Journal of Ornithology
https://doi.org/10.1007/s10336-024-02144-5

ORIGINAL ARTICLE

Guidelines for appropriate use of BirdNET scores and other detector outputs

Connor M. Wood[1] · **Stefan Kahl**[1,2]

Received: 16 August 2023 / Revised: 13 December 2023 / Accepted: 2 January 2024

Abstract

Machine learning tools capable of identifying animals by sound have proliferated, making the challenge of interpreting their outputs much more prevalent. These tools, like their predecessors, quantify prediction uncertainty with scores that tend to resemble probabilities but are actually unitless scores that are (generally) positively related to prediction accuracy in species-specific ways. BirdNET is one such tool, a freely available animal sound identification algorithm capable of identifying > 6,000 species, most of them birds. We describe two ways in which BirdNET "confidence scores"—and the output scores of other detector tools—can be used appropriately to interpret BirdNET results (reviewing them down to a user-defined threshold or converting them to probabilities), and provide a step-by-step tutorial to follow these suggestions. These suggestions are complementary to common performance metrics like precision, recall, and receiver operating characteristic. BirdNET can be a powerful tool for acoustic-based biodiversity research, but its utility depends on the careful use and interpretation of its outputs.

Keywords Machine learning · Passive acoustic monitoring · Bioacoustics · Detector

Zusammenfassung

Leitlinien für die geeignete Auswertung von BirdNET- Konfidenzwerten

Die zunehmende Verbreitung von maschinellen Lernwerkzeugen zur Identifikation von Tieren anhand ihrer Geräusche stellt neue Herausforderungen bei der Interpretation der Ergebnisse dar. Diese modernen Tools, wie etwa BirdNET, nutzen zur Vorhersage von Tierarten sogenannte "Konfidenzwerte", die Wahrscheinlichkeiten ähneln, jedoch tatsächlich einheitenlose Zahlen sind. Diese Werte korrelieren üblicherweise positiv mit der Genauigkeit der Vorhersagen, variieren aber je nach Tierart. BirdNET, ein kostenfrei verfügbarer Algorithmus, kann über 6.000 Tierarten erkennen, wobei der Fokus hauptsächlich auf Vögeln liegt. Der Artikel stellt zwei Ansätze zur geeigneten Nutzung der BirdNET-Konfidenzwerte und anderer ähnlicher Detektor-Werkzeuge vor: die Verifikation bis zu einem benutzerdefinierten Schwellenwert und die Umwandlung der Ausgabewerte in Wahrscheinlichkeiten. Zudem bietet er eine detaillierte Anleitung zur Anwendung dieser Methoden. Diese Ansätze ergänzen traditionelle Metriken wie Precision und Recall. Der Artikel verdeutlicht, dass BirdNET ein wirkungsvolles Instrument für die akustische Biodiversitätsforschung sein kann, dessen Nutzen jedoch stark von der sorgfältig en Anwendung und Interpretation seiner Ergebnisse abhängt.

Introduction

Machine learning tools capable of identifying animals by sound have proliferated, making the challenge of interpreting their outputs much more prevalent. These tools, like their predecessors (e.g., template detectors based on spectrogram cross-correlation), generally quantify their predictions in a way that resembles a probability, and indeed for any good tool, prediction accuracy should be positively related to score. However, the fact that these scores are often bounded

Communicated by T. S. Osiejuk.

✉ Connor M. Wood
cmw289@cornell.edu

[1] K. Lisa Yang Center for Conservation Bioacoustics, Cornell Lab of Ornithology, Cornell University, Ithaca, NY, USA

[2] Chemnitz University of Technology, Chemnitz, Germany

Published online: 14 February 2024

between [0–1] causes them to resemble probabilities, which they are not. We use the BirdNET algorithm (Kahl et al. 2021), a freely available animal sound identification tool that is rapidly gaining traction in ornithology and ecology, to illustrate several solutions to the challenge of working with unitless prediction scores. Although technical specifics are unique to BirdNET, the underlying principles are readily transferrable to other sound identification tools. Knight et al. (2017) offer six recommendations for users of animal sound detectors; our suggestions here largely fall under the heading of their guidance on score thresholds and metrics.

BirdNET is a convolutional neural network that was initially designed to identify 984 European and North American Birds; a 2023 release of v2.4 brought the species count to > 6,000 (https://github.com/kahst/BirdNET-Analyzer). BirdNET analyzes audio in 3-s segments, making a prediction for every class (i.e., species) on which it has been trained. By default, predictions with a "confidence score" < 0.01 are suppressed, and users are presented with a "confidence score" spanning 0.01–1.0. Users of the free BirdNET app (Wood et al. 2022) are presented with qualitative outputs (e.g., "likely" or "almost certain") which are derived from the output scores; the original quantitative scores are saved so that researchers can work directly with the data. As noted in a recent review (Pérez-Granados 2023), BirdNET's prediction accuracy increases with confidence score, but substantial confusion remains about how to interpret BirdNET's confidence scores. We briefly describe what BirdNET scores are not, what they are, and several ways to use them.

What BirdNET scores are not and what they are

BirdNET confidence scores are *not* probabilities and they are not related to confidence intervals. BirdNET confidence scores are *not* transferrable among species; the same score may have very different performance outcomes (e.g., precision, recall, pr(true positive)) for any two species. Indeed, scores are not necessarily transferrable among studies of the same species.

Therefore, what *is* BirdNET's confidence score? The confidence score is a unitless, numeric expression of BirdNET's "confidence" in its prediction, and it has been scaled to the range [0–1] via a sigmoid activation function:

$$\text{Confidence score} = \frac{1}{1 + e^{\text{logit score} * \text{sensitivity}}}$$

The "logit score" is the output of a linear classifier that uses feature embeddings to derive an unbounded linear class "score"; "sensitivity" determines the distribution of prediction scores with a value < 1 leads to more higher and lower scoring predictions and a value > 1 leads to more intermediate-scoring predictions[1] Feature embeddings, in turn, are a vector of floating point values that represent the algorithm's characterization of that 3-s audio snippet. BirdNET scores, expressed as "confidence score" or the original logit score, is not the result of a probabilistic model, but rather the final outcome of the deep neural network itself.

BirdNET summarizes much more about input audio than just biological sounds because it analyzes the entirety of a 3-s snippet. Animal sounds, wind, rain, and other sounds are reflected in the feature embeddings and thus its predictions—and so too are characteristics of those sounds as manifested by recording choices like sample rate or microphone quality. Thus, BirdNET scores—and likely those of any CNN—will be sensitive to recording hardware rendering outcome–score relationships non-transferrable. Recording the same birdsong on different recording units, or even the same recording unit but with different sample rates, will not yield identical BirdNET scores. The sensitivity of scores to recording equipment strongly incentivizes careful project planning, as well as coordination among acoustic monitoring projects in the same ecosystem (or operated by the same entity in different ecosystems).

BirdNET's logit scores tend to range ~ -4– ~ 7, though they are technically unbounded. The challenge posed by the logit scores is that they defy fast and easy interpretation (e.g., is 2.61 a "good score"?); scaling them [0–1] and presenting them as "confidence scores" allows for fast and superficially simple interpretation. Yet, that simplicity is deceptive. The confidence scores are not probabilities, they just look like them. BirdNET scores are a unitless expression of the algorithm's prediction for a given class. After analyzing the performance of nearly 1000 species, we have yet to encounter a class in BirdNET for which prediction accuracy does not increase with confidence score. While there is generally a positive relationship between prediction accuracy and confidence score, the shape of this relationship varies across classes, making it important to evaluate the meaning of BirdNET scores on a class-by-class basis.

[1] The technical description of sensitivity is that changes the sigmoid activation curve which determines how much neuron activation (aka less "signal" or "evidence" for a species) is required to reach higher output (confidence) scores. Lower sensitivity yields a steeper curve and increasingly binary outputs; higher sensitivity yields a shallower curve and a more uniform distribution of scores.

Suggestions for working with BirdNET scores

Translating BirdNET confidence scores, or the output scores of any detector, to prediction accuracy requires manual evaluation of model performance on a subset of the data. Familiar signal classification metrics precision and recall can be calculated for any desired BirdNET score threshold. Precision is the proportion of predictions that are correct, or 1—the false positive rate; recall is the proportion of target signals that were correctly identified as such, or 1—the false negative rate. Precision and recall are valuable metrics, but their relative importance depends entirely on the application. For example, low recall could be highly problematic if vocalization counts are being used to estimate animal density or assess behavior (e.g., see review by Pérez-Granados 2023), while low recall may not be problematic at all if animal observations are used to generate detection-corrected population estimates (e.g., occupancy modeling). In the latter case, recall at the level of whole bouts of vocalization, rather than individual sounds, may be a more salient metric (e.g., Fig. 3 in (Wood et al. 2019)).

Using scores as they are

To browse predictions by score, compile all predictions, sort them by score, and begin manually reviewing, starting with the highest scores and moving to lower scores. This approach will be most tractable with only a few species at a time. The lower confidence score limit of the prediction review and validation process will be determined by some balance of the amount of time that can be allocated to manual review and the quality of the data relative to research objectives. For example, Kelly et al. (2023) reviewed > 30,000 BirdNET-based predictions for two different species based on species-specific confidence score thresholds that were determined based on species-specific analyses of precision and recall across a range of scores in a large test dataset. Developing post-processing tools that incorporate ecological and spectral information associated with detector predictions can massively improve precision with a negligible cost to recall (Knight et al. 2020; Leseberg et al. 2020).

While it is good practice to report score thresholds, scores tend to be sensitive to the characteristics of the input audio and are unique to each species, meaning that the numerical score will have little practical value to other users of the same tool. More important than the threshold is the process used to determine it: assessments of precision, recall, and analyst time. If animal observations derived from a detector tool will be used in a model that accounts for imperfect detection (e.g., Brunk et al. 2023; Kelly et al. 2023 and many others), low recall may not be problematic, while behavior-oriented analyses may require high recall.

BirdNET itself can be used to generate a standalone compilation of species-specific predictions above a given confidence score threshold via the “segments” tab of the graphical user interface (GUI) or executing the “segments.py” script via the command line interface (CLI). The result is a folder of validation files, a subfolder for each species that BirdNET has predicted to be present above a user-defined score threshold (specified during the preceding analysis step), and then a user-defined quantity of randomly selected predictions. Specifying at least $^{s}/_{3}$ predictions, where s is the total number of seconds in the audio dataset, will ensure that all predictions above that score are provided. The predictions are saved as ≥ 3-s audio clips, which can rapidly be reviewed to determine if the prediction is correct. This process can be repeated iteratively at progressively lower scores as needed.

Alternatively, BirdNET predictions (and those of other detectors) can be reviewed in Raven Pro (K. Lisa Yang Center for Conservation Bioacoustics). BirdNET-generated selection tables can be added to their source audio files and reviewed using the Selection Review feature, though this tool is presently limited to displaying results within a single selection table. BirdNET has been incorporated into Raven Pro version 1.6.5 + such that one can analyze large batches of audio (i.e., many audio files added together via a “list file), yielding a single selection table per job (e.g., many hundreds of audio files), an approach that is well suited to Raven’s selection review.

Converting scores to probabilities

A second and more quantitative approach to using BirdNET scores and, again, the scores of any sound identification tool, is to link the score directly to the probability that the prediction is correct. For applications in which all predictions cannot be manually confirmed and must, therefore, be treated as putative observations, a probabilistic threshold, rather than unitless one, is extremely important. Examples include single-species but massive-scale surveys entailing many hundreds of thousands of hours of audio (e.g., Brunk et al. 2023), or studies with many tens of species (and possibly also hundreds of thousands of hours of audio) (e.g., Wood et al. 2024). An effective method for generating probabilistic output scores is to review predictions across a range of confidence scores and then using logistic regression to relate prediction outcome to the probability that the prediction is correct, an approach that we have briefly described in the context of occupancy modeling (Wood et al. 2023a) and in a BirdNET-specific tutorial (Symes et al. 2023). For BirdNET users, the “segments” tool described above was designed for this purpose.

First, for each desired species or class, generate a random selection of BirdNET predictions spanning the full range of BirdNET scores. If the validation samples have been randomly selected from all possible predictions, the distribution of BirdNET scores in the sample will mirror the distribution of scores for the species; for infrequently encountered species (or species with intrinsically low classification performance), this can skew the validation dataset towards low-scoring predictions, which are generally unlikely to be the target sound. In such cases, it can be valuable to generate a second set of randomly selected validation samples from a higher range of scores and add these to the original dataset, providing additional resolution in the range of scores where false positives transition to true positives.

Second, review these predictions and indicate whether each prediction is correct or incorrect. When the “segments” feature exports the predictions as.wav files, it appends the confidence score associated with each prediction to the beginning of the file name, enabling users to link the score with the outcome. Thus, sorting the audio files enables one to associate the binary outcome of a BirdNET prediction (correct or incorrect) with continuous BirdNET confidence scores. The process of reviewing BirdNET predictions, either randomly generated ones or starting from high scores and progressing down, can also be a valuable opportunity to understand which acoustic signals BirdNET is assigning to a given class. At least for North American and European birds, BirdNET can often identify species via multiple signals, such as a song and one or more calls, as well as woodpecker drumming. Conducting intraspecific sound classification has many exciting possibilities that are beyond the scope of this paper but which we have introduced elsewhere (McGinn et al. 2023).

Third, use logistic regression to relate the binary outcome of the validation process to the continuous BirdNET score, where p = the probability that a BirdNET prediction is correct.

$$\log it(p) = \log \frac{p}{1-p} = \beta_0 + \beta_{\text{BirdNET\_score}}.$$

Critically, the logistic regression model can then be solved for any desired p:

$$\text{BirdNET score threshold} = \frac{\ln\left(\frac{p}{1-p}\right) - \beta_0}{\beta_{\text{BirdNET\_score}}}.$$

These equations can be implemented easily in Program R (R Core Development Team 2020). The following R code, an expanded version of which is provided in the Supplementary Materials 1 and 2, assumes that the validated predictions have been imported as a data frame called ‘validation’ with columns ‘outcome’ (containing 1 s and 0 s for correct and incorrect predictions, respectively) and ‘score’ (containing the corresponding BirdNET confidence scores).

```
> your_model <- glm(outcome ~ score, family = "binomial", data = validation)
> p <- 0.9 # your desired p (probability of true positive)
> threshold <- (log(p/(1-p))- your_model$coefficients[1]) / your_model$coefficients[2]
```

Using the logistic regression approach, BirdNET scores are no longer unitless: they can be described in terms of probabilities.

Probabilistic score thresholds will be applicable to the range of conditions encompassed by data from which they were drawn. Sampling the same location but in a different season, or a different location in the same conditions could warrant another round of validation to test the transferability of score threshold. If monitoring projects grow to encompass multiple ecoregions, continued validation efforts are warranted. The generalizability of the logistic regression-based outcome–score relationship can be tested explicitly by simply including spatial and/or temporal prediction covariates, such as:

$$\log it(p) = \beta_0 + \beta_{\text{BirdNET\_score}} + \beta_{\text{season}} + \beta_{\text{latitude}} + \beta_{\text{ecosystem}}.$$

The support for competing logistic regression models can be evaluated with Akaike’s Information Criterion (Burnham and Anderson 2010), yielding a more detailed understanding of how BirdNET performance may vary across space and time.

Note that we did not specify “confidence score” or logit score in the preceding three equations. Either can be used, but we suggest transforming confidence scores back into the original logit scale, because the highest scores tend to be compressed in the [0–1] scale. Unbounded logit scores have an intrinsically broader distribution and are thus more amenable to logistic regression. We suggest doing the conversion process in the R environment after importing the validation data:

$$\text{Logit score} = \ln\left(\frac{\text{confidencescore}}{1-\text{confidencescore}}\right) \Big/ \text{sensitivity}$$

It is important to note that BirdNET analyzer scores can be back-transformed, but BirdNET app observations cannot be back transformed, because app observations are subject to a pooling calculation that averages multiple scores when users submit recordings > 3s.

Probabilistic thresholds are quite valuable, because dataset can be described in consistent, quantitative terms (e.g., “for all 50 species, we only included putative observations with pr(tp) ≥ 0.95”). If all putative observations cannot be validated, it is very important to understand (and disclose) the estimated error rates. The potentially serious implications of false positives have been studied extensively (Clare et al. 2021); fortunately, there are multiple approaches to addressing this challenge. First, users can apply both a probabilistic threshold to the BirdNET results and a temporal one, such that, for example, observations with a pr(tp) ≥ 0.95 are required on two or more days of a survey in order to consider a species "present" (Wood and Peery 2022; Brunk et al. 2023). Second, observation-specific probabilities are compatible with population models designed to accommodate false positives (e.g., Wood et al., unpublished). However, the same probabilistic threshold may yield different recall across species (Wood et al. 2023b), so it is not perfect solution to standardization. Thus, applications in which vocalization counts are used in interspecific comparisons, recall may still be a more important metric.

Conclusion

In summary, we have outlined the technical origins of BirdNET scores and several approaches to using these scores appropriately. Although much of the text relates to BirdNET, the principles we outline for generating probabilistic scores can be applied to the output of any detector. We have also created a step-by-step tutorial to help users implement these suggestions (Supplementary Materials 1 and 2). These approaches can facilitate the appropriate interpretation and use of BirdNET scores, but they cannot necessarily reveal which score should be used in any given application. Standard detector performance metrics such as precision, recall, and receiver operating characteristic (ROC) curves are still informative and potentially necessary; probabilistic cores are not necessarily a substitute for these metrics. For a more expansive treatment of the implications of different detector tools and score thresholds, see (Knight et al. 2017).

For each 3-s snippet of audio, BirdNET uses its feature embeddings to generate a numeric prediction on an unbounded logit scale for each species; these logit scores are converted to "confidence scores" to constrain all values [0–1] and, by default, predictions with a confidence score < 0.01 are suppressed. The greater the confidence score, the more likely a prediction is to be correct, but confidence scores are not probabilities. Confidence scores are also not transferrable among species or recording settings. Thus, (*i*) species-specific evaluation is required, and (*ii*) recording settings should be as consistent as possible to avoid repeating the validation process. Maintaining standardized test datasets for the evaluation process can also be valuable when comparing performance across models. Species-specific validation can take the form of manual verification of results down to a user-defined confidence score threshold. For applications that would be enhanced by probabilistic scores, as opposed to unitless confidence (or logit) scores, logistic regression can be used to relate the accuracy of a random set of BirdNET predictions to BirdNET score, preferably, the logit score rather than the confidence score. BirdNET can be a powerful tool for acoustic-based biodiversity research, but its utility depends on the careful use and interpretation of its outputs.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10336-024-02144-5) contains supplementary material, which is available to authorized users.

Acknowledgements We thank Laurel Symes and Larissa Sugai for insightful comments that improved this manuscript. The BirdNET project is supported by Jake Holshuh (Cornell class of '69) and The Arthur Vining Davis Foundation. Our work in the K. Lisa Yang Center for Conservation Bioacoustics is made possible by the generosity of K. Lisa Yang to advance innovative conservation technologies to inspire and inform the conservation of wildlife and habitats. The German Federal Ministry of Education and Research is funding the development of BirdNET through the project "BirdNET+" (FKZ 01|S22072). Additionally, the German Federal Ministry of Environment, Nature Conservation and Nuclear Safety is funding the development of BirdNET through the project "DeepBirdDetect" (FKZ 67KI31040E).

Author contributions SK created BirdNET, and both authors contribute directly to its ongoing development.

Data availability A step-by-step tutorial, R code, and sample data are available in the Supplementary Material.

Declarations

Conflict of interest The authors declare no competing financial interests.

References

Brunk KM, Gutiérrez RJ, Peery MZ, Cansler CA, Kahl S, Wood CM (2023) Quail on fire: changing fire regimes may benefit mountain quail in fire-adapted forests. Fire Ecol 19:19

Burnham KP, Anderson DR (2010) Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York

Clare JDJ, Townsend PA, Zuckerberg B (2021) Generalized model-based solutions to false-positive error in species detection/non-detection data. Ecology 102:e03241

Kahl S, Wood CM, Eibl M, Klinck H (2021) BirdNET: A deep learning solution for avian diversity monitoring. Eco Inform 61:101236

Kelly KG, Wood CM, McGinn K, Kramer HA, Sawyer SC, Whitmore S, Reid D, Kahl S, Reiss A, Eiseman J, Berigan W, Keane JJ, Shaklee P, Gallagher L, Munton TE, Klinck H, Gutiérrez RJ, Peery MZ (2023) Estimating population size for California spotted owls and barred owls across the Sierra Nevada ecosystem with bioacoustics. Ecol Ind 154:110851

Knight E, Hannah K, Foley G, Scott C, Brigham R, Bayne E (2017) Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. Avian Conserv Ecol 12:14

Knight EC, Sòlymos P, Scott C, Bayne EM (2020) Validation prediction: a flexible protocol to increase efficiency of automated acoustic processing for wildlife research. Ecol Appl 30:e02140

Leseberg NP, Venables WN, Murphy SA, Watson JEM (2020) Using intrinsic and contextual information associated with automated signal detections to improve call recognizer performance: A case study using the cryptic and critically endangered Night Parrot Pezoporus occidentalis. Methods Ecol Evol 11:1520–1530

McGinn K, Kahl S, Peery MZ, Klinck H, Wood CM (2023) Feature embeddings from the BirdNET algorithm provide insights into avian ecology. Eco Inform 74:101995

Pérez-Granados C (2023) BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165:1068–1075

R Core Development Team (2020) R: a language and environment for statistical computing

Symes L, Sugai LS, Gottesman B, Pitzrick M, Wood CM (2023) Acoustic analysis with BirdNET and (almost) no coding: practical instructions. 10.5281/zenodo.8357176

Wood CM, Peery MZ (2022) What does 'occupancy' mean in passive acoustic surveys? Ibis 164:1295–1300

Wood CM, Popescu VD, Klinck H, Keane JJ, Gutiérrez RJ, Sawyer SC, Peery MZ (2019) Detecting small changes in populations at landscape scales: a bioacoustic site-occupancy framework. Ecol Ind 98:492–507

Wood CM, Kahl S, Rahaman A, Klinck H (2022) The machine learning–powered BirdNET App reduces barriers to global bird research by enabling citizen science participation. PLoS Biol 20:e3001670

Wood CM, Barceinas Cruz A, Kahl S (2023a) Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modeling of an endangered primate. Am J Primatol 85:e23507

Wood CM, Kahl S, Barnes S, Van Horne R, Brown C (2023b) Passive acoustic surveys and the BirdNET algorithm reveal detailed spatiotemporal variation in the vocal activity of two anurans. Bioacoustics 32:532–543

Wood CM, Socolar J, Kahl S, Zachariah Peery M, Chaon P, Kelly K, Koch RA, Sawyer SC, Klinck H (2024) A scalable and transferable approach to combining emerging conservation technologies to identify biodiversity change after large disturbances. J Appl. Ecol. https://doi.org/10.1111/1365-2664.14579

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.