An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (Poephila cincta cincta)

Author

van Osta, John M, Dreis, Brad, Meyer, Ed, Grogan, Laura F, Castley, J Guy

Published

2023

Journal Title

Ecological Informatics

Version

Version of Record (VoR)

DOI

10.1016/j.ecoinf.2023.102233

An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the southern black-throated finch (*Poephila cincta cincta*)

Authors

John M. van Osta[ab*]

Brad Dreis[ab]

Ed Meyer[b]

Laura F. Grogan[a]

J. Guy Castley[a]

[a] Centre for Planetary Health and Food Security, Griffith University, Gold Coast Campus, Queensland, 4222, Australia

[b] E2M Pty Ltd, 19 Lang Parade, Milton, Queensland, 4064, Australia

* Corresponding author: john.vanosta@griffithuni.edu.au

Abstract

The application of machine learning methods has led to major advances in the development of automated recognisers used to analyse bioacoustics data. To further improve the performance of automated call recognisers, we investigated the development of efficient data annotation strategies and how best to address uncertainty around ambiguous vocalisations. These challenges present a particular problem for species whose vocalisations are rare in field recordings, where collecting enough training data can be problematic and a species' vocalisations may be poorly documented.

We provide an open access solution to address these challenges using two strategies. First, we applied an active learning framework to iteratively improve a convolutional neural network (CNN) model able to automate call identification for a target rare bird species, the southern black-throated finch (*Poephila cincta cincta*). We collected 9,093 hours of unlabelled audio recordings from a field study in the Desert Uplands Bioregion of Queensland, Australia, and used active learning to prioritise human annotation effort towards data that would best improve model fit. Second, we progressed methods for managing ambiguous vocalisations by applying machine learning methods more commonly used in medical image analysis and natural language processing. Specifically, we assessed agreement among human annotators and the CNN model (i.e. inter-annotator agreement) and used this to determine realistic performance outcomes for the CNN model and to identify areas where inter-annotator agreement may be improved. We also applied a classification approach that allowed the CNN model to classify sounds into an 'uncertain' category, which replicated a requirement of human-annotation and facilitated the comparison of human-model annotation performance.

We found that active learning was an efficient strategy to build a CNN model where there was limited labelled training data available, and target calls were extremely rare in the unlabelled data. As few as five active learning iterations, generating a final labelled dataset of 1,073 target calls and 5,786 non-target sounds, were required to train a model to identify the target species with comparable performance to experts in the field.

Assessment of inter-annotator agreement identified a bias in our model to align predictions most closely with those of the primary annotator and identified significant differences in inter-annotator agreement among subsets of our acoustic data. Our results highlight the use of inter-annotator agreement to understand model performance and identify areas for improvement in data annotation. We also show that excluding ambiguous vocalisations during data annotation results in an overestimation of model performance, an important consideration for datasets with inter-annotator disagreement.

Keywords

Bioacoustics; Machine learning; Annotator agreement; Call recognition; Active learning

1 Introduction

The use of low-cost acoustic sensors in ecological research has rapidly expanded, enabling new approaches to the detection and monitoring of species (Hervás *et al.*, 2017; Sugai *et al.*, 2019; Ross *et al.*, 2023). The increased adoption of acoustic sensors has led to an exponential increase in the volume of data produced, thereby positioning bioacoustics within the domain of 'big data' (Sugai *et al.*, 2019).

Accurate automated recognition of bird calls in real-world acoustic data remains a key challenge limiting practical applications of passive acoustic monitoring (Priyadarshani, Marsland and Castro, 2018; Stowell, 2022). Acoustic analyses of bird species and communities typically rely on the development of individual species call recognisers, particularly for threatened species (LeBien *et al.*, 2020; Ruff *et al.*, 2021; Teixeira *et al.*, 2022), or use soundscape indices to draw comparisons among sites (Colonna, Carvalho and Rosso, 2020; Dema *et al.*, 2020). Classification of bird species from real-world environmental data has proven challenging due to unavoidable environmental noise overlap with target bird calls, attenuation of bird calls at distance, overlapping calls, and intra-specific call variation (Priyadarshani, Marsland and Castro, 2018), but also the lack of sufficient training data to build species-specific recognisers (Gibb *et al.*, 2019).

Deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks, have proven highly effective in the classification of individual bird species in complex acoustic data (Stowell, 2022) and have been used in a variety of automated call recognition studies (Fairbrass *et al.*, 2019; Stowell *et al.*, 2019; Allen *et al.*, 2021; Kahl *et al.*, 2021; Ruff *et al.*, 2021). Deep learning methods have outperformed alternative detection and classification methods for the identification of specific bird species provided sufficient training data are available to avoid model overfitting, which is a common constraint of deep learning (Gibb *et al.*, 2019).

While the application of machine learning (ML) methods to detect wildlife calls is a well-established and active field of research (Stowell *et al.*, 2019), the majority of existing studies have used large, labelled datasets to train classifiers using supervised learning approaches. Examples include the use of publicly available wildlife call datasets (Sankupellay and Konovalov, 2018; Stowell *et al.*, 2019), annotation by citizen scientists (Mac Aodha *et al.*, 2018) or existing labelled datasets (Towsey *et al.*, 2012; Bergler *et al.*, 2019). However, these approaches may not be suitable for the detection of a species where labelled training data are not available and are expensive to obtain due to the cost of human-generated labels. The expense of creating labelled datasets limits the ability to train state-of-the-art deep learning models (Van Engelen and Hoos, 2020), and provides a barrier to developing recognisers for rare or cryptic species, as well as more common species that vocalise infrequently.

Rare species are more sensitive to human-induced impacts such as habitat loss and climate change than common species and have a higher risk of extinction (Davies, Margules and Lawrence, 2004; Sekercioglu *et al.*, 2008; Loiseau *et al.*, 2020). Environmental legislation and policies have broadly struggled to reverse the decline of rare species, in part due to the difficulty gaining knowledge on these species' distributions, habitat use and ecological interactions (Bland *et al.*, 2015). At the problem's most severe, data deficient species may not be recognised in threatened species lists that drive conservation actions, such as International Union for Conservation of Nature's red list (Howard and Bickford, 2014) and the Australian Government's *Environment and Biodiversity Protection Act 1999* threat categories (Braby, 2018). Furthermore, when these species are recognised, the data to guide conservation actions may not be available (Howard and Bickford, 2014; Braby, 2018). Extending bioacoustic methods to improve the application for rare and difficult to survey species may be a powerful tool to limit data deficiencies and thus improve species' conservation outcomes (Wood *et al.*, 2023).

One such method that has the potential to improve the applicability of bioacoustics to rare and difficult to survey species is active learning. Active learning is a ML paradigm that incorporates a human-machine interface to optimise the annotation of training data (Van Engelen and Hoos, 2020; Monarch, 2021). In a recent review and roadmap of computational bioacoustics, Stowell (2022) identified the development of active learning techniques as highly important to the future of bioacoustics. Active learning is particularly suitable for applications where limited labelled data exist and the cost of obtaining a large labelled dataset is high (Van Engelen and Hoos, 2020). For example, labelled data may be unavailable due to the high cost of human-annotation (Zhu and Goldberg, 2009), or due to the rarity of an event/record making the target signal difficult to find in unlabelled data, thus reducing sample size or increasing the cost of human-annotation (Shi *et al.*, 2019).

The collection of labels for bioacoustics research is often conducted through manual annotation by experts and trained annotators with expertise in the species of interest (Stowell, 2022). The goal is to achieve near 100% accuracy in the annotation of target calls. However, in practice, even highly-trained and experienced experts can have difficulty achieving high annotation reliability from real-world recordings. In two studies of auditory detection at avian point counts, both Simons *et al.* (2007) and Alldredge, Simons and Pollock (2007) found annotators' abilities to detect species' calls varied by 28% to 42% among varying treatments of ambient noise and distance. Furthermore, Mortimer and Greene (2017) found that agreement among annotators processing audio recordings of New Zealand avifauna ranged between 23.5% and 85.1%, depending on the species being annotated. All three studies identified that the distinctiveness of the bird species call compared to other calls in the environment was a key factor influencing the reliability of call detection and annotation.

Assessment of inter-annotator agreement, also known as inter-rater reliability, is widely recommended to determine the reliability of manual annotation (Reidsma and Carletta, 2008; Gwet, 2014; Duc *et al.*, 2021). Inter-annotator agreement is a measure of the consistency and agreement among multiple annotators in their annotation of a dataset (Gwet, 2014). While inter-annotator disagreement is a well-documented issue for bioacoustics data (Duc *et al.*, 2021), relatively few studies incorporate methods to account for imperfect inter-annotator agreement into the design and evaluation of resulting ML models. Instead, all annotations regardless of annotator or species are typically treated as certain, which Cabitza *et al.* (2020) and Campagner *et al.* (2021) describe as the 'elephant in the machine' in the context of medical ML applications. Cabitza *et al.* (2020) also argue that considering uncertainty in the annotation process is important to develop models that generalise to real-world scenarios.

We present here an open access, novel method to develop a call detection model for a threatened bird species, the southern black-throated finch (SBTF), *Poephila cincta cincta*, that lacks an existing labelled call dataset. Our method applies an active learning framework to prioritise scarce human annotation resources towards data that are the most likely to improve the model. Our work builds on the 'standard recipe' for bioacoustic classification described by Stowell (2022) and applies an active learning framework to iteratively improve the model. We also apply a novel method to include inter-annotator agreement and label uncertainty, which are rarely considered in bioacoustics, into the development and evaluation of call detection models. This work has broad application for many species that have limited existing training data available and/or for species whose calls may not be annotated with certainty.

2 Methods

2.1 Target species and call

The southern subspecies of the black-throated finch (SBTF), *Poephila cincta cincta*, is a grassfinch (family Estrildidae) endemic to north-eastern Australia typically inhabiting open grassy woodlands (Higgins, Peter and Cowling, 2006; Shephard, Pridham and Forshaw, 2012). The SBTF is listed as 'Endangered' under the Australian *Environment Protection and Biodiversity Conservation Act 1999* since its extent of occurrence declined by up to 59% in the ten years prior to 2005 (Threatened Species Scientific Committee (TSSC), 2005).

Black-throated finch have a complex vocal repertoire, with 12 calls described (Shephard, Pridham and Forshaw, 2012). Of these, the 'long'/'pew' call is the farthest carrying and most frequently heard call (Higgins, Peter and Cowling, 2006; Shephard, Pridham and Forshaw, 2012) and as such is the priority for automated detection. Other calls described for the species include a complex of short-range communication and alarm calls, which are heard less

frequently and are not as far carrying as the 'long'/'pew' call (Shephard, Pridham and Forshaw, 2012).

2.2 Audio data and study area

Recordings of the SBTF were made in in the Desert Uplands Bioregion of Queensland, Australia, near the southern limit of the species current distribution, from 25 August 2020 to 24 April 2021. Recordings were collected as part of an ongoing SBTF research program for Bravus Mining and Resources (Bravus) on a 75,000 ha section of the Moray Downs property, which includes areas managed as environmental offset by Bravus. The study area is largely vegetated, with remnant vegetation, comprising 24 discrete vegetation communities, occurring over 79% of the study area (Department of Environment and Science, 2022) (Figure 1). The study area contains a diverse assemblage of woodland birds, with a total of 119 bird species recorded in the SBTF habitat within the study area (Author, unpublished data, 2023). Among the local avifauna are other Estrildid finches, such as the double-barred finch (*Taeniopygia bichenovii*), plum-headed finch (*Neochmia modesta*), and zebra finch (*Taeniopygia castanotis*). Of these, the double-barred finch has a call that is most similar to the target SBTF call.

Two types of acoustic recording unit were used within this study. Solar-powered bioacoustic recorders (BARs; Frontier Labs; Roe *et al.*, 2021) were installed within known and suspected SBTF habitat for periods of 6 to 12 months per unit. Solar-powered BARs are suited to long-term deployments, with a weight of 4 kg, no battery limitations, 4 SD card slots and a high-fidelity microphone (signal-to-noise ratio [SNR] = 80 dB; Open Acoustic Devices, n.d.). We also installed Audiomoth recorders (Hill *et al.*, 2019) near known and suspected SBTF nests for a duration of 2 to 4 weeks per unit (Figure 1). Audiomoths are suited to short-term and opportunistic deployments due to their light-weight (80 g, including 3 AAA batteries) and low-cost (Hill *et al.*, 2019). Recording quality is comparatively lower on Audiomoths compared to the solar-powered BARs, with a SNR of 44.2 dB (Hill *et al.*, 2019).

A total of 9,098 hours of audio data, comprising 8,266 hours from solar-powered bioacoustic recorders from nine independent sites, and 832 hours from Audiomoth recorders from ten known and suspected nest sites, were used for this study. Audio data were separated into training/validation (Section 2.3) and test datasets (Section 2.4) prior to the commencement of CNN development. The training/validation dataset comprised 6,363 hours of audio from ten independent sites (five solar-powered bioacoustic recorders and five Audiomoth recorders) and the test dataset comprised 2,735 hours of audio from nine independent sites (four solar-powered bioacoustic recorders and five Audiomoth recorders) that were different from the training/validation dataset. All audio data were unlabelled at the commencement of the study.

All recorders were set to record 10-minute files continuously between 6 am and 6 pm daily, which reflected the diurnal activity patterns of SBTF. We stored audio data as FLAC files at a sampling rate of 44.1 kHz for the solar-powered bioacoustic recorders and WAV files at a 32 kHz sampling rate for the Audiomoth recorders.

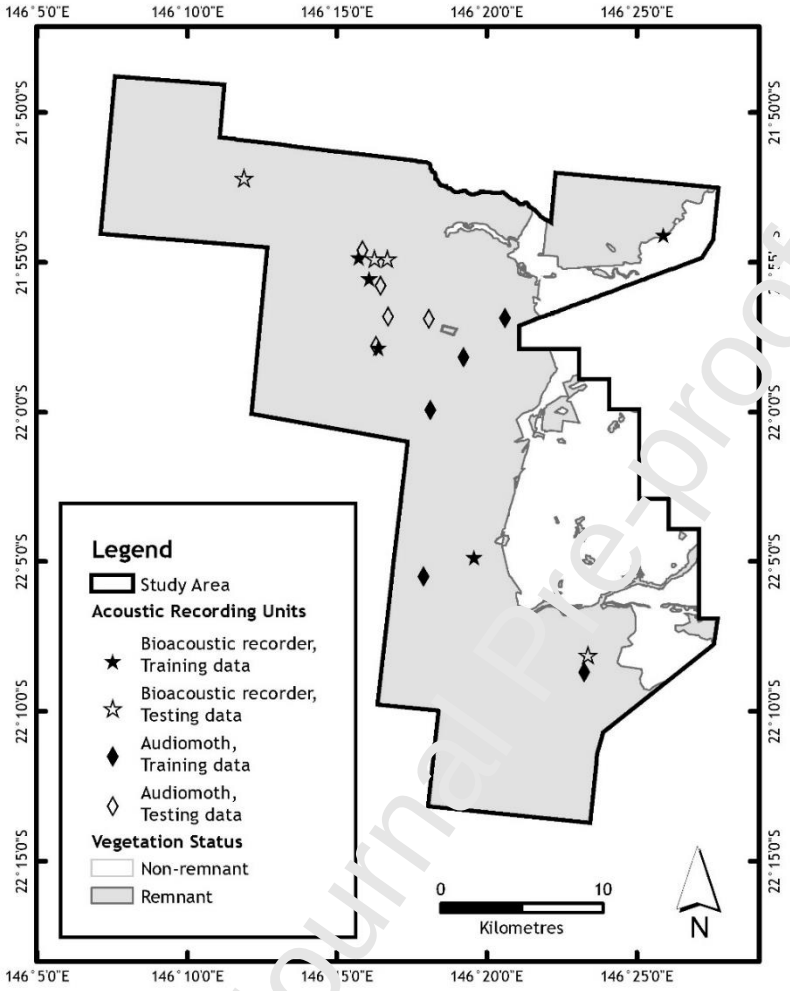

Figure 1. Spatial distribution of acoustic recording units within the Study Area. Filled and hollow symbols represent recordings used for training and testing data, respectively. Recorders were placed within remnant woodlands within the Study Area, focussing on areas that comprised broadly suitable habitat for southern black-throated finch.

2.3 CNN development

2.3.1 Data pre-processing

We split audio data into 1.8 second audio frames, which is approximately double the maximum length of the target call (Higgins, Peter and Cowling, 2006). We used a 50% overlap between frames, i.e. a sliding window approach, which ensured target calls would be entirely included within at least one audio frame (Kahl *et al.*, 2021). We then transformed audio frames into mel-scaled spectrogram images using a short-time Fourier transform, with a Hann

window length of 2048, a 50% overlap between segments and 128 mel filter banks, following standard methods such as those applied by LeBien *et al.* (2020).

We chose a frequency bandwidth of 1.5 MHz to 5 MHz, which ensured the dominant frequency of the target call was included within the frame, while avoiding excess noise at lower frequencies and higher frequencies (e.g. cicadas). We found that this frequency bandwidth typically included a fundamental harmonic and/or a higher harmonic of the dominant frequency; however, these harmonics were not the focus of the bandwidth selection due to their rapid attenuation at lower sound pressure levels (Koehler *et al.*, 2017). An example frame used as input into the CNN, following data pre-processing steps, is depicted within Figure 2.

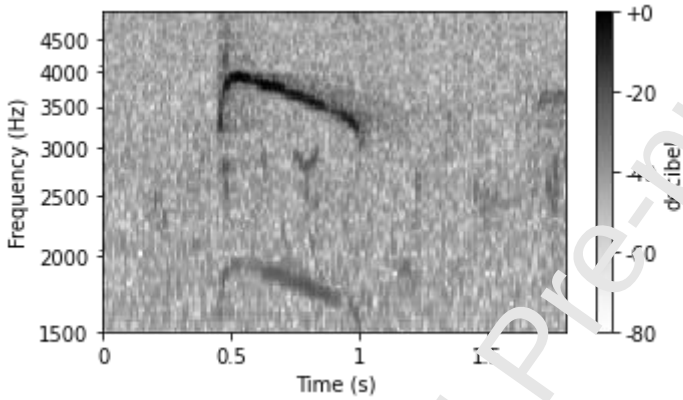

Figure 2. Example audio frame following data pre-processing steps that was used as input into the convolutional neural network. The audio frame shows a SBTF call with a dominant frequency (stronger signal) and a fundamental harmonic (weaker signal) included in the frame.

2.3.2 **CNN architecture**

Application of CNNs to audio recognition tasks is well established, with CNNs forming part of the 'standard recipe' for bioacoustic classification (LeBien *et al.*, 2020; Allen *et al.*, 2021; Stowell, 2022). We used a data pre-processing pipeline and CNN architecture that have been widely applied for bioacoustic classification tasks (Christin, Hervet and Lecomte, 2019; Stowell *et al.*, 2019; Stowell, 2022).

We compiled a CNN using Python (version 3.6.9, Python Foundation), within Google Colaboratory, to interface with Pytorch (version 1.10), an open source ML library (Paszke *et al.*, 2019). We used a ResNet-34 model with pre-trained weights for the CNN architecture. ResNet models typically achieve high-performance in image and audio recognition tasks (He *et al.*, 2016; Stowell *et al.*, 2019; Bergler *et al.*, 2022) and have been widely applied for automated wildlife image and call recognition (Sankupellay and Konovalov, 2018; Kahl *et al.*, 2021; Stowell, 2022). Our training dataset was imbalanced, with a lower number of audio frames containing a SBTF call than no-SBTF call. We followed the recommendations of Buda,

Maki and Mazurowski (2018) and oversampled the SBTF audio frames with the WeightedRandomSampler function in Pytorch. We used an Adam optimiser algorithm with an exponential learning rate decay function, which is a common method of learning rate optimisation used for CNN training (Kingma and Ba, 2014). We then applied a sigmoid activation to the output layer of the CNN to generate predictions between 0 and 1.

Active learning iterations used a batch size of 64, ten epochs and learning rate of 0.001. We used a grid search technique (Mohri, Rostamizadeh and Talwalkar, 2018) to tune hyperparameters of the final model including the number of epochs, batch size and learning rate.

2.3.3 Active learning framework

We applied an active learning approach to iteratively train and improve the CNN model. The active learning approach is depicted within Figure 3 and described below.

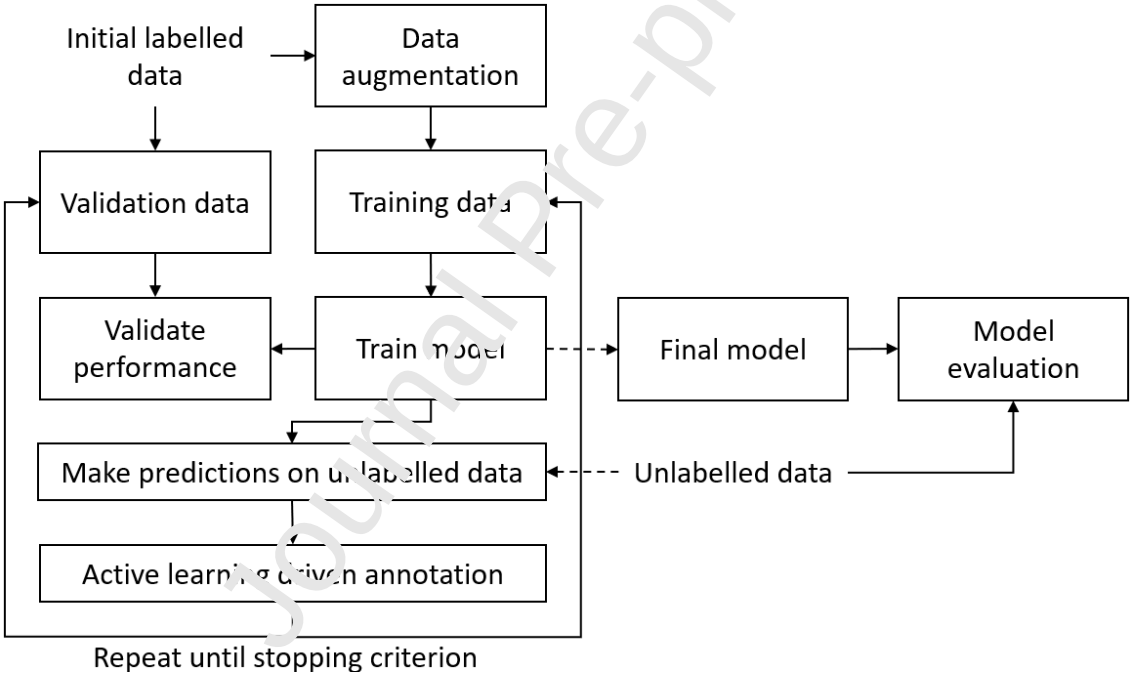

Figure 3. CNN model development pipeline that incorporates active learning iterations to develop a model for identifying calls of the southern black-throated finch.

We created an initial labelled dataset using unlabelled audio data from one solar-powered bioacoustic recorder (841 hours) and one Audiomoth recorder (108 hours). We clustered similar audio signals within these unlabelled data using the cluster analysis feature of Kaleidoscope Pro, which applied a hidden Markov model (Wildlife Acoustics, 2019). Clusters were manually reviewed and all identified target calls were labelled (n = 254). We also labelled non-target sounds that were distributed throughout the clusters (n = 661). We

augmented the initial training data by creating triplicate versions of each target call shifted at random horizontal (time domain) positions, which is a common technique to artificially increase the size of a training set (Stowell, 2022). An initial model was trained from this dataset.

Following the creation of an initial labelled dataset we developed the model using a series of active learning iterations. For each iteration we used the model to make predictions on a new set of unlabelled data, which comprised audio data from one solar-powered bioacoustic recorder (between 833 and 1402 hours) and one Audiomoth recorder (between 80 and 133 hours). Model predictions were in the form of a logit, on a scale of 0 to 1 (Monarch, 2021), where 0 represented the lowest probability of being the target call and 1 represented the greatest probability. Human annotators then labelled all predictions with a logit of greater than 0.5. The 0.5 logit cut-off was selected to prioritise the annotation of data likely to include target calls, which are rare within the unlabelled data, while also annotating signals that the model identified with the least certainty and would gain the greatest amount of new information, i.e. around the 0.5 logit; (Roh, Heo and Whang, 2019; Monarch, 2021). Through this active learning driven annotation process we grew the number of audio frames in the training data to 1,073 audio frames containing target calls and 5,786 non-target audio frames within five iterations (Table 1). All labels were reviewed by a primary annotator (J.V.O.) with 5 years' experience with SBTF prior to the next iteration of model training.

Each successive model iteration was trained on 70% of the training data and validated against 30% of the training data. The train/validation split approach minimises potential overfitting of the model on small samples sizes, when compared to a cross-validation approach (Vabalas *et al.*, 2019). We chose F_1 score as our validation metric, which is a standard performance metric that includes information on both model precision and recall (Mohri, Rostamizadeh and Talwalkar, 2018,; Stowell, 2022). Iterative model training stopped when the stopping criterion was reached, which was when successive model iterations did not reduce the F_1 score. A final model was then trained using the combined training and validation data. A separate test dataset was used for final testing/evaluation of the model, which is described in Section 2.4.1.

Table 1. The number of target and non-target audio frames used within each model iteration.

Iteration	Number of target audio frames	Number of non-target audio frames	Train:validation split
1	254	661	70:30
2	314	2642	70:30
3	1044	4,877	70:30
4	1073	5786	70:30
Final model	1073	5786	100:0

2.4 Model evaluation

2.4.1 Test dataset

Evaluation of models trained on data labelled through an active learning framework require evaluation designs beyond the standard method of testing the model on labelled data that are reserved from the initial labelled dataset, i.e. the 'hold-out test set' (Mohri, Rostamizadeh and Talwalkar, 2018; Stowell, 2022). This is because all labelled data available were identified using either hidden Markov model clustering in the Kaleidoscope Pro software (Wildlife Acoustics, 2019), or through the active learning process, which biases data to the model and may not be representative of the broader unlabelled data (Settles and Craven, 2008; Roh, Heo and Whang, 2019).

Therefore, to evaluate the final model, we used an unlabelled dataset that contained 2,735 hours of audio recorded from four sites with solar-powered bioacoustic recorders and five sites with Audiomoth recorders. These sites were independent of those used for the model training. The final model was run over the entire test dataset to predict if each audio frame (1.8 seconds in duration, with a 50% overlap between frames) contained a target call, resulting in c. 10.9 million predictions. We created a test dataset using a stratified random sampling approach. We selected a random sample of 500 audio frames from each 0.01 increment of prediction scores. Prediction scores were in the form of logits on the scale of 0 to 1 (refer to Section 2.3.2). When 0.01 logit increments had less than 500 audio frames, we included all audio frames within that logit increment. In total, the test dataset included 12,278 audio frames. The primary annotator manually labelled all audio frames in the test dataset.

2.4.2 Managing uncertain labels

Automated call recognition relies on the accurate identification of calls within the sample data. As noted previously, even highly trained experts can have difficulty achieving high annotation reliability from real-world recordings. In our study the accuracy of annotations is limited by factors such as similar vocalisations of co-occurring species, attenuation of the target calls at distance and overlap with other environmental noise.

Experience of annotators within our study suggested certain sounds cannot be definitively labelled as either 'SBTF' or 'Not SBTF'. These include calls that were substantially attenuated by distance from the microphone or were obscured by other sounds in the same frequency range, as well as short contact calls of SBTF, which are acoustically similar to other estrildid finches within the Study Area. We therefore allowed annotators to label audio frames into three categories, 'SBTF', 'Uncertain' and 'Not SBTF', following rules supplied to each annotator (provided in the Supplementary Data).

For our model, we used a 'classification with a reject option approach', which allowed audio frames to not be classified (i.e. labelled as uncertain) when the model predictions were ambiguous (Bishop, 2006; Thakur et al., 2019). Classification with rejection uses a threshold value that determines the threshold above which the model's predictions are labelled as 'uncertain'. For the model output of our CNN architecture, prediction scores closest to the logit of 0.5 have the highest uncertainty. The threshold value determined the distance away from the logit 0.5 that we included in the 'uncertain' category. We chose this threshold value to match the proportion of uncertain audio frames in the final model's output to the proportion of uncertain audio frames labelled by the primary annotator. Thus, the final model achieved the same likelihood of classifying an audio frame as uncertain (the reject option) as an annotator of labelling an audio frame as uncertain. This approach allowed us to map the CNN predictions into the same three categories used by the annotators and facilitate a direct evaluation of annotation performance among the model and annotators.

2.4.3 Model performance

We evaluated the performance of the final model and all preceding active learning iterations against the test dataset. We converted all labels and model predictions to an ordinal scale, with 'Not SBTF' assigned a value of 1, 'Uncertain' a value of 2, and 'SBTF' a value of 3. We evaluated model performance using macro-averaged mean absolute error (MMAE). MMAE is a common evaluation metric for ordinal classification (Cardoso and Sousa, 2011) and accounts for imbalanced datasets (Baccianella, Esuli and Sebastiani, 2009). We did not use the 'standard metrics' of performance recommended by Stowell (2022), which include precision, recall, F-score and area under the curve metrics, as MMAE accounts for the ordinal nature of our model predictions and labels compared to categorical data that is assumed by the

'standard metrics' (Cardoso and Sousa, 2011). We chose an analysis suitable for ordinal data as the 'uncertain' label is considered closer to the label 'SBTF' and 'Not SBTF' than 'SBTF' and 'Not SBTF' are to each other. We estimated confidence intervals of the MMAE using bootstrapping (n = 10,000). We conducted all analyses using the imblearn version 0.8 package in Python.

2.4.4 Inter-annotator agreement

Two experts (B.D. and E.M.) independently annotated a random subset of 9.5% (n = 1,165) of the test dataset as secondary annotators. Both experts had over 10 years of experience working on the species. The annotation process was blind, with each expert not receiving annotations from the other expert or primary annotator (J.V.O.).

We assessed inter-annotator agreements using the agreement coefficient Krippendorff's alpha (α) (Krippendorff, 2011; Monarch, 2021). We then bootstrapped (n = 10,000) the distribution of each agreement coefficient and calculated 95% confidence intervals, as recommended by Hayes and Krippendorff (2007). It is generally regarded that values above 0.8 represent good agreement among annotators, while values within the range of 0.667 to 0.8 represent tolerable agreement among annotators (Reidsma and Carletta 2008).

We assessed agreement among the final model, the primary annotator and expert secondary annotators using pairwise combinations of α and tested for differences among the pairwise combinations using a Kruskal-Wallis Test. We tested the effect of excluding 'uncertain' labels during model evaluation by calculating agreement coefficients with 'uncertain' labels retained and removed using a Wilcoxon rank-sum test. We also tested for a difference between the agreement coefficients of data recorded on solar-powered bioacoustic recorders compared to Audiomoth devices using a Wilcoxon rank-sum test. We undertook all significance tests using Scipy version 1.10.0.

3 Results

3.1 CNN model development with active learning

The CNN model performance stabilised within five active learning iterations (Figure 4). The majority of model performance improvement was achieved within the first two iterations, with a 34.3% decrease in macro-averaged MAE between iterations one and three, compared to a 12.1% decrease in macro-averaged MAE between iterations three and the final model (Figure 4). Most of the training data added between iterations one and three consisted of non-target calls (Table 1), which led to rapid improvements in the model's ability to differentiate non-target sounds, resulting in a reduction in false positives.

Out of a total of 10,920,849 audio frames of unlabelled data used for testing, the final model predicted that 3,024 audio frames contained a SBTF call (0.028%), 4,581 audio frames were uncertain (0.042%), while the remaining 10,913,244 audio frames (99.9%) did not contain a SBTF call. Southern black-throated finch calls were detected from all locations; however, the frequency of SBTF calls within the audio data varied among recorder locations, ranging from 0.002% to 0.060% of the audio frames.

3.2 Inter-annotator agreement

Overall, the results showed good agreement among the model and human-annotators, with agreement coefficients ranging from 0.760 α to 0.878 α (Figure 5). The highest agreement coefficients values were observed between the primary annotator and expert two (0.878 α) and between the primary annotator and the final model (0.859 α). The agreement coefficients were similar between the final model and the two experts (0.760 α and 0.795 α respectively for experts 1 and 2) and between the two experts (0.808 α). Overall, the agreement coefficients show that the final model labelled data within the range of annotation uncertainty that was present among the primary annotator and experts. The final model was more consistent with the primary annotator, who annotated the training data, compared to the experts who independently annotated test data only.

The removal of audio frames labelled as 'uncertain' by the primary annotator during model evaluation caused a small but significant increase in the agreement between the final model and the primary annotator (mean increase 0.016 α, $p<0.01$, Figure 6). There was reasonable agreement between the audio frames labelled by both the final model and the primary annotator as 'uncertain, with 63.4% of the audio frames labelled 'uncertain' by the primary annotator also being labelled 'uncertain' by the final model. This is compared to the final model assigning an 'uncertain' label to 12.3% and 7.55% of the audio frames which the primary annotator labelled as 'SBTF' and 'not SBTF', respectively (Figure 6).

Inter-annotator agreement was greater for data recorded on solar-powered bioacoustic recorders than Audiomoths (0.824 compared to 0.639; Figure 7).

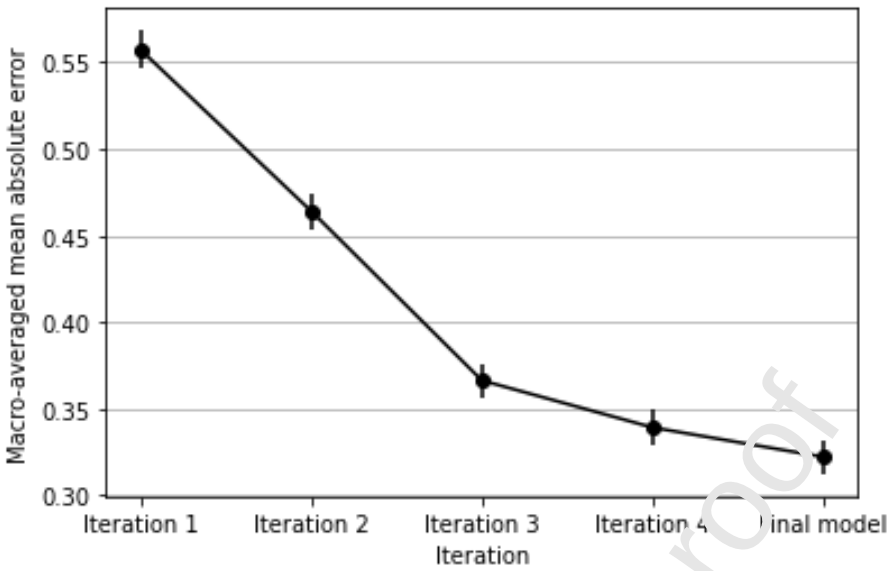

Figure 4. Evaluation performance of each model iteration. Error bars show 95% confidence intervals.

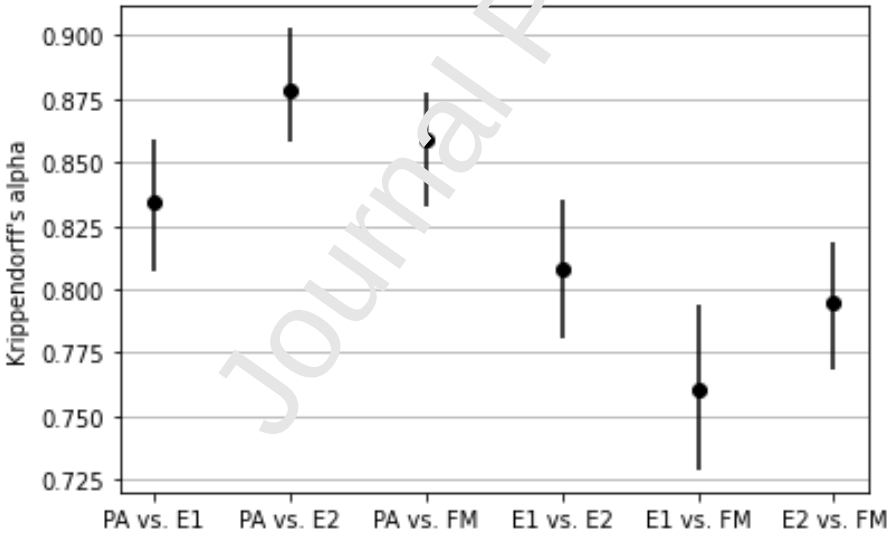

Figure 5. Pairwise comparison of inter-annotator agreement among the primary annotator (PA), experts one and two (E1 and E2) and the final model (FM). Error bars show 95% confidence intervals.

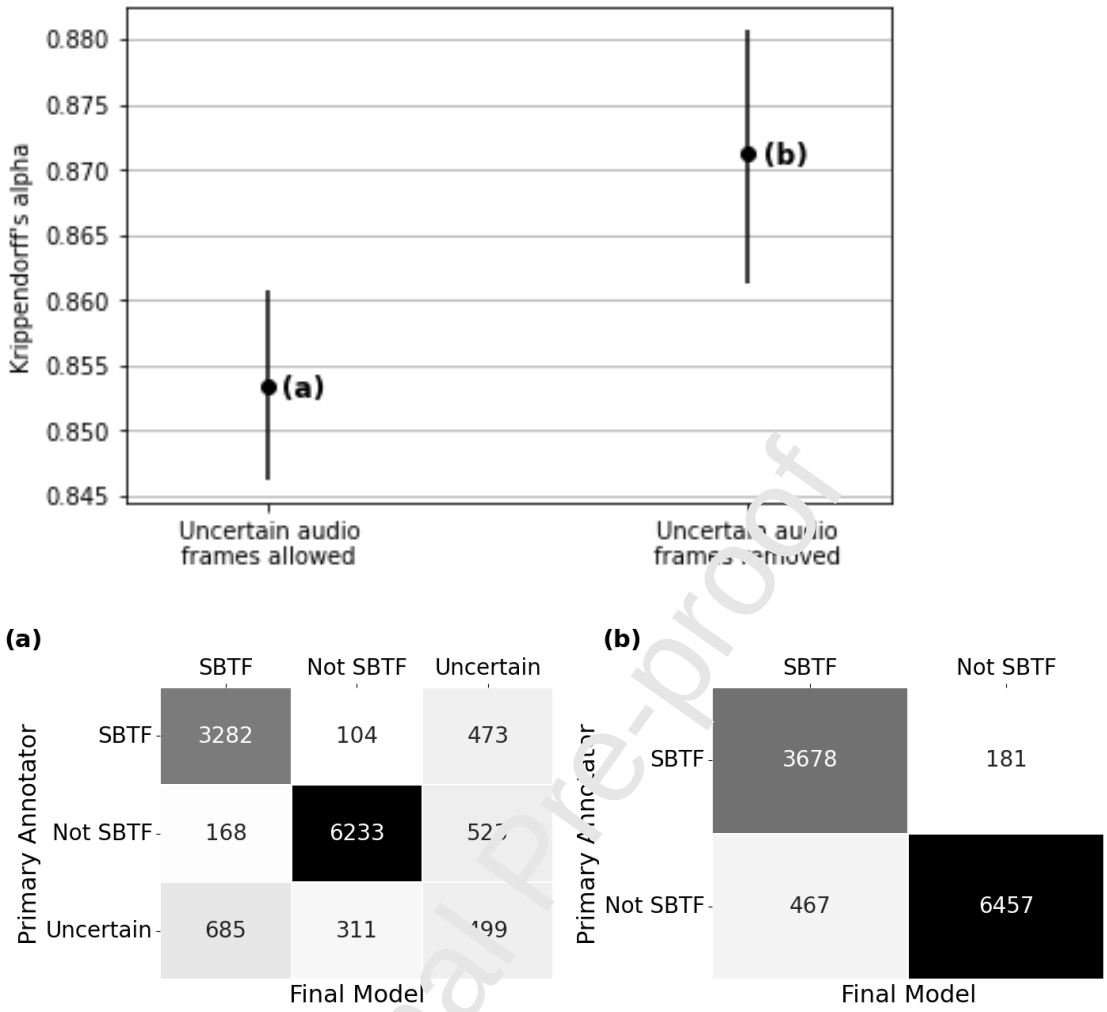

Figure 6. Comparison of agreement between the primary annotator and the final model when the evaluation process has allowed 'uncertain' audio frames (a), compared to when 'uncertain' audio frames, as assessed by the primary annotator, have been excluded (b). Error bars show 95% confidence intervals. Confusion matrixes are presented below each option as heat plots. The heat plot shading and numbers represent the count of audio frames labelled in each class by the primary annotator compared to the final model.

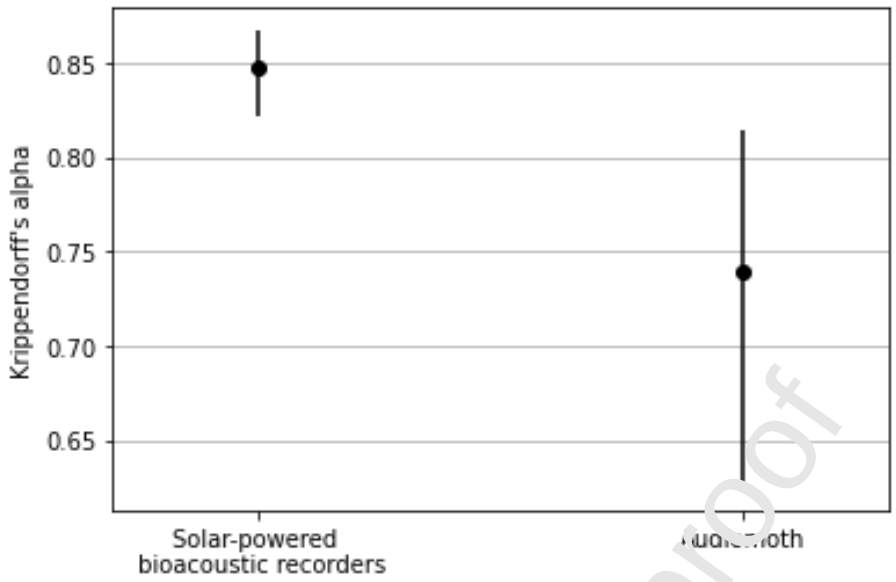

Figure 7. Agreement among all annotators, including the primary annotator and two experts, for audio frames captured using solar-powered bioacoustic recorders (Frontier Labs) and Audiomoths (Hill *et al.*, 2019). Error bars show 95% confidence intervals.

4 Discussion

Our study demonstrates that active learning is an effective strategy for building machine learning (ML) classifiers for species with limited labelled training data. Using as few as five active learning iterations, we generated a final labelled dataset of 1,073 target calls and 5,786 non-target sounds. This was sufficient to create a model with classification abilities comparable to experts familiar with the species. Active learning focusses scarce human resources to annotate data that are most valuable for model performance (Van Engelen and Hoos, 2020), primarily records with greater uncertainty. In our study, we found that the active learning framework selected audio frames for annotation that included non-target calls with similarity to the target call, as well as sounds that were dissimilar to the target call due to over-fitting in early model iterations. Over-fitting occurs when the model learns noise in the training data instead of the underlying pattern of the target call (Bishop, 2006). Annotation of these uncertain calls had a measurable impact on model performance. As such, active learning can reduce the time and effort required to develop labelled training data for building classifiers (Priyadarshani, Marsland and Castro, 2018; Stowell, 2022).

Active learning reduces the cost of developing labelled datasets, which form the foundation of successful call detection models (Van Engelen and Hoos, 2020). We estimate that 0.028% of our unlabelled data comprises the target call, making manual screening of our unlabelled data resource intensive. Manually obtaining an equivalent number of target calls to our final

dataset (n = 1,073) by reviewing our unlabelled audio data would have required the review of approximately 1,916 hours of audio data. Active learning provided an approach to overcome this barrier to developing a call detection model for the target species. As many wildlife species do not have existing labelled datasets, an active learning approach has general applicability to the field of bioacoustics, particularly for species whose calls are rare in long-term field recordings (Ricci, Rokach and Shapira, 2022). Development of open access methods to improve call recognisers for these rare and difficult to survey species, many of which are also threatened (Loiseau *et al.*, 2020), will contribute to research and monitoring that removes data deficiencies for these species and thus improves policy and conservation outcomes (Davies, Margules and Lawrence, 2004; Sekercioglu *et al.*, 2008; Loiseau *et al.*, 2020).

Active learning is a powerful method for bioacoustic deep learning, but it poses certain challenges that deviate from the 'standard recipe' of bioacoustic deep learning described by Stowell (2022). One challenge is that the active learning process biases labelled data to the model since the model's predictions guide the annotation process. Resulting labels therefore cannot be used in the model's evaluation (Roh, Heo and Whang, 2019; Ricci, Rokach and Shapira, 2022). Chambert *et al.* (2018) and Ruff *et al.* (2020) addressed this issue by selecting audio frames for review post-processing based on their model's predictions. We extended this approach to account for a highly imbalanced dataset. Evaluation of a highly imbalanced dataset through random selection alone would require unfeasibly large numbers of audio frames to be manually reviewed to ensure sufficient target calls were captured to give reliable evaluation metrics (Raeder, Forman and Chawla, 2012). For example, where target calls make up 0.028% of the unlabelled data, which is the case for our unlabelled data, approximately 357,000 audio frames would require manual review to capture 100 target calls. To overcome this challenge, we applied a random selection approach that was stratified across the model's predictions (logits). While this approach substantially reduced class imbalance within our evaluation data and allowed for the calculation of reliable evaluation metrics (Raeder, Forman and Chawla, 2012), the nature of this approach alters the distribution of data and prohibits evaluation metrics being generalised to the unlabelled data. Additional research is needed to investigate more appropriate evaluation methods for highly imbalanced and unlabelled test data.

An active learning strategy requires consideration of biases that may be introduced during annotation (Monarch, 2021). While annotation bias is a consideration for all ML datasets, the active learning framework allows ML models to be generated from smaller datasets and potentially developed by fewer annotators compared to, for example, a publicly available call database such as Xeno-canto (Vellinga and Planqué, 2015). In our study, the final model's predictions most closely reflected the primary annotator's labels, compared to the expert's

labels, which suggests that the model is replicating biases of the training data (Roh, Heo and Whang, 2019; Koenecke *et al.*, 2020). Performance improvements may therefore be achieved through the inclusion of labelled training data from additional annotators experienced with the species, or further investigation of the vocal repertoire of the target species compared to potential false positives. Broadly, our results highlight the importance of considering annotation biases, particularly for datasets created by a small group of annotators.

The use of multiple annotators to label training data requires consideration of inter-annotator agreement and how this affects ML model performance. Classifying wildlife vocalisations with certainty is unfeasible for many species. While ML models may theoretically be able to outperform inter-annotator agreement, the level of inter-annotator agreement remains a useful benchmark to assess model performance (Boguslav and Cohen, 2017; Richie, Grover and Tsui, 2022). In the field of natural language processing, where labels are often subjective, inter-annotator agreement metrics are a common tool to assess label quality and model performance (Pustejovsky and Stubbs, 2012). Our results support these arguments with the congruence among annotators providing a useful benchmark to assess model performance and identify potential areas of improvement in the training data.

The standard approach to build a bioacoustic classification model with ML often explicitly or implicitly assumes that labels are accurate, with limited scope for the inclusion of ambiguous vocalisations (Cabitza *et al.*, 2020; Otani *et al.*, 2020; Campagner *et al.*, 2021). Our results show that exclusion of ambiguous vocalisations from the evaluation dataset significantly inflated the model's evaluation metrics. Our findings therefore agree with those of Cabitza *et al.* (2020) from the medical literature, that model performance is overestimated if inter-annotator agreement is not accounted for. While the effect size for our data was small, due to a good agreement among annotators, the overestimation of model performance is theoretically negatively correlated with inter-annotator agreement (Cabitza *et al.*, 2020). We recommend that the development of models for species with ambiguous vocalisations account for inter-annotator agreement in design. We demonstrated how this can be accomplished using multiple annotators and adoption of a classification with a 'reject' option approach (i.e. the 'uncertain' category), which allows the model results to be directly compared to those of the annotators (Bishop, 2006; Campagner, Cabitza and Ciucci, 2019). Other approaches applied within the natural language processing and medical imaging fields include the development of a 'gold-standard' set of labels through the use of multiple annotators or label cleaning (Pustejovsky and Stubbs, 2012; Karimi *et al.*, 2020), or incorporating annotator or label uncertainty into the machine learning process (Nguyen, Valizadegan and Hauskrecht, 2014; Hüllermeier and Waegeman, 2021). These approaches all require the evaluation of inter-annotator agreement within the study design, which is supported by the results of our study.

Assessment of inter-annotator agreement allows investigation of the sources of disagreement and potential improvements to annotator accuracy and label reliability (Duc *et al.*, 2021; Monarch, 2021). In our study, inter-annotator agreement was generally high for the entire unlabelled dataset, however inter-annotator agreement for data recorded on Audiomoth devices was substantially lower when compared to solar-powered bioacoustic recorders. Two factors may explain this difference: the placement of Audiomoth devices in proximity to nests and differences in the recording quality of Audiomoth and bioacoustics recorders.

Having been placed near active SBTF nests, Audiomoths would likely have captured a broader range of SBTF calls than bioacoustics recorders placed at foraging and drinking sites, including numerous softer calls unlikely to be heard at distance (see Shephard, Pridham and Forshaw, 2012). Annotators would likely differ in their familiarity with these softer calls, resulting in lower inter-annotator agreement on Audiomoth recordings when compared with recordings from bioacoustic recorders (which are more likely to have captured the more commonly heard, louder and distinctive 'long'/'pew' call of SBTF).

Differences in the volume/amplitude and quality of recordings captured by bioacoustic recorders and Audiomoths may also have affected annotators assessment of calls, with bioacoustics recorders having a signal-to-noise ratio (SNR) of 80 dB compared with 42 dB in Audiomoths (Roe *et al.*, 2021; Open Acoustic Devices, n.d.). The higher SNR of bioacoustic recorders would afford greater sound fidelity and less background noise, thereby providing recordings that align more closely with the annotators' field experience (Turgeon, Van Wilgenburg and Drake, 2017; Darras *et al.*, 2020).

5 Conclusions

Our study investigated the efficacy of active learning as a framework for building deep learning models in a setting where limited training data were available and the cost to obtain training data was high. Using field recordings of a threatened species, the southern black-throated finch, we successfully developed and demonstrated the value of an open access active learning method that considered imbalanced and unlabelled data and ambiguous vocalisations, which are common barriers to constructing call recognition models for rare or cryptic species, and other species that vocalise infrequently. Our results demonstrate the utility of these methods in developing effective call recognisers for rare and difficult to survey species.

Our results also support the assessment of inter-annotator agreement during the development of call recognition models, with the assessment of inter-annotator agreement allowing for more accurate evaluation of model performance and identification of annotator biases, as well helping identify areas for improvement in training data.

6 Author contributions

John van Osta: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization, Writing – Review & Editing. Brad Dreis: Conceptualization, Data Curation, Project Administration, Funding acquisition. Ed Meyer: Conceptualization, Writing – Review & Editing. Laura Grogan: Writing – Review & Editing, Supervision. Guy Castley: Writing – Review & Editing, Supervision.

7 Acknowledgements

Funding: This work was supported by Bravus Mining and Resources and E2M Pty Ltd. The authors thank Jessica Hogg and Cameron Davey for their contribution to data annotation and model implementation and Hector Pople for reviewing the steps to reproduce this study's findings within the Supplementary Material.

8 Declaration of competing interests

None.

9 References

Alldredge, M.W., Simons, T.R. and Pollock, K.H. (2007) 'Factors affecting aural detections of songbirds', *Ecological Applications*, 17(3), pp. 948–955.

Allen, A.N. *et al.* (2021) 'A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset', *Frontiers in Marine Science*, 8, p. 607321.

Baccianella, S., Esuli, A. and Sebastiani, F. (2009) 'Evaluation measures for ordinal regression', in. *2009 Ninth international conference on intelligent systems design and applications*, IEEE, pp. 283–287.

Bergler, C. *et al.* (2019) 'ORCA-SPOT: An automatic killer whale sound detection toolkit using deep learning', *Scientific Reports*, 9(1), pp. 1–17.

Bergler, C. *et al.* (2022) 'ANIMAL-SPOT enables animal-independent signal detection and classification using deep learning', *Scientific Reports*, 12(1), p. 21966.

Bishop, C.M. (2006) *Pattern recognition and machine learning*. New York: Springer.

Bland, L.M. *et al.* (2015) 'Predicting the conservation status of data-deficient species', *Conservation Biology*, 29(1), pp. 250–259.

Boguslav, M. and Cohen, K.B. (2017) 'Inter-annotator agreement and the upper limit on machine performance: Evidence from biomedical natural language processing.', *Studies in health technology and informatics*, 245, pp. 298–302.

Braby, M.F. (2018) 'Threatened species conservation of invertebrates in Australia: an overview', *Austral Entomology*, 57(2), pp. 173–181.

Buda, M., Maki, A. and Mazurowski, M.A. (2018) 'A systematic study of the class imbalance problem in convolutional neural networks', *Neural networks*, 106, pp. 249–259.

Cabitza, F. *et al.* (2020) 'The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability', *Applied Sciences*, 10(11), p. 4014.

Campagner, A. *et al.* (2021) 'Ground truthing from multi-rater labeling with three-way decision and possibility theory', *Information Sciences*, 545, pp. 771–790.

Campagner, A., Cabitza, F. and Ciucci, D. (2019) 'Three-way classification: Ambiguity and abstention in machine learning', in. *Rough Sets: International Joint Conference, IJCRS 2019, Debrecen, Hungary, June 17–21, 2019, Proceedings*, Springer, pp. 280–294.

Cardoso, J.S. and Sousa, R. (2011) 'Measuring the performance of ordinal classification', *International Journal of Pattern Recognition and Artificial Intelligence*, 25(08), pp. 1173–1195.

Chambert, T. *et al.* (2018) 'A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing', *Methods in Ecology and Evolution*, 9(3), pp. 560–570.

Christin, S., Hervet, É. and Lecomte, N. (2019) 'Applications for deep learning in ecology', *Methods in Ecology and Evolution*, 10(10), pp. 1632–1644.

Colonna, J.G., Carvalho, J.R. and Rosso, O.A. (2020) 'Estimating ecoacoustic activity in the Amazon rainforest through Information Theory quantifiers', *PloS One*, 15(7), p. e0229425.

Darras, K.F. *et al.* (2020) 'High microphone signal-to-noise ratio enhances acoustic sampling of wildlife', *PeerJ*, 8, p. e9955.

Davies, K.F., Margules, C.R. and Lawrence, J.F. (2004) 'A synergistic effect puts rare, specialized species at greater risk of extinction', *Ecology*, 85(1), pp. 265–271.

Dema, T. *et al.* (2020) 'Acoustic detection and acoustic habitat characterisation of the critically endangered white-bellied heron (Ardea insignis) in Bhutan', *Freshwater Biology*, 65(1), pp. 153–164.

Department of Environment and Science (DES) (2023) 'Biodiversity status of 2019 remnant regional ecosystems - Queensland - Version 12.2'. Queensland Government.

Duc, P.N.H. *et al.* (2021) 'Assessing inter-annotator agreement from collaborative annotation campaign in marine bioacoustics', *Ecological Informatics*, 61, p. 101185.

Fairbrass, A.J. *et al.* (2019) 'CityNet—Deep learning tools for urban ecoacoustic assessment', *Methods in Ecology and Evolution*, 10(2), pp. 186–197.

Gibb, R. *et al.* (2019) 'Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring', *Methods in Ecology and Evolution*, 10(2), pp. 169–185.

Gwet, K.L. (2014) *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hayes, A.F. and Krippendorff, K. (2007) 'Answering the call for a standard reliability measure for coding data', *Communication Methods and Measures*, 1(1), pp. 77–89.

He, K. *et al.* (2016) 'Deep residual learning for image recognition', in. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hervás, M. *et al.* (2017) 'An FPGA-based WASN for remote real-time monitoring of endangered species: A case study on the birdsong recognition of *Botaurus stellaris*', *Sensors*, 17(6), p. 1331.

Higgins, P.J., Peter, J.M. and Cowling, S.J. (2006) *Handbook of Australian, New Zealand and Antarctic Birds*. Melbourne: Oxford University Press (Boatbill to Starlings).

Hill, A.P. *et al.* (2019) 'AudioMoth: A low-cost acoustic device for monitoring biodiversity and the environment', *HardwareX*, 6. Available at: https://doi.org/10.1016/j.ohx.2019.e00073.

Howard, S.D. and Bickford, D.P. (2014) 'Amphibians over the edge: silent extinction risk of Data Deficient species', *Diversity and distributions*, 20(7), pp. 837–846.

Hüllermeier, E. and Waegeman, W. (2021) 'Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods', *Machine Learning*, 110, pp. 457–506.

Kahl, S. *et al.* (2021) 'BirdNET: A deep learning solution for avian diversity monitoring', *Ecological Informatics*, 61, p. 101236.

Karimi, D. *et al.* (2020) 'Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis', *Medical Image Analysis*, 65, p. 101759.

Kingma, D.P. and Ba, J. (2014) 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* [Preprint].

Koehler, J. *et al.* (2017) 'The use of bioacoustics in anuran taxonomy: theory, terminology, methods and recommendations for best practice', *Zootaxa*, 4251(1), pp. 1–124.

Koenecke, A. *et al.* (2020) 'Racial disparities in automated speech recognition', *Proceedings of the National Academy of Sciences*, 117(14), pp. 7684–7689.

Krippendorff, K. (2011) *Computing Krippendorff's alpha-reliability*. Working Paper. University of Pennsylvania. Available at: https://repository.upenn.edu/asc_papers/43.

LeBien, J. *et al.* (2020) 'A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network', *Ecological Informatics*, 59, p. 101113.

Loiseau, N. *et al.* (2020) 'Global distribution and conservation status of ecologically rare mammal and bird species', *Nature Communications*, 11(1), p. 5071.

Mac Aodha, O. *et al*. (2018) 'Bat detective—Deep learning tools for bat acoustic signal detection', *PLoS Computational Biology*, 14(3), p. e1005995.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018) *Foundations of machine learning*. MIT Press.

Monarch, R.M. (2021) *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Mortimer, J.A. and Greene, T.C. (2017) 'Investigating bird call identification uncertainty using data from processed audio recordings', *New Zealand Journal of Ecology*, 41(1), pp. 126–133.

Nguyen, Q., Valizadegan, H. and Hauskrecht, M. (2014) 'Learning classification models with soft-label information', *Journal of the American Medical Informatics Association*, 21(3), pp. 501–508.

Open Acoustic Devices (n.d.) *Audio Quality*. Available at: https://www.openacousticdevices.info/audio (Accessed: 19 March 2023).

Otani, N. *et al*. (2020) 'Binary classification with ambiguous training data', *Machine Learning*, 109, pp. 2369–2388.

Paszke, A. *et al*. (2019) 'Pytorch: An imperative style, high-performance deep learning library', *Advances in Neural Information Processing Systems*, 32, pp. 8026–8037.

Priyadarshani, N., Marsland, S. and Castro, I. (2018) 'Automated birdsong recognition in complex acoustic environments: a review', *Journal of Avian Biology*, 49(5). Available at: https://doi.org/10.1111/jav.01447.

Pustejovsky, J. and Stubbs, A. (2012) *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Raeder, T., Forman, G. and Chawla, N.V. (2012) 'Learning from imbalanced data: Evaluation matters', *Data Mining: Foundations and Intelligent Paradigms: Volume 1: Clustering, Association and Classification*, pp. 315–331.

Reidsma, D. and Carletta, J. (2008) 'Reliability measurement without limits', *Computational Linguistics*, 34(3), pp. 319–326.

Ricci, F., Rokach, L. and Shapira, B. (2022) *Recommender systems handbook*. Third edition. New York, NY: Springer. Available at: https://doi.org/10.1007/978-1-0716-2197-4.

Richie, R., Grover, S. and Tsui, F.R. (2022) 'Inter-annotator agreement is not the ceiling of machine learning performance: Evidence from a comprehensive set of simulations',

in. *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 275–284.

Roe, P. *et al.* (2021) 'The Australian acoustic observatory', *Methods in Ecology and Evolution*, 12(10), pp. 1802–1808.

Roh, Y., Heo, G. and Whang, S.E. (2019) 'A survey on data collection for machine learning: a big data-ai integration perspective', *IEEE Transactions on Knowledge and Data Engineering*, 33(4), pp. 1328–1347.

Ross, S.R. *et al.* (2023) 'Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions', *Functional Ecology* [Preprint].

Ruff, Z.J. *et al.* (2020) 'Automated identification of avian vocalizations with deep convolutional neural networks', *Remote Sensing in Ecology and Conservation*, 6(1), pp. 79–92.

Ruff, Z.J. *et al.* (2021) 'Workflow and convolutional neural network for automated identification of animal sounds', *Ecological Indicators*, 124, p. 107419.

Sankupellay, M. and Konovalov, D. (2018) 'Bird call recognition using deep convolutional neural network, ResNet-50', in. *Proceedings of Acoustics*.

Sekercioglu, C.H. *et al.* (2008) 'Climate change, elevational range shifts, and bird extinctions', *Conservation biology*, 22(1), pp. 140–150.

Settles, B. and Craven, M. (2008) 'An analysis of active learning strategies for sequence labeling tasks', in. *proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 1070–1079.

Shephard, M., Pridham, A. and Forshaw, J.M. (2012) *Grassfinches in Australia*. Melbourne: CSIRO Publishing.

Shi, B. *et al.* (2019) 'Semi-supervised acoustic event detection based on tri-training', in. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 750–754.

Simons, T.R. *et al.* (2007) 'Experimental analysis of the auditory detection process on avian point counts', *The Auk*, 124(3), pp. 986–999.

Stowell, D. *et al.* (2019) 'Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge', *Methods in Ecology and Evolution*, 10(3), pp. 368–380.

Stowell, D. (2022) 'Computational bioacoustics with deep learning: a review and roadmap', *PeerJ*, 10, pp. e13152–e13152. Available at: https://doi.org/10.7717/peerj.13152.

Sugai, L.S.M. *et al.* (2019) 'Terrestrial passive acoustic monitoring: review and perspectives', *BioScience*, 69(1), pp. 15–25.

Teixeira, D. *et al.* (2022) 'Fledge or fail: Nest monitoring of endangered black-cockatoos using bioacoustics and open-source call recognition', *Ecological Informatics*, 69, p. 101656.

Threatened Species Scientific Committee (TSSC) (2005) *Commonwealth listing advice on southern black-throated finch (Poephila cincta cincta)*. Canberra, ACT.

Towsey, M. *et al.* (2012) 'A toolbox for animal call recognition', *Bioacoustics*, 21(2), pp. 107–125.

Turgeon, P., Van Wilgenburg, S. and Drake, K. (2017) 'Microphone variability and degradation: implications for monitoring programs employing autonomous recording units', *Avian Conservation and Ecology*, 12(1).

Vabalas, A. *et al.* (2019) 'Machine learning algorithm validation with a limited sample size', *PLoS One*, 14(11), p. e0224365.

Van Engelen, J.E. and Hoos, H.H. (2020) 'A survey on semi-supervised learning', *Machine Learning*, 109(2), pp. 373–440.

Vellinga, W.-P. and Planqué, R. (2015) 'The xeno-canto collection and its relation to sound recognition and classification.', in. *CLEF (Working Notes)*.

Wildlife Acoustics (2019) 'Kaleidoscope Pro Analysis Software'.

Wood, C.M. *et al.* (2023) 'Challenges and opportunities for bioacoustics in the study of rare species in remote environments', *Conservation Science and Practice*, p. e12941.

Zhu, X. and Goldberg, A.B. (2009) 'Introduction to semi-supervised learning', *Synthesis lectures on artificial intelligence and machine learning*, 3(1), pp. 1–130.

Appendix A. Supplementary material

Supplementary data to this article can be found online at Figshare DOI: 10.6084/m9.figshare.23053382.

Note: the above DOI will be hyperlinked to the final dataset. A draft dataset for peer review can be accessed through this link: https://figshare.com/s/b1377829938b276b17ea. The dataset will be finalised and published following peer review comments.

Table 2: Audio annotation rules provided to annotators

Label	Description
SBTF (southern black-throated finch)	You have evaluated the detection to contain a SBTF call. This may include any type of SBTF call. You need to be certain, or confident, that a call in the detection was made by one (or multiple) SBTF.
Not SBTF	You have evaluated the detection to not contain a SBTF call. You need to be certain, or confident, that no calls in the detection were made by SBTF.
Uncertain	You are not confident that one (or multiple) call/s within the detection were made by a SBTF or were made by another species.

Highlights

Title: An active learning framework and assessment of inter-annotator agreement facilitate automated recogniser development for vocalisations of a rare species, the black-throated finch

Manuscript number: ECOINF-D-23-00645

- Active learning is an efficient data annotation strategy for rare calls.
- Assessing inter-annotator agreement benefits model evaluation and bias detection.
- Excluding ambiguous vocalisations artificially inflates model performance.
- We demonstrate an approach to manage ambiguous vocalisations in model development.