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

Running head: Machine-assisted invertebrate classification

Jarrett Blair[1], Michael D. Weiser[2], Kirsten de Beurs[2], Michael Kaspari[2], Cameron Siler[2,3], Katie E. Marshall[1]

[1] Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

[2] Department of Biology, University of Oklahoma, Norman, OK 73019-0235, USA

[3] Sam Noble Oklahoma Museum of Natural History, University of Oklahoma, 2401 Chautauqua Ave., Norman, OK 73072-7029, USA

Corresponding author: Jarrett Blair, Departments of Zoology 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada, blair@zoology.ubc.ca

## Abstract

Despite growing concerns over the health of global insect populations, the spatiotemporal breadth of insect population data is severely lacking. Machine-assisted classification has been proposed as a potential solution to quickly gather large amounts of data, but previous studies have often used unrealistic or idealized datasets to train their models. In this study, we describe a practical methodology for including machine learning in ecologic data acquisition pipelines. Here we train and test machine learning algorithms to classify over 56,000 bulk terrestrial invertebrate specimens from image data. All specimens were collected in pitfall traps by the National Ecological Observatory Network (NEON) at 27 locations across the United States. Image data was extracted as feature vectors using ImageJ. When classifying specimens that were known and seen by our models, we reached an accuracy of 74.7% at the lowest taxonomic level. We also classified invertebrate taxa that the model was not trained on using zero-shot classification, with an accuracy of 42.1% on these taxa. The general methodology outlined here represents a realistic approach to how machine learning may be used as a tool for ecological studies.

## Introduction

Several recent studies have suggested that terrestrial invertebrates may be suffering drastic population and diversity losses (Dirzo et al., 2014; Welti, Roeder, De Beurs, Joern, & Kaspari, 2020; Wepprich, Adrion, Ries, Wiedmann, & Haddad, 2019). However, these losses are likely not universal globally or across all invertebrate taxa (Guzman, Johnson, Mooers, & M'Gonigle, 2021; van Klink et al., 2020). Generalizations about global invertebrate diversity and abundance trends require a solid data foundation, yet invertebrates remain poorly-sampled relative to their diversity and abundance (Høye et al., 2021; van Klink et al., 2020). While there

are many invertebrate biomonitoring programs operating around the world, the data collected from these studies are often spatiotemporally coarse and taxonomically shallow (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, Hartop, Forshage, Jaschhof, & Ronquist, 2020; van Klink et al., 2020). Perhaps the most time- and resource-intensive aspect of a invertebrate biomonitoring project is sorting and classifying specimens (Karlsson et al., 2020). While this process is normally performed by human taxonomists and parataxonomists, advances in machine learning and computer vision have made machine-assisted classification a possibility for many groups of invertebrates (Ärje et al., 2020; Blair et al., 2020; Marques et al., 2018; Mayo & Watson, 2007). Machine-assisted classification has the potential to increase the efficiency, output, and accessibility of biomonitoring programs (Peters et al., 2014; Thessen, 2016). In a time when more data on invertebrate diversity and abundance is desperately needed, machine learning may offer a transformative solution to change the way we conduct biomonitoring research and invertebrate biology generally.

A common theme in the literature involving computer vision for taxonomic classification is the use of unrealistic or impractical case studies. Specifically, the datasets used in these studies often 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 use large, real ecological datasets accept that the long tail of low abundance taxa must be sacrificed for the benefit of their model's performance (Marques et al., 2018; Mayo & Watson, 2007). While these studies can still provide valuable insights about ecological uses for machine learning, significant work is necessary improve the practicality of computer vision for

specimen classification if machine learning can be integrated into ecological data acquisition pipelines.

One practical application of machine learning for ecological purposes is the web & mobile application iNaturalist (Van Horn et al., 2018). iNaturalist is a community science social network in which users can share their own biodiversity observations as well as comment on others. In 2018, iNaturalist implemented an automated identification tool, whereby a computer vision algorithm recommends taxonomic classifications to users' submissions. Predecessors of iNaturalist such as LeafSnap and the Merlin Bird ID app had similar implementations of automated classifications using computer vision, but iNaturalist was the first to do so with the goal of being able to classify *any* macro-organism. However, as a community-science focussed app, iNaturalist has limitations that make it unsuitable for other macroecological uses. For example, iNaturalist submissions tend to be biased towards charismatic taxa, which causes its automated identification tool to fumble uncharismatic fauna such as ground beetles (Blair et al., 2020). Additionally, the current build of iNaturalist is not suited for large-scale bulk identifications that are needed for many macroecological studies. In many instances, custom built automated classifiers are still necessary instead of off-the-shelf applications like iNaturalist.

While there have been some promising results when using machine learning to classify biological specimens (Blair et al., 2020; Ding & Taylor, 2016; Spiesman et al., 2021; Van Horn et al., 2018), ecological datasets pose several challenges for machine learning algorithms. For example, most ecological datasets are uneven, with relatively few taxa that are highly abundant, while many others have relatively low abundance (i.e. the 'long tail' of a rank abundance curve; McGill et al., 2007). This causes two problems for machine learning algorithms. First, machine learning algorithms tend to be biased towards predicting overabundant classes. This is because a

model may achieve higher accuracies by simply predicting the most abundant taxon every time instead of making informed predictions (Dinga, Penninx, Veltman, Schmaal, & Marquand, 2019). This can lead to relatively accurate but useless models that never predict low abundance taxa (Dinga et al., 2019). Second, if the taxa in the long tail are very scarce, there may not be enough data to properly train a machine learning algorithm to identify those rare taxa. These taxa are 'unknown and unseen' by the machine learning model and can never be specifically predicted by the model (Deng, Krause, Berg, & Fei-Fei, 2012). In order for machine learning to be a practical solution for use in ecological studies, these problems must be addressed.

Another challenge facing some ecological datasets (especially large, diverse ones) is a 'ragged edge' of taxonomic specification: in most real ecological datasets, identifications are often made at multiple taxonomic levels dependent on the taxonomist. Some may be identified down to species, while others are only identified to some higher level (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 vs specificity (Deng et al., 2012). Two potential ways to address this problem are lowest identified taxonomic level (LITL) models and single-level models. LITL models assign labels and measure accuracy based on the LITL label assigned to each individual, while single-level models assign labels and measure accuracy all at one specified taxonomic level. 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). Single-level models prevent nested labels by classifying all individuals at a single taxonomic level, but can result in excluded individuals that labelled at higher taxonomic levels. 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) and an 11.2% increase (68.6% → 79.8%) when classifying North American birds (Berg et al., 2014). Additionally, the hierarchical structure of taxonomic relationships 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, Schimel, Hargrove, & Hoffman, 2008). Part of NEON's ecological monitoring involves collecting and processing thousands of invertebrate specimens every year from arrays of pitfall traps that are run for 14 days (Thorpe et al., 2016; Hoekman et al., 2017). 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 away

unused. 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 classification methods to the large terrestrial invertebrate dataset collected by NEON with the goal of addressing the challenges taxonomic classification poses for machine learning while also taking advantage of the benefits provided by ecological datasets. 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 morphological data? (2) How can we improve classification in the real-world conditions of long tails and inconsistent taxonomic resolution? We predict that models trained with datasets that include contextual metadata will outperform models that only contain morphometric data. We also predict that if the presence of overrepresented groups adversely affects model performance, increasing the evenness in the training data will improve model performance, especially when measured by F1 score. 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 classification to overcome the 'long tail' problem.

## Methods

### Contextual Metadata

Here, we define contextual metadata as any non-morphometric and non-taxonomic data included in the invertebrate dataset. Our contextual metadata can broadly be categorized as spatiotemporal, temperature, and precipitation data. We extracted the daily minimum and maximum temperature as well as the daily precipitation for each location from the Daymet V3 dataset. Daymet V3 provides gridded estimates of daily climatological and weather variables at 1000m by 1000m spatial resolution for North America. The gridded datasets are developed using ground-based observation and statistical modeling techniques (Thornton et al., 2021). Spatiotemporal data includes longitude, latitude, and the ordinal day of the sampling event. Our temperature and precipitation data were measured daily and reported as the minimum, maximum, mean, and standard deviation during the preceding two weeks before a sampling event. All contextual metadata values were the same for all specimens collected in a given trap event. Between our morphometric data and contextual metadata, 46 descriptive variables were used as our training data (36 morphometric variables, 10 contextual variables).

## Machine Learning

### *Preparation*

All machine learning methods were developed using R version 4.0.5 (R Core Team, 2021). To prepare the data for machine learning training and testing, we first removed all taxa with fewer than 100 observations from the dataset (Figure 1), 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. This process was repeated 10 times, resulting in 10 separate training and testing datasets (Figure S1) to ensure we had robust performance measurements for each model. Finally, we normalized the data by centering and scaling each predictor variable such that they all had a mean of 0 and standard deviation of 1. After pre-processing the data, it was ready

to be used for training in a machine learning model. We contrasted five types of machine learning algorithms: K-nearest neighbours (KNN; Cover & Hart, 1967), linear discriminant analysis (LDA; Mika, Ratsch, Weston, Scholkopf, & Muller, 1999), naïve Bayes (NB; Mika et al., 1999), eXtreme Gradient Boosting (XGBoost; Chen & He, 2014), and artificial neural networks (ANN; Haykin, 2008). We chose these five algorithms as they use a wide range of machine learning techniques and increase our likelihood that at least one algorithm will reach a high classification accuracy. Our methods for optimizing these algorithms are described in Table 1.

*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) (Figure S2a). 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 (Figure S2b). 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.

*Contextual Metadata*

To determine the effects and importance of contextual metadata, we trained and tested models of each algorithm type using LITL 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.

*Adjusting for sampling bias*

To combat the unevenness of taxonomic group representation (e.g. overabundance of ant classes), we trained and tested several configurations of binary and multiclass classifiers (Figure S3). The objective of the binary classifiers was to increase evenness among classes by removing/separating out overabundant classes. Our first binary classifier was an "adult invertebrate" classifier. When creating the invertebrate dataset, some non-adult invertebrate objects (e.g. juveniles, body parts, abiotic particles, etc.) were included. The goal of this classifier was to identify adult invertebrates and classify them as "keep" and classify everything else as "ignore". Specimens labelled as "keep" were kept in the dataset and passed on to subsequent classifiers, while "ignore" specimens were discarded. The "keep" specimens were passed on to one of two classifiers depending on the classifier configuration. One option was for the specimens to simply be passed to a multiclass model to assign the final taxonomic classification. Alternatively, the specimens could be passed to a second binary classifier to split the data into "ants" and "not ants", where "ants" are any member of the family Formicidae and "not ants" are every label not part of the "ant" or "ignore" groups. Of the 46 total LITLs, 17 were related to ants (i.e. Formicidae or lower). We predicted that this may cause the model to be biased towards ant-related classifications, and designating them their own model would reduce that bias. Once classified, the "ant" and "not ant" groups were passed on to their own respective

multiclass models to assign the final taxonomic classification. All binary and multiclass models were trained using XGBoost with morphometric and contextual metatdata.

To account for overabundance of individual classes, we downsampled the LITL training dataset such that no label had more than 1000 specimens in the training dataset. Individuals from the downsampled taxa were removed randomly. Overrepresented taxa were not downsampled from the testing dataset, as this maintained their natural occurrence frequencies and allowed for a 1:1 performance comparison with the other models. All downsampled models were trained using XGBoost with morphometric and contextual metadata.

*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 87 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 belong to the order Coleoptera, which is a label known by the model. 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 measured the accuracy of zero-shot classifications in two ways (Figure S4). In the first method, hereon known as the comprehensive method, we measured accuracy using all zero-shot specimens at every taxonomic level, regardless of if they had a known taxonomic label. For example, in the order-level models, a specimen with the LITL of “Arthropoda” would still be included when calculating the class-level zero-shot accuracy despite having zero chance of being correctly classified by the model. In the second zero-shot accuracy measurement, hereon known as the limited method, we only calculated accuracy using specimens that had a known taxonomic group. Using the limited method, specimens with the LITL of Arthropod would only be included in the accuracy calculation at the phylum-level. When measuring zero-shot accuracy, specimens labelled as any non-intact adult group (i.e. “Ignore”, “Larva”, etc.) were classified as the species level in the LITL models and at the order level in the order level models. We also used zero-shot accuracy to estimate LITL and order-level accuracy across the entire invertebrate dataset. We extrapolated the accuracies measured from the testing datasets to the training datasets to get an estimated accuracy for all ‘known’ specimens (43,025 LITL specimens and 43,390 order-level specimens). We then combined this estimated accuracy with our zero-shot accuracies to get an estimated accuracy for the entire dataset.

## Results

<u>LITL</u>

The entire NEON invertebrate dataset contained 56,153 specimens with 153 LITLs and 29 different taxonomic orders (Figure 1). After we removed uncommon LITL labels (<100 individuals per label), our LITL training datasets contained between 36,024 and 36,100 specimens across 46 LITL labels while our testing datasets contained between 15,422 and 15,501 specimens (Figure 1). Our training datasets varied in size due to the random sampling method we used when splitting the datasets. The 13,128 uncommon specimens we removed from the training and testing datasets were used for zero-shot classifications. The artificial neural network model was only able to converge once out of the 10 training iterations, so its performance measurements were only taken from that iteration. The performance metrics of all other models were taken as the average across the 10 iterations.

When predicting specimens down to their LITL, we found models that contained metadata (i.e. location, date, temperature, and precipitation data) always had better performance metrics than their non-metadata counterparts across all algorithms (Figure 2). Of the models we tested, XGB always had the highest scores, with an average top-1 accuracy of 73.7%, top-3 accuracy of 90.4%, and F1 score of 0.636 when including metadata. XGBoost also had the greatest accuracy increase between the metadata and no-metadata configurations with an average boost of 9.1%. The average accuracy boost when we included metadata in the training dataset was 7.2% across all models. The F1 score of KNN was improved the most when metadata was added (+0.219), with an average increase of 0.121 across all models. We found the metadata variables with the highest importance (mean decrease in accuracy) for the XGB models were latitude and longitude, ranking sixth and ninth out of 46 predictor variables respectively (Figure SX).

When we tried to increase evenness in taxonomic representation (by splitting ‘ants’ into their own model) and individual abundance (downsampling overrepresented classes), we found that performance of the models were lower compared to the unaltered model. Down sampling resulted in the lowest accuracy at 69.2%, while only splitting the data once (using a binary classifier to classify specimens as intact adults or not) resulted in the lowest F1 score (0.629). Splitting the dataset twice resulted in no change in accuracy, but resulted in an F1 score of 0.644.

We performed zero-shot classification on unseen and unknown specimens by classifying them as seen and known classes at higher taxonomic levels. When classified to the closest possible level, we found zero shot classification had a top-1 accuracy of 42.1%. The zero-shot accuracy increased as the taxonomic resolution of classifications decreased, with an accuracy of 85.1% at phylum (Figure S5). When we combined regular classifications and zero-shot classifications, the average overall accuracy of the XGB model across the entire NEON dataset was 65.0% when classifying to the lowest possible label with metadata included.

Order level

After we removed uncommon orders and specimens with LITLs above order, our training datasets contained between 30,321 and 30,446 specimens across 16 orders (including “Ignore”, “Larva”, and “Nymph” groups) while our testing datasets contained between 12,945 and 13,069 specimens. The 12,763 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, XGB still performed the best with an average top-1 accuracy of 85.9%, top-3 accuracy of 96.0%, and F1 score of 0.700 (Figure 4). Zero-shot classification of the order level models was only possible at class and

phylum, reaching a top-1 accuracy of 6.0% and 72.8% at each level respectively. Zero-shot accuracy was 12.4% at the class level when specimens that did not have known labels at the class level were excluded (Figure S5). When we combined regular classifications and zero-shot classifications, the average overall accuracy of the XGB models across the entire NEON dataset was 76.6% when classifying at the order level or higher.

## 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-biological and fragmented specimens, and was highly uneven in label representation. Despite this, our best performing model (XGBoost) reached a LITL accuracy of 73.7% and order-level accuracy of 85.9%. In addition to overcoming the aforementioned challenges, we also exploited qualities of ecological datasets to improve our models' performances beyond what would normally be expected using standard computer vision techniques. This included using contextual metadata to boost accuracy by 9.2% in or LITL XGBoost models and use the hierarchical structure of taxonomic relationships to perform zero-shot classification on unseen and unknown taxa. Overall, we show that even under unfavourable circumstances, machine learning can offer a practical solution for bulk specimen classification.

We show that in a macroecological context, adding contextual metadata can substantially increase classification accuracy. Location data was especially important, with latitude and longitude ranking sixth and ninth out of 46 variables respectively (Figure SX). Other contextual metadata such as date, temperature, and precipitation were also more important than most morphometric variables (Figure SX). However, the effect of these variables may have been

dampened by the fact that no site had more than 4 sampling events and all sampling events took place in the same year. Such a small sample size means that variables like temperature and precipitation are unlikely to provide anymore information than location and time. For example, if a remote location only has one sampling event, any contextual metadata other than location would be redundant, as all specimens taken from that site would have the same contextual metadata. The fact that contextual metadata variables like temperature and precipitation still show some importance in our models is a testament 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.

Our techniques for overcoming label and taxa overrepresentation did not lead to improved performance, measured either by accuracy or F1 score. Additionally, taxa with low prevalence in our training dataset did not have significantly lower F1 scores than taxa with high prevalence (linear regression, $F = 2.22$, $df = 1, 44$, $p > 0.05$, $R^2 = 0.026$) (Figure SX). This leads us to conclude that, at least in our models, taxonomic overrepresentation did not negatively impact model performance. In some use cases, unmodified taxonomic prevalence may actually be beneficial, as introducing some bias to more prevalent taxa can increase accuracy. An example would be if studies that prioritize accuracy in relative abundance, especially if the model can still detect low abundance taxa to some degree. Issues surrounding bias towards high prevalent taxa arise when detecting low prevalent taxa is a higher priority than relative abundance accuracy. Models whose purpose is to measure taxonomic richness may be impacted more by prevalence bias, as low abundance taxa might be completely missed by the model.

Ultimately, the importance of prevalence bias is dependant on the model's purpose and the impact of prevalence bias on model performance.

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. This all leads to an overwhelming number of options to measure and interpret model performance. In order 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-shot accuracy would be the most informative. Conversely, a practical measurement of accuracy that uses all possible input data, regardless of if it is seen or known by the model, including zero-shot accuracy would be best.

We also recommend researchers consider the fit of their chosen model to their research question as well. In our study, we compared our model's order-level accuracy to the order-level accuracy of iNaturalist when predicting our specimens. While iNaturalist performed comparably to our model (84.6% vs 85.9%, respectively), there are several other factors to consider. For example, iNaturalist's options for bulk image uploads are limited, and there is currently no way to automatically generate labels without the user manually selecting from iNaturalist's suggestions. iNaturalist asks that no more than 10,000 lines of metadata be submitted at a time, and when we tried submitting ~5,000 bulk images, the website became unresponsive. At the very

least, a dataset like NEON's would have to be fragmented and submitted in pieces. This might be fixable as iNaturalist's web application is open source, so there is the possibility that enhanced bulk identification features could be added. Additionally, iNaturalist currently has no way of dealing with non-biological or fragmented specimens (i.e. specimens we listed as "Ignore"). iNaturalist assumes every submission contains at least one specimen, and its model is forced to make a taxonomic classification. This can result in inaccurate classification data if images are not manually vetted for "Ignore" specimens. Finally, as a community science focused app, iNaturalist tends to be biased towards charismatic taxa. Due to this, iNaturalist performs poorly when classifying uncharismatic taxa at low taxonomic levels (Blair et al., 2020). Unlike the bulk classification issue, this problem is not easily solvable, as iNaturalist's machine learning algorithms are not open source. While iNaturalist can be an excellent tool for macroecological research, especially when used in a community science setting, other flexible options such as our models here are useful particularly for non-charismatic species. There is currently no one-size-fits-all solution for machine learning in an ecological context, and that tools that are custom built for a particular use case are likely to be the most effective.

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're far less accurate than normal classifications (31.6% less accurate in our LITL XGBoost model), must be at a lower taxonomic specificity than their original label, and are restricted to the seen and known taxonomic groups of the model (Figure SX). This means that while the long tail of uncommon species (Figure 1) can be classified to some degree, there currently is unavoidable information loss when making these classifications. Solving this

problem of unavoidable information loss and restricted labelling would be a major breakthrough for zero-shot classification in ecological datasets.

## 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 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. Despite this progress, limitations on machine-assisted classification of relatively rare species remain, and even large invertebrate datasets will contain a large number of species that cannot currently be well-classified.

Acknowledgements

## References

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 in Ecology and Evolution*. 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. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.259

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. *Ecology and Evolution*. https://doi.org/10.1002/ece3.6905

Chen, T., & He, T. (2014). xgboost: Extreme Gradient Boosting. *R Package Version 0.4-2*.

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information 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. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2012.6248086

Ding, W., & Taylor, G. (2016). Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture*, *123*, 17–28. https://doi.org/10.1016/j.compag.2016.02.003

Dinga, R., Penninx, B. W. J. H., Veltman, D. J., Schmaal, L., & Marquand, A. F. (2019). Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines. *BioRxiv*. https://doi.org/10.1101/743138

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

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. *Biological Conservation*, *257*. https://doi.org/10.1016/j.biocon.2021.109141

Haykin, S. (2008). Neural Networks and Learning Machines. In *Pearson Prentice Hall New Jersey USA 936 pLinks*. https://doi.org/978-0131471399

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. *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 118. https://doi.org/10.1073/PNAS.2002545117

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. *Ecological Informatics*, *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. *Biodiversity Data Journal*, *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. *Frontiers in Ecology and the Environment*, *6*(5), 282–284. https://doi.org/10.1890/1540-9295(2008)6[282:ACSFTN]2.0.CO;2

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. *Knowledge-Based Systems*, *20*(2), 195–202. https://doi.org/10.1016/j.knosys.2006.11.012

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., … White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, Vol. 10. https://doi.org/10.1111/j.1461-0248.2007.01094.x

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Muller, K. R. (1999). Fisher discriminant analysis with kernels. *Neural Networks for Signal Processing - Proceedings of the IEEE Workshop*, 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*. https://doi.org/10.1890/ES13-00359.1

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. *Scientific Reports*, *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 in Ecology and Evolution*, *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 Ecosystem*, *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. *Scientific Data*. 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. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 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

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. *Proceedings of the National Academy of Sciences of the United States of America*, *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

## Figures

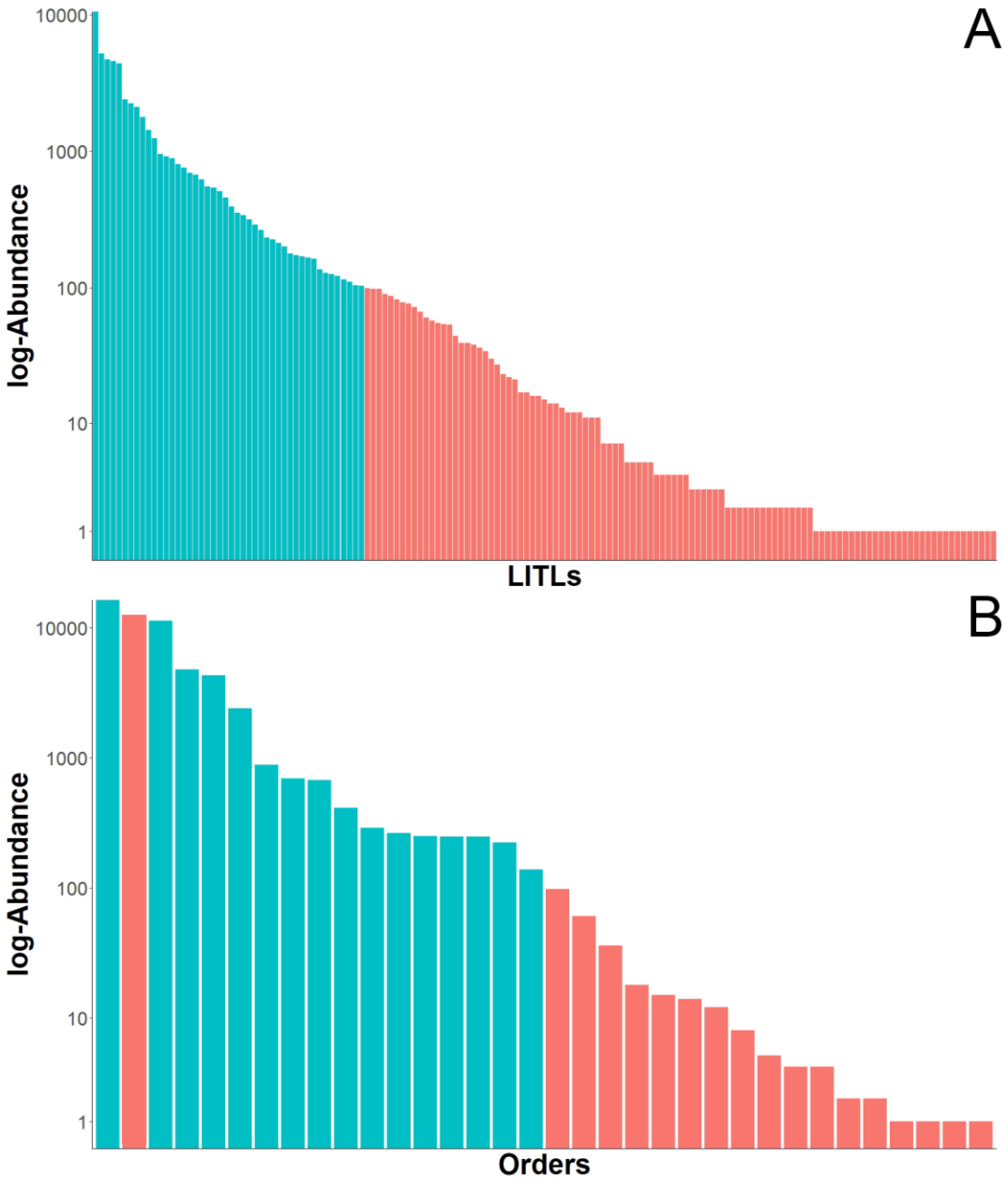


**Figure 1.** Log-scale rank abundance plots of invertebrates in the NEON dataset. (A) Each bar represents a lowest identified taxonomic label (LITL), as identified by MDW. Teal bars represent groups that had equal to or greater than 100 individuals and were included in the training dataset, while red bars represent groups that had fewer than 100 individuals and were removed from the training dataset. (B) Each bar represents an order as identified by MDW. Teal bars represent groups that had equal to or greater than 100 individuals and were included in the training dataset, while red bars represent groups that had fewer than 100 individuals or had a LITL above order (e.g. class or phylum) and were removed from the training dataset.

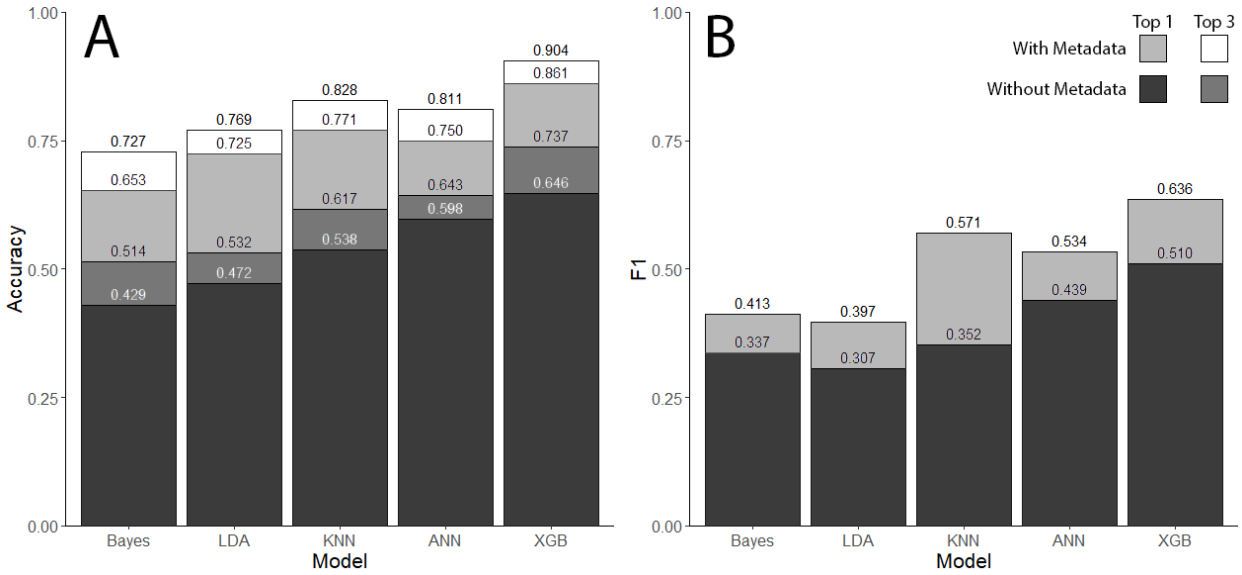


**Figure 2**. Invertebrate bycatch classification performance metrics for models trained with and without contextual metadata (Naïve Bayes (Bayes); Linear discriminant analysis (LDA); K-Nearest Neighbours (KNN); Artificial Neural Network (ANN); Extreme Gradient Boosting (XGB)). All results are shown as the average across 10 models for each algorithm, except for ANN which only has one model. (A) Top 1 and top 3 accuracy. (B) F1 score.

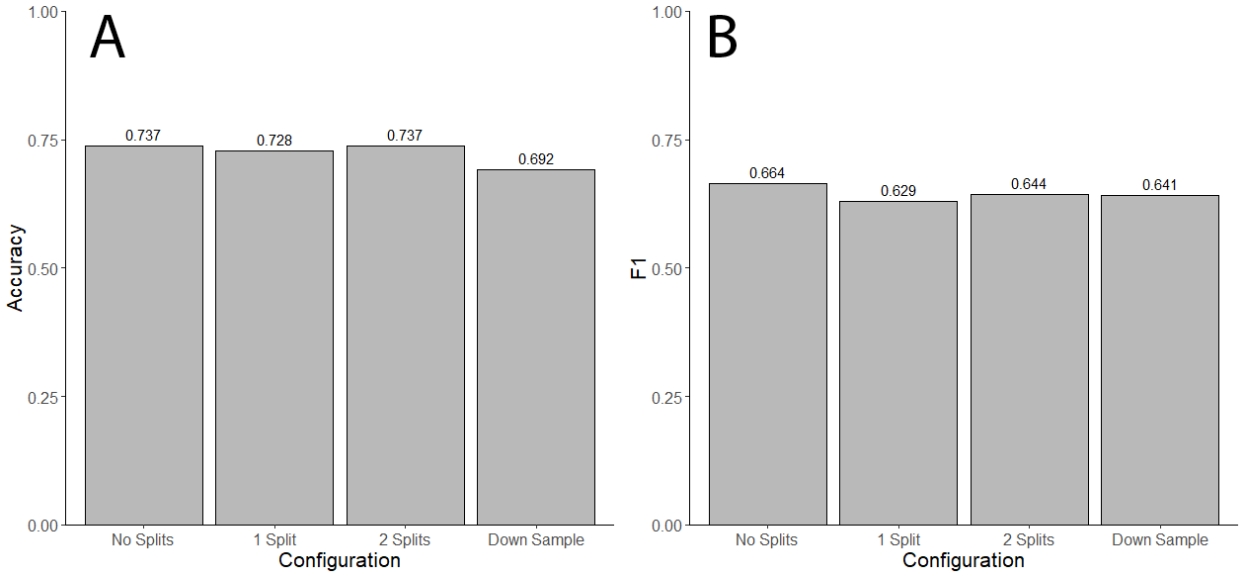


**Figure 3.** Invertebrate bycatch classification performance metrics of the extreme gradient boosting (XGB) model across various model configurations. All results are shown as the average across 10 XGB models trained with contextual metadata. 'No splits' is the performance of the XGB models with no binary classifiers. '1 Split' is the performance of the XGB models when 1 binary classifier is used to separate intact adult specimens from everything else before passing the observations on to a multiclass model. '2 Split' uses the same binary classifier as '1 Split' but passes the observations on to a second binary classifier to separate ant specimens from non-ant specimens before passing the observations on to one of two multiclass models (an ant model or a non-ant model). 'Down Sample' uses no binary classifiers, but the most abundant classes in the training dataset were down-sampled so they had <1000 specimens. (A) Accuracy results of each XGB model configuration. (B) F1 score of each XGB model configuration.

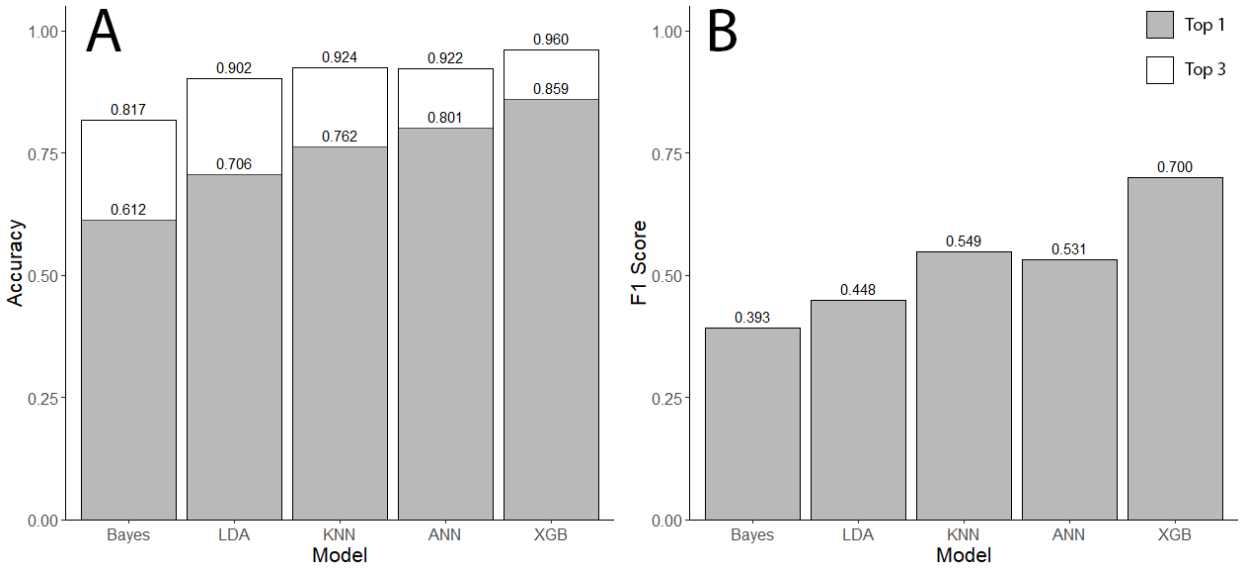


**Figure 4.** Invertebrate bycatch classification performance metrics for models trained at the order level (Naïve Bayes (Bayes); Linear discriminant analysis (LDA); K-Nearest Neighbours (KNN); Artificial Neural Network (ANN); Extreme Gradient Boosting (XGB)). All results are shown as the average across 10 models for each algorithm. (A) Top 1 and top 3 accuracy. (B) F1 score.

## Figures and Tables

| **Algorithm Name** | **Optimization Method** |
|---|---|
| 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. |
| Random Forests | *mtry* values (the number of variables used at each tree node) of all integers between 7 and 15 were tested. The *mtry* value resulting in the highest test dataset accuracy was selected. All random forests contained 1000 trees. |
| eXtreme Gradient Boosting | *max.depth* values (maximum depth of a decision tree) between 7 and 15 were tested at 1000 *nrounds* (max number of boosting iterations). |
| Artificial Neural Network | Each model was trained with a hidden layer containing 25 nodes. |