TITLE

Enhancing Avian Sound Recognition Models Detection Precision via Logistic Regression of Large Acoustic Datasets: A Case Study of the European Robin

AUTHORS AND AFFILIATIONS

Bethany Shackleton, Luiza Passos*, Ross Macleod

School of Biological and Environmental Sciences, Faculty of Health, Innovation, Technology and Science. Liverpool John Moores University, Liverpool, UK

Email Addresses of the Co-authors:

Bethany Shackleton	b.r.shackleton@2023.ljmu.ac.uk
Ross Macleod	r.c.macleod@ljmu.ac.uk

*Corresponding author:

Luiza Passos	l.figueiredopassos@ljmu.ac.uk

SUMMARY

The goal of this protocol is to determine species- and site-specific confidence score thresholds using logistic regression to improve detection precision in large acoustic datasets processed with automated acoustic recognition software.

ABSTRACT

Passive acoustic monitoring (PAM) has become an invaluable tool for biodiversity research, enabling the non-invasive collection of vast datasets. However, a significant challenge remains in efficiently and reliably processing this large volume of data to extract species-specific information across varying locations. This paper presents a detailed, step-by-step protocol to address this challenge using a machine learning detector module within a bioacoustics analysis software. The methodology is designed to accurately and confidently identify and validate bird vocalisations from raw acoustic recordings.

Our protocol details the process from initial data collection using autonomous recording units (ARUs) to the final generation of a high-quality annotated dataset. Key steps include configuring the machine learning detector module to generate initial detections, a manual validation procedure to calculate precision tables, and a logistic regression analysis to determine a species-specific and, where appropriate, a location-specific confidence score threshold. This statistically derived threshold is then used to refine the detector's output, tested on two overlap configurations (0 s and 2 s). We show that applying the derived optimal confidence score thresholds substantially improves the machine learning-based avian sound recognition models detection precision across sites. For the three sites used to illustrate the process (Liverpool Park, Cairngorms, and Glasgow Suburban) precision increased by 26.1%, 17.7%, and 17% for an overlap of 0 s, and by 28.77%, 16.87%, and 15% for an overlap of 2 s. We suggest the resulting methodology is superior to manual counting methods in both speed and reliability. In summary, this paper provides a reproducible framework that facilitates the accessible and effective use of machine learning approaches in bioacoustics, enabling researchers to confidently leverage large acoustic datasets for ecological studies and parameter analysis.

INTRODUCTION

Animal vocalisations offer valuable insights into the natural world and act as powerful biological tools allowing researchers to investigate various aspects of animals lives[1]. The field of bioacoustics has become increasingly important in biodiversity monitoring and conservation, aiding in species identification and in assessing population trends and ecosystem health[2,3,4]. This field of research also contributes to our understanding of animal behaviour, for example, in communication and habitat use[5,6]. More recently, it has been demonstrated as a promising approach for assessing animal welfare[7,8].

One of the key advantages of using acoustic analysis as a research tool is the recent technological advances in autonomous recording units (ARUs), which enable researchers to non-invasively and remotely monitor animal vocalisations for extended periods of time, a technique known as passive acoustic monitoring (PAM)[9]. This has allowed for the collection of data from cryptic species without human intervention[10,11]. Additionally, ARUs are cost-effective and minimise the need for trained researchers to be in the field[9]. A study on the cryptic European nightjar (*Caprimulgus europaeus*) demonstrated these advantages, finding that the use of ARUs was significantly more effective than human observers in detecting the species[12]. Large-scale bioacoustics studies are increasingly becoming more popular because they can provide insights into a range of ecological and behavioural aspects, they are also highly applicable and well-suited for field studies. Yet ARUs running for long periods mean researchers are inevitably faced with the challenge of efficiently and reliably processing vast amounts of acoustic data[13].

More recently, researchers have begun combining PAM with the use of automatic recognition tools to detect species within large datasets, enhancing and speeding up the overall processing of data. These techniques automatically recognise and classify species based on vocalisation type[13], often using machine learning and deep learning models such as artificial neural networks, which have gained traction in recent years[14]. The reliability of automatic recognition software is often superior to manual methods of species detection, as demonstrated in studies on Giant Panda *(Ailuropoda melanoleuca)* vocalisations[15] and Malaysian frog species[16]. Importantly, one of the primary drivers behind adoption of these approaches is processing speed, particularly as the volume of acoustic data generated through PAM continues to grow[13]. Manual analysis of large acoustic datasets is inefficient and impractical at scale, yet automated recognition tools can process recordings in a fraction of the time required by a human observer. For example, one study reported that 257 h of manual visual scanning was equivalent to 1 h of processing time using BirdNET (machine learning–based avian sound recognition model)[17].

This machine learning–based avian sound recognition model is one such tool that has recently gained traction for its use in ornithology and other acoustics studies, it is an accessible automated software for bird species identification that relies on a deep learning model[18,19]. Most studies to date have used the machine learning–based avian sound recognition model to generate species lists for a given area, and with human validation it has been shown to detect 89% of species in a community, while requiring less than a quarter of the time used by visual scanning alone[17]. This machine learning–based avian sound recognition model has also been shown to effectively identify the presence of a specific target species in an area, such as the Eurasian Bittern (*Botaurus stellaris*), with a 93.7% success rate[20]. However, this model often has a higher success rate when combined with manual validation and isn't yet reliable as a fully automatic detection tool[21]. There are also other challenges to consider according to Wood and Kahl[21] regarding its production of unitless predictions.

The machine learning detector module assigns each detected bird vocalisation a quantitative confidence score between 0 and 1, which indicates how certain the software is that the detected sound has been correctly identified as a particular species[18,19,21]. Precision, defined as the proportion of detections that are true positives (i.e. correct identifications), is a key metric influenced by this score[22]. Consequently, a higher confidence score may produce less detections, but with higher precision, yet miss detections especially in noisy environments, and a lower score may produce more detections with less certain precision[22,23]. The relationship between the confidence scores and precision is complex, it is often found that some high confidence scores can be false positives, for example in locations with high levels of background noise, or with species that are likely to mimic other species vocalisations[23,24]. Due to the variability in precision among species and location, it is possible during the configuration process to assign individual confidence score thresholds to each bird species, whereby only detections with scores above that threshold will be labelled. A high universal confidence score threshold could be applied across all species and locations, however, doing so often increases the number of false negatives (missed identifications)[22,25,26], while the number of false positives vary between species and with environmental differences between recordings.

Several studies and best-practice guidelines outline various methods for determining an optimal confidence score threshold at which to run the machine learning detector module for specific species and study locations, enabling detections to be accepted with greater confidence as true positives[21,24,26–28]. One approach involves calculating F-scores across a range of confidence thresholds and selecting the threshold that maximises the F-score, balancing both precision and recall[29,30]. An alternative approach, which allows for probability calibration, involves manually validating a random subset of predictions and fitting a logistic regression to establish a relationship between the binary validation outcome (true/false detection) and the associated confidence score[21]. This method can then be used to determine context-specific thresholds, which may be adjusted across species and environmental conditions (e.g., varying background noise)[19]. Such methods have been applied in passive acoustic monitoring studies on Yucatán black howler monkeys (*Alouatta pigra*), gray wolves (*Canis lupus*) and coyotes (*C. latrans*)[27,28].

The aim of this methods paper is to demonstrate and validate a scalable methodology for accurately and reliably extracting species-specific vocal data from large acoustic datasets using the machine learning–based avian sound recognition model within Raven Pro-bioacoustics analysis software (ver 1.6)[19,31]. This approach incorporates the calculation of appropriate confidence score thresholds following Wood and Kahl's[21] best-practice guidelines. While these guidelines[21] provide a comprehensive framework for applying confidence scores within the machine learning–based avian sound recognition model[19], they are primarily presented as best-practice recommendations. Few studies have operationalised these procedures across distributed ARU study systems under realistic monitoring constraints, where acoustic conditions and background noise vary spatially[22]. This protocol is particularly suitable for large-scale passive acoustic monitoring studies involving multiple sites, variable acoustic environments, or target species with high misclassification rates. Furthermore, the implementation of these guidelines within the bioacoustics analysis software (ver 1.6)[19,31], a more accessible and user-friendly platform compared to coding-based environments[32,33] has not been previously demonstrated.

To address this gap, we apply and validate the protocol using large-scale datasets recorded with AudioMoths - ARUs (ver 1.2.0)[34,35], across a diverse range of habitats. Our focal species the European Robin (*Erithacus rubecula*), exhibits high song variability and structural complexity[36,37], and demonstrated a susceptibility to misidentification in initial pilot analyses. By applying species- and site-specific confidence score thresholds to the machine learning–based avian sound recognition model configuration within the bioacoustics analysis software[19,31], we evaluate the protocol under varying acoustic conditions, incorporating differing background noise levels and species assemblages, and demonstrate that tailoring threshold optimisation can substantially improve detection precision in comparison to default settings. This approach is designed to ultimately facilitate accessible, streamlined, and efficient application of the machine learning–based avian sound recognition model[19] for in large scale PAM studies, with minimal coding.

PROTOCOL

Ethics statement

This protocol follows the ethical guidelines of Liverpool John Moores University. No animals were handled, disturbed, or experimentally interfered with during data collection. The study involved passive acoustic monitoring only and did not include the handling, disturbance, or experimental manipulation of animals. No hazardous procedures were required to deploy the ARU units. Autonomous recording units were deployed in accordance with institutional research and fieldwork guidelines, and appropriate signage was installed at recording sites to inform the public of acoustic monitoring. However, check weather forecasts before conducting field activity and to conduct a risk assessment for each site

Software, licensing and usage

NOTE: Process ARU units data using the machine learning–based avian sound recognition model within the bioacoustics analysis software (v1.6)[19,31] and same software (v1.6)[31] for detector output processing. This is a licensed software and is not freely available, use of the workflow described here therefore requires access via an institutional or individual license. In addition, the machine learning–based avian sound recognition model within the software (v1.6)[19,31], does not currently run on Apple devices with ARM-based processor architecture (M1 or M2 chips), however this protocol is compatible with macOS-based computers with Intel CPUs and Windows operating systems.

1. Collecting audio data

1.1. Prepare ARUs for deployment.

1.1.1. Insert batteries into the ARUs (rechargeable alkaline batteries were used during this experiment).

1.1.2. Insert appropriately sized removable memory cards relative to the amount of data to be collected (2 GB removable memory cards were used during this experiment).

Note: For recording schedules producing 108, 1 min files per day over a 30 day period (see step 1.2.5.2), each file was approximately 5760 kB, resulting in a total data volume of 622 MB. With an estimated daily energy consumption of 26 mAh. Therefore, removable memory cards with a capacity of 2 GB were used for the representative results to provide sufficient storage. Additionally, rechargeable alkaline batteries retained usable power for the longest period, as opposed to lithium AA batteries and rechargeable NiMH, under the test conditions.

1.2. Configure each ARU for deployment

1.2.1. Download the 'ARU device firmware update application'[39] and the 'ARU device configuration application'[40] from the '**App Download Page'** on the ARU website[41].

1.2.2. Connect the ARUs individually to a computer using a data transfer cable.

1.2.3. Ensure the device is switched to the **USB/Off** mode.

1.2.4. Use the ARU device firmware update application'[39] to update the devices to the latest firmware.

1.2.5. Use the 'ARU device configuration application'[40] to set the ARUs to record at a target time of day for a selected number of days. Determine active recording periods and sleep intervals.

1.2.5.1. Identify the time windows when target species are most active (e.g., dawn chorus for diurnal birds), and schedule recording periods accordingly.

1.2.5.2. Set an appropriate duty cycle, specifying how long the ARUs are actively recording versus in sleep mode.

Example: For these representative results, ARUs were programmed to record daily between 4:00–9:00 AM and 5:00–9:00 PM from 1st June-30th June 2023. These time windows were selected to capture both the dawn and dusk choruses. During active periods, devices recorded 1 min WAV files followed by a 4 min sleep interval, resulting in a total of 108 1 min recordings per day.

Note: ARUs can record continually, but this rapidly increases battery consumption and removable memory card usage.

1.2.5.3. Use the 'ARU device configuration application'[42] to set the devices to record at a sampling rate of 48 kHz as recommended for capturing a full range of bird vocalisations[34,35] and with medium gain (typically left at default unless recording in noisy environments[34,35]). Other sampling rates can be used depending on the frequency of target species vocalisations, for example for ultrasonic wildlife, such as bats or amphibians, use a sampling rate of 384 kHz[35].

1.2.6. After configuring all ARUs, switch each device to the **CUSTOM** mode.

1.3. Device protection and labelling

1.3.1. Insert each ARU device into an Antistatic Ziplock bag. Use additional waterproof tape as needed to prevent water damage. Place Silica Gel packets inside the bags to absorb moisture.

1.3.2. Label each ARU by number and its location.

1.4. ARU deployment

1.4.1. Determine locations for ARUs recorders to be deployed within the target species habitat.

1.4.2. Deploy ARUs at spacing distances appropriate to the study requirements and target species.

Note: Detection distance depends on habitat, background noise, and species' call volume. While loud species (e.g., *Eurasian bittern*) can be detected by ARUs up to 800 m, recordings degrade at longer ranges, limiting fine scale measurements[20]. Here, one ARU per site was analysed, with units about 100 m apart to ensure good spatial coverage and clear vocalisations. Adjust spacing according to species, habitat, and project goals.

2. Using the machine learning detector module in the bioacoustics analysis software[19,31] for automatic species detection

   2.1. Transfer audio data

      2.1.1. After the deployment period has ended, remove removable memory cards from the ARUs, keeping note of which removable memory cards correspond to each numbered ARU.

      2.1.2. Transfer all WAV files from each removable memory cards onto a computer using a card reader.

      2.1.3. Organise all WAV files generated from each ARU into folders labelled with the corresponding ARU number and location.

      2.1.4. Save back up versions of all data onto additional devices. Pause point: Protocol can be paused here.

      2.1.5. Restrict access to this data to authorised project researchers and ensure data is password protected to maintain confidentiality, particularly where human activity may have been inadvertently recorded.

   2.2. Configure the machine learning detector module.

      2.2.1. Open the **bioacoustics analysis software**[31] and under the drop-down menu select **Tools** > **Detector** > **Learning Detector.**

      2.2.2. In the **Choose Detector Inputs** window, select **Browse**. In the subsequent **Open Sound Files** window, select all the WAV files within the relevant ARU folder and click **Open** (**Supplementary Files S1 and S2**).

      2.2.3. In the Configure New Sound Window that opens, under **Paging > Page Sound**, enter the appropriate Page Size for the length of each recording (e.g., 60 s for 1 min WAV files) (**Supplementary File S3**).

      2.2.4. In the same **Configure New Sound Window,** under **Multiple Files** select **Open** as file sequence in one window, then click **OK** (**Supplementary File S3**).

      2.2.5. In the returning Choose **Detector Inputs window**, select **Waveform** under **Available Signals** and **Views** on the right and click the **<< arrows** button to move it over to Required Inputs. Confirm that the status under the **Required Inputs box** reads Ready!, and click **OK** (**Supplementary File S4**).

2.2.6. In the **Configure Machine Learning Detector window** that appears, under the Inputs tab, select the desired model from **Select Model**. When analysing bird species use the global avian classification model, which among various updates include over 3000–6000 bird species[19] (**Supplementary File S5**).

2.2.7. Under the Inputs tab, leave the Overlap at the default of 0 s to prevent any interference with the subsequent confidence score threshold analysis (**Supplementary File S5**).

2.2.8. Under the **Outputs tab**, select the drop-down menu **Output Class File** and select the output list that relates to the geographical location of the study (**Supplementary File S6**).

2.2.9. Under the **Outputs tab** there will be a list of bird species relevant to the chosen **Output Class File.** Lower the **Threshold** to 0.1 and select **Apply All,** make sure that all values under **Threshold** are now at 0.1 in the list (**Supplementary File S6**).

2.2.10. Select the **Suppress All Species** option and ticks will appear in all boxes in the species list. Then locate the target species and un-tick it in the **Suppress** column (**Supplementary File S6**).

2.2.11. Select **OK** to begin the Learning Detector, this will process the files and create selections where it has detected the target species vocalisations.

2.2.12. Monitor the **Progress Manager** to track **Percentage Completed** and **Time Remaining** (**Supplementary File S7**).

2.3. Save the **Learning Detector** output

Note: A **Learning Detector** table will be generated during processing and displayed beneath the spectrogram view. The numbered selections in this table correspond to the detector's identified selections on the spectrogram view. Each selection is fixed at 3 s in length and may not always be repeated over the entire vocalisation being detected. Each selection is assigned an identification **Label** and **Score** in the selection table (**Supplementary File S8**). These scores are a unitless prediction and will be used in the logistic regression analysis in subsequent stages.

2.3.1. Select **File** > **Save Selection Table "Learning Detector" As** and save the file using the ARU name and an identifier such as: AM1_LearningDetectorTable_Original.txt.

3. Create precision tables for target species at a representative ARU from each study location

3.1. Prepare the validation dataset

3.1.1. Open the **Learning Detector** selection table that was saved in step 2.3.1 and transfer to a spreadsheet software.

3.1.2. Confirm that only the target species appears under the **Label** column.

3.1.3. Delete all columns apart from the **Selection** and **Score** column**.**

3.1.4. Randomise the rows using the spreadsheet's randomisation function (e.g., add a column using =RAND(), sort this column, then delete the randomisation column).

3.1.5. Retain the first 300 randomised detections, if fewer than 300 exist, retain all detections.

Note: Selecting 300 detections for validation is standard in similar studies using 1 month of acoustic data[24]. This sample size estimates precision and mislabelling rates reliably, captures a range of detection scores, and keeps workload manageable. More selections may be needed for longer datasets.

3.1.6. Save the validation dataset using the ARU name being analysed and an identifier such as: AM1_ValidationDataset.xlsx.

3.2. Prepare bioacoustics analysis software for manual validation

3.2.1. Open the software[31], select **File** > **Open Sound Files**… and in the **Open Sound Files window** select all the corresponding **WAV files** for the ARU being analysed, and click **Open**.

3.2.2. In the **Configure New Sound Window** which opens, under **Paging**, select **Page Sound** and enter **the Page size** used in step 2.2.3. Under **Multiple Files** select **Open** as file sequence in one window and select **OK.**

3.2.3. Navigate to **File > Open Selection Table…** and open the **Learning Detection Table** for the ARU being analysed saved in step 2.3.1.

3.2.4. In the Panel on the left of the Sound view, under the **Layouts tab**, in Views:, unselect the **Waveform** view so all files are displayed in **Spectrogram view** to enable visual confirmation of vocalisation structure.

3.3. Validate randomised selections

3.3.1. Alongside the bioacoustics analysis software[31] with the prepared data ready for manual validation, open the validation spreadsheet saved in step 3.1.6.

3.3.2. Create a new column in the spreadsheet labelled **Correctly Detected**.

3.3.3. Locate each selection number from the randomised list within the spectrogram view and input whether it has been correctly detected (1) or incorrectly detected (0) into the **Correctly Detected** column.

3.4. Calculate detector precision

3.4.1. Calculate precision of the detector (prior to the logistic regression) for the ARU and target species using the formula:

3.4.2. Precision (%) = (Number of correct detections/Total validated detections) x 100

3.4.3. For datasets with precision of 90% or greater, retain current detector settings.

3.4.4. For datasets with precision lower than 90%, carry onto section 4 to fit a logistic regression analysis to determine the optimal threshold to set the detector to run at.

Note: When precision falls below 90%, high mislabelling rates make reliable extraction of vocalisations difficult without threshold adjustment. A 90% cut-off balances detection reliability with adequate sample size. Previous machine learning–based avian sound recognition model calibration studies also regard precision rates of 90% and above as being sufficiently reliable[42].

4. Perform a logistic regression analysis on the validation tables to generate an optimal confidence score threshold

4.1. Generate confidence score thresholds, utilising Wood & Kahl's[21] guidelines

4.1.1. Use the software R[32] to import the validation dataset created in section 3.

4.1.2. Run a logistic regression using the generalized linear modelling function with the family as binomial on the dataset.

4.1.3. Set the desired probability of precision as 0.9, this means that the probability that a detection is a true positive will be 90% or above.

Note: Setting the desired probability of precision to 0.9 ensures that most detections are true positives, however, even at this level false positives can occur in large acoustic datasets and users should be aware of this limitation. A 90% precision threshold allows users to maintain high numbers of correct detections, while retaining enough data for a robust analysis, aiming for 100% would limit usable data.

4.1.4. Calculate the optimal threshold by solving the fitted logistic regression equation for the predictor value (confidence score) that yields a predicted probability of 0.9 using the Logistic Regression Analysis Script below:

$$\text{Threshold} = \frac{\log\left(\frac{p}{1-p}\right) - \beta_0}{\beta_1}$$

p = 0.9, and $\beta0$ and $\beta1$ are the intercept and slope from the fitted model

Logistic Regression Analysis Script for statistical computing software

```
# Load validation table
ThresholdRobin <-read.csv('ValidationTableRobin.csv')
# Run Logistic Regression
Model1<-glm(Correctly.Detected~Score, data=ThresholdRobin, family='binomial')
Model1
# Desired probability above 90% precision
p <- 0.9 # the desired p (probability of true positive)
threshold <- (log(p/(1-p))- Model1$coefficients[1]) / Model1$coefficients [2]
```

Note: Instead of using the Logit Score as in previous studies[24,27] as the independent variable, use the raw confidence scores as the independent variable. In Wood and Kahl's[21] guidelines it is mentioned that either the raw confidence scores or back-transformed versions can be used.

4.2. Apply the optimal confidence score

4.2.1. Open the **bioacoustics analysis software**[31].

4.2.2. Re-run the **Learning Detector** on the ARU being investigated using the steps in section 2. However, this time in the configuration process, the **Threshold** for the target species being detected must be changed to the threshold determined in section 4.1 of the protocol.

4.2.3. **(Optional)** Increase **Overlap** to 2 s to increase detection resolution.

Note: The Overlap value determines how much adjacent detected segments overlap. At 0 s vocalisations on segment borders might be missed (false negatives). Increasing overlap improves detection resolution and predictive power but slows analysis[19,29]. Here, for the logistic regression, a default overlap of 0 s was used. With the representative results and confidence score thresholds calculated in the logistic regression analysis, an overlap of both 0 and 2 s is tested.

4.2.4. A new list of detections will be created that now should be more accurate, with less false positives (see Results).

4.2.5. Save the new **Learning Detector output** with the list of detections that were generated using the derived confidence score threshold by selecting **File** > **Save Selection Table "Learning Detector" As** and save the file using the ARU name and an identifier such as: AM1_LearningDetectorTable_OptimisedThreshold.txt.

4.2.6. This file represents the final validated dataset generated by the protocol. Use the saved optimised learning detector selection table as the final output of this protocol. Please note that subsequent analyses, such as annotation of vocalisations and acoustic parameter extraction, fall outside the scope of this method.

RESULTS

To illustrate the effectiveness of the described protocol, we present representative results from the analysis of acoustic data for the European Robin (*Erithacus rubecula*) collected at three sites of varying environments and avian species assemblages, including a park in Liverpool (ARU 1), a naturally regenerating woodland site within The Cairngorms National Park (ARU 2), and a suburban site in Glasgow (ARU 3). These examples demonstrate how applying statistically derived confidence score thresholds can improve detection precision across varying acoustic environments. In this study, a successful outcome was defined as a detection precision ≥ 90% following threshold optimisation, values below this threshold were considered suboptimal and indicative of continued misidentification of species.

Liverpool park - ARU 1

The machine learning detector module[19] was initially run on the full 1-month dataset recorded at the Liverpool Park site, with a default confidence score threshold of 0.1 and an overlap of 0 s, which resulted in 299 detections. All of these were manually validated as per section 3 of the protocol, yielding an initial detection precision of 66.56%, below the desired 90% benchmark, indicating the need for a logistic regression analysis to determine a more accurate threshold.

Using the statistical computing software script provided in the protocol (step 4.1.4), a logistic regression model was fitted to the validation data to determine the relationship between the detector's confidence score and the probability of a correct detection. The model's coefficients were used to calculate a new, optimal confidence score threshold of 0.54 to achieve a precision rate of 90% or higher (**Figure 1**). Logistic regression analysis indicated that the machine learning detector modules optimised confidence scores significantly predicted correct classification at the Liverpool Park site (β = 5.73, SE = 1.04, z = 5.53, $p < 0.001$). The model showed improved fit relative to the null model (residual deviance = 329.42 vs. null deviance = 381.10; AIC = 333.42; n = 299).

[Place **Figure 1** here].

The machine learning detector module[19] was re-run with this new threshold of 0.54 on the same full 1-month dataset recorded at the Liverpool Park site, following step 4.2 of the protocol, with and without an additional increase in overlap (0 s and 2 s). A new decreased set of 109 detections was generated for the default overlap of 0 s, whereas 354 detections were generated with the increased overlap of 2 s. A subsequent validation of these refined datasets confirmed a substantial improvement in precision. The post-regression precision with the default overlap of 0 s increased to 92.66% and with the increased overlap of 2 s the precision increased to 95.33%, both above the desired 90% threshold.

Cairngorms naturally regenerating woodland - ARU 2

The machine learning detector module[19] was initially run on the full 1-month dataset recorded at the Cairngorms Naturally Regenerating Woodland site, with a default confidence score threshold of 0.1 and overlap of 0 s, which resulted in 113 detections. All of these were manually validated as per section 3 of the protocol, yielding an initial detection precision of 82.3%, below the desired 90% benchmark, indicating the need for a logistic regression analysis to determine a more accurate threshold.

Using the statistical computing software script provided in the protocol (step 4.1.4), a logistic regression model was fitted to the validation data to determine the relationship between the detector's confidence score and the probability of a correct detection. The model's coefficients were used to calculate a new, optimal confidence score threshold of 0.23 to achieve a precision rate of 90% or higher (**Figure 2**). Logistic regression analysis indicated that the machine learning detector modules optimised confidence score significantly predicted correct detection at the Cairngorms Naturally Regenerating Woodland site (β = 14.02, SE = 5.78, z = 2.43, p = 0.015). The model showed improved fit relative to the null model (residual deviance = 88.04 vs. null deviance = 105.50; AIC = 92.04; n = 114).

[Place **Figure 2** here].

The machine learning detector module[19] was re-run on the full 1-month dataset recorded at the Cairngorms Naturally Regenerating Woodland site, with the new threshold of 0.23, following step 4.2 of the protocol, with and without an additional increase in overlap (0 s and 2 s). A new decreased set of 34 detections was generated for the default overlap of 0 s, whereas 121 detections were generated with the increased overlap of 2 s. A subsequent validation of these refined datasets confirmed a substantial improvement in precision. The post-regression precision with the default overlap of 0 s increased to 100% and with the increased overlap of 2 s the precision increased to 99.17%, both well above the desired 90% threshold.

Glasgow suburban - ARU 3

The machine learning detector module[19] was initially run on the full 1-month dataset recorded at the Glasgow Suburban site, with a default confidence score threshold of 0.1 and overlap of 0 s, which resulted in 1240 detections. A randomised sample of 300 of these detections was manually validated as per section 3 of the protocol, yielding an initial detection precision of 72.33%. This precision value, falling below the desired 90% benchmark, indicated the need for a logistic regression analysis to determine a more accurate threshold.

Using the statistical computing software script provided in the protocol (step 4.1.4), a logistic regression model was fitted to the validation data to determine the relationship between the detector's confidence score and the probability of a correct detection. The model's coefficients were used to calculate a new, optimal confidence score threshold of 0.36 to achieve a precision of 90% or higher (**Figure 3**). Logistic regression analysis showed that the machine learning detector modules optimised confidence score was a significant predictor of correct classification ($\beta = 9.71 \pm 1.79$ SE, $z = 5.42$, $p < 0.001$). The model showed improved fit relative to the null model (residual deviance = 268.88 vs. null deviance = 353.87; AIC = 272.88; n = 300).

[Place **Figure 3** here].

The machine learning detector module[19] was re-run on the full 1-month dataset recorded at the Glasgow Suburban site, with this new threshold of 0.36, following step 4.2 of the protocol, with and without an additional increase in overlap (0 and 2 s). A new decreased set of 346 detections was generated for the default overlap of 0 s, whereas 1003 detections were generated with the increased overlap of 2 s. A subsequent validation of these refined datasets confirmed a substantial improvement in precision. The post-regression precision with the default overlap of 0 s increased to 89.33%, and with the increased overlap of 2 s the precision increased to 87.33%. This outcome demonstrates that threshold optimisation improves precision but may not always achieve predefined targets in acoustically complex environments.

These representative results demonstrate that while a default threshold may provide a reasonable starting point, the protocol's use of a species and location-specific logistic regression analysis is crucial for generating a highly accurate and reliable dataset for further ecological analysis. For the three sites used to illustrate the process (Liverpool Park, Cairngorms, and Glasgow Suburban), precision increased by 26.1%, 17.7%, and 17% for an overlap of 0 s, and by 28.77%, 16.87%, and 15% for an overlap of 2 s. Overall precision improved by 15–28.77% across sites following the optimisation of confidence scores, and final precision ranged from 87.33–100%, demonstrating that protocol performance varies depending on acoustic environment and species assemblages.

FIGURE AND TABLE LEGENDS

Figure 1: Logistic regression showing the probability of correct detection as a function of the machine learning–based avian sound recognition model confidence scores for Robin detections at ARU 1 at the Liverpool Park site. Points represent manually validated detections (0 = false positive, 1 = true positive). The solid blue line represents the fitted logistic regression model, and

the shaded region represents the 95% confidence interval. The vertical green dashed line marks the derived confidence score threshold corresponding to 90% predicted precision.

Figure 2: Logistic regression showing the probability of correct detection as a function of the machine learning–based avian sound recognition model confidence scores for Robin detections at ARU 2 at the Cairngorms site. Points represent manually validated detections (0 = false positive, 1 = true positive). The solid blue line represents the fitted logistic regression model, and the shaded region represents the 95% confidence interval. The vertical green dashed line marks the derived confidence score threshold corresponding to 90% predicted precision.

Figure 3: Logistic regression showing the probability of correct detection as a function of the machine learning–based avian sound recognition model confidence scores for Robin detections at ARU 3 at the Glasgow site. Points represent manually validated detections (0 = false positive, 1 = true positive). The solid blue line represents the fitted logistic regression model, and the shaded region represents the 95% confidence interval. The vertical green dashed line marks the derived confidence score threshold corresponding to 90% predicted precision.

Table 1: Comparison of Robin detection metrics before and after applying the calculated confidence score threshold. Precision (%) was calculated as the proportion of manually validated true positives relative to the total validation detection count. Precision (%) was tested at both a 0 s overlap (ol 0) and 2 s overlap (ol 2), which were applied during detector configuration.

DISCUSSION

The use of the protocol set out in this paper provides a reliable, streamlined method for generating species-specific and accurately classified datasets of bird vocalisations, which has proven useful when handling large audio datasets. We chose to test the protocol on the European Robin due to its well-known structural and frequency variability within its song repertoires[43]. Meaning initially, detections were found to have high mislabelling (false positive) rates (**Table 1**). We also validated the use of these methods with post-protocol testing to calculate increases in precision rates using the newly derived confidence score thresholds. The representative results use manual validation of Robin vocalisations and subsequent logistic regression analysis to obtain confidence score thresholds to increase detection precision when using the machine learning detector module[19]. The aim was to increase precision to at least 90%, which was achieved for two of the three location-based datasets. Precision for ARU 2 in the Cairngorms increased to 100% (overlap 0) and 99.17% (overlap 2). Precision for ARU 1 in Liverpool Park increased to 92.66% (overlap 0) and 95.33% (overlap 2). Critical steps in this protocol that strongly influenced detection precision, include randomisation and correct manual validation of detections, accurate fitting of the logistic regression model to the validation datasets, and application of the newly derived confidence score threshold during reconfiguration of the detector.

Precision did increase for ARU 3 at the Glasgow Suburban site, to 89.33% (overlap 0) and 87.33% (overlap 2), when using the confidence score thresholds derived from the protocol, representing increases of 17% and 15% in precision. However, these values failed to reach the statistically determined 90% precision rate. The false positives observed here were primarily associated with mislabelling of acoustically similar species and noise as Robin vocalisations. Song Thrush (*Turdus philomelos)* song and the introductory phrases of Blue Tit songs (*Cyanistes caeruleus*) were often mislabelled as Robin vocalisations, and non-biological sounds such as rustling and rain were mislabelled as Robin 'tic' alarm calls. In some instances, these false positives were given higher confidence scores than some true positives. Further increase of the confidence score may improve precision slightly, however this would result in a reduction in recall. The persistence of false positives due to acoustically similar species and noise are the limitations of the model under specific environmental conditions. In practice, if precision remains below the desired threshold, users may consider increasing the validation sample size, reassessing the model fit, adjusting the overlap settings, or examining recurring and persistent sources of misclassification within the audio data before reapplying the threshold optimisation. Although the precision target was not reached for this site, these findings emphasise the variance in learning detector performance for the same species across different locations with varying acoustic environments. This supports previous observations that calculated optimal confidence score thresholds are environment-specific[19], often varying dramatically depending on species assemblages and background noise levels.

Additionally, as shown in **Figures 1**, **2**, and **3,** there are detections that fall below the newly derived thresholds that were manually identified as correct. This highlights one key trade-off with this protocol. Although this process reduces false positives, it may also exclude true positives and ultimately reduce recall. This effect has been demonstrated in other research looking at the effects of increasing the confidence score threshold[26]. However, given the large volumes of data produced by ARUs, comprehensive manual identification of species is impractical at scale and prone to human error[3,24]. In these representative results, manual review was used as a reference standard for evaluating the machine learning–based avian sound recognition model detections[21]. However, this does not imply that this method of review is fully precise, rather it serves as an independent benchmark to assess the protocol. In this context, statistically selected thresholds offer a practical compromise to allow for the processing of large datasets while still allowing for substantial numbers of detections. Where recall is of priority, a possible alternative method that incorporates this, involves calculating F-scores across a range of confidence thresholds and selecting the threshold that maximises the F-score[29,30].

The representative results here for the Glasgow and Cairngorms sites yield a higher precision rate using the newly generated confidence scores with the default overlap of 0 s, compared to 2 s. However, the 0 s overlap setting resulted in fewer total detections (**Table 1**), indicating reduced recall (i.e., a greater number of missed detections). In contrast, increasing the overlap to 2 s increased the total number of detections but also increased false positives, reducing overall precision. One explanation is that increased temporal overlap increases the likelihood that non-target, mislabelled vocalisations are segmented multiple times, inflating the false positive rate. This finding contradicts previous studies reporting that an increase in overlap will increase detection resolution and precision[19,29]. Our results suggest that the effect of overlap may be context dependent and influenced by call structure and background noise. Researchers should therefore consider the trade-off between precision and recall when selecting an overlap. If maximising recall is priority, adopting a larger overlap may be justified, provided the increase in manual validation time is accounted for. However, for this study each ARU recorded 1-month of data, and large numbers of accurate detections were still produced after adjusting the confidence score threshold, even with the lower 0 s overlap (**Table 1**).

The use of a statistically derived, species- and location-specific confidence score, is likely superior to techniques of simply increasing thresholds to reduce false positives. As demonstrated by Tseng et al[26] higher thresholds yield a large number of false negatives (missed identifications). If higher thresholds were applied to the sites represented in this study, by theoretically shifting the threshold lines forwards in **Figures 1, 2** and **3**, a loss of a large proportion of true positives would occur. Conversely, applying a threshold derived for one site to a different acoustic environment may result in the opposite problem. A threshold that gives precise detections for a quieter site may be too low when transferred to a noisier site with more complex sounds and species assemblages, resulting in inflated numbers of false positives.

Furthermore, the variability observed between the statistically derived thresholds in this study (0.23–0.54) for the same species across varying sites, indicates that a universal threshold may not be suitable. As this variation is reflective of differences in environmental noise and species assemblages among sites.

The representative results here support the use of this protocol in improving detection precision when utilising the machine learning detector module[19] to identify vocalisations within analysis software[31]. This process will allow users to efficiently filter large datasets, reducing the time and effort required for full manual validation of detections. The application of the logistic regression derived threshold means users can be confident in producing higher volumes of accurate detections. These detections can be used for a multitude of research purposes, such as for annotation and parameter extraction, for monitoring of species presence, creating training datasets, and for conducting vocal behaviour analysis. The differences in the precision and thresholds derived from each location for the same species also highlight the importance of location based and acoustic environmental context. Reinforcing the importance of utilising this process when using audio data from multiple locations, with varying acoustic environments.

ACKNOWLEDGMENTS
We thank Liverpool John Moore's University for guidance and resources. We also thank the Wild Animal Initiative (WAI) for funding support, this research was funded by WAI grants (S-2023–00038).

DISCLOSURES
Authors have no competing financial interest.

Author's contribution
Bethany Shackleton: Conceptualisation, methodology, data analysis and manuscript writing.
Luiza Passos: Methodology development and manuscript writing.
Ross Macleod: Methodology development and manuscript writing.

REFERENCES

1. Briefer E. F. Vocal expression of emotions in mammals: Mechanisms of production and evidence. *J Zool*. **288** (1), 1–20 (2012). https://doi.org/10.1111/j.1469-7998.2012.00920.x
2. Changapur M. W., Seema S., Sowmya B. J. Bioacoustics monitoring to improve conservation efforts for endangered species [Conference Paper]. Presented at: 2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS); *IEEE.* (2023). Available from: https://doi.org/10.1109/CSITSS60515.2023.10334168
3. Sebastián-González E., Pang-Ching J., Barbosa J. M., Hart P. Bioacoustics for species management: two case studies with a Hawaiian forest bird. *Ecol Evol*. **5** (20), 4696–4705 (2015). https://doi.org/10.1002/ece3.1743
4. Raich X. et al. Use of bioacoustics in species identification: Piranhas from genus *Pygocentrus* (Teleostei: Serrasalmidae) as a case study. *PLOS One*. **15** (10), e0241316 (2020). https://doi.org/10.1371/journal.pone.0241316
5. Pillay R. et al. Bioacoustic monitoring reveals shifts in breeding songbird populations and singing behaviour with selective logging in tropical forests. *J Appl Ecol*. **56** (11), 2482–2492 (2019). https://doi.org/10.1111/1365-2664.13492
6. Teixeira D., Maron M., van Rensburg B. J. Bioacoustic monitoring of animal vocal behaviour for conservation. *Conserv Sci Pract.* **1** (8), e72 (2019). https://doi.org/10.1111/csp2.72
7. Coutant M., Villain A. S., Briefer E. L. A scoping review of the use of bioacoustics to assess various components of farm animal welfare. *Appl Anim Behav Sci*. **275**, 106286 (2024). https://doi.org/10.1016/j.applanim.2024.106286
8. Mcloughlin M. P., Stewart R., McElligott A. G. Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. *J R Soc Interface*. **16** (155), 20190225 (2019). https://doi.org/10.1098/rsif.2019.0225
9. Shonfield J., Bayne E. M. Autonomous recording units in avian ecological research: Current use and future applications. *Avian Conserv Ecol*. **12** (1), 14 (2017). https://doi.org/10.5751/ACE-00974-120114
10. Newson S. E., Bas Y., Murray A., Gillings S. Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. *Methods Ecol Evol*. **8** (9), 1051–1062 (2017). https://doi.org/10.1111/2041-210X.12720
11. Wheeldon A., et al. Comparison of acoustic and traditional point count methods to assess bird diversity and composition in the Aberdare National Park, Kenya. *Afr J Ecol*. **57** (2), 168–176 (2019). https://doi.org/10.1111/aje.12596
12. Zwart M. C., Baker A., McGowan P. J. K., Whittingham M. J. The Use of Automated Bioacoustic Recorders to Replace Human Wildlife Surveys: An Example Using Nightjars. *PLoS ONE.* **9** (7), e102770 (2014). https://doi.org/10.1371/journal.pone.0102770

13. Xie J. et al. Review of automatic recognition technology for bird vocalizations in the deep learning era. *Ecol Inform*. **73**, 101927 (2023). https://doi.org/10.1016/j.ecoinf.2022.101927
14. Yang F., Jiang Y., Xu Y. Design of bird sound recognition model based on lightweight. *IEEE Access*. **10**, 85189–85198 (2023). https://doi.org/10.1109/ACCESS.2022.3198104
15. Liao Z., et al. Automatic recognition of giant panda vocalizations using wide spectrum features and deep neural network. *Math Biosci Eng.* **20** (8), 15456–15475 (2023). https://doi.org/10.3934/mbe.2023690
16. Jaafar H., Ramli D. A. Automatic syllables segmentation for frog identification system [Conference Paper]. Presented at: *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*; Kuala Lumpur, Malaysia. 224–228 (2013). https://ieeexplore.ieee.org/document/6530046.
17. Ware L., Mahon C. L., McLeod L., Jetté J. F. Artificial intelligence (BirdNET) supplements manual methods to maximize bird species richness from acoustic data sets generated from regional monitoring. *Can J Zool*. **101** (12), 1031–1051 (2023). https://doi.org/10.1139/cjz-2023-0044
18. Kahl S., Wood C. M., Eibl M., Klinck H. BirdNET: A deep learning solution for avian diversity monitoring. *Ecol Inform.* **61**, 101236 (2021). https://doi.org/10.1016/j.ecoinf.2021.101236
19. Kahl K., Wood C., Klinck M. Cornell Lab of Ornithology. *BirdNET: Bird sound identification* [computer program]. Ithica (NY): The Cornell Lab of Ornithology. (2021). https://birdnet.cornell.edu
20. Manzano-Rubio R., et al. Low-cost open-source recorders and ready-to-use machine learning approaches provide effective monitoring of threatened species. *Ecol Inform.* **72**, 101910 (2022). https://doi.org/10.1016/j.ecoinf.2022.101910
21. Wood C. M, Kahl S. Guidelines on the appropriate use of BirdNet scores and other detector outputs. *J Ornithol*. **165**, 777–782 (2024). https://doi.org/10.1007/s10336-024-02144-5.
22. Pérez-Granados C. BirdNET: applications, performance, pitfalls and future opportunities. *Ibis.* **165** (3), 1068–1075 (2023). https://doi.org/10.1111/ibi.13193.
23. Singer D., et al. Aggregated time-series features boost species-specific differentiation of true and false positives in passive acoustic monitoring of bird assemblages. *Remote Sens Ecol Conserv*. **10** (4), 517–530 (2024). https://doi.org/10.1002/rse2.385
24. Bota G., et al. Hearing to the Unseen: AudioMoth and BirdNET as a Cheap and Easy Method for Monitoring Cryptic Bird Species. *Sensors.* **23** (16), 7176 (2023). https://doi.org/10.3390/s23167176.
25. Amorós-Ausina A., Schuchmann K-L. Marques M. I., Pérez-Granados, C. Living together, singing together: Revealing similar patterns of vocal activity in two tropical songbirds applying BirdNET. *Sensors.* **24** (17), 5780 (2024). https://doi.org/10.3390/s24175780
26. Tseng S., Hodder D. P, Otter K. A. Setting BirdNET confidence thresholds: species-specific vs. universal approaches. *J Ornithol.* **166** (4), 1123–1135 (2025). https://doi.org/10.1007/s10336-025-02260-w

27. Wood C. M., Cruz A. B., Kahl S. Pairing a user-friendly machine-learning animal sound detector with passive acoustic surveys for occupancy modelling of an endangered primate. *Am J Primatol*. **85** (8), e23507 (2023). httsp://doi.org/10.1002/ajp.23507.
28. Sossover D., Burrows K., Kahl S., Wood, C. M. Using the BirdNET algorithm to identify wolves, coyotes, and potentially their interactions in a large audio dataset. *Mamm Res*. **69**, 159–165 (2024). https://doi.org/10.1007/s13364-023-00725-y
29. Funosas D., et al. Assessing the potential of BirdNET to infer European bird communities from large-scale ecoacoustic data. *Ecol Indic*. **164**, 112146 (2023). https://doi.org/10.1101/2023.12.06.570351
30. Ziegenhorn M. A., et al. ArcticSoundsNET: BirdNET embeddings facilitate improved bioacoustic classification of Arctic species. *Ecol Inform*. **90**, 103270 (2025). https//doi.org/10.1016/j.ecoinf.2025.103270
31. Cornell Lab of Ornithology. Raven Pro: Interactive sound analysis software [Computer program] (Version 1.6). *The Cornell Lab of Ornithology.* (2025). https://ravensoundsoftware.com/
32. R Core Team. *R: A language and environment for statistical computing* [computer program]. *R Foundation for Statistical Computing.* (2024). https://www.R-project.org/
33. Python Software Foundation. *Python Language Reference* [Internet] (Version 3.14.3). *Python Software Foundation*. (2026). https://www.python.org
34. Hill A. P., et al. AudioMoth: Evaluation of a Smart open Acoustic Device for Monitoring Biodiversity and the Environment. *Methods in Ecol and Evol*. **9** (5), 1199–1211 (2017). https://doi.org/10.1111/2041-210X.12955
35. Hill A. P, et al. AudioMoth: A low-cost acoustic device for monitoring biodiversity and the environment. *HardwareX*. **6**, e00073 (2019). https://doi.org/10.1016/j.ohx.2019.e00073
36. Bremond J. C. Recherche sur la semantique et les elements vecteurs d'information dans les signaux aeoustiques du rouge-gorge (*Erithacus rubecula*). *Terre Vie*. **22**, 109–220 (1968). https://hal.science/hal-03531635v1
37. Hoelzel A. R. Song characteristics and response to playback of male and female robins, *Erithacus rubecula*. *Ibis*. **128**, 115–127 (1968). https://doi.org/10.1111/j.1474-919X.1986.tb02098.x
38. Sandbrook C. et al. Principles for the socially responsible use of conservation monitoring technology and data. *Conserv Sci Pract*. **3** (5), e374 (2021). https://doi.org/10.1111/csp2.374
39. Open Acoustic Devices. *AudioMoth Flash App* [computer program] (Version 1.7.0.) *Open Acoustic Devices* (2025). https://www.openacousticdevices.info/applications.
40. Open Acoustic Devices. *AudioMoth Configuration App* [computer program] (Version 1.12.1.) *Open Acoustic Devices* (2025). https://www.openacousticdevices.info/applications.
41. Open Acoustic Devices. Applications [Internet]. *Open Acoustic Devices* (2025). https://www.openacousticdevices.info/applications.
42. Sethi S. S. et al. Large-scale avian vocalization detection delivers reliable global biodiversity insights. *Proc Natl Acad Sci U S A*. **121** (33), e2315933121 (2024). https://doi.org/10.1073/pnas.2315933121

43. Verboom W. C. Bird vocalizations: songs of the European robin (*Erithacus rubecula*). *JunoBioacoustics.* 202107 (2021). https://doi.org/10.13140/RG.2.2.29357.82406