

# Machine learning and deep learning—A review for ecologists

Maximilian Pichler | Florian Hartig

Theoretical Ecology, University of Regensburg, Regensburg, Germany

**Correspondence**

Maximilian Pichler  
 Email: [maximilian.pichler@biologie.uni-regensburg.de](mailto:maximilian.pichler@biologie.uni-regensburg.de)

**Funding information**

Bavarian Ministry of Science and the Arts in the Context of Bavarian Climate Research Network

**Handling Editor:** Arthur Porto

## Abstract

1. The popularity of machine learning (ML), deep learning (DL) and artificial intelligence (AI) has risen sharply in recent years. Despite this spike in popularity, the inner workings of ML and DL algorithms are often perceived as opaque, and their relationship to classical data analysis tools remains debated.
2. Although it is often assumed that ML and DL excel primarily at making predictions, ML and DL can also be used for analytical tasks traditionally addressed with statistical models. Moreover, most recent discussions and reviews on ML focus mainly on DL, failing to synthesise the wealth of ML algorithms with different advantages and general principles.
3. Here, we provide a comprehensive overview of the field of ML and DL, starting by summarizing its historical developments, existing algorithm families, differences to traditional statistical tools, and universal ML principles. We then discuss why and when ML and DL models excel at prediction tasks and where they could offer alternatives to traditional statistical methods for inference, highlighting current and emerging applications for ecological problems. Finally, we summarize emerging trends such as scientific and causal ML, explainable AI, and responsible AI that may significantly impact ecological data analysis in the future.
4. We conclude that ML and DL are powerful new tools for predictive modelling and data analysis. The superior performance of ML and DL algorithms compared to statistical models can be explained by their higher flexibility and automatic data-dependent complexity optimization. However, their use for causal inference is still disputed as the focus of ML and DL methods on predictions creates challenges for the interpretation of these models. Nevertheless, we expect ML and DL to become an indispensable tool in ecology and evolution, comparable to other traditional statistical tools.

**KEY WORDS**

artificial intelligence, big data, causal inference, deep learning, machine learning

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Methods in Ecology and Evolution* published by John Wiley & Sons Ltd on behalf of British Ecological Society.

## 1 | INTRODUCTION

In recent years, machine learning (ML), artificial intelligence (AI) and deep learning (DL) have revolutionized almost all areas of science (Jordan & Mitchell, 2015). Early ML algorithms emerged together with the first computers in the '50s, and co-evolved with advances in computing power ever since. During the '90s, the ML field experienced its first bloom, when a wave of fundamental concepts and algorithms such as boosting, bagging, shrinkage estimation and random forest (RF) were discovered. These algorithms challenged, for the first time, the supremacy of classical probability-based statistical models for data analysis and predictions. In the last decade, a second revolution occurred with the rediscovery of deep neural networks, fueled by the availability of graphics processing units (GPUs; 'graphic cards') which made applying these large neural networks practical for the first time. Famous breakthroughs of DL include playing Go (AlphaGo Zero; see Silver et al., 2017), natural language processing (NLP, e.g. GPT-2; see Radford et al., 2019), detecting and identifying objects in images (Mask R-CNN; see He et al., 2017), and predicting protein structures (AlphaFold; see Jumper et al., 2021).

Research in ecology and evolution (E&E) has eagerly adopted both waves of innovation. Several reviews have highlighted the potential of the recent advances in DL (Borowiec et al., 2022; Christin et al., 2019; Tuia et al., 2022; Wäldchen & Mäder, 2018), particularly for processing ecological data such as species recognition from video and audio analysis (Aodha et al., 2018; e.g. Fritzler et al., 2017; Gray et al., 2019; Guirado et al., 2018; Lasseck, 2018; Tabak et al., 2019) or for extracting trait or behavioural information (Dunker et al., 2020; Graving et al., 2019; Mathis et al., 2018; Ott & Lautenschlager, 2021; Pereira et al., 2019). A second area where both traditional ML and DL approaches are already widely used in E&E is predictive modelling. Examples include filling missing links in ecological networks (e.g. Desjardins-Proulx et al., 2017), as part of or in conjunction with traditional mechanistic models (Rammer & Seidl, 2019; Reichstein et al., 2019), for approximating differential equations (Chen et al., 2019; Rackauckas et al., 2021), or for species distribution models (Chen et al., 2018; Elith & Leathwick, 2009; Harris et al., 2018; Wilkinson et al., 2019).

However, despite the rising popularity and attention, the principles and inner workings of ML and DL algorithms are often still perceived as opaque, and their relationship to more classical tools of data analysis, in particular statistical models, remains debated. Trained ML and DL models are often described as a "black box" because their complexity makes it difficult to understand what they have learned. Explainable AI (xAI) methods address this problem and try to understand how trained ML or DL models make predictions (Ribeiro et al., 2016; Ryo et al., 2021). Moreover, a pervasive concern is that ML models are trained for prediction, but the best predictive model does not necessarily correspond to the causal model (Breiman, 2001b; Pearl, 2019, 2021; Box 4). Many researchers thus assume that ML and DL are unable to generate ecological understanding and can only be used as predictive tools (but see

Zhao & Hastie, 2021). This view, however, neglects that there is active research to expand ML and DL methods also to causal inference (Chernozhukov et al., 2018; Schölkopf, 2019; Zhao & Hastie, 2021), which is the classical domain of inferential (causal inference, confirmatory and similar) statistics.

A second reason for confusion about the field is the wealth of algorithms that have been developed in recent years. Most recent reviews on ML have exclusively focused on DL (Borowiec et al., 2022; Christin et al., 2019; Wäldchen & Mäder, 2018). These algorithms differ considerably from simpler, more traditional ML algorithms such as k-nearest-neighbour or boosted regression trees (BRT), and not all statements that are made with respect to DL algorithms apply across the field of ML algorithms in general. For example, image based tasks such as automatic species identification (e.g. Ferreira et al., 2020; Tabak et al., 2019) profit from the use of DL algorithms because they can process spatial patterns better than other ML algorithms (LeCun et al., 2015), whereas traditional ML algorithms often cope better with lower number of observations (e.g. Pichler et al., 2020) or structured (tabular) data (cf. Arik & Pfister, 2020).

Third, a too narrow focus on specific algorithms often prevents researchers from appreciating the general principles that apply across all ML and DL algorithms. For example, the general principles of regularization via shrinkage and model averaging form the backbone of nearly all ML and DL algorithms. Other principles must be relearned when moving from classical ML to DL. For example, the bias-variance tradeoff classically predicts that increasing model complexity reduces systematic model error (bias) at the cost of increasing stochastic error (variance) of the parameters (Box 3). For DL models, however, it was shown that beyond a certain point, variance decreases again with model complexity, thus helping very large networks to achieve low generalization error (Frankle & Carbin, 2019; Huh et al., 2021; Zhang, Wang, et al., 2021). The question of why deep neural networks do not suffer from overparameterization, but instead even depend on it for making accurate predictions is still heavily debated (Sejnowski, 2020), and we will comment on this later.

In the remainder of the review, we will expand on these ideas. Our aim is to provide a comprehensive overview of the principles of ML and DL, starting with the historical development of this field, how ML and DL algorithms differ from traditional statistical tools and how they can be applied for predictive and explanatory modeling (for recent reviews on specific methods, see e.g. DL: Borowiec et al., 2022; Christin et al., 2019; Wäldchen & Mäder, 2018; Computer vision: Lürig et al., 2021 or for specific application areas of ML, see e.g. Tuia et al., 2022 [wildlife management]). After discussing the algorithmic ideas, history and general properties of ML and DL algorithms, we focus on understanding the mechanisms that make ML and DL excel in certain predictive and analytical tasks and how this can be used by ecologists. We also discuss the current limitations of ML and DL and where traditional statistical methods are preferable and highlight current and emerging applications for ecological problems.

## 2 | HISTORY OF ML AND DL AND ITS RELATION TO STATISTICS

### 2.1 | Statistics as the starting point for ML

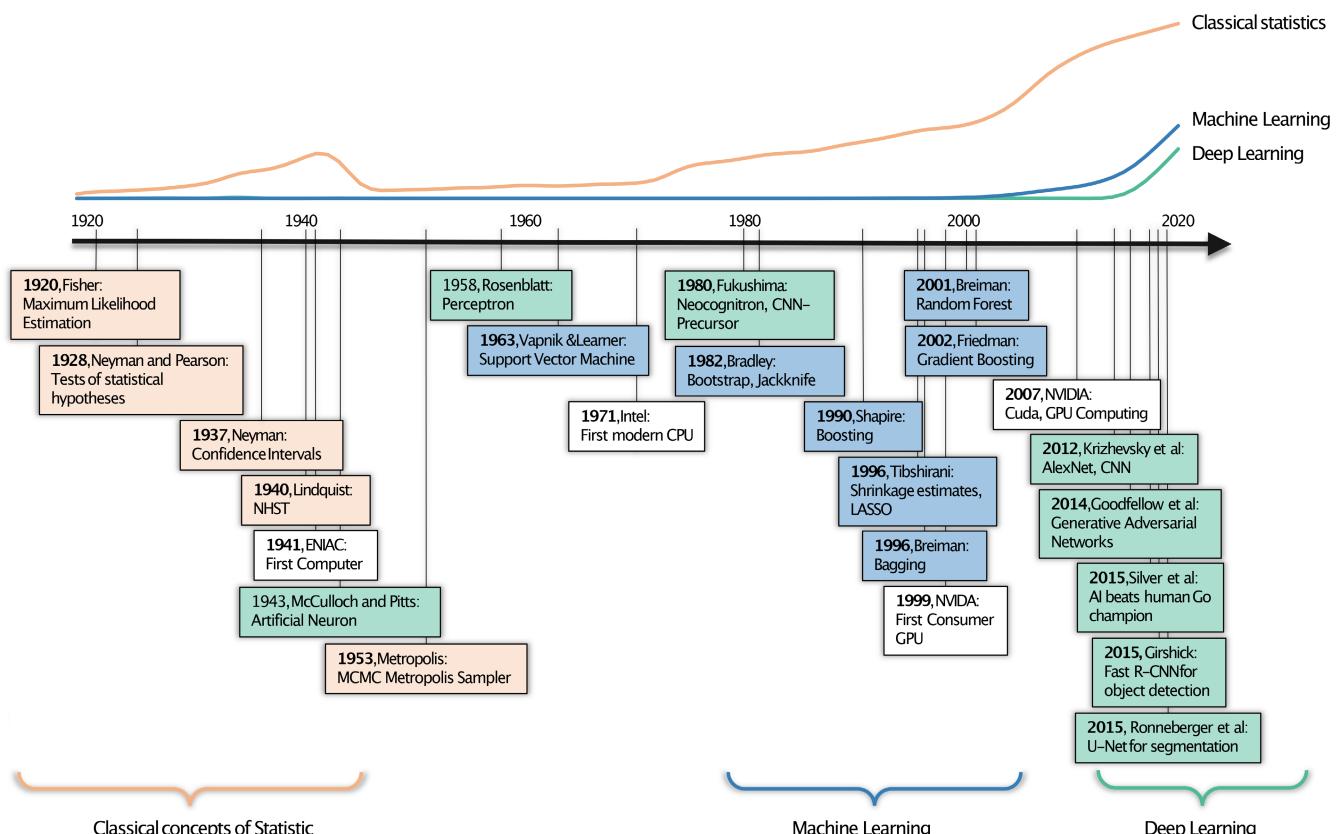
The roots of ML and DL go back a long way, and the development of this field is tightly linked to the development of modern statistics. Apart from Bayesian statistics, many foundational statistical principles such as the maximum likelihood estimation (MLE) or null hypothesis significance testing (NHST) were established in the first half of the 20th century (Figure 1, left). The core of these classical parametric methods is the idealization of a data-generating model, which allows the calculation of the probability of making certain observations, given the model assumptions and parameters. Based on this, eminent statisticians such as Fisher, Neyman and Pearson developed the theory and practice of estimating model parameters with confidence intervals (CI) and calculating *p*-values that has dominated data analysis in E&E to this day (but see Dushoff et al., 2019; Gelman & Loken, 2014; Hartig & Barraquand, 2022; Muff et al., 2022).

Initially, the data-generating model underlying these methods had to be relatively simple to make the calculations of the involved probabilities tractable. The emergence of the first computers (Figure 1), supported by the discovery of new numerical algorithms (e.g.

Markov-chain-Monte-Carlo (MCMC), Metropolis et al., 1953), allowed for a substantial increase in the complexity of parametric statistical models, a development reflected in ecological analyses (Clark, 2005). Even so, when considering the known complexity of the natural world (Grimm et al., 2005), statistical models tend to be rather simple and rigid, due to the mathematical difficulties involved in calculating likelihoods for more complex or flexible models, and it remains an important caveat of traditional statistical methods that the quality of their inference is conditional on those simplified model assumptions (Breiman, 2001b).

### 2.2 | Machine learning

The rising availability of computers around the 1980s allowed not only more refined numerical solutions for classical statistical methods but also the development of alternative modelling approaches for data analysis and predictions that we collectively refer to as “machine learning”. Although these approaches differ in their details, we see their commonality in that they abandoned the idea of a probabilistic data-generating model (associated with the ability to calculate *p*-values, CIs, and all that) in favour of generic algorithmic structures that are trained to perform certain tasks (for general ML principles, see Box 1) with the goal of minimizing a general

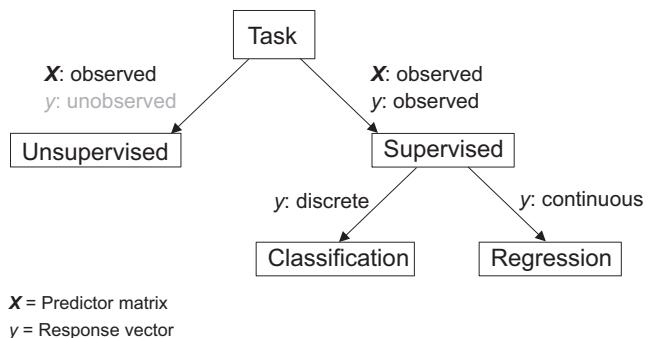


**FIGURE 1** The three eras of statistical learning. The classical concept of statistics, such as the maximum likelihood estimation, null hypothesis significance testing (NHST), or the Markov-chain-Monte-Carlo (MCMC) metropolis sampler were developed in the 1920s–1940s. Common machine learning algorithms or techniques such as boosting, random forest, or the LASSO were discovered between 1980 and the early 2000s. While the theoretical foundation for Deep learning was postulated in the '60s, it has only gained popularity in recent years. The trend lines above the timeline correspond to the frequency of the occurrence of classical statistics (orange), machine learning (blue), and deep learning (green) terms in the scientific literature (see Appendix S1.1 for more details). CNN, convolutional neural network; GPU, graphical processing unit.

## BOX 1 BASICS OF ML

### GENERAL OBJECTIVE OF ML

The objective of ML is to build a good predictive model. By “good”, we mean that the model should predict well for new data. Sometimes ML models make almost no errors on the data they were trained on, but fail for new data (we say the model overfits). A more complex and flexible model has a higher risk of overfitting. The trade-off between complexity and flexibility can be depicted by the bias-variance tradeoff (Box 3). The general ideal of ML algorithms is thus to take a certain algorithmic structure and then adjust their parameters to the data (training), while simultaneously adapting its complexity by optimizing the bias-variance trade-off so that the fitted model generalizes well to new data.



**FIGURE B1** Decision tree to assist in task identification. Given feature matrix  $X$  and a response vector  $y$ , the first decision is to choose between unsupervised (outcome  $y$  is unobserved) and supervised (outcome  $y$  is observed) learning. In the case of supervised learning, if  $y$  is discrete (e.g. species classes), it is a classification task, and if  $y$  is a continuous variable (e.g. biomass), it is a regression task.

### Tasks and learning situations

In ML, the different use cases for the algorithms are called tasks. In supervised learning, examples of the “correct” execution of the task are presented to the algorithm, and the model is trained to minimize the differences between its own actions and the “correct” actions. Common supervised tasks are classification (e.g. labeling of images) and regression (predicting a numerical variable). In contrast to that, unsupervised learning refers to tasks where no examples are supplied, and the algorithms optimize some general loss function (e.g. genomic species delimitation, see Derkarabetian et al., 2019). Finally, in reinforcement learning, the ML algorithm is trained by interacting with a (virtual) environment. Reinforcement learning is used in tasks where the learning depends on executed actions and their produced consequences, for instance, playing strategy computer games such as DOTA (Berner et al., 2019) or Starcraft (Vinyals et al., 2019).

### Model classes and architectures

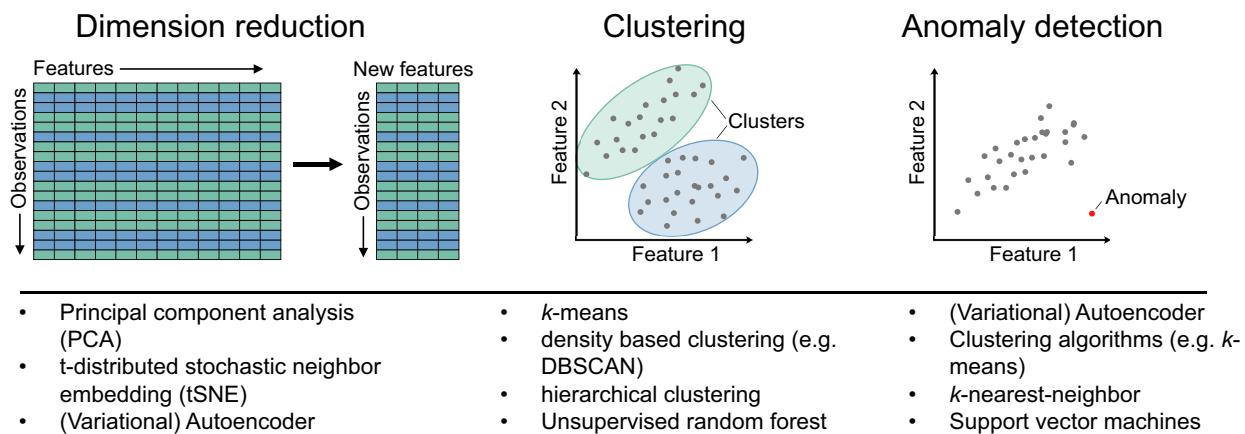
In principle, any algorithm that makes predictions for a given task can be used for ML. In practice, for supervised learning, the most commonly used model classes and architectures can be broadly divided into neural networks, which mimic the functioning of a brain, regression and classification trees, and distance-based method (see Section 3). In unsupervised learning, model classes can be broadly divided into agglomerative hierarchical methods and methods where the number of clusters must be specified a priori (e.g. k-means; Box 2).

### Training the models

In supervised and reinforcement learning, training a model consists of two steps. The first step is to define a loss function that measures the current score (performance) of the algorithm in solving a certain task. The loss function differs for classification and regression tasks (e.g. mean squared error and categorical cross-entropy are common loss functions for regression and classification tasks). The second component is the optimizer, which updates the parameters of the algorithm with the goal to improve its performance. In unsupervised learning a common approach is to use similarities between observations to decide whether or not to group observations together.

## BOX 2 UNSUPERVISED LEARNING IN E&E

Unsupervised learning algorithms (for definitions, see Box 1) also have interesting applications in E&E. In most cases, the goal is to find patterns in the feature space, for example to reduce the dimensionality of the data, to find clusters of similar data, or to detect anomalies (Figure B2).



**FIGURE B2** The three main tasks and their algorithms in unsupervised learning. Dimension reduction techniques reduce the dimensionality of the data by discarding redundant or non-task relevant information. Clustering algorithms try to identify patterns in the data which correspond to mechanistic processes. Anomaly detection is used to identify observations that may have originated from a different data generation process.

Examples of algorithms that perform dimension reductions that are well-known to ecologists include ordination methods such as principal component analysis (PCA) or t-distributed stochastic neighbour embeddings (tSNE), but also DL algorithms such as variational autoencoders. The latter also works on more complex data such as images. The same is true for clustering and anomaly detection tasks: in addition to simple methods such as k-means, which should be familiar to many ecologists, there are now deep-learning methods available that generally have advantages when the data is highly structured, such as in images.

loss function that is not necessarily tied to the probability of the observations (Breiman, 2001b; Shmueli, 2010). Examples of early ML algorithms include neural networks (McCulloch & Pitts, 1943), RF (Breiman, 2001a), and BRT (Friedman, 2001; more on these in the Section 3).

The algorithmic nature of the new ML models lacked the necessary distributional assumptions for calculating p-values and CIs and fueled the development of non-parametric approaches for estimating model uncertainty. A famous example is the bootstrap (Efron, 1992), a resampling technique that is often used for estimating CIs on the parameters and predictions of statistical or ML models. Another example is cross-validation, where a part of the data is used to train the model and the other part of the data is used to evaluate the error (Stone, 1974; see Roberts et al., 2017 for cross-validation strategies for structured ecological data). Since either of these methods require repeated evaluations of the model, their application would be unimageable without computers and even today, they can be computationally challenging for complex models (more on this in Section 4).

## 2.3 | Deep learning

The co-evolution of computational resources and ML algorithms reached a final peak with the emergence of DL algorithms in the last decade. DL algorithms are neural networks (McCulloch & Pitts, 1943) that differ from classical artificial neural networks (ANNs) mainly by their size. While many algorithms and network architectures that are used today were already described in the '80s and '90s (e.g. Fukushima, 1980; Lecun et al., 1998), their practical application was prevented by the lack of computing power at the time. This changed with the emergence of GPUs in the '90s (Figure 1). Although GPUs were originally developed for computer games or other graphical rendering tasks, it was quickly realized that they are often far more efficient than CPUs for certain numerical and linear algebra tasks. Krizhevsky et al. (2017) ushered in the new era of DL when they demonstrated that their competition-winning neural network could be trained on a GPU within hours instead of days or weeks on a CPU. Today, large DL models trained on GPUs with hundreds of millions parameters dominate the

competition for many complex ML tasks, and their behaviour is often markedly different from that of simple ML algorithms (see Section 4).

### 3 | IMPORTANT ML AND DL ALGORITHMS IN MORE DETAIL

Considering that ML branched off from classical statistical models with the goal of increasing model flexibility and complexity while at the same time abandoning the idea of a probabilistic model, it seems obvious to discuss the advantages and disadvantages of this decision. We will do so in Section 4.

Before that, however, it will be useful to explain the most important ML algorithm in more detail. In the main text, we focus on algorithms for supervised learning (see Box 1 for definitions of ML tasks) but we also provide a short overview about unsupervised learning in Box 2. Note that classical statistical models such as linear and logistic regression models can also be used for supervised regression and binary classification tasks, respectively. Arguably, they provide a baseline that ML models should be able to beat. However, because we assume that ecologists are aware of these models, and because our very aim here is to understand why ML algorithms can beat these models, we do not describe them in this section. R, Python, and Julia code examples for all ML and DL algorithms that are discussed (Table 1) are available in Appendix S1.2 or at <https://maximilianpi.github.io/Pichler-and-Hartig-2022>.

#### 3.1 | Support vector machines

A support-vector machine (SVM) is a binary classifier (which can be extended to multiclass and regression tasks, Table 1) that separates the available classes by a hyperplane in the feature (predictor) space (see SVM in Table 1). A predecessor of the SVM, the generalized portrait algorithm, was proposed already by Vapnik and Lerner (1963). The generalized portrait algorithm was computationally cheap (which was important at the time), but as the perceptron, the initial predecessor of ANNs, it was unable to solve non-linear tasks. Boser et al. (1992) overcame this obstacle by using a non-linear feature space transformation (the kernel-trick) to make the task linearly separable (the modern form of the generalized portrait algorithm, the SVM). Because of their computational efficiency for dealing with high dimensional data and relatively low data requirements (compared to DL), SVMs were the most common method for image classification in E&E, particularly in remote sensing (e.g. Gualtieri & Cromp, 1999; Melgani & Bruzzone, 2004; Mountrakis et al., 2011), prior to the success of DL.

#### 3.2 | Ensemble models

Apart from SVMs, ensemble models are the other central ML paradigm that emerged in the 90s: Schapire (1990) showed that ensembles of weak learners (typically simple models such as linear regression models or classification and regression trees, see Friedman, 2001)

often have a low prediction error when their predictions are averaged, even if each individual model has large prediction errors. This principle of generating ensembles of “weak learners” gave rise to two prominent ML techniques: boosting and bagging.

Boosting is an ensemble modelling approach in which weak models are trained sequentially, either by training the next model to correct the errors of the previous model (high-weighting of misclassified observations, AdaBoost, see Freund & Schapire, 1997), or by sequentially optimizing a general differentiable objective (cost) function, gradient boosting (Friedman, 2001) with the latter being the state-of-the-art today (e.g. BRT for species distribution models, see Elith et al., 2008; Elith & Leathwick, 2009; Table 1).

In bagging (bootstrap aggregation), an ensemble of independent weak models is created by training models on bootstrap samples (Breiman, 1996). A famous representative is the RF algorithm which additionally subsamples the features in each node of the decision trees (Table 1; Breiman, 2001a).

Ensemble models are based on an important ML principle that is still valid today: simple algorithms or statistical models can be transformed into more complex algorithms by creating ensembles, which are more difficult to interpret, but often have low prediction errors (see Section 4). BRT and RF are still widely applied, mostly for structured tabular data (Table 1), also because they cope better with smaller datasets than comparable DL models. Examples of recent applications of ensemble models in E&E include predictions in ecological networks (Pichler et al., 2020), linking gene variation to phenotypes (Brieuc et al., 2018), species distribution models (Elith & Leathwick, 2009), and various applications in remote sensing (Belgium & Drăguț, 2016).

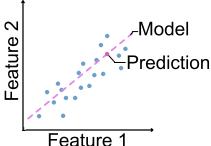
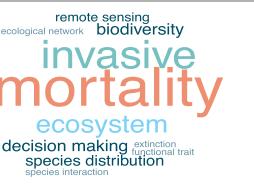
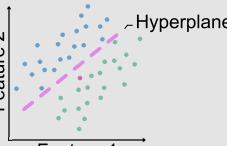
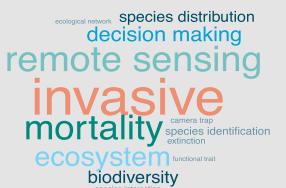
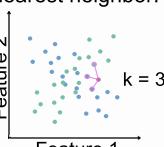
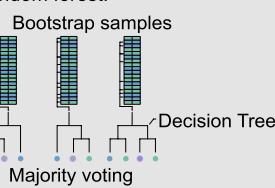
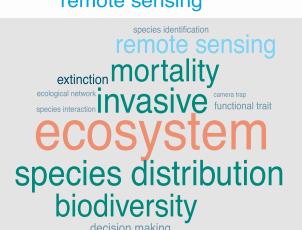
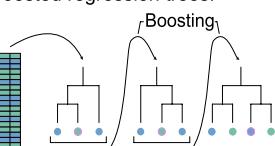
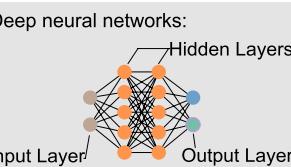
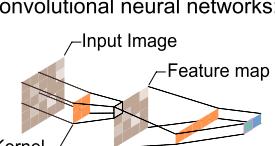
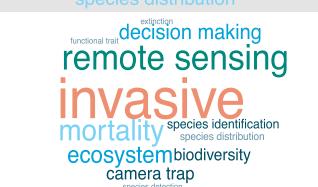
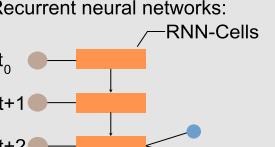
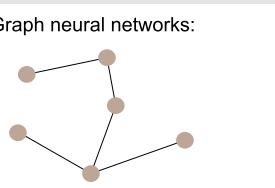
#### 3.3 | Neural networks

Artificial neural network (ANN), inspired by the architecture of our brains, are arguably the most iconic ML architecture, reflecting the long-held dream of building intelligence into a computer. The first fully functional ANN was described by Rosenblatt (1958). This “perceptron algorithm” was a binary classifier that connected the input neurons (one for each input variable = feature) to an output neuron (response). If the signal in the output crossed a certain threshold (activation function), the predicted class changed (e.g. from ‘0’ to ‘1’). However, because of its limited flexibility and particularly its inability to represent nonlinear relationships, the perceptron fell into oblivion for many years until it was discovered that additional layers between the input and output neurons (so-called ‘hidden’ layers) made it possible to approximate any functional form (see subsection 3.4). The potential of ANNs for ecological applications was recognized early (Foody, 1995; French & Recknagel, 1970; Simpson et al., 1992), although to date they have largely been replaced in E&E by the more advanced Deep Neural Networks.

#### 3.4 | Deep learning

Deep learning models represent the latest methodological advance in ML (Figure 2). DL algorithms are neural networks which differ

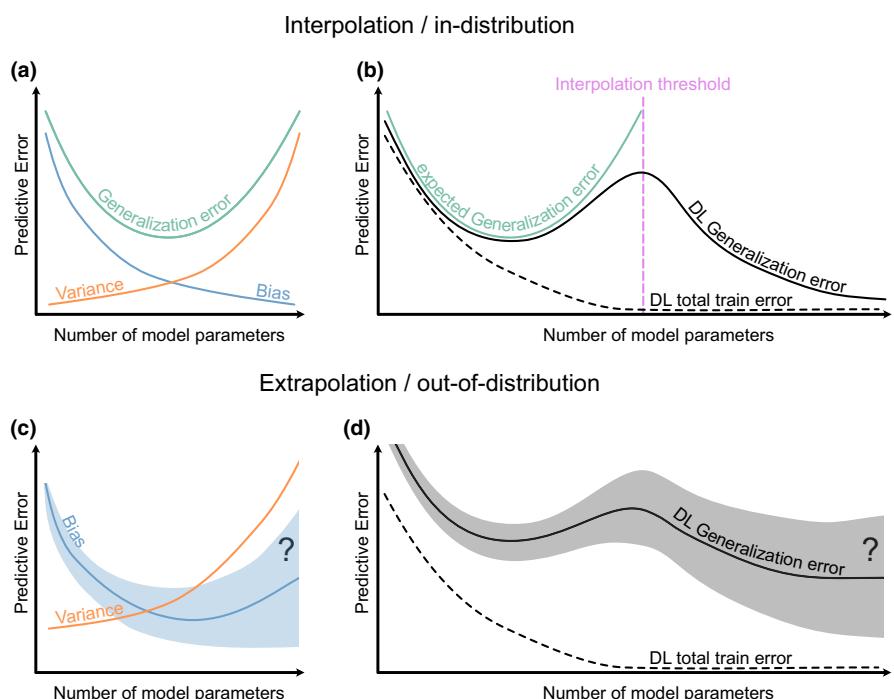
**TABLE 1** Overview of common supervised machine learning algorithms and their most common application areas. Word clouds were created by searching abstracts and titles in the ecology and evolution literature within the specific machine learning algorithms for ecological keywords, the size of the words corresponds to their frequency (see Appendix S1.1).

Machine learning algorithms	Description	Data Type	Application areas
LASSO, Ridge regression: 	Regression models with regularized coefficients (Appendix S1.2.1): + highly interpretable + few observations - limited flexibility	Tabular data: - Classification - Regression	
Support vector machines: 	Hyperplane is optimized to separate response classes (Appendix S1.2.2): + fast and memory efficient + high dimensional data - kernel dependent - no probabilities	Tabular data: - Classification - Regression	
k-nearest neighbor: 	k nearest neighbors in feature space decide response (e.g. by majority voting) (Appendix S1.2.3): + simple + no training - scales poorly - high dimensional data	Tabular data: - Classification - Regression	
Random forest: 	N decision (regression) trees are fitted on bootstrap samples. Split variable is selected from random subset of variables (Appendix S1.2.4): + flexible + robust (e.g. outliers) + few hyper-parameters (+ variable importance - scales poorly	Tabular data: - Classification - Regression	
Boosted regression trees: 	N trees are fitted sequentially to minimize an overall loss function (Appendix S1.2.5): + flexible (+ variable importance - many hyper-parameters - high complexity	Tabular data: - Classification - Regression	
Deep neural networks: 	Input (features) are passed through many hidden layers. Last layer maps into response space (Appendix S1.2.6): + flexible + adaptive to different tasks - many hyper-parameters - computationally expensive	Tabular data: - Classification - Regression	
Convolutional neural networks: 	Small kernels (filters) processes images before passing it to fully connected layers (Appendix S1.2.7): + flexible + detecting shapes and edges - many hyper-parameters - computationally expensive	Images: - Classification - Object detection	
Recurrent neural networks: 	RNN-Cells (e.g. Long short term memory cells) process input sequences and hidden states are recycled (Appendix S1.2.8): + flexible - long-term dependencies - many hyper-parameters - computationally expensive	Sequences (e.g. temporal): - Classification - Regression	
Graph neural networks: 	GNN operate directly on the nodes and their edges. Possible tasks are node or edge classification (Appendix S1.2.9): + flexible + non-Euclidean data - many hyper-parameters - computationally expensive - complex data type	Graphs: - Classification - Regression	

### BOX 3 GENERALIZATION ERROR AND THE BIAS-VARIANCE TRADEOFF

By making models more complex, one can make the prediction error on the training data (*in-sample error*) arbitrarily small. What we really care about, however, is the ability of a model to predict new (out-of-sample data). The discrepancy between model predictions and observations on independent data (e.g. generated by an appropriate cross-validation, see Roberts et al., 2017) is called the *generalization error*.

When minimizing the generalization error, there exists a fundamental tradeoff between variance (parameter uncertainty) and bias. More complex models have higher variance, but also lower bias because (Figure B3a). The two counteracting errors usually lead to a sweet spot of the generalization error at intermediate model complexity. Interestingly, DL models show a double sloping curve of generalization error (Figure B3b), suggesting that the variance of deep neural networks does not increase for very wide and deep networks. The reasons for this are still being discussed in the literature.



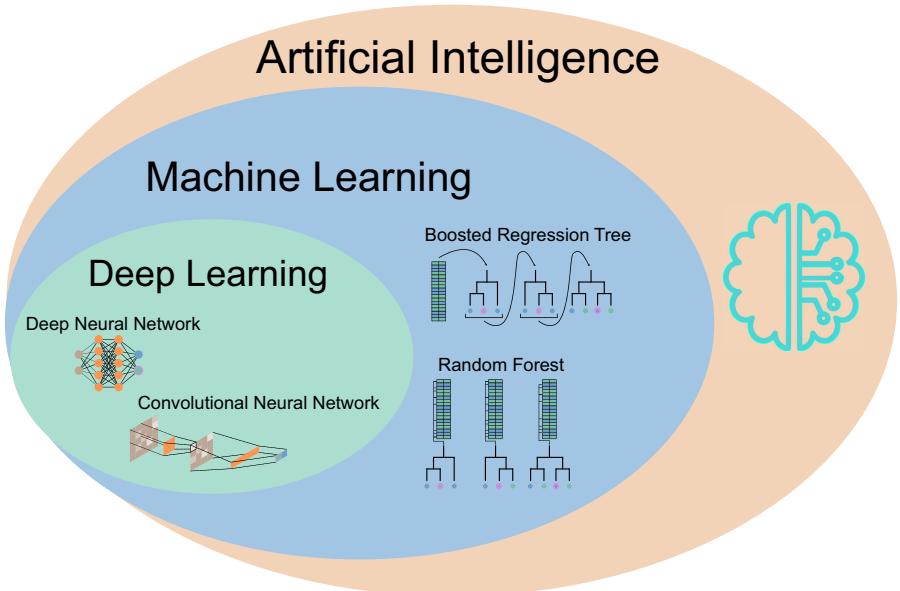
**FIGURE B3** Typical bias-variance tradeoffs in classical machine learning (left panels) and deep learning (right panels) models for interpolation (in-distribution) and extrapolation (out-of-distribution) tasks. In contrast to the classical bias-variance trade-off in panel (a), the bias-variance trade-off for DL in panel (b) shows that after the interpolation threshold (pink dotted line) the training loss is constant (i.e. bias is not improved by increasing model complexity), but the test loss (and thus variance) can be still reduced by increasing the model size. Note also that the total generalization error in extrapolation tasks (panels c and d) is usually higher and the optimal model complexity lower, as the bias will not go to zero with increasing model complexity (depending on the similarity between training and test data).

*Overfitting/Underfitting* describes a situation where the generalization error is higher than necessary or expected. In interpolation (in-distribution, Figure B3a,b) tasks, overfitting/underfitting is usually associated with too high/low model complexity, which leads to a poor compromise between bias and variance. In extrapolation (out-of-sample, see Figure B3c,d) tasks, the reasons for overfitting are often more rooted in bias problems, meaning that the patterns learned in the training data do not generalize to the test data (see, e.g. Yang et al., 2020 for an example in vision tasks).

from simple ANN in the larger number of hidden layers (Borowiec et al., 2022; LeCun et al., 2015) and the often more complicated connection between the neurons (=architecture). Complex task-specific architectures, often with millions of parameters and specific

structures, evolved over the years (for example residual neural networks, He et al., 2016, see also Table 1).

Although DL is based on the same ideas and principles as all other ML algorithms, it is commonly treated as a new field because



**FIGURE 2** Relationship between artificial intelligence (AI), machine learning (ML), and deep learning (DL). AI refers to algorithms that are capable of achieving similar to human-like performance in specific decision or recognition tasks. This is sometimes contrasted with the pursuit of Artificial general intelligence, which refers to AI algorithms that can perform a wide range of tasks and may display human-like abilities in cognitive tasks such as reasoning, logic or common sense. ML algorithms serve as a tool for AI systems to learn from data and make a decision based on data. There are many different ML algorithms such as boosted regression trees or random forest. Within ML, a family of ML algorithms based on artificial neural networks emerged in recent years. Due to their similarities in the way they work and their backbone, DL is considered as a family of its own.

of its distinct principles (see Section 4) and the task-specific architectures that do not resemble traditional ML models. For example image-based tasks (e.g. species identification, see Borowiec et al., 2022) are typically handled by convolutional neural networks (CNNs), a special architecture that uses kernels (convolution matrices) to detect certain shapes and which is used, for example, in identifying species (Ferreira et al., 2020), automatic monitoring of species (Norouzzadeh et al., 2018; Tuia et al., 2022), or landscape classification (Stupariu et al., 2022). Recurrent neural networks (RNNs) are another architecture that is applied for time series tasks (Table 1; Christin et al., 2019; LeCun et al., 2015). In ecology, for example, RNNs have been used to predict population dynamics (Joseph, 2020) or animal movements (Rew et al., 2019; see Borowiec et al., 2022 for more details on different DL algorithms). DL algorithms have also been used to synthesize taxonomic information from literature (Le Guillarme & Thuiller, 2022) or to predict species interactions in ecological networks (Strydom et al., 2021), or to predict species distributions (Deneu et al., 2021). In the following, we treat DL as a subfield of ML and only mention DL when relevant differences to classical ML algorithms are involved.

#### 4 | WHY DOES ML WORK?

When considering current DL algorithms with millions of parameters, researchers trained in classical statistics often struggle to understand why they should work at all. A statistical model with a similar number of parameters could likely not even be fit (e.g. in a linear regression

model, if the number of parameters is greater than the number of observations, there are no degrees of freedom and the equation system is underdetermined). And even if it were possible to fit the model, the bias-variance tradeoff that is fundamental to both statistics and ML (Box 3; Figure B3a) predicts that the optimal compromise between systematic model error (bias) and error due to variance (parameter uncertainty) is at intermediate model complexity (Boxes 3 and 4). Excessively large models should therefore overfit the data and generalize badly.

Despite that, the practical experience shows that ML models converge and generalize well to new data, suggesting that they do not overfit (at least for in-distribution predictions; extrapolation beyond the data domain is as challenging for ML as for other approaches). Even more surprisingly, it was observed that for deep neural networks, the bias-variance trade-off actually reverts, following a double-descent curve (Box 3; Figure B3b; see Belkin et al., 2019; Nakkiran et al., 2019): beyond a certain model size, making deep neural networks even more deeper and wider decreases generalization error, suggesting that model size, contrary to our general expectation, can actually be beneficial for reducing total prediction error (Arora, Du, et al., 2019; Huh et al., 2021; Nakkiran et al., 2019; Novak et al., 2019; Shwartz-Ziv & Alemi, 2020).

The reason for this superficially perplexing behaviour is that practically all ML approaches, despite formally having a very high number of parameters and the associated ability to model complex input-output relationships, perform implicit complexity adjustments that limit their flexibility and avoid overfitting. As a result, especially for DL models, the number of parameters is a poor measure of effective model complexity (Birdal et al., 2021), which is confirmed by more appropriate

complexity measurements (Box 3; Figure B3b; see Birdal et al., 2021; Nakkiran et al., 2019). We divide the underlying mechanisms that adjust model complexity in ML algorithms into two categories: internal (algorithmic) and external (optimization) based approaches.

#### 4.1 | Internal (algorithmic) complexity optimization

By internal (algorithmic) complexity optimization, we understand algorithmic structures that lead to a self-adaptation of model complexity.

One basic mechanism for generating this behaviour is that many ML algorithms implicitly or explicitly generate ensemble predictions. An ensemble model may formally include many parameters, but its effective complexity is by no means the sum of the complexity of each ensemble member. Rather, the complexity of an ensemble model is typically related to the average complexity of its ensemble members and the differences between them, which in turn affect error and variance of the ensemble estimator (Bernardo & Smith, 2009; Dietterich, 2000; Dormann et al., 2018; Ganaie et al., 2021).

Because the ensemble members are fit to the data, the data can influence the difference between the ensemble members and thus the complexity of the entire ensemble estimator. To support this behaviour, many ML algorithms include (tunable) mechanisms to increase heterogeneity in the ensemble. For example, bagging decreases the similarity between ensemble members by bootstrapping the data (Sagi & Rokach, 2018). RF goes one step further by using a random subset of the features in each node, which further diversifies and decorrelates the individual models (Breiman, 2001a) and reduces the variance of the ensemble (Breiman, 2001a). In gradient boosting (see Friedman, 2001), the subsequently trained models depend on the previous model but they are uncorrelated because the following members are forced to compensate for the errors of the previous model (Sagi & Rokach, 2018).

#### 4.2 | External (optimization) adaptation of model complexity

On top of internal mechanisms to adopt model complexity, most practical ML pipelines apply an additional optimization step where hyperparameters of the model are optimized under cross-validation (or simply into training, evaluation, and test splits for large DL models).

Hyperparameters are parameters that do not directly control predictions, but rather the architecture (e.g. number of nodes in a hidden layer of a neural network or the number of trees in a RF) or the learning behaviour of ML algorithms. Some ML algorithms have few (e.g., RF), others many (BRT or DL) hyperparameters. Hyperparameters are usually tuned via a nested cross-validation setup, that is, an outer cross-validation to estimate generally the prediction error of the model and an inner cross-validation to control the tuning (see Table 2 for ML frameworks).

A particularly important class of hyperparameters are regularization parameters, which control the flexibility of the algorithms. In general, regularization means imposing constraints on an algorithm to

limit its flexibility. The type and strength of the regularization depends on the task, the data and the algorithm but the most common regularization type is a so-called shrinkage penalty which biases parameter estimates to a certain value, typically zero. For example, L1 (LASSO, Tibshirani, 1996) and L2 (Ridge; Hoerl & Kennard, 1970), or elastic-net when combined (Zou & Hastie, 2005), intentionally biases the estimates to zero. Shrinkage penalties were originally developed to estimate complex statistical models (e.g. when number of observations << number of predictors) such as linear or logistic regression models but have since been adopted in ML models. In tree-based methods (e.g. RF), hyperparameters such as the depth of the trees have regularizing effects, whereas in DL a range of regularization techniques is used, such as L1 or L2 on weights and dropout (where random parts of the network are set to zero during training; see Srivastava et al., 2014).

#### 4.3 | Open questions regarding model complexity in DL

While the principles of internal and external complexity adoption are central to both classical ML and DL algorithms, it is often conjectured that, they alone are not sufficient to explain the success of the highly complex DL algorithms, in particular the puzzling double-decent behaviour where generalization loss improves with model size even after the training loss has reached a value close to zero (Figure 2b), a behaviour that is not observed in simpler ML algorithms.

One hypothesis to explain the discrepancy between simple and deep neural networks is that overparameterization combined with stochastic training of the networks (stochastic gradient descent) leads to an implicit regularization (Arora, Cohen, et al., 2019; Huh et al., 2021; Li et al., 2021). This would explain why deep neural networks display a bias towards simpler functions (De Palma et al., 2019; Valle-Pérez et al., 2019) that increases with the depth of the networks (Huh et al., 2021). It was also observed that often over 90% of the trained networks' parameters can be set to zero with little or no loss of generalization accuracy (Frankle & Carbin, 2019), suggesting that there is a considerable amount of redundancy and possibly ensemble behaviour in deep neural networks. Such a pruning can reduce the computational cost of the model and reduce the generalization error (Bartoldson et al., 2020) or identify robust models (Kuhn et al., 2021). It was also suggested that the random initialization of a large DNN is more likely to create a good subnetwork (Frankle & Carbin, 2019; Zhang, Wang, et al., 2021), which is then identified by training, regularization or pruning (Zhang, Wang, et al., 2021). Moreover, most modern DL models consist of a mix of different architectures and techniques which can even include common ML concepts such as boosting or bagging. For example, dropout training can be interpreted as generating a large number of subnetworks, similar to an ensemble model (Srivastava et al., 2014), and deep residual networks (He et al., 2016) for image-based tasks resemble boosting and thus ensemble models (Veit et al., 2016). None of this does fully answers the question of DL's superiority but it does at least provide conjectures that need to be followed up by future research.

Name	Description	Language	Link
ranger	Random Forest algorithm	R	<a href="https://github.com/imbs-hl/ranger">https://github.com/imbs-hl/ranger</a>
xgboost	Boosted Machine framework	R, Python	<a href="https://github.com/dmlc/xgboost">https://github.com/dmlc/xgboost</a>
lightGBM	Boosted Machine framework	R, Python	<a href="https://github.com/microsoft/LightGBM">https://github.com/microsoft/LightGBM</a>
caret	ML framework for hyper-parameter tuning and cross-validation. Supports different ML algorithms	R	<a href="https://topepo.github.io/caret/">https://topepo.github.io/caret/</a>
mlr3	ML framework for hyper-parameter tuning and cross-validation. Supports different ML algorithms	R	<a href="https://mlr3.mlr-org.com/">https://mlr3.mlr-org.com/</a>
tidymodels	(ML) framework for hyper-parameter tuning and cross-validation. Supports different (ML) algorithms	R	<a href="https://www.tidymodels.org/">https://www.tidymodels.org/</a>
Scikit-learn	ML framework for hyper-parameter tuning and cross-validation. Supports different ML algorithms	Python	<a href="https://scikit-learn.org/">https://scikit-learn.org/</a>
TensorFlow	Deep Learning framework	R, Python	<a href="https://www.tensorflow.org/">https://www.tensorflow.org/</a>
Keras	Higher-level deep learning framework	R, Python	<a href="https://keras.io/">https://keras.io/</a>
PyTorch	Deep Learning framework	R, Python	<a href="https://pytorch.org/">https://pytorch.org/</a>
PyTorch Geometric	Graph Neural Network (GNN) framework. Supports different GNN algorithms	Python	<a href="https://github.com/pyg-team/pytorch_geometric">https://github.com/pyg-team/pytorch_geometric</a>
Flux	Deep learning framework	Julia	<a href="https://fluxml.ai/Flux.jl/stable/">https://fluxml.ai/Flux.jl/stable/</a>
MLJ	ML framework. Supports different ML algorithms	Julia	<a href="https://alan-turing-institute.github.io/MLJ.jl/stable/">https://alan-turing-institute.github.io/MLJ.jl/stable/</a>

**TABLE 2** Common machine learning and deep learning libraries and frameworks.

## 5 | EMERGING TRENDS IN ML (IN E&E)

In the last section of this paper, we will look at the current practice and emerging trends in ML and speculate on how they will impact the field of E&E.

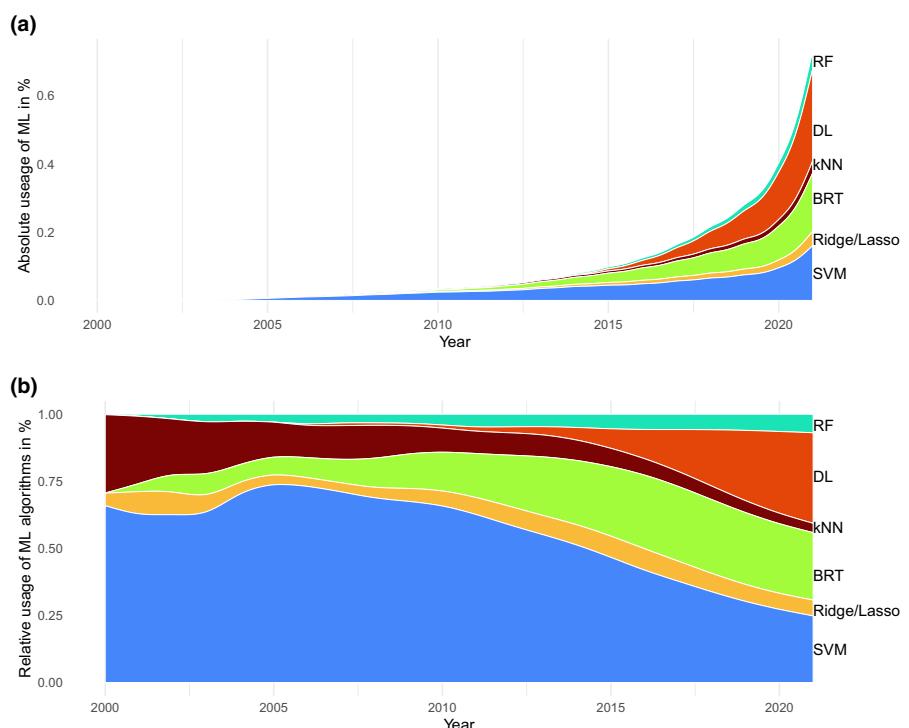
### 5.1 | Trends in algorithm use in E&E

As a basis for this discussion, we performed a text analysis of the E&E literature over the last decades (for details see Appendix S1.1). Our results show that the use of both ML and DL methods in E&E

increased sharply over the last decade (Figure 3). Classical ML methods still dominate in practical applications. Of those, SVMs were the most popular algorithm in the early 2000s, but lost their dominance since then. BRTs became popular in the mid 2010s (Figure 3b), and more recently, neural networks (including DL) are rising in popularity (Figure 3b). The increase in publications using DL approaches explains why these algorithms receive so much attention in recent reviews, but our analysis (Figure 3b) also highlights that classical ML methods still account for a proportionally larger share of all applications.

We anticipate that classical ML will remain important in the future, as many tasks in E&E are more naturally approached with

**FIGURE 3** Development usage of ML algorithms (RF, DL, kNN, BRT, Ridge/Lasso, SVM) in literature from the E&E field (see Appendix S1.1 for more details about the trend analysis). Panel (a) shows the absolute change in their usage in percent and panel (b) shows the relative change in their usage in percent. The overall usage of ML algorithms increased strongly over the last 20 years and especially DL attracted a lot of attention in the last 10 years. BRT, boosted regression trees; DL, deep Learning; E&E, ecology and evolution; kNN, k-nearest neighbour; ML, machine learning; RF, random forest; Ridge/Lasso, ridge or lasso or elastic-net (ridge and lasso) regression; SVM, support vector machines.



simpler ML algorithms. In particular, there is little evidence that DL can outperform classical ML algorithms in supervised learning tasks with limited structured (tabular) data (cf. Strydom et al., 2021). The higher flexibility of DL algorithms tends to be advantageous only when the data is large and complex enough. One would therefore expect that classical ML algorithms will continue to be used for tasks such as species distribution modelling (Beery et al., 2021; Elith & Leathwick, 2009), with subsequent applications for identifying conservable or restorable areas (Cheng et al., 2018; Duhart et al., 2019; Kwok, 2019; Moradi et al., 2019), forest management (Lauer et al., 2020), ecosystem service management (Dietterich et al., 2012; Scowen et al., 2021), wildlife management (Humphries et al., 2018) and conservation (see Tuia et al., 2022), assessing the risk of invasive species (Barbet-Massin et al., 2018; Jensen et al., 2020), and biodiversity assessments (Distler et al., 2015). Other tasks where classical ML will likely remain competitive include filling (knowledge) gaps in datasets (Penone et al., 2014) or in ecological networks (e.g. food webs, Desjardins-Proulx et al., 2017; plant-pollinator networks, Pichler et al., 2020; host-parasite networks, Dallas et al., 2017, 2021), and predicting potential wildlife hosts of zoonotic diseases (Albery et al., 2021; see Becker et al., 2022; Han et al., 2015; Wardeh et al., 2021).

Deep learning algorithms, on the other hand, will likely continue to gain popularity for analyzing complicated and unstructured data in E&E, such as species identification in aerial images (Ferreira et al., 2020; Gray et al., 2019; Guirado et al., 2018; Torney et al., 2019) or camera (trap) images (Aodha et al., 2018; Beery et al., 2020; Fairbrass et al., 2018; Ferreira et al., 2020; Fritzler et al., 2017; Lasseck, 2018; Mäder et al., 2021; Norouzzadeh et al., 2021; Stowell et al., 2018; Tabak et al., 2019; Van Horn et al., 2018; Willi et al., 2019).

For clustering and ordination tasks, which have a long tradition in ecology and ML algorithms for unsupervised learning tasks (Box 2), classical ML algorithms such as k-means or t-distributed stochastic neighbour embedding algorithms are and will remain important, for example for species delimitation (Derkarabetian et al., 2019), outlier detection, identification of eco-provinces (Sonnewald et al., 2020) or operational taxonomic units (OTUs) in metabarcoding (Deiner et al., 2017). DL-based approaches (e.g. based on [variational] autoencoders), on the other hand, are gaining popularity into certain data-dependent tasks, such as image-based tasks in remote sensing (Zerrouki et al., 2021) or (genomic) sequences (Wang & Gu, 2018).

As more and more data will become available in the future (Albery et al., 2021), and ecologists are just beginning to grasp the applications of DL, we anticipate that these applications, based on unstructured data such as images or sequences, will continue to grow faster than classical ML algorithms, until the time that a methodological equilibrium is reached.

## 5.2 | New applications for ML in E&E

Apart from improving the quality of classical prediction and classification tasks, there are many novel applications that could be addressed in particular by the more advanced DL algorithms. For example, Davies et al. (2021) demonstrated that DL can aid researcher by generating new hypotheses which were tested afterwards, or it was shown that modern DL models can achieve human-like performance in text generation (Brown et al., 2020). Generative models may play an increasing role in the coming years; however, it is currently difficult to predict where they will be used in ecological research, for example, whether they will help in data-collection or in

subsequently answering the research questions themselves (e.g. by generating new hypotheses).

Another interesting field for ML is simulations and simulation-based inference. For stochastic simulations or big process-based models, likelihoods are often intractable or computationally expensive to evaluate (e.g. phylogenetic analyses). ML and DL algorithms can support simulation-based inference by generating new summary statistics (e.g. Hauenstein et al., 2019), by being incorporated into process-based models for computational gains (e.g. Rammer & Seidl, 2019), or by emulating them (Wang et al., 2019). Moreover, ML can also be used to predict the parameters of complex stochastic models (Roy et al., 2022; Voznica et al., 2022), and thus act as a likelihood-free calibration method, similar to approximate Bayesian computation (Hartig et al., 2011).

Moreover, in the era of cheap sensors and other data collection sources, the dimensionality of the data is often difficult to handle with traditional methods. Unsupervised learning algorithms can help to reduce the dimensionality of the data and detect patterns and trends, for example, before the data is used in downstream supervised learning tasks (Strydom et al., 2021; Zerrouki et al., 2021), to handle the data itself (Alves de Oliveira et al., 2021), or to detect anomalies in the data (Zhang, Xu, et al., 2021).

### 5.3 | Rethinking the data collection process in the light of the new methods

The wide availability of DL algorithms could also have a strong impact on data collection in E&E. Image recognition methods can reduce labor costs and thus help to generate much larger datasets. DL can identify species in different data types (Ferreira et al., 2020; Gray et al., 2019; Guirado et al., 2018; Mäder et al., 2021; Tabak et al., 2019; Torney et al., 2019; Willi et al., 2019) or extract information such as traits from raw data (Dunker et al., 2020). Moreover, technical advances in eDNA and other sensor data allow the collection of much larger datasets (e.g. see Pimm et al., 2015) which can then be processed and combined by ML and DL algorithms (see Tuia et al., 2022).

One of the advantages of establishing such machine-assisted data collection and processing pipelines is that they can be reused by many researchers, similar to the development in sequencing technology. For example, once an image-based species recognition pipeline is established, it can be reused without requiring the time of taxonomic experts for data analysis. So far, there are few examples of such ready-to-use pipelines for realistic data collection tasks, and those that exist do not always perform well and generalize well to new situations. However, we believe that the field should develop ML models for data collection that are available to everyone (McIntire et al., 2022) and do not need to be retrained by experts. The NLP community demonstrate how this could be done: Model hubs with many different pre-trained models and a simple and common interface that can be used by everyone (e.g. Mäder et al., 2021; Wolf et al., 2019; cf. Ott & Lautenschlager, 2021).

### 5.4 | Making ML work with small datasets

A pervasive problem in the application of many modern ML algorithms for E&E is that their training requires data sizes that are rarely available. New DL techniques such as few-shot learning (see Wang et al., 2020) or transfer learning can greatly reduce the necessary amount of data, and are thus of particular interest to ecologists. As an example, most DL-based image classifiers consist of two stages: first, they identify edges and shapes in the images, and second, they classify the shapes (LeCun et al., 2015). The first stage makes up a major part of the model and is both data and resource intensive. Research has shown that this part of the network is relatively generic, and usually only the second stage needs to be retrained when a network is adopted for a new task (Weiss et al., 2016; Zhuang et al., 2021). Thus, many large model architectures can be downloaded pre-trained and can then be fine-tuned for a new task (transfer learning).

Options such as transfer learning, however, are mainly applicable to vision-based tasks and DL, and not to classical structured tabular data. In the latter, the response is often explained by specific relationships with a particular feature, which rarely generalize to other tasks. In such a situation, there is little to be gained by applying transfer learning, which may also partly explain why DL rarely outperforms traditional ML algorithms on small classical structured tabular data (Pichler et al., 2020; Schwartz et al., 2020; cf. Arik & Pfister, 2020).

Small datasets are common in E&E because observations are often difficult to obtain (e.g. for ecological networks, see Magliaresi et al., 2014; Strydom et al., 2021); and because due to the change in ecological patterns across scales (Poisot et al., 2015), datasets are difficult to combine. Here, E&E researchers can benefit from the wide range of different ML algorithms (Figure 4): SVMs or kNN can handle sparse datasets well (Como et al., 2017; Drake et al., 2006) and LASSO, Ridge, and elastic-net regressions are well suited for datasets with more features than observations (Zou & Hastie, 2005). On top of the data dimensions, the nature of the signal and the interpretability can influence the choice of the modelling approach (e.g. Pichler et al., 2020). However, these trade-offs are difficult to predict a priori, which explains the common practice of comparing different algorithms for a given task (Faisal et al., 2010; Norberg et al., 2019; Pichler et al., 2020).

### 5.5 | Transparency and bias of decision based on ML and DL models

Predictions and research in E&E often intersect with policy and decision-making (de Groot et al., 2010; Sofaer et al., 2019). As ML models are increasingly used in this context, for example for conservation planning (Huettmann, 2018), management decisions (Humphries et al., 2018), agricultural management (e.g. crop management; Liakos et al., 2018), and disease control (e.g. Romero et al., 2021), we anticipate that problems of bias and transparency

will emerge, as they have in other fields (e.g. Hardt et al., 2016; Vayena et al., 2018).

Transparency refers to the problem that stakeholders may question why certain predictions are being made by the algorithm and whether they can be trusted. Without satisfactory answers to these questions, ML decisions may be subject to legal challenges. While it is not impossible to answer these questions for ML models (see the next subsection on xAI), it is undoubtedly more challenging to understand and communicate the logic of ML decisions, compared to simple statistical models (Figure 4).

This lack of transparency also makes it difficult to understand if an algorithm exhibits bias. In the context of ML and AI, bias is understood more broadly than in statistics, and includes both the use of non-representative or socially undesirable training data and the use of features that should not normally be used in decisions (e.g. gender, race). The former occurs when training data was disproportionately collected in different groups or regions (Zou & Schiebinger, 2018) and is not representative of what the algorithm should learn. A common example from the social sciences is that language models trained on classical literature or texts often learn gender-biased word associations, such as a preference for doctors to be male. Biased data may be a significant problem for E&E, as geographic (Martin et al., 2012; Meyer et al., 2016) and taxonomic (Pyšek et al., 2008; Trimble & van Aarde, 2012) sampling biases are common. The use of undesired features describes situations where the data may be representative, but certain features should not be used for ethical reasons. The challenge here is that these features may be implicitly encoded by other features and thus be used in ML and DL algorithms, even if they are not explicitly provided in the data (e.g. ethnic background can be inferred from the people's urban districts, Caliskan et al., 2017; Feuerriegel et al., 2020; Hardt et al., 2016).

To understand these issues and to find solutions, the field of responsible and trustworthy AI has formed at the intersection between AI and social science disciplines (sociology, psychology, law). The focus of responsible AI is on creating fair and sustainable ML and DL models and avoiding their misuse or misinterpretation (e.g. Barredo Arrieta et al., 2020; Wearn et al., 2019). For example, it is possible to algorithmically detect biases (Cirillo et al., 2020) and correct the models accordingly (e.g. Alvi et al., 2018; Kim et al., 2019) by using fairness metrics to guide training so that underrepresented groups are not neglected (e.g. Liu & Vicente, 2021).

## 5.6 | Explainable AI as peephole in black-box models

The previously mentioned transparency issues are amplified by the fact that ML algorithms become increasingly difficult to interpret as model complexity increases (Figure 4; Breiman, 2001b). Although some algorithms provide metrics for feature importance (e.g. Breiman, 2001a; Friedman, 2001), ML algorithms

usually do not provide simple effect estimates, nor do they provide measures of certainty, such as CIs or *p*-values. This poses a problem not only for ethical transparency but also for researchers that want to understand why predictions are made for scientific reasons.

To address this problem, the field of explainable AI (xAI) has emerged that develops tools to understand how ML models make their predictions. Most xAI methods are post-hoc, meaning that the model is trained first and then investigated (Barredo Arrieta et al., 2020; Lucas, 2020). xAI differs from other similar-sounding approaches such as identifying predictive trait profiles (domain expertise is used to group features and by including and excluding them from the model, their predictive capabilities are estimated (e.g. Han et al., 2015)), in that the goal is to understand the model itself. Global xAI methods try to summarize the models by generating variable importance (similar to the natural variable importances from RF and BRT; Fisher et al., 2018) or simplify the models by approximating the original model with interpretable models (Molnar, 2020). Local xAI methods attempt to explain individual predictions (Ribeiro et al., 2016). In E&E, for example xAI methods are already being used to assess trust of predictions from SDMs (Ryo et al., 2021). Although there are still many questions about the reliability of these methods, especially under collinearity of features (Hooker & Mentch, 2019; Yu et al., 2021, but see Apley & Zhu, 2020), xAI is becoming an indispensable tool for working with ML models, and presumably there will be specialized xAI methods for ecological applications.

## 5.7 | Causal inference with ML

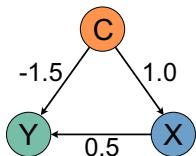
Finally, it is not uncommon for models to give us the right answer for the wrong reason. Sometimes, it is even easier to get good predictions for the wrong reasons. An example is severe feature collinearity, where including all features increases the uncertainties of the estimates (Greenland, 2003; Lederer et al., 2018), whereas removing features, even if causally connected to the response, can improve predictive performance (Arif & MacNeil, 2022; Dormann et al., 2013; Hoerl & Kennard, 1970). Interestingly, some of the regularization techniques now widely used in ML were originally developed to reduce collinearity problems (Hoerl & Kennard, 1970; Lederer et al., 2018), although that does not mean that they necessarily always 'select' the causal one from two collinear features (Zou & Hastie, 2005).

Optimizing ML algorithms for predictive performance means that, generally, we should not assume that ML algorithms will learn the correct causal dependencies between the input features and the response (Box 4). It also means that we should not interpret xAI metrics as "effect estimates"—if a certain feature is strongly used by the ML model, it could be because this feature has a biological or ecological effect on the response, but it could also easily be because it correlates with other features (e.g. Genuer et al., 2010).

#### BOX 4 PREDICTIVE VERSUS CAUSAL MODEL BUILDING STRATEGIES

The best predictive model need not be the true causal model. To demonstrate this, we created a simulated dataset, where the response variable  $Y$  is affected by the feature  $X$  with a causal effect of 0.5 and by a second feature  $C$  with the causal effect of -1.5 (Figure B4).  $C$  also has a causal effect of 1.0 on  $X$ .  $X$  and  $C$  are thus highly correlated ( $>0.9$  Pearson correlation factor). We fitted different models (full model, model with only  $X$  or  $C$ , and full model with a ridge regularization ( $\lambda = 0.01$ ), Figure B3) on 20 observations and estimated the prediction error (root-mean-square-error [RMSE]) on 480 observations of the holdout. We repeated the simulation and the model fitting 10,000 times.

In causal inference, the objective is to obtain correct effect estimates, which means that sometimes otherwise uninteresting collinear features must be included to control for confounding (variable  $C$ , Figure B4), while other structures, such as collider, must be excluded to obtain correct estimates. Thus, in causal analysis, the focus is to establish a correct hypothesis about the causal structure to obtain correct effect estimates (first model, Figure B5). The causal structure can be based on logical considerations or causal discovery algorithms. Importantly, minimizing predictive error is not the primary goal of the analysis, and controlling confounders often increases uncertainties of the parameter estimates that propagate through the model and negatively affect the predictive error (note that the true causal model shows the highest RMSE, Figure B5).

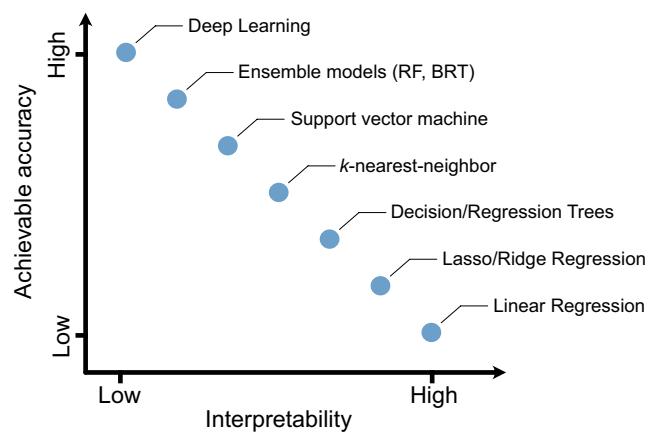


**FIGURE B4** Small example. We are interested how  $X$  effects  $Y$ .  $C$  is a confounder, that is, affecting  $X$  and  $Y$ .  $C$  and  $X$  are highly correlated ( $>0.9$  Pearson correlation factor). Numbers show the true effect estimates.

Model	X-estimate	C-estimate	RMSE on holdout
$Y \sim X + C$	0.5	-1.5	0.109
$Y \sim X$	-1.0	.	0.107
$Y \sim C$	.	-1.0	0.106
Ridge	-0.32	-0.59	0.106

**FIGURE B5** Results of different model specifications. We simulated 10,000 times from our small example (Figure B1) and fitted four different models on 20 observations: (a) The full model with the Confounder ( $C$ ) and our variable of interest ( $X$ ), (b) only our variable of interest, (c) only the confounder, and (d) the full model with a ridge regularization ( $\lambda = 0.01$ ). We evaluated the prediction error of the models by calculating the root-mean-square error of the predictions for the holdout (480 observations).

In predictive modelling, our goal is to minimize the prediction error of our model. A common strategy is to provide the model with all the variables and use methods such as regularization or AIC selection to reduce model complexity and find the sweet spot of the bias variance tradeoff (see Figure B5 last model). In such an approach, collinear features are often removed because they increase uncertainties while contributing relatively little to the prediction, given that their effects can be “emulated” by other features. In our simulation (Figure B4), we see that the true causal model has the highest prediction error, but correct effect estimates. The other models, have incorrect estimates but smaller prediction errors (Figure B5).



**FIGURE 4** Conceptual illustration of the trade-off between achievable accuracy and interpretability of machine learning and deep Learning (DL) algorithms. DL can achieve the highest accuracy but shows the lowest interpretability. BRT, boosted regression trees; RF, random forest.

Nevertheless, there are interesting ideas for exploiting ML for (causal) data analysis. For example, if we have a high dimensional dataset that would be difficult to analyse with conventional statistical tools, we could use ML and xAI to identify interesting patterns or features (Lucas, 2020; Pichler et al., 2020), and test the later in a confirmatory analysis.

Moreover, based on a causal analysis, we could pre-select features that are consistent with ideas of causal inference. When selected in such a way, ML algorithm can achieve more exact control for confounders due to their greater flexibility (Tank et al., 2021; Wein et al., 2021; Zečević et al., 2021), and they can be used to estimate causal effects of treatments (Chernozhukov et al., 2018; Wager & Athey, 2018). Another interesting approach is to combine statistical models with ML and DL algorithms (Joseph, 2020; Masahiro & Rillig, 2017; Tank et al., 2021).

A third idea is to incorporate physical laws as constraints in the learning of neural networks. Such physics induced (or informed) neural networks (PINN, see Karniadakis et al., 2021) were originally developed to improve predictions but could also mark an important milestone in combining ML and causal inference by forcing the models to adhere to known physical or biological laws. This could be of interest for a field as E&E (Wesselkamp et al., 2022) that has acquired a lot of knowledge about various ecological systems over the years.

Finally, there is active research to develop ML methods that directly achieve causal discovery. For example, DL research has shown promising progress in symbolic regression where equations for systems are automatically inferred (Cardoso et al., 2020; d'Ascoli et al., 2022).

In summary, causal inference or causal discovery with ML is challenging, but there are various ways to combine ML methods with causal inference and based on the interest in this topic in the ML and DL field, but also the interest of ecologists to understand causal relationships in their data, we believe that the importance of this topic will increase in the coming years.

## 6 | CONCLUSION

ML and DL algorithms are powerful and very general tools for predictive modelling and data analysis. Currently, DL algorithms have conquered image-based and similar tasks on complex, unstructured data, while classical ML algorithms such as RF and BRT still excel on structured data. The superior performance of ML and DL algorithms compared to statistical models can be explained by their higher flexibility and automatic data-dependent complexity optimization. For example, ML algorithms such as RF and BRT balance complexity by combining uncorrelated weak models into an ensemble, and DL uses a combination of indirect regularization through overparameterization and stochastic gradient descent. Compared to classical statistical tools in E&E, ML methods are rather optimized for prediction, and caution is advised with their causal interpretations (Box 4).

A common challenge for both statistical models and ML alike is making predictions that extrapolate outside the feature space used to train the model (out-of-distribution predictions, see Beery et al., 2018; Koh et al., 2021; Box 3). The reason why such predictions fail is that predictive models often learn to use non-causal proxies but also that relationships do not necessarily remain linear outside the area of data. Efforts have been made to identify such potentially unreliable predictions *a priori* (e.g. Meyer & Pebesma, 2021) or to correct for them (e.g. Tseng et al., 2022). Possibly, extending causal principles to ML (Zhao & Hastie, 2021), or incorporating ecological or mechanistic understanding, for example as additional model constraints, could help to improve model generalizability.

Much research in recent years has focused on making ML algorithms more transparent and bridging the gap between the properties of classical statistical tools and ML tools (e.g. xAI). New methods such as Bayesian neural networks have paved the way to obtain uncertainty and prediction intervals for DL models, bringing ML algorithms closer to statistical models (Ashukha et al., 2021; Loquercio et al., 2020). Despite that, their use for (causal) statistical inference is still controversial, not least because it is still unclear if ML can separate causal effects under feature collinearity, which is a basic requirement for causal inference.

Finally, despite the well-deserved attention, ML does not offer a free lunch. We have seen that the focus of ML methods on minimizing prediction comes at a cost elsewhere, such as in data requirements, interpretability or runtime. Even more strongly than statistical models, ML depends on the quality and the quantity of the data. Because of this, we should carefully consider whether the application of ML or even DL is necessary or promising for a task when simpler models with the advantages of higher interpretability, higher statistical power, and lower computational costs (and thus a better CO<sub>2</sub> footprint, Schwartz et al., 2020) could do the job (Mignan & Broccardo, 2019). Nevertheless, we expect ML to become an indispensable tool in E&E, comparable to other traditional statistical tools such as linear regression models or analysis of variance models that have been used for many years.

## AUTHOR CONTRIBUTIONS

Maximilian Pichler and Florian Hartig jointly conceived and designed the study. Both authors contributed equally to the writing and preparation of the manuscript.

## ACKNOWLEDGEMENTS

Maximilian Pichler received funding from the Bavarian Ministry of Science and the Arts in the Context of Bavarian Climate Research Network (bayklif). We thank Daniel Rettelbach and Tankred Ott, and two anonymous reviewers for their valuable comments and suggestions. Open Access funding enabled and organized by Projekt DEAL.

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/2041-210X.14061>.

## DATA AVAILABILITY STATEMENT

Code chunks for the different ML algorithms for different programming languages (R, Python, and Julia) can be found in the Supporting Information S1. Trend analysis (including the data) and scripts for reproducing the simulations and figures can be found in Pichler and Hartig (2022).

## ORCID

Maximilian Pichler  <https://orcid.org/0000-0003-2252-8327>

Florian Hartig  <https://orcid.org/0000-0002-6255-9059>

## REFERENCES

- Albery, G. F., Becker, D. J., Brierley, L., Brook, C. E., Christofferson, R. C., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A., Farrell, M. J., Glennon, E., Guth, S., Joseph, M. B., Mollentze, N., Neely, B. A., Poisot, T., Rasmussen, A. L., Ryan, S. J., Seifert, S., ... Carlson, C. J. (2021). The science of the host–virus network. *Nature Microbiology*, 6(12), 12. <https://doi.org/10.1038/s41564-021-00999-5>
- Alves de Oliveira, V., Chabert, M., Oberlin, T., Poulliat, C., Bruno, M., Latry, C., Carlan, M., Henrot, S., Falzon, F., & Camarero, R. (2021). Reduced-complexity end-to-end Variational autoencoder for on board satellite image compression. *Remote Sensing*, 13(3), 3. <https://doi.org/10.3390/rs13030447>
- Alvi, M., Zisserman, A., & Nellaaker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. [https://openaccess.thecvf.com/content\\_eccv\\_2018\\_workshops/w5/html/Alvi\\_Turning\\_a\\_Blind\\_Eye\\_Explicit\\_Removal\\_of\\_Biases\\_and\\_Variation\\_ECCVW\\_2018\\_paper.html](https://openaccess.thecvf.com/content_eccv_2018_workshops/w5/html/Alvi_Turning_a_Blind_Eye_Explicit_Removal_of_Biases_and_Variation_ECCVW_2018_paper.html)
- Aodha, O. M., Gibb, R., Barlow, K. E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G. R., Newson, S. E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., & Jones, K. E. (2018). Bat detective—Deep learning tools for bat acoustic signal detection. *PLOS Computational Biology*, 14(3), e1005995. <https://doi.org/10.1371/journal.pcbi.1005995>
- 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. <https://doi.org/10.1111/rssb.12377>
- Arif, S., & MacNeil, A. (2022). Predictive models aren't for causal inference. *Ecology Letters*, 25, 1741–1745. <https://doi.org/10.1111/ele.14033>
- Arik, S. O., & Pfister, T. (2020). TabNet: Attentive interpretable tabular learning. *ArXiv:1908.07442 [Cs, Stat]*. <http://arxiv.org/abs/1908.07442>
- Arora, S., Cohen, N., Hu, W., & Luo, Y. (2019). Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32. <https://proceedings.neurips.cc/paper/2019/hash/c0c783b5fc0d7d808f1d14a6e9c8280d-Abstract.html>
- Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., & Wang, R. (2019). On exact computation with an infinitely wide neural net. *ArXiv:1904.11955 [Cs, Stat]*. <http://arxiv.org/abs/1904.11955>
- Ashukha, A., Lyzhov, A., Molchanov, D., & Vetrov, D. (2021). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *ArXiv:2002.06470 [Cs, Stat]*. <http://arxiv.org/abs/2002.06470>
- Barbet-Massin, M., Rome, Q., Villemant, C., & Courchamp, F. (2018). Can species distribution models really predict the expansion of invasive species? *PLoS One*, 13(3), e0193085. <https://doi.org/10.1371/journal.pone.0193085>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Bartoldson, B. R., Morcos, A. S., Barbu, A., & Erlebacher, G. (2020). The generalization-stability tradeoff In neural network pruning. *ArXiv:1906.03728 [Cs, Stat]*. <http://arxiv.org/abs/1906.03728>
- Becker, D. J., Albery, G. F., Sjodin, A. R., Poisot, T., Bergner, L. M., Chen, B., Cohen, L. E., Dallas, T. A., Eskew, E. A., Fagre, A. C., Farrell, M. J., Guth, S., Han, B. A., Simmons, N. B., Stock, M., Teeling, E. C., & Carlson, C. J. (2022). Optimising predictive models to prioritise viral discovery in zoonotic reservoirs. *The Lancet Microbe*, 3, e625–e637. [https://doi.org/10.1016/S2666-5247\(21\)00245-7](https://doi.org/10.1016/S2666-5247(21)00245-7)
- Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species distribution modeling for machine learning practitioners: A review. ACM SIGCAS Conference on Computing and Sustainable Societies, 329–348. <https://doi.org/10.1145/3460112.3471966>
- Beery, S., van Horn, G., & Perona, P. (2018). Recognition in Terra Incognita (arXiv:1807.04975). arXiv. <https://doi.org/10.48550/arXiv.1807.04975>
- Beery, S., Wu, G., Rathod, V., Votol, R., & Huang, J. (2020). Context R-CNN: Long term temporal context for per-camera object detection. [https://openaccess.thecvf.com/content\\_CVPR\\_2020/html/Beery\\_Context\\_R-CNN\\_Long\\_Term\\_Temporal\\_Context\\_for\\_Per-Camera-Object\\_Detection\\_CVPR\\_2020\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2020/html/Beery_Context_R-CNN_Long_Term_Temporal_Context_for_Per-Camera-Object_Detection_CVPR_2020_paper.html)
- Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine learning practice and the bias-variance trade-off. *ArXiv:1812.11118 [Cs, Stat]*. <http://arxiv.org/abs/1812.11118>
- Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.
- Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., Pinto, H. P. O., Raiman, J., Salimans, T., ... Zhang, S. (2019). Dota 2 with large scale deep reinforcement learning. *ArXiv:1912.06680 [Cs, Stat]*. <http://arxiv.org/abs/1912.06680>
- Birdal, T., Lou, A., Guibas, L. J., & Simsekli, U. (2021). Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34, 6776–6789.

- Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022). Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13, 1640–1660. <https://doi.org/10.1111/2041-210X.13901>
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory*. <https://doi.org/10.1145/130385.130401>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1), 5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Brieuc, M. S. O., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, 18(4), 755–766. <https://doi.org/10.1111/1755-0998.12773>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- Cardoso, P., Branco, V. V., Borges, P. A. V., Carvalho, J. C., Rigal, F., Gabriel, R., Mammola, S., Cascalho, J., & Correia, L. (2020). Automated discovery of relationships, models, and principles in ecology. *Frontiers in Ecology and Evolution*, 8. <https://doi.org/10.3389/fevo.2020.530135>
- Chen, D., Xue, Y., & Gomes, C. P. (2018). End-to-end learning for the deep multivariate Probit model. *ArXiv:1803.08591 [Cs, Stat]*. <http://arxiv.org/abs/1803.08591>
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2019). Neural ordinary differential equations. *ArXiv:1806.07366 [Cs, Stat]*. <http://arxiv.org/abs/1806.07366>
- Cheng, S. H., Augustin, C., Bethel, A., Gill, D., Anzaroot, S., Brun, J., DeWilde, B., Minich, R. C., Garside, R., Masuda, Y. J., Miller, D. C., Wilkie, D., Wongbusarakum, S., & McKinnon, M. C. (2018). Using machine learning to advance synthesis and use of conservation and environmental evidence. *Conservation Biology*, 32(4), 762–764. <https://doi.org/10.1111/cobi.13117>
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters*. Oxford University Press.
- Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10), 1632–1644. <https://doi.org/10.1111/2041-210X.13256>
- Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., & Mavridis, N. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digital Medicine*, 3(1), 1. <https://doi.org/10.1038/s41746-020-0288-5>
- Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, 8(1), 2–14. <https://doi.org/10.1111/j.1461-0248.2004.00702.x>
- Como, F., Carnesecchi, E., Volani, S., Dorne, J. L., Richardson, J., Bassan, A., Pavan, M., & Benfenati, E. (2017). Predicting acute contact toxicity of pesticides in honeybees (*Apis mellifera*) through a k-nearest neighbor model. *Chemosphere*, 166, 438–444. <https://doi.org/10.1016/j.chemosphere.2016.09.092>
- Dallas, T., Park, A. W., & Drake, J. M. (2017). Predictability of helminth parasite host range using information on geography, host traits and parasite community structure. *Parasitology*, 144(2), 200–205. <https://doi.org/10.1017/S0031182016001608>
- Dallas, T., Ryan, S. J., Bellekom, B., Fagre, A., Christofferson, R., & Carlson, C. (2021). Predicting the tripartite network of mosquito-borne disease. *EcoEvoRxiv*. <https://doi.org/10.32942/osf.io/xzmp8>
- d'Ascoli, S., Kamienny, P.-A., Lample, G., & Charton, F. (2022). Deep symbolic regression for recurrent sequences. *ArXiv:2201.04600 [Cs]*. <http://arxiv.org/abs/2201.04600>
- Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., Lackenby, M., Williamson, G., Hassabis, D., & Kohli, P. (2021). Advancing mathematics by guiding human intuition with AI. *Nature*, 600(7887), 70–74. <https://doi.org/10.1038/s41586-021-04086-x>
- de Groot, R. S., Alkemade, R., Braat, L., Hein, L., & Willemen, L. (2010). Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. *Ecological Complexity*, 7(3), 260–272. <https://doi.org/10.1016/j.ecocom.2009.10.006>
- De Palma, G., Kiani, B. T., & Lloyd, S. (2019). Random deep neural networks are biased towards simple functions. *ArXiv:1812.10156 [Cond-Mat, Physics:Math-Ph, Physics:Quant-Ph, Stat]*. <http://arxiv.org/abs/1812.10156>
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, 26(21), 5872–5895. <https://doi.org/10.1111/mec.14350>
- 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. <https://doi.org/10.1371/journal.pcbi.1008856>
- Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., & Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*, 139, 106562. <https://doi.org/10.1016/j.ympev.2019.106562>
- Desjardins-Proulx, P., Laigle, I., Poisot, T., & Gravel, D. (2017). Ecological interactions and the Netflix problem. *PeerJ*, 5, e3644. <https://doi.org/10.7717/peerj.3644>
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Springer. [https://doi.org/10.1007/3-540-45014-9\\_1](https://doi.org/10.1007/3-540-45014-9_1)
- Dietterich, T. G., Dereszynski, E. W., Hutchinson, R. A., & Sheldon, D. (2012). *Machine learning for computational sustainability*. IGCC.
- Distler, T., Schuetz, J. G., Velásquez-Tibatá, J., & Langham, G. M. (2015). Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography*, 42(5), 976–988. <https://doi.org/10.1111/jbi.12479>
- Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., ... Hartig, F. (2018). Model averaging in ecology: A review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, 88(4), 485–504. <https://doi.org/10.1002/ecm.1309>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Drake, J. M., Randin, C., & Guisan, A. (2006). Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, 43(3), 424–432. <https://doi.org/10.1111/j.1365-2664.2006.01141.x>

- Duhart, C., Dublon, G., Mayton, B., Davenport, G., & Paradiso, J. A. (2019). Deep learning for wildlife conservation and restoration efforts. In *36th International conference on machine learning, Long Beach* (Vol. 5).
- Dunker, S., Motivans, E., Rakosy, D., Boho, D., Mäder, P., Hornick, T., & Knight, T. M. (2020). Pollen analysis using multispectral imaging flow cytometry and deep learning. *New Phytologist*, 229, 593–606. <https://doi.org/10.1111/nph.16882>
- Dushoff, J., Kain, M. P., & Bolker, B. M. (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*, 10(6), 756–759. <https://doi.org/10.1111/2041-210X.13159>
- Efron, B. (1992). Bootstrap methods: Another look at the jackknife. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: Methodology and distribution* (pp. 569–593). Springer. [https://doi.org/10.1007/978-1-4612-4380-9\\_41](https://doi.org/10.1007/978-1-4612-4380-9_41)
- 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(1), 677–697. <https://doi.org/10.1146/annurev.ecolsys.110308.120159>
- Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813. <https://doi.org/10.1111/j.1365-2656.2008.01390.x>
- Fairbrass, A. J., Firman, M., Williams, C., Brostow, G. J., Titheridge, H., & Jones, K. E. (2018). City net—Deep learning tools for urban eco-acoustic assessment. *Methods in Ecology and Evolution*, 10, 186–197. <https://doi.org/10.1111/2041-210X.13114>
- Faisal, A., Dondelinger, F., Husmeier, D., & Beale, C. M. (2010). Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5(6), 451–464. <https://doi.org/10.1016/j.ecoinf.2010.06.005>
- Ferreira, A. C., Silva, L. R., Renna, F., Brandl, H. B., Renoult, J. P., Farine, D. R., Covas, R., & Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9), 1072–1085. <https://doi.org/10.1111/2041-210X.13436>
- Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI. *Business & Information Systems Engineering*, 62(4), 379–384. <https://doi.org/10.1007/s12599-020-00650-3>
- Fisher, A., Rudin, C., & Dominici, F. (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. ArXiv E-Prints.
- Foody, G. M. (1995). Land cover classification by an artificial neural network with ancillary information. *International Journal of Geographical Information Systems*, 9(5), 527–542. <https://doi.org/10.1080/02693799508902054>
- Frankle, J., & Carbin, M. (2019). The lottery ticket hypothesis: Finding sparse, trainable neural networks. ArXiv:1803.03635 [Cs]. <http://arxiv.org/abs/1803.03635>
- French, M., & Recknagel, F. (1970). Modeling of algal blooms in freshwaters using artificial neural networks. *WIT Transactions on Ecology and the Environment*, 6, 87–94.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. <https://doi.org/10.1006/jcss.1997.1504>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.
- Fritzler, A., Koitka, S., & Friedrich, C. M. (2017). *Recognizing bird species in audio files using transfer learning*. LEF (working notes), 14.
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202. <https://doi.org/10.1007/BF00344251>
- Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. (2021). Ensemble deep learning: A review. ArXiv:2104.02395 [Cs]. <http://arxiv.org/abs/2104.02395>
- Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—A “garden of forking paths”—Explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6), 460.
- Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236. <https://doi.org/10.1016/j.patrec.2010.03.014>
- Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., & Couzin, I. D. (2019). Fast and robust animal pose estimation. *BioRxiv*, 620245. <https://doi.org/10.1101/620245>
- Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bézy, V. S., Lohmann, K. J., & Johnston, D. W. (2019). A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution*, 10(3), 345–355. <https://doi.org/10.1111/2041-210X.13132>
- Greenland, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3), 300–306.
- Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, 310(5750), 987–991. <https://doi.org/10.1126/science.1116681>
- Gualtieri, J. A., & Cromp, R. F. (1999). Support vector machines for hyperspectral remote sensing classification. 27th AIPR workshop: Advances in computer-assisted recognition, 3584, 221–232. <https://doi.org/10.1117/12.339824>
- Guirado, E., Tabik, S., Rivas, M. L., Alcaraz-Segura, D., & Herrera, F. (2018). Automatic whale counting in satellite images with deep learning. *BioRxiv*. <https://doi.org/10.1101/443671>
- Han, B. A., Schmidt, J. P., Bowden, S. E., & Drake, J. M. (2015). Rodent reservoirs of future zoonotic diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 112(22), 7039–7044. <https://doi.org/10.1073/pnas.1501598112>
- Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29. <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>
- Harris, D. J., Taylor, S. D., & White, E. P. (2018). Forecasting biodiversity in breeding birds using best practices. *PeerJ*, 6, e4278. <https://doi.org/10.7717/peerj.4278>
- Hartig, F., & Barraquand, F. (2022). The evidence contained in the P-value is context dependent. *Trends in Ecology & Evolution*, 37, S0169–S5347.
- Hartig, F., Calabrese, J. M., Reineking, B., Wiegand, T., & Huth, A. (2011). Statistical inference for stochastic simulation models—Theory and application: Inference for stochastic simulation models. *Ecology Letters*, 14(8), 816–827. <https://doi.org/10.1111/j.1461-0248.2011.01640.x>
- Hauenstein, S., Fattebert, J., Grüebler, M. U., Naef-Daenzer, B., Pe'er, G., & Hartig, F. (2019). Calibrating an individual-based movement model to predict functional connectivity for little owls. *Ecological Applications*, 29(4), e01873. <https://doi.org/10.1002/eaap.1873>
- He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. [http://openaccess.thecvf.com/content\\_iccv\\_2017/html/He\\_Mask\\_R-CNN\\_ICCV\\_2017\\_paper.html](http://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html)
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- Hooker, G., & Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. ArXiv:1905.03151 [Cs, Stat]. <http://arxiv.org/abs/1905.03151>
- Huettmann, F. (2018). Machine learning for ‘strategic conservation and planning’: Patterns, applications, thoughts and urgently needed global Progress for sustainability. In G. Humphries, D. R. Magness, & F. Huettmann (Eds.), *Machine learning for ecology and sustainable natural resource management* (pp. 315–333). Springer International Publishing. [https://doi.org/10.1007/978-3-319-96978-7\\_16](https://doi.org/10.1007/978-3-319-96978-7_16)

- Huh, M., Mobahi, H., Zhang, R., Cheung, B., Agrawal, P., & Isola, P. (2021). The low-rank simplicity bias in deep networks. *ArXiv:2103.10427 [Cs]*. <http://arxiv.org/abs/2103.10427>
- Humphries, G. R., Magness, D. R., & Huettmann, F. (2018). *Machine learning for ecology and sustainable natural resource management*. Springer.
- Jensen, T., Seerup Hass, F., Seam Alkbar, M., Holm Petersen, P., & Jokar Arsanjani, J. (2020). Employing machine learning for detection of invasive species using Sentinel-2 and AVIRIS data: The case of kudzu in the United States. *Sustainability*, 12(9), 9. <https://doi.org/10.3390/su12093544>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Joseph, M. B. (2020). Neural hierarchical models of ecological populations. *Ecology Letters*, 23(4), 734–747. <https://doi.org/10.1111/ele.13462>
- 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 Alpha Fold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 422–440. <https://doi.org/10.1038/s42254-021-00314-5>
- Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. 9012–9020. [https://openaccess.thecvf.com/content\\_CVPR\\_2019/html/Kim\\_Learning\\_Not\\_to\\_Learn\\_Training\\_Deep\\_Neural\\_Networks\\_With\\_Biased\\_CVPR\\_2019\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html)
- Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., ... Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th international conference on machine learning* (pp. 5637–5664). <https://proceedings.mlr.press/v139/koh21a.html>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. <https://doi.org/10.1145/3065386>
- Kuhn, L., Lyle, C., Gomez, A. N., Rothfuss, J., & Gal, Y. (2021). Robustness to pruning predicts generalization in deep neural networks. *ArXiv:2103.06002 [Cs, Stat]*. <http://arxiv.org/abs/2103.06002>
- Kwok, R. (2019). AI empowers conservation biology. *Nature*, 567(7746), 133–134. <https://doi.org/10.1038/d41586-019-00746-1>
- Lasseck, M. (2018). Audio-based bird species identification with deep convolutional neural networks. Working Notes of CLEF.
- Lauer, C. J., Montgomery, C. A., & Dietterich, T. G. (2020). Managing fragmented fire-threatened landscapes with spatial externalities. *Forest Science*, 66(4), 443–456.
- Le Guillarme, N., & Thuiller, W. (2022). TaxoNERD: Deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3), 625–641. <https://doi.org/10.1111/2041-210X.13778>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. <https://doi.org/10.1109/5.726791>
- Lederer, D. J., Bell, S. C., Branson, R. D., Chalmers, J. D., Marshall, R., Maslove, D. M., Ost, D. E., Punjabi, N. M., Schatz, M., Smyth, A. R., Stewart, P. W., Suissa, S., Adjei, A. A., Akdis, C. A., Azoulay, É., Bakker, J., Ballas, Z. K., Bardin, P. G., Barreiro, E., ... Vincent, J.-L. (2018). Control of confounding and reporting of results in causal inference studies. Guidance for authors from editors of respiratory, sleep, and critical care journals. *Annals of the American Thoracic Society*, 16(1), 22–28. <https://doi.org/10.1513/AnnalsATS.201808-564PS>
- Li, Z., Luo, Y., & Lyu, K. (2021). Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *ArXiv:2012.09839 [Cs, Stat]*. <http://arxiv.org/abs/2012.09839>
- Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, 18(8), 8. <https://doi.org/10.3390/s18082674>
- Liu, S., & Vicente, L. N. (2021). *The Sharpe predictor for fairness in machine learning*. <https://arxiv.org/abs/2108.06415v1>
- Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2), 3153–3160. <https://doi.org/10.1109/LRA.2020.2974682>
- Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, 90(4), e01422. <https://doi.org/10.1002/ecm.1422>
- Lürig, M. D., Donoughue, S., Svensson, E. I., Porto, A., & Tsuoboi, M. (2021). Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in Ecology and Evolution*, 9. <https://doi.org/10.3389/fevo.2021.642774>
- Mäder, P., Boho, D., Rzanny, M., Seeland, M., Wittich, H. C., Deggelmann, A., & Wäldchen, J. (2021). The Flora incognita app—Interactive plant species identification. *Methods in Ecology and Evolution*, 12(7), 1335–1342. <https://doi.org/10.1111/2041-210X.13611>
- Magliaresi, M. A., Blüthgen, N., Böhning-Gaese, K., & Schleuning, M. (2014). Morphological traits determine specialization and resource use in plant–hummingbird networks in the neotropics. *Ecology*, 95(12), 3325–3334. <https://doi.org/10.1890/13-2261.1>
- Martin, L. J., Blossey, B., & Ellis, E. (2012). Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment*, 10(4), 195–201. <https://doi.org/10.1890/110154>
- Masahiro, R., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, 8(11), e01976. <https://doi.org/10.1002/ecs2.1976>
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9), 1281–1289. <https://doi.org/10.1038/s41593-018-0209-y>
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133. <https://doi.org/10.1007/BF02478259>
- McIntire, E. J. B., Chubaty, A. M., Cumming, S. G., Andison, D., Barros, C., Boisvenue, C., Haché, S., Luo, Y., Micheletti, T., & Stewart, F. E. C. (2022). PERFECT: A Re-imagined foundation for predictive ecology. *Ecology Letters*, 25, 1345–1351. <https://doi.org/10.1111/ele.13994>
- Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(8), 1778–1790. <https://doi.org/10.1109/TGRS.2004.831865>
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087–1092. <https://doi.org/10.1063/1.1699114>
- Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, 19(8), 992–1006. <https://doi.org/10.1111/ele.12624>
- Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9), 1620–1633. <https://doi.org/10.1111/2041-210X.13650>
- Mignan, A., & Broccardo, M. (2019). One neuron versus deep learning in aftershock prediction. *Nature*, 574(7776), E1–E3. <https://doi.org/10.1038/s41586-019-1582-8>

- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- Moradi, S., Sheykhi Ilanloo, S., Kafash, A., & Yousefi, M. (2019). Identifying high-priority conservation areas for avian biodiversity using species distribution modeling. *Ecological Indicators*, 97, 159–164. <https://doi.org/10.1016/j.ecolind.2018.10.003>
- Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3), 247–259. <https://doi.org/10.1016/j.isprsjprs.2010.11.001>
- Muff, S., Nilsen, E. B., O'Hara, R. B., & Nater, C. R. (2022). Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*, 37(3), 203–210. <https://doi.org/10.1016/j.tree.2021.10.009>
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2019). Deep double descent: Where bigger models and more data hurt. *ArXiv:1912.02292 [Cs, Stat]*. <http://arxiv.org/abs/1912.02292>
- Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., ... Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3), e01370. <https://doi.org/10.1002/ecm.1370>
- Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, 12(1), 150–161. <https://doi.org/10.1111/2041-210X.13504>
- Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), E5716–E5725. <https://doi.org/10.1073/pnas.1719367115>
- Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., & Schoenholz, S. S. (2019). Neural tangents: Fast and easy infinite neural networks in python. *ArXiv:1912.02803 [Cs, Stat]*. <http://arxiv.org/abs/1912.02803>
- Ott, T., & Lautenschlager, U. (2021). *GinJinn2: Object detection and segmentation for ecology and evolution*. <https://doi.org/10.1101/2021.08.20.457033>
- Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3), 54–60. <https://doi.org/10.1145/3241036>
- Pearl, J. (2021). Radical empiricism and machine learning research. *Journal of Causal Inference*, 9(1), 78–82. <https://doi.org/10.1515/jci-2021-0006>
- Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, 5(9), 961–970. <https://doi.org/10.1111/2041-210X.12232>
- Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature Methods*, 16(1), 117–125. <https://doi.org/10.1038/s41592-018-0234-5>
- Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M., & Hartig, F. (2020). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution*, 11(2), 281–293. <https://doi.org/10.1111/2041-210X.13329>
- Pichler, M., & Hartig, F. (2022). Maximilian Pi/Pichler-and-Hartig-2022: Publication. Zenodo. <https://doi.org/10.5281/zenodo.7433226>
- Pimm, S. L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R., & Loarie, S. (2015). Emerging technologies to conserve biodiversity. *Trends in Ecology & Evolution*, 30(11), 685–696. <https://doi.org/10.1016/j.tree.2015.08.008>
- Poisot, T., Stouffer, D. B., & Gravel, D. (2015). Beyond species: Why ecological interaction networks vary through space and time. *Oikos*, 124(3), 243–251. <https://doi.org/10.1111/oik.01719>
- Pyšek, P., Richardson, D. M., Pergl, J., Jarošík, V., Sixtová, Z., & Weber, E. (2008). Geographical and taxonomic biases in invasion ecology. *Trends in Ecology & Evolution*, 23(5), 237–244. <https://doi.org/10.1016/j.tree.2008.02.002>
- Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., & Edelman, A. (2021). Universal differential equations for scientific machine learning. *ArXiv:2001.04385 [Cs, Math, q-Bio, Stat]*. <http://arxiv.org/abs/2001.04385>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.
- Rammer, W., & Seidl, R. (2019). A scalable model of vegetation transitions using deep neural networks. *Methods in Ecology and Evolution*, 10, 879–890. <https://doi.org/10.1111/2041-210X.13171>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Rew, J., Park, S., Cho, Y., Jung, S., & Hwang, E. (2019). Animal movement prediction based on predictive recurrent neural network. *Sensors*, 19(20), 20. <https://doi.org/10.3390/s19204411>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). <https://doi.org/10.1145/2939778.72.2939778>
- 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. <https://doi.org/10.1111/ecog.02881>
- Romero, M. P., Chang, Y.-M., Brunton, L. A., Prosser, A., Upton, P., Rees, E., Tearne, O., Arnold, M., Stevens, K., & Drewe, J. A. (2021). A comparison of the value of two machine learning predictive models to support bovine tuberculosis disease control in England. *Preventive Veterinary Medicine*, 188, 105264. <https://doi.org/10.1016/j.prevetmed.2021.105264>
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and Organization in the Brain. *Psychological Review*, 65, 65–386.
- Roy, A., Fablet, R., & Bertrand, S. L. (2022). Using generative adversarial networks (GAN) to simulate central-place foraging trajectories. *Methods in Ecology and Evolution*, 13(6), 1275–1287. <https://doi.org/10.1111/2041-210X.13853>
- Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, 44(2), 199–205. <https://doi.org/10.1111/ecog.05360>
- Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4), e1249. <https://doi.org/10.1002/widm.1249>
- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2), 197–227.
- Schölkopf, B. (2019). Causality for machine learning. *ArXiv:1911.10500 [Cs, Stat]*. <http://arxiv.org/abs/1911.10500>
- Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. <https://doi.org/10.1145/3381831>
- Scowen, M., Athanasiadis, I. N., Bullock, J. M., Eigenbrod, F., & Willcock, S. (2021). The current and future uses of machine learning in ecosystem service research. *Science of the Total Environment*, 799, 149263. <https://doi.org/10.1016/j.scitotenv.2021.149263>

- Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48), 30033–30038. <https://doi.org/10.1073/pnas.1907373117>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Shwartz-Ziv, R., & Alemi, A. A. (2020). Information in infinite ensembles of infinitely-wide neural networks. In *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference* (pp. 1–17). <https://proceedings.mlr.press/v118/shwartz-ziv20a.html>
- 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. <https://doi.org/10.1038/nature24270>
- Simpson, R., Williams, R., Ellis, R., & Culverhouse, P. F. (1992). Biological pattern recognition by neural networks. *Marine Ecology Progress Series*, 79(3), 303–308.
- Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., Edwards, T. C., Guala, G. F., Howard, T. G., Morisette, J. T., & Hamilton, H. (2019). Development and delivery of species distribution models to inform decision-making. *Bioscience*, 69(7), 544–557. <https://doi.org/10.1093/biosci/biz045>
- Sonnewald, M., Dutkiewicz, S., Hill, C., & Forget, G. (2020). Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces. *Science. Advances*, 6(22), eaay4740. <https://doi.org/10.1126/sciadv.aay4740>
- 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.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>
- Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2018). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, 10(3), 368–380. <https://doi.org/10.1111/2041-210X.13103>
- Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L., & Poisot, T. (2021). A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1837), 20210063. <https://doi.org/10.1098/rstb.2021.0063>
- Stupariu, M.-S., Cushman, S. A., Pleșoianu, A.-I., Pătru-Stupariu, I., & Fürst, C. (2022). Machine learning in landscape ecological analysis: A review of recent approaches. *Landscape Ecology*, 37(5), 1227–1250. <https://doi.org/10.1007/s10980-021-01366-9>
- Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., Vercauterent, K. C., Snow, N. P., Halseth, J. M., Salvo, P. A. D., Lewis, J. S., White, M. D., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., ... Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4), 585–590. <https://doi.org/10.1111/2041-210X.13120>
- Tank, A., Covert, I., Foti, N., Shojai, A., & Fox, E. (2021). Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 4267–4279. <https://doi.org/10.1109/TPAMI.2021.3065601>
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., Kohi, E. M., & Hopcraft, G. C. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10(6), 779–787. <https://doi.org/10.1111/2041-210X.13165>
- Trimble, M. J., & van Aarde, R. J. (2012). Geographical and taxonomic biases in research on biodiversity in human-modified landscapes. *Ecosphere*, 3(12), art119. <https://doi.org/10.1890/ES12-00299.1>
- Tseng, G., Kerner, H., & Rolnick, D. (2022). *TIML: Task-informed meta-learning for agriculture* (arXiv:2202.02124). arXiv. <https://doi.org/10.48550/arXiv.2202.02124>
- Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewart, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature. Communications*, 13(1), 1. <https://doi.org/10.1038/s41467-022-27980-y>
- Valle-Pérez, G., Camargo, C. Q., & Louís, A. A. (2019). Deep learning generalizes because the parameter-function map is biased towards simple functions. *ArXiv:1805.08522 [Cs, Stat]*. <http://arxiv.org/abs/1805.08522>
- Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The INaturalist species classification and detection dataset. [https://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Van\\_Horn\\_The\\_INaturalist\\_Species\\_CVPR\\_2018\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2018/html/Van_Horn_The_INaturalist_Species_CVPR_2018_paper.html)
- Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774–780.
- Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, 15(11), e1002689. <https://doi.org/10.1371/journal.pmed.1002689>
- Veit, A., Wilber, M. J., & Belongie, S. (2016). Residual networks behave like ensembles of relatively shallow networks. *Advances in Neural Information Processing Systems*, 29. <https://proceedings.neurips.cc/paper/2016/hash/37bc2f75bf1bcfe8450a1a41c200364c-Abstract.html>
- Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., ... Silver, D. (2019). Grandmaster level in star craft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354. <https://doi.org/10.1038/s41586-019-1724-z>
- Voznica, J., Zhukova, A., Boskova, V., Saulnier, E., Lemoine, F., Moslonka-Lefebvre, M., & Gascuel, O. (2022). Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1). <https://doi.org/10.1038/s41467-022-31511-0>
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, 9(11), 2216–2225. <https://doi.org/10.1111/2041-210X.13075>
- Wang, D., & Gu, J. (2018). VASC: Dimension reduction and visualization of single-cell RNA-seq data by deep Variational autoencoder. *Genomics, Proteomics & Bioinformatics*, 16(5), 320–331. <https://doi.org/10.1016/j.gpb.2018.08.003>
- Wang, S., Fan, K., Luo, N., Cao, Y., Wu, F., Zhang, C., Heller, K. A., & You, L. (2019). Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nature Communications*, 10(1), 1–9. <https://doi.org/10.1038/s41467-019-12342-y>
- Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3), 1–34.
- Wardeh, M., Baylis, M., & Blagrove, M. S. C. (2021). Predicting mammalian hosts in which novel coronaviruses can be generated. *Nature*

- Communications*, 12(1), 1. <https://doi.org/10.1038/s41467-021-21034-5>
- Wearn, O. R., Freeman, R., & Jacoby, D. M. P. (2019). Responsible AI for conservation. *Nature Machine Intelligence*, 1(2), 72–73. <https://doi.org/10.1038/s42256-019-0022-7>
- Wein, S., Malloni, W. M., Tomé, A. M., Frank, S. M., Henze, G.-I., Wüst, S., Greenlee, M. W., & Lang, E. W. (2021). A graph neural network framework for causal inference in brain networks. *Scientific Reports*, 11(1), 8061. <https://doi.org/10.1038/s41598-021-87411-8>
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1), 9.
- Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J., & Dormann, C. F. (2022). Process-guidance improves predictive performance of neural networks for carbon turnover in ecosystems (arXiv:2209.14229). arXiv. doi:[10.48550/arXiv.2209.14229](https://doi.org/10.48550/arXiv.2209.14229)
- Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence-absence data. *Methods in Ecology and Evolution*, 10(2), 198–211. <https://doi.org/10.1111/2041-210X.13106>
- Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, 10(1), 80–91. <https://doi.org/10.1111/2041-210X.13099>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Davison, J. (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv Preprint ArXiv:1910.03771.
- Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In *Proceedings of the 37th international conference on machine learning* (pp. 10767–10777). <https://proceedings.mlr.press/v119/yang20j.html>
- Yu, Q., Ji, W., Prihodko, L., Ross, C. W., Anchang, J. Y., & Hanan, N. P. (2021). Study becomes insight: Ecological learning from machine learning. *Methods in Ecology and Evolution*, 12(11), 2117–2128. <https://doi.org/10.1111/2041-210X.13686>
- Zečević, M., Dhami, D. S., Veličković, P., & Kersting, K. (2021). Relating graph neural networks to structural causal models. ArXiv:2109e.04173 [Cs, Stat]. <http://arxiv.org/abs/2109.04173>
- Zerrouki, Y., Harrou, F., Zerrouki, N., Dairi, A., & Sun, Y. (2021). Desertification detection using an improved Variational autoencoder-based approach through ETM-Landsat satellite data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 202–213. <https://doi.org/10.1109/JSTARS.2020.3042760>
- Zhang, J., Xu, Y., Zhan, T., Wu, Z., & Wei, Z. (2021). Anomaly detection in hyperspectral image using 3D-convolutional Variational auto-encoder. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS* (pp. 2512–2515). <https://doi.org/10.1109/IGARSS47720.2021.9554184>
- Zhang, S., Wang, M., Liu, S., Chen, P.-Y., & Xiong, J. (2021). Why lottery ticket wins? A theoretical perspective of sample complexity on pruned neural networks. *ArXiv Preprint ArXiv:2110.05667*
- Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1), 272–281. <https://doi.org/10.1080/07350015.2019.1624293>
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76. <https://doi.org/10.1109/JPROC.2020.3004555>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.
- Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—It's time to make it fair. *Nature*, 559(7714), 324–326. <https://doi.org/10.1038/d41586-018-05707-8>

## SUPPORTING INFORMATION

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

### Appendix S1. Supporting Information.

**How to cite this article:** Pichler, M., & Hartig, F. (2023). Machine learning and deep learning—A review for ecologists. *Methods in Ecology and Evolution*, 14, 994–1016. <https://doi.org/10.1111/2041-210X.14061>