Bird Census News 2025, 38/1–2: 14–25

Applying Passive Acoustic Monitoring for Dutch common bird species: evaluating detection and classifier performance

Maja Roodbergen, Chiel Boom and Niels van Harten

Sovon Dutch Centre for Field Ornithology, Postbus 6521, 6503 GA Nijmegen, The Netherlands
maja.roodbergen@sovon.nl

Abstract. To evaluate the usability of audio-recordings for monitoring, we recorded audio data during 232 ten-minute point counts in agricultural habitat. All recordings were automatically classified, using four classifiers, and 70 recordings were annotated manually. Finally, two new classifiers were developed by training an existing classifier using annotated recordings of 326 species occurring in the Netherlands, with and without use of secondary species labels. Here we show the results of comparisons between 1) the field data of the point counts vs the manually annotated audio data, and 2) the automatic classifications by the four existing and the two new classifiers.

Introduction

Being a densely populated country, with many interested and skilled birders, monitoring of bird numbers and distribution is extensive in the Netherlands. Monitoring schemes are coordinated by Sovon Dutch Centre for Field Ornithology and carried out by c. 10,000 volunteers, which go out in the field to count birds. Though extensive, some species, habitats and/or periods are underrepresented, due to (e.g.) detectability, accessibility and attractiveness issues. Technological advances in low-cost automatic audio recording devices in combination with developments in AI facilitating automatic sound recognition, have opened the potential for acoustic monitoring (e.g. Browning et al. 2017, Shonfield & Bayne 2017).
However, being a relatively new method, exploring the comparability of acoustic and regular existing monitoring methods is pivotal to assess the applicability of acoustic monitoring and to avoid methodological trend breaks. In addition, classifying audio recordings manually is time consuming. Automatic classification using classifiers, developed and trained for this purpose (e.g. Kahl et al. 2021), could greatly improve efficiency. However, thus far, classifiers have been trained using species-specific recordings and may perform inadequately for soundscapes with multiple species vocalising simultaneously as happens during spring choruses. Analyses using classifiers which miss and/ or misidentify vocalising local species produce incorrect presence-absence data. Moreover, species differ in their detectability and identifiability, making between species comparisons difficult. Validating present, freely available classifiers is therefore a prerequisite for their practical implementation.
A habitat type that is typically underrepresented in Dutch monitoring schemes is intensively used farmland, as it contains few and usually only common species and therefore is of little interest to birders. As it is dynamic and constitutes a large part of the land surface of the Netherlands, improving coverage of this habitat type is desirable. Breeding bird monitoring in agricultural sites in the Netherlands to a large degree consists of point counts (Teunissen et al. 2019). Data from audio recordings from a static location can best be compared to point counts (e.g. Klingbeil & Willig 2015, Van Wilgenburg et al. 2017). Therefore, while performing point counts in the agricultural landscape of the province of Noord-Brabant, field observers simultaneously recorded bird sounds using a recorder. We used these data to answer the questions: 1) 'How do presence-absence data collected in agricultural landscapes using audio recorders differ from data collected using point counts?'; 2) 'What is the performance of four freely available bird sound classifiers when it comes to confirming the occurrences of farmland bird species?' and

3) 'Can the performance of the best of these four classifiers be improved for Dutch bird species?'

Methods

Data collection

In 2022, four different field workers performed a total of 232 point counts in 13 agricultural sites with Agri-Environment Schemes in the province of Noord-Brabant, with 5 points per site, repeated 4 times during the breeding season. Counts were performed between sunrise and 5 hours after sunrise and lasted 10 minutes, during which all individuals tied to the location (i.e. excluding fly-overs) within a radius of 300 m around the point were registered on a map, noting species and breeding code (Teunissen et al. 2019).
During the point counts, bird sounds were recorded using an AudioMoth located at observer height and set at a sample rate of 48 kHz and medium gain.

Evaluation data set

For the comparison of audio data with point counts and the performance tests of the classifiers, 70 audio recordings were selected (Table 1). We aimed at including all sites and all four rounds, but prioritised two sites where species diversity was relatively high (Maasheggen and Schijndel). The selected audio-recordings were annotated manually by three expert field workers from Sovon Dutch Centre for Field Ornithology, by composing a list of all species heard during the 10-minute recordings.
The 70 recordings used in the analyses contain 759 minutes of audio with 864 annotations for 73 species, after removal of all uncertain annotations. The Common Chaffinch *Fringilla coelebs* was the most frequently recorded species appearing in 63 out of 70 recordings, while 13 species only occurred in a single recording. The soundscapes included between 4 and 22 annotations, averaging 12 annotations per soundscape. All audio had a bit rate of 768 kbps, though the quality of the recordings varied. Some soundscapes suffered from disturbances from wind, traffic or field observers.
The evaluation dataset used to answer the first question (comparing audio with point counts) contained 1332 records (unique species-point-visit combinations, Table 1) from 91 species that were present on the audio-recording and/or in the point count dataset.

Table 1. Number of records (unique species-point-visit-combinations) in the evaluation dataset per point and visit (records from both point counts and audio-recordings).

		Visit			
Site	Point	1	2	3	4
De Bleken	dblk1	0	0	16	17
Gastelse Heide	gsth1	0	12	20	13
	gsth2	0	17	0	0
Keent	kent1	32	31	26	0
	kent2	35	22	0	0
	kent4	26	0	0	0
	kent5	22	0	0	0
Lage Zwaluwe	lgzw1	0	0	22	14
Made Noord	mdnd1	0	0	11	18
Maasheggen	mshg1	10	22	16	17
	mshg2	21	22	22	21
	mshg3	20	21	19	20
	mshg4	20	21	21	19
	mshg5	21	19	16	14
Rucphen Heikant	rcph1	0	0	25	19
Rielsche heide	rlhd1	0	0	0	8
Reek-Schaijk	rsch1	20	19	18	0
	rsch2	0	19	0	0
Schijndel	schd1	25	22	20	15
	schd3	18	17	18	9
	schd4	18	20	23	18
	schd5	16	20	15	12
	schd7	18	20	25	16
Steenbergen Noord	stbn1	0	0	16	10
Zeeland	zlnd1	20	20	17	0
	zlnd2	0	20	0	0

Model descriptions

We compared four bird sound classification models: BirdNET, the Google Bird Vocalization
Classifier (GBVC), AvesEcho and Aquila. BirdNET and GBVC are global models, whereas
AvesEcho focuses on Europe and Aquila on the Netherlands.
BirdNET (v2.4, Kahl et al. 2021) is a neural network classifying over 6500 bird species. It operates on three-second-long audio segments. The model is trained using recordings from Xeno-Canto, the Macaulay Library of Natural Sounds and most likely additional unknown sources as the model has been updated after publishing. We

used an overlap of two seconds resulting in one prediction for every second of audio. We did not use the available post-processing mask to filter results based on location and time of the year in our evaluation as it might exclude species which could be present. Instead, we used a selection of 326 bird species occurring in the Netherlands.

The Google Bird Vocalization Classifier (GBVC, v4) is trained on Xeno-Canto recordings and classifies over 10,000 bird species based on 5-second-long audio segments. We inputted segments with an overlap of 4 seconds, resulting in one prediction for every second of the recording. In our evaluation, we again limited the model's output to the same 326 species.

Researchers at the Naturalis Biodiversity Center, led by Burooj Ghani, developed a classification model targeted at European bird vocalisations: AvesEcho (Ghani et al. 2023). The model used in this study was still in its development phase (v0) and was a single-label multi-species classification model, meaning it produced a single species prediction given an input segment of three seconds.

Aquila was developed by Aquila Ecology, a small Dutch company developing technological solutions for ecological research. Their classification model performed well for bats (pers. comm A. Krediet) and also classifies 332 bird species that occur in the Netherlands. This species list excludes 28 species which we distinguish for BirdNET, GBVC and AvesEcho. Most of these are rare species which were not present in the evaluation dataset, apart from five species which were present in the evaluation set (45 annotations) but not included in the Aquila classifier: Red Junglefowl *Gallus gallus,* Mandarin Duck *Aix galericulata,* Pheasant *Phasianus colchicus,* Egyptian Goose *Alopochen aegyptiaca* and Greater Canada Goose *Branta canadensis.*

Model evaluation

We evaluated the four models using our evaluation dataset consisting of 70 soundscapes. The models predicted scores for 3 to 10-second-long segments of a soundscape while we required a single score for each species per soundscape. Therefore, we took the maximum score to obtain a single score for each species given a soundscape. To get a single score for the performance of each model on the soundscapes we used the following approach: 1) We calculated the area under the ROC curve (AUC, Hanley & McNeil 1982) for each species present in the evaluation set. The ROC curve is produced by plotting the true positive rate (sensitivity) on the y-axis against the false positive rate (false alarm rate) on the x-axis at varying threshold settings (Hoo et al. 2017). The larger the area under this curve, the better its performance. An AUC score of 0.5 is as good as random guessing while a score of 1.0 describes a perfect classifier for the data. 2) We averaged the AUC scores calculated in step 1 to obtain a single score per model. We call this metric AUC-mean as it is the mean value of the species' AUC scores. This method ignores species absent from the evaluation set as those have no true positive rate. To not ignore the cost of false positives for other species, we used a second metric we call AUC-binary. Here, we first binarized the problem by accumulating the scores for the bird species in the evaluation set into one set, and the scores for all other species into a second. This representation can be used to plot a ROC curve using all scores and obtain AUC-binary by calculating the area under this ROC curve. We also created a detection error trade-off (DET, Martin et al. 1997) plot, which is similar to a ROC curve but uses the false negative rate (miss rate) instead of the true positive rate and the cumulative normal distribution to scale the curve into a more linear representation. The AUC-binary measure is affected by calibration errors as all species share the same threshold setting. Therefore, it measures a combination of the performance for all species and how well the species are calibrated. To get a more detailed picture, we also plot the ROC curve for a few species. The code used for evaluation can be found in our GitHub repository (van Harten 2023).

Model training

Besides comparing existing models using our evaluation set, we trained the Google Bird Vocalization Classifier, the best performing classifier for our data, for Dutch usage. We replaced the default classifier of GBVC, predicting 10932 classes, with our own classifier predicting 326 species and called this model NLC. First, we obtained training data for this classifier from Xeno-Canto and pre-processed it. Subsequently, we trained the classifier, and in the end, we evaluated the model's performance. The code needed to train and evaluate the classifier can be found in our repository (van Harten 2023). We optimized our model trying to maximize the AUC-mean score for the validation split while also preferring sim-

pler approaches and approaches known to help generalization (like drop-out, label smoothing and up-sampling).
We trained a third model called NLC-NoSec. This model is similar to NLC but ignores the secondary labels of Xeno-Canto. We were interested to see how the model performs being ignorant of these background vocalisations during training.

Training data

To train a model fine-tuned on Dutch data, we used both recordings containing bird sounds for all 326 species and recordings without bird sounds. We obtained bird sound recordings from Xeno-Canto, with a maximum of 100 recordings per species. As non-event data, we used the Warblrb10k development dataset for DCASE 2018 Bird Audio Detection task 3, which consists of 8,000 short smartphone recordings from around the UK and includes weather noise, traffic noise, human speech and even human bird imitations. We used nearly 2,000 recordings that do not contain bird vocalisations.
The methods used for data preprocessing, the model architecture and the training procedure are described in the Appendix. The resulting models were again evaluated using AUC-mean, AUC-binary and a DET-plot (see above).

Results

Comparison audio vs point counts

Of the 91 species present on the audio-recordings (manual annotation) and/or in the point count dataset, 65 were recorded with both methods (though not always at the same point-visit combination), 19 were only observed during point counts and 7 only in audio. Point counts resulted in 1112 records, while audio recordings resulted in 857 records (species-point-visit combinations). Of these, 637 records overlapped, being present in both point count and audio data, while 475 records occurred exclusively in the first, and 220 exclusively in the latter.
Excluding species with less than 10 records resulted in 44 species, of which 18 overlapped in at least 50% of the records in which they were observed, with Common Chaffinch (95%), Blackcap *Sylvia atricapilla* and Common Chiffchaff *Phylloscopus collybita* (both 81%) showing most overlap (Fig. 1); Greater Canada Goose (0%), Grey Heron *Ardea cinerea* (0%) and Eurasian Jay *Garrulus glandarius* (7%) showed least overlap, though all three were observed in both point counts and audio-recordings, if not simultaneously. Sixteen species were most often recorded in point counts only, while only 4 species were most often recorded in audio only (Greater Canada Goose, Oystercatcher *Haematopus ostralegus*, Greylag Goose *Anser anser* and Jackdaw *Corvus monedula*).

Table 2. Mean performances (AUC-mean and AUC-binary) of the four evaluated classifiers BirdNET, GBVC, AvesdEcho and Aquila.

Model	AUC-mean	AUC-binary
BirdNET	0.834	0.911
GBVC	0.836	0.919
AvesEcho	0.832	0.825
Aquila	0.741	0.774

Evaluating classifiers

Comparing existing classifiers

The mean value of all present species' AUC-mean scores was highest for the Google Bird Vocalization Classifier (Table 2), however, the values for both BirdNET and AvesEcho were close. The performance of the Aquila model stayed behind. When looking at the AUC-binary, the performance gap between AvesEcho and both GBVC and BirdNET increased. Looking at the DET plot (Fig. 2), we again see BirdNET and GBVC close together. AvesEcho performs worse at lower thresholds compared to the other models and the performance of Aquila lags.
Figure 4 shows the AUC-mean scores (for examples see Fig. 3) for the 37 most frequent species. AvesEcho seemed unable to classify the Common Starling *Sturnus vulgaris* for the evaluation set, and also seemed to have problems identifying Egyptian Goose, Song Thrush *Turdus philomelos*, Eurasian Coot *Fulica atra* and Common Blackbird *Turdus merula* (AUC < 0.7). The two latter species were also challenging for GBVC and BirdNET, which however performed better for Egyptian Goose, Song Thrush and Eurasian Coot. Aquila could not classify Common Pheasant, Egyptian Goose and Greater Canada Goose, as these species were not included in their species list. However, it performed even worse than random for Eurasian Coot and Common Blackbird, and had problems identifying Willow Warbler *Phylloscopus trochilus*, Greylag Goose, Common Linnet *Linaria cannabina* and European Robin *Erithacus*

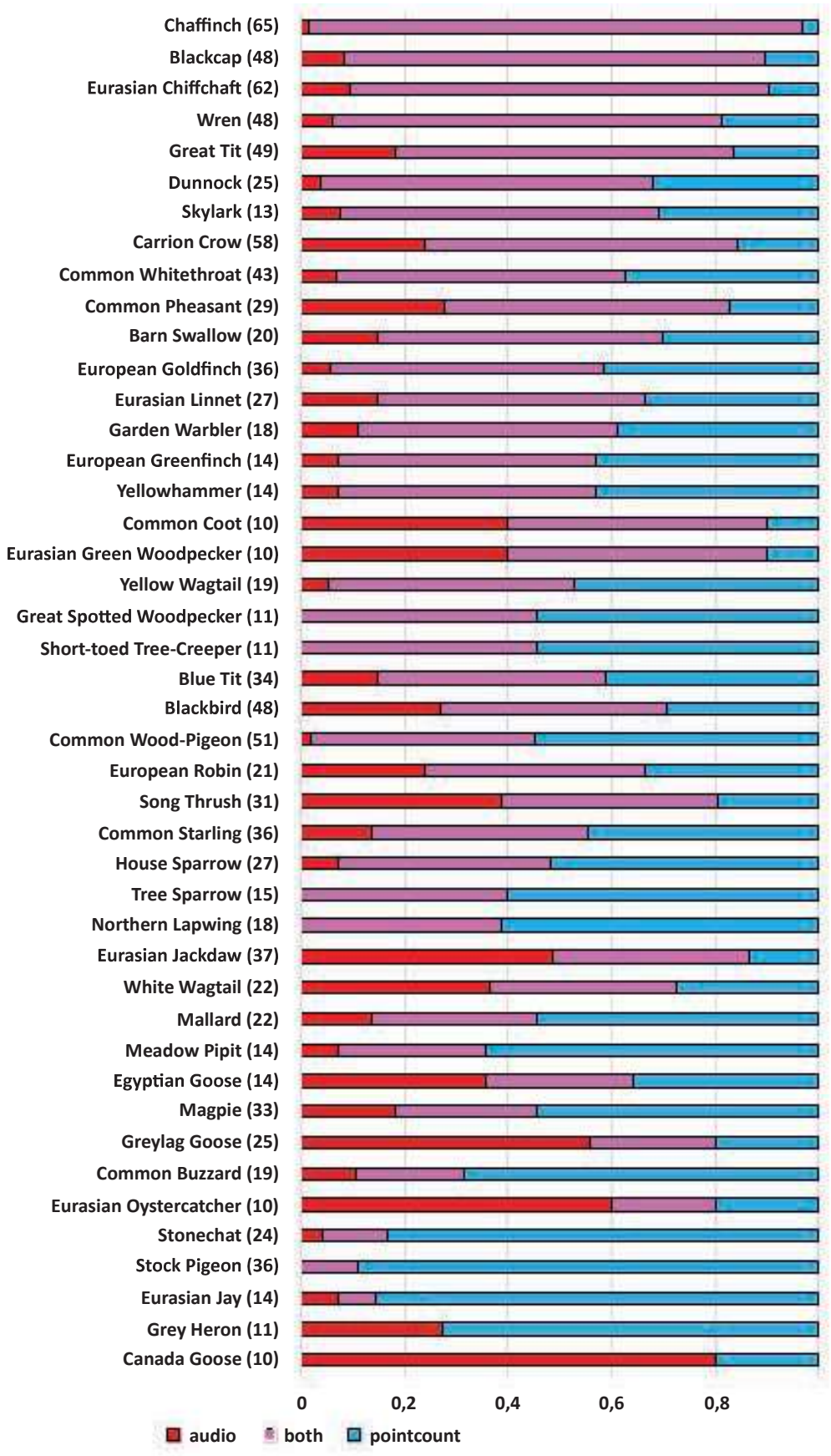

Fig. 1. Proportions of species-point-visit combinations in which a species was observed from the audio-recording only (red), during the point count only (blue) or during both (purple) for 44 species with 10 or more records (in audio-recordings and point counts combined). Number of records given between brackets. The species are ordered by the degree of overlap.

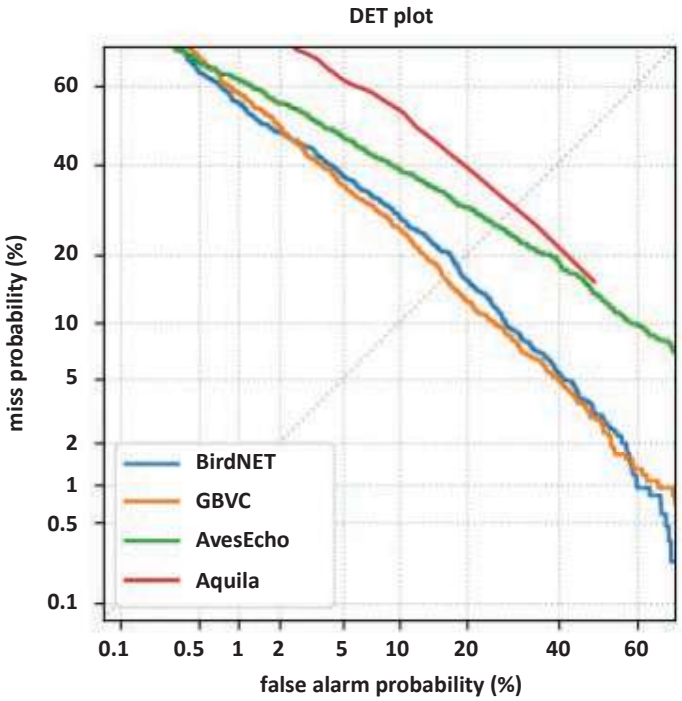

Fig. 2. Detection Error Trade-off curve for BirdNET, GBVC, AvesEcho and Aquila, accumulating targets (bird vocalisations) and non-targets (other sounds).

rubecula, species for which the other models performed relatively well.

Training and evaluating new classifiers

Our approach using a custom classifier for Dutch application (NLC) improves results on both AUC metrics (Table 3). NLC-NoSec scores close to GBVC for AUC-mean but worse for AUC-binary. Looking at the AUC scores for the 37 most frequent species individually (Fig. 6), results vary for different species. Outliers seem to be the European Green Woodpecker, Greater Canada Goose, Song Thrush and Common Blackbird. For the first species, both NLC and NLC-NoSec perform better than GBVC while for the other species NLC performs best but NLC-NoSec worst. In general, NLC performs better than GBVC over the entire range (Table 3) while NLC-NoSec performs competitively for high thresholds but falls behind when lowering the threshold (Fig. 5).

Table 3. Mean performances (AUC-mean and AUC-binary) of the two new classifiers NLC and NLC-NoSec and GBVC.

Model	AUC-mean	AUC-binary
GBVC	0.836	0.919
NLC	0.867	0.944
NLC-NoSec	0.838	0.834

Discussion

Audio recordings vs point counts

An important difference between point count data and data collected using (simple) audio recorders is that the first contains information on presence, abundance and density, while the latter mainly on presence, though analyses techniques are available which enable estimation of abundance from audio-data, using Time To Detection (Strebel et al. 2020), or removal models (van Wilgenburg et al. 2017). However, for these analysis techniques to work properly, the audio-recorders used need to be calibrated for the species present and location/circumstances, to be able to determine detection probabilities.

In 48% of species-point-visit combinations, species were both observed during the point count and heard on the audio-recording. Species showing good overlap between the two methods were all locally occurring songbirds of closed habitats, such as Common Chaffinch, Blackcap, Eurasian Wren *Troglodytes troglodytes* and Great Tit *Parus major*. These species are usually observed by their vocalisations when performing point counts.

Most species were recorded more often during the point counts than on the audio-recordings. This can largely be attributed to visual observations of birds that did not vocalise during the 10-minute count. These were often large and/or conspicuous and relatively silent species of open habitats, such as Stock Dove *Columba oenas*, Eurasian Jay, Stonechat *Saxicola rubicola*, Grey Heron, Common Buzzard *Buteo buteo* or Lapwing *Vanellus vanellus*.

Some species (e.g. Greater Canada and Greylag Goose, Eurasian Oystercatcher and Jackdaw) were observed more often from the audio-recordings than during point counts. These were often species with loud vocalisations and a large action radius and were probably individuals occurring outside the 300 m radius or flying over, and therefore deliberately omitted during the point count. However, it is likely that some individuals/species were truly missed during the point count, e.g. while the observer was focussing on other species; an advantage of using audio-recordings

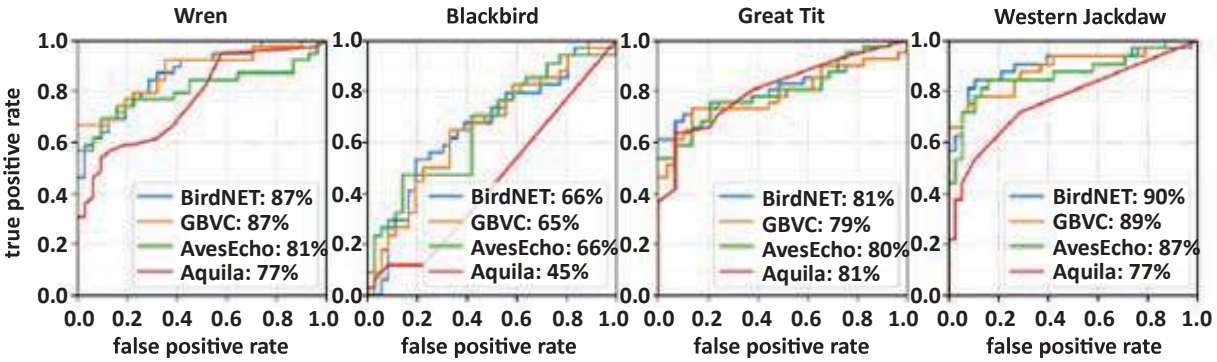

Fig. 3. Plots showing the ROC curves and AUC scores for BirdNET, GBVC, AvesEcho and Aquila for four common species. The area under the ROC curve (AUC) for each species, shown in Fig. 4, was calculated using these curves.

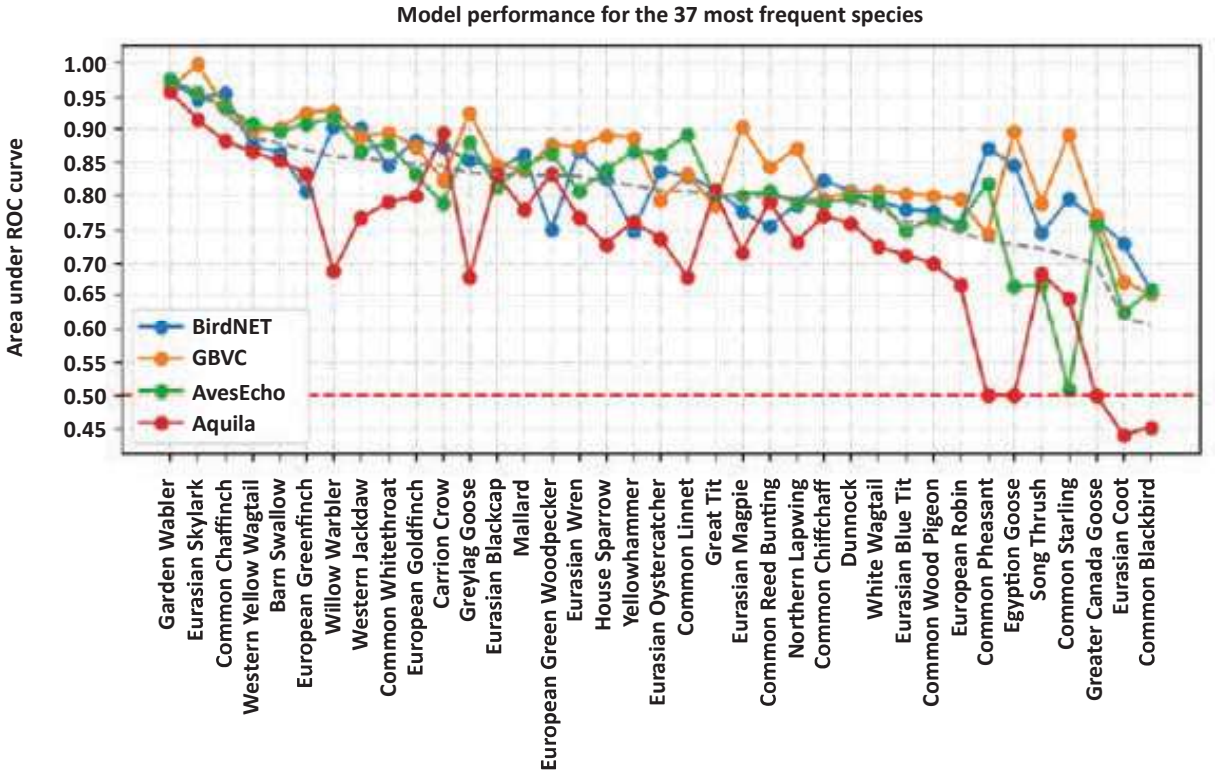

Fig. 4. Comparison of AUC scores for BirdNET, GBVC, AvesEcho and Aquila for 37 species most frequent in the evaluation set.

is that one can replay the recording when many species are vocalising simultaneously, to be able to identify and register every species.

Using acoustic observations only, (at least for now) it is not possible to distinguish between individuals inside or outside a specific detection radius, nor between individuals tied to a location or flying over. It is important to keep in mind such methodological differences, which come on top of the discrepancies in the mode of observation (sight and audio vs audio-only).

Differences in observation probabilities due to the lack of visual cues can be partly overcome by 1) applying ARU's in habitats/situations where visual detection is generally less used (forest, nighttime) and 2) increasing recording effort (duration). For the latter to be time and cost-effective, classification should be done automatically (Venier et al. 2017). However, automatic classification is associated with other issues, as discussed in the next section.

Model evaluation

The main limitations of classification model evaluation are generally the size and characteristics of the evaluation set. Only 73 out of 326 target

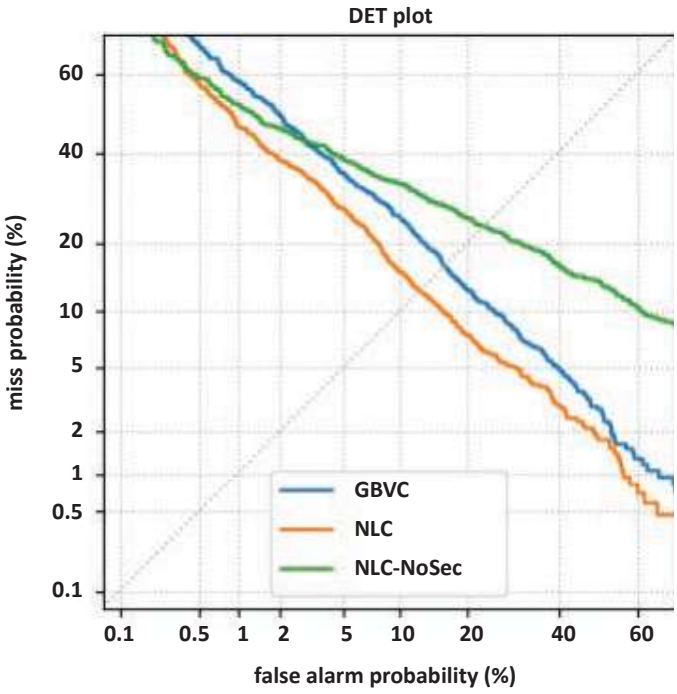

Fig. 5. Detection Error Trade-off curve for GBVC and the new classifiers NLC and NLC-NoSec, accumulating targets (bird vocalisations) and non-targets (other sounds).

species are present in the used evaluation dataset, so model performance could not be assessed for species not present in the evaluation dataset. Any conclusions on model performance are therefore limited to this set of 73, mainly agricultural and common species.

Moreover, for the species for which there were annotations, there are only twelve on average. Ideally, one would make annotations for every five seconds of the recording instead of per 10-minute recording. In this way, there would have been 120 times as many targets (bird sounds) and non-targets (other sounds) combined, which could be used to get a more accurate performance measure. Despite these limitations, the evaluation dataset provided an impression of how the different models perform.

When using a model in practice, one has to pick a confidence threshold, discarding all predictions with a score lower than the chosen threshold. A major difficulty is that the optimal threshold differs per species. So, one could best vary the threshold setting per species. Ideally, these thresholds are set using the results from a large and representative evaluation set. One could also use a manual approach in which the system first uses a very low confidence threshold for all species. Over time, the user could increase the threshold setting for a species if the system returns many false alarms for that species. A downside of this method is that it is not systematic and is mostly based on the number of false alarms.

Performance of classifiers

All evaluated models performed far from perfect on the evaluation set. When allowing for a false alarm probability of 1%, the miss probability is around 60% (Fig. 2). As there are 864 targets (vocalisations) and 21,956 non-targets in the dataset, this results in 220 false alarms while missing around 6 out of every 10 targets. Though classifiers are improving rapidly (GBVC at the point of writing has evolved to version 8 and AvesEcho to version 1), for most purposes it is pivotal to perform a manual validation of results.

In general, the performances of the classifiers, excluding Aquila, showed comparable patterns for the most frequent species, suggesting that species-specific vocalisation characteristics to a large degree determine their identifiability, regardless of the classifier used. Alternatively, if the classifiers used similar training sets for these species (e.g. all of the models evaluated use Xeno-Canto as (one of) their primary training data source), the correlation may reflect species-specific differences in the quality of the training data.

The overall performance of BirdNET and GBVC on our evaluation dataset is comparable, but differs for some species and at different thresholds. Choosing one model over the other should therefore ideally be done by evaluating the candidate models on a dataset similar to the intended application.

AvesEcho seems to be competitive at higher thresholds, but performs worse at lower thresholds, resulting in more false positives, while yielding fewer additional true positives. It could be that calibration is worse for AvesEcho than for BirdNET and GBVC, as unlike AUC-binary, the AUC-mean score for the three models is similar. Moreover, the prototype of the model we evaluated is single-label. The performance for the newer version of the model (Aves Echo v1) will probably have improved, as it is now multi-label. It would be interesting to see whether this newer version of the model can match or even out-

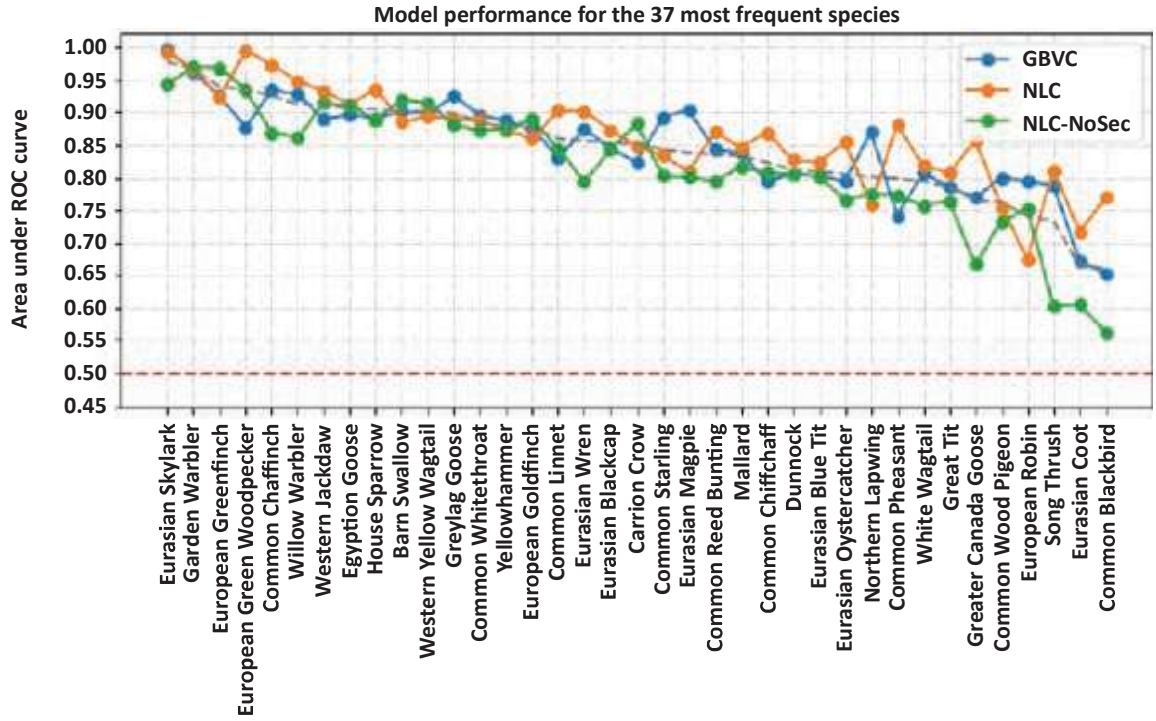

Fig. 6. Comparison of AUC scores for GBVC, and the new classifiers NLC and NLC-NoSec for 37 species most frequent in the evaluation set.

compete state-of-the-art models when applied to soundscapes.

The classifier of Aquila had the lowest performance on our evaluation dataset. The dataset on which the model was trained did not include five of the species present in our evaluation dataset. This can be seen from Figure 4, where Common Pheasant and Egyptian and Greater Canada Goose, three of those species, had an AUC = 0.5, exactly equalling performance when classified randomly. Including these species in the comparison will thus result in a lower overall performance score. Removing these species on which the model was not trained, only slightly increased the AUC value from 0.741 to 0.757. The Aquila classifier was mainly developed for insect and bat sounds, for which it performs markedly better (personal communication A. Krediet). Possibly the model is more suited for sounds of these species groups rather than for bird sounds (since the sounds of these species groups are more distinctive than bird sounds), or the classifier was trained with a smaller set of bird data than the other classifiers.

All of the models evaluated use Xeno-Canto as (one of) their primary data source to train the model on bird vocalisations. However, the recordings from Xeno-Canto have multiple characteristics that make models trained on this data likely less suitable for analysing soundscapes: weak labels (we know which bird is vocalising but not when in the recording), noisy labels (not all species vocalising in the background are labelled (Vellinga and Planqué 2015) which is particularly harmful in the context of soundscapes where we want to be able to detect background vocalisations) and focal recordings (most recordings are focal, meant to capture the labelled bird as clearly as possible). Another drawback is that not all species are equally well-represented.

The four models we evaluated predict scores for 3 (BirdNET, AvesEcho), 5 (GBVC) or 10-second-long (Aquila) segments of a soundscape. As a consequence, species with longer, more elaborate vocalisations may be harder to classify. An example would be the Blackbird, which proved difficult to classify for all of the models assessed. Another possible consequence is that one long (multi-segment) vocalisation could be attributed to multiple species. We have seen this occur with, again, Blackbird song, of which segments were misclassified as Mistle Thrush. This could be solved if classifiers would use longer segments, or if they would use information from adjacent segments. Remarkably, Aquila, the model using the longest segments (10 s) performed worst for the Blackbird, even worse than random.

Performance of new classifiers

Our results indicate the usefulness of creating a custom classifier to improve performance for a specific use case. Based on try-outs with the Google Bird Vocalization Classifier, we suspect to be able to improve in two ways:

• As we transfer to a Dutch setting, we only have to predict 3% of the species of the global setting. This matters most if the vocalisation of a target species is very similar to that of another species not in our target set. By removing the other species, the model will be more confident of the target species. It looks like an example of this can be seen in the case of the European Green Woodpecker *Picus viridis*. Whereas GBVC tries to distinguish the Iberian Green Woodpecker *Picus sharpie*, our model does not, resulting in a higher AUC score for the species, even for NLC-NoSec.

• As we switch from focal recordings to soundscapes, we need to increase our sensitivity to background vocalisations. We did not dispose of annotated multispecies recordings for training the model, but to check for the effect of background vocalisations, we decided to ignore the secondary labels present in some Xeno-Canto recordings. Ignoring secondary labels, NLC-NoSec performs worse on the evaluation set than NLC at lower thresholds. Two possible reasons for this drop are that the model is calibrated worse or that it has more difficulty with fainter/overlapping vocalisations. Given that the AUC-mean score is also lower than for NLC, it is not only a calibration issue. Therefore, we suspect that using secondary labels when training helps the model to predict fainter/overlapping vocalisations. This would also mean that an important step to improve classifiers would be to train them using annotated multispecies recordings.

For some species, the GBVC still performed better than the NLC, possibly because GBVC uses a larger training dataset than NLC, for which a maximum of only 100 recordings were selected. The performance of NLC could therefore probably be further improved by increasing the training dataset per species.

Conclusions

Passive Acoustic Monitoring at present is only/mainly useful for collecting presence data, not for trends in numbers or densities. Applicability for absence information is likely limited to vocal species. That said, it can be a useful addition for distribution mapping, including atlases, especially for nocturnal, vocalising species, and underrepresented habitats or sites. In the future, new analyses techniques, in combination with improved (and calibrated!) ARUs may enable abundance estimation for a subset of species on a local scale, though not without a considerable effort in the field set-up and validation/calibration of results. Furthermore, it is important to keep in mind that data collected using PAM differs systematically from data collected by point counts or other field methods, especially in open landscapes and for less vocal species. Therefore, it is important to always beware of a potential methodological trend break and always label data collected by PAM as such.

The quality of recordings greatly determines their usefulness for species classification. Though effects have not been analysed here, some recordings were excluded as they contained too much background noise and were deemed useless for the evaluation. Minimizing background noise is therefore a first step in improving audio monitoring results. This also highlights a potential risk of using low-cost ARU's, which might use low quality microphones with considerable within device variation, which can result in highly variable performance among devices in the field.

To improve a model's performance on soundscapes, it is pivotal to use annotated soundscapes as (additional) training data. Annotations should contain the species name and start and end time for each vocalisation. In addition, including recordings without bird sounds in the training dataset could improve the distinction between bird vocalisations and background noise. We expect models could also be greatly improved if they would also use the information from adjacent sound fragments. Furthermore, the species-specific performance of models can relatively easily be improved by varying and optimizing the threshold setting per species.

Finally, we suggest using a hybrid approach in which a model's predictions are validated by an expert. The validated data can then be used for training, which can further improve the model (human-in-the-loop).

Acknowledgements

We would like to thank Vincent Kalkman and Burooj Ghani (AvesEcho) and Anne Krediet (Aquila)

for kindly offering us their classifiers for evaluation; Fred de Boer for providing the AudioMoths and Eveline Schaar for deploying them in the field; Frank Majoor, Jesse Keijzer and Bas Hissel for performing the point counts; Willem van Manen, Joost van Bruggen and Bas Hissel for performing the manual annotations of the audio-recordings.

References

Browning, E., R. Gibb, P. Glover-kapfer & K. E. Jones. 2017. Passive acoustic monitoring in ecology and conservation. WWF Conservation Technology Series 1 (2). WWF-UK, Woking, United Kingdom.

Ghani, B., T. Denton, S. Kahl & H. Klinck. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports,* 13: 1–14.

Google Bird Vocalization Classifier https://www.kaggle.com/models/google/bird-vocalization-classifier. Accessed: 2023-11-20.

Hanley, J. A. & B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* 143: 29–36.

van Harten, N. 2023. NL-Bird-PAM. https://github.com/nielsvharten/NL- Bird- PAM/.

Hoo, Z. H., J. Candlish & D. Teare. 2017. What is an ROC curve? *Emergency Medicine Journal,* 34: 357–359.

Kahl, S., C. M. Wood, M. Eibl & H. Klinck. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics,* 61: 101236.

Klingbeil, B. T. & M. R. Willig. 2015. Bird biodiversity assessments in temperate forest: The value of point count versus acoustic monitoring protocols. *PeerJ,* 3: e973.

Martin, A., G. Doddington, T. Kamm, M. Ordowski & M. Przybocki. 1997. the Det Curve in Assessment of Detection Task Performance. Pages 1895–1898 5th European Conference on Speech Communication and Technology, EUROSPEECH 1997.

Shonfield, J. & E. M. Bayne. 2017. Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology,* 12: 14.

Strebel, N., C. J. Fiss, K. F. Kellner, J. L. Larkin, M. Kéry & J. Cohen. 2021. Estimating abundance based on time-to-detection data. *Methods in Ecology and Evolution,* 12: 909–920.

Teunissen, W. A., P. Wiersma, A. de Jong, E. Kleyheeg & J. Vergeer. 2019. Handleiding Meetnet Agrarische Soorten (MAS).

Vellinga, W. P. & R. Planqué. 2015. The Xeno-canto collection and its relation to sound recognition and classification. Page CEUR Workshop Proceedings.

Venier, L. A., M. J. Mazerolle, A. Rodgers, K. A. McIlwrick, S. Holmes & D. Thompson. 2017. Comparison of semiautomated bird song recognition with manual detection of recorded bird song samples. *Avian Conservation and Ecology,* 12: 1.

Van Wilgenburg, S. L., P. Sólymos, K. J. Kardynal & M. D. Frey. 2017. Paired sampling standardizes point count data from humans and acoustic recorders. *Avian Conservation and Ecology,* 12: 13.

Xeno-Canto https://xeno-canto. org/about/xeno-canto. Accessed: 2023-11-17.

Appendix

The methods used for data preprocessing, the model architecture and the training procedure for the new classifiers NLC and NLC-NoSec

Data preprocessing

We trained a classifier that uses GBVC embeddings of 1280 features as input. To generate these embeddings, all audio was resampled to 32 kHz as GBVC is trained using this sample rate. To increase the diversity of our training data, we used an overlap of 4 seconds, ignoring the last three embeddings, unless the recording was shorter than 4 seconds, in which case we only used the first embedding. As the input length for GBVC is 5 seconds, this resulted in 30 embeddings for a 30-second-long recording. We also used the model's predictions for all training data to filter which embeddings of a recording to use. We discarded embeddings for which the GBVC prediction (after softmax) was less than 0.02 for the species labelled in the recording. If no segments had a prediction of at least 0.02, we kept the first embedding hoping that the target species vocalises within the first five seconds of the recording.

Model architecture

Our classifier is a simple, fully connected neural network having one hidden layer of size 512. First, we applied dropout on our input. After the first fully connected layer, we used batch normalization and ReLU activation followed by dropout. After the second fully connected layer, we only used sigmoid activation. Given the 1280 input and 326 output dimensions, the classifier has 823k trainable parameters.

Training procedure

We split our training data in a train and validation split of 80% and 20% respectively. We upsampled species with fewer recordings by repeating samples using the number of embeddings as their probability. This way, recordings for which there are more embeddings are more likely to be sampled multiple times. To diversify inputs and improve the model's ability to learn overlapping vocalisations, we employed E-stitchup. This approach combines two embeddings into one by randomly selecting the value at each index of the combined embedding from the value at that index for one of the two original embeddings (Cameron & Lundgaard 2019). The probability to sample from one embedding is determined by a value lambda, for each combination drawn from a beta distribution with $\alpha = \beta = 1$. Given that Xeno-Canto allows specifying secondary labels of species also vocalising in a recording, we give these classes a non-zero probability p. After experimentation, we found $p = 0.9$ to work best. Besides, we use label softening, subtracting 0.1 from all classes with non-zero probability and adding 0.1 to all other classes.
As our optimizer, we use ADAM (Kingma & Ba 2014) with an initial learning rate of 0.001 and a cosine annealing schedule gradually decreasing the learning rate to zero over 30 epochs. We train in batches of 32 samples where each sample is constructed out of two embeddings randomly selected from two different recordings using E-stitchup. We use a dropout probability of 0.1.

Cameron R.W. and K.T. Lundgaard. 2019. "E-stitchup: Data augmentation for pretrained embeddings". In: *arXiv preprint arXiv*: 1912.00772.
Kingma P.D. and J. Ba. 2014. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv*: 1412.6980.

Received: 14th July 2025
Accepted: 12th December 2025