BirdNET is as good as experts for acoustic bird monitoring in a European city

Andrew J. Fairbairn[1*], Josija-Simeon Burmeister[1,2], Wolfgang W. Weisser[1], Sebastian T. Meyer[1].

[1] Terrestrial Ecology Research Group, Department of Life Science Systems, School of Life Sciences, Technical University of Munich, 85356 Freising, Germany

[2] Faculty of Forest Sciences and Forest Ecology, Georg August University of Göttingen, 37077 Göttingen, Germany

* Corresponding Author:

Email: afairbainr85@googlemail.com

ORCID: 0000-0002-2826-3465

Abstract

BirdNET has become a leading AI for recognising bird species in audio recordings. However, the effect of parameter settings on the algorithm's performance is not well understood, leaving questions about its applicability in ecological research. We compared the species identification of BirdNET against expert identification using an acoustic dataset consisting of 930 minutes of recordings collected in Munich, Germany. The performance of BirdNET was evaluated using three performance metrics: precision, recall, and F1-score. The metrics were calculated using 24 combinations of the parameters week, sensitivity, and overlap at four temporal aggregations, to also test the effects of recording length. We found that BirdNET closely matched expert identification, particularly when given more data (higher temporal aggregation) and when including the parameters week of the year, a suitable sensitivity, and an overlap of one to two seconds. Thus, our study shows that BirdNET can provide species lists and richness comparable to that of expert ornithologists in acoustic surveys in urban environments. We provide a set of recommendations and parameter settings for BirdNET.

Keywords:

BirdNET, acoustic monitoring, biodiversity monitoring, urban ecology

Statements and declarations

Competing interests

The authors declare no competing interests.

Introduction

BirdNET has emerged as a leading tool for automatically identifying birds in acoustic recordings. Currently, BirdNET version 2.4 has global coverage of over 6,500 avian and non-avian classes (Kahl et al. 2021). In Germany,

BirdNET covers 407 of the 527 bird species tracked by the German Ornithologists Society (Barthel and Krüger 2019) including rare and vagrant species. BirdNET is already widely used in industry and is becoming more prevalent in scientific literature (e.g. Manzano-Rubio 2022, Cole et al. 2022, Bota et al. 2023). However, using this technology in ecological research requires rigorous testing to understand its performance. Currently, how BirdNET parameter settings affect its performance may be perplexing to an ecologist or downright unknown (Pérez-Granados 2023) questioning the efficacy of passive acoustic monitoring with automatic identification of bird songs for ecological research.

In an urban environment, we compared the performance of BirdNET to expert identification. Specifically, we: a) tested how varying BirdNET parameters affects its performance, b) investigated the effects of using different temporal resolutions on performance. Finally, we provide guidelines for using BirdNET to conduct acoustic bird surveys in cities.

Methods

Acoustic recording

We placed a single Frontier Labs BAR on the roof of a housing complex in Munich, Laim, between May and October 2021. We recorded in week-long blocks, with a minimum of one week between recording periods. We recorded one minute every 10 minutes from two hours before sunrise to three hours after sunrise (Abrahams 2018) to keep the amount of manual identification required to a manageable level resulting in a total of 15.5 hours of recordings over 60 days. Recordings were taken at a sample rate of 48kHz, a bit depth of 16 and a gain of 40dB.

Species Identification

Two experts (A. Fairbairn, J.S. Burmeister) identified all bird calls in each recording visually and aurally using Wildlife Acoustics Kaleidoscope (Wildlife Acoustics 2021). Next, we ran BirdNET on the same recordings, producing a list of species detections for each one-minute recording.

Analysis

Parameter effects

We tested four parameter settings. First, BirdNET can use the *week* of the year of recording in conjunction with the location to filter what species are likely to occur at a location at that time of the year. We ran all analyses with or without week included. Second, the *detection sensitivity* (range 0.5 to 1.5, default 1.0) affects how sensitive BirdNET is to faint or background calls. We ran all analyses with three sensitivity levels (0.5, 1.0, 1.5). Third, BirdNET works on three-second audio segments for analysis. The *overlap* determines how much of the previous three-second segment is “overlapped” (default 0.0s). We ran all analyses with four levels of overlap (0, 1, 2, 2.9 seconds). Finally, BirdNET provides a *confidence level*, the confidence that BirdNET has in its own predictions (Wood and Kahl 2024), for each detection. Setting a minimum confidence level in BirdNET causes all detections with a lower confidence to be removed from the results. To identify the best settings, including the best minimum confidence level, we ran all analyses with the lowest minimum confidence level to get a full list of detections that

we could filter afterwards for the analyses. To understand the effects of each parameter, we created a mixed effects model including the temporal aggregation (see below), sensitivity, overlap, minimum confidenc as fixed and each run of BirdNET as a random effect (*lme* from the *nlme* package (Pinheiro et al. 2021); Supplement 1).

Temporal aggregation

To test if the amount of available recordings affects BirdNET performance, we ran the analyses on four levels of temporal aggregation: minute (no aggregation), day, week, and the whole dataset (Table 1). We aggregated expert and BirdNET results by creating a list of all unique species at each aggregation level. For example, species identified during any of the minutes on a day resulted in a presence of the species at the day aggregation.

BirdNET vs expert

Compared to expert identification, the BirdNET results can be: true positives (correctly identified species), false positives (incorrectly identified species), or false negatives (species occurring that were not identified). Based on these, we calculated precision, recall, and F1 to test the performance of BirdNET. Precision (Eq. 1) measures the proportion of species identified by BirdNET that are correct. Recall (Eq. 2) measures how many species identified by the expert are correctly identified by BirdNET. Because of a trade-off between recall and precision, i.e., maximising one decreases the other, we use F1-scores as the primary measure of performance, which provide a balanced integration of precision and recall (Eq. 3).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{1}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Confirmation test

In the above test of BirdNET performance, we assumed the expert identification to be perfect, but incorrect identification or missed species may occur despite our best efforts. Therefore, we conducted a confirmation test. Following Sethi et al. (2021), we randomly selected 50 detections for each species (or all detections if a species was detected less than 50 times) from the best BirdNET results for the dataset aggregation level and listened to them, confirming them as correct or incorrect.

Results

We recorded 930 minutes over five weeks and expertly identified 9466 calls of 23 different bird species (Supplement Table 2-1). With default settings, and including week of the year, BirdNET identified 93 species (Supplement Table 2-2).

BirdNET vs expert

Comparing BirdNET with expert identification, temporal aggregation and all tested parameters (minimum confidence, week, sensitivity, and overlap) significantly affected BirdNET performance (F1 scores ; Fig. 1, Supplementary Table 1). Including the *week* of the year consistently provided better results (Supplement Fig. 3-1). Minimum confidence had the strongest effect on F1 score ($p < 0.001$) and significantly interacted with temporal aggregation, sensitivity, and overlap (Fig. 1, Supplement Table 1-1). As aggregation increased from minute resolution to the whole dataset, maximum F1 scores improved from 0.62 to 0.84, respectively, when using optimal combinations of minimum confidence, sensitivity, and overlap (Fig. 1, Table 2). With the exception of high confidence levels for the minute and day aggregations, decreasing overlap provided higher maximum F1 scores (Fig. 1, Supplement Table 4-1).

Default BirdNET parameter settings performed poorly when considering F1 scores (Table 2). However, default settings provided higher recall, maximising the number of true positives while inflating false positives.

Confirmation test

In our confirmation test using the best parameter settings identified above for data aggregated to the whole dataset, BirdNET detected 20 species, two of which (*Delichon urbicum*, Western house martin, *Turdus philomelos*, Song thrush) were false positives, i.e., species not occurring in the original expert identification. However, upon listening to these BirdNET detections, we determined the 14 Song thrush detections to be correct, while the single detection for Western house martin was indeed a false positive (Fig. 2), resulting in an adjusted F1 score for the dataset resolution of 0.86 (Precision= 0.95, Recall=0.792).

Discussion

Current acoustic monitoring practices recommend short recordings during peak activity periods (Abrahams 2018; Metcalf et al. 2023), to reduce manual validation effort. However, artificial intelligence eliminates the need to listen to entire recordings, allowing for longer and more frequent data collection. As highlighted by our temporal aggregation results, BirdNET performs better with longer recordings, which also enhance diversity monitoring by capturing species with varying activity patterns (Robbins 1981).

While the best settings for BirdNET parameters varied depending on the temporal aggregation level, we can deduce some generalisations for running BirdNET. Since our best-performing parameter settings always included the week of the year, we recommend for general monitoring to include week. Overlap had a more significant impact on performance than sensitivity. We recommend using an overlap between one and two seconds for short one to five-

minute recording schemes. With longer or continuous recordings, an overlap may not be necessary as calls that may be unidentifiable due to being cut up are likely to occur again. Default sensitivity (1.0) generally produced the same results as higher or lower sensitivities while maintaining higher minimum confidence levels. We, therefore, recommend using the default sensitivity. For short recordings of 30 minutes or less, we suggest filtering by a minimum confidence of 0.54 or higher. We recommend filtering by a minimum confidence of 0.74 for recordings around an hour and an even higher minimum confidence for longer recordings.

When using BirdNET in ecological studies, we recommend validating the results. Only manually checking odd or uncommon species is sufficient if the researcher is familiar with what is likely to occur on their study site. If confirming the presence of the species at a location (i.e. producing a species list) is the research goal, validation can be done easily by manually checking the top results for each species/site, as only one confirmed detection is needed.

To increase the effectiveness of BirdNET, we suggest some avenues for further investigation. One approach would be using a sound separation algorithm which can distinguish bird calls from environmental or anthropogenic noise and separate them from the recordings. After this, these calls can be run through BirdNET, potentially reducing the likelihood of BirdNET misidentifying non-avian sounds as birds. In addition, we suggest investigating individual species' confidence thresholds rather than using a universal threshold. Going one step further, probability curves could be calculated to determine the likelihood that an individual detection is correct (Wood and Kahl 2024).

Here, we showed that BirdNET's performance can be optimised using adapted parameter settings and recording for longer periods. With these, passive acoustic monitoring using BirdNET is suitable for monitoring avian biodiversity in urban areas in Germany, offering new research opportunities in ecology and ornithology.

Acknowledgements

We thank the German Research Foundation (DFG Research Training Group 2679 - Urban Green Infrastructure) for funding this research. We would also like to thank Julia Windl and Lisa Maier for assisting with identifying the bird recordings. Additionally, we would like to thank Münchener Wohnen for permitting our recording.

Tables

Table 1. Breakdown of how the data was aggregated to get the different temporal resolutions.

aggregation	length	total comparisons
minute	1 minute	930
day	31 minutes	30
week	155 minutes*	5
dataset	930 minutes	1

* Two weeks varied slightly by the number of recordings.
155 minutes is the average

Table 2. Best settings resulted in the highest F1 scores for each temporal aggregation level from 1,944 BirdNET and expert identification comparisons.

aggregation	week	sensitivity	overlap	minConf	precision	recall	F1 score	true positives	false positives	false negat
minute*	yes	1	0	0.1	0.35	0.69	0.46	1196	2267	546
minute	**yes**	**0.5**	**2**	**0.56**	**0.70**	**0.56**	**0.62**	**975**	**426**	**767**
minute	yes	1	2	0.54	0.70	0.56	0.62	975	426	767
minute	yes	1.5	2	0.52	0.70	0.56	0.62	975	426	767
day*	yes	1	0	0.1	0.38	0.88	0.53	328	538	44
day	**yes**	**0.5**	**1**	**0.57**	**0.73**	**0.75**	**0.74**	**278**	**102**	**94**
week*	yes	1	0	0.1	0.34	0.96	0.50	92	181	4
week	yes	0.5	1	0.83	0.77	0.74	0.76	71	21	25
week	**yes**	**1**	**1**	**0.74**	**0.77**	**0.74**	**0.76**	**71**	**21**	**25**
week	yes	1.5	1	0.63	0.77	0.74	0.76	71	21	25
dataset*	yes	1	0	0.1	0.25	1.00	0.40	23	70	0
dataset	yes	1.5	0	0.79	0.90	0.78	0.84	18	2	5
dataset	**yes**	**1.5**	**0**	**0.8**	**0.90**	**0.78**	**0.84**	**18**	**2**	**5**

* Default BirdNET settings with week included. Bold are the best settings we recommend

Figures

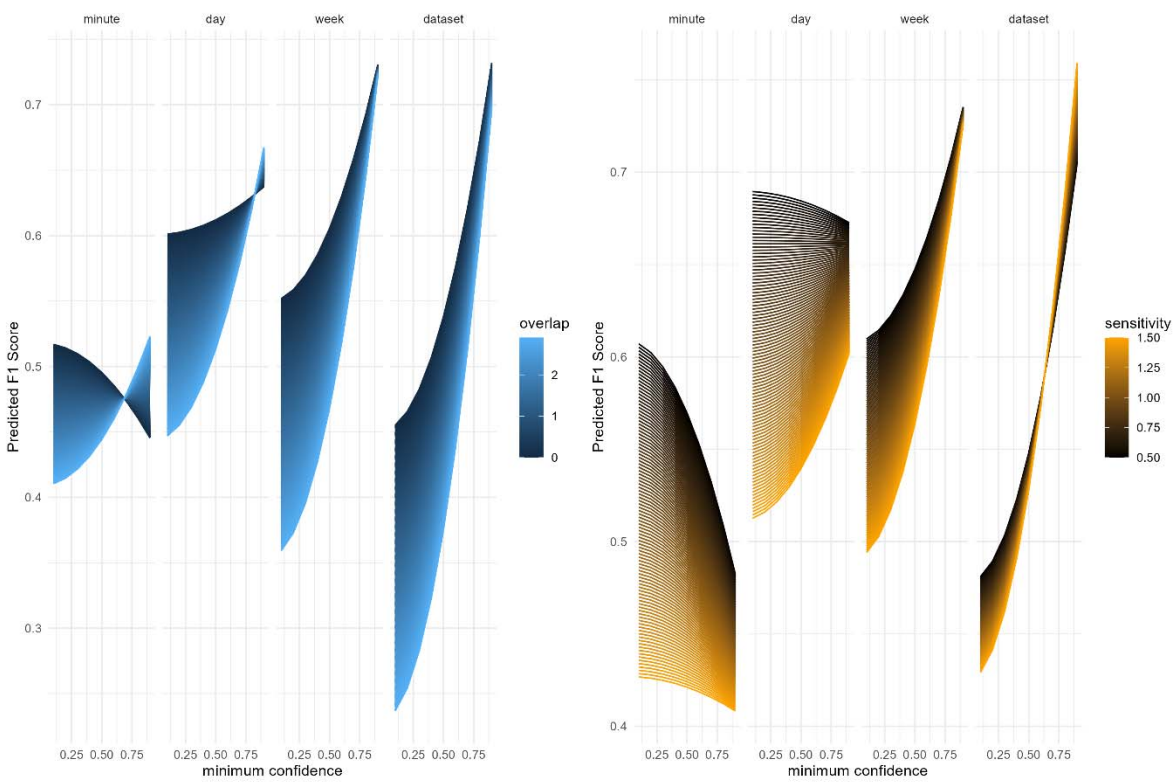

Figure 1. Predicted F1 score as a function of data aggregation level, overlap, sensitivity and minimum confidence from a liner mixed-effects model based on 1,944 comparison tests between BirdNET and an expert ornithologist. Each panel shows the variation in predicted F1 scores across each aggregation level, minimum confidence and overlap (left) or sensitivity (right).

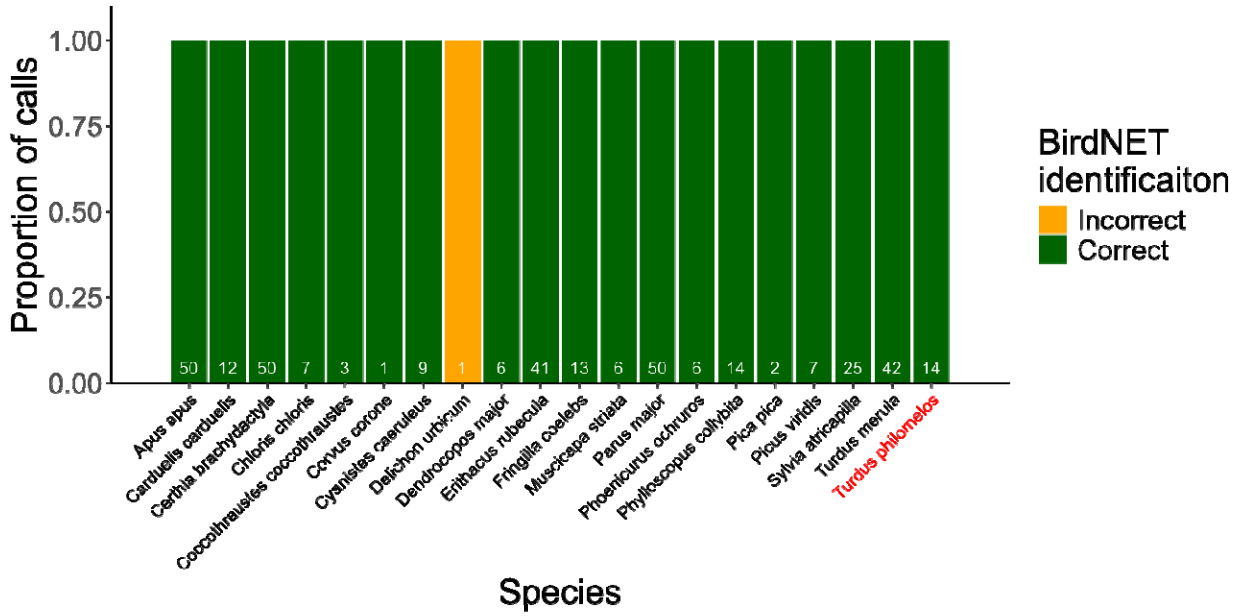

Figure 2. The proportion of BirdNET detections from the dataset resolution that were manually checked by an expert and determined to be correct or incorrect. A maximum of 50 random detections per species were checked. If a species was detected less than 50 times, all detections were checked. White numbers are the number of detections checked. The red label denotes the species missed in our expert identification that BirdNET identified.

References

Abrahams C (2018) Bird Bioacoustic Surveys – developing a standard protocol. In Practice 20–23

Barthel PH, Krüger T (2019) Liste der Vögel Deutschlands: Version 3.2

Bota G, Manzano-Rubio R, Catalán L, et al (2023) Hearing to the Unseen: AudioMoth and BirdNET as a Cheap and Easy Method for Monitoring Cryptic Bird Species. Sensors 23:7176. https://doi.org/10.3390/s23167176

Cole JS, Michel NL, Emerson SA, Siegel RB (2022) Automated bird sound classifications of long-duration recordings produce occupancy model outputs similar to manually annotated data. Ornithological Applications 124:1–15. https://doi.org/10.1093/ornithapp/duac003

Kahl S, Wood CM, Eibl M, Klinck H (2021) BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics 61:. https://doi.org/10.1016/j.ecoinf.2021.101236

Manzano-Rubio R (2022) Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species. Ecological Informatics 72:. https://doi.org/10.1016/j.ecoinf.2022.101910

Metcalf O, Abrahams C, Ashington B, et al (2023) Good practice guidelines for long-term ecoacoustic monitoring in the UK. The UK Acoustics Network

Pérez Granados C (2023) BirdNET: applications, performance, pitfalls and future opportunities. Ibis 165:1068–1075. https://doi.org/10.1111/ibi.13193

Pinheiro J, Bates D, DebRoy S, et al (2021) nlme: Linear and nonlinear mixed effects models

Robbins CS (1981) Effect of time of day on bird activity. Studies in Avian Biology 6:275–286

Wildlife Acoustics (2021) Kaleidoscope Pro 5

Wood CM, Kahl S (2024) Guidelines for appropriate use of BirdNET scores and other detector outputs. J Ornithol 165:777–782. https://doi.org/10.1007/s10336-024-02144-5