Machine learning and deep learning—A review for ecologists

Maximilian Pichler | Florian Hartig

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.

Keywords

artificial intelligence, big data, causal inference, deep learning, machine learning
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 modelling (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
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.
loss function that is not necessarily tied to the probability of the observa-
tions (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 nec-
essary 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 estimat-
ing CIs on the parameters and predictions of statistical or ML mod-
els. 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 mod-
els trained on GPUs with hundreds of millions parameters dominate the

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).

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 embedding (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.
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. Guatieri & 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 (Belgiu & Drăguţ, 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
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; Arora et al., 2019; Huh et al., 2021; 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 regularization 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.
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...
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; Måder 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; Mader 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 (Mctire 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. Mader 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 Maglianesi 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 models and avoiding their misuse or misinterpretation (e.g. Barredo Arrieta et al., 2016; 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).
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).

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).
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., 2020) 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 CO₂ 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., Rettelbach, Daniel, A., Ryan, S. J., Seifert, S., … Carlson, C. J. (2021). Bat detective—Deep learning tools for bat acoustic signal detection. *Remote Sensing*, 13(3), e1093085. https://doi.org/10.1017/journal.pone.0193085

Bartoldsson, B. R., Morcos, A. S., Barbu, A., & Erlebacher, G. (2020). The generalization-stability tradeoff in neural network pruning. *ArXiv*:1906.03728 [Stat]. http://arxiv.org/abs/1906.03728

Becker, D. J., Albery, G. F., Sjödin, A., 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

Beery, S., Cole, E., Parker, J., Perona, P., & Winner, K. (2021). Species distribution modelling for machine learning practitioners: A review. *ACM SIGCAS Conference on Computing and Sustainable Societies*, 32–348. https://doi.org/10.1145/3460123.3471966

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., Radath, V., Votel, R., & Huang, J. (2020). Context R-CNN: Long term temporal context for per-camera object detection. https://openaccess.thecvf.com/content_ICCV_2020/html/Beery_Context_R-CNN_Long_Temporal_Context_for_Per-Camera_Object_Detection_ICCV_2020_paper.html

Belgu, M., & Dra guţ, 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 [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 [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.
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., Kalwekt, 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

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., Veldhuis, 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., Daïri, 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 autoencoder. 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