). Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint*, arXiv:1909.06312. https://doi.org/10.48550/arXiv.1909.06312

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.-Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206–219.

Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). How convolutional neural network see the world—A survey of convolutional neural network visualization methods. *arXiv preprint*, arXiv:1804.11191. https://doi.org/10.48550/arXiv.1804.11191

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13–17, 2016* (pp. 1135–1144).

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., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*, 99–121.

Scheibenreif, L., Hanna, J., Mommert, M., & Borth, D. (2022). Self-supervised vision transformers for land-cover segmentation and classification. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1422–1431).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017a). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision* (pp. 618–626).

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017b). Grad-cam: Visual explanations from deep networks via gradient-based localization. *2017 IEEE international conference on computer vision (ICCV)* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein

structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710.
Shiv, V., & Quirk, C. (2019). Novel positional encodings to enable tree-based transformers. In *Advances in neural information processing systems* (p. 32). ICLR.
Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *International conference on machine learning* (pp. 3145–3153).
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.
Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.
Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556
Somodi, I., Lepesi, N., & Botta-Dukát, Z. (2017). Prevalence dependence in model goodness measures with special emphasis on true skill statistics. *Ecology and Evolution*, *7*(3), 863–872.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.
Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). Integration *k*-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: Materials science and engineering* (vol. 336, p. 012017).
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning* (pp. 6105–6114).
Thuiller, W., Lafourcade, B., Engler, R., & Araújo, M. B. (2009). Biomod–a platform for ensemble forecasting of species distributions. *Ecography*, *32*(3), 369–373.
Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H. V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., Moutakanni, T., Bojanowski, P., Johns, T., White, B., Tiecke, T., & Couprie, C. (2024). Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, *300*, 113888. https://doi.org/10.1016/j.rse.2023.113888
Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques* (pp. 242–264). IGI Global.
Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International conference on machine learning* (pp. 10347–10357).
Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., & Jegou, H. (2022). Three things everyone should know about vision transformers. *arXiv preprint*, arXiv:2203.09795. https://doi.org/10.48550/arXiv.2203.09795
Touvron, H., Cord, M., & Jégou, H. (2022). DeiT III: Revenge of the ViT. *European conference on computer vision* (pp. 516–533).
Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.
Venkatesan, R., & Er, M. J. (2016). A novel progressive learning technique for multi-class classification. *Neurocomputing*, *207*, 310–321.
Vivanco Cepeda, V., Nayak, G. K., & Shah, M. (2024). Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, *36*, 8690–8701.
Wan, S., Pan, S., Yang, J., & Gong, C. (2021). Contrastive and generative graph convolutional networks for graph-based semi-supervised learning. *Proceedings of the AAAI conference on artificial intelligence* (vol. 35, pp. 10049–10057).
Wu, K., Peng, H., Chen, M., Fu, J., & Chao, H. (2021). Rethinking and improving relative position encoding for vision transformer. *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10033–10041).
Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 8934–8954.
Yunusa, H., Qin, S., Chukkol, A. H. A., Yusuf, A. A., Bello, I., & Lawan, A. (2024). Exploring the synergies of hybrid CNNs and ViTs architectures for computer vision: A survey. *arXiv preprint*, arXiv:2402.02941. https://arxiv.org/abs/2402.02941
Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2020). Improving prediction of rare species' distribution from community data. *Scientific Reports*, *10*(1), 12230.
Zhu, X., & Goldberg, A. B. (2022). *Introduction to semi-supervised learning*. Springer Nature.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1: Details on the case study.

How to cite this article: Hu, Y., Si-Moussi, S., & Thuiller, W. (2025). Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution*, *16*, 228–246. https://doi.org/10.1111/2041-210X.14466