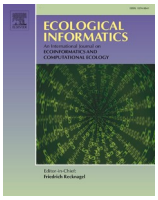
Ecological Informatics 77 (2023) 102256

Contents lists available at ScienceDirect

Ecological Informatics

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



Improving deep learning acoustic classifiers with contextual information for wildlife monitoring

Lorène Jeantet [a,b,c,*], Emmanuel Dufourq [a,b,c]

[a] African Institute for Mathematical Sciences, South Africa
[b] Stellenbosch University, South Africa
[c] National Institute for Theoretical and Computational Sciences, South Africa

ARTICLE INFO

Keywords:
Bioacoustics
Deep learning
Convolutional neural networks
Passive acoustic monitoring
Species identification
Birds
Hainan gibbons

ABSTRACT

Bioacoustics, the exploration of animal vocalizations and natural soundscapes, has emerged as a valuable tool for studying species within their habitats, particularly those that are challenging to observe. This approach has broadened the horizons of biodiversity assessment and ecological research. However, monitoring wildlife with acoustic recorders produces large volumes of data that can be labor-intensive to analyze. Deep learning has recently transformed many computational disciplines by enabling the automated processing of large and complex datasets and has gained attention within the bioacoustics community. Despite the revolutionary impact of deep learning on acoustic detection and classification, attaining both high detection accuracy and low false positive rates in bioacoustics remains a significant challenge. An intriguing yet unexplored avenue for enhancing deep learning in bioacoustics involves the utilization of contextual information, such as time and location, to discern animal vocalizations within acoustic recordings. As a first case study, a multi-branch Convolutional Neural Network (CNN) was developed to classify 22 different bird songs using spectrograms as a first input, and spatial metadata as a secondary input. A comparison was made to a baseline model with only spectrogram input. A geographical prior neural network was trained, separately, to estimate the probability of a species occurring at a given location. The output of this network was combined with the baseline CNN. As a second case study, temporal data and spectrograms were used as input to a multi-branch CNN for the detection of Hainan gibbon (*Nomascus hainanus*) calls, the world's rarest primate. Our findings demonstrate that adding metadata to the bird song classifier significantly improves classification performance, with the highest improvement achieved using the geographical prior model (F1-score of 87.78% compared to 61.02% for the baseline model). The multi-branch CNNs also proved efficient (F1-scores of 76.87% and 78.77%) and simpler to use than the geographical prior. In the second case study, our findings revealed a decrease in false positives by 63% (94% of the calls were detected) when the metadata was used by the multi-branch CNN, and an increase of 19% in gibbon detection. This study has uncovered an exciting new avenue for improving classifier performance in bioacoustics. The methodology described in this study can assist ecologists, wildlife management teams, and researchers in reducing the amount of time spent analyzing large acoustic datasets obtained from passive acoustic monitoring studies. Our approach can be adapted and applied to other calling species, and thus tailored to other use cases.

1. Introduction

The study of animal vocalization and natural soundscape, bioacoustics, has proven to be a valuable source of data for both a better understanding of animal behavior and for biodiversity monitoring. Technological advances in the field of digital sound recorders have led to increased battery and memory capacities, allowing for miniaturization of the devices while remaining affordable. Therefore, since 2017, the number of bioacoustics studies has steadily grown, alongside the amount of recorded data (Stowell, 2022; Mutanu et al., 2022). Nowadays, long-term deployment of multiple sound records in remote areas is both feasible and commonly done to study species that are difficult to observe (Gibb and Browning, 2019; Sugai et al., 2019). By detecting the presence of a species of interest, acoustic monitoring enables researchers to address fundamental ecological questions, encompassing aspects like occupancy, spatial distribution, and abundance trends, with significant

* Corresponding author at: African Institute for Mathematical Sciences, South Africa.
E-mail address: lorene.jeantet@hotmail.fr (L. Jeantet).

https://doi.org/10.1016/j.ecoinf.2023.102256
Received 2 June 2023; Received in revised form 7 August 2023; Accepted 7 August 2023
Available online 15 August 2023

conservation challenges for endangered species (Gibb and Browning, 2019; Ross et al., 2023; Miller et al., 2015). However, long-term deployments come with a cost in that large files are created which are time-consuming to manually analyze. Kohlsdorf et al. (2020) stated that a one-hour field record can require up to ten hours of manual analysis. The automation within the analysis phase of bioacoustics research has become an issue and unequivocally, there is a need to adopt modern analysis methods to facilitate rapid processing.

Over the past few years, deep learning has revolutionized several research fields such as bioinformatics (Li et al., 2020) and medicine (Piccialli et al., 2021) by enabling automated processing of large and complex datasets. Considered a branch of machine learning, deep learning refers to algorithms, commonly called deep neural networks, able to automatically detect very complex and highly discriminating patterns in data (Chollet, 2018). The succession of processing layers performing linear and non-linear transformations allows the neural networks to learn representations of data with multiple levels of abstraction (LeCun et al., 2015). This ability makes deep learning particularly relevant for solving complex problems such as speech recognition, object detection, computer vision, and many other domains (Taigman et al., 2014; Hinton et al., 2012). Naturally, these practices have spread to bioacoustics with the development and adaptation of deep learning methods to automatically process acoustic recordings.

Deep learning in bioacoustics is a relatively recent development, and there are not many studies available on its application prior to 2017. As per the review by Stowell (2022), the earliest known application of deep learning in bioacoustics can be traced back to the 2014 LifeCLEF bird identification challenge, also known as BirdCLEF (Goëau et al., 2014). However, at that time, only one participant attempted a deep neural network, and the performance was substantially lower than the winning team. The resurgence of deep learning in bioacoustics was marked in 2016, as highlighted in the title of the third edition report of LifeCLEF, titled "LifeCLEF Bird Identification Task 2016: The arrival of Deep learning" (Goëau et al., 2016). Three teams utilized deep learning techniques trained on a dataset comprising 999 species, achieving the best performance with a mean average precision of 0.55 for the top-performing team. The year 2017 witnessed the emergence of several conference papers focusing on deep learning-based classification and detection models, primarily for birds, but also extending to frogs, mice, and whales (Hassan et al., 2017; Grill and Schlüter, 2017; Dorian et al., 2017; Smith and Kristensen, 2017). Currently, deep learning is gaining popularity in bioacoustics research, with studies encompassing a diverse range of species, ranging from giant marine mammals such as whales to the wing beats of mosquitoes (Khalighifar et al., 2021; Zhong et al., 2020; Bermant et al., 2019). Irrespective of species size, deep learning studies in bioacoustics tend to focus on either finely tuning models for precise applications or inversely expanding them to include as many species as possible to achieve optimal generalization (Kahl et al., 2021). However, despite recent advancements, achieving high detection accuracy or low false positive rates continues to be a significant challenge in the use of deep learning for bioacoustics. The annual BirdCLEF event is a testament to this, as it aims to improve the accuracy of deep learning models in bird sound classification every year. In 2021, the leading submission for detecting and classifying 397 species from South and North America achieved an F1-score of 0.69 (Joly et al., 2021). Therefore, there is significant room for improvement in deep learning in bioacoustics to make it a reliable complementary tool for monitoring wildlife.

The predominant paradigm for enhancing deep learning applications in bioacoustics has centered on refining neural network architectures, in conjunction with optimizing pre- and post-processing methodologies but few have considered adding contextual data. Despite the availability of metadata in BirdCLEF, utilization of this contextual information to enhance bird sound classification has been infrequently used by competing teams (Joly et al., 2021; Joly et al., 2022). However, it stands to reason that an experienced human practitioner in sound recognition would rely on a diverse array of contextual cues to accurately identify bird species from acoustic recordings. Moreover, the most common approach in deep learning in bioacoustics is to convert the raw acoustic recordings into a succession of spectrograms or mel-spectrograms – a visual representation of the frequency spectrum of sound over time – and treat it as an image classification problem (Stowell, 2022). These spectrograms generally give a fixed image in time of the sounds and deprive them of their context. Although not inherently present in the spectrograms, metadata such as time, date and location can offer valuable information that is correlated with the recordings. This can be particularly useful in distinguishing between bird species with similar vocalizations, yet from geographically distinct distributions (Fig. 1). As it has become common practice to record the metadata associated with each sound recording, there is significant interest in leveraging this contextual information to enhance the performance of deep learning classifiers. Therefore, we hypothesize that one way to improve deep learning in species classification is to provide the algorithms with contextual information via metadata.

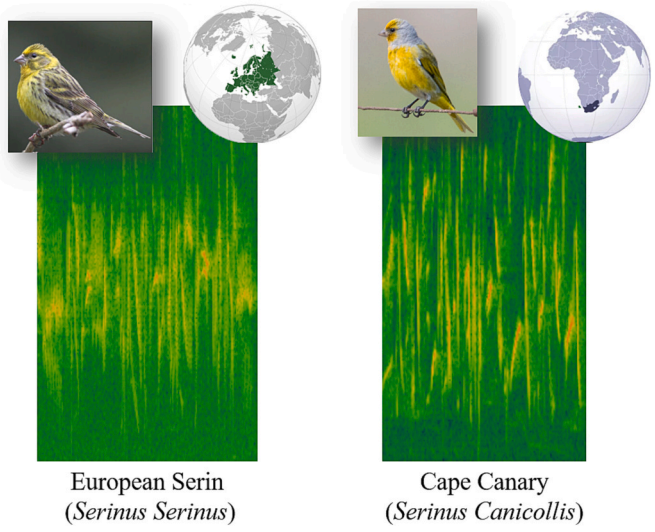


Fig. 1. Example of spectrograms recorded for two species, European serin (*Serinus serinus*) and Cape Canary (*Serinus canicollis*), occurring in different countries.

Our investigation contributes to the field of deep learning and bioacoustics by highlighting the potential for improved classification performance through the incorporation of contextual information such as time and location. Leveraging established techniques from image classification, we applied our approach to two distinct case studies. First, we developed a subset dataset derived from Xeno-canto to classify 22 different bird songs that included location metadata as input alongside the spectrogram. Our findings indicated a notable improvement in classification performance when using this auxiliary data. In the second case study, we focused on Hainan gibbons *Nomascus hainanus*, the world's rarest primate, which are known to be especially vocal during the morning hours. This knowledge, is however not known to a neural network. We trained the neural network on both spectrogram input and temporal metadata, again resulting in improved performance compared to using spectrograms alone. The rationale behind this study is to present efficient solutions that leverage simple model architectures and additionally, to introduce methods of varying complexity to incorporate temporal and location information that can easily be reused and adapted to other bioacoustics studies. We demonstrate the versatility and potential for adaptation of these solutions by applying them in two distinct cases: bird song classification and Hainan gibbon vocalisation detection. Their applicability extends beyond these specific cases, and could be applied to other calling species, offering valuable insights and opportunities for future investigations.

2. Related literature

Despite the potential for improved performance in bioacoustics deep learning classifiers by integrating contextual information, this approach remains underexplored in the scientific literature. For the purpose of our study and within the scope of this literature review, we only focused on bioacoustics studies that used deep learning techniques for the classification or detection of animal calls. To the best our of knowledge, only three studies have attempted to include contextual information in bioacoustics classifiers (Lostanlen et al., 2019; Madhusudhana et al., 2021; Roch et al., 2021). Lostanlen et al. (2019) and Madhusudhana et al. (2021) added long-term temporal context to a CNN to detect bird calls and whale songs, respectively. In the first study, the authors employed a conventional CNN to detect avian flight calls from 60 ms windows, while simultaneously training an auxiliary network on long-term summary statistics of 30 min windows to learn the representation of acoustic noise. The outcomes of these two networks were merged using various tested equations to acquire the predicted probability of presence. The authors were able to improve the performance of the classic CNN by incorporating this auxiliary model. It helped minimize spatial fluctuations in background noise that arise from using sensors at different locations. In the whale song detection study by Madhusudhana et al. (2021), a two-stage approach was used. First, a CNN was trained to detect song notes in short audio segments of 4 s. Then, a recurrent network was trained on the output sequences of the CNN in the second stage, with a duration of 111 s. By combining the two models, a hybrid model was created that takes a spectrogram as input and outputs the probability of the presence or absence of whale songs. This approach enabled the model to not only consider short-duration features, but also the temporal pattern between song notes, resulting in improved accuracy in detecting whale songs. In a distinct approach, Roch et al. (2021) trained a neural network consisting of two dense layers to detect whale echolocation clicks directly from the raw sound waveforms with a duration of 500 μs. In order to incorporate context, they computed the signal-to-noise ratio for each segment and used the time series appended to its signal-to-noise ratio as input. In the three studies, the authors concentrated on utilizing the contextual information embedded in the soundscape, which was present in the model input data but required distinct processing and integration techniques to be effectively employed by the model.

In recent years, several applications have emerged that allow the general public to identify bird, frog, or bat songs using smartphones (e.g., BirdNET (Kahl et al., 2021), Merlin Bird ID,[1] Warblr (Stowell and Plumbley, 2014), WhatFrog (Tomasini et al., 2017), Echo Meter Touch Bat Detector (Wildlife Acoustics), Bat Recorder,[2] FrogID,[3] etc.). While these applications aim to be versatile for widespread use, it is not uncommon to find that some of them have started to request location information in addition to acoustic recording. However, many of these applications are not openly accessible, and the details of how they utilize metadata remain undisclosed. In cases where the application is freely accessible or published, it seems that location information is often used as a pre or post processing filter rather than being incorporated into the classifier itself. For instance, WhatFrog utilizes GPS coordinates obtained from the smartphone to select a K-nearest neighbors (K-NN) algorithm from a set of K-NN classifiers, each trained specifically for frogs from each state of the United States (Tomasini et al., 2017). Similarly, BirdNET has recently integrated metadata such as time and location as input to their algorithm, although they have not provided detailed information about this aspect. Upon further investigation of their publicly available source code, it appears that a separate model has been trained on location and time of year to generate a list of bird species likely to be encountered at a given location. However, this information is not directly integrated into the classifier, and the results from the metadata model are not utilized to improve identification performance. Furthermore, while it is possible to access the architecture of the metadata model, no information regarding the training process or the data used is provided. There is currently an underutilization of metadata in the field of bioacoustics, despite the increased availability of such data through citizen science platforms (Pocock et al., 2015; Silvertown, 2009; Swanson et al., 2015). While methods have been developed to tackle this issue in image classification (Tang et al., 2015; Aodha et al., 2019; Terry et al., 2020), no clear and practical method has been proposed for bioacoustics. Therefore, this study aims to fill this gap by exploring various methods to incorporate metadata, such as time and location, into deep learning algorithms. Throughout this study, a focus has been placed on keeping the models simple and computationally efficient, ensuring their reusability and applicability to other research cases. Our open-source code is available on GitHub, and the data used is freely accessible on Zenodo.

3. Materials and methods

3.1. Data collection

To test if spatial–temporal information can enhance deep learning classifier, we created a bird song classification task on a dataset for which location can help to distinguish species. Xeno-canto is a well-known website created with the aim of sharing wildlife sounds from all over the world.[4] Founded in 2005, it now stores a particularly large dataset of bird songs recorded around the world (Vellinga and Planqué, 2015). In addition, each recording is associated with metadata such as the country of recording, latitude and longitude information, and time of recording. We therefore used this database with the primary purpose of creating a bird song classification task with species carefully selected to share similar vocal characteristics but from distinct geographical distributions. We only considered the recordings of category 'A', corresponding to loud and clear recordings of the calls.[4] The rationale was to construct an annotated dataset of good quality recordings that could be used in other studies. Furthermore, it is easier to downgrade (such as by artificially adding noise) audio files than it is to improve them (from low quality to clear signals). Thus, in our work we were able to produce audio examples of lower quality from the high quality recordings. To ensure reproducibility, we have made efforts to establish a replicable protocol for the selection of species that were included in our dataset. In the following sections, we will describe our approach in detail.

Firstly, we selected the ten most recorded families in the Passeriformes order, the most represented order in the Xeno-canto database. From each of the ten families, we again sub-sampled the ten most recorded genera. For each genus, we observed the countries of the recordings and the number of available recordings per species and country. From the information gathered, and by visually analyzing the spectrograms, we conducted a self-selection process of genera that comprised species with similar songs recorded in different regions. Our aim was to ensure that there were sufficient recordings available for each species and country, allowing us to form a comprehensive dataset. In the end, 5 genera were selected containing 22 species (Table 1). Due to the significant variation in the number of available recordings across different species, we needed to determine a suitable allocation of segments for each species. To address this, we calculated the average number of records per species and per country. For species/country pairs with a higher number of recordings than this average, we set an upper limit on the number of assigned segments to this average value. The recordings were downloaded from the Xeno-canto database in.wav format and each recording was manually annotated by labelling the start

[1] https://merlin.allaboutbirds.org.
[2] https://digitalbiology.com/bat-recorder.
[3] https://www.museum.qld.gov.au/learn-and-discover/apps/frogid-app
[4] http://www.xeno-canto.org/.

Table 1
Details regarding the dataset used in this study. The audio recordings were obtained from Xeno-canto.

Family	Species	Total number of recordings	Average number of recordings per country	Country with records
Muscicapidae	Saxicola gutturalis	8	8	Indonesia
	Saxicola rubetra	87	11	France, Germany, Ireland, Norway, Poland, Russian, Sweden, United Kingdom
	Saxicola rubicola	70	10	Belgium, France, Germany, Netherlands, Poland, Portugal, Spain
	Saxicola tectes	12	12	France
	Saxicola torquatus	8	8	South Africa
Thamnophilidae	Hypocnemis cantator	32	11	Brazil, Suriname, Venezuela
	Hypocnemis hypoxantha	29	10	Brazil, Ecuador, Peru
	Hypocnemis peruviana	33	11	Brazil, Ecuador, Peru
	Hypocnemis striata	11	11	Brazil
Fringillidae	Serinus canicollis	14	14	South Africa
	Serinus serinus	83	14	France, Germany, Italy, Poland, Portugal, Spain, Netherlands, Ireland
Turdidae	Catharus aurantiirostris	68	14	Colombia, Costa Rica, Hondura, Mexico, Panama
	Catharus bicknelli	6	6	United States
	Catharus fuscater	29	7	Colombia, Costa Rica, Ecuador, Panama
	Catharus fuscescens	24	12	Canada, United States
	Catharus guttatus	37	12	Canada, United States, Mexicos
	Catharus minimus	16	16	United States
	Catharus ustulatus	33	16	Canada, United States
Troglodytidae	Troglodytes aedon	120	20	Brazil, Chile, Colombia, Ecuador, Mexico, United States
	Troglodytes hiemalis	39	19	Canada, United States
	Troglodytes pacificus	35	17	Canada, United States
	Troglodytes troglodytes	173	19	Belgium, France, Germany, Ireland, Netherlands, Poland, Portugal, Spain, United Kingdom

and stop time for every vocalisation occurrence using Sonic Visualiser (Suppl. Fig. 1, Cannam et al. (2010)). In total, we obtained 6,537 occurrences of bird songs of various lengths from 967 file recordings (Table 1).

3.2. Data pre-processing

Each recording was downsampled to 22,050 Hz and converted into a mel-scale spectrogram to be used as an input image to a 2-D Convolutional Neural Network (CNN). We performed a Hann analysis window size of 46 ms (1,024 samples) with a hop size of 12 ms (256 samples) and 128 mel frequency bins. Following Kahl et al. (2021), we restricted the frequency range of the spectrograms between 150 Hz and 15 kHz as most bird vocalizations occur between 250 Hz and 8.3 kHz (Kahl et al., 2021; Hu and Cardoso, 2009). Using our manual annotations, the songs were extracted from the spectrograms and divided into equal 3 s segments using a sliding window with an overlap of 1 s. Each spectrogram was thus an image of size 128×259. Additionally, each had corresponding metadata based on the recording, this included the latitude, longitude, date, time, and country. Examples of spectrograms for each species can be visualized in the supporting information (Suppl. Fig. 2).

We obtained an average of 129 segments per species and country, with a large standard deviation (std $= 119$, min $= 7$, max $= 721$). Our dataset was therefore highly imbalanced, thus, it was important to balance it as imbalanced classes can have an impact on the performance of the classifier (Johnson and Khoshgoftaar, 2019). To achieve the class balance, we either applied data augmentation or data reduction so that 150 segments were available per species and country. The reduction involved random sampling so that only 150 segments remain. For data augmentation, we artificially added new samples to reach 150 segments using five different methods, namely, time shifting, blending, adding noise, time and frequency masking. Time shifting involved randomly selecting a time point within the segment, and shifting the start of the segment to that time point, wrapping back on itself so that it remained 3 s. Blending involved randomly selecting two segments from the same species and country, and blending them together using the following formula: $\alpha \times x_{s1} + (1-\alpha) \times x_{s2}$, where x_{s1} and x_{s2} are the two randomly selected segments, and α is an adjustable weight. We set $\alpha = 0.4$ through preliminary experimentation. To add noise to a segment, we generated random samples from a normal distribution (mean of 0 and standard deviation of 1) and added them to the original segment with a separate factor of 0.009 – again this parameter was determined through preliminary experimentation. Time masking involved applying a mask, $t \in U(0, 100)$, to a spectrogram for which the start of the mask is randomly selected as $t_\theta \in [0, L-t]$, where L is the length of the spectrogram. Thus, $[t_\theta, t_{\theta+t})$ is masked out and replaced with a value of zero. Finally, frequency masking is similar to time masking, however, it is applied on the frequency channels. All five augmentation data methods were applied at least once with distinct randomly selected segments for each method and were all repeated if necessary to reach 150 segments.

3.3. Incorporating location information into bird song classifier

3.3.1. Case I: Baseline

To assess the impact of incorporating metadata into our neural network, we established a baseline model without spatial–temporal information. The rationale behind this study is to prioritize simplicity in the models, necessitating minimal neural network parameters, thereby enabling their utilization on computationally constrained systems. This is crucial as access to graphical processing units is not always feasible. Additionally, using a CNN with a large number of network parameters, such as ResNet with millions of parameters, can lead to overfitting when

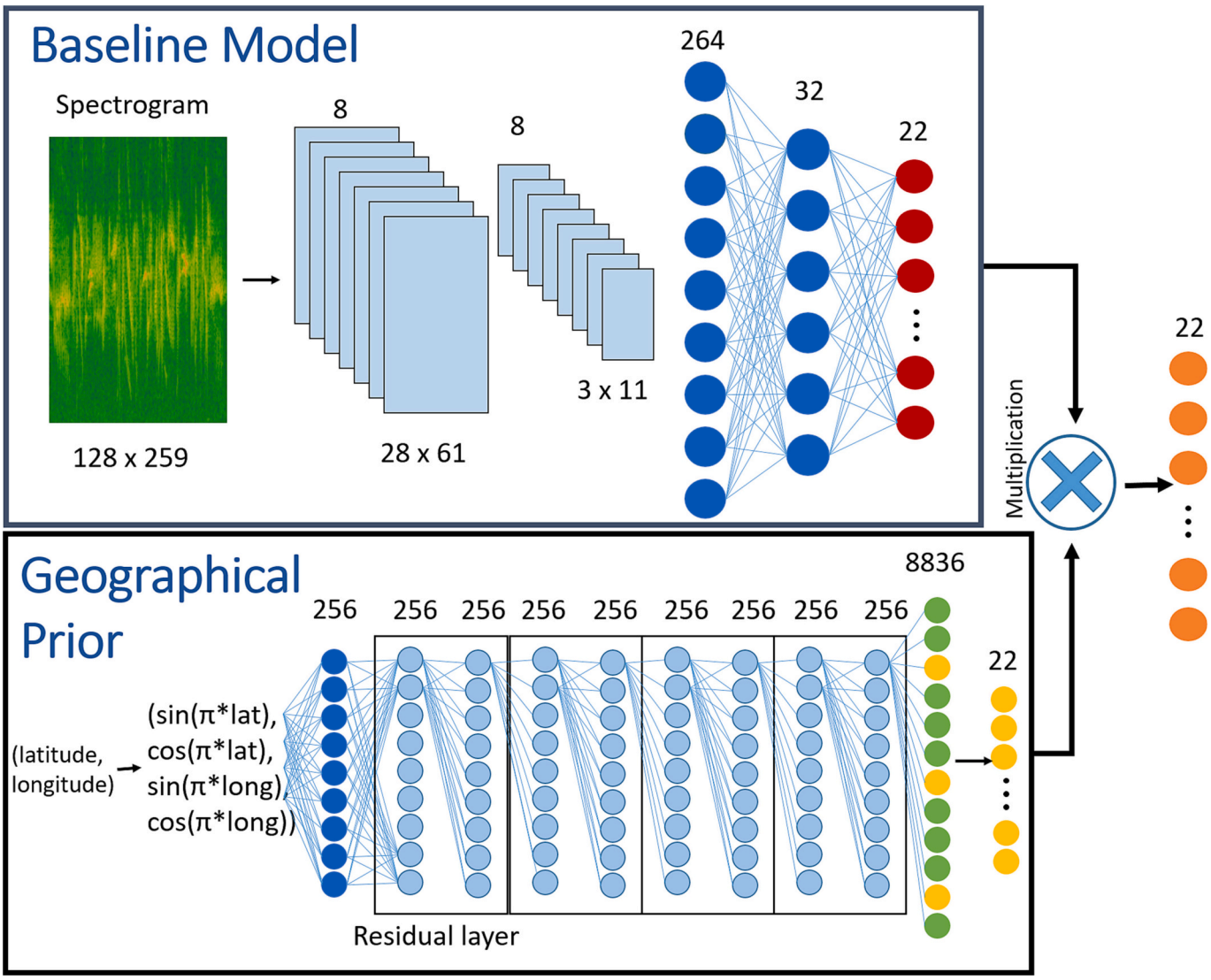


Fig. 2. Architecture of the model for Case IV; the baseline model (top) and the geographical prior (bottom) trained separately. The numbers indicate the size of each layer. From the probabilities obtained for the 8,836 species (represented in green), only the 22 species involved in our study (represented in yellow) are kept and multiplied with the corresponding outputs of the baseline model (represented in red).

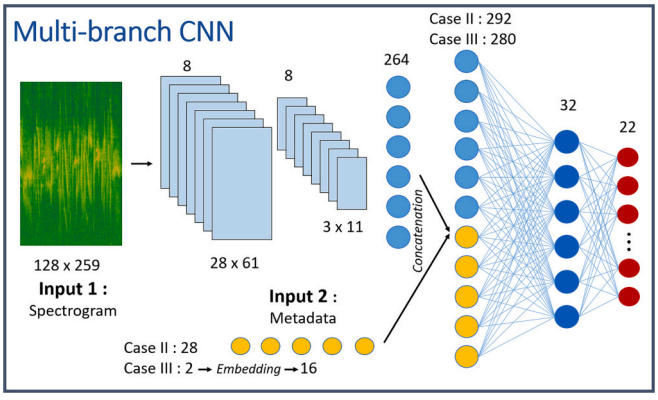


Fig. 3. Architecture of the multi-branch CNN used with different pre-processing of the location metadata. Case II uses the one-hot encoded vectors as input, and case III uses the metadata embeddings. The numbers indicate the size of each layer. For Case III, the country names are represented by 2 integers. The embedding dimension is 8, thus producing 16 output values after flattening.

the network is trained on a relatively small number of examples, which is often the case in ecology datasets. We conducted preliminary experiments on enhancing deep learning with metadata using the Hainan gibbon case study and the network developed by Dufourq et al. (2021). Our findings revealed that this architecture, applied to the bird classification task, resulted in good results and thus we adopted this straightforward approach. The proposed methodology can be adapted to other CNN architectures.

The baseline model comprised a simple CNN architecture that had two convolutional layers (8 filters of size 16×16, with ReLU activations), followed by max pooling (4×4), a flattening operation, and two fully-connected layers (32 ReLU and 22 softmax units respectively) (see Fig. 2). Given only a spectrogram input, the baseline model produced a probability for each of the 22 species.

3.3.2. *Case II: One-hot encoding*

Different methods were explored and tested to enhance the neural network with contextual information. One straightforward approach to incorporate additional information into a deep learning classifier, while maintaining model simplicity, is to utilize a multi-branch CNN. As a result, we investigate this approach as the initial method to integrate

contextual information into a deep learning classifier. Therefore, in Case II, we tested a two-branch CNN that receives two inputs, namely a spectrogram and the associated metadata as second input (Fig. 3). We had to encode the metadata in a manner that would enable it to be processed by the CNN. Thus, we assigned a unique number to each country used in this study ($n = 28$) and converted the number into a one-hot encoded vector. This encoding signifies a binary vector representation of categorical variables, in which the position of the number 1 in the vector corresponds to the index of the associated variable, and the remaining values are assigned to 0. We used the same architecture from Case I for the main CNN branch which takes spectrogram input. For the second metadata input branch, we concatenated the one-hot encoded vectors with the flattened output of the convolution layers from the main branch (Fig. 3). The new concatenated vectors were then used as input to the fully-connected layers.

3.3.3. Case III: Metadata embedding

While encoding metadata into one-hot vectors is simple, it can result in high-dimensional vectors with an increasing number of represented countries. Word embeddings have gained attention with the development of natural language processing (i.e. the understanding, manipulation and generation of natural language by machines). It consists in mapping the words to continuous number vectors and allows for a reduction in the dimensionality of the categorical variables while keeping meaningful information in the transformed space. As other studies may include a larger number of countries and therefore generate a high-dimensional metadata vector, we explored the embedding process to reduce the dimensionality of input 2. This involved incorporating a neural network embedding layer. Within this layer, trainable parameters are fine-tuned to acquire an optimal understanding of the embedding space between words throughout the training phase.

In Case III, the network takes two inputs, namely a spectrogram along with its corresponding country name as a second input. To preprocess country names, we started by assigning a unique numerical value between 0 and 50 to each word present in our country list. Since some countries are composed of two words (e.g. South Africa, United Kingdom, …), this results in a vector of size 2, with a 0 in the second position for countries with only one word (e.g. Belgium, Venezuela, …). We subsequently incorporated an embedding layer that mapped each value into an 8-dimensional transformed space, resulting in a vector of size $[2, 8]$ for each country. This dimension 8 of the embedded vector was determined through preliminary experimentation. The output from the embedding is flattened, producing 16 output values, and concatenated with the flattened output of the first branch (Fig. 3, see Suppl. Fig. 3 for a schematic representation of the country name embedding).

3.3.4. Case IV: Geographical prior

The idea of using a presence-only geographical prior was developed by Aodha et al. (2019) to enhance image classifiers. From a dataset containing images associated with their time and location of recording, two models were trained separately. Firstly, a CNN is trained on images to predict classes (e.g. cat/dog, …). Secondly, a geographical prior model is trained on time and location metadata, for which the objective of this model is to estimate the probability, for each class, that it occurs at a given location. Finally, an element-wise multiplication of the probabilities of the image classifier with those of the geographical prior is performed to obtain the final probability that an object is represented in the image knowing its location (Aodha et al., 2019).

In case IV, we replicated this method and applied it to our bird song classifier. We created a separate dataset, which we refer to as the metadata dataset, to train the geographical prior based on Xeno-canto containing only metadata; we downloaded the metadata information from all category 'A' recordings for the training dataset, representing n = 214,365 recordings of 8,994 different species, and from all category 'B' recordings for the validation dataset (n = 160,458 of 6,309 species). However, 158 species were associated with no latitude and longitude information and have been removed. We took latitude and longitude as input, which we treated similarly to Aodha et al. (2019). Assuming the latitude is denoted as x and longitude as y, then we calculated four values, namely $\sin(\pi x), \cos(\pi x), \sin(\pi y), \cos(\pi y)$, resulting in a vector of dimension four for each input and preserving the continuity of the geographic coordinates all around the earth (Aodha et al., 2019). As output, the model produces the probability of presence (close to 1) or absence for each category, with a weighted binary cross entropy as a loss function. The architecture of the geographical prior was exactly the same as proposed by Aodha et al. (2019). The first fully connected layer had 256 ReLU units, followed by four residual layers (as defined in He et al. (2016)), and a final fully connected layer of 8 836 softmax output units (Fig. 2). To balance our training metadata dataset, for each epoch, we randomly selected a maximum number of 50 data points for the over-represented categories. We used the Adam optimizer with a learning rate of 0.0005 and trained the geographic prior on 10 epochs with a batch size of 32 – obtained via preliminary experimentation.

3.4. Training and testing the models

To train and evaluate the efficiency of each method, the bird song dataset was split into a training, validation and testing dataset. For each species and country, we randomly selected 70% of the downloaded recordings for the training dataset and kept the remaining 30% for testing. Data augmentation was only applied to the training dataset, and the segments obtained were randomly distributed between the training and validation sets with a ratio 0.8/0.2. In each case, the model was trained for 40 epochs with a batch size of 8 segments and a learning rate of 0.001 using the Adam optimizer. The training process was performed ten times to account for the effect of random weight initialization in neural networks. In each execution, we applied the model to the testing dataset and recorded the confusion matrix, number of True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). From that we calculated the accuracy $((TP + TN)/(TP + TN + FP + FN))$, sensitivity or recall $(TP/(TP + FN))$, specificity $(TN/(TN + FP))$, precision $(TP/(TP + FP))$ and F1-score (harmonic mean between precision and recall). For each metric, we computed the average across ten executions to facilitate comparisons between the four methods. For case IV, the geographical prior model was trained once and the resulting probabilities were multiplied by the probabilities obtained from the baseline model. Using this single result from the geographical prior, we performed this multiplication for each output of the basic model obtained from the ten training processes and calculated the average for each metric in the same way as the other cases. The models were implemented in Python 3 using the TensorFlow and Keras libraries (Abadi et al., 2015; Chollet, 2015), and the audio processing and spectrogram construction were performed using Librosa (McFee et al., 2020).

3.5. Incorporating temporal information – a second case study

To extend our method to a new context, we investigated the effectiveness of incorporating temporal information in improving the automatic detection of vocalizations from the critically endangered Hainan gibbon (*Nomascus hainanus*, Fig. 4). Passive acoustic monitoring was employed to study the population size and dynamics of Hainan gibbons in Bawangling National Nature Reserve, China, as described by Dufourq et al. (2021). The study developed a binary classifier for automatic detection of the Hainan gibbon calls and revealed their calls were predominantly recorded in the morning with peak activity observed between 6 am to 10 am. The model produced good performance, however, there were a number of false positives that were produced on overnight recordings which contradicts the fact that these gibbons vocalise primarily in the morning.

As human practitioners, we are able to learn about this vocalising pattern, while traditional CNNs are not designed to achieve this naturally. Assuming that the environmental noise cannot be used to predict

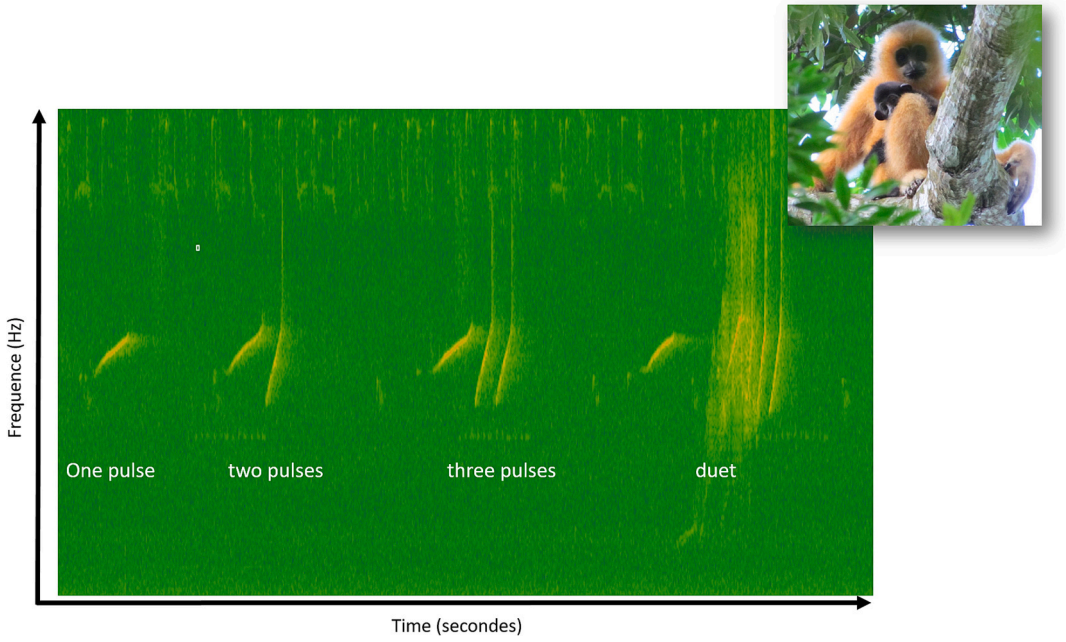


Fig. 4. Spectrogram of the Hainan gibbon calls visualized on Sonic Visualiser. The vocalizations of the Hainan gibbon consist of pulses, which are depicted in the figure. A duet occurs when both a female and a male gibbon are singing together. Hainan gibbon photo credit: Jessica Bryant, Zoological Society of London.

the time of day, then a CNN trained on spectrogram input will not be able to learn that Hainan gibbons vocalise in the morning. To explore this, we modified Case II so that the first branch takes spectrogram input and the second branch takes the time of the recording. Hence, as a second case study, we examined the effectiveness of incorporating temporal information to reduce false positives whilst preserving high precision in the passive acoustic monitoring of Hainan gibbons. The first branch of the CNN was adapted from the architecture developed by Dufourq et al. (2021). We used two convolutional layers (8x8 kernels) each followed by max pooling, one fully connected layer with 32 ReLU units, and a final fully connected layer with two softmax units.

From the original Hainan gibbon dataset (Dufourq et al., 2021), we constructed a subset such that we could train and validate the model on audio files recorded at different times throughout any given day. Using 54 audio files, we obtained 23,319 windows (each of 4 s and processed into spectrograms as described in Dufourq et al. (2021) recorded between 00:00 to 15:00 and 18:00 to 00:00. We randomly selected 16,414 training spectrograms (equal class balance between presence and absence), each of size of 128×76. The remaining spectrograms were used as a validation set to optimise hyperparameters. Similarly to how the location information was used in our birdsong case study, we modified this so that the recording times were converted into one-hot encoded vectors of size 24 (i.e. for the 24 h in a day), and the vectors were concatenated with the flattened layer from the main branch.

To determine the efficacy of our approach, we applied the model to day (10 audio files independent of the training and validation set, approximately 32 h) and overnight recordings (25 files, approximately 25 h) and compared the baseline to the enhanced model's confidence in predicting the gibbon class. We hypothesized that the enhanced model would be more confident that the overnight recordings would not contain gibbon calls. Considering the alteration made to the network, a concern arose that the network might learn that predicting Hainan gibbon calls at night is never correct. To eliminate this possibility, we meticulously designed our test set. We created a standard test set, but we deliberately added specific instances containing gibbon vocalizations (recorded in the morning), which we manipulated to appear as though they occurred at night by changing the timestamp. Consequently, if the network indeed learned that Hainan gibbon calls should never occur at night, then these particular test instances would be misclassified. Model training and implementation details were similar as described in Section 3.4.

Table 2
Average evaluation measures over ten independent executions for the four different methods. Incorporating geographical information into the CNN improves the model performance. The best result for each metric is highlighted in bold.

	Case I: Baseline	Case II: One-hot encoding	Case III: Embeddings	Case IV: Geographical prior
Accuracy	97.62% (±0.02)	98.52% (±0.01)	98.43% (±0.01)	**99.19%** (±0.01)
Sensitivity	61.34% (±0.30)	84.31% (±0.14)	81.81% (±0.17)	**86.96%** (±0.13)
Specificity	98.72% (±0.02)	99.23% (±0.01)	99.17% (±0.01)	**99.57%** (±0.01)
Precision	70.21% (±0.29)	80.20% (±0.23)	78.69% (±0.24)	**91.06%** (±0.10)
F1-score	61.02% (±0.27)	78.77% (±0.18)	76.87% (±0.19)	**87.78%** (±0.08)

4. Results

4.1. Adding location information to the bird song classifier

The four classifiers were evaluated for the classification of bird songs from 22 different species. The F1-score ranged from 61.02% to 87.78% and the addition of metadata improved classification performance in each case (Table 2). The highest accuracy and F1-score were obtained using the geographical prior (Case IV) with values of 99.19% and 87.78%, respectively. The utilization of the geographical prior model resulted in a considerable improvement in both sensitivity (25.62% points) and precision (20.85% points) as compared to the baseline model (Table 2). Species that have low F1-scores in the baseline model and are linked to only a few countries of record exhibit a notable rise in F1-scores when location information is included; for example *Catharus bicknelli* associated only with United States, *Saxicola gutturalis* with Indonesia, *Saxicola torquatus* and *Serinus canicollis* with South Africa

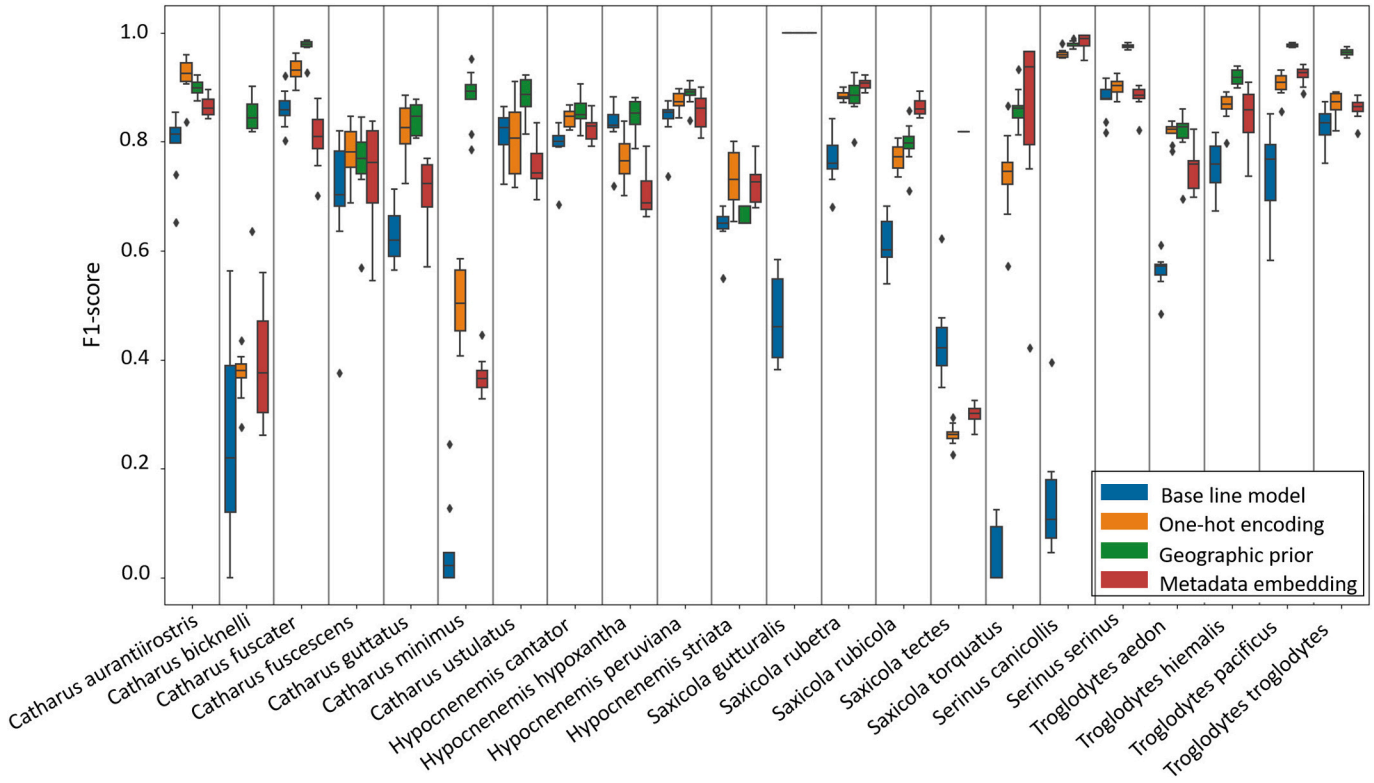


Fig. 5. Box-plot of F1-scores obtained over the ten training runs for each species according to the model architecture.

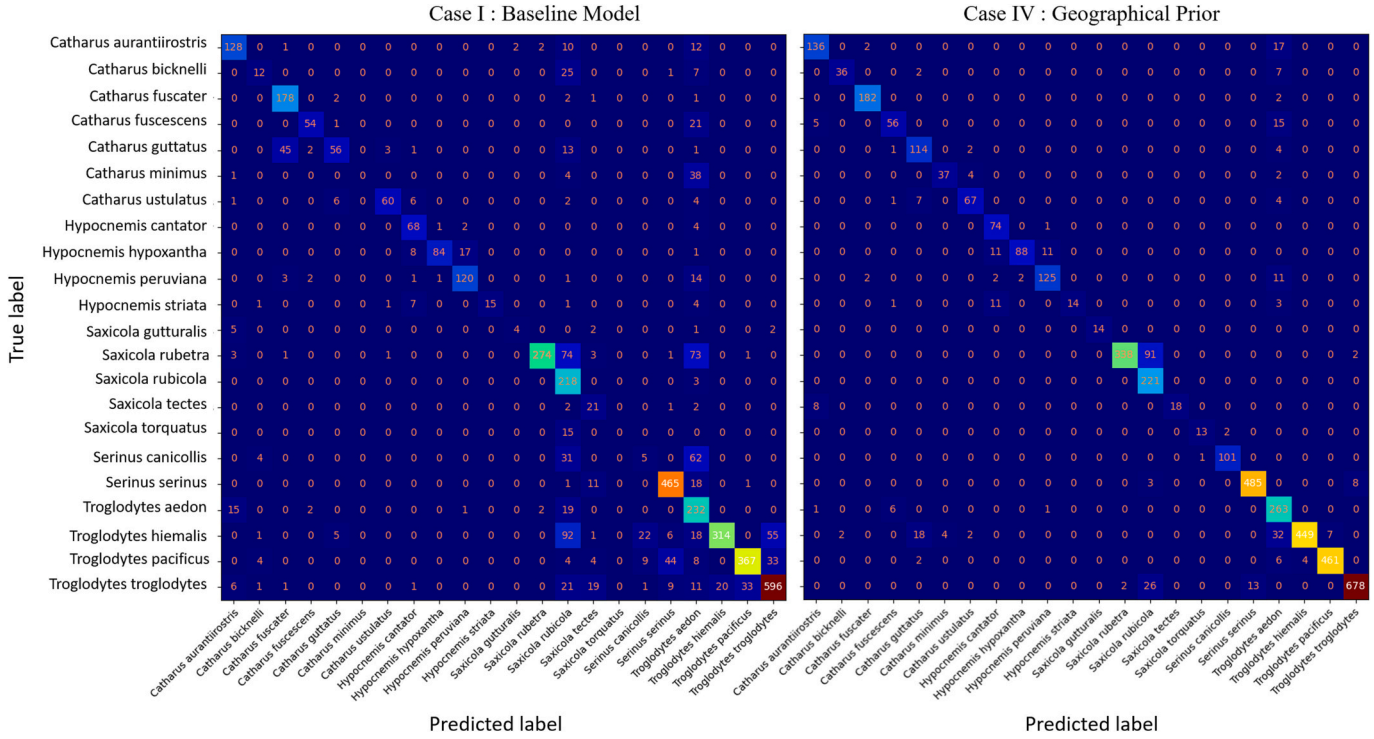


Fig. 6. Confusion matrix associated with the predictions of the baseline model (Case I, left) and the geographical prior (Case IV, right).

(Fig. 5, Table 1). Only the F1-score of *Saxicola tectes* decreased in Case II and Case III due to an increase in misclassification with *Serinus serinus*, a species found in the same country (Fig. 6, Table 1).

The use of the geographical prior model led to an important reduction in misclassifications compared to the baseline model, as evidenced by the confusion matrix (Case IV, Fig. 6). Notably, misclassifications occurred primarily between bird species with overlapping distribution areas, such as *Saxicola rubetra* and *Troglodytes troglodytes*, which were both misclassified as *Saxicola rubicola* in the European region. Similarly, species such as *Catharus aurantiirostris*, *Catharus fuscescens*, and

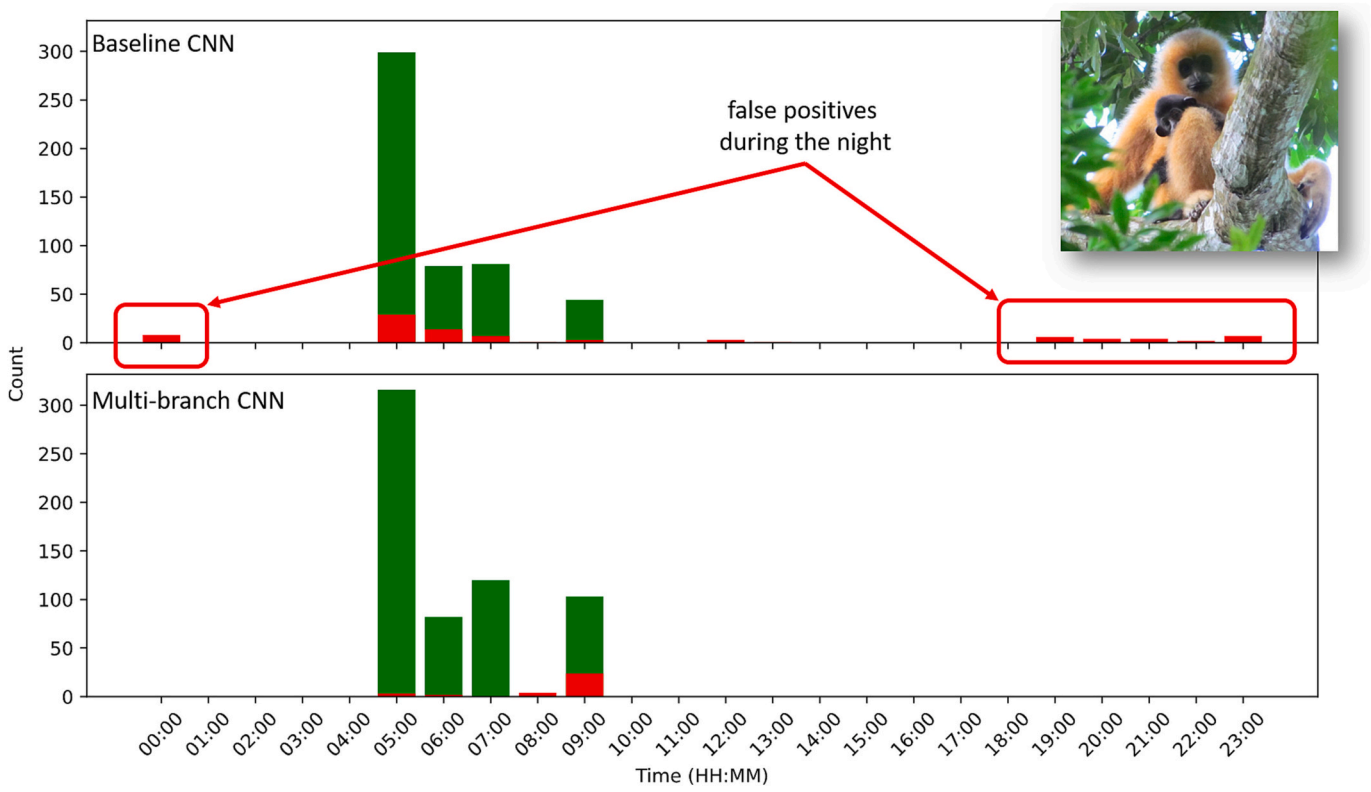


Fig. 7. Comparison between the baseline (top) and the multi-branch CNN (bottom). True detected gibbon calls are shown in green, and false positives are shown in red. The multi-branch CNN does not produce any false positives during the night, results in an overall reduction of false positives by 63% and increase in detected gibbon calls by 19%.

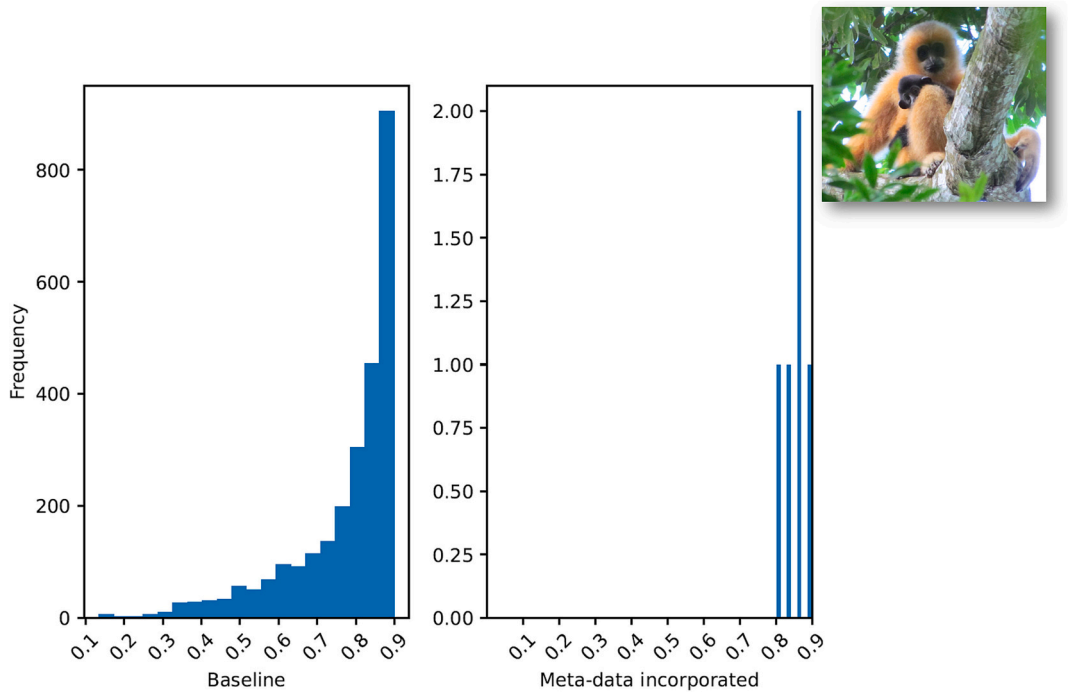


Fig. 8. The cumulative probabilities obtained from the softmax function are displayed for the absence of the Hainan gibbon calls class. A baseline (left) and enhanced model (right) were applied to overnight recordings. A higher probability value denotes strong confidence that a segment did not contain a call. Probabilities greater than 0.9 are not displayed. The baseline model showed a large number of spectrogram segments for which it was not strongly confident that there was an absence of Hainan gibbon vocalization. On the other hand, for the enhanced model, there were only 5 spectrograms for which the model's confidence for the absence class was 0.8, and thus the remaining probabilities were all greater than 0.9. Hainan gibbon photo credit: Jessica Bryant, Zoological Society of London.

Troglodytes hiemalis were misclassified as *Trogolodytes aedon*, as their distribution areas overlap in the region from Brazil to the United States.

The baseline model made a number of misclassifications where the calls, for some genera, were classified as different genera. This observation was reduced in the geographical prior model where the majority of the misclassifications were observed within the correct genus. For the *Catharus* genus, the baseline misclassified 153 calls across 7 incorrect genera, whereas the geographical prior misclassified 51 calls across 1 incorrect genus (*Troglodytes aedon*). A similar observation is made for calls from the *Hypocnemis* genus, the baseline model misclassified 32 calls as 6 incorrect genera, and again the geographical prior misclassified 17 calls as 3 incorrect genera. Similar observations were made for *Saxicola*, *Serinus* and *Troglodytes*, where the baseline made mistakes with 7, 5 and 12 genera respectively, and the geographical prior made mistakes with 3, 3 and 10 genera respectively.

4.2. Adding temporal information to the Hainan gibbon call detector

The baseline and the multi-branch model which used temporal information were applied in three separate cases. Firstly, on audio files recorded only in the day (from 05:00 am onwards), audio files recorded overnight (from 20:00 onwards), and the same day audio files which were renamed as overnight recordings. In the first case, both the baseline and the enhanced model obtained the same performance, with an F1-score of 0.9. In the second case, the baseline produced 31 false positive bouts on the overnight recordings, whereas the enhanced model did not produce any false positive bouts. Finally, in the third case, both models produced the same performance, with an F1-score of 0.9. A comparison between the baseline and the multi-branch model is presented in Fig. 7 grouped across the testing data. In the end, over the whole testing dataset (10 daytime audio files and 25 overnight), the multi-branch model correctly detected 19% more gibbon calls and reduced the false positive by 63%. Fig. 8 shows the cumulative probabilities for the absence class when the two models were applied to the overnight recordings, giving an indication of the confidence of the model in identifying windows without gibbon calls. Probabilities greater than 0.9 are not displayed as we were interested in cases where the model is not completely confident in the classification. There were 2,629 segments for which the baseline had a probability of less than 0.9, whereas the enhanced model only had 5 segments. Thus, in general, the baseline model is less confident than the enhanced model in identifying the absence of Hainan gibbon vocalization, hence the baseline model produced a higher number of false positives. There were 182 segments for which the baseline probability was less than 0.5, and thus the baseline would have incorrectly predicted that a gibbon was present.

5. Discussion

The objective of this study was to assess whether deep learning algorithms can utilize contextual information for improved animal call detection and classification. The investigation of contextual metadata incorporation in deep learning classifiers for bioacoustics has been limited in prior literature, despite the potential for significant improvement. Our study demonstrated that techniques commonly used in image processing can be adapted and effectively applied in bioacoustics for this purpose. Based on a dataset built specifically for this purpose, we found that augmenting spatial information with spectrograms improved the accuracy of bird song classification, with higher sensitivity and precision indices than the baseline model (Table 2). Specifically, we achieved a significant improvement in the F1-score of bird song classification of 22 species from 61.02% (baseline) to 87.78% by separately training a geographical prior on longitude and latitude data (to predict the probability of species occurrence at a given location) and multiplying it with the output of a spectrogram classifier. We also tested several methods based on multi-branch CNNs, which yielded improved results and are simpler to implement than the geographical prior. Additionally, our results on the gibbon dataset indicated that a multi-branch CNN enhanced with temporal information reduced the number of false positives for Hainan gibbon call detection by 63%.

The classification of bird songs is challenging, due to the fact that certain bird species possess rich and diverse song repertoires featuring highly complex sound sequences (Suppl. Fig. 2, Samotskaya et al. (2016)). For the baseline model, the main errors mostly concerned a multitude of species wrongly labelled as *Saxicola rubicola* and/or *Troglodytes aedon* (Fig. 6). An analysis and visualization of the spectrograms did not provide an understanding of the origin of these errors; further studies are warranted to fully understand the nature of these errors. Nevertheless, the confusion matrix associated with the geographical prior (Case IV) showed confusion only for species with countries in common in their spatial distribution (Fig. 6 and Table 1). The misclassification for species found in distinct regions is no longer present which allows us to claim that the complementary information of the location is at the origin of this improvement. This is also the case for the multi-branch CNN (Case II and Case III) with an increase of F1-score, precision and specificity index and a reduction in misclassification between species found within a distinct spatial distribution. However, in Cases II and III, we observed a higher rate of misclassification between species that share the same countries of recording, as compared to the baseline model. For example, we can notice an increase in misclassification of *Catharus minimus*, *Troglodytes aedon*, *Troglodytes hiemalis*, *Troglodytes pacificus*, wrongly classified as *Catharus bicknelli*, in case II and case III compared to the baseline model (Suppl. Fig. 4). And all these five species can be heard in the United States (Table 1). It is reasonable to assume that the model places greater weight on metadata when we merge the information into a single model. However, we anticipate this weight to be secondary as deliberate modifications to the metadata during the gibbon call detection task demonstrated that the model was still able to detect calls during times when gibbons typically do not vocalize. This confirms that the metadata was utilized as supplementary information and that the spectrograms remained the primary source of information for detection.

The geographical prior was originally developed to improve image classification by adding spatial–temporal information. We showed in this study that it can be suitably adapted to realise the same task in bioacoustics classification. It is therefore a valuable methodology that can be used in different contexts with several advantages. Firstly, the geographical prior has the advantage of being distinct from the classifier and can therefore be run independently. This implies that we can add new species to the classifier without re-training the geographical prior; our geographical prior was trained on a larger dataset ($n = 8{,}836$ species for this study). Furthermore, an acoustic recording without metadata can still be classified. Secondly, the geographical prior allows to exploit a large amount of contextual data collected all over the world, available but generally underutilized. In this study, we trained the algorithm on the citizen-based dataset Xeno-canto for which more than 374,823 metadata instances were available. As the geographical prior is only trained on location and does not require the associated acoustic recording, a wide variety of databases could be used. As it has become common to associate metadata information with images, image collections are other potential sources to acquire training data. Therefore, iNaturalist (130,714,554 images of 421,616 species, Berg et al. (2014)) and BirdSnap (49,829 images from 500 bird species), are examples of databases that can be used for this purpose. As scientists are increasingly resorting to citizen science projects (Pocock et al., 2015; Silvertown, 2009; Swanson et al., 2015), where volunteers contribute to the collection, the number of available databases should increase. Thirdly, the geographical prior can be used to visualise species distribution (Fig. 9). Species distribution modelling relies largely on the observation of species at a given location. Thus, each observation of a species by the citizens and recorded within this citizen-based dataset with the location is proof of the presence of the species in this area and can be valued by the geographical prior. Temporal information can also be added as

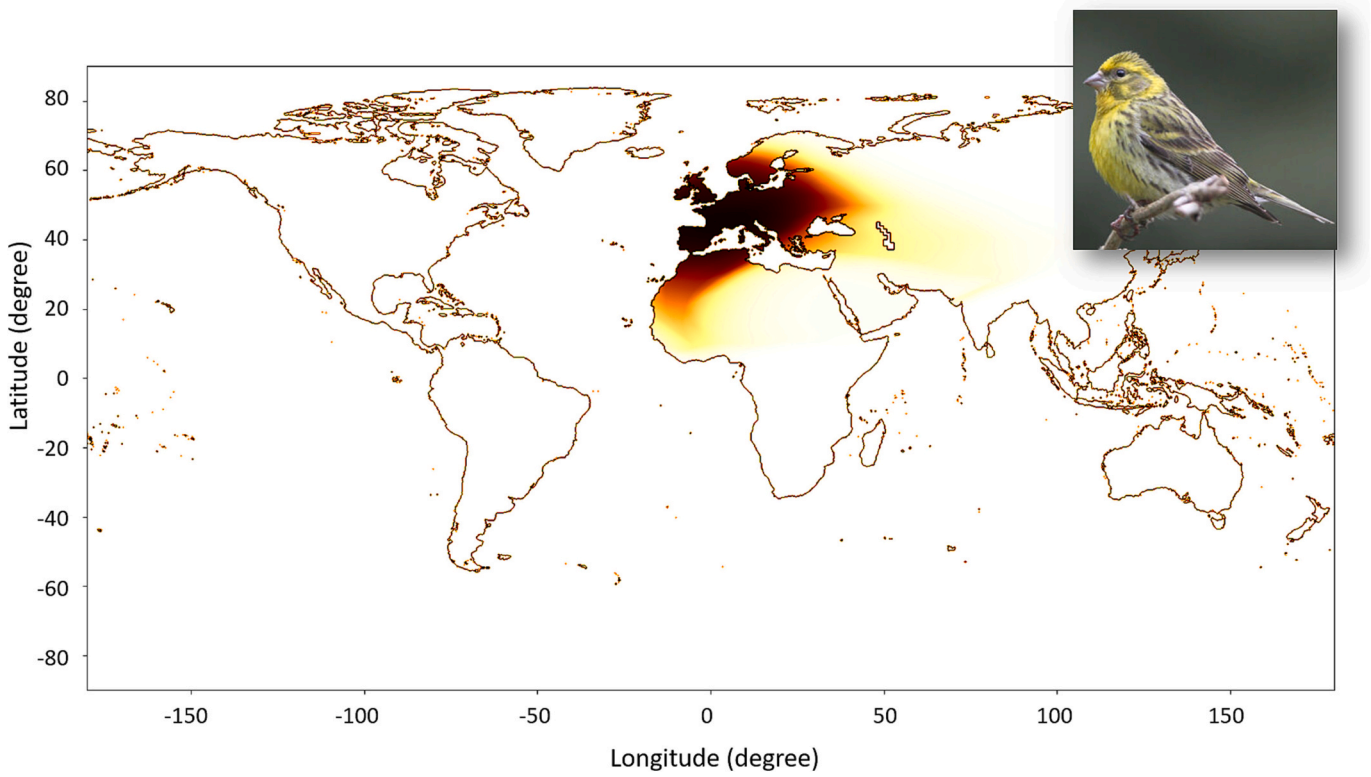


Fig. 9. Predicted distribution of *Serinus Serinus* by the geographical prior trained on the entire Xeno-canto database (quality 'A' and 'B'). Dark colors indicate locations with a high probability of species occurrence based on the observations collected on Xeno-canto.

metadata to the geographical prior and allows for the integration and visualization of migratory events (Aodha et al., 2019).

Therefore, the geographical prior is a valuable tool that can be used to enhance information stored in citizen-based datasets. However, these databases based on citizen collection may also present biases in representativeness of the species as some rare and/or difficult to record species may be underrepresented or even absent (Tulloch and Szabo, 2012; Boakes et al., 2016). Added to this is the fact that the number of recordings is not equally distributed across all countries with the majority of recordings coming from America and Europe and far fewer from Africa, Asia and Australasia (see the total number of recordings per region in xeno-canto.org). In another context, the application of the geographical prior can be limited by the availability of a training dataset. Birds are among the most represented species in citizen-based datasets with large available data sets but this is not the case for all taxa (Troudet et al., 2017). Moreover, the geographical prior is particularly useful for the classification task but can be more difficult to adapt to the detection task. In these contexts, other methods of metadata integration must be implemented.

We tested two multi-branch CNN that used the name of the country of the recording as the second input. Multi-branch CNNs are widely used methods in various domains, as they result in improved classification by adding as much additional information as desired in a simple way (Aslani et al., 2019; Georgakilas et al., 2020; Yan et al., 2022). For example, Yan et al. (2022) obtained better classification results for three types of ion channel peptide binders with a multi-branch CNN, than an ensemble method based on thirteen traditional machine learning algorithms. It appears therefore to be a relevant solution to integrate metadata into acoustic classifiers. In bioacoustics, to the best of our knowledge, this is the first time that multi-branch CNNs are investigated as a means to improve acoustic classification. In our case, multi-branch CNNs (Case II and Case III) have the advantage of being easy to implement while enabling a simple CNN to be improved whilst also keeping the number of network parameters low (Case I: 27 654 parameters, Case II: 28 550, Case III: 28 556 and Case IV: 2 817 286). However, they cannot classify data without associated metadata and must be retrained each time new data or species are added.

To incorporate spatial information into the multi-branch CNNs, we used the country name as metadata and evaluated two different methods: one-hot encoding (Case II) and embeddings (Case III). Case II was implemented first, as it was a simple way to add spatial information. This approach improved the accuracy and F1-score by 0.9% and 17.7% points, respectively (Table 2). While simple, it can result in a high-dimensional vector with an increasing number of represented countries. Our case study includes only 28 different countries, resulting in a metadata input vector size of 28. However, since the intention is to apply the method to a broader context, the size of the metadata vector can quickly expand to over 100, depending on the study scale and the number of countries considered. This would significantly increase the number of inputs and the corresponding number of trainable weights in the subsequent fully-connected layers. However, in the field of ecology, data scarcity poses a common challenge, often leading to the preference for simple models over larger ones. Large models typically require extensive training sets to mitigate the risk of overfitting (Hoffmann et al., 2022). To address this, we investigated Case III as a next step. In this case, the model learns the best representation of the country names in a user-defined dimension space through an embedding layer. Preprocessing the country names using embeddings reduced the dimension of the categorical variables independently of the number of countries used in the study (vector size of the metadata input is 28 in Case II and 16 in Case III, Fig. 3). Additional testing may be necessary to identify the optimal dimension size.

While using the country name as a second input in the multi-branch CNN is a simple and straightforward way of adding contextual information, neither the embedding nor the one-hot encoded vectors take into account the proximity between countries. A particular species may be observed in two neighboring countries, however, the algorithm lacks awareness of their close proximity, and thus treats the two countries as

independent entities with no understanding of the geographical proximity. On a finer scale, various species may exhibit similar songs despite evolving in different regions within the same country. For instance, the Northern Mockingbird (*Mimus polyglottos*) and the Brown Thrasher (*Toxostoma rufum*) are two bird species in the United States known for their mimicry abilities. Individuals of both species occasionally sing apparent imitations of each other's songs (Boughey and Thompson, 1975). However, the Brown Thrasher's habitat is limited to the eastern part of the country (Cavitt and Haas, 2020). Therefore, working at the country name scale alone would provide limited additional information for these two species, whereas latitude and longitude would enable us to work at a more fine-grained scale and discern them in certain situations. In Case IV, the geographical prior, trained on latitude and longitude, has a better overview of species distribution and relationships between countries (Fig. 9). While a geographical prior is more complex to apply than the multi-branch CNN, future work could investigate embedding methods applied directly to latitude and longitude metadata to learn geographical relationships. As such, the embedding would serve both to perform dimensionality reduction and to learn relationships between locations.

Adding temporal information to the Hainan gibbon classifier completely removed the number of false positives produced during the overnight recordings compared to the baseline model, and previous research (Dufourq et al., 2021), which generated false positives on overnight recordings. The ability to detect Hainan gibbon calls was not impacted by adding temporal information – the F-1 score remained the same for audio files which were recorded from the morning onwards. We observed an improvement in gibbon detection. It was assumed that perhaps the enhanced model would simply convert any potential false positive during the overnight recordings into a true negative using the extra branch. However, our experiments revealed that this was not the case and that the features generated from the spectrogram branch were still used to predict the presence of a gibbon call even if we manipulated the time of the recording to appear as though it occurred at night. When we applied the enhanced model to the overnight recordings, the model was very confident that there were no gibbon calls, with nearly all of the spectrograms yielding a probability of 0.9 or greater (Fig. 8). Our findings reveal that incorporating temporal information was useful in this case study to reduce the number of false positives. While it might not be the case that all species only vocalize within certain times such as the Hainan gibbons, our experiments reveal that by using human knowledge, we were able to adapt our network to learn patterns that human practitioners would use and enable the network to leverage this to improve classification performance.

Incorporating metadata into deep learning algorithms is a promising approach for improving bioacoustics classification and detection performance, with wider ecological applications beyond the scope of detection, such as occupancy patterns and spatial distribution, among others (Gibb and Browning, 2019; Ross et al., 2023; Kvsn et al., 2020). Our study demonstrated the effectiveness of adding spatial information to improve classification performance on 22 bird species, and temporal information to reduce false positives in the detection of Hainan gibbon calls. Although bioacoustics detectors tend to be sensitive and overestimate the number of calls, reducing false positives is critical for accurate analyses of population dynamics and habitat use. While checking false positives is generally easier than checking false negatives, it can still be a time-consuming task, highlighting the importance of decreasing false positives. Furthermore, the success of this new approach in two distinct study cases suggests its potential for easy generalization to other species.

6. Conclusion

Machine learning models in ecology often struggle with generalization. This difficulty is mainly due to the scarcity of training data from wild animals, which often lack variability within the dataset and are collected from specific areas. Consequently, detection algorithms trained on a specific area may perform poorly on new data from different areas, leading to confusion with unknown background sounds. This limits the usage of trained models and requires each study to build its own training dataset and model. In this study, we demonstrated that adding location information can reduce confusion between species found in distinct areas, leading to more general multi-area models capable of discerning the context of different areas. While location information is particularly relevant for bird classification, other metadata can be considered for different case studies. This new possibility of adding contextual information to classifiers/detectors in bioacoustics opens up exciting opportunities for improving model generalization. We encourage researchers and practitioners to explore further modifications to these networks in order to incorporate prior knowledge about the environment and animals.

Author contributions

LJ and ED conceived the project and designed the methodology; LJ generated and analysed the dataset; LJ performed the analysis; LJ and ED led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

LJ: Conceptualization, Methdology, Investigation, Data curation, Formal analysis, Writing - original draft, Writing - review & editing ED: Conceptualization, Methdology, Validation, Writing - original draf, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All code for training and testing the neural networks is available at https://github.com/AIMS-Research/research_za/tree/main/bioacoustics_contextual_analysis. A subset of acoustic recordings, including training and testing labels, has been stored on Zenodo and can be accessed via https://doi.org/10.5281/zenodo.7828148.

Acknowledgments

ED is supported by a research chairship from the African Institute for Mathematical Sciences South Africa. This work was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada, www.idrc.ca, and with financial support from the Government of Canada, provided through Global Affairs Canada (GAC), www.international.gc.ca. We thank the School for Data Science and Computational Thinking at Stellenbosch University for providing computational resources for certain aspects of this study. Computations were performed using the University of Stellenbosch's HPC2:http://www.sun.ac.za/hpc.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecoinf.2023.102256.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved fromhttps://www.tensorflow.org/.

Aodha, O.M., Cole, E., Perona, P., 2019. Presence-only geographical priors for fine-grained image classification. Proc. IEEE Int. Conf. Comput. Vis. 2019-Octob, 9595–9605. https://doi.org/10.1109/ICCV.2019.00969.

Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M.A., Sona, D., 2019. Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage 196, 1–15. https://doi.org/10.1016/j.neuroimage.2019.03.068.

Berg, T., Liu, J., Lee, S.W., Alexander, M.L., Jacobs, D.W., Belhumeur, P.N., 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In: 2014 ieee conference on computer vision and pattern recognition, pp. 2019–2026. doi: 10.1109/CVPR.2014.259.

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), 1–10. https://doi.org/10.1038/s41598-019-48909-4.

Boakes, E.H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D.B., Haklay, M., 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. Sci. Rep. 6 (August), 1–11. https://doi.org/10.1038/srep33051.

Boughey, M.J., Thompson, N.S., 1975. Species Specificity and Individual Variation in the Songs of the Brown Thrasher (Toxostoma Rufum) and Catbird (Dumetella Carolinensis). Behaviour 57 (1–2).

Cannam, C., Landone, C., Sandler, M., 2010. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In: Proceedings of the 18th acm international conference on multimedia. Association for Computing Machinery, New York, NY, USA, pp. 1467–1468. https://doi.org/10.1145/1873951.1874248.

Cavitt, J.F., Haas, C.A., 2020. Brown Thrasher (Toxostoma rufum). In: Birds of the world. Cornell Lab of Ornithology, Ithaca, NY, USA. doi: 10.2173/bow.brnthr.01.

Chollet, F., 2015. Keras. Retrieved fromhttps://keras.io.

Chollet, F., 2018. Deep learning with Python. Manning Publications Co., Shelter Island, NY11964.

Dorian, C., Lefort, R., Bonnel, J., Zarader, J.L., Adam, O., 2017. Bi-class classification of humpback whale sound units against complex background noise with deep convolution neural network.

Dufourq, E., Durbach, I., Hansford, J.P., Hoepfner, A., Ma, H., Bryant, J.V., Turvey, S.T., 2021. Automated detection of Hainan gibbon calls for passive acoustic monitoring. Remote Sens. Ecol. Conserv. 7 (3), 475–487. https://doi.org/10.1002/rse2.201.

Georgakilas, G.K., Grioni, A., Liakos, K.G., Chalupova, E., Plessas, F.C., Alexiou, P., 2020. Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. Sci. Rep. 10 (1), 1–10. https://doi.org/10.1038/s41598-020-66454-3.

Gibb, R., Browning, E., 2019. Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring. Methods Ecol. Evol. 2019 (10), 169–185. https://doi.org/10.1111/2041-210X.13101.

Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., Joly, A., 2016, September. LifeCLEF Bird Identification Task 2016: The arrival of Deep learning. In: CLEF: Conference and Labs of the Evaluation Forum, vol. CEUR Workshop Proceedings. Évora, Portugal, pp. 440–449.

Goëau, H., Glotin, H., Vellinga, W.P., Planqué, R., Rauber, A., Joly, A., 2014. LifeCLEF Bird Identification Task 2014. In: CLEF: Conference and Labs of the Evaluation Forum, vol. CEUR Workshop Proceedings. Sheffield, United Kingdom, pp. 585-597.

Grill, T., Schlüter, J., 2017. Two convolutional neural networks for bird detection in audio signals. In: 2017 25th european signal processing conference (eusipco), pp. 1764–1768. doi: 10.23919/EUSIPCO.2017.8081512.

Hassan, N., Ramli, D.A., Jaafar, H., 2017. Deep neural network approach to frog species recognition. In: 2017 ieee 13th international colloquium on signal processing & its applications (cspa), pp. 173–178. doi: 10.1109/CSPA.2017.8064946.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/CVPR.2016.90.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97. https://doi.org/10.1109/MSP.2012.2205597.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Sifre, L., 2022. An empirical analysis of compute-optimal large language model training. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (Eds.), Advances in neural information processing systems, vol. 35. Curran Associates, Inc., pp. 30016–30030

Hu, Y., Cardoso, G.C., 2009. Are bird species that vocalize at higher frequencies preadapted to inhabit noisy urban areas? Behav. Ecol. 20 (6), 1268–1273. https://doi.org/10.1093/beheco/arp131.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1). https://doi.org/10.1186/s40537-019-0192-5.

Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., Müller, H., 2021. Overview of LifeCLEF 2021: an evaluation of Machine-Learning based Species Identification and Species Distribution Prediction. In: Candan, K.S., et al. (Eds.), CLEF 2021–12th International Conference of the CLEF Association, vol. LNCS. LNISA - 12880. Springer International Publishing, Virtual Event, France, pp. 371–393. https://doi.org/10.1007/978-3-030-85251-1_24.

Joly, A., Goëau, H., Kahl, S., Picek, L., Lorieul, T., Cole, E., Hrúz, M., 2022. Overview of LifeCLEF 2022: An Evaluation of Machine-Learning Based Species Identification and Species Distribution Prediction. In: Barrón-Cedeño, A., et al. (Eds.), CLEF 2022–13th International Conference of the CLEF Association, vol. LNCS-13390. Springer International Publishing, Bologna, Italy, pp. 257–285. https://doi.org/10.1007/978-3-031-13643-6_19.

Kahl, S., Wood, C.M., Eibl, M., Klinck, H., 2021. BirdNET: A deep learning solution for avian diversity monitoring. Ecol. Inform. 61 (December 2020), 101236 https://doi.org/10.1016/j.ecoinf.2021.101236.

Khalighifar, A., Jiménez-García, D., Campbell, L.P., Ahadji-Dabla, K.M., Aboagye-Antwi, F., Ibarra-Juarez, L.A., Peterson, A.T., 2021. Application of deep learning to community-science-based mosquito monitoring and detection of novel species. J. Med. Entomol. 59, 355–362.

Kohlsdorf, D., Herzing, D., Starner, T., 2020. An auto encoder for audio dolphin communication.

Kvsn, R.R., Montgomery, J., Garg, S., Charleston, M., 2020. Bioacoustics data analysis-a taxonomy, survey and open challenges. IEEE Access 8, 57684–57708. https://doi.org/10.1109/ACCESS.2020.2978547.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Li, H., Tian, S., Li, Y., Fang, Q., Tan, R., Pan, Y., Gao, X., 2020. Modern deep learning in bioinformatics. J. Mol. Cell Biol. 12 (11), 823–827.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., Bello, P., 2019. Robust sound event detection in bioacoustic sensor networks. PLoS ONE 14 (10), e0214168. https://doi.org/10.1371/journal.pone.0214168.

Madhusudhana, S., Shiu, Y., Klinck, H., Fleishman, E., Liu, X., Nosal, E.M., Roch, M.A., 2021. Improve automatic detection of animal call sequences with temporal context. J. R. Soc. Interface 18 (180). https://doi.org/10.1098/rsif.2021.0297.

McFee, B., Lostanlen, V., McVicar, M., Metsai, A., Balke, S., Thomé, C., Weiss, A., 2020. Librosa. 10.5281/ZENODO.3606573.

Miller, B.S., Barlow, J., Calderan, S., Collins, K., Leaper, R., Olson, P., Double, M.C., 2015. Validating the reliability of passive acoustic localisation: A novel method for encountering rare and remote Antarctic blue whales. Endanger. Species Res. 26 (3), 257–269. https://doi.org/10.3354/esr00642.

Mutanu, L., Gohil, J., Gupta, K., Wagio, P., Kotonya, G., 2022. A review of Automated bioacoustics and general acoustics classification research. Sensors 22 (8361). https://doi.org/10.3390/s22218361.

Piccialli, F., Di Somma, V., Giampaolo, F., Cuomo, S., Fortino, G., 2021. A survey on deep learning in medicine: Why, how and when? Inf. Fusion 66, 111–137.

Pocock, M.J., Roy, H.E., Preston, C.D., Roy, D.B., 2015. The Biological Records Centre: A pioneer of citizen science. Biol. J. Linn. Soc. 115, 475–493. https://doi.org/10.1111/bij.12548.

Roch, M.A., Lindeneau, S., Aurora, G.S., Frasier, K.E., Hildebrand, J.A., Glotin, H., Baumann-Pickering, S., 2021. Using context to train time-domain echolocation click detectors. J. Acoust. Soc. Am. 149 (5), 3301–3310. https://doi.org/10.1121/10.0004992.

Ross, S.R., O'Connell, D.P., Deichmann, J.L., Desjonquères, C., Gasc, A., Phillips, J.N., Burivalova, Z., 2023. Passive acoustic monitoring provides a fresh perspective on fundamental ecological questions. Funct. Ecol. 37 (4), 959–975. https://doi.org/10.1111/1365-2435.14275.

Samotskaya, V.V., Opaev, A.S., Ivanitskii, V.V., Marova, I.M., Kvartalnov, P.V., Opaev, A.S., Kvartalnov, P.V., 2016. Syntax of complex bird song in the large-billed reed warbler (Acrocephalus orinus). Bioacoustics. https://doi.org/10.1080/09524622.2015.1130648.

Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. 24 (9), 467–471. https://doi.org/10.1016/j.chemosphere.2018.03.203.

Smith, A.A., Kristensen, D., 2017. Deep learning to extract laboratory mouse ultrasonic vocalizations from scalograms. In: 2017 ieee international conference on bioinformatics and biomedicine (bibm), pp. 1972–1979. doi: 10.1109/BIBM.2017.8217964.

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

Stowell, D., Plumbley, M.D., 2014. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. PeerJ 2, e488.

Sugai, L.S.M., Silva, T.S.F., Ribeiro, J.W., Llusia, D., 2019. Terrestrial Passive Acoustic Monitoring: Review and Perspectives. Bioscience 69 (1), 5–11. https://doi.org/10.1093/biosci/biy147.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C., 2015. Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. Sci. Data 2 (150026), 1–14. https://doi.org/10.1038/sdata.2015.26.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. DeepFace: Closing the gap to human-level performance in face verification. In: Proceedings of the ieee computer society conference on computer vision and pattern recognition. IEEE Computer Society, pp. 1701–1708. https://doi.org/10.1109/CVPR.2014.22.

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., Bourdev, L., 2015. Improving image classification with location context. Proc. IEEE Int. Conf. Comput. Vis. 2015 Inter, 1008–1016. https://doi.org/10.1109/ICCV.2015.121.

Terry, J.C.D., Roy, H.E., August, T.A., 2020. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. Methods Ecol. Evol. 11 (2), 303–315. https://doi.org/10.1111/2041-210X.13335.

Tomasini, M., Smart, K., Menezes, R., Bush, M., Ribeiro, E., 2017. Automated robust anuran classification by extracting elliptical feature pairs from audio spectrograms.

In: 2017 ieee international conference on acoustics, speech and signal processing (icassp), pp. 2517–2521.
Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R., Legendre, F., 2017. Taxonomic bias in biodiversity data and societal preferences. Sci. Rep. 7 (1), 1–14. https://doi.org/10.1038/s41598-017-09084-6.
Tulloch, A.I.T., Szabo, J.K., 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu - Austral Ornithol. 112 (4), 313–325. https://doi.org/10.1071/MU12009. Retrieved fromhttps://doi.org/10.1071/MU12009.
Vellinga, W.P., Planqué, R., 2015. The Xeno-canto collection and its relation to sound recognition and classification. In: Proceedings of the 2015 clef. Toulouse, France.
Yan, J., Zhang, B., Zhou, M., Kwok, H.F., Siu, S.W., 2022. Multi-branch-cnn: Classification of ion channel interacting peptides using multi-branch convolutional neural network. Comput. Biol. Med. 147, 105717 https://doi.org/10.1016/j.compbiomed.2022.105717.
Zhong, M., Castellote, M., Dodhia, R., Lavista Ferres, J., Keogh, M., Brewer, A., 2020. Beluga whale acoustic signal classification using deep learning neural network models. J. Acoust. Soc. Am. 147 (3), 1834–1841. https://doi.org/10.1121/10.0000921.