NEW ZEALAND JOURNAL OF ECOLOGY

RESEARCH

An invasive species model and dataset for bioacoustic monitoring of common brushtail possum

Ben McEwen[1,2]*, Andrew Bainbridge-Smith[1], Stefanie Gutschmidt[2], Richard Green[1], James Atlas[1]

[1]Computer Science and Software Engineering, University of Canterbury, 20 Kirkwood Ave, Christchurch, 8140, Canterbury, New Zealand
[2]Mechanical Engineering, University of Canterbury, 20 Kirkwood Ave, Christchurch, 8140, Canterbury, New Zealand
*Author for correspondence (Email: ben.mcewen@pg.canterbury.ac.nz)

Published online: 18 July 2024

Abstract: Passive acoustic monitoring (PAM) is a critical tool in the monitoring and conservation of native species but until now its use in the detection of invasive species has been under-utilised. We present the first publicly available dataset of invasive common brushtail possum (*Trichosurus vulpecula*) vocalisations including 3500 annotated field recording segments. This study presents an automatic classification model designed and fine-tuned to detect the presence/absence of possums, achieving 98.4% test set accuracy and F1 score of 0.983. To our knowledge, this is the first model of its kind applied to the target taxa. We also discuss the development of computational tools in the context of invasive species detection, conservation potential, and critical challenges such as vocalisation frequency and feature sparsity. All data and code are publicly available.

Keywords: audio classification, bioacoustics, dataset, machine learning

Introduction

Bioacoustic monitoring is commonly applied to avian species in New Zealand (Priyadarshani et al. 2016; Mortimer & Greene 2017; Bombaci & Pejchar 2018; Jahn et al. 2022), with results demonstrating performance benefits compared to human observations (Darras et al. 2019; Hoefer et al. 2023). Bioacoustic monitoring is also applied worldwide (Stowell et al. 2018) to monitor the biodiversity of ecosystems and inform conservation decisions, with the majority of terrestrial studies focused on bats and birds (Sugai et al. 2019). In New Zealand, bioacoustic monitoring tools have not yet been applied to invasive mammalian species. This is in part due to biological constraints (Ross et al. 2023), as the target species is not highly vocal, and, the resource costs and human labour required to collect and analyse raw field recordings. Over recent years passive acoustic monitoring (PAM) tools and the broader field of computational bioacoustics have progressed significantly. Computational bioacoustics includes automatic segmentation techniques, used to identify features of interest, and classification techniques that aid in the analysis of acoustic data. Computational bioacoustics recent progress is in part due to the availability of affordable PAM devices and progress in machine learning (Stowell 2021). PAM presents a cost-effective tool for landscape-scale bioacoustic monitoring of challenging species that previously may have been considered infeasible. This presents an opportunity to extend current monitoring tools for the detection of challenging invasive species such as possums. PAM has the potential to provide a cost-effective monitoring and detection and automated classification can improve the scalability of data collection and analysis.

We present the first publicly available dataset of common brushtail possum (*Trichosurus vulpecula*) vocalisations as well as a classification model developed to detect the presence/absence of possum vocalisation in long term field recordings. We also discuss the challenges and opportunities associated with acoustic monitoring of typically less vocal invasive species. We hope that these contributions provide a useful resource to accelerate the development of improved bioacoustic monitoring tools in New Zealand.

Methods

Data collection

Audio was primarily collected from the Manaaki Whenua Landcare Research, Animal Facility in Lincoln, Canterbury New Zealand. Secondary recordings were collected from Governors Bay, Canterbury at Living Springs and were provided by The Cacophony Project. Secondary recordings were only used for spectral profiling and to inform study design (microphone testing, recording duration) at the Lincoln recording location. Field recording collection began in June 2021 and continued until January 2023. Three AudioMoth microphones (Hill et al. 2019) were used with recording bandwidths up to 48 kHz and with a recording interval varying from 7pm to 7am in the winter and 9pm to 5am in the summer. Field recordings are five minutes in length and recorded at

DOI: https://doi.org/10.20417/nzjecol.48.3566

five-minute intervals using a 50% recording duty cycle (Fig. 1). Audio was collected from multiple locations within the facility spaced roughly 50 m apart. The position of the microphones was constrained by the location of target species, initially possums, mustelids, rats, and cats. Microphones remained at these locations for the duration of the study. Initially, devices were set up at the Animal Facility at a recording bandwidth of 250 kHz for a period of two weeks. Spectral profiling of target species was completed, and the bandwidth was adjusted to 48 kHz for subsequent data collection. The format of all audio files was standardised. All raw field recordings are single-channel and sampled at 48 kHz. Field recordings were stored using waveform audio file format (WAV). This is an uncompressed format and was selected for its higher quality and standardised support across a number of platforms. WAV format does result in larger file sizes, each five-minute field recording is 28.8 MB. Annotated segments have been downsampled to 16 kHz for the application of possum detection.

Long-term data collection found that the target species are mostly active (vocalising) during the evening and morning demonstrating crepuscular characteristics. Given changes in daylight hours across the year, recording periods were adjusted accordingly. These observations are not necessarily representative of the wider population given the small sample size and captive environment. We also observed higher vocalisation rates at the Lincoln recording location compared to data collected at other field locations such as Governors Bay, Canterbury as well as recordings shared by the Cacophony Project from various locations.

Dataset

All raw field recordings and annotated segments are available through the public Kaggle (McEwen et al. 2024). In total, 1236 five-minute field recordings are included in the dataset. These recordings were collected between June 2022 and January 2023. We comment on audio collected outside of this time frame (Governors Bay and Cacophony Project audio) but it is not included due to differences in bandwidth, microphone configuration and location. We use a wavelet-based segmentation approach to extract features of interest from the raw field recordings resulting in 3500 5 second segments. This segmentation approach is based on a wavelet packet decomposition (WPD) approach (Priyadarshani et al. 2020) and applied using the Listening Lab annotator (McEwen et al. 2023), an open-source platform developed to analyse sparse acoustic features. Each of these segments were manually analysed and labelled as possum or noise. Wavelet packet decomposition is used as an efficient data reduction step hence recall is prioritised, as the true detection of positive features outweighs the introduction of false positives which are reanalysed later using the classification model. This results in precision/recall results of 0.365/0.861 respectively. Note that annotations are for binary classification e.g. possum absence/presence therefore these classes, particularly absence, contain unlabelled sub-classes such as wind, rain, traffic, and non-target vocalisations (birds, farm animals). The 3500 segments are separated into training (3000 samples) and test (500 samples) sets. The training set contains 1130 possum vocalisations and 1870 noise sources. The test set contains 187 possum vocalisations and 313 noise sources. Both the training and test sets contain imbalanced data that reflects the output of the segmentation stage. The occurrence of possum vocalisations across the duration of the study was not temporally constant hence the training and test sets have been uniformly sampled across the entire set to achieve a consistent proportion of possum to non-possum segments.

Model evaluation

Using the invasive species dataset we fine-tune a pre-trained transformer-based model Audio Spectrogram Transformer (AST) (Gong et al. 2021a). Hyperparameter selection is based on the authors' recommendations (Gong et al. 2021b). The model was trained for 25 epochs with model achieving the highest validation accuracy stored. An Adam optimiser with an initial learning rate of 1e-5 was used and halved every five epochs after the first ten epochs. Model training is fully-supervised using the training set and evaluated using the test set. AST is pre-trained on 527 audio classes and therefore the default output is an embedding of 527 elements. This model takes a spectrogram as an input. To satisfy the model input requirements and due to the bandwidth of the target species, input audio is downsampled to 16 kHz. As we are training the model for a binary classification task, the embeddings of AST are fed into a secondary fully connected network (FCN) consisting of two layers with 100 and 2 neurons respectively as a simple transfer learning task (Fig. 2).

We also evaluate a commonly applied pre-trained ResNet-50 architecture (He et al. 2016). Dufourq et al. (Dufourq et al. 2022) also demonstrate the strong classification performance and low data requirements of ResNet-50 when comparing alternative pre-trained Convolutional neural network -based models. With the rapid development of transformer-based models, we also evaluate two transformer models AST (Gong et al. 2021a) which, like ResNet, operates

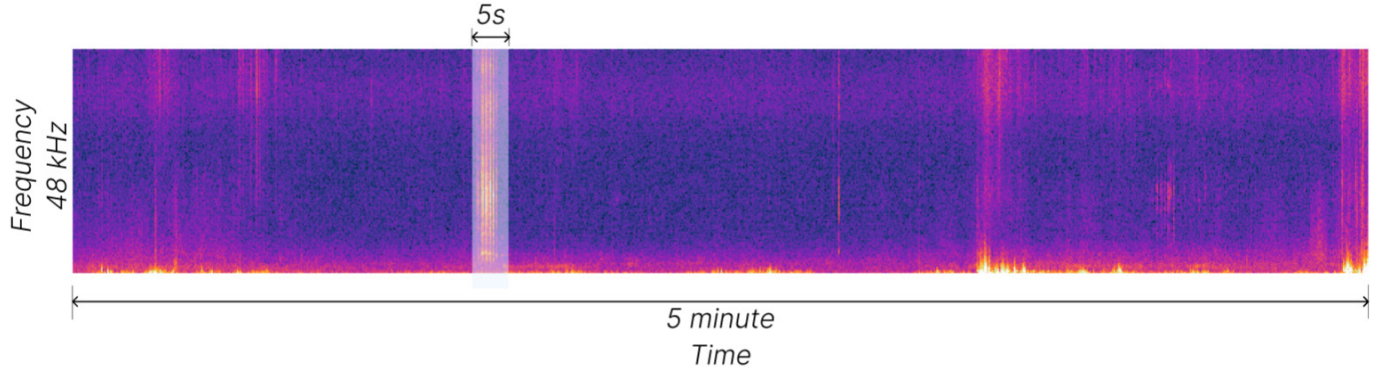

Figure 1. Example of five-minute field recording

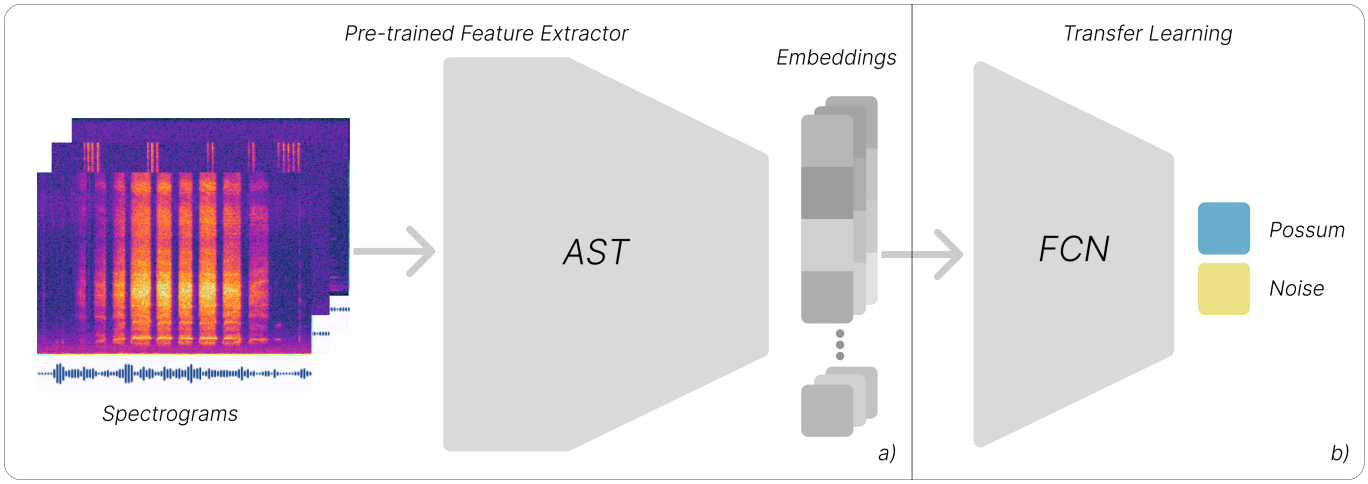

Figure 2. Classification pipeline developed for possum detection using pre-trained Audio Spectrogram Transformer (AST) model and fully connected neural network (FCN) for transfer learning.

on the spectrogram. We also evaluate HuBERT (Hsu et al. 2021) which operates directly on the 1-D waveform. Hyperparameter selection for ResNet-50 (He et al. 2016) and HuBERT (Hsu et al. 2021) was based on the recommendations of the model authors. Due to the imbalanced datasets, F1 score as well as precision and recall metrics are provided when evaluating classification models.

Results

We present observations from the collected field recordings including spectral profiling of common possum vocalisations and other invasive species. We also discuss the potential conservation benefits and challenges of applying bioacoustic monitoring to invasive species in New Zealand.

Spectral profiling

Initial testing using a maximum bandwidth of 250 kHz enabled the analysis of the target species' full spectral band. Spectral profiles of three common vocalisation types were found. The spectrograms and spectral profiling show temporal and spectral features up to 24 kHz (sampling at 48 kHz; Fig. 3). A spectrogram is a time-frequency representation generated using the short-time Fourier transform (Sejdić et al. 2009) displaying signal amplitude over time and frequency. The spectral profile is generated using a Fourier transform and displays amplitude (decibels) at logarithmically spaced frequency (Hz). Both spectrograms and spectral profiles are generated using Audacity® (Audacity 2023).

The most common vocalisation, chitter (Fig. 3a), has consistent peaks at 2 and 8 kHz and ranges in length from 1–5 s. Based on observations (Kean 1967) this vocalisation corresponds to social calls. This call contains distinct and repeated impulses, 0.2–0.5 s in length. The next most common vocalisation, screech (Fig. 3b), was also common and is characterised by a distinct peak at 2 kHz that taper off at higher frequencies. These vocalisations were shorter, ranging from 1–2 s. These vocalisations align with Kean's observations of aggressive behaviour (Kean 1967). We also note a third category, grunt (Fig. 3c). This vocalisation is characterised by a short, transient feature, less than 0.5 s in length. We observed features extending up to 24 kHz. This vocalisation may also correspond to aggression or disturbance (Kean 1967).

There is limited literature characterising the vocalisations of possums. Researchers have observed hearing sensitivity increasing from 2–15 kHz and continuing up to 35 kHz (Osugi et al. 2011). These observations align with microelectrode mapping (Gates & Aitkin 1982). The results of spectral profiling agree with the literature with key features occurring from 2–15 kHz.

Classification

We demonstrate the use of this dataset to fine-tune a pre-trained model in a simple binary classification transfer learning task. Convolutional neural network (CNN) based audio classification remains a common approach. Audio representations such as spectrograms or mel-frequency cepstral coefficients (Abdul 2022) can be represented as a single-channel image. We evaluate the models on a two-class binary output presence/absence of possum. Using the test dataset, we compare the test accuracy for each model. Due to the test set being unbalanced F1 score is referenced. Both transformer models outperformed ResNet-50 which achieved an F1 score of 0.945. AST marginally outperformed HuBERT with an F1 score of 0.983 and 0.945 respectively (Table 1).

Vocalisation three (grunt) was commonly misidentified. For evaluation of AST on the test set, 44% of incorrectly classified segments containing this vocalisation subclass. Other misidentification included false-positive identification of bird song (30%). The remaining misidentifications contained low signal-to-noise ratio features with either high background noise or low-volume possum vocalisations.

Discussion

Potential and challenges

Computational bioacoustics is a rapidly advancing field, allowing monitoring of cryptic species, previously infeasible to monitor. PAM devices are generally an affordable monitoring option compared other mode such as manual monitoring and

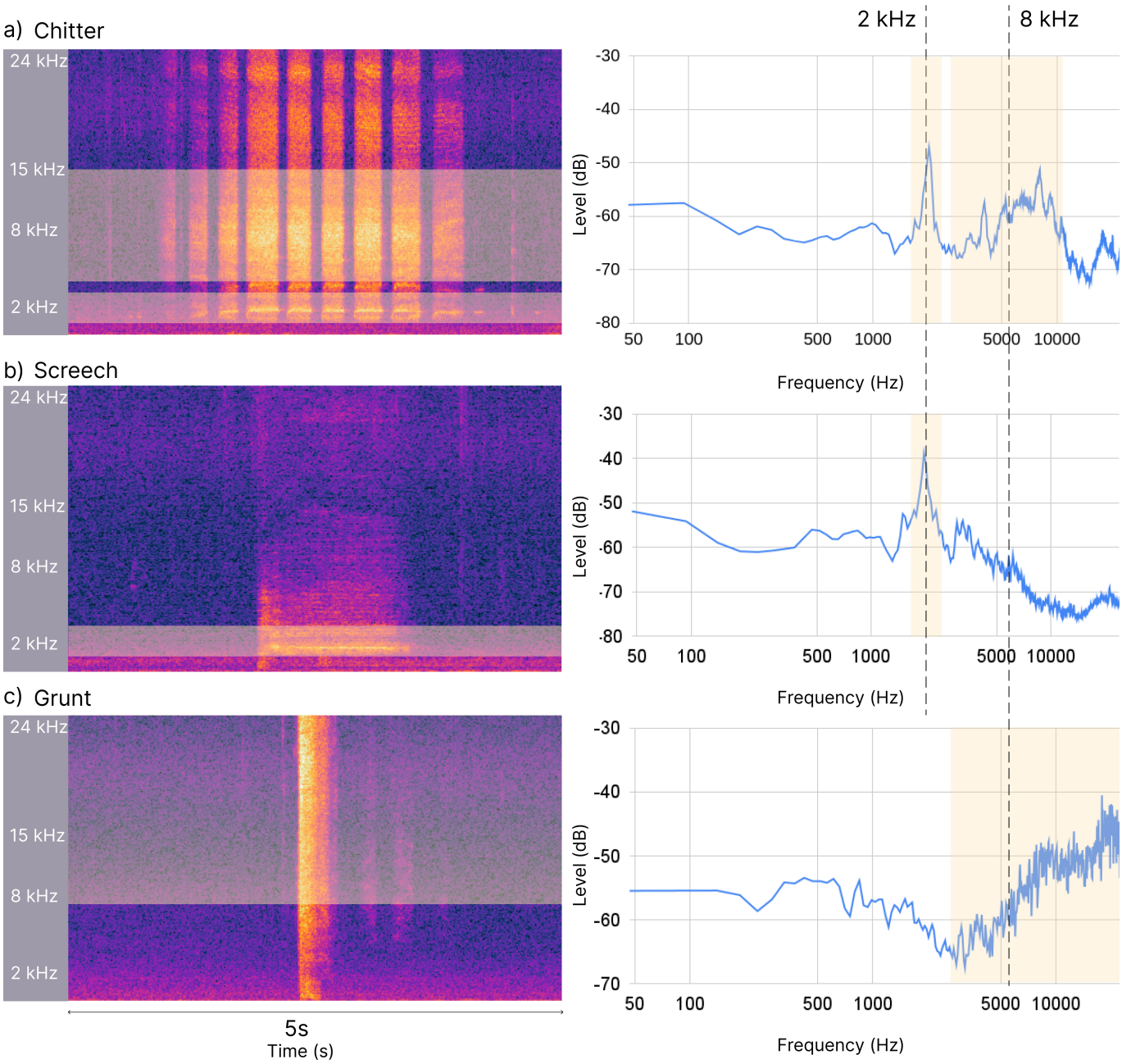

Figure 3. Spectral profiling of common possum vocalisations showing spectrogram (left) and spectral profile (right) a) chitter (2–10 kHz) b) screech (2 kHz) c) grunt (2–24 kHz). The yellow regions of the spectral profile denote corresponding spectral features, and the dashed line represents the frequency at which the feature occurs.

camera-based detection. However, there are key challenges that affect bioacoustic monitoring feasibility including bandwidth and feature sparsity.

Bandwidth

Vocalisations of possums have key spectral features within human hearing ranges 16–22 kHz. There is potential for standard PAM tools to be used for monitoring using commonly used bandwidths. This is not the case for all invasive species. Other species such as rats vocalise at higher frequency than other common sound profiles and harmonics. This reduces spectral overlap with other sound sources but comes with practical challenges. Recording at higher frequencies requires higher bandwidths, above that of many affordable devices. Recording at this level results in higher power consumption and larger data sizes (more storage requirements). In addition to this, existing pre-trained models are generally developed for natural language processing (NLP) applications and therefore biased towards human-perceivable sounds limited to 16–22 kHz. This is therefore a challenge when applying pre-trained models to bioacoustic applications with higher bandwidth audio.

Feature sparsity

Another key challenge for this application and other bioacoustic applications focused on less vocal species is feature sparsity.

Table 1. Evaluation of models ResNet-50, audio spectrogram transformer (AST) and HuBERT using the test dataset.

Model	Accuracy (%)	F1	Precision	Recall
ResNet-50	90.2	0.892	0.882	0.908
AST	98.4	0.983	0.982	0.984
HuBERT	94.8	0.945	0.948	0.942

At the primary recording location with captive animals, only 0.4% of the audio collected contains features of interest (invasive species vocalisations). For non-captive animals at lower densities, this is likely to be significantly lower. Feature sparsity combined with monitoring at a landscape scale accentuates the data imbalance challenges such as the amount of raw audio collection and analysis required for meaningful information to be extracted (i.e. occupancy, abundancy etc). Understanding how often features occur within data is useful when considering data collection requirements, human labour/analysis costs, and modelling feasibility. Understanding the point at which a challenging bioacoustics application becomes an infeasible one is an important but challenging question. The feasibility of bioacoustic monitoring needs to be evaluated with spatial, temporal and behavioural considerations in mind as well as resource and human labour costs. Not all species are well-suited to bioacoustic monitoring and other monitoring tools are available.

Applying computational tools, such as automatic segmentation and classification, can improve the feasibility of monitoring even at low densities. For example, we demonstrate model training using a fully-supervised learning approach. We applied low-data requirement computational tools such as WPD segmentation (Priyadarshani et al. 2020) and few-shot learning approaches (McEwen et al. 2023) to aid in analysis. These approaches are currently under-utilised (Hoefer et al. 2023) with only 17% of studies applying computational approaches.

Conclusion

We present the development of the first publicly available audio dataset of the common brushtail possum (*Trichosurus vulpecula*) as well as the development of an automatic classification model. Evaluation of this model demonstrates high classification performance, this model can be applied to aid in the detection of sparse possum vocalisations within long-term field recordings. The development of computational tools for, typically less vocal, invasive species comes with a number of challenges including feature sparsity and recording frequency. The feasibility of bioacoustic monitoring needs to be accompanied by spatial, temporal (seasonal), and behavioural (across age groups and sexes) considerations including further research into distance-to-detection and call attenuation. The development of alternative datasets containing labelled possum vocalisations would also allow the model to be evaluated in terms of generalisability. We hope that this new dataset and method will aid the broader conservation community in the continued development of improved detection tools, particularly for species that may be too costly to monitor previously.

Acknowledgements

We thank the Cacophony Project for their support and also Manaaki Whenua, Landcare Research for providing access to their facilities and staff support. We also thank University of Canterbury students Kaspar Soltero, Isaac Cone, and Mikayla Franco for their support with data collection and annotation. Many thanks to the editor and reviewers for their thorough reviews and suggestions.

Additional information and declarations

Conflicts of interest: The authors declare no conflicts of interest.

Funding: This project was made possible through the Predator Free 2050 Limited Capabilities Development Funding as well as the Forest and Bird Stocker Scholarship.

Ethics: No ethics approval was required for this study.

Data availability: All data is provided under the Creative Commons (CCBY) Attribution 4.0 licence. The classification model, initially developed by Gong et al. (2021a), was modified for this application under the BSD 3-Clause and retains this license. This allows all data and the model to be shared and adapted by the public for free. All data is accessible through the public Kaggle dataset https://www.kaggle.com/datasets/benmcewen1/invasive-species. We also provide a notebook that presents the classification model implementation and demonstrates use of the dataset https://www.kaggle.com/datasets/benmcewen1/invasive-species/code.

Author Contributions: Ben McEwen: dataset preparation, analysis, writing. Andrew Bainbridge-Smith, Stefanie Gutschmidt, James Atlas, Richard Green: technical support, supervision and review of the manuscript.

References

Abdul Z, Abdulbasit AK 2022. Mel-frequency cepstral coefficient and its applications: A review. IEEE Access 10: 122136–122158

Audacity Team 2023. Free audio editor and recorder [computer application]. version 3.5.1 https://audacityteam.org/ (Accessed June 2022).

Bombaci S, Pejchar L 2018. Using paired acoustic sampling to enhance population monitoring of New Zealand's forest birds. New Zealand Journal of Ecology, 43(1): 3356.

Darras, KF, Batáry P, Furnas BJ, Grass I, Mulyani Y, Tscharntke T 2019. Autonomous sound recording outperforms human observation for sampling birds: a systematic map and user guide. Ecological applications 29(6): e01954.

Davidson NB, Hurst JL 2019. Testing the potential of 50 kHz

rat calls as a species-specific rat attractant. PLoS ONE 14(4): e0211601.
Dufourq E, Batist C, Foquet R, Durbach I 2022. Passive acoustic monitoring of animal populations with transfer learning. Ecological Informatics 70: 101688.
Gates GR, Aitkin LM 1982. Auditory cortex in the marsupial possum *Trichosurus vulpecula*. Hearing research 7(1): 1–11.
Gong Y, Chung YA, Glass J 2021a. AST: Audio spectrogram transformer. Interspeech Proceedings 2021: 571–575.
Gong Y, Chung YA, Glass J 2021b. PSLA: Improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29: 3292–3306.
He K, Zhang X, Ren S, Sun J 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 770–778.
Hill A, Prince P, Snaddon J, Doncaster C, Rogers A 2019. AudioMoth: A low-cost acoustic device for monitoring biodiversity and the environment. HardwareX 6: e00073.
Hoefer S, McKnight DT, Allen-Ankins S, Nordberg EJ, Schwarzkopf L 2023. Passive acoustic monitoring in terrestrial vertebrates: a review. Bioacoustics 32(5): 506–531.
Hsu WN, Bolte B, Tsai YH, Lakhotia K, Salakhutdinov R, Mohamed A 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29: 3451–3460.
Jahn P, Ross J, MacKenzie D, Molles L 2022. Acoustic monitoring and occupancy analysis: cost-effective tools in reintroduction programmes for roroa-great spotted kiwi. New Zealand Journal of Ecology 46(1): 3466.
Kean RI 1967. Behaviour and territorialism in *Trichosurus vulpecula* (Marsupialia). Proceedings of the New Zealand Ecological Society 14: 71–78.
McEwen B, Soltero K, Gutschmidt S, Bainbridge-Smith A, Atlas J, Green R 2023. An improved computational bioacoustic monitoring approach for detecting sparse features. The Journal of the Acoustical Society of America 154: 143–143.
McEwen B, Bainbridge-Smith A, Gutschmidt S, Green R, Atlas J 2024. Invasive Species NZ, Kaggle. https://www.kaggle.com/datasets/benmcewen1/invasive-species.
Mortimer JA, Greene T 2017. Investigating bird call identification uncertainty using data from processed audio recordings. New Zealand Journal of Ecology 41(1): 126–133.
Osugi M, Foster TM, Temple W, Poling A 2011. Behavior-based assessment of the auditory abilities of brushtail possums. Journal of the Experimental Analysis of Behavior 96(1): 123–138.
Priyadarshani N, Marsland S, Castro I, Punchihewa A 2016. Birdsong denoising using wavelets. PLOS ONE 11(1): e0146790.
Priyadarshani N, Marsland S, Juodakis J, Castro I, Listanti V 2020. Wavelet filters for automated recognition of birdsong in long-time field recordings. Methods in Ecology and Evolution 11(3): 403–417.
Ross S, O'Connell D, Deichmann J, Desjonquères C, Gasc A, Phillips J, Burivalova Z 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Functional Ecology 37(4): 959–975.
Sejdić E, Djurović I, Jiang J 2009. Time-frequency feature representation using energy concentration: An overview of recent advances. Digital Signal Processing 19(1): 153–183.
Stowell D 2021. Computational bioacoustics with deep learning: a review and roadmap. PeerJ 10: e13152.
Stowell D, Stylianou Y, Wood M, Pamula H, Glotin H 2018. Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge. Methods in Ecology and Evolution 10(3): 368–380.
Sugai LS, Silva TS, Ribeiro Jr JW, Llusia D 2019. Terrestrial passive acoustic monitoring: review and perspectives. BioScience 69(1): 15–25.

Received: 21 November 2023; accepted: 22 February 2024
Editorial board member: Zachary Carter