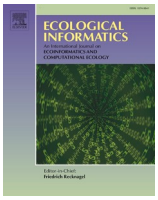
Ecological Informatics 83 (2024) 102851

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Improving acoustic species identification using data augmentation within a deep learning framework

Jennifer MacIsaac [a,b,*], Stuart Newson [a], Adham Ashton-Butt [a,c,d], Huma Pearce [a], Ben Milner [b]

[a] British Trust for Ornithology, The Nunnery, Thetford, Norfolk IP24 2PU, UK
[b] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, Norfolk NR4 7TJ, UK
[c] School of Biology, Univ. of Leeds, Leeds, UK
[d] School of Environmental Sciences, University of East Anglia, Norwich, UK

ARTICLE INFO

Keywords:
Bioacoustics
Convolutional neural networks
Data augmentation
Deep learning
Passive acoustic monitoring
Small mammals

ABSTRACT

Convolutional neural networks (CNNs) are effective tools for acoustic classification tasks such as species identification. Large datasets of labelled recordings are required to develop CNN classifiers which can be difficult to obtain, particularly if species are rare or vocalise infrequently. Additionally, data often requires manual labelling which can be time consuming requiring expert analysis. Artificially generating data using augmentation can address these challenges, however the impact of data augmentation on CNN performance is poorly understood and often omitted in bioacoustic studies. Here, we empirically test the impact of CNN architecture and 20 data augmentation methods on classifier performance. We use acoustic identification of 18 small mammal species as a case study of a species group that can be effectively surveyed by acoustic monitoring, but recordings for training data are scarce and difficult to collect. Networks that achieved the highest accuracy across all sample sizes was a 10-layer CNN (96.43 %) and a pre-trained ResNet50 model (96.37 %). Overall, all augmentation effects improved ResNet50 model performance and 17 effects improved Conv10 performance, increasing relative change in accuracy (RCA) by 0.021–0.641. Three augmentation effects negatively impacted Conv10 RCA by −0.042 to −0.182. We also show that adding augmented data when the number of original samples is low has the greatest positive impact on accuracy and this effect was larger with ResNet50 models. Our work demonstrates that using data augmentation where few original samples are available can considerably improve model performance and highlights the potential of augmentation in developing acoustic classifiers for species where data are limited or difficult to obtain.

1. Introduction

Passive acoustic monitoring (PAM) is a valuable tool for ecology and conservation to detect vocalising species that are difficult to observe visually and for surveying at greater geographical and temporal scales, reducing the need for trained personnel and time spent in the field (Gibb et al., 2019; Newson et al., 2017; Thomas et al., 2017). Recent improvements in acoustic hardware in terms of power requirements and memory capacity have increased the accessibility of acoustic sensors and therefore the scope of acoustic surveys, enabling the implementation of long term, landscape scale monitoring programs (Gibb et al., 2019; Prince et al., 2019). Surveys of this magnitude can generate large quantities of data comprising many hours of recordings; a single landscape-scale acoustic survey can generate several terabytes of data. Manual data analysis would be labour intensive and time consuming, eliminating many of the advantages that PAM offers (Florentin et al., 2020). Developing partially or fully automated data processing and analysis pipelines is therefore key to ensuring that PAM remains viable for large scale ecological surveying.

During the last decade machine learning algorithms, for example support vector machines and Random Forest, have been used to develop acoustic classifiers capable of identifying a range of species' vocalisations for primates (Clink and Klinck, 2021; Heinicke et al., 2015), shrews (Zsebők et al., 2015), bats (Ayala-Berdon et al., 2020), bush crickets (Newson et al., 2017) birds (Sebastián-González et al., 2015) and elephants (Zeppelzauer et al., 2015). Deep learning, a subfield of machine learning that uses neural networks to extract features to identify patterns and relationships within highly dimensional data (LeCun

* Corresponding author at: British Trust for Ornithology, The Nunnery, Thetford, Norfolk IP24 2PU, UK.
E-mail address: jennifer.macisaac@bto.org (J. MacIsaac).

https://doi.org/10.1016/j.ecoinf.2024.102851
Received 30 April 2024; Received in revised form 7 October 2024; Accepted 7 October 2024
Available online 15 October 2024

et al., 2015) has gained renewed interest in recent years following the development of graphics processing units (GPUs) capable of training complex neural networks which has extended into ecological studies (Christin et al., 2019).

Convolutional neural networks are perhaps the most widely used neural network in ecological studies (Borowiec et al., 2022; Christin et al., 2019) and have been developed to support a number of ecological applications including habitat classification (Abrams et al., 2019), animal behaviour prediction (Browning et al., 2018) and tree species identification (Allen et al., 2023). Although CNNs are primarily used in computer vision tasks, they have proven to be effective in acoustic classification tasks using spectrograms as input (Bermant et al., 2019; Chen et al., 2020; Mac Aodha et al., 2018; Ravaglia et al., 2023; Zhong et al., 2021). An advantage of neural networks compared to classic machine learning algorithms is that it does not require identification and description of features relevant for the classification task (feature extraction) by the developer prior to the training process (Chollet, 2018). Manual feature extraction can be a time consuming and subjective process often requiring expert analysis (Borowiec et al., 2022). Instead, features are extracted during training using weighted filters to create feature maps. These weights are then adjusted during training to reduce loss, (the difference between the true class and the predicted class), increasing classifier performance by highlighting features relevant to the task (Chollet, 2018). In order to ensure that features relevant for the classification task are extracted and unwanted signals ('noise') are ignored, a large number of training examples are required. In PAM studies involving common species that vocalise frequently at high amplitudes it may be straightforward to obtain recordings particularly if their vocal repertoire is well known and easy to distinguish (Zhong et al., 2020). In studies involving rare or elusive species this can prove challenging as vocalising individuals may be difficult to locate, requiring extensive field time to record a range of individuals to ensure there is both a sufficient number of recordings and variation in the training data (Zhong et al., 2021). Additionally, there may be no prior knowledge of the target species vocalisations making them difficult to isolate in recordings without other means of identification and it may be necessary to record individuals in captivity (Newson et al., 2020).

A solution commonly employed in computer vision tasks is to use data augmentation to artificially generate data where there are insufficient samples in a class. Data augmentation takes the original sample and applies a transformation to produce an additional training sample that is an altered version of the original whilst retaining characteristics relevant to that class (Nanni et al., 2020; Shorten and Khoshgoftaar, 2019). Augmenting visual data usually involves image manipulation (e.g. rotation, reflection) however for audio processing, adjustments to the spectrogram can be made using time shifting or masking to generate augmented data. Additionally, sound manipulation techniques such as pitch shifting and stretching can be used in the time-domain, before spectrogram extraction, potentially increasing the range of variation that can be added to training data. (Nanni et al., 2020; Shorten and Khoshgoftaar, 2019). Although data augmentation can increase the amount of training data available, some methods may not be suitable for bioacoustic studies as features required for classification may be masked (Salamon and Bello, 2017).

Several CNN based acoustic studies have used a range of data augmentation techniques to boost sample sizes, however detailed analyses of the impact of augmentation methods on bioacoustic classification tasks is often lacking or limited in scope. Augmentation methods commonly used include pitch and time shifting, time stretching, frequency and time masking, noise addition (e.g. background and white noise) and combining sample spectrograms (Dufourq et al., 2021; Goussha et al., 2022; Lostanlen et al., 2019; Nanni et al., 2020; Nshimiyimana, 2024; Sprengel et al., 2016). Sprengel et al. (2016) used a combination of time and pitch shifting, background noise addition and spectrogram combination which increased mean average precision of bird species detection by 7.4 %. Similarly Lostanlen et al. (2019) developed a classifier for detecting bird migration calls and used pitch shifting and time stretching to improve precision and recall. Dufourq et al. (2021) augmented data using time shifting and background noise addition during the development of a Hainan gibbon (*Nomascus hainanus*) CNN acoustic classifier and found that using augmented data significantly improved accuracy, specificity and precision however sensitivity was reduced by *c.* 2 %. A study of augmentation methods for bioacoustic data investigated the impact of four augmentation protocols on mean accuracy (Nanni et al., 2020). Three augmentation protocols improved mean accuracy, however protocols that involved standard image transformations such as image rotation and reflection, resulted in decreased model performance (Nanni et al., 2020). Although this study highlighted the importance of selecting augmentation methods suitable for acoustic data, the overall impact of each augmentation technique was not investigated. Manriquez et al., (2024) and Su et al., (2024) also found that using data augmentation to increase the number of training samples improved accuracy. Contrastingly, both studies found conflicting results regarding Gaussian noise; Manriquez et al. (2024) demonstrated that adding noise improved accuracy but adding more than 15 % of the recordings' amplitude negatively impacted performance. Conversely, Su et al. (2024) showed that reducing the signal to noise ratio by adding Gaussian noise corresponded to a decrease in model performance. Nshimiyimana (2024) evaluated CNN performance using data augmented using five different methods; the results indicated that adding pink noise improved performance even when the number of original samples was low but adding random noise had the greatest negative impact. It is clear from these studies that data augmentation can improve classifier performance but the results are conflicting in terms of which augmentation method is the most effective.

Here we use small mammal species as a case study to investigate the use of data augmentation to supplement training data for species where the availability of acoustic data are limited or difficult to obtain. We define small mammals according to the International Biological Programme as mammal species weighing up to 5 kg. Small mammal species perform a range of key functions in terrestrial ecosystems including seed dispersal and pollination as well as being an important food resource for predators (Benedek et al., 2021). Certain small mammals species e.g. *Muscardinus avellanareous* can also serve as indicator species due to their sensitivity to habitat loss and fragmentation (Goodwin et al., 2017). Popular survey techniques include live trapping, sign surveys (e.g. the Great Nut Hunt), nest box surveys and foot print tunnels however these methods can lack adequate time resolution and may also be invasive and resource intensive in terms of field time and trained personnel (Goodwin et al., 2017; Mills et al., 2016). Live trapping can also carry a high mortality risk particularly for species with fast metabolisms (e.g. *Sorex* spp., (Shonfield et al., 2013)). This limits the scalability of most small mammal survey methods, however many small mammal species produce a range of vocalisations which provides an opportunity for PAM methods to be developed for this taxa. (Ancillotto et al., 2014; Ancillotto and Russo, 2016; Middleton et al., 2023; Newson et al., 2020; Zsebők et al., 2015). Futhermore, the availability of a multi-species acoustic classifier would enable the implementation of large scale and long term PAM programmes which could provide baseline data and valuable insights into the spacial ecology of small mammals species.

Small mammals have a varied vocal repertoire both in terms of function and acoustic features (Ancillotto et al., 2017; Middleton et al., 2023; Newson et al., 2020). Additionally, many species emit species-specific vocalisations that can be used for identification due to the differences in call properties (e.g. frequency, duration and bandwidth) (Ancillotto et al., 2017; Middleton et al., 2023; Newson et al., 2020; Zsebők et al., 2015). Vocalisations can be audible, for example *Apodemus spp* distress calls (Ancillotto et al., 2017), however the species in this study emit predominantly ultrasonic calls (> 20 kHz), which are inaudible to humans (see supplementary information Figs. 1–4 for example spectrograms). An exception is *Glis glis* which vocalise in lower frequencies, producing a range of audible calls, particularly during mating

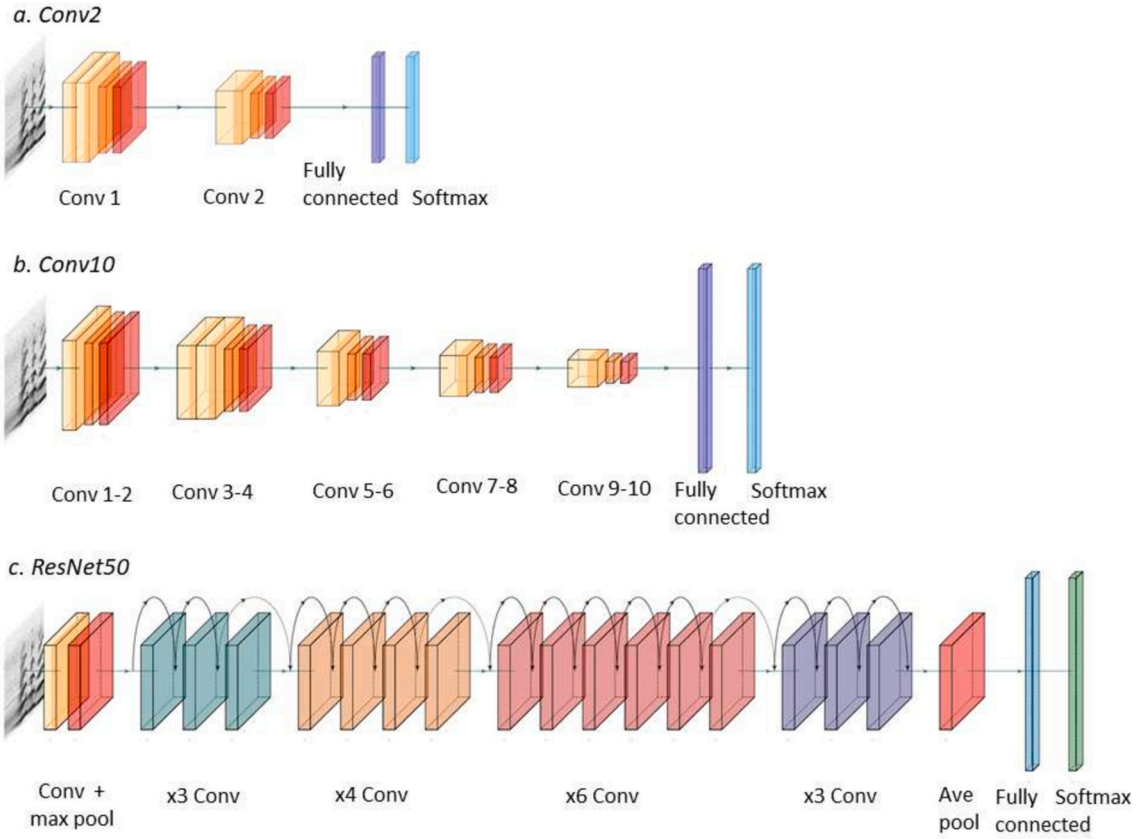


Fig. 1. Schematic representations of CNN architecture: a) Conv2 network consists of two blocks of a single convolution layer proceeded by a max pooling layer, 0.25 dropout is applied before the next convolutional block, the final two layers are fully connected (dense) layers; b) the Conv10 model consists of five blocks of 2 convolution layers followed by max pooling and dropout layers; c) ResNet50 model; the first layer is a convolutional layer with 64 7 × 7 filters proceeded by a max pool layer. Each layer represents a residual block of three convolution layers and the arrows represent the shortcut connections between residual blocks. The final two layers are fully connected layers.

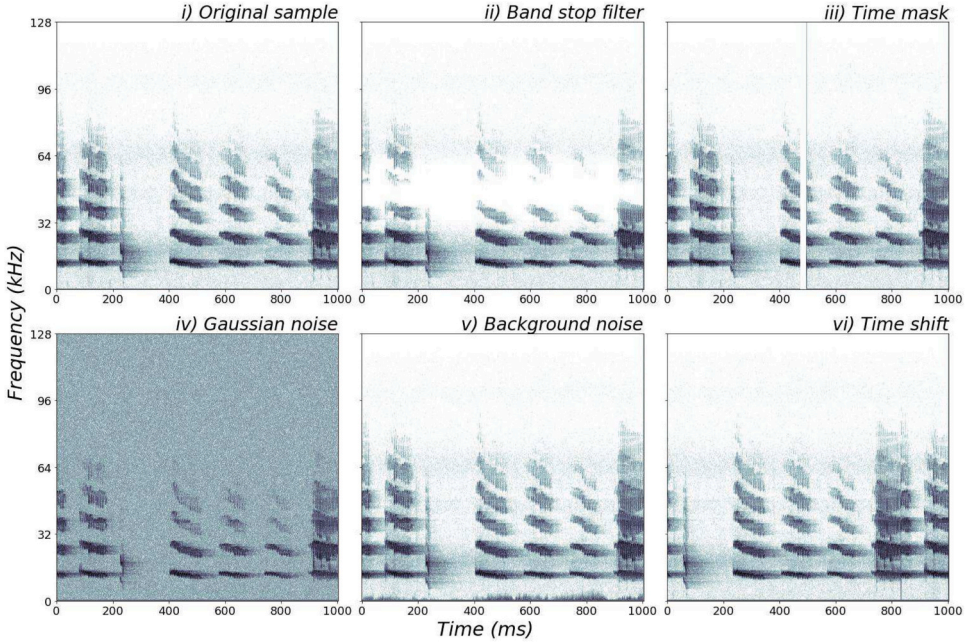


Fig. 2. Spectrograms showing a 1 s recording of *S. araneous*. *i.* shows the spectrogram of the unaltered recording. *ii.* shows the effect of band stop filter which masks a random frequency band; *iii.*) shows the impact of time mask which removes a section of the spectrogram along the time axis. Spectrogram *iv.* has Gaussian noise added, *v.* shows the effect of background noise addition which is the combination of the original wav file with another 1 s wav file containing non-target sounds, spectrogram *vi.* shows how time shift alters the image by shifting the spectrogram along the x-axis.

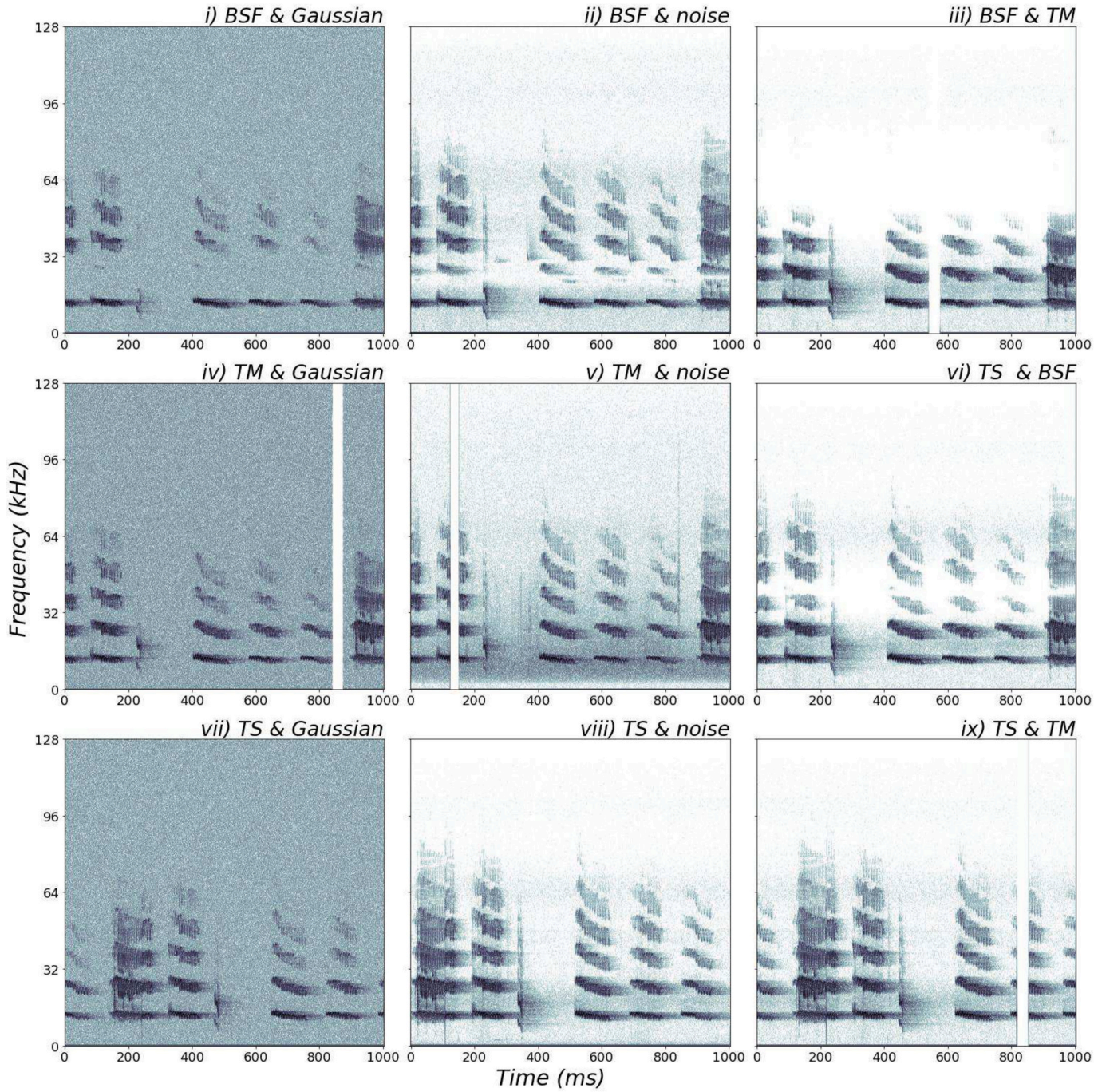


Fig. 3. Spectrograms of a 1 s recording of *S. araneous* with two augmentation methods applied. *i.* the original spectrogram with a band stop filter applied and Gaussian noise added; *ii.* a band stop filter is applied after background noise is added; *iii.* a band stop filter and time mask are added so that sections of the spectrogram on the x and y axes are removed; *iv.* Gaussian noise is added followed by a time mask; *v.* background noise is added before applying a time mask; *vi.* the spectrogram is shifted along the time axis and a band stop filter applied; *vii.* Gaussian noise is added followed by time shift; *viii.* Background noise is added and before time shift; *ix.* time shift is applied and time mask added.

season (Middleton et al., 2023). Although the vocal behaviour of laboratory rodents has been well studied, including recent works involving the development of deep learning acoustic classifiers (Coffey et al., 2019; Goussha et al., 2022), there have been fewer studies focusing on wild species.

Our objectives were three-fold; first we investigate the impact of network structure and training sample size on the acoustic classification of small mammal species. We achieve this by training CNNs with increasing numbers of convolutional layers and training samples. Secondly, we address the issue of limited training data availability by developing 20 data augmentation techniques. The impact of each augmentation method on classifier performance is determined by training CNNs using augmented data and comparing model accuracy. We also test the effect of increasing the proportion of augmented training data on overall classifier performance by conducting a series of experiments in which CNNs are trained using data where novel samples are replaced by an increasing proportion of augmented samples. Thirdly, we investigate the impact of using data augmentation to balance an acoustic dataset of 18 small mammal species with an uneven number of recordings per species by comparing overall model F1 scores for CNNs

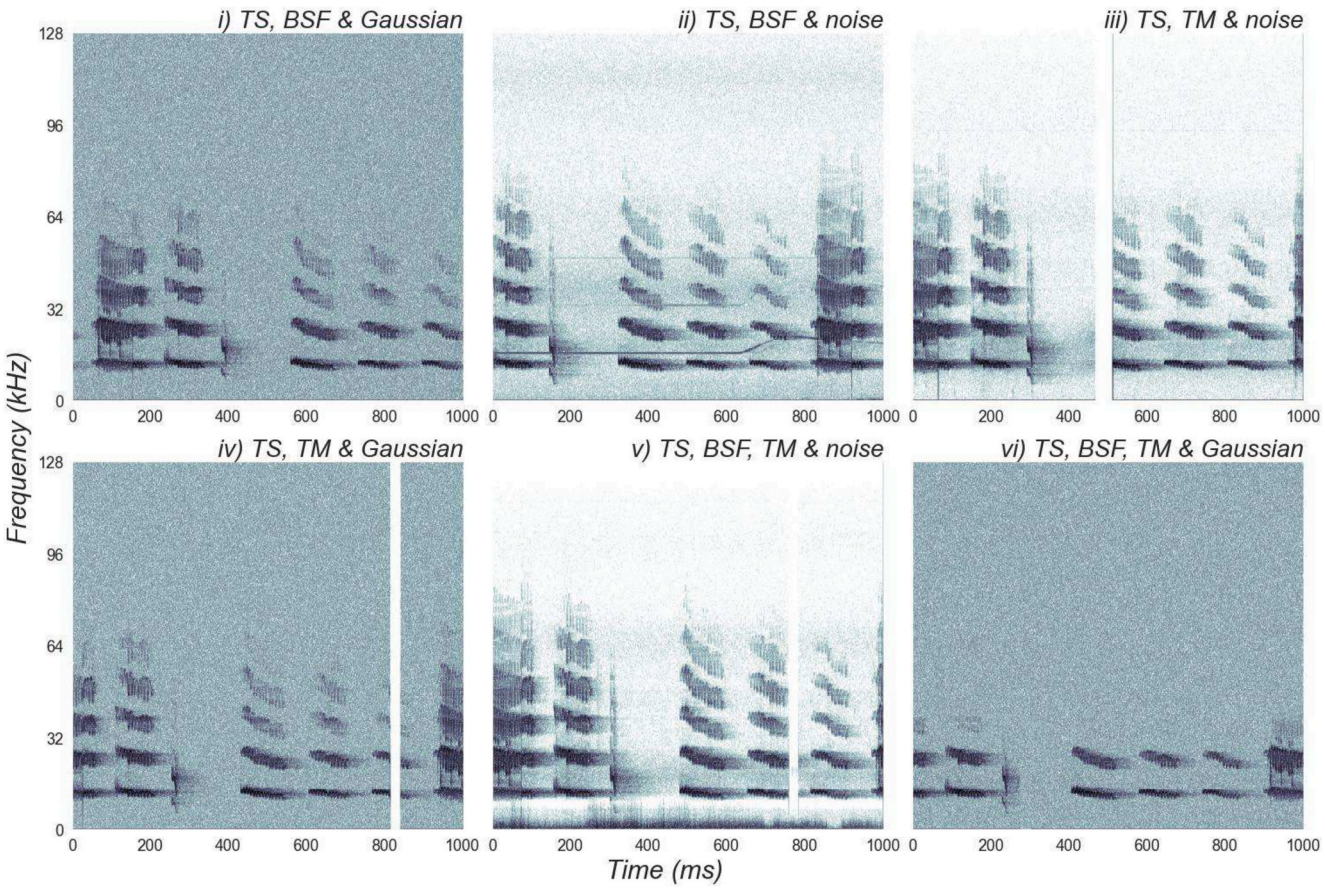


Fig. 4. Augmented spectrograms showing a 1 s recording of *S. araneous* where three or more augmentation effects are applied. *i.* the spectrogram is shifted along the time axis, a band stop filter is applied followed by Gaussian noise addition; *ii.* Similar to the previous spectrogram but background noise is applied instead of Gaussian noise; *iii.* a time shift transformation is applied, followed by adding background noise and a time mask; *iv.* A time shift transformation is applied followed by Gaussian noise and a time mask; *v.* background noise is added before applying a band stop filter and time mask; *vi.* a time shift transformation is applied followed by a band stop filter and adding Gaussian noise.

trained with an unbalanced dataset and a dataset balanced using data augmentation.

2. Materials and methods

2.1. Data collection

Sound recordings of small mammal species found in the UK were obtained from a number of sources of wild and captive individuals (Newson et al., 2020). Additional recordings of three species (*Apodemus agrarius*, *Dryomys nitedula* and *Glis glis*) were obtained during field visits to Turov Meadow, Pripyat Forest and Białowieża National Park, Belarus in July 2021, Budos-Pravieniškių miškų nature reserve, Lithuania in August 2022 and Hockeridge Wood, UK in August and September 2022. In Belarus, target species were captured using baited traps and transferred to a terrarium overnight before being released at their capture location. Vocalisations were recorded using SM4 Bat FS bat detectors (Wildlife acoustics, MA, USA) using a sampling rate of 256 kHz. Vocalisations were identified by visually inspecting sound files using SonoBat Viewer (UK). Recordings were obtained from four *Apodemus agrarius* individuals and a single *Dryomys nitedula*. Further *Dryomys nitedula* recordings were obtained in Budos-Pravieniškių miškų nature reserve, Lithuania by deploying acoustic sensors (SM4 minbat, (sampling rate 256 kHz), Audio Moths, (250 kHz)) and camera traps next to four nest boxes and two feeding stations. Vocalisations were identified using the timestamp of camera trap videos of *Dryomys nitedula* activity to isolate the corresponding acoustic file which was subsequently reviewed in Sonobat. Similarly *Glis glis* recordings were obtained in Hockeridge Wood during August and September 2022 using camera traps and acoustic sensors set up at three nest boxes. A total of 18 small mammal species and two non-target classes were included in the acoustic reference library with 467 to 7033 recordings per species (Table 1). Bat social calls from a broad range of over 45 European species were included as a separate class as they can be acoustically similar to small mammal vocalisations and a noise class containing background noise to minimise the occurrence of false positive identifications.

2.2. Data pre-processing

The acoustic library consists of files in wav format with varying lengths and sample rates, which requires pre-processing to ensure the input spectrograms are uniform in bandwidth and duration. Recordings with sample rates greater than 250 kHz were resampled to 256 kHz; files with sample rates less than 250 kHz were discarded. Each recording was sliced into one second segments using the labels created using Tadarida (Bas et al., 2017) which contain the start time of the call in each file. Small mammal vocalisations are predominantly short in duration and one second was considered sufficient to retain the defining characteristics whilst also ensuring that the maximum number of samples were extracted.

Over 130 h of recordings from more than 90,000 wav files were collected during 2021 fieldwork therefore to reduce the time required for manual processing a semi-automated process was employed. First, a subsection of recordings were visually inspected using SonoBat. Sonobat

Table 1
Small mammal acoustic dataset; number of one sample recordings per species extracted from the acoustic library. Species in bold were included in training data for the experiments in sections 2.5.1–2.5.3. Data for species marked with * were collected during 2021–2022 fieldwork.

Scientific Name	English name	No. 1 s samples	Species code
Rattus rattus	Brown rat	7033	Ratrat
Microymys minutus	Harvest mouse	5530	Micmin
Rattus norvegicus	Black rat	5303	Ratnor
Apodemus flavicollus	Yellow necked mouse	5187	Apofla
Apodemus sylvaticus	Wood mouse	2762	Aposyl
Mus Musculus	House mouse	2605	Musmus
Sorex araneus	Common shrew	2587	Sorara
Arvicola amphibius	Water vole	2491	Arvamp
Sorex minutus	Pygmy shrew	2152	Sormin
Muscardinus avellanarius	Hazel dormouse	2063	Musave
Crocidura leucodon	Bicoloured shrew	1921	Croleu
*Apodemus agrarius**	Striped field mouse	1760	Apoagr
Microtus agrestis	Field vole	1639	Micagr
Crocidura russula	Greater white-toothed shrew	1379	Crorus
*Glis glis**	Edible dormouse	1252	Gligli
Crocidura suaveolens	Lesser white-toothed shrew	978	Crosua
Neomys fodiens	Water shrew	851	Neofod
*Dryomus nitedula**	Forest dormouse	467	Drynit
Bat social calls	n/a	4718	Batsoc
Non-target sounds	n/a	3000	Noise

is designed for bat call analysis but can be used to view recordings of lower frequency sounds. Three or four vocalisations were selected as templates to mine the remaining data using the R package monitoR (Hafner and Katz, 2018). MonitoR was developed to detect and identify animal sounds using spectrogram template matching to identify target sounds within acoustic survey files (Hafner and Katz, 2018). Vocalisations used for templates were selected to represent the acoustic variation in each small mammal species' calls. The spectrogram template is used as a sliding window to score the survey file. Each score that exceeds a predefined threshold constitutes a positive detection and is recorded with the start time and filename. Although a high number of false identifications were generated, a low threshold value was selected (0.5) to minimise the risk of missing vocalisations in the recordings. The detection data were used to extract one second files containing potential vocalisations which were manually verified and added to the reference library. The reference library comprising one second samples for 18 small mammal species and two non-target classes (Table 1) was randomly split into training, validation and test datasets.

2.3. CNN architecture

The development of deep convolutional neural networks has transformed image classification tasks with deeper networks (networks with greater numbers of layers) generally outperforming shallower networks (He et al., 2016; Simonyan and Zisserman, 2015; Szegedy et al., 2015). Deep networks with many layers can be difficult to train due increased computational load and the 'vanishing gradient' problem (He et al., 2016; Szegedy et al., 2015). Several networks have overcome these issues; ResNet models can contain in excess of 50 layers (48 convolution layers, one average and one max pooling layer) and eliminate the vanishing gradient issue using residual blocks which skip layers by taking the output of one layer and passing it to another layer deeper in the network (He et al., 2016). Another important consideration during acoustic classifier development for PAM is its ability to generalise on unseen data. The use of deep CNNs may perform well in test conditions, achieving high performance metrics but could potentially lead to overfitting. Deeper networks can extract higher level features due to the additional convolutional layers, which can lead the model to learn features that are only present in the training data (Zhang et al., 2021). This could have negative consequences on its applicability for PAM where survey conditions may differ markedly from test conditions due to differences in environmental variables (e.g. temperature, vegetation cover, anthropogenic noise). In this study we use a ResNet50 model pretrained using the ImageNet dataset which consists of *c.* 14 million labelled images of everyday objects (Abadi et al., 2015) in addition to a ResNet50 model without pretraining, and networks consisting of increasing numbers of convolutional layers from 2 to 10 layers (Conv2 - Conv10) to determine the optimal CNN architecture and to assess whether simpler models (i.e. fewer convolutional layers) are better at generalising compared to a ResNet50 model (the number of trainable parameters for each model is provided in the Supplementary materials, Table 1.). The Conv models consist of blocks of one or two convolution layers followed by max pooling and 0.25 % dropout layers with two fully connected layers as the output layers (Fig. 1). As the depth of the Conv model increases, the number of filters for each convolution layer increases. For example, the first layer in Conv2 has 64 filters and the second layer has 128 filters (Fig. 1a). All filters have a kernel size of 3 × 3. The deepest network Conv10, starts with 32 filters in the first layer and increases to 256 filters in the final layer (Fig. 1b). The increase in the number of filters serves to extract more fine grained features (Chollet, 2018). ResNet50 consists of four groups of residual units comprising three convolutional layers (Fig. 1c). The residual units within each group has the same configuration in terms of filters and kernel size, with the final group containing 2048 1 × 1 filters (for further details see (He et al., 2016)). Models were trained for 100 epochs using an Adam optimizer with a learning rate of 0.0001. Relu activation functions were used with convolution layers and a Softmax activation function was used in the final fully connected layer.

2.4. Data augmentation

Many data augmentation methods used in computer vision tasks (e.g. image inversion, rotation etc) are unsuitable for acoustic classifiers as they distort the defining characteristics of the signal (Nanni et al., 2020; Zhong et al., 2021). Similarly, methods used to augment non-biological sounds (e.g. pitch shifting, time-stretching) may also alter the signal sufficiently to change features that are used for species identification such as frequency and duration. However, there are a number of augmentation techniques that can generate data to introduce variability without losing defining features. In this study we used five augmentation techniques; *i.* band stop filter; *ii.* background noise; *iii.* Gaussian noise; *iv.* time mask and *v.* time shift (Table 2). We applied each technique singly and then in combinations of up to four effects (Figs. 2–4). With the

Table 2
Augmentation effects: Description of each data transformation method and its abbreviation.

Augmentation effect	Method	Abbreviation
Background noise	Selects a random one second segment from a directory of audio files containing non-target sounds and adds this to the original sample. Non-target sounds include animal vocalisations and anthropogenic sounds such as cars.	Noise
Bandstop filter	masks a frequency band based on a randomly selected central frequency. The filter gradient is also randomized.	BSF
Gaussian noise	adds Gaussian noise at an amplitude between 0.01 and 0.015	Gaussian
Time mask	masks a time band in the spectrogram for 0.01–0.2 s	TM
Timeshift	shifts the time axis of the spectrogram by up to +/− 0.5 s. Samples that are shifted beyond the start or end of the spectrogram are moved to the opposite end	TS

exception of background noise all augmentation methods were implemented using the Python library Audiomentations (Jordal et al., 2019).

2.5. Testing the impact of network configuration and augmentation on classifier performance

To investigate the effects of increasing network depth and the number of training samples on classifier performance, we conducted a series of experiments using a dataset of 12 classes that each contained a minimum of 2000 novel (unaugmented) recordings (classes highlighted in bold in Table 1). Two sets of 200 samples per species were randomly selected for the validation and test datasets, of the remaining recordings a further 1600 were randomly selected for training. We used the same dataset to determine the most effective augmentation methods and the impact of substituting novel samples with augmented samples. Our final set of experiments included all 20 classes (Table 1) and investigated the impact of uneven classes in relation to balanced classes supplemented with augmented data (see supplementary information, Table 2 for a summary of experiments).

2.5.1. Network architecture and sample size

We assessed network architecture and configuration by training 12 different models: 10 models using greyscale input images (1 channel; nine CNNs with 2–10 convolutional layers and a ResNet50 model), and two networks with RGB input images (three channels); a 10-layer CNN and a ResNet50 model (Fig. 1). The RGB ResNet model was initialised using weights trained using Imagenet data, the greyscale ResNet50 did not use pre-trained weights. The effect of training sample size was tested by training each CNN using datasets of 100 to 1600 novel (unaugmented) samples per species in 100 sample increments. Each model configuration was trained for 100 epochs. We repeated this process five times in order to calculate mean accuracy and to account for any variation in results due to the stochastic nature of the training process. The two models achieving the highest overall mean accuracy were selected for the remaining experiments.

2.5.2. Assessing the impact of each augmentation effect

The effect of each augmentation method on overall model accuracy was assessed by training Conv10 (GS) and ResNet50 (RGB) models with a dataset containing 12 classes that contained a minimum of 2000 samples (Fig. 1). Models were first trained using a baseline dataset of 160 samples per species and then using a 'gold standard' dataset of 1600 novel samples per species to give probable lower and upper bounds for performance. Subsequent models were trained using 160 original samples that were augmented using one or a combination of the augmentation effects listed in Table 2 to create an augmented dataset containing 10 % original samples and 90 % augmented data, so that the amount of training data was the same as the gold standard dataset. Models were evaluated five times and mean accuracy calculated. Augmented model performance was compared by calculating the proportion of the gold standard accuracy achieved in relation to the baseline accuracy referred to as relative change in accuracy (RCA). This metric was used instead of baseline accuracy change as it gives a measure of the impact of data augmentation on model performance compared to models training using novel data only i.e. the gold standard dataset. Relative change in accuracy (RCA) was computed using the equation below:

$$RCA = \frac{ACC_{Augmented} - ACC_{Baseline}}{ACC_{Gold} - ACC_{Baseline}} \quad (1)$$

Where $ACC_{Augmented}$, $ACC_{Baseline}$ and ACC_{Gold} are the mean test accuracies of the augmented model under test, the baseline model trained on 160 samples and the gold standard model trained on 1600 samples.

2.5.3. Augmented data vs original samples

To understand how model performance is affected by increasing the proportion of augmented data and reducing the number of novel samples compared to using novel samples only, a series of experiments was conducted using the dataset which included 12 species with more than 2000 samples available. In each iteration we increased the number of novel samples by 100, starting from 100 up to 1600 samples per species. Novel samples were augmented using the five methods with the greatest RCA (Fig. 3) in equal proportion to increase the number of samples to 1600 per species. Both models (Conv10 (GS) and ResNet50 (RGB)) were evaluated five times to compute mean accuracy.

2.5.4. Unbalanced data

Training models with unbalanced classes (i.e. an unequal number of samples per species) can lead to bias towards the better represented classes (Borowiec et al., 2022). In our reference library there is a significant difference between the species with the lowest number of recordings (*D. nitedula* 467 recordings) and the greatest number of recordings (*Rattus rattus*, 7033). To examine this, we first trained Conv10 (GS) and ResNet50 (RGB) models using the complete, unbalanced reference library listed in Table 1 to obtain the overall mean F1 score. The F1 score is a metric commonly used in machine learning and is the harmonic mean between precision (the number of predictions that are correct) and recall (the number of correct predictions as a proportion of the total number of examples). It is calculated using the following formula:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Next we trained Conv 10 (GS) and ResNet50 (RGB) models using balanced datasets of 250 to 4000 samples per species, increasing the training datasets by 250 samples per species during each iteration. As the size of the training dataset increased, we used data augmentation to balance the classes that had insufficient novel samples.

3. Results

Our analyses begin with the network architecture and sample experiments, we find that accuracy increases as the number of convolutional layers increases (Fig. 5). Next, we compare the RCA for each of the 20 data augmentation methods, the results show that all augmentation techniques improve RCA with the exception of three effects that give negative RCA scores to Conv10 mean accuracy (Fig. 6). The third part of our analyses is represented in Fig. 7 which shows the change in accuracy as the proportion of augmented data increases versus using novel samples only. Finally, we evaluate the difference in model accuracy when the number of training samples increased, using augmentation to balance uneven classes compared to a model trained using an unbalanced dataset of novel samples only. We determine that the number of training samples is more important for improving accuracy than balancing the classes (Fig. 8).

3.1. Network configuration

The models achieving the two highest mean accuracy scores across all training sample sizes were ResNet50 (RGB) (96.37 %) and Conv10 (GS) (96.43 %). The lowest performing models were Conv2 (95.0 %) and Conv4 (95.0 %) (Fig. 5).

The difference in mean accuracy was greater for models trained with fewer training samples; for models trained with 100 samples per species the difference between the highest and lowest accuracy scores was 6.82 % points compared to 1.33 % points for models trained using 1600 samples per species (Fig. 5). Training Conv10 models using RGB images resulted in a poorer performance, reducing accuracy by 0.15 to 4.56 % compared to models trained with greyscale images (Fig. 5). Conversely, using RGB images to train ResNet50 models increased accuracy by 0.42 to 5.76 % compared to models using greyscale input (Fig. 5).

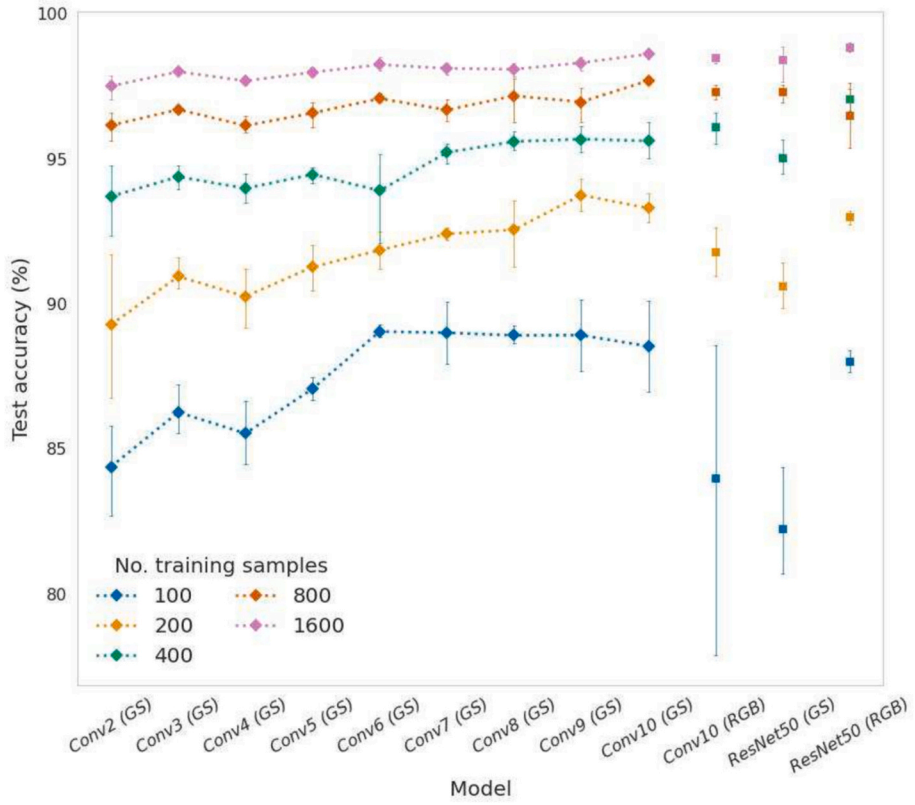


Fig. 5. Mean model accuracy for Conv2–10 and ResNet50 greyscale (GS) networks and Conv10 and ResNet50 (RGB) networks trained using 100, 200, 400, 800 and 1600 samples per species. The bars represent 95 % confidence intervals.

3.2. Augmentation effects

Augmentation effects were tested using the two CNN models that achieved the greatest accuracy in section 3.1 (Conv10 (GS) and ResNet50 (RGB)). Overall, data augmentation had a greater positive impact on ResNet50 RCA compared to Conv10; ResNet50 performance was not negatively impacted by any augmentation method (Fig. 6). Relative change in accuracy was higher in models trained with data augmentation protocols that contained time shift (Table 3, Fig. 6). Three protocols negatively affected Conv10 performance; time mask and Gaussian noise (−0.041 RCA), Gaussian noise (−0.094 RCA) and band stop filter with Gaussian noise (−0.182 RCA).

3.3. Augmented data vs original samples

Fig. 7. shows mean test accuracy for models trained using novel samples only and models trained with datasets of 1600 samples per species with increasing proportions of augmented data moving from left to right. Model accuracy significantly improved when augmented data was used to increase the number of training samples, particularly when the number of original samples was low. Model accuracy improved by 3.92 % (Conv10 (GS)) and 8.37 % (ResNet50 (RGB)) with the addition of 90 % augmented data compared to training with 100 samples per species only (Fig. 7). Although model accuracy increased when data augmentation was used to increase the number of training samples to 1600 per species, it was not a replacement for using 1600 novel samples per species which achieved the best accuracy scores. Model accuracy decreased when novel data was substituted with augmented data particularly with when the proportion of augmented data constituted more than 50 % of training data (Fig. 7). This effect had a greater impact with Conv10 (GS) networks compared to a significantly smaller effect with ResNet50 (RGB) (Fig. 7).

3.4. Unbalanced data

Fig. 8 shows mean test accuracy using data with unbalanced classes was greater for Conv10 (GS) (98.61 %) compared to ResNet50 (RGB) models (97.91 %). For models trained using datasets with balanced classes, sample sizes of 1000 or greater were needed in order to surpass ResNet50 (RGB) unbalanced mean accuracy however the number of training samples per classes required for mean accuracy to exceed unbalance Conv10 (GS) accuracy was 2500 indicating that the number of training examples is more important than balancing classes where the amount of training data are limited, particularly with ResNet50 (RGB) models. Model accuracy for both networks began to plateau at 2250 samples per species.

The confusion matrices (Fig. 9) show that when the classes are unbalanced the two species that had the highest classification errors were *D. nitedula* (8 %) and *G. glis* (4 %). These species were generally misclassified as noise. Similarly, the species with the highest classification error for the Conv10 (GS) model trained with a balanced dataset were *D. nitedula* (4 %) and *G. glis* (3 %) with the addition of *M. avellanarius* (4 %). The confusion matrix for ResNet50 (RGB) trained using balanced data showed that *G. glis* has a significantly higher classification error (17 %). These errors were caused by incorrectly classifying *G. glis* calls as Noise. In all four models errors were also generated by vocally similar species being misidentified for example *A. flavicollus* and *A. sylvaticus* and both *Sorex* species.

4. Discussion

A key challenge during CNN acoustic classifier development is the availability of large numbers of labelled recordings, particularly for elusive or endangered species (Borowiec et al., 2022; Zhong et al., 2021). We show that although increasing the number of real training samples is the best way to improve CNN accuracy, data augmentation

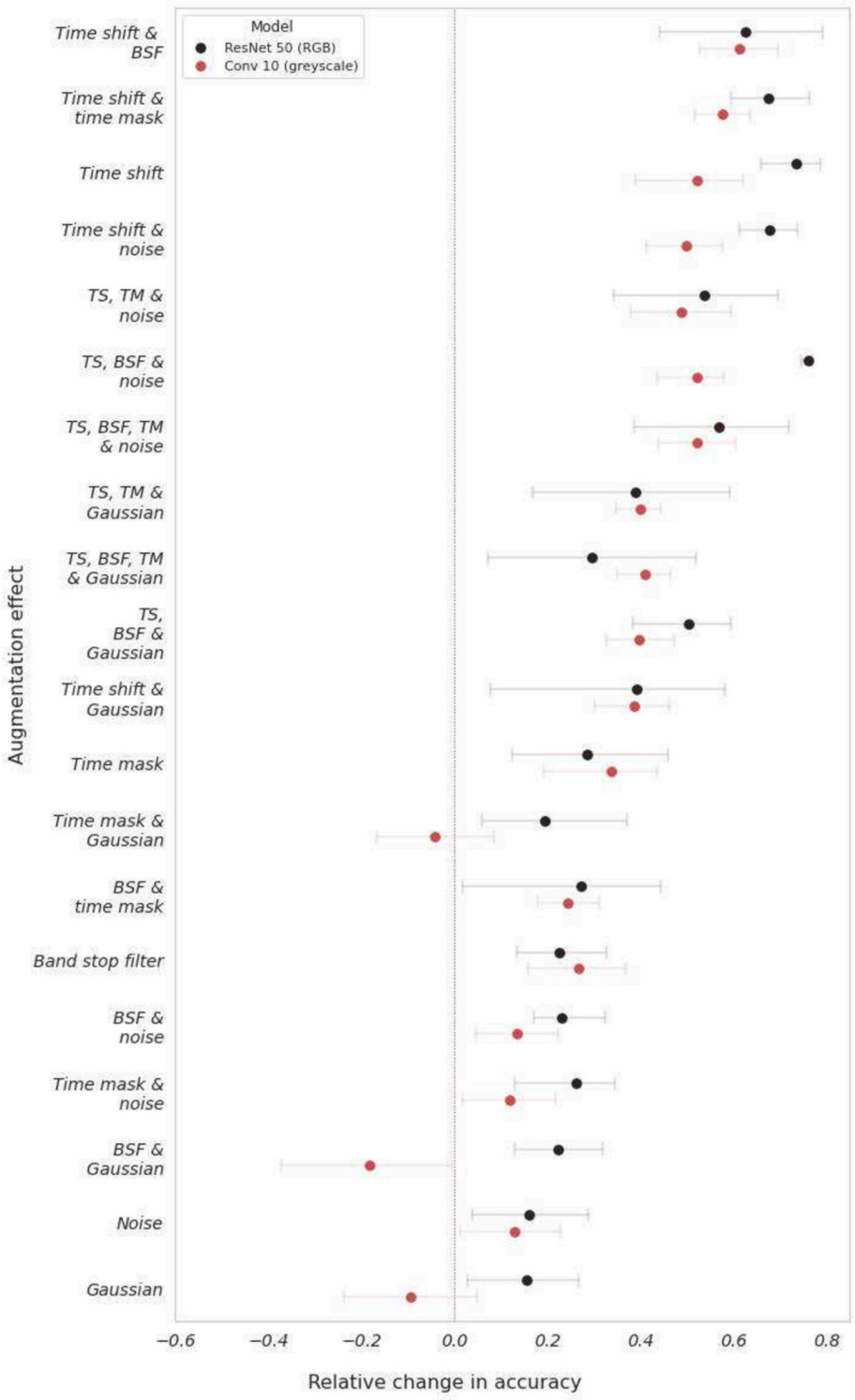


Fig. 6. Relative change in accuracy for each augmentation protocol for ResNet50 and Conv10 models. The bars represent 95 % confidence intervals.

can substantially improve accuracy where the amount of training data is low. This has important implications in PAM studies for species where labelled data are unavailable or difficult to obtain.

Our network configuration tests demonstrate that a very shallow neural network with two convolution layers can achieve a high level of model accuracy identifying 12 species of small mammals with mean accuracy of 94.95 % (Fig. 1). Model accuracy was further improved with the addition of convolutional layers although this effect was less pronounced with larger datasets suggesting that increasing the number of training samples had a greater impact on accuracy compared to network depth (Fig. 1). Generally, deeper networks require more training data to prevent overfitting (Zhang et al., 2021) but we did not find this effect here with the exception of the greyscale ResNet50 model. Deeper networks can overfit with small datasets due to the lack of variation in training data coupled with the need to train additional layers that may extract fine grained features that are only present in the training data (Zhang et al., 2021). This is reflected in the accuracy scores for ResNet50 (GS), which is not pre-trained, that are markedly lower compared to the

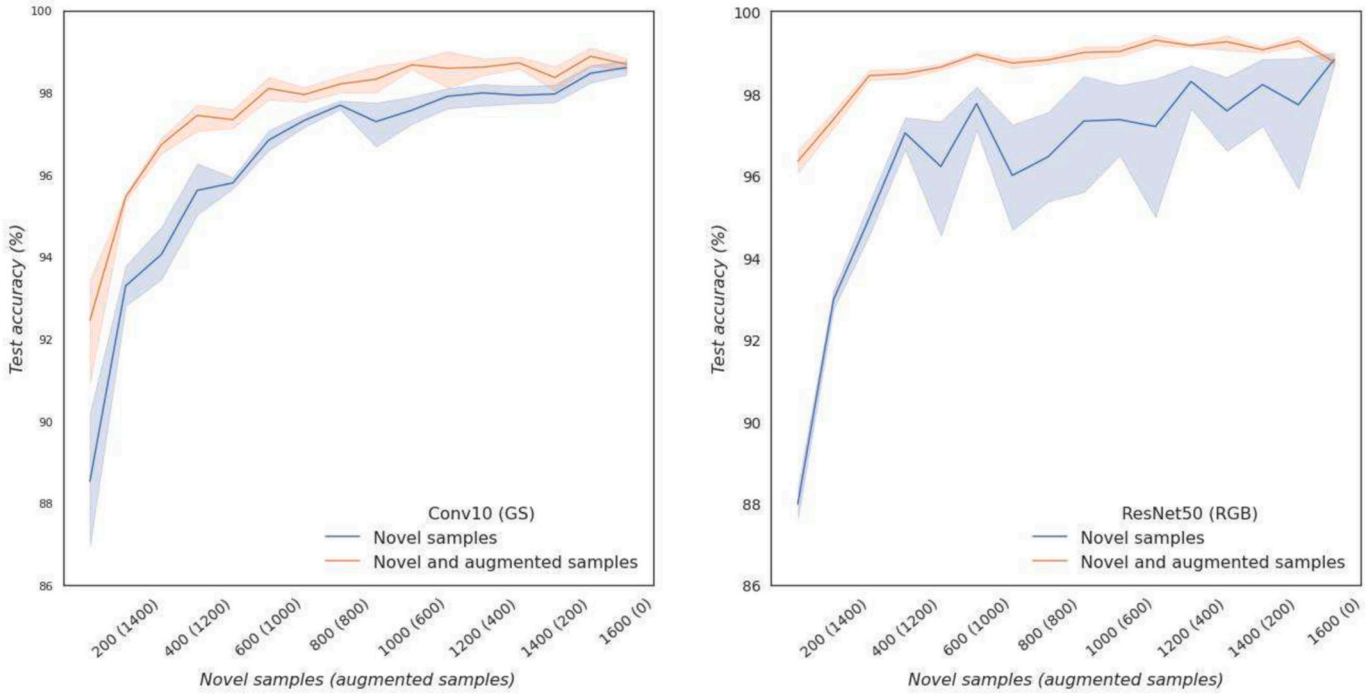


Fig. 7. Test accuracy of models trained with increasing numbers of novel samples vs models with trained using novel samples and the addition of augmented data. The right panel represents mean test accuracy for Conv10 (GS)networks and the left panel shows test accuracy for ResNet50 (RGB) models.

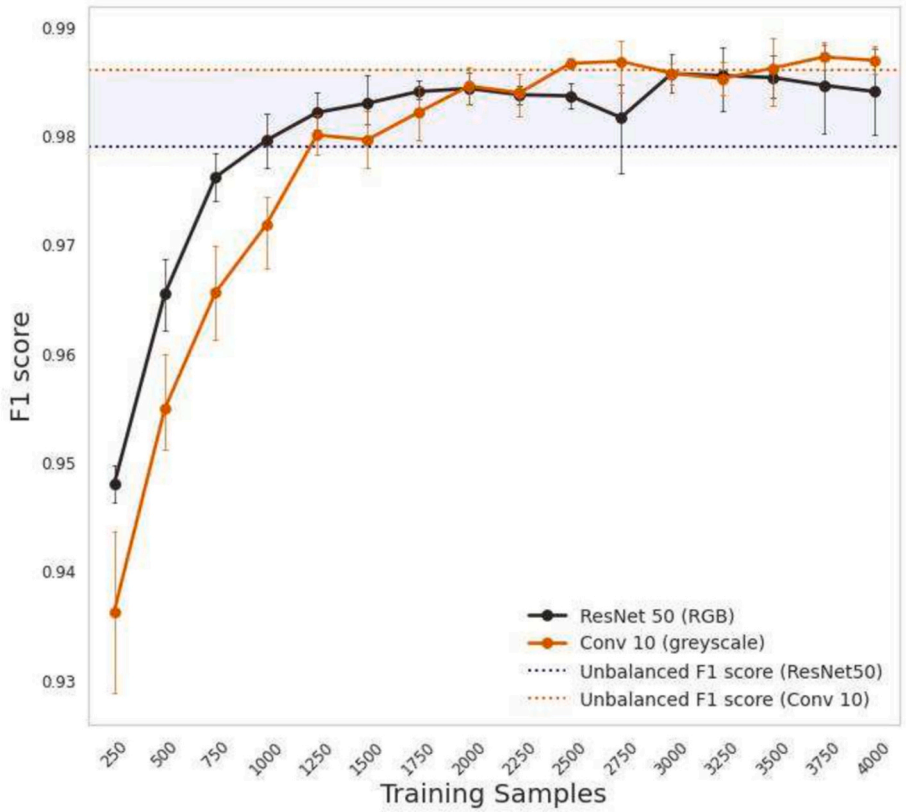


Fig. 8. F1 score of models trained with increasing training sample sizes. Data augmentation was used to increase sample sizes for species with insufficient novel samples. The dotted lines represent the F1 score of each model trained using data containing unbalanced classes of unaugmented samples.

pre-trained ResNet50 (RGB) model. The difference is particularly evident when networks were trained using 100 samples per species (Fig. 1).

Apart from ResNet50 (GS) classifiers, deeper models outperformed networks with fewer convolutional layers when the amount of training data was reduced. To address overfitting concerns, training accuracy

Table 3
Augmentation effects with the five highest and lowest combined relative change in accuracy (RCA). Combined RCA is calculated as the mean of the RCA values for ResNet50 and Conv10 classifiers.

Augmentation effect	combined RCA
Time shift, band stop filter and noise	0.641
Time shift	0.629
Time shift and time mask	0.626
Time shift and band stop filter	0.620
Time shift and background noise	0.588
Background noise	0.184
Band stop filter and background noise	0.146
Time mask and Gaussian noise	0.077
Gaussian noise	0.0311
Band stop filter and Gaussian noise	0.0213

and loss curves were assessed; the accuracy and loss curves for Conv10 (GS) models trained with 100 samples per species indicate that although validation accuracy and loss improve, values fluctuate significantly which could indicate some overfitting (Supplementary materials, Fig. 5a). In comparison, adding further training data shows the accuracy and loss curves to be more stable (Supplementary materials, Fig. 4b and c). In fact, adding as few as another 100 samples per species suggests that overfitting can be reduced with a relatively modest increase in training data (Supplementary materials, Fig. 4c). Due to the hierarchical nature of CNNs, shallow networks lack the capacity to extract higher level features which may be required to reduce classification errors (LeCun et al., 2015). Our results suggest that there is an optimal network depth for small datasets which contains sufficient convolutional layers for effective feature extraction but also reduces the risk of overfitting.

ResNet50 (RGB) models outperformed ResNet50 models trained with greyscale images which can be attributed to the addition of pre-trained weights, partially mitigating the issue of overfitting seen with small datasets. Dufourq et al. (2022) also highlights the potential of deep networks for acoustic classification by demonstrating that CNN classifiers with more than 50 convolutional layers could achieve F1 scores of up to 82 % when using as few as 25 training examples (Dufourq et al., 2022).

Convolutional neural networks require large datasets to achieve high accuracy scores (Christin et al., 2019) and this is illustrated in Fig. 7; model accuracy rapidly increases initially when additional training examples are added. Accuracy starts to plateau at 1600 samples per species suggesting that increasing the number of training samples further would have a negligible impact on overall accuracy. Fig. 7 also highlights the impact of adding augmented data. Augmentation was most beneficial when the number of original samples was low; adding augmented data had a greater impact on ResNet50 (RGB) performance. Deep models are prone to over fitting particularly when there are few training data (Zhang et al., 2021) however our study shows this can be mitigated with the addition of augmented data. This has important implications for PAM studies involving rare or elusive species where it is difficult to obtain sufficient recordings; when relatively few recordings are available sample sizes can be boosted using augmentation in order to achieve accuracy comparable to unaugmented datasets.

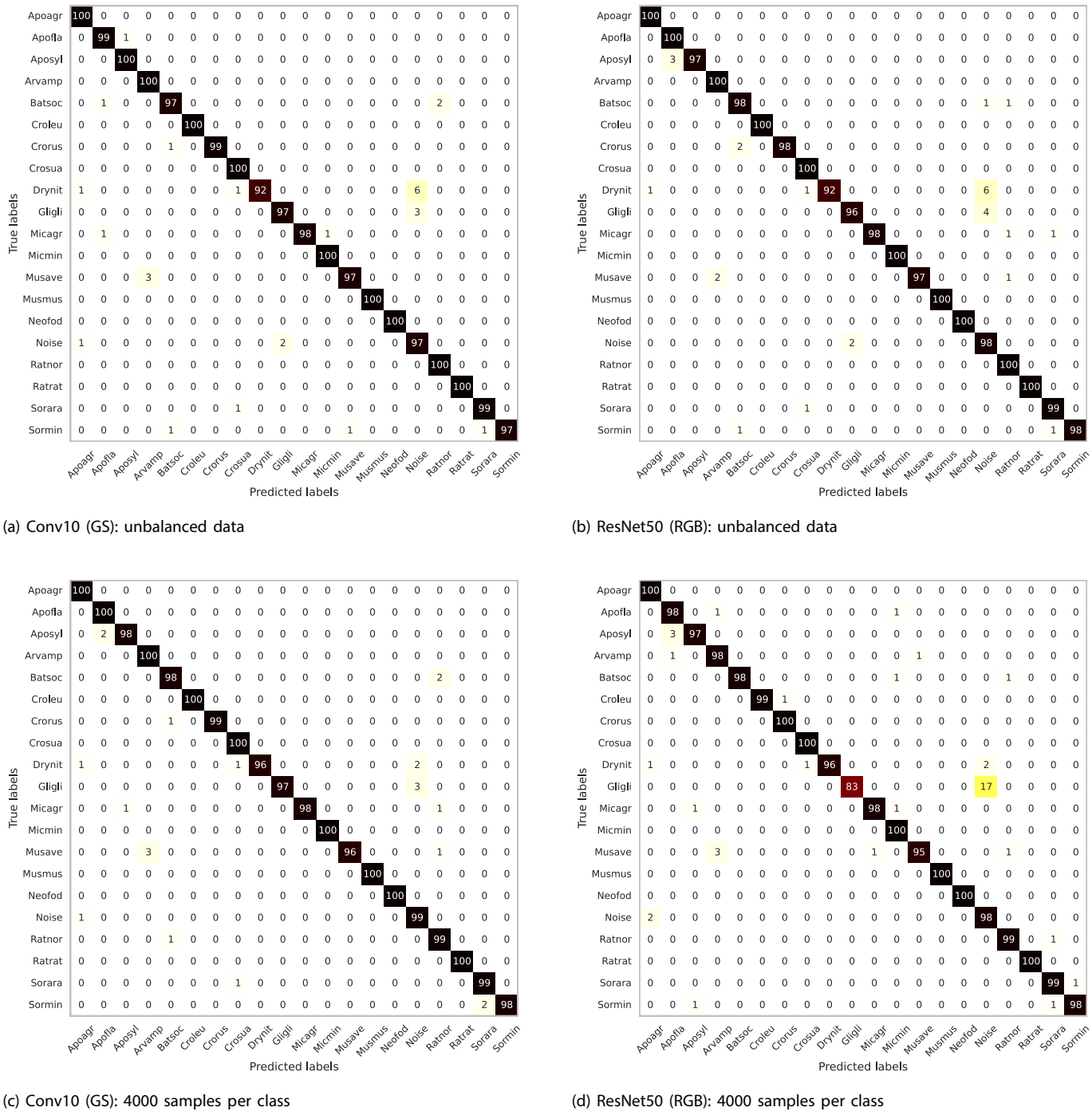
Augmentation methods that achieved the highest relative change in accuracy included time shift with the single effect time shift achieving the third highest RCA for both models (Fig. 6). Using a combination of augmentation effects compared to single augmentation effects generally had a greater positive impact on RCA which is likely due to the addition of a time shift transformation. Time shift transformations introduce a significant amount of variation whilst maintaining the key features of call structure in terms of frequency and duration. Combining time shift with one or more effects adds another layer of variation which in this study has resulted in an increase in RCA. Conversely the three augmentation methods that had the greatest negative impact on RCA were two combined effects band stop filter and Gaussian noise, time mask and background noise, and the single effect Gaussian noise (Fig. 3). The reference library contains recordings with varying signal to noise ratios (SNR) and it is possible that the addition of Gaussian or background noise masked vocalisations resulting in poorer model performance. This is supported in the works presented by Su et al. (2024) which demonstrated that adding Gaussian noise reduced F1 scores and Nshimiyimana (2024) which showed that adding random noise also negatively impacted performance. Manriquez et al., (2024) determined that adding noise below 15 % of total amplitude was beneficial which suggests that SNR needs to be carefully considered when using Gaussian noise addition. Models trained with data augmentation methods including band stop filter had mixed results in terms of RCA. Frequency is a key characteristic in small mammal species identification (Newson et al., 2020) and applying a band stop filter could potentially eliminate the frequency band containing the vocalisation resulting in a reduced impact on model accuracy. Additionally, adding a time mask could remove part of the spectrogram containing the vocalisation although time mask had a greater positive impact on model performance compared to band stop filter. Given that masking part of the spectrogram could remove key features required for classification it may seem contradictory that both effects combined improved RCA. Improvements to RCA could simply be due to the increased number of training examples irrespective of augmentation method however three augmentation protocols decreased model performance. Another potential reason that adding band stop filters and time masks improves RCA is that they could act in a similar way to dropout layers added to the network architectures. Adding dropout layers to CNNs helps to reduce over fitting by randomly setting a proportion of the convolutional layer's outputs to zero i.e. omitting a percentage of the layer's output which has the effect of breaking up patterns identified by the network that are not relevant to the classification task (Srivastava et al., 2014).

The issue of unbalanced datasets is a common concern in machine learning due to the potential bias towards classes with disproportionately large numbers of training examples (Borowiec et al., 2022). Both models trained using unbalanced datasets achieved high mean F1 scores which appears to contradict this statement (Fig. 8). However using training data with unbalanced classes had a greater impact on ResNet50 (RGB) model performance compared to Conv10 (GS) performance indicating that complex models are more susceptible to class imbalance. Deeper networks are prone to over fitting when there are insufficient training data (Srivastava et al., 2014; Zhang et al., 2021) which may have contributed to the poorer performance with the unbalanced dataset where there were as few as 467 recordings for *D. nitedula*. Class imbalance had little impact on Conv10 (GS) model. Although overall F1 score did not improve significantly, balancing class reduced classification errors for *D. nitedula*, which had the lowest number of novel samples (Fig. 9) for both models. The confusion matrix for the balanced ResNet50 (RGB) network had a significantly high classification error rate for *G. glis* (17 %). *G. glis* is the only species in this study that vocalises primarily in the audible frequency range and it is possible that the network has overfit the Noise class resulting in the misclassification of *G. glis* calls as noise. Our analyses demonstrates that the number of training samples is more important than balanced classes in terms of overall F1 score however increasing the number of samples for poorly represented classes can improve class specific accuracy.

5. Recommendations

The results of our study highlights several key areas that could benefit researchers developing CNN acoustic classifiers and we suggest the following recommendations:

- When training data is limited (e.g. 100 samples per species), increasing the number of convolutional layers can improve model accuracy by up to 6.82 %; very deep networks such as ResNet50



(a) Conv10 (GS): unbalanced data

(b) ResNet50 (RGB): unbalanced data

(c) Conv10 (GS): 4000 samples per class

(d) ResNet50 (RGB): 4000 samples per class

Fig. 9. Confusion matrices showing test data predictions (%). The top row shows species predictions using a Conv10 (GS) model trained using data containing 20 uneven class (a) and data containing 20 classes with 4000 samples of novel and augmented data (b). The bottom row shows predictions using a ResNet50 (RGB) model trained with unbalanced data (c) and classes balanced using data augmentation (d).

benefit from pretraining to reduce overfitting and can increase accuracy by more than 5 % compared to models without pretraining.

- Classifier accuracy can be improved by *c.* 3.92–8.36 % using data augmentation to generate 90 % more training data when there are as few as 100 recordings per species.
- Data augmentation methods that include time shift are preferable to methods based on band stop filters; avoid adding Gaussian noise or background noise where the signal to noise ratio is poor to reduce the risk of masking vocalisations.
- Although balancing classes using augmentation can make modest improvements on overall accuracy, (0.12–0.75 %), it can significantly reduce species-specific classification errors for under-represented classes.

6. Conclusion

Our study demonstrates that although the number of training samples has the greatest impact on classifier performance, data augmentation can improve models when training data are scare. Obtaining recordings of animal vocalisations is challenging and time consuming (Newson et al., 2020) and the availability of labelled recordings can be limited. We show that data augmentation can be used to supplement

training datasets with as few as 100 samples per species to improve classifier performance. This highlights the potential of data augmentation as an efficient means of generating additional data where acoustic data collection is difficult and to mitigate challenges associated with unbalanced or small training datasets.

CRediT authorship contribution statement

Jennifer MacIsaac: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stuart Newson:** Writing – review & editing, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Adham Ashton-Butt:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Huma Pearce:** Investigation, Data curation. **Ben Milner:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors have no conflict of interest to declare.

Data availability

The source code for this paper is available on GitHub: https://github.com/j-macisaac/bioacoustic-classification-and-data-augmentation.git

Acoustic datasets used during this study are available from the corresponding author on reasonable request.

Acknowledgements

This work was supported by the Natural Environment Research Council and the ARIES Doctoral Training Partnership [grant number NE/S007334/1], the Endangered Landscapes and Seascapes Programme, managed by the Cambridge Conservation Initiative in partnership with Arcadia and Frankfurt Zoological Society. Data collection was facilitated by APB and Anton Kuzmickij in Belarus, Kaunas Tadas Ivanauskas Museum of Zoology in Lithuania and Roger Trout in the UK. The research presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102851.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from. tensorflow.org.

Abrams, J.F., Vashishtha, A., Wong, S.T., Nguyen, A., Mohamed, A., Wieser, S., Kuijper, A., Wilting, A., Mukhopadhyay, A., 2019. Habitat-net: segmentation of habitat images using deep learning. Eco. Inform. 51, 121–128.

Allen, M.J., Grieve, S.W., Owen, H.J., Lines, E.R., 2023. Tree species classification from complex laser scanning data in mediterranean forests using deep learning. Methods Ecol. Evol. 14 (7), 1657–1667.

Ancillotto, L., Russo, D., 2016. Individual vs. non-individual acoustic signalling in African woodland dormice (Graphiurus murinus). Mamm. Biol. 81 (4), 410–414.

Ancillotto, L., Sozio, G., Mortelliti, A., Russo, D., 2014. Ultrasonic communication in Gliridae (Rodentia): the hazel dormouse (Muscardinus avellanarius) as a case study. Bioacoustics 23.

Ancillotto, L., Mori, E., Sozio, G., Solano, E., Bertolino, S., Russo, D., 2017. A novel approach to field identification of cryptic apodemus wood mice: calls differ more than morphology. Mammal Rev. 47 (1), 6–10.

Ayala-Berdon, J., Medina-Bello, K.I., López-Cuamatzi, I.L., Vázquez-Fuerte, R., MacSwiney, G., Orozco-Lugo, L., Iñiguez-Dávalos, I., Guillén-Servent, A., Martínez-Gómez, M., 2020. Random forest is the best species predictor for a community of insectivorous bats inhabiting a mountain ecosystem of Central Mexico. Bioacoustics 1–21.

Bas, Y., Bas, D., Julien, J.-F., 2017. Tadarida: a toolbox for animal detection on acoustic recordings. J. Open Res. Softw. 5, 6.

Benedek, A.M., Sîrbu, I., Lazăr, A., 2021. Responses of small mammals to habitat characteristics in Southern Carpathian forests. Sci. Rep. 11 (1), 12031.

Bermant, P.C., Bronstein, M.M., Wood, R.J., Gero, S., Gruber, D.F., 2019. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. Sci. Rep. 9 (1), 12588.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., White, A.E., 2022. Deep learning as a tool for ecology and evolution. Methods Ecol. Evol. 13 (8), 1640–1660.

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., Freeman, R., 2018. Predicting animal behaviour using deep learning: GPS data alone accurately predict diving in seabirds. Methods Ecol. Evol. 9 (3), 681–692.

Chen, X., Zhao, J., Chen, Y.-H., Zhou, W., Hughes, A.C., 2020. Automatic standardized processing and identification of tropical bat calls using deep learning approaches. Biol. Conserv. 241, 108269.

Chollet, F., 2018. Deep Learning with Python. Manning Publications Co.

Christin, S., Hervet, E., Lecomte, N., 2019. Applications for deep learning in ecology. Methods Ecol. Evol. 10 (10), 1632–1644.

Clink, D.J., Klinck, H., 2021. Unsupervised acoustic classification of individual gibbon females and the implications for passive acoustic monitoring. Methods Ecol. Evol. 12 (2), 328–341.

Coffey, K.R., Marx, R.G., Neumaier, J.F., 2019. DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. Neuropsychopharmacology 44 (5), 859–868.

Dufourq, E., Durbach, I., Hansford, J.P., Hoepfner, A., Ma, H., Bryant, J.V., Stender, C.S., Li, W., Liu, Z., Chen, Q., et al., 2021. Automated detection of Hainan gibbon calls for passive acoustic monitoring. Remote Sens. Ecol. Conserv. 7 (3), 475–487.

Dufourq, E., Batist, C., Foquet, R., Durbach, I., 2022. Passive acoustic monitoring of animal populations with transfer learning. Eco. Inform. 70, 101688.

Florentin, J., Dutoit, T., Verlinden, O., 2020. Detection and identification of European woodpeckers with deep convolutional neural networks. Eco. Inform. 55, 101023.

Gibb, R., Browning, E., Glover-Kapfer, P., Jones, K.E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 10 (2), 169–185.

Goodwin, C.E.D., Hodgson, D.J., Al-Fulaij, N., Bailey, S., Langton, S., Mcdonald, R.A., 2017. Voluntary recording scheme reveals ongoing decline in the United Kingdom hazel dormouse *Muscardinus avellanarius* population. Mammal Rev. 47 (3), 183–197.

Goussha, Y., Bar, K., Netser, S., Cohen, L., Hel-Or, Y., Wagner, S., 2022. HybridMouse: a hybrid convolutional-recurrent neural network-based model for identification of mouse ultrasonic vocalizations. Front. Behav. Neurosci. 15.

Hafner, S.D., Katz, J., 2018. monitoR: Acoustic template detection in R. R package version 1.0.7.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heinicke, S., Kalan, A.K., Wagner, O.J.J., Mundry, R., Lukashevich, H., Kühl, H.S., 2015. Assessing the performance of a semi-automated acoustic monitoring system for primates. Methods Ecol. Evol. 6 (7), 753–763.

Jordal, I., Tamazian, A., Chourdakis, E., Angonin, C., 2019. Audiomentations: A Python Library for Audio Data Augmentation.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, J.P., 2019. Robust sound event detection in bioacoustic sensor networks. PLoS One 14 (10), e0214168.

Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B., Kinsey, L., Mead, G.R., Newson, S.E., Pandourski, I., Parsons, S., Russ, J., Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G., Jones, K.E., 2018. Bat detective—deep learning tools for bat acoustic signal detection. PLoS Comput. Biol. 14 (3), e1005995.

Manriquez, P, Kotz, S.A., Ravignani, A., De Boer, B., 2024. Bioacoustic classification of a small dataset of mammalian vocalisations using deep learning. Bioacoustics 33 (4), 354–371.

Middleton, N., Newson, S., Pearce, H., 2023. Sound Identification of Terrestrial Mammals of Britain & Ireland. Pelagic Publishing Ltd.

Mills, C.A., Godley, B.J., Hodgson, D.J., 2016. Take only photographs, leave only footprints: novel applications of non-invasive survey methods for rapid detection of small, arboreal animals. PLoS One 11 (1), e0146142.

Nanni, L., Maguolo, G., Paci, M., 2020. Data augmentation approaches for improving animal audio classification. Eco. Inform. 57, 101084.

Newson, S., Bas, Y., Murray, A., Gillings, S., 2017. Potential for coupling the monitoring of bush-crickets with established large-scale acoustic monitoring of bats. Methods Ecol. Evol. 8 (9), 1051–1062.

Newson, S., Middleton, N., Pearce, H., 2020. The acoustic identification of small terrestrial mammals in britain. Br. Wildl. 32, 186–194.

Nshimiyimana, A., 2024. Acoustic data augmentation for small passive acoustic monitoring datasets. Multimed. Tools Appl. 1–19.

Prince, P., Hill, A., Piña Covarrubias, E., Doncaster, P., Snaddon, J., Rogers, A., 2019. Deploying acoustic detection algorithms on low-cost, open-source acoustic sensors for environmental monitoring. Sensors 19 (3), 553.

Ravaglia, D., Ferrario, V., De Gregorio, C., Carugati, F., Raimondi, T., Cristiano, W., Torti, V., Hardenberg, A.V., Ratsimbazafy, J., Valente, D., et al., 2023. There you

are! Automated detection of indris' songs on features extracted from passive acoustic recordings. Animals 13 (2), 241.
Salamon, J., Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. 24 (3), 279–283.
Sebastián-González, E., Pang-Ching, J., Barbosa, J.M., Hart, P., 2015. Bioacoustics for species management: two case studies with a Hawaiian forest bird. Ecol. Evol. 5 (20), 4696–4705.
Shonfield, J., Do, R., Brooks, R.J., McAdam, A.G., 2013. Reducing accidental shrew mortality associated with small-mammal livetrapping I: an inter- and intrastudy analysis. J. Mammal. 94 (4), 745–753.
Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6 (1), 60.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: 3rd International Conference on Learning Representations (ICLR 2015).
Sprengel, E., Jaggi, M., Kilcher, Y., Hofmann, T., 2016. Audio based bird species identification using deep learning techniques. LifeCLEF 2016, 547–559.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Su, S., Gu, D., Lai, J.Y., Arcilla, N., Su, T.Y., 2024. A novel deep learning-based bioacoustic approach for identification of look-alike white-eye (Zosterops) species traded in wildlife markets. Ibis.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
Thomas, L., Jaramillo-Legorreta, A., Cardenas-Hinojosa, G., Nieto-Garcia, E., Rojas-Bracho, L., Ver Hoef, J.M., Moore, J., Taylor, B., Barlow, J., Tregenza, N., 2017. Last call: passive acoustic monitoring shows continued rapid decline of critically endangered vaquita. J. Acoust. Soc. Am. 142 (5). EL512–EL517.
Zeppelzauer, M., Hensman, S., Stoeger, A.S., 2015. Towards an automated acoustic detection system for free-ranging elephants. Bioacoustics 24 (1), 13–29.
Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 64 (3), 107–115.
Zhong, M., LeBien, J., Campos-Cerqueira, M., Dodhia, R., Lavista Ferres, J., Velev, J.P., Aide, T.M., 2020. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. Appl. Acoust. 166, 107375.
Zhong, M., Taylor, R., Bates, N., Christey, D., Basnet, H., Flippin, J., Palkovitz, S., Dodhia, R., Lavista Ferres, J., 2021. Acoustic detection of regionally rare bird species through deep convolutional neural networks. Eco. Inform. 64, 101333.
Zsebők, S., Czabán, D., Farkas, J., Siemers, B.M., von Merten, S., 2015. Acoustic species identification of shrews: twittering calls for monitoring. Eco. Inform. 27, 1–10.