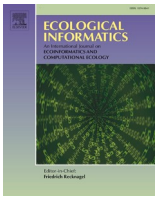
Ecological Informatics 72 (2022) 101896

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



# Embracing imperfection: Machine-assisted invertebrate classification in real-world datasets

Jarrett Blair [a,*], Michael D. Weiser [b], Kirsten de Beurs [b], Michael Kaspari [b], Cameron Siler [b,c], Katie E. Marshall [a]

[a] Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[b] Department of Biology, University of Oklahoma, Norman, OK 73019-0235, USA
[c] Sam Noble Oklahoma Museum of Natural History, University of Oklahoma, 2401 Chautauqua Ave., Norman, OK 73072-7029, USA

ARTICLE INFO

*Keywords:*
Machine learning
Computer vision
Image classification
Macroecology
Terrestrial invertebrates

ABSTRACT

Despite growing concerns over the health of global invertebrate diversity, terrestrial invertebrate monitoring efforts remain poorly geographically distributed. Machine-assisted classification has been proposed as a potential solution to quickly gather large amounts of data; however, previous studies have often used unrealistic or idealized datasets to train and test their models.

In this study, we describe a practical methodology for including machine learning in ecological data acquisition pipelines. Here we train and test machine learning algorithms to classify over 72,000 terrestrial invertebrate specimens from morphometric data and contextual metadata. All vouchered specimens were collected in pitfall traps by the National Ecological Observatory Network (NEON) at 45 locations across the United States from 2016 to 2019. Specimens were photographed, and two separate machine learning paradigms were used to classify them. In the first, we used a convolutional neural network (ResNet-50), and in the second, we extracted morphometric data as feature vectors using ImageJ and used traditional machine learning methods to classify specimens. Issues stemming from inconsistent taxonomic label specificity were resolved by making classifications at the lowest identified taxonomic level (LITL). Taxa with too few specimens to be included in the training dataset were classified by the model using zero-shot classification.

When classifying specimens that were known and seen by our models, we reached a maximum accuracy of 72.7% using eXtreme Gradient Boosting (XGBoost) at the LITL. This nearly matched the maximum accuracy achieved by the CNN of 72.8% at the LITL. Models that were trained without contextual metadata underperformed models with contextual metadata. We also classified invertebrate taxa that were unknown to the model using zero-shot classification, reaching a maximum accuracy of 65.5% when using the ResNet-50, compared to 39.4% when using XGBoost.

The general methodology outlined here represents a realistic application of machine learning as a tool for ecological studies. We found that more advanced and complex machine learning methods such as convolutional neural networks are not necessarily more accurate than traditional machine learning methods. Hierarchical and LITL classifications allow for flexible taxonomic specificity at the input and output layers. These methods also help address the 'long tail' problem of underrepresented taxa missed by machine learning models. Finally, we encourage researchers to consider more than just morphometric data when training their models, as we have shown that the inclusion of contextual metadata can provide significant improvements to accuracy.

## 1. Introduction

Several recent studies have suggested that terrestrial invertebrates may be suffering drastic population and diversity losses (Dirzo et al., 2014; Welti et al., 2020; Wepprich et al., 2019). However, these losses are not distributed equally across the planet nor across taxonomic diversity (Guzman et al., 2021; van Klink et al., 2020). Generalizations about trends in global invertebrate diversity and abundance require a solid data foundation, yet invertebrates remain significantly poorly-sampled relative to their diversity and abundance (Høye et al., 2021;

* Corresponding author at: Departments of Zoology, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada.
*E-mail address:* blair@zoology.ubc.ca (J. Blair).

https://doi.org/10.1016/j.ecoinf.2022.101896
Received 22 March 2022; Received in revised form 31 October 2022; Accepted 1 November 2022
Available online 5 November 2022

van Klink et al., 2020). While there are many invertebrate biomonitoring programs operating around the world, the data collected from these programs are often spatiotemporally coarse and lack taxonomic diversity (Høye et al., 2021). Large-scale invertebrate biomonitoring efforts require vast funding and human resources, reducing their accessibility to developing nations and further limiting the geographic scope of biodiversity data (Karlsson et al., 2020; van Klink et al., 2020). One of the most time and resource intensive aspects of any invertebrate biomonitoring program relates to specimen sorting and taxonomic identification (Karlsson et al., 2020). While these processes are traditionally performed by trained taxonomists, often in coordination with parataxonomists (i.e. individuals who are not expertly-trained taxonomists; Krell, 2004), advances in machine learning and computer vision have made machine-assisted classification a possibility for many groups of insects (Ärje et al., 2020; Blair et al., 2020; Marques et al., 2018; Mayo and Watson, 2007). Not only does machine-assisted classification have the potential to increase the efficiency, output, and accessibility of biomonitoring programs (Peters et al., 2014; Thessen, 2016), but also, in a time when more data on invertebrate diversity and abundance is desperately needed, such an approach offers a transformative solution to the way we monitor global invertebrate diversity in a changing climate.

Excitement regarding the possibilities of machine learning for taxonomic classification has led to many publications on the topic over the last decade. However, a common theme in this literature is the use of unrealistic or idealized case studies. Specifically, many datasets used in these studies have low species richness, uniform taxonomic resolution, and no geographic/temporal component (Ärje et al., 2020; Blair et al., 2020; Joutsijoki et al., 2014). Additionally, nearly all studies that have used large ecological datasets accept that low abundance taxa must be sacrificed for the benefit of model performance (Marques et al., 2018; Mayo and Watson, 2007). In most natural communities, most taxonomic diversity is comprised of low-abundance taxa (i.e. the "long tail" of rank abundance curves; Preston, 1948; Verberk, 2012; Whittaker, 1965), so simply ignoring them is not feasible if machine learning tools are to be widely used in ecological studies. We note that the previously-mentioned machine learning studies still provide foundational knowledge about machine learning's potential uses for taxonomic classification. However, significant work remains to improve the practicality of computer vision for specimen classification if it is to be integrated into ecological data acquisition pipelines.

While there have been some promising results developing models for classifying biological specimens (Blair et al., 2020; Ding and Taylor, 2016; Spiesman et al., 2021; Van Horn et al., 2018), ecological datasets pose several challenges for machine learning. One challenge that is effectively universal among macroecological datasets are taxa that are 'unknown' to the model. These are taxa that the model was not trained to classify, but that the model may need to classify in the future. Most commonly, these taxa will be hidden in the aforementioned 'long tail' of uncommon species that are excluded from model training due to their low prevalence. For machine learning to be a practical solution for use in ecological studies, this problem must be addressed.

Another challenge facing the classification of ecological datasets (especially large, diverse ones) is a 'ragged edge' of taxonomic specification: in many ecological datasets, identifications are made at varied taxonomic resolution dependent on the taxonomist's knowledge (Jansen et al., 2018; Schmidt-Kloiber and Nijboer, 2004). For example, some specimens may be identified down to species, while others are identified to a higher taxonomic level only (e.g. family, order, etc.). This then poses the question of what a 'correct' classification is, and machine learning algorithms may need to weigh the benefits of accuracy versus specificity (Deng et al., 2012). Here we explored two potential ways to address this problem: lowest identified taxonomic level (LITL) models and single-level models. Single-level models assign labels to specimens and measure accuracy all at one specified taxonomic level, while LITL models assign labels and measure accuracy based on the LITL label assigned to each individual. LITL labels allow for maximum inclusion and specificity, but result in nested labels (i.e. labels belonging to related taxa at different taxonomic levels, such as Formicidae and Insecta; Fig. 1). Single-level models prevent nested labels by classifying all individuals at a single taxonomic level but can result in excluded individuals that were labelled at higher taxonomic levels in the training dataset. Additionally, single-level models restrict taxonomic specificity as classification cannot be made below the specified taxonomic level. From an ecologist's perspective, there is likely no definitive "best" solution as the importance of taxonomic specificity will vary depending on the use case.

Despite the challenges ecological datasets pose, they also provide many unique opportunities for improving machine learning performance. For example, they often have contextual metadata about the collection site (e.g., location, time, temperature, etc.) that can be used to inform computer vision. This contextual metadata can significantly improve accuracy, as seen in an accuracy increase of 9.1% (48.1% ➔ 57.3%) when classifying ladybeetles (Coleoptera: Coccinellidae; Terry et al., 2020), an 11.2% increase (68.6% ➔ 79.8%) when classifying North American birds (Berg et al., 2014), and a significant improvement when classifying marine plankton (Ellen et al., 2019). Additionally, the hierarchical structure of taxonomies can be used to perform 'zero-shot classification', which allows unknown and unseen classes to be classified by making predictions at the higher taxonomic rank that is known by the model (Blair et al., 2020; Deng et al., 2012). Deng et al. (2012) shows that while 'flat' classifiers (i.e. single-level classifiers) have zero accuracy and zero information gain when classifying unknown and unseen classes, hierarchical classifiers can often predict these classes accurately. Strategies such as these may greatly increase the practicality of incorporating computer vision methods in large scale ecology projects.

One such large scale ecology project is the National Ecological Observatory Network (NEON), which collects open access ecological data at a continental scale in the United States (Keller et al., 2008). Part of NEON's ecological monitoring involves collecting and processing thousands of invertebrate specimens every year from arrays of pitfall traps that are collected and pooled every 14 days (Hoekman et al., 2017; Thorpe et al., 2016) (Fig. S1). NEON's current workflow for these samples is to have a local parataxonomist separate and identify all ground beetles (Family: Carabidae), while the remaining unsorted invertebrate bycatch is stored in a biorepository. While NEON's decision to focus on carabids is understandable given the human resources it would require to properly sort the invertebrate bycatch, it also means they are sitting on a large and as-of-yet untapped source of North American terrestrial invertebrate data. This makes NEON's invertebrate bycatch a prime target for machine assisted specimen classification so that researchers in a wide variety of disciplines can make use of this extensive data source.

Here we apply machine learning and deep learning classification methods to the large terrestrial invertebrate dataset collected by NEON. Our primary goals are to address the challenges taxonomic classification poses for machine learning while also taking advantage of the benefits provided by ecological datasets. We also compare the advantages and drawbacks of traditional machine learning algorithms (e.g. k-nearest neighbour, decision trees) and deep learning algorithms (e.g. convolutional neural networks (CNNs)). We did this by training a variety of algorithm types on several configurations of the NEON dataset to determine the optimal algorithm and dataset configuration. Using the NEON dataset, we ask: (1) How does the inclusion of contextual metadata impact model performance and what is its relative importance compared to morphometric data? (2) How can we improve classification in the real-world conditions of low-abundance taxa and inconsistent taxonomic resolution? (3) Are deep learning algorithms universally superior to traditional machine learning algorithms, or are there merits to each? Here we show that the inclusion of contextual metadata improves classification performance while unevenness in label representation is relatively unimportant. We also compare the performance and practicality of LITL vs a single-level (order) model and use zero-shot

A

| # | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|
| 1 | Annelida | -- | -- | -- | -- | -- |
| 2 | Arthropoda | Insecta | Coleoptera | Staphylinidae | -- | -- |
| 3 | Arthropoda | Insecta | Hymenoptera | Formicidae | *Formica* | -- |
| 4 | Mollusca | Gastropoda | Stylommatophora | Zonitidae | *Ventridens* | *ligera* |
| 5 | Arthropoda | Insecta | Coleoptera | -- | -- | -- |

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| Annelida | Staphylinidae | *Formica* | *ligera* | Coleoptera |

B

| # | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|
| 1 | Annelida | -- | -- | -- | -- | -- |
| 2 | Arthropoda | Insecta | Coleoptera | Staphylinidae | -- | -- |
| 3 | Arthropoda | Insecta | Hymenoptera | Formicidae | *Formica* | -- |
| 4 | Mollusca | Gastropoda | Stylommatophora | Zonitidae | *Ventridens* | *ligera* |
| 5 | Arthropoda | Insecta | Coleoptera | -- | -- | -- |

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| -- (zero-shot) | Coleoptera | Hymenoptera | Stylommatophora | Coleoptera |

**Fig. 1.** Lowest Identified Taxonomic Level (LITL) and order-level labelling methods. (A) The top table shows the taxonomic names (from phylum to species) for five specimens. Cells filled with "–" indicate that our labeller was unable to make a classification for that specimen at that taxonomic level. The cells highlighted in green show which label was chosen for each specimen, which are summarized in the bottom table. Labels were chosen based on the LITL for each specimen. (B) The methodology is the same as (A), except that all labels were recorded at the order-level, regardless of the LITL identified by our labeller. No order-level label was given to specimen #1, so it can only be identified using zero-shot classification. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

classification to overcome the 'long tail' problem (See Table 2 for a glossary of terms). Finally, we show that traditional machine learning algorithms can outperform deep learning algorithms, even when given less data. Taken together, we outline potential approaches for the challenges of machine-assisted classification in "real world" ecological datasets.

## 2. Methods

### 2.1. Invertebrate bycatch collection

We used terrestrial invertebrates collected from pitfall traps by the National Ecological Observatory Network (NEON, see Hoekman et al., 2017 for collection methodology). All specimens were collected from 2016 to 2019. NEON sets the pitfall traps to collect ground beetles (Coleoptera: Carabidae) and separates them from the rest of the invertebrates immediately, so the ground beetles are not considered here (but see Blair et al., 2020).

**Table 2**
Classification algorithms and their respective parameter settings used for classifying the NEON invertebrate bycatch.

| Algorithm Name | Parameter settings |
|---|---|
| Linear Discriminant Analysis | No tuning parameters to optimize. |
| Naïve Bayes | No tuning parameters to optimize. |
| K-Nearest Neighbours | *k* values of all integers between 1 and 25 were tested, and the *k* value with the lowest error in the training dataset was selected. |
| eXtreme Gradient Boosting | max.*depth* was set as 10, *eta* was set at 0.1, and *nrounds* was set at 120. |
| Artificial Neural Network | 1 hidden dense layer with 128 nodes and ReLU activation function.<br>Optimizer: Adam<br>Loss function: Categorical cross-entropy<br>Epochs: 10 |
| Convolutional Neural Network | Base model: ResNet50 with 'imagenet' weights (He et al., 2016)<br>2 hidden dense layers (1024 and 128 nodes, respectively), both with batch normalization, 0.3 dropout, and ReLU activation function.<br>Optimizer: Adam<br>Loss function: Categorical cross-entropy<br>Epochs: 10 |

### 2.2. Imaging

We made bulk digital images of the invertebrates (mostly Arthropoda, Mollusca and Annelida) at a resolution of 729 pixels per $mm^2$ against a white background (for complete methods, see Weiser et al., 2021; Fig. S1). Using the FIJI implementation of ImageJ (Schindelin et al., 2012), for each individual organism we extracted 21 morphological measures (e.g., major and minor axis, perimeter, image area) and eight statistics (e.g., mean, skew, and kurtosis) for the distribution of values from each of the three RGB (Red, Green, and Blue) colour layers.

### 2.3. Contextual metadata

Here, we define contextual metadata as any non-morphometric and non-taxonomic data included in the invertebrate dataset. We sourced our contextual metadata from NEON (spatiotemporal data such as location and elevation) (NEON, 2022), NASA's Land Processes Distributed Active Archive Center (LP DAAC; gross primary productivity and evapotranspiration) (Running et al., 2015, 2017), and the Daymet V4 dataset (Thornton et al., 2021). A complete list of all variables used and their definitions can be found in this table (Table S1). All contextual metadata values were the same for all specimens collected in a given trap event. Between our morphometric data and contextual metadata, 49 descriptive variables were used as our training data (36 morphometric variables, 13 contextual variables).

### 2.4. Machine learning

#### 2.4.1. Preparation

To prepare the data for machine learning training and testing, we first removed all taxa with fewer than 100 observations from the dataset (Fig. 2), as these taxa contained too little data to properly train the models. We then randomly split the dataset in training and testing datasets at a ratio of 70:30 (Fig. S2). Each dataset was randomly shuffled to prevent overfitting. We normalized the feature vector data by centering and scaling each predictor variable such that they all had a mean of 0 and standard deviation of 1. In the image dataset, all images were scaled to 224 × 224 pixels and randomly augmented (i.e. sheared, zoomed, flipped horizontally and vertically).

#### 2.4.2. Algorithm types and architecture

After pre-processing the data, it was ready to be used for training in a machine learning model. We contrasted four types of 'traditional' machine learning algorithms: K-nearest neighbours (KNN; Cover and Hart, 1967), linear discriminant analysis (LDA; Mika et al., 1999), naïve Bayes (NB; Mika et al., 1999), and eXtreme Gradient Boosting (XGBoost; Chen and He, 2015). We chose these four algorithms as they use a wide range of machine learning techniques. Our methods for optimizing these algorithms are described in Table 1 and Fig. S3. All traditional machine learning methods were developed using R version 4.1.1 (R Core Team, 2021).

We also trained artificial neural network (ANN) models using Python version 3.9.12. Unlike the traditional machine learning models, which only took input data in the form of feature vectors, we designed our ANNs to take data as feature vector-only, image-only, or image and

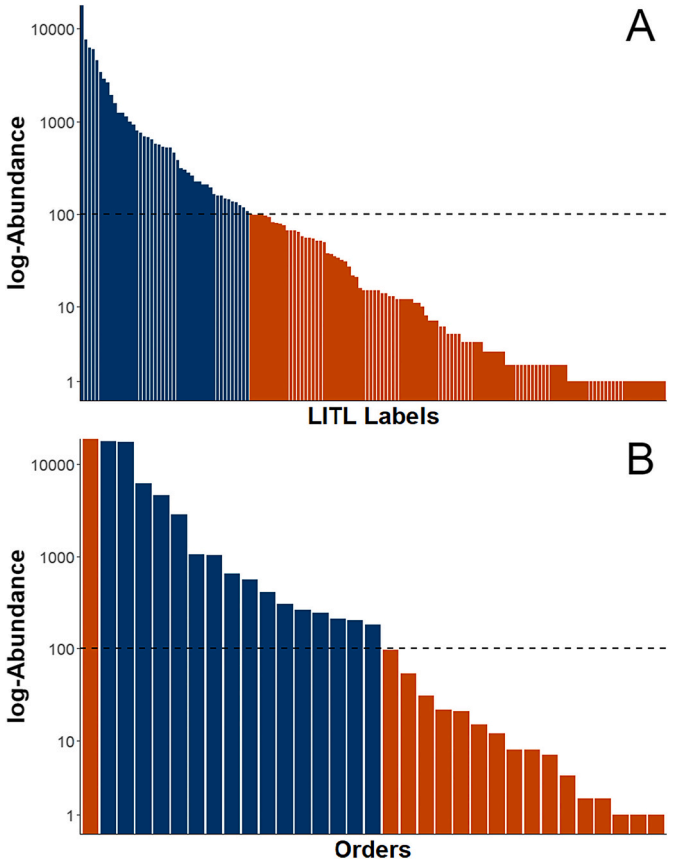



**Fig. 2.** Log-scale rank abundance plots of invertebrates in the NEON dataset. (A) Each bar represents a lowest identified taxonomic label (LITL), as identified by MDW. Blue bars represent groups that had ≥100 individuals and were included in the training dataset; orange bars represent groups with <100 individuals and were removed from the training dataset. A cut off of 100 individuals (represented by a horizontal line) was used to ensure the models had sufficient data to classify each LITL label. (B) Each bar represents a taxonomic order of invertebrates, as identified by MDW. Blue bars represent groups that had ≥100 individuals and were included in the training dataset; orange bars represent groups with <100 individuals or had a LITL above the level of order (e.g. class or phylum) and therefore were removed from the training dataset. A cut off of 100 individuals (horizontal line) was used to ensure the models had sufficient data to classify each LITL label. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

**Table 1**
Glossary of terms.

| Term | Definition |
|---|---|
| 'Long tail' | Refers to the 'long tail' of rank abundance curves in which most taxa have low abundance compared to the relatively few high abundance taxa. |
| Label | A categorical label given to a specimen (e.g. Formicidae, *Canthon viridis*, etc.). Synonymous with the machine learning definition of 'class'. |
| (Un)known & (Un)seen | Known labels are those that were input into the model. Unknown labels are those that were not input into the model. For example, a model trained using the labels "Cat", "Dog", and "Bird" knows the labels "Cat", "Dog", and "Bird" but does not know the label "Pigeon".<br>Seen labels are those that are represented in the model via data. Seen labels may be known by the model, but it is not required. Using the same example as above, if images of pigeons were used to train the model to recognize birds, pigeons would be seen but not known by the model. If images of pigeons were not in the training dataset, pigeons would be unseen and unknown. |
| Lowest Identified Taxonomic Level (LITL) | A specimen's LITL is the level at which the specimen's most specific label is assigned. E.g. the LITL of a specimen labelled *Canthon* sp. is genus and it's LITL label would be *Canthon*. |
| LITL Model | A LITL model is a machine learning model that is trained using LITL labels, usually consisting of multiple LITLs. |
| Contextual Metadata | Non-morphometric data that provides context about the time, location, and environment a specimen was collected in. |
| Zero Shot Classification | The classification of unknown and unseen specimens by making classifications at a level where the specimen's label is known. E.g. if the beetle family Elateridae is unknown and unseen by the model, but the order Coleoptera (beetles) is known and seen by the model, Elateridae specimens can be zero shot classified at the order level. |
| Hierarchical Classification | Hierarchical classification can infer labels at higher levels from a single, base classification. These classifications can be used to calculate hierarchical accuracy at multiple levels. Zero shot classification is an example of hierarchical classification. |
| Training and Testing Datasets | The training dataset is used as the input layer of the machine learning model. It is used as the reference data to train the model how to classify each label. The testing dataset is composed of entirely separate specimens from the training dataset and is used to measure a trained model's performance metrics (Fig. S3). |
| Comprehensive Zero Shot Accuracy | Zero shot accuracy using all zero shot specimens at every taxonomic level, regardless of if they had a known taxonomic label for any given level. |
| Limited Zero Shot Accuracy | Zero shot accuracy measured only using specimens that had a known label at any given taxonomic level. |
| Top 1 Accuracy | Accuracy measured using only the label with the highest probability as measured by the model. |
| Top 3 Accuracy | Accuracy measured using the labels with the three highest probabilities as measured by the model. If any of the three labels match the actual label, the classification is deemed correct. |
| F1 score | When referring to a model's F1 score (i.e. macro F1 score): $macroF1 = \frac{1}{n}\sum_{i=1}^{n}\frac{True\ Positives_n}{True\ Positives_n + 0.5(False\ Positives_n + False\ Negatives_n)}$<br>When referring to a label's F1 score (i.e. micro F1 score): $microF1 = \frac{True\ Positives}{True\ Positives + 0.5(False\ Positives + False\ Negatives)}$ |
| Precision & Recall | $Precision = \frac{True\ Positives}{True\ Postivies + False\ Positives}$<br>$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$ |

feature vector data simultaneously. We first trained and tested our single-input-type models independently to optimize their training parameters, then combined them in our multi-input model. Our feature vector model was constructed using a single dense layer and softmax classifier layer. For the image model, we trained a CNN using the ResNet-50 architecture for feature extraction (He et al., 2016). We then added one global average pooling layer, two dense layers, and a softmax classifier layer. To combine the feature vector and image models, we removed the classifier layers from each, concatenated their output layers, added one dense layer and one softmax classifier layer (Fig. S4). The model was trained over 10 epochs, the point at which loss was minimized (Fig. S3).

All machine learning and deep learning code is available on GitHub (Blair, 2022).

#### 2.4.3. LITL & order-level labels

Due to the uneven taxonomic resolution of our dataset (i.e. some individuals were classified to lower taxonomic levels than others), specimens in the training dataset were labelled at their lowest identified taxonomic level (LITL; Fig. 1a). We treated nested taxonomic labels (e.g. Staphylinidae and Coleoptera) as non-equivalent and mutually exclusive. For example, if a specimen with a LITL label of Staphylinidae is classified as Coleoptera by a model, or vice-versa, that classification would be deemed incorrect despite Staphylinidae being a family within Coleoptera. We also created a separate training dataset in which all specimens were labelled at the order level (Fig. 1b). Any specimens with an LITL label below the order level were relabelled as their corresponding order (e.g. Staphylinidae would be relabelled as Coleoptera). Conversely, any specimens with an LITL label above order level were removed from the training dataset. These datasets were pre-processed separately from the LITL datasets and included orders with 100 or more individuals in the NEON dataset. Performance of the LITL and order-level models were measured using accuracy and F1 score.

#### 2.4.4. Contextual metadata

To determine the effects and importance of contextual metadata, we trained and tested models of each algorithm type using the LITL and order-level datasets that contained both contextual metadata and morphometric data as well as with datasets that only contained morphometric data. Differences in performance were measured as the net change in accuracy, and variable importance was measured using 'mean decrease accuracy' in the XGBoost model.

#### 2.4.5. Zero-shot classification

We performed zero-shot classification by taking taxa that had too few specimens to be included in our training datasets ("unseen" taxa) and classifying them at taxonomic levels where they belonged to a common group that was included in the training dataset ("known" taxa). For example, there were only 92 specimens from the family Elateridae (click beetles) in our LITL dataset, and thus this family was not included in the training dataset. This makes the label "Elateridae" unknown and unseen by the models, which means that the models cannot classify this label at the family level. However, Elateridae belongs to the order Coleoptera, which is a label known by the model. The hierarchical structure of taxonomy allows us to make classifications at multiple taxonomic levels simultaneously, as long as these levels are not more specific than the original classification (e.g. you cannot infer species from a genus level classification). This allowed us to make a zero-shot classification for Elateridae at the order level and above by classifying them as Coleoptera. When measuring zero-shot accuracy, any classification made of a label belonging to the known group would be deemed correct. For example, if "Elateridae" was classified as "Staphylinidae" instead of "Coleoptera", it would still be considered accurate at the order level despite "Staphylinidae" being treated as mutually exclusive from "Coleoptera" in the LITL models. We note that in practical situations, uncommon groups could be added to the training dataset by labelling them at their known taxonomic level, but they were left out of our training dataset so we could measure zero-shot classification performance. We also performed zero-shot classification using the order level models by classifying unincluded taxa at the class and phylum levels.

We also used zero-shot accuracy to estimate LITL and order-level accuracy across the entire invertebrate dataset. We weighted the accuracies measured from the testing datasets to the training datasets to get an estimated accuracy for all 'known' specimens (70,466 LITL specimens and 53,683 order-level specimens). We then combined this estimated accuracy with our zero-shot accuracies to get an estimated accuracy for the entire dataset.

We measured the taxonomic specificity of the LITL and order level models by averaging the taxonomic level of each label in the training and testing datasets, as well as the first known levels for zero shot specimens. Six taxonomic levels (species, genus, family, order, class, and phylum) were used. Each taxonomic level was assigned a numerical value in ascending order (i.e. species = 1, phylum = 6). If the taxonomic level of a label was between one of the six measured levels, its value was rounded up (e.g. superfamily labels were measured as order level). Specimens labelled as "Ignore", "Larva", "Nymph", and "Juvenile" were not assigned a numerical value and were removed from the measurement.

## 3. Results

### 3.1. LITL

Our entire NEON invertebrate dataset contained 72,678 specimens with 160 LITLs and 30 taxonomic orders, sampled across 45 sites and 323 sampling events (i.e. 323 pitfall trap collections) (Fig. 2). After we removed uncommon LITL labels (<100 individuals per label), our training datasets contained 49,337 specimens across 46 LITL labels while our testing datasets contained 21,129 specimens (Fig. 2). The 2212 uncommon specimens we removed from the training and testing datasets were used for zero-shot classifications.

We found models that contained contextual metadata always performed better than their non-metadata counterparts on all metrics across all algorithms when predicting specimens down to their LITL (Fig. 3). Of the 'traditional' models we tested, XGBoost always performed best, with an average top-1 accuracy of 72.7%, top-3 accuracy of 89.8%, and F1 score of 0.661 when including metadata. XGBoost also had the greatest accuracy increase between the metadata and no-metadata configurations (+10.2%; 62.5%➔ 72.7%). The average accuracy boost when we included metadata in the training dataset was 8.7% across all models. We also found the F1 score of XGBoost was improved the most when metadata was added (+0.173), with an average increase of 0.132 across all models. Latitude and longitude had the highest importance (mean decrease in accuracy) among metadata variables for the XGB models, ranking sixth and seventh out of 46 predictor variables respectively (Fig. 4).

Of our artificial neural network configurations, the CNN with all three data types (images, morphometric data, and contextual metadata) performed best across all metrics (top-1 accuracy 72.8%, top-3 accuracy 91.2%, F1 score 0.623) (Fig. 5). Conversely, the metadata-only model performed the worst with a top-1 accuracy of 36.9% (Fig. 5). Overall, XGBoost and the best performing CNN returned similar results. XGBoost had relatively higher top-1 accuracy and F1 score (+0.7% and + 0.047), while the CNN had higher top 3 accuracy (+0.6%). XGBoost also performed slightly better on taxa with low abundance (Fig. S5, Fig. S6), but neither had a significant difference in F1 score relative to abundance (XGBoost: $F = 1.016$, $df = 1, 44$, $p > 0.05$, $R^2 = 0.023$; CNN: $F = 1.716$, $df = 44, 1$, $p > 0.05$, $R^2 = 0.038$; Fig. S5).

We performed zero-shot classification on unseen and unknown specimens by classifying them as known classes at higher taxonomic levels using XGBoost with metadata, our best-performing model. When classified to the lowest possible level, we found XGBoost zero shot

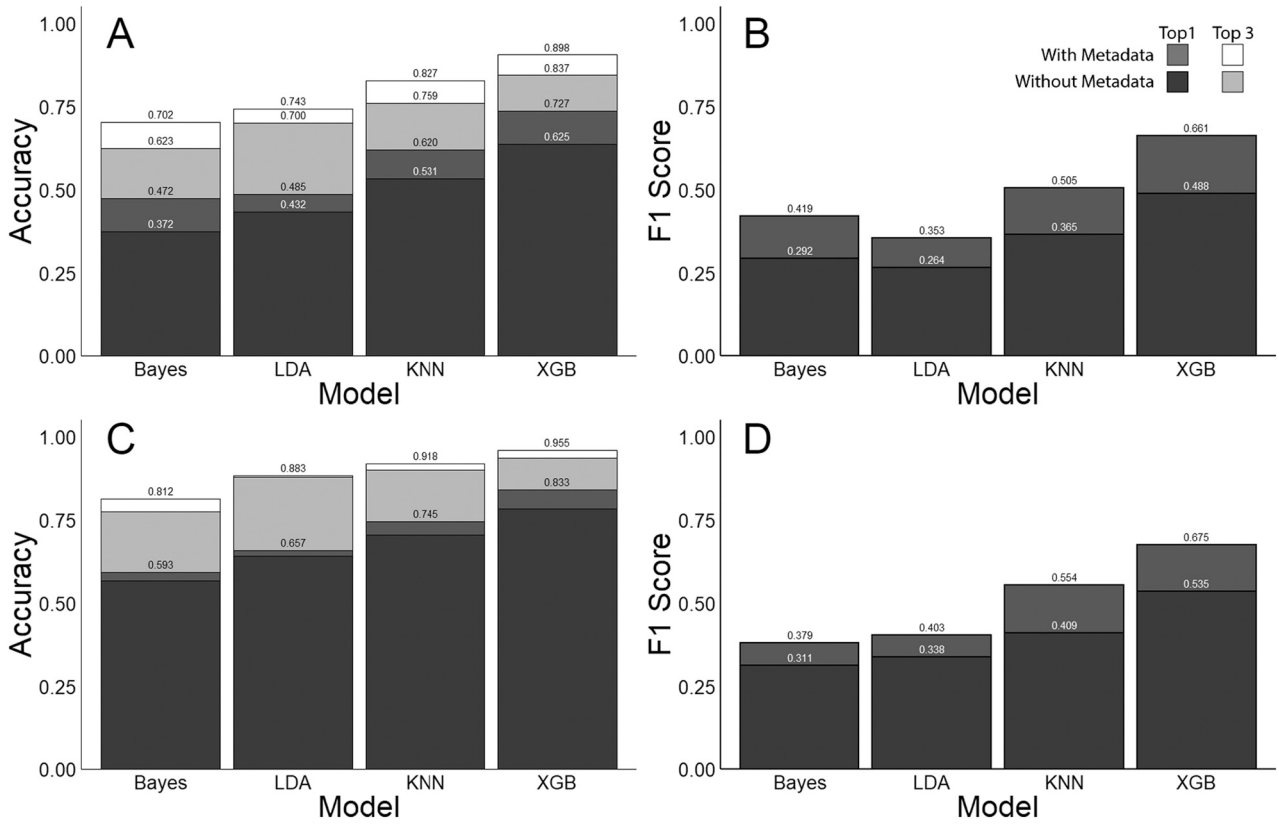



**Fig. 3.** (A, B, C, D) Invertebrate bycatch classification performance metrics for 'traditional' models (i.e. not deep learning) trained with and without contextual metadata. Top 1, Top 3, and F1 score are defined in the Glossary. (A) Top 1 and top 3 accuracy at LITL. (B) F1 score at LITL. (C) Top 1 and top 3 at order level. (D) F1 score at order level. (Naïve Bayes [Bayes]; Linear discriminant analysis [LDA]; K-Nearest Neighbours [KNN]; Artificial Neural Network [ANN]; Extreme Gradient Boosting [XGB]).

classification had a top-1 accuracy of 39.4%. When we combined regular classifications and zero-shot classifications, the average overall top-1 accuracy of the XGBoost model across the entire NEON dataset was 71.7% when classifying to the lowest possible label with metadata included. Comparatively, the top performing CNN had a top-1 zero shot accuracy of 65.5%, resulting in a whole-NEON-dataset accuracy of 72.6%. The average specificity for the LITL labeling scheme was 3.2 (i.e. specimens were generally classified either as family or order).

### 3.2. Order level

After we removed uncommon orders and specimens with LITLs above order (Fig. 2), our training datasets contained 37,578 specimens across 16 orders (including "Ignore", "Larva", and "Nymph" groups) while our testing datasets contained 16,105 specimens. The 18,995 specimens we removed from the training and testing datasets were used for zero-shot classifications.

When we trained and tested the models at the order level, XGBoost still performed the best out of the 'traditional' models with an average top-1 accuracy of 83.3%, top-3 accuracy of 95.5%, and F1 score of 0.675 (Fig. 3). Zero-shot accuracy was 15.1% when classified to the lowest possible level resulting in a whole dataset accuracy of 65.5%. The neural network with the highest accuracy used all three data types (images, morphometric data, and contextual metadata), and returned a top-1 accuracy of 86.5%, top 3 accuracy of 97.1%, and F1 score of 0.723 (Fig. 5). The CNN's zero shot accuracy was 34.7%, resulting in a whole dataset accuracy of 73.0%. Unlike the LITL models, the CNN outperformed XGBoost across all performance metrics at the order level. Both the CNN and XGBoost models greatly overpredicted the 'Ignore' label when making zero-shot classifications, with 'Ignore' comprising 65.3% of classifications in the XGBoost model and 58.1% in the CNN model. The average specificity for the order level model was 4.6 (i.e. most specimens were classified as either order or class).

## 4. Discussion

Here, we explore practical approaches for using machine learning specimen classification on a challenging terrestrial invertebrate dataset. Our dataset is continental in scope, has varying levels of taxonomic specificity, spans three phyla of terrestrial invertebrates, contains non-invertebrate and fragmented specimens, and had a very long tail distribution rank abundance curve—in short, it looks much like many real-world ecological datasets. Despite this, our best performing model (CNN) reached an overall LITL accuracy of 72.6%, and an order-level accuracy of 73.0%. These accuracies include zero-shot classifications in which the models identified taxa belonging to labels not included in the training dataset. When used in a practical setting, machine learning models will have to frequently overcome common challenges of ecological datasets while maintaining high performance standards. We propose our methods and results presented here be used as a foundation to be built and improved upon as we strive to increase the robustness and practicality of ecological machine learning models.

### 4.1. The challenges and opportunities of ecological datasets

Ecological datasets have a wide range of challenges and opportunities compared with the idealized datasets frequently examined in machine learning classification studies. Due to data complexities like inconsistent taxonomic resolution and the long tail of diversity, our classification accuracies are significantly lower than in other machine learning studies (Árje et al., 2020; Blair et al., 2020; Marques et al., 2018; Mayo and Watson, 2007; Terry et al., 2020). But we also propose

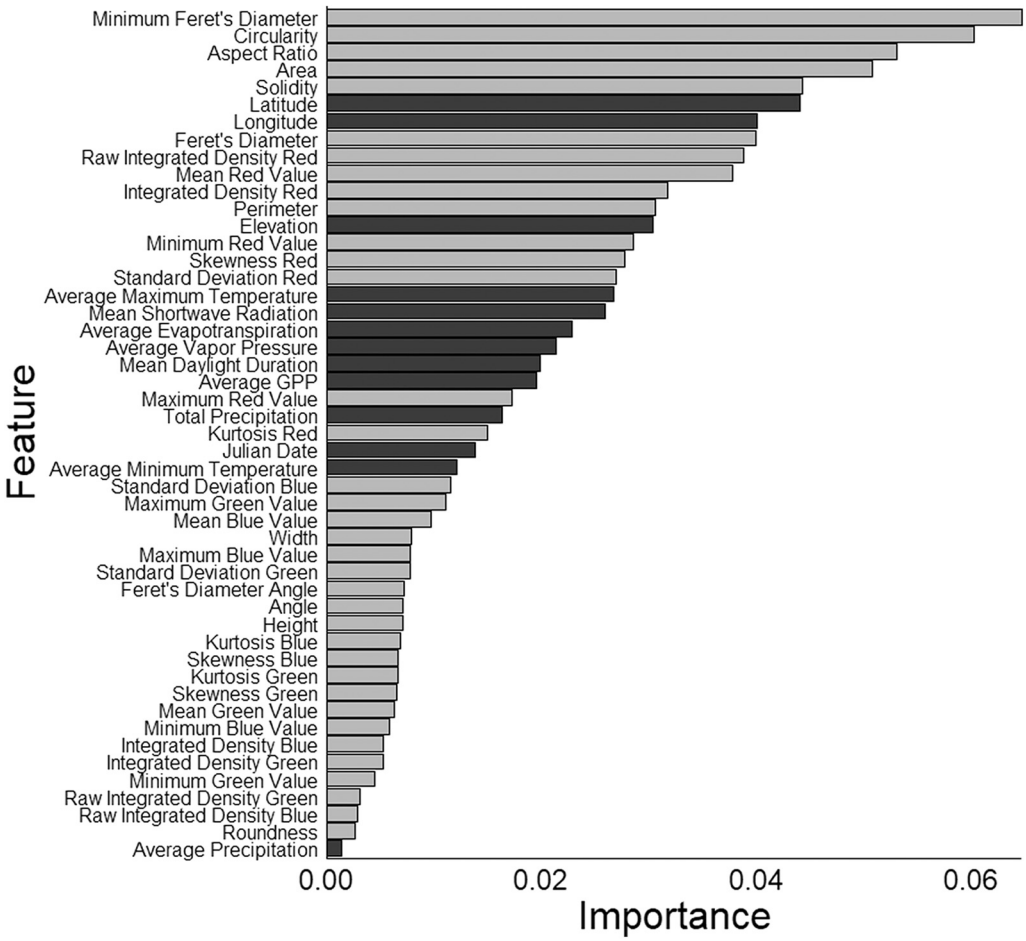



**Fig. 4.** Predictor variable importance for the Lowest Identified Taxonomic Level (LITL) XGBoost multiclass classifier. "Importance" is defined in the xgboost R package as the contribution of each feature (predictor variable) to the model (Chen & He, 2014). Bars of variables representing contextual metadata (i.e. data about the sampling event, such as location and weather) are shaded darker.

solutions for these more real-world datasets. In datasets with inconsistent taxonomic resolution, we find that LITL classification has two main advantages over single-level classification. First, LITL classification allows for more data to be seen by the model. In the LITL models, 97.0% of all specimens were included in either the training or testing dataset, compared to only 73.9% of specimens being included in the order-level models. This resulted in only a 0.2% minimum difference between the CNN LITL model's accuracy with and without zero-shot specimens (72.8% vs 72.6%), while the order-level classifier had a minimum 13.5% gap (86.5% vs 73.0%). LITL classification also increase the model's specificity, with our LITL models having a specificity score of 3.2 (between family and order) compared to the order-level's specificity score of 4.6 (between order and class).

We have also shown here that contextual metadata is one example of how common features of ecological datasets can be advantageous to model performance. Previous studies have already shown adding contextual metadata can improve performance in deep learning models (Ellen et al., 2019; Terry et al., 2020), but similar research in traditional machine learning models had not been explored to our knowledge. Location data was especially important, with latitude and longitude ranking sixth and ninth out of 46 variables respectively (Fig. 4). Other contextual metadata such as date, temperature, and precipitation were also more important than most morphometric variables (Fig. 4). This is despite the possibility that the effect of these variables may have been dampened by the fact that data was only collected over a three year period, with 92.9% of the data coming from 2016 alone. Without a larger timeframe to allow trends to emerge, some metadata like temperature and precipitation are unlikely to provide any additional information than location and time. The fact that contextual metadata variables like temperature and precipitation still show some importance in our models testifies to their potential value in datasets spanning several years with more sampling events. In such datasets, patterns between taxa occurrence and contextual metadata could be much stronger, leading to greater variable importance and possibly higher overall accuracy.

### 4.2. *Challenges of measuring accuracy*

Throughout the course of this study, we discovered that defining a model's accuracy is more subjective and contextual than intuition may suggest. For our XGBoost models, accuracy can be measured as top-x (e.g. top-1, top-3, etc.), at the order-level or LITL, with zero-shot, without zero-shot, or only zero-shot. As shown in our zero-shot accuracy as well as in other studies (Blair et al., 2020), the hierarchical structure of taxonomies can also be used to measure accuracy at multiple taxonomic levels, even if the model was trained to be used at one taxonomic level. Model performance can also be measured on a per-label basis using precision, recall, and confusion matrices (Fig. S6). This all leads to an overwhelming number of options to measure and interpret model performance.

To filter through this data deluge, we recommend choosing accuracy metrics that are reflective of the research question your model is addressing. For example, to estimate a model's accuracy when it sees and knows all input taxa, an accuracy measurement that excludes zero-

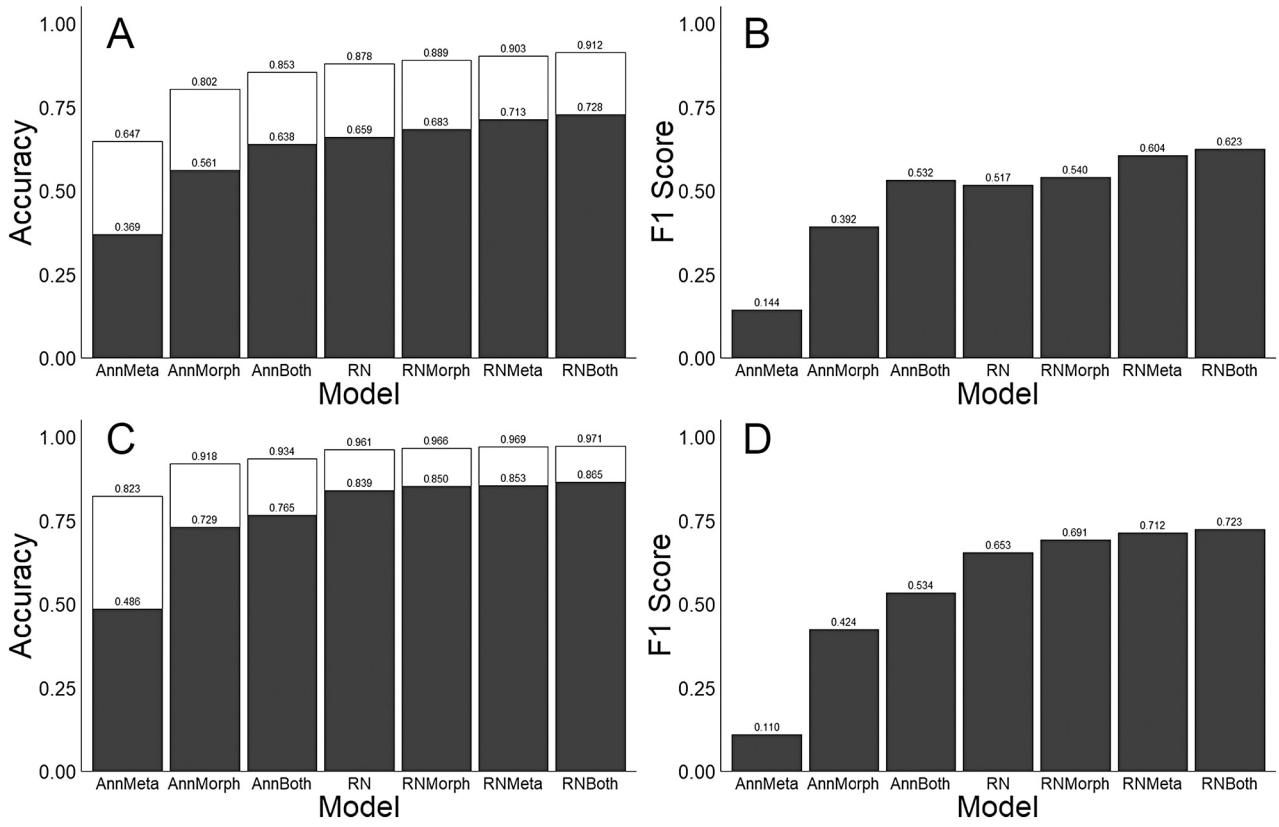



**Fig. 5.** (A, B, C, D) Invertebrate bycatch classification performance metrics for different configurations of artificial neural network (ANN) models trained. Top 1, Top 3, and F1 score are defined in the Glossary. Filled bars represent top 1 results, hollow bars represent top 3 results. (A) Top 1 and top 3 accuracy at LITL. (B) F1 score at LITL. (C) Top 1 and top 3 at order level. (D) F1 score at order level. (ANN trained with only metadata [AnnMeta]; ANN trained with only morphometric data [AnnMorph]; ANN trained with metadata and morphometric data [AnnBoth]; ResNet-50 trained using only images [RN]; Combined ResNet-50 with ANN trained only using morphometric data [RNMorph], Combined ResNet-50 with ANN trained only using metadata [RNMeta], Combined ResNet-50 with ANN trained using both metadata and morphometric data [RNBoth]).

shot accuracy would be the most informative. Conversely, including zero-shot accuracy would be best for a practical measurement of accuracy that uses all possible input data, regardless of if it is seen or known by the model.

We also recommend researchers consider the fit of their chosen model to their research question and study design. While researchers might be eager to use 'bleeding edge' technology with deep learning models, we have shown that traditional machine learning models are still deserving of consideration. Deep learning methods offer convenience to the end user as they accept the most types of input data, and do not necessarily require images to be standardized. In cases when the imaging conditions are expected to be variable (e.g. in the field), deep learning models should be favoured. Deep learning models also tend to see greater benefit compared to traditional machine learning as the amount of training data increases (Zappone et al., 2019). However, when the training dataset is relatively small, the performance differences between deep learning and traditional machine learning are not as pronounced, and may actually favour the latter. In cases where imaging conditions can be precisely controlled (such as a lab setting), traditional machine learning models warrant consideration. Traditional machine learning models have the benefit of simplicity, both in terms of interpretability and required programming skills. This simplicity also leads to accessibility benefits, as the hardware requirements to run traditional machine learning models are markedly lower than deep learning models. If part of the motivation behind automating specimen classification is to remove barriers to entry for large scale biomonitoring efforts, the development of accessible methods should be encouraged.

### 4.3. Next steps

We show that zero-shot classification is highly practical for ecological purposes, as it allows otherwise unknown taxa, such as those with too few observations or with broadly specific labels to still be classified by the model. However, zero-shot classifications performed this way are still imperfect. They are less accurate than normal classifications (7.3% less accurate in our LITL CNN model: 65.5% vs 72.8%), must be at a lower taxonomic specificity than their original label, and are restricted to the seen and known taxonomic groups of the model (Fig. S7). It is also currently impossible to know with certainty which classifications should be considered zero-shot in a practical setting where the correct labels are not known a priori. This means that while the long tail of uncommon species (Fig. 2) can be classified to some degree, there currently is unavoidable information loss and severe practicality issues when making these classifications. Solving these problems of unavoidable information loss and restricted labelling would be a breakthrough for zero-shot classification in ecological datasets.

## 5. Conclusion

Automated classifications via machine learning has the potential to transform the way ecologists conduct large scale monitoring programs. However, if machine learning classification is to be accepted as a standard data collection tool for ecologists, our assessments of model performance must be demonstrative of real-life scenarios. While there is a tendency to want to be 'fair' to our models, fairness means very little in practice. If our models do not meet the performance demands of ecologists under field conditions, that simply means our models must be

adapted and improved. Here we showed that the inclusion of contextual metadata can greatly improve classification accuracy, particularly when the specificity of classifications is high. We also explored experimental techniques in zero shot classification to address the challenge of classifying under-represented taxa. Finally, we contrasted traditional machine learning methods with newer deep learning techniques and showed that traditional machine learning methods may still warrant consideration in some use cases. Overall, we demonstrated methods in which models can be assessed practically while also describing methods in which classification performance can be improved in the face of challenges posed by ecological datasets.

## Declaration of Competing Interest

Jarrett Blair is the CEO and cofounder of Luna ID Inc., a mobile application development company that specializes in developing apps to identify insects from images using machine learning. However, the contents of this manuscript are purely academic, and the methods herein have no commercial purpose. All other authors have no competing interests.

## Data availability

All R code and data is available on GitHub (https://github.com/Jarrett-Blair/EcoVision).

## Acknowledgements

This work was supported by a NSERC Discovery grant to K.E.M. as well as NSF DEB 1702426 to M.D.W., M.K., C.D.S., and K.E.M. We thank Drs Michelle Tseng, Leonid Sigal, and Rachel Germain for their fruitful discussions, as well as the National Ecological Observation Network for allowing us to use their specimens. We thank Tanner Ortery for help in developing the imaging pipeline as part of the NSF REU program. We also thank the anonymous reviewer and the editor, Dr. George Arhonditsis, for their helpful feedback.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2022.101896.

## References

(NEON) National Ecological Observatory Network, 2022. Explore Field Sites. Retrieved August 11, 2022, from. https://www.neonscience.org/field-sites/explore-field-sites.

Ärje, J., Melvad, C., Jeppesen, M.R., Madsen, S.A., Raitoharju, J., Rasmussen, M.S., Høye, T.T., 2020. Automatic image-based identification and biomass estimation of invertebrates. Methods Ecol. Evol. 11 (8), 922–931. https://doi.org/10.1111/2041-210X.13428.

Berg, T., Liu, J., Lee, S.W., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N., 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011–2018. https://doi.org/10.1109/CVPR.2014.259.

Blair, J., 2022. EcoVision. https://github.com/Jarrett-Blair/EcoVision.

Blair, J., Weiser, M.D., Kaspari, M., Miller, M., Siler, C., Marshall, K.E., 2020. Robust and simplified machine learning identification of pitfall trap-collected ground beetles at the continental scale. Ecol. Evol. 10 (23), 13143–13153. https://doi.org/10.1002/ece3.6905.

Chen, T., He, T., 2015. xgboost: Extreme Gradient Boosting. R Package Version 0.4-2, 1 (4), pp. 1–4.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Deng, J., Krause, J., Berg, A.C., Fei-Fei, L., 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3450–3457. https://doi.org/10.1109/CVPR.2012.6248086.

Ding, W., Taylor, G., 2016. Automatic moth detection from trap images for pest management. Comput. Electron. Agric. 123, 17–28. https://doi.org/10.1016/j.compag.2016.02.003.

Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J.B., Collen, B., 2014. Defaunation in the Anthropocene. Science 345 (6195), 401–406. https://doi.org/10.1126/science.1251817.

Ellen, J.S., Graff, C.A., Ohman, M.D., 2019. Improving plankton image classification using context metadata. Limnol. Oceanogr. Methods 17 (8), 439–461. https://doi.org/10.1002/lom3.10324.

Guzman, L.M., Johnson, S.A., Mooers, A.O., M'Gonigle, L.K., 2021. Using historical data to estimate bumble bee occurrence: variable trends across species provide little support for community-level declines. Biol. Conserv. 257, 109141 https://doi.org/10.1016/j.biocon.2021.109141.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.90.

Hoekman, D., Levan, K.E., Ball, G.E., Browne, R.A., Davidson, R.L., Erwin, T.L., Work, T., 2017. Design for ground beetle abundance and diversity sampling within the National Ecological Observatory Network. Ecosphere 8 (4), e01744. https://doi.org/10.1002/ecs2.1744.

Høye, T.T., Ärje, J., Bjerge, K., Hansen, O.L.P., Iosifidis, A., Leese, F., Raitoharju, J., 2021. Deep learning and computer vision will transform entomology. Proc. Natl. Acad. Sci. U. S. A. 118 (2) https://doi.org/10.1073/PNAS.2002545117.

Jansen, J., Hill, N.A., Dunstan, P.K., Eléaume, M.P., Johnson, C.R., 2018. Taxonomic resolution, functional traits, and the influence of species groupings on mapping Antarctic seafloor biodiversity. Front. Ecol. Evol. 6, 81. https://doi.org/10.3389/fevo.2018.00081.

Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Ärje, J., Juhola, M., 2014. Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. Ecol. Inform. 20, 1–12. https://doi.org/10.1016/j.ecoinf.2014.01.004.

Karlsson, D., Hartop, E., Forshage, M., Jaschhof, M., Ronquist, F., 2020. The Swedish malaise trap project: a 15 year retrospective on a countrywide insect inventory. Biodivers. Data J. 8, e47255 https://doi.org/10.3897/bdj.8.e47255.

Keller, M., Schimel, D.S., Hargrove, W.W., Hoffman, F.M., 2008. A continental strategy for the national ecological observatory network. Front. Ecol. Environ. 6 (5), 282–284. https://doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2.

Krell, F.T., 2004. Parataxonomy vs. taxonomy in biodiversity studies - pitfalls and applicability of "morphospecies" sorting. Biodivers. Conserv. https://doi.org/10.1023/B:BIOC.0000011727.53780.63.

Marques, A.C.R., Raimundo, M.M., Cavalheiro, E.M.B., Salles, L.F.P., Lyra, C., Von Zuben, F.J., 2018. Ant genera identification using an ensemble of convolutional neural networks. PLoS One 13 (1), e0192011. https://doi.org/10.1371/journal.pone.0192011.

Mayo, M., Watson, A.T., 2007. Automatic species identification of live moths. Knowl.-Based Syst. 20 (2), 195–202. https://doi.org/10.1016/j.knosys.2006.11.012.

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., Muller, K.R., 1999. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing - Proceedings of the IEEE Workshop, pp. 41–48. https://doi.org/10.1109/nnsp.1999.788121.

Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5 (6), 1–15. https://doi.org/10.1890/ES13-00359.1.

Preston, F.W., 1948. The commonness, and rarity, of species. Ecology 29 (3), 254–283. https://doi.org/10.2307/1930989.

Running, S., Mu, Q., Zhao, M., 2015. MOD17A2H MODIS/Terra Gross Primary Productivity 8-Day L4 Global 500m SIN Grid V006 (NASA EOSDIS Land Processes DAAC).

Running, S., Mu, Q., Zhao, M., 2017. MOD16A2 MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V006 (NASA EOSDIS Land Processes DAAC).

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Cardona, A., 2012. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9 (7), 676–682. https://doi.org/10.1038/nmeth.2019.

Schmidt-Kloiber, A., Nijboer, R.C., 2004. The effect of taxonomic resolution on the assessment of ecological water quality classes. Hydrobiologia 516 (1–3), 269–283. https://doi.org/10.1023/B:HYDR.0000025270.10807.10.

Spiesman, B.J., Gratton, C., Hatfield, R.G., Hsu, W.H., Jepsen, S., Mccornack, B., Wang, G., 2021. Assessing the potential for deep learning and computer vision to identify bumble bee species from images. Sci. Rep. 11 (1), 1–10. https://doi.org/10.1038/s41598-021-87210-1.

Team, R.C., 2021. R: A Language and Environment for Statistical Computing. Retrieved from. https://www.r-project.org/.

Terry, J.C.D., Roy, H.E., August, T.A., 2020. Thinking like a naturalist: enhancing computer vision of citizen science images by harnessing contextual data. Methods Ecol. Evol. 11 (2), 303–315. https://doi.org/10.1111/2041-210X.13335.

Thessen, A.E., 2016. Adoption of machine learning techniques in ecology and earth science. One Ecosyst. 1, e8621 https://doi.org/10.3897/oneeco.1.e8621.

Thornton, P.E., Shrestha, R., Thornton, M., Kao, S.C., Wei, Y., Wilson, B.E., 2021. Gridded daily weather data for North America with comprehensive uncertainty quantification. Sci. Data 8 (1), 1–17. https://doi.org/10.1038/s41597-021-00973-0.

Thorpe, A.S., Barnett, D.T., Elmendorf, S.C., Hinckley, E.L.S., Hoekman, D., Jones, K.D., Thibault, K.M., 2016. Introduction to the sampling designs of the National Ecological Observatory Network Terrestrial Observation System. Ecosphere 7 (12), e01627. https://doi.org/10.1002/ecs2.1627.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Belongie, S., 2018. The iNaturalist species classification and detection dataset. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8769–8778. https://doi.org/10.1109/CVPR.2018.00914.

van Klink, R., Bowler, D.E., Gongalsky, K.B., Swengel, A.B., Gentile, A., Chase, J.M., 2020. Meta-analysis reveals declines in terrestrial but increases in freshwater insect abundances. Science 368 (6489), 417–420. https://doi.org/10.1126/SCIENCE.AAX9931.

Verberk, W., 2012. Explaining general patterns in species abundance and distributions. Nat. Educ. Knowl. 3 (10), 38.

Weiser, M.D., Marshall, K.E., Siler, C.D., Kaspari, M., 2021. Batch Extraction of Morphological and Color Metrics from Invertebrate Samples. https://doi.org/10.17504/protocols.io.byt4pwqw.

Welti, E.A.R., Roeder, K.A., De Beurs, K.M., Joern, A., Kaspari, M., 2020. Nutrient dilution and climate cycles underlie declines in a dominant insect herbivore. Proc. Natl. Acad. Sci. U. S. A. 117 (13), 7271–7275. https://doi.org/10.1073/pnas.1920012117.

Wepprich, T., Adrion, J.R., Ries, L., Wiedmann, J., Haddad, N.M., 2019. Butterfly abundance declines over 20 years of systematic monitoring in Ohio, USA. PLoS One 14 (7), e0216270. https://doi.org/10.1371/journal.pone.0216270.

Whittaker, R.H., 1965. Dominance and diversity in land plant communities: numerical relations of species express the importance of competition in community function and evolution. Science (New York, N.Y.) 147 (3655). https://doi.org/10.1126/science.147.3655.250.

Zappone, A., Di Renzo, M., Debbah, M., 2019. Wireless networks design in the era of deep learning: model-based, aI-based, or both? IEEE Trans. Commun. 67 (10), 7331–7376. https://doi.org/10.1109/TCOMM.2019.2924010.