Comparing Automated Species Identification Classifiers for Acoustic Bat Data from Anabat and AudioMoth Detectors

Authors: Perks, Samantha J., O'Connell, Mark J., and Goodenough, Anne E.

Source: Acta Chiropterologica, 28(1)

Published By: Museum and Institute of Zoology, Polish Academy of Sciences

URL: https://doi.org/10.3161/15081109ACC2026.28.1.011

Acta Chiropterologica 2026, vol. 28, 11: 1–11
e-ISSN 1733-5329, © Museum and Institute of Zoology Polish Academy of Sciences

DOI 10.3161/15081109ACC2026.28.1.011

Comparing automated species identification classifiers for acoustic bat data from Anabat and AudioMoth detectors

Samantha J. Perks[1] ✉, Mark J. O'Connell[1], Anne E. Goodenough[1]

[1]School of Education and Science, University of Gloucestershire, Francis Close Hall, Swindon Road, GL50 4AZ, Cheltenham, United Kingdom
✉ sperks2@glos.ac.uk

Abstract. Passive Acoustic Monitoring (PAM) is becoming an increasingly popular method for surveying and monitoring bats within conservation and scientific research contexts, as well as for legislative complacence. Although PAM protocols have the potential to be standardisable and scalable, they typically produce vast acoustic datasets that require considerable time and ecological skills to analyse manually. With the continually evolving capabilities of Artificial Intelligence (AI), several automated classifiers now exist to classify bat guilds, which have the potential to streamline analysis workflows considerably. However, uncertainty remains regarding the reliability of such classifiers, particularly as regards how their outputs differ from one another and how their performance is impacted by differential recording quality from different types of detector. Here, agreement between three commonly used classifiers (Kaleidoscope Pro, BatClassify in Anabat Insight, and the British Trust for Ornithology Acoustic Pipeline) is investigated for acoustic recordings from commercial bat detectors (Anabat Swift; n = 13,582 recordings) and open-source acoustic recorders (AudioMoth; n = 30,574 recordings). Habitat type (riparian, woodland, wood pasture, arable) and bat species both affected the level of agreement between detectors. Disagreement was most prevalent on recordings from more structurally complex habitats such as woodland and where a species within the genus *Myotis* was classified by at least one classifier. There were also differential interactions between habitat and taxonomy depending on detector type. For Anabat recordings, disagreement was high for *Barbastella* and *Plecotus* in woodland and wood pasture relative to riparian and arable; for AudioMoth recordings disagreement was high for *Nyctalus/Eptesicus* in woodland and *Rhinolophus* in riparian habitat. These findings suggest that classifiers must be used with caution and in conjunction with manual auditing carried out by suitable skilled practitioners, especially when errors in classification have consequences within formal legislative frameworks.

Keywords: PAM, bats, automated detection, Anabat, AudioMoth, bioacoustics, automated classification, automated identification.

INTRODUCTION

Passive Acoustic Monitoring (PAM) is becoming an increasingly important method in ecological research (Browning *et al.* 2017; Gibb *et al.* 2018; Teixeira *et al.* 2019). It is especially useful for compiling species inventories and monitoring spatiotemporal change in practitioner and conservation contexts (Wrege *et al.* 2017; López-Bosch *et al.* 2022), as well as ensuring compliance with protected species legislation (Collins 2023). Because acoustic methods are especially important in the study of bats, they are usually considered to be the terrestrial taxa for which acoustic analysis is most developed (Sugai *et al.* 2019). Despite this, there are ongoing challenges when using PAM for bats. Firstly, there is potential for detector choice and deployment to affect what species are detected, and recorded activity levels (Kunberger and Long 2023; Starbuck *et al.* 2024; Perks *et al.* 2025). Secondly, and possibly even more pressingly, diminishing detector costs and the growing prominence and scalability of PAM means that ensuring time-efficient, cost-effective and reliable analysis of large datasets is becoming ever-more challenging (Browning *et al.* 2017; Gibb *et al.* 2018; Sugai *et al.* 2019; Brinkløv *et al.* 2023). Such challenges are compounded by manual processing being potentially impacted by user subjectivity and error (Gibb *et al.* 2018). These issues have driven the development of automated tools that aim to improve the efficiency, reliability and objectivity with which species-specific data can be extracted from acoustic recordings (Stowell 2022).

Initial processing of acoustic bat data typically starts by practitioners using filtering to remove files containing sound recordings that are unlikely to contain bat vocalisations. This is done using pre-defined criteria, usually based on upper and lower frequency thresholds (Mac Aodha *et al.* 2018). For open-source, non-bat-specific detectors that lack on-board triggers (or have triggers that have not been empirically tested), filtering is especially important as the entire soundscape is recorded. For detectors that have an on-board threshold trigger,

Recd 27 Nov. 2024, **Acc.** 06 June 2026, **Pub.** 30 June 2026
Corresponding editor: M. Corrie Schoeman

filtering is primarily done by the detector in the field as noises outside specified sound parameters are not recorded. However, all sounds within range will be retained, including ultrasonic vocalisations of species other than bats (e.g., high-frequency vocalisations of small mammals, birds, and insects) or false triggers (e.g., rain, wind, or man-made sounds), which need to be removed either before or during the identification of species within the recordings (Brinkløv *et al.* 2023).

Automated classifiers for acoustic bat data first became available in the 1990s. These were based on a broad range of underlying analytical approaches, including Artificial Neural Networks (ANNs) and machine learning algorithms such as Random Forest techniques (Zamora-Gutierrez *et al.* 2021). Modern classifiers, including those using Artificial Intelligence frameworks, involve species being identified by algorithms developed using known call parameters (e.g., amplitude, frequency) and/or matching sonograms to a pre-verified call library (Gibb *et al.* 2018). Several such classifiers are now available to practitioners and researchers, with more currently in development. These include desktop-based programmes that are integrated into analysis software, for example the suite of customisable regional classifiers in Kaleidoscope Pro (Wildlife Acoustics, Maynard, MD, USA). Alternatively, recent technological advances have led to the launch of cloud-based platforms, for example, the British Trust for Ornithology (BTO) Acoustic Pipeline (BTO, Thetford, UK), which launched in 2021 and employs machine-learning algorithms trained by a dataset of > 90,000 of verified recordings to classify bat echolocation/social calls, and feeding buzzes (British Trust for Ornithology 2025). In some instances, classifiers are freely available to end-users, or at least for recordings made using specific detectors. For example, BatClassify, a tree-based ensemble classifier developed by Scott and Altringham (2014*a*) specifically for UK bats in woodland, is available both open-access as a standalone classifier, and in the freeware version of the Anabat Insight analysis software (Titley Scientific, Brendale, QLD, Australia) for files recorded on Titley Scientific devices; in this instance, data recorded on open-source units can only be analysed if a licence is purchased to unlock additional functionality. Other tools, such as the classifiers built into Kaleidoscope Pro and the standalone BTO Acoustic Pipeline, allow limited datasets to be processed for free, regardless of device. However, purchase of a software licence (or 'credits') is necessary for long-term access or to process large datasets. It should be noted that even where use of an automated classifier incurs a financial cost (e.g., approx. 310 GBP for Kaleidoscope Pro per year, 435 GBP for the full version of Anabat Insight in perpetuity; September 2024 costings), this can sometimes be more cost efficient for organisations compared to the staff-hours required for the manual analysis of large datasets (Adams *et al.* 2010). In theory, therefore, classifiers can provide accessible and scalable PAM processes for resource-limited conservation organisations, as well as for citizen science initiatives such as the British Bat Survey (BbatS) in the UK (Bat Conservation Trust 2024).

Despite the potential benefits of automated analysis, multiple facets of classifiers remain largely untested, such that an end user cannot necessarily be confident in the output. For example, although there has been continual development of the technology and an expansion of the call libraries used as training data, significant uncertainties remain regarding variation and error rates between different classifiers, as well as how this might be impacted by recording quality or situations where recordings are complicated by background noise or detector self-noise (Gibb *et al.* 2018; Brinkløv *et al.* 2023). Algorithms can also struggle to recognise variations in calls of individual bat species, which can occur due to differences in habitat structure (especially in 'cluttered' habitats such as woodland) and weather conditions, as well as bat-specific characteristics including age and sex (López-Baucells *et al.* 2019). Taken together, these factors mean inaccurate classifications can occur, and, importantly, the level of risk remains unknown and could be context dependent (Mac Aodha *et al.* 2018; Barré *et al.* 2019). Although a small number of previous studies have compared classifier performance, these have largely been undertaken using recordings from North American bat guilds (e.g., Lemen *et al.* 2015; Nocera *et al.* 2019; Goodwin and Gillam 2021). There has seemingly been just one study in Europe: Rydell *et al.* (2017) compared the performance of three classifiers: SonoChiro (Biotope, Mèze, France), Kaleidoscope and BatClassify, on a bat guild in Sweden. Such comparisons are yet to be conducted for newer classifiers, such as the BTO Acoustic Pipeline, and classification of recordings from open-source acoustic recorders with lower-quality microphones, such as AudioMoth units, has also yet to be tested. Without a robust understanding of the error associated with automated workflows as a function of the detector type and classifier used, an element of manual auditing is still considered best practice. This limits the time and cost advantages that should be inherent in automating the species identification process (Barré *et al.* 2019; López-Baucells *et al.* 2019; Collins 2023).

Here, we compare a suite of automated classifiers (Kaleidoscope Pro, BatClassify in Anabat Insight, BTO Acoustic Pipeline) on two large full spectrum bat datasets collected across four habitats in the UK: riparian, woodland, wood pasture and arable (Perks *et al.* 2025). Each dataset was recorded using a different type of detector, either Anabat Swift (Titley Scientific, Brendale, QLD, Australia), or AudioMoth (Open Acoustic Devices, UK). Our first objective is to assess the level of agreement between a pair of classifiers on each of the datasets (Kaleidoscope Pro versus BatClassify for Anabat Swift recordings; Kaleidoscope Pro versus BTO Acoustic Pipeline for AudioMoth recordings), factoring in taxonomy, habitat, and the interaction between these factors. Our second objective is to consider each pair of classifiers to understand the frequency with which specific species are confused.

MATERIAL AND METHODS

Study design and data collection

We collected acoustic data at four sites in the county of Worcestershire, UK, over a 16-week period between June 2022 and October 2022 (see Perks *et al.* 2025) for more details of site location). Each site supported a different single habitat that collectively represented a spectrum of suitability for bats from high-quality (riparian; woodland) to medium-quality (wood pasture) and low-quality (arable) (Collins 2023). We used two full spectrum detector types, firstly commercial units (Anabat Swift, $n = 2$) and secondly open-source units (AudioMoth, $n = 10$). Details of detector configuration are given in Fig. 1; all detectors (regardless of type) were located at least 150 m apart from any other detector to ensure that recordings were independent. The only difference in configuration was that, for the Anabat Swift units, the on-board trigger was used with default settings to trigger recordings in the frequency range 10–150 kHz. At the time of study, the AudioMoths lacked a reliable on-board trigger, so were continually recording all sounds in the frequency range 0–125 kHz on a 5-second record ('wake') cycle 20-second sleep cycle. Acoustic data were recorded over a total of 80 nights (20 nights × 4 sites). At a site level, all recordings from both Anabat Swift detectors were pooled to create a single set of recordings (termed 'Anabat recordings'); the same process was undertaken for the AudioMoths by pooling recordings from all 10 units per site to create another single set of recordings (termed 'AudioMoth recordings'). Our data processing and analysis is detailed more fully below, but the study design was fundamentally based on comparing bat identifications made by two different classifiers per recording on the same detector (Fig. 1). This meant that each individual Anabat recording was classified using two classifiers (Anabat Insight and Kaleidoscope), while each individual AudioMoth recording was classified using two classifiers (BTO Acoustic Pipeline and Kaleidoscope). Although there were data from multiple units of the same detector type in the pooled data for each site, the inter-detector distance meant that the same call would not have been recorded multiple times, such that each line of data was independent and not pseudoreplicated. We never compare classifications made on different detector types and detectors were not 'paired' either in the field or within analysis.

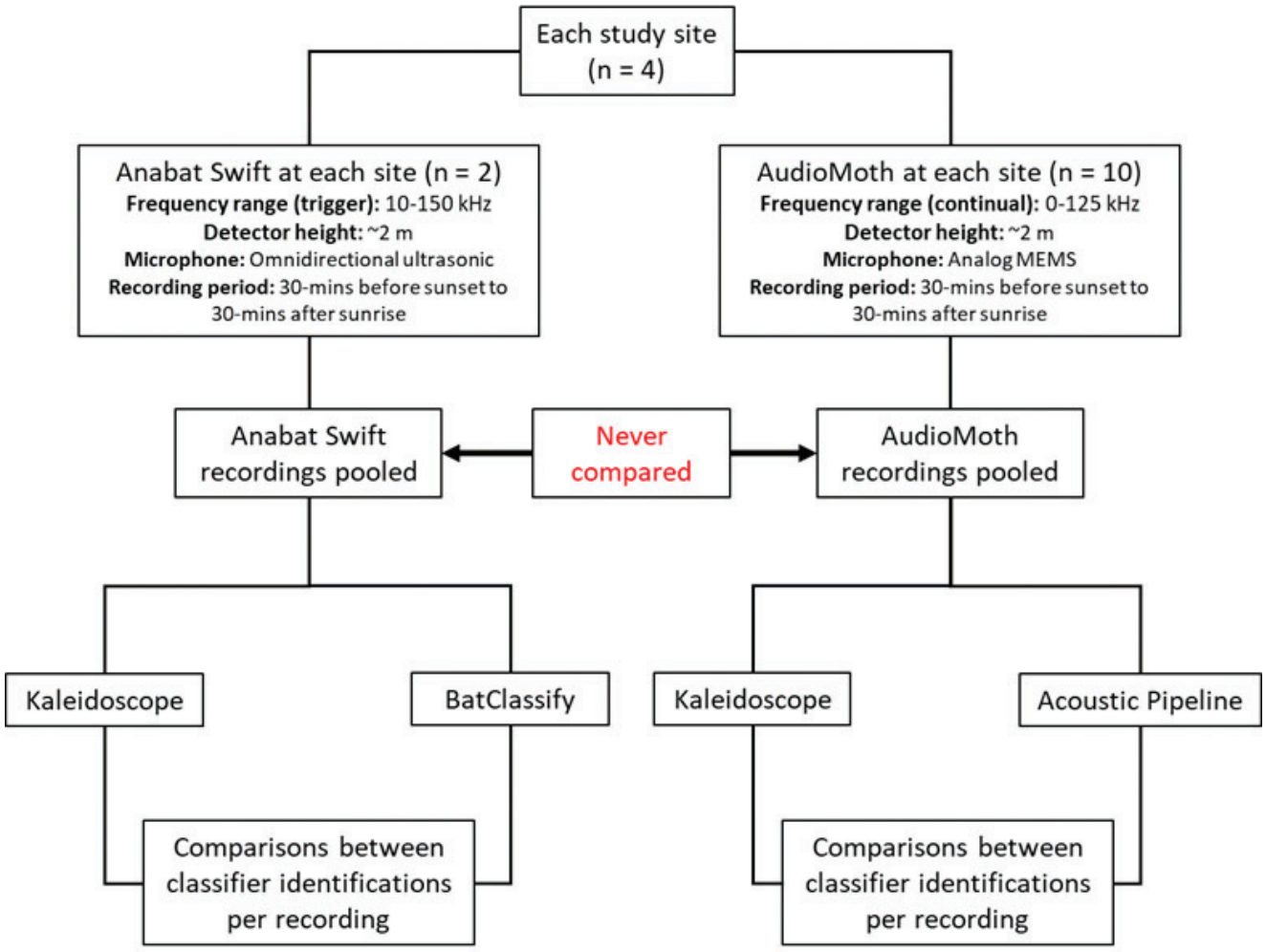

Figure 1. Data collection and classification workflow (MEMS = micro-electrical mechanical systems).

Data processing and classification

Prior to automated classification, we passed all Anabat and AudioMoth recordings through a broad frequency filter ('All bats' in Anabat Insight) to remove recordings of sounds < 4 kHz or > 300 kHz. This removed fewer Anabat recordings (where the on-board trigger had been used in-situ) relative to AudioMoth recordings when the entire soundscape was recorded during the wake part of the sleep:wake cycle. Regardless of detector/recording type, we classified all files that remained after the broad frequency filter had been applied using the Bats of Europe classifier (v 5.4.0) in Kaleidoscope Pro (v 5.4.8). This gave a single species classification for each recording that contained a 'bat pass' or 'call sequence', as well as a classifier-determined match ratio (confidence value) based on the number of calls in the sequence that matched the reference library (range 0–1; higher = better). To allow direct comparison between classifiers, we additionally classified the Anabat recordings using BatClassify integrated into Anabat Insight as a second classifier that was optimised for Anabat devices. Each classification was given a confidence value (range 0–100%; higher = better) by the classifier. A similar process was used for AudioMoth recordings, which, after being classified by Kaleidoscope, we additionally classified using the BTO Acoustic Pipeline as a second classifier. This contained numerous verified reference calls recorded on AudioMoth devices (Stewart Newson, Acoustic Pipeline Developer, personal communication). Each classification was given a probability (range 0–1; higher = better reliability) by the classifier.

After processing through the classifiers (Kaleidoscope and BatClassify for Anabat recordings; Kaleidoscope and Acoustic Pipeline for AudioMoth recordings), we undertook a screening process to refine the datasets before statistical analysis. Firstly, only those recordings with a match ratio (Kaleidoscope), probability (Acoustic Pipeline), or confidence value (BatClassify) ≥ 0.5 or 50%, were retained (henceforth 'confidence' as a percentage, is used in all cases for ease of reporting). This threshold, although arbitrary, was consistent with current recommendations for classifier-assisted analysis of acoustic datasets (Barre *et al.* 2019; British Trust for Ornithology 2024), and followed the protocols used in numerous previous studies (Braun de Torrez *et al.* 2017; Springall *et al.* 2019; Smith *et al.* 2021; Taillie *et al.* 2021). Secondly, we screened all recordings classified using BatClassify and Acoustic Pipeline for instances where multiple species had been identified. Although these two classifiers can, in theory, recognise calls from multiple species within the same recording, Kaleidoscope was restricted to assigning a single classification per recording. To avoid systematic bias due to differential classification frameworks confounding statistical analysis, we removed any files classified as containing calls from multiple species by BatClassify or by the Acoustic Pipeline, along with the corresponding files classified by Kaleidoscope. This ensured that all comparisons were based recordings that only contained a single species. Finally, we created a single combined ('matched') dataset for Anabat data that contained only files classified ≥ 50% by both Kaleidoscope and BatClassify (n = 13,582). We used the same process for AudioMoth data, where a single matched dataset was created that contained only files classified ≥ 50% by both Kaleidoscope and the Acoustic Pipeline (n = 30,574).

Statistical analysis

Initial data exploration considered differences in classifier confidence, firstly in relation to habitat and secondly in relation to taxonomy, by comparing each species/taxonomic group individually. Then, to analyse patterns statistically, we created a new ordinal variable for each matched dataset. This summarised, on a recording-by-recording basis, whether the two classifiers agreed on species identification (or not), and, if they did, how confident the classifiers were in the identification provided using an objective, rule-based 5-point scale, as detailed in Table 1. The increased granularity provided by the ordinal variable, versus a simple agreement/disagreement binary variable, allowed more nuanced and meaningful

Table 1. Criteria used to score the level of agreement between two classifiers on a recording-by-recording basis to create a new ordinal variable for the Anabat dataset and the AudioMoth dataset. Note that the ordinal scale ran between 1 (no match) and the 2–6 (match with increasing confidence) rather than 0 (no match) and 1–5 (match with increasing confidence) because the statistical approach adopted could not be used with zero data.

Level of agreement between classifiers	Criteria	Anabat		AudioMoth	
		n	%	*n*	%
1	Species do not match	692	5.09	2,919	9.55
2	Species match, both classifiers 50–74% confident	70	0.52	114	0.37
3	Species match, one classifier 50–74% confident and the other classifier ≥ 75% confident	1,702	12.53	2,941	9.62
4	Species match, both classifiers 75–98% confident	3,784	27.86	5,624	18.39
5	Species match, one classifier 75–98% confident and the other classifier ≥ 99% confident	6,588	48.51	12,640	41.34
6	Species match, both classifiers ≥ 99% confident	746	5.49	6,336	20.72

analysis to be conducted. Because the classifier outputs included both a categorical species label and an associated confidence value, it was important to represent situations where the two classifiers agreed partially (e.g., one high confidence and one low confidence classification of the same species), in addition to cases of strong mutual confidence, which would be considered equal under a binary approach. Match strength was scored for each individual recording at species level across all possible species in the dataset.

To explore the effects of taxonomic group and habitat on the probability of obtaining a particular level of classifier agreement, we used a Cumulative Link Model (CLM) with a logit link function for the matched Anabat data (comparing Kaleidoscope and BatClassify) in R 4.2.2 (R Core Team 2022) using the 'ordinal' package (Christensen 2023). We used a second model for the matched AudioMoth data (comparing Kaleidoscope and Acoustic Pipeline). In each model, we used the ordinal variable summarising pairwise classifier agreement (Table 1) as the dependent variable, with two predictor variables being added as fixed factors: habitat (riparian, woodland, wood pasture, arable) and taxonomic group, as well as the interaction between them. We used CLMs and an ordinal response variable to allow for greater granularity and statistical power that would have been possible using binary data and logistic regression models, as the ordinal variable allowed us the degree of agreement between classifiers. We used AIC scores to assess model fit relative to a null model (Burnham and Anderson 2002). To avoid overparameterizing the model, we combined individual species into one of six taxonomic groups rather than each individual species being used, thus: (1) *Barbastella* comprising Western barbastelle (*Barbastella barbastellus*) only; (2) *Myotis* comprising Alcathoe (*Myotis alcathoe*), Brandt's (*M. brandtii*), Bechstein's (*M. bechsteinii*), Daubenton's (*M. daubentonii*), Natterer's (*M. nattereri*) and whiskered (*M. mystacinus*); (3) *Nyctalus*/*Eptesicus* comprising Leisler's (*Nyctalus leisleri*), common noctule (*N. noctule*) and serotine (*Eptesicus seotinus*); (4) *Pipistrellus* comprising common pipistrelle (*Pipistrellus pipistrellus*), Nathusius' pipistrelle (*P. nathusii*) and soprano pipistrelle (*P. pygmaeus*); (5) *Plecotus* comprising brown long-eared (*Plecotus auratus*) and grey long-eared (*P. austriacus*); and (6) *Rhinolophus* comprising greater horseshoe (*Rhinolophus ferrumequinum*) and lesser horseshoe (*R. hipposideros*). These groups were determined by genus in all cases except for the combined *Nyctalus*/*Eptesicus* grouping, which was necessitated by BatClassify grouping these 'big bat' species together. Finally, we considered species-specific disagreement within each classifier combination (Kaleidoscope versus BatClassify for Anabat; Kaleidoscope vs. Acoustic Pipeline for AudioMoth) diagrammatically.

RESULTS

From a total of 211,145 Anabat recordings, Kaleidoscope Pro initially classified 40,367 (19%) as containing bat calls that could be identified with a confidence value ≥ 50%, and BatClassify initially classified 68,257 (32%). Of these, there were 13,582 recordings that were classified with ≥ 50% confidence by both Kaleidoscope and BatClassify (matched data). From a total of 258,731 AudioMoth recordings, Kaleidoscope Pro initially classified 39,701 (15%) as containing bat calls that could be identified with confidence value ≥ 50%, and the BTO Acoustic Pipeline initially classified 84,357 (33%). Of these, there were 30,574 recordings that were classified with ≥ 50% confidence by both Kaleidoscope and the Acoustic Pipeline (matched data). The Acoustic Pipeline and BatClassify classified a higher percentage of the recordings in their respective datasets with a confidence value ≥ 50% than Kaleidoscope; this trend occurred in all habitats (Table 2).

Species in the genus *Rhinolophus* were the most confidently classified species (lesser horseshoe for Kaleidoscope and BatClassify, greater horseshoe for the Acoustic Pipeline). Conversely, species in the genus *Myotis* tended to be classified with the least confidence, with the Acoustic Pipeline not classifying any recordings with ≥ 50% confidence for Alcathoe and Bechstein's (Fig. 2).

Agreement between Kaleidoscope and BatClassify for Anabat recordings

The optimal CLM predicting the level of agreement between these two classifiers on a recording-by-recording basis was significant overall (χ^2 = 6,465.3, *d.f.* = 23,

Table 2. Number and percentage of Anabat and AudioMoth recordings where bat calls were identified with a ≥ 50% confidence value, according to habitat and classification method.

Habitat	Anabat					AudioMoth				
	Total recordings	Kaleidoscope Pro		BatClassify		Total recordings	Kaleidoscope Pro		BTO Acoustic Pipeline	
		n	%	*n*	%		*n*	%	*n*	%
Riparian	63,083	20,402	32.34	26,060	41.31	94,438	24,759	26.22	50,142	53.10
Woodland	28,603	6,477	22.65	15,966	55.82	43,387	4,625	10.66	14,199	32.73
Wood pasture	40,027	10,779	26.93	17,224	43.03	51,001	7,521	14.75	14,279	28.00
Arable	79,432	2,709	3.41	9,007	11.34	69,905	2,796	4.00	5,737	8.21
Total	211,145	40,367	19.12	68,257	32.33	258,731	39,701	15.35	84,357	32.60

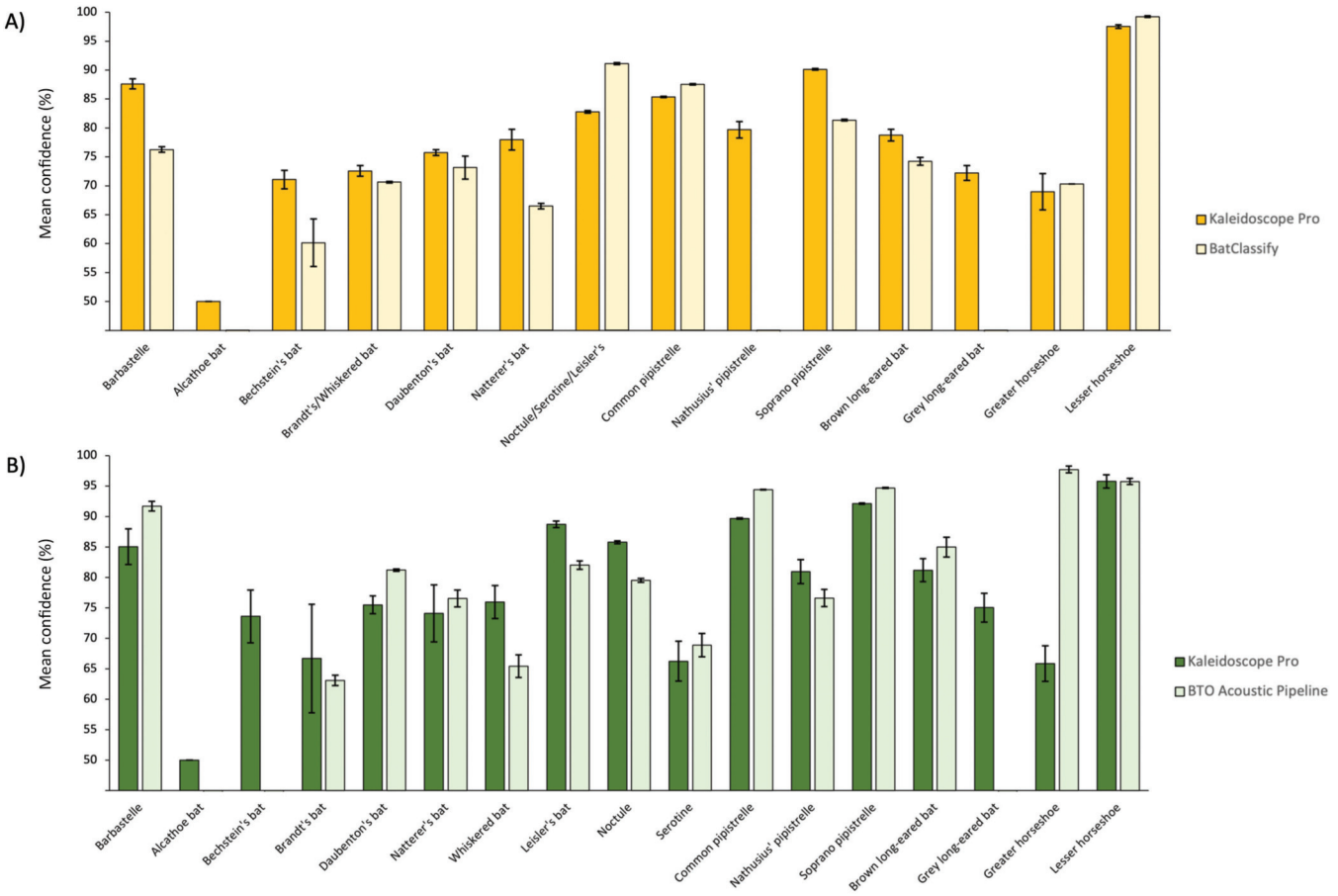

Figure 2. Mean confidence values ≥ 50% for each species/species group assigned by each classifier, for (A) the Anabat and (B) AudioMoth datasets. Error bars show SEM.

$p < 0.001$) and included habitat, taxonomic group, and the interaction term as predictors, all of which were individually statistically significant (habitat: $\chi^2 = 8{,}514.4$, *d.f.* = 3, $p < 0.001$, taxonomic group: $\chi^2 = 1{,}2896$, *d.f.* = 5, $p < 0.001$, habitat x taxonomic group interaction: $\chi^2 = 2{,}760.1$, *d.f.* = 15, $p < 0.001$). The model was a substantially better fit (AIC = 64,513) than the null model (AIC = 70,933).

Habitat type had a relatively small but significant effect on agreement between classifiers. There was basic agreement (ordinal scores 2–6; Table 1) for 92% of all recordings in riparian habitat (lowest) compared to 99% of recordings in wood pasture (highest). We observed more substantial variation when taxonomic group was considered, with the classifiers showing agreement for > 90% of recordings for *Barbastella*, *Eptesicus*/*Nyctalus*, *Pipistrellus* or *Rhinolophus* groups versus < 10% for *Myotis*. When considering the impacts of both habitat and taxonomic group on classifier agreement using the probabilities of each level of agreement for each habitat/taxa combination (Fig. 3), levels of agreement were uniformly high (ordinal scores 4–6) across all habitats for *Nyctalus*/*Eptesicus*, *Pipistrellus* and *Rhinolophus* recordings. This contrasted with *Myotis* recordings where classifier disagreement (ordinal score 1 — Table 1) was the most likely outcome in all habitats, ranging from a probability of 0.61 in woodland to 0.96 in riparian (Fig. 3). For *Barbastella* and *Plecotus* recordings, the level of agreement was more variable between habitats (i.e., these taxonomic groups drove the significant interaction between habitat and taxonomic group in the CLM), being higher in wood pasture and woodland than in the arable and riparian habitats (Fig. 3).

Agreement between Kaleidoscope and Acoustic Pipeline for AudioMoth recordings

The optimal CLM predicting the level of agreement between the classifiers on a recording-by-recording basis was significant overall ($\chi^2 = 3{,}877.1$, *d.f.* = 15, $p < 0.001$) and included habitat, taxonomic group, and the interaction term as predictors, all of which were individually statistically significant (habitat: $\chi^2 = 24451.4$, *d.f.* = 3, $p < 0.001$, taxonomic group: $\chi^2 = 32{,}677.8$, *d.f.* = 3, $p < 0.001$, habitat x taxonomic group interaction: $\chi^2 = 3{,}261.5$, *d.f.* = 9, $p < 0.001$). In this dataset, we omitted recordings involving *Barbastella* or *Plecotus* classifications, as these occurred too infrequently for meaningful

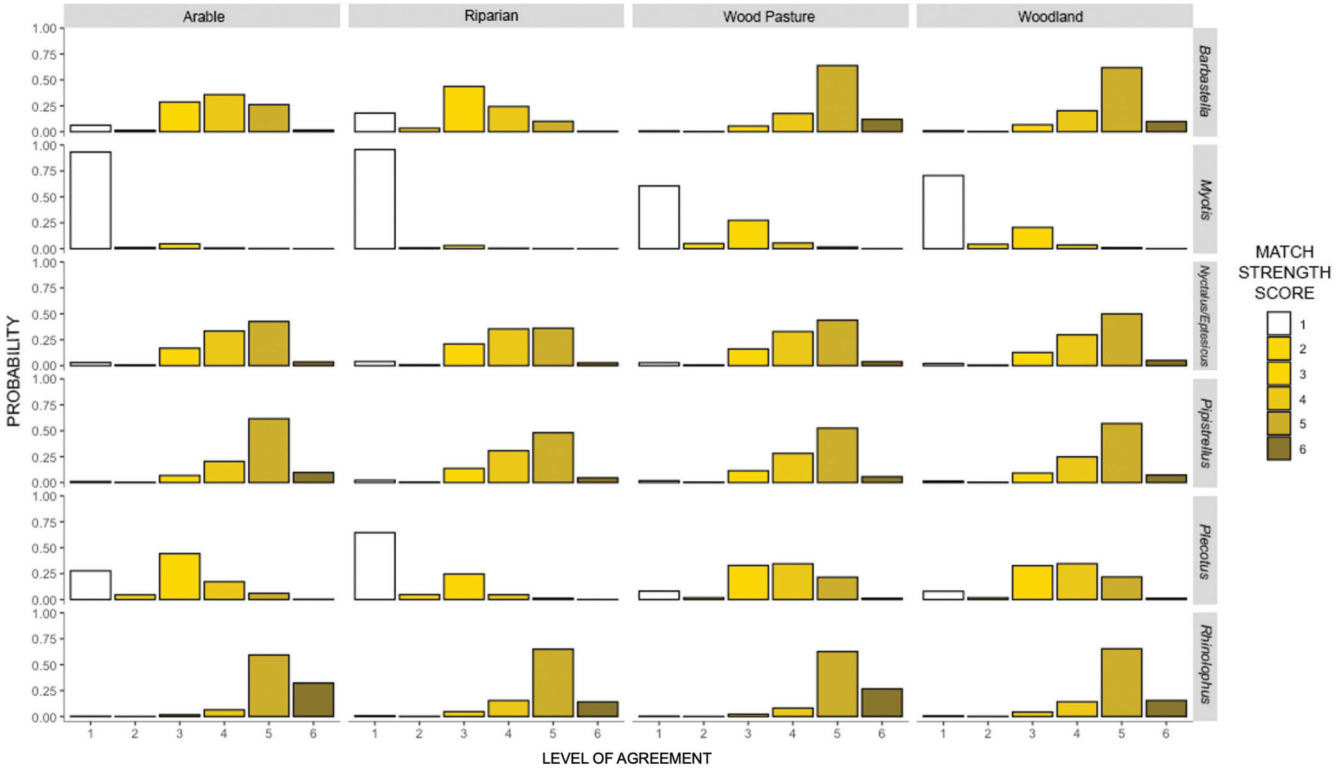

Figure 3. Model-predicted probabilities of obtaining each match-strength score (with a higher score indicating a stronger match), for each interaction between habitat and taxonomic group, using the Anabat Swift dataset (see Table 1 for definitions of match strength scores).

statistical analysis (Supplementary Table S1). Thus, there were four levels in the taxonomic group fixed factor rather than six. The model was a substantially better fit (AIC = 175,626) than the null model (AIC = 179,473).

The effect of habitat type on agreement between classifiers was relatively small but significant. There was basic agreement (ordinal scores 2–6 — Table 1) for 90% of recordings in riparian habitat (lowest) compared to 93% of recordings in arable habitat (highest). However, taxonomic group had a more substantial effect. There was basic agreement between classifiers > 90% of *Pipistrellus* recordings, and ca. 75% of *Eptesicus*/*Nyctalus* and *Rhinolophus* recordings but this dropped to below 50% for recordings of *Myotis*. When considering impacts of both habitat and taxonomic group using the probabilities of each level of agreement for each habitat/taxa combination (Fig. 4), classifier agreement was uniformly high (ordinal scores 4–6) for *Pipistrellus* across all habitats and for *Rhinolophus* in arable, wood pasture and woodland habitats. Conversely, classifier disagreement (ordinal score 1 — Table 1) was the most likely outcome in all habitats for *Myotis*, peaking at a probability of 0.66 in woodland (Fig. 4). Classifier disagreement was also high for *Nyctalus*/*Eptesicus* in woodland (0.47), and for *Rhinolophus* in riparian habitat (0.42) (Fig. 4); these differences likely drove the significant interaction between habitat and taxonomic group in the CLM.

Species-specific disagreement

For the Anabat recordings, species-specific disagreement between BatClassify and Kaleidoscope most usually occurred for species in the genus *Myotis*. Indeed, recordings being classified as Brandt's/Whiskered by BatClassify but as Daubenton's by Kaleidoscope accounted for 50% of all instances of disagreement in this dataset (Fig. 5A). Conversely, for the AudioMoth dataset, classifier disagreement was predominantly driven by recordings being classified as common pipistrelle by the Acoustic Pipeline but as a soprano pipistrelle by Kaleidoscope, which accounted for 53% of all instances of disagreement in this dataset (Fig. 5B). Additionally, there were cross-genera disagreements in the AudioMoth data whereby recordings classified as either common or soprano pipistrelle by the Acoustic Pipeline were classified as noctule by Kaleidoscope (12% and 17% of all instances of disagreement, respectively).

DISCUSSION

There were similarities in the influence of habitat and taxonomic group on level of agreement between classifiers in both datasets. Moreover, the species most frequently confused in instances of disagreement were predominately within the same genus, but there was

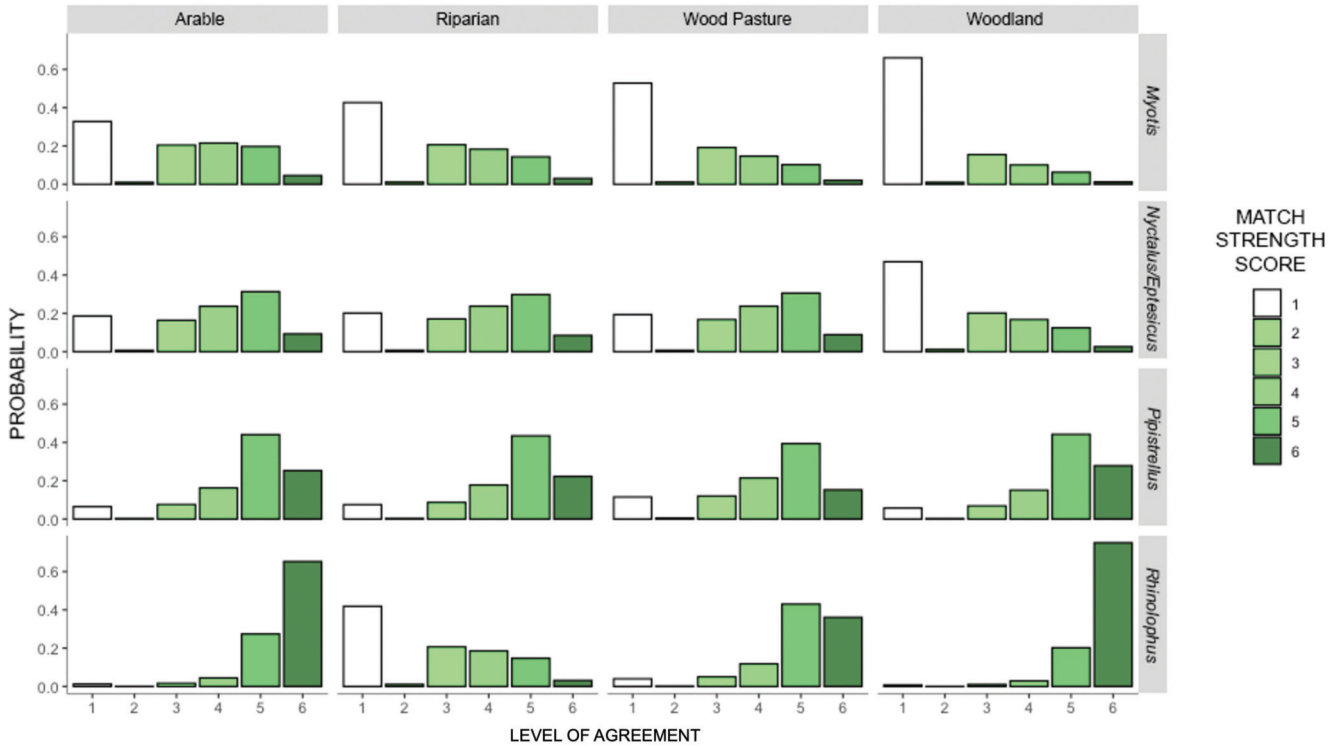

Figure 4. Model-predicted probabilities of obtaining each match-strength score (with a higher score indicating a stronger match), for each interaction between habitat and taxonomic group, using the AudioMoth dataset (see Table 1 for definitions of match-strength scores).

between-genera confusion between pipistrelles and noctule between Kaleidoscope and the Acoustic Pipeline. However, there were notable differences in the overall performance of the classifiers. Kaleidoscope Pro classified fewer recordings above the 50% confidence threshold than BatClassify (for Anabat recordings) or the BTO Acoustic Pipeline (for AudioMoth recordings).

The metrics and thresholds used to define confidence vary between classifiers (Lemen *et al.* 2015). The threshold adopted in any study is, to an extent, arbitrary and might be species- and context-dependent (Barre *et al.* 2019). To minimise false positives and provide a generic threshold for ease and consistency of workflow, the BTO Acoustic Pipeline (British Trust for Ornithology 2024) has adopted a confidence value of 0.5 (50%) as the recommended threshold below which recordings should be discarded. This threshold was used throughout the current study, but it was notable that the proportion of recordings meeting this threshold varied substantially based on original detector (commercial Anabat Swift versus open-source AudioMoth) and classifier (Kaleidoscope vs. BatClassify, or Kaleidoscope vs. Acoustic Pipeline).

The effect of habitat was significant statistically when predicting classifier agreement for both Anabat and AudioMoth recordings, but relatively low in terms of effect size. The best performing and least performing habitats differed by seven percentage points for Anabat data and four percentage points for AudioMoth data. Within the Anabat dataset, there was higher agreement between classifiers in woodland and wood pasture habitats compared to arable and riparian habitats. The principal drawback of conducting pairwise analysis on unverified field recordings, is that it is impossible to determine if one classifier is preforming more accurately than the other (Lemen *et al.* 2015); the focus of analysis is agreement/disagreement rather than correct/incorrect. However, BatClassify was originally developed for a Department of Food and Rural Affairs (DEFRA) commissioned project which aimed to devise robust and efficient protocols for monitoring woodland bat species, which are threatened by ongoing habitat loss and degradation (Scott and Altringham 2014*b*; Titley Scientific 2024). Thus, it is likely optimised for classifying recordings in these types of habitats, resulting in stronger classifier matches. Within the AudioMoth dataset, levels of agreement were highest for the arable habitat, which was the most homogeneous and open of the four habitats surveyed. This lack of habitat complexity and clutter is likely to increase detectability and improve the overall quality of the recordings, which is especially necessary for AudioMoths given the lower quality of the microphone in these open-source inexpensive units compared to those used in commercial detectors such as Anabat devices (Mac Aodha *et al.* 2018; Perks *et al.* 2025).

Taxonomic group had a significant and substantial effect on the level of agreement between classifiers. Species in the *Myotis* genus were least confidently classified by all classifiers, which is likely driven by similarities in call parameters and peak frequency overlaps. Calls attributed to the genus *Myotis* are notoriously challenging to classify to species level, even by skilled practitioners

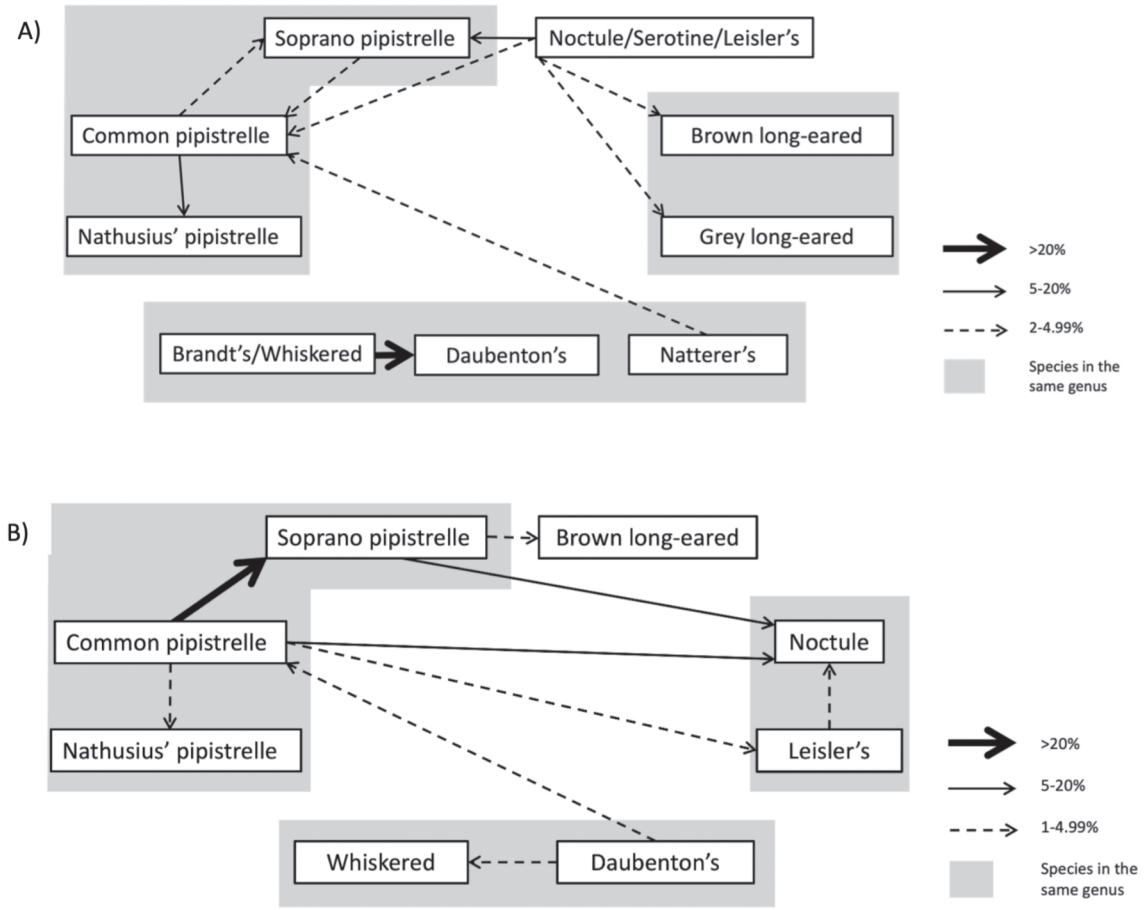

Figure 5. Species classifications most commonly involved in a lack of consensus between classifiers in analysing (A) Anabat Swift recordings (origin of arrow = BatClassify classification, tip of arrow = Kaleidoscope classification) and (B) AudioMoth recordings (origin of arrow = BTO Acoustic Pipeline classification, tip of arrow = Kaleidoscope classification).

(Vaughan *et al.* 1997; Russ 2012; Gibb *et al.* 2018), and the findings presented indicate that this is also the case for automated analysis. In the Anabat dataset, disagreement was largely driven by recordings classified as Brandt's/Whiskered by BatClassify being classified as Daubenton's by Kaleidoscope, likely due to interspecific overlap in call parameters (Rydell *et al.* 2017; Gibb *et al.* 2018). Moreover, Kaleidoscope classified some recordings in the dataset as Alcathoe, a nationally rare *Myotis* species within the UK, which was never classified by BatClassify. Although such occurrences were infrequent, the occurrence of classifications for species outside their expected range warrants consideration. Whilst these cannot be formally considered misclassifications without manual auditing, geographically implausible classifications should be interpreted carefully. These findings are consistent with those of previous studies (Rydell *et al.* 2017; Thomas and Davison 2022) which have reported limitations in the use of Kaleidoscope and/or BatClassify in identifying *Myotis* calls to species level, urging caution. Conversely, recordings classified in the *Rhinolophus* and *Pipistrellus* taxonomic groups were associated with the highest levels of agreement between classifiers in both datasets. *Rhinolophus* calls have distinctive call shapes and high frequencies, and this likely drove accurate classifications by all classifiers despite calls being short range and prone to attenuation, while *Pipistrellus* calls also have parameters are unlikely to overlap with other taxonomic groups (Russ 2012). However, within the AudioMoth dataset, confusion within the *Pipistrellus* genus (i.e., between common and soprano pipistrelle) accounted for over half of all instances of disagreement. Although calls from the two species are typically distinguished by differences in peak frequency, the plasticity of *Pipistrellus* calls can vary substantially depending on environmental or behavioural factors (Montauban *et al.* 2021). This can create considerable overlap in characteristic call frequencies, which likely caused classifier disagreement.

The findings presented here further highlight the need for automated classifiers to be used with caution, and if possible, in conjunction with manual auditing by skilled practitioners (Russo and Voigt 2016; Rydell *et al.* 2017; Brabant *et al.* 2018; Solick *et al.* 2024). Despite improvements in classifier accuracy in recent years with the expansion of reference call libraries, the pairwise disagreement between widely used classifiers on the same recordings shown here, highlights that there is still considerable potential for misidentification. Although adopting a confidence threshold into analysis workflows

can reduce false positives, these cannot currently be eliminated. Therefore, where the impacts of identification errors are substantial, for example when establishing the presence/absence of species for impact assessments (Barre *et al.* 2019), manual auditing is still necessary.

Automated classification of bat calls can improve the efficiency of analysis workflows, saving substantial resources. However, standardising workflows remains a major challenge. Complexities include the continually expanding array of classifiers that are available, and variation in the quality of the recordings produced by different types of acoustic recorders (Brinkløv *et al.* 2023). As highlighted here, different classifiers use a range of classification techniques from tree-based algorithms to machine learning. Thus, the results parameters also differ, including the confidence metric reported, and crucially whether multiple species can be classified within a file. Moreover, with the increasingly wider use of open-source acoustic recorders such as AudioMoths, consideration also needs to be given to the relative quality of recordings obtained by different detectors. The call libraries used to train classifiers typically contain example calls of known species, and there may be a tendency to select high quality calls for inclusion in the library (Lemen *et al.* 2015). These may not relate well to recordings from complex habitats or lower quality microphones, resulting in misidentification.

SUPPORTING INFORMATION

Contents: Supplementary Table S1. Numbers of recordings classified into each taxonomic group by each classifier in each dataset. Supplementary Information is available exclusively on BioOne.

AUTHOR CONTRIBUTIONS

SJP: research concept and design, collection and/or assembly of data, data analysis and interpretation, writing, and final approval of the article; AEG: research concept and design, critical revision, and final approval of the article; MJOC: critical revision, and final approval of the article.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

ACKNOWLEDGEMENTS

The authors wish to thank the landowners who kindly granted access permissions for data collection.

REFERENCES

Adams MD, Law BS, Gibson MS. 2010. Reliable automation of bat call identification for eastern New South Wales, Australia, using classification trees and AnaScheme software. Acta Chiropterologica, 12: 231–245.

Barré K, Le Viol I, Julliard R, Pauwels J, Newson SE, Julien JF, Claireau F, Kerbiriou C, Bas Y. 2019. Accounting for automated identification errors in acoustic surveys. Methods in Ecology and Evolution, 10: 1171–1188.

Bat Conservation Trust. 2024. British Bat Survey (BBatS). Available at: https://www.bats.org.uk/our-work/science-research/passive-acoustic-surveys/british-bat-survey.Accessed: 15 August 2021.

Brabant R, Laurent Y, Dolap U, Degraer S, Poerink BJ. 2018. Comparing the results of four widely used automated bat identification software programs to identify nine bat species in coastal Western Europe. Belgian Journal of Zoology, 148: 119-128.

Braun de Torrez EC, Ober HK, McCleery RA. 2017. Critically imperiled forest fragment supports bat diversity and activity within a subtropical grassland. Journal of Mammalogy, 99: 273–282.

Brinkløv SM, Macaulay J, Bergler C, Tougaard J, Beedholm K, Elmeros M, Madsen PT. 2023. Open-source workflow approaches to passive acoustic monitoring of bats. Methods in Ecology and Evolution, 14: 1747–1763.

British Trust for Ornithology. 2025. Ultrasonic classifiers. Available at: https://www.bto.org/data/tools-products/acoustic-pipeline/ultrasonic-classifiers. Accessed: 3 November 2025.

British Trust for Ornithology. 2024. BTO Acoustic Pipeline User Manual. Available at: https://www.bto.org/sites/default/files/bto_acoustic_pipeline_user_manual_0.pdf. Accessed: 28 June 2024.

Browning E, Gibb R, Glover-Kapfer P, Jones KE. 2017. Passive acoustic monitoring in ecology and conservation. WWF UK, Woking, UK.

Burnham KP, Anderson DR. 2002. Multimodel avoiding pitfalls when using information-theoretic methods. Journal of Wildlife Management, 66: 912–918.

Christensen R. 2023. Ordinal — regression models for ordinal data. R package version 2023.12-4.1, https://CRAN.R-project.org/package=ordinal.

Collins J (ed.). 2023. Bat surveys for professional ecologists: good practice guidelines, 4th edition. Bat Conservation Trust, London.

Gibb R, Browning E, Glover-Kapfer P, Jones KE. 2018. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods in Ecology and Evolution, 10: 169–185.

Goodwin KR, Gillam EH. 2021. Testing accuracy and agreement among multiple versions of automated bat call classification software. Wildlife Society Bulletin, 45: 690–705.

Kunberger JM, Long AM. 2023. A comparison of bat calls recorded by two acoustic monitors. Journal of Fish and Wildlife Management, 14: 171–178.

Lemen C, Freeman PW, White JA, Andersen BR. 2015. The problem of low agreement among automated identification programs for acoustical surveys of bats. Western North American Naturalist, 75: 218–225.

López-Baucells A, Torrent L, Rocha R, Bobrowiec PE, Palmeirim

JM, Meyer CF. 2019. Stronger together: combining automated classifiers with manual post-validation optimizes the workload vs reliability trade-off of species identification in bat acoustic surveys. Ecological Informatics, 49: 45–53.

López-Bosch D, Rocha R, López-Baucells A, Wang Y, Si X, Ding P, Gibson L, Palmeirim AF. 2022. Passive acoustic monitoring reveals the role of habitat affinity in sensitivity of sub-tropical East Asian bats to fragmentation. Remote Sensing in Ecology and Conservation, 8: 208–221.

Mac Aodha O, Gibb R, Barlow KE, Browning E, Firman M, Freeman R, Harder B, Kinsey L, Mead GR, Newson SE, Pandourski I. 2018. Bat detective — Deep learning tools for bat acoustic signal detection. PLoS Computational Biology, 14: e1005995.

Montauban C, Mas M, Tuneu-Corral C, Wangensteen OS, Budinski I, Martí-Carreras J, Flaquer C, Puig-Montserrat X, López-Baucells A. 2021. Bat echolocation plasticity in allopatry: a call for caution in acoustic identification of *Pipistrellus* sp. Behavioral Ecology and Sociobiology, 75: 1–15.

Nocera T, Ford WM, Silvis A, Dobony CA. 2019. Let's agree to disagree: Comparing auto-acoustic identification programs for northeastern bats. Journal of Fish and Wildlife Management, 10: 346–361.

Perks, SJ, O'Connell, MJ, Goodenough, AE. 2025. Comparing Passive Acoustic Monitoring data from commercial and open-source detectors- evidence to support best practice. Ecological Solutions and Evidence. 6: e70103.

Russ J. 2012. British Bat Calls: A Guide to Species Identification. Pelagic Publishing: Exeter.

Russo D, Voigt CC. 2016. The use of automated identification of bat echolocation calls in acoustic monitoring: A cautionary note for sound analysis. Ecological Indicators, 66: 598–602.

Rydell J, Nyman S, Eklöf J, Jones G, Russo D. 2017. Testing the performances of automated identification of bat echolocation calls: A request for prudence. Ecological Indicators, 78: 416–420.

Scott C, Altringham J. 2014*a*. Bat classify software. Available at: https://bitbucket.org/chrisscott/batclassify/downloads. Accessed: 9 August 2024.

Scott C, Altringham J. 2014*b*. Developing effective methods for the systematic surveillance of bats in woodland habitats in the UK. Final Report to DEFRA. EC1015. University of Leeds, Leeds, UK.

Smith TN, Furnas BJ, Nelson MD, Barton DC, Clucas B. 2021. Insectivorous bat occupancy is mediated by drought and agricultural land use in a highly modified ecoregion. Diversity and Distributions, 27: 1152–1165.

Solick DI, Hopp BH, Chenger J, Newman CM. 2024. Automated echolocation classifiers vary in accuracy for northeastern US bat species. PLoS ONE, 19: e0300664.

Springall BT, Li H, Kalcounis-Rueppell MC. 2019. The in-flight social calls of insectivorous bats: species specific behaviors and contexts of social call production. Frontiers in Ecology and Evolution, 7: 441.

Starbuck CA, DeSchepper LM, Hoggatt ML, O'Keefe JM. 2024. Tradeoffs in sound quality and cost for passive acoustic devices. Bioacoustics, 33: 58–73.

Stowell D. 2022. Computational bioacoustics with deep learning: a review and roadmap. PeerJ, 10: e13152.

Sugai LSM, Silva TSF, Ribeiro Jr JW, Llusia D. 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience, 69: 15–25.

Taillie PJ, de Torrez EB, Potash AD, Boone IV WW, Jones M, Wallrichs MA, Schellenberg F, Hooker K, Ober HK, McCleery RA. 2021. Bat activity response to fire regime depends on species, vegetation conditions, and behavior. Forest Ecology and Management, 502: 119722.

Teixeira D, Maron M, van Rensburg BJ. 2019. Bioacoustic monitoring of animal vocal behavior for conservation. Conservation Science and Practice, 1: e72.

Thomas RJ, Davison SP. 2022. Seasonal swarming behavior of *Myotis* bats revealed by integrated monitoring, involving passive acoustic monitoring with automated analysis, trapping, and video monitoring. Ecology and Evolution, 12: e9344.

Titley Scientific. 2024. FAQs. Available at: https://www.titley-scientific.com/support/faqs/. Accessed: 10 July 2024.

Vaughan N, Jones G, Harris S. 1997. Identification of British bat species by multivariate analysis of echolocation call parameters. Bioacoustics, 7: 189–207.

Wrege PH, Rowland ED, Keen S, Shiu Y. 2017. Acoustic monitoring for conservation in tropical forests: examples from forest elephants. Methods in Ecology and Evolution, 8: 1292–1301.

Zamora-Gutierrez V, MacSwiney G MC, Martínez Balvanera S, Robredo Esquivelzeta E. 2021. The evolution of acoustic methods for the study of bats. Pp. 43–59, in 50 Years of Bat Research: Foundations and New Frontiers (Lim BK, Fenton MB, Brigham RM, Mistry S, Kurta A, Gillam EH, Russell A, Ortega J, eds.). Springer International Publishing, Berlin.