Whale-VAD: Whale Vocalisation Activity Detection

Christiaan M. Geldenhuys
cmgeldenhuys@sun.ac.za

Günther Tonitz
gtonitz@sun.ac.za

Thomas R. Niesler
trn@sun.ac.za

Department of Electrical and Electronic Engineering,
University of Stellenbosch, South Africa

***Abstract*—We present a lightweight sound event detection (SED) system focused on the discovery of whale calls in marine audio recordings. Our proposed architecture uses a hybrid CNN-BiLSTM architecture with an added residual bottleneck and depthwise convolutions to perform coherent per-frame whale call event detection. We discover that, for this task, the inclusion of spectral phase information among the input features notably improves performance. We also evaluate the effectiveness of negative batch undersampling and the inclusion of a focal loss term. As part of the 2025 BioDCASE challenge (Task 2), we compare our system to ResNet-18 and YOLOVv11 models, as well as to our own baseline. All models are trained exclusively on the same subset of the publically available ATBFL dataset. Our proposed whale call event detector improves on the development set performance of all models, including the top performing YOLOv11, achieving an F1-score of 0.44.**

***Index Terms*—Whale Call Detection, Computational Bioacoustics, Sound Event Detection, Hybrid CNN-BiLSTM**

1. INTRODUCTION

Passive acoustic monitoring (PAM) enables researchers to monitor species in remote locations using non-invasive and relatively low-cost methods. However, large volumes of data are generated, typically with low signal-to-noise ratio (SNR) [1]. This makes manual review, which requires trained experts, both time-consuming and expensive.

To address this, many automated algorithms have been developed to detect and classify signals of interest. One focus has been the detection and classification of blue and fin whale vocalisations. Blue whales were driven nearly to extinction by 20th-century whaling, and as a result are still considered endangered today. Together with fin whales, they are considered vulnerable by the IUCN red list [2, 3]. Population densities for both species remain difficult to estimate with confidence due to limited data availability [4].

In this paper, we focus on developing a lightweight whale call activity detector based on a convolutional bidirectional long short-term memory neural network (CNN-BiLSTM). Our work is inspired by a voice activity detection (VAD) framework originally proposed for speech [5]. We develop an optimised architecture and explore input features that all contribute to substantial gains over our own baseline. Among our key findings are that incorporating phase information from the short time Fourier transform (STFT) strongly enhances model performance, as does the inclusion of a residual bottleneck and depthwise convolutions in the classifier architecture.

2. BACKGROUND

Sound event detection (SED) refers to the task of recognising a sound or sequence of sounds in a long, continuous audio signal that may be polluted with noise from interfering sources. In bioacoustic tasks, these are typically the vocalisations of different species, or inter-species call types. A sequence of sound events can also form an amalgamated call or song, depending on the taxa of interest.

Automatic SED algorithms are typically supervised, and thus require a labelled dataset describing the times at which the events of interest occur in the audio signal. In bioacoustic call detection, these annotations either identify the start and/or end of a call or sub-call, or simply indicate the presence of a call within a longer audio signal without specifying its exact location.

3. LITERATURE REVIEW

The use of bioacoustic data for detecting and classifying animal vocalisations is well established, with early research focusing on birds [6], bats [7] and insects [8], before attention turned to marine animals [9]. Early approaches were based on the identification of regions of high spectral energy within specific frequency bands [10, 11] or the application of template matching techniques using prototypical examples of species-specific calls or sub-calls [9, 12]. However, these approaches tend to perform poorly in low SNR conditions and often require call signatures that are invariant across varying exogenous factors such as season or geographic region.

Dugan et al. [13] incorporated an artificial neural network (ANN) into a suite of parallel detectors for classification of North Atlantic right whale *upcalls*, while Pourhomayoun et al. [14] used image processing techniques to identify regions of interest in the spectrograms before applying the ANN.

More recently, deep neural networks have emerged as a powerful approach in bioacoustics. Many of these employ convolutional neural networks (CNNs), which take spectrograms of audio signals as input and output labels indicating the presence of specific calls or species. While conventional CNNs are widely used, convolutional recurrent neural networks (CRNNs) which incorporate recurrent layers after the convolutional layers to improve sequential modelling have also become popular [15]. CNN-based models have shown strong performance across different taxa, including birds [16] and marine mammals [17].

4. DATA

As part of the 2025 BioDCASE (Task 2) challenge, a dataset consisting of strongly labelled blue- and fin-whale calls in the Antarctic region of the Southern Ocean was released. The data was originally obtained by the Antarctic Blue and Fin Whale Acoustic Trends Project (ATP) as part of the International Whaling Commission's Southern Ocean Research Partnership (IWC-SORP).

The Acoustic Trends Blue Fin Library (ATBFL) consists of 11 site-year datasets recorded around the Antarctic, in the period 2005 to 2017. Each dataset was manually annotated, in both the time and frequency domain, using data collection and annotation procedures described in [4]. The challenge excludes three site-year datasets as a develeopment set, while the remaining eight site-year datasets form the training set, as set out in Table 1.

The ATBFL library includes annotations for seven different call types, of which four are produced by blue whales (*BmA*, *BmB*, *BmZ* and *BmD*) and three by fin whales (*BpD*, *Bp20* and *Bp20plus*). During the final evaluation, these seven calls are collapsed to a three class problem, as set out by the organisers: `bmabz`, `d`, and `bp`.

Table 1: Summary of the ATBFL site-year training and development sets, indicating total recording duration for each set (hours) and how much of the set has been annotated to contain whale calls (hours). The test set annotations are not publically available.

Dataset	Recording (h)	Annotated Whale Call (h)
`ballenyisland2015`	204	2.8
`casey2014`	194	14.2
`elephantislands2013`	187	16.1
`elephantislands2014`	216	28.1
`greenwich2015`	32	2.1
`kerguelen2005`	200	3.5
`maudrise2014`	83	5.7
`rosssea2014`	176	0.1
Total training set	**1292**	**72.6**
`casey2017`	185	6.1
`kerguelen2014`	200	11.4
`kerguelen2015`	200	7.4
Total development set	**585**	**24.9**
`kerguelen2020`	198	—
`ddu2021`	206	—
Total test set	**404**	**—**

5. EXPERIMENTAL STRUCTURE

Figure 1 illustrates the proposed whale call activity detection system. First, the audio is segmented and preprocessed. Each segment is provided to the model as input, with frame-level classification targets determined from the boundary annotation file. Each model consists of two parts: a *feature extractor*, that produces high-dimensional vectors representative of the information contained in the audio segment; and a *classification model*, tasked with producing a class membership probability from feature vectors, obtained at each time instant.

5.1. Preprocessing

The training data consists of long continuous audio recordings, typically the result of PAM. Due to computational constraints, these were subdivided into shorter intervals, referred to as segments.

Each segment corresponds to the audio between the start and end points of a human annotation, indicating the occurrence of a particular call type. The segments were extended to include additional audio before the start and at the end of the call, referred to as a collar. The length of the collar was independently and randomly sampled from a uniform distribution for both the start and end of each segment, ensuring the call does not always appear in the centre of the segment.

An associated discrete classification target vector was constructed from the annotations for each segment. In all experiments, one model classification was computed every 20 ms. As there may be overlapping annotations, the problem is treated as multi-class multi-label. Consequently, a binary label was assigned for each respective class at each discrete time instant, independent of the other classes. When a human annotation boundary intersects completely with the classification target vector at a time instant, the label is *true*, indicating the presence of a particular call; otherwise, it remains *false*.

Additionally, segments without any vocalisation annotations were included (Section 5.5). In such cases, all classification targets were set to *false*, indicating that no vocalisation has occurred.

The variable-length segments were gathered into a batch of fixed length during preprocessing. To achieve this, each segment and the associated classification target was zero padded to the length of the longest segment in the batch. This padding was removed from each segment during loss calculation and model weight backpropagation.

During evaluation, human annotations are not available. Therefore, the continuous audio recordings were subdivided into regularly spaced segments with a fixed length of 30 s and a 2 s overlap. The postulated model classification probabilities were averaged over this overlap.

5.2. Feature extractor

The feature extraction model is provided with an audio segment as obtained from the continuous audio recording (Section 5.1). In this work, we only consider spectral and cepstral features.

First, a spectrogram representation was computed using a $\sim$1 s frame length and 20 ms stride between frames. The long frame length was motivated by the low fundamental frequency of the whale calls. The frame stride was dictated by the desired classification resolution, which was fixed at 20 ms, to allow for direct comparison of loss figures between experiments (Section 5.1). A Hanning window was applied to each frame, without additional zero padding. A subsequent 256-point fast Fourier transform (FFT) resulted in 128 frequency bins and the DC component. This power spectrum was compressed into 64 bins using a bank of triangular filters with a mel-scale spacing. Finally, mel frequency cepstral coefficients (MFCCs) were obtained by applying the discrete cosine transform (DCT) to the resulting binned spectrum, and retaining the lower 20 coefficients.

After the features were obtained, mean spectral and cepstral subtraction was performed, respectively. The mean was computed independently for each frequency bin over the duration of the segment.

5.3. Baseline classification model

From the sequence of features obtained from the feature extractor, the call posterior probabilities were computed using a classification model with sigmoid activation functions at the outputs.

A bidirectional long short-term memory network (BiLSTM) model was chosen and configured with between one and four hidden layers; a hidden dimension size of 64, 128, or 256; and layer dropout of between 20 % and 50 %. We found that the recurrent model was prone to overfit, but that increased dropout reduced this risk. The posterior call probabilities produced by these models aligned well with the call segments. Informally, it was observed that both the start and end boundaries produced by the recurrent models closely matched those of the human annotators.

5.4. Whale vocalisation activity detector

We propose a whale call activity detection system inspired by the AVA-VAD system first presented in [5]. We alter the AVA-VAD system by introducing a residual bottleneck network and depthwise convolutions. Furthermore, instead of using a mel spectrogram as input, we utilise the spectrogram features directly and apply a linear convolutional layer to act as a learnable filterbank. Figure 2 provides an overview of the model architecture.

The spectrogram is computed using the configuration described in Section 5.2. However, during experimentation, we found that including phase information substantially improves detection performance. Thus, instead of the power spectrum typically utilised as feature, we provide the model with a three-dimensional representation of each complex spectral component (z) as follows:

$$z = r(\cos\theta + i\sin\theta); \qquad \boldsymbol{x}_k^{(n)} = \begin{bmatrix} r \\ \cos\theta \\ \sin\theta \end{bmatrix}$$

where r is the spectral magnitude and θ is the phase and $\boldsymbol{x}_k^{(n)}$ is the model input at time instant n and discrete frequency k.

The learnable filterbank consists of a linear convolutional layer. The one-dimensional kernel is convolved with each vector of energies

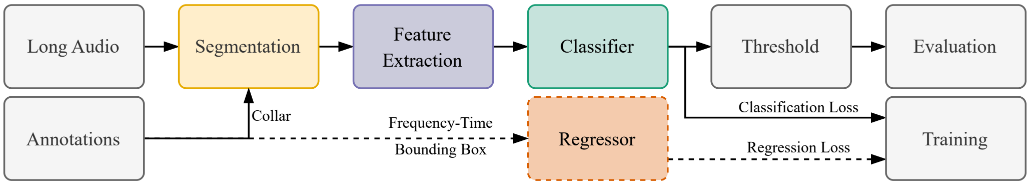

Fig. 1: Illustration depicting the whale call activity detection system overview and experimental setup.

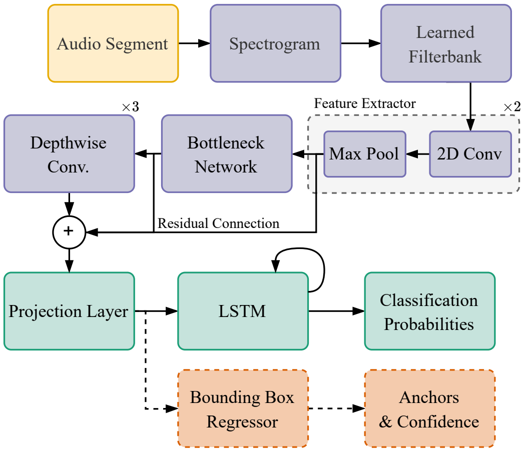

Fig. 2: Illustration of the Whale-VAD system proposed for whale call activity detection.

constituting the spectrogram. This output is then passed through a two layer CNN with max pooling, GELU activation and batch normalisation. During model development several architectural variations were considered. We found the addition of a residually connected bottleneck network and depthwise convolutional network to improve model performance, inspired by [18, 19, 20]. Each bottleneck network consists of three convolutional layers, each with GELU activation, compressing the features to a lower dimensional representation. This representation is passed through three depthwise convolutional layers with intermediate drop out. Table 2 provides a summary of the CNN layer configuration employed by the final Whale-VAD model.

The residual output connection is reduced to 64 dimensions by a linear layer. Padding is applied to each CNN layer to ensure the number of output activations remains consistent with the number of input frames. These latent features are then recurrently processed by a BiLSTM network. Finally, a linear layer with sigmoid activations produces the model call probabilities.

We further investigate two input regularisation techniques: spectral augmentation [21] and noise perturbation. For noise perturbation, we inject Gaussian noise into the audio signal such that the resulting SNR of the original signal to the perturbed signal is 10 dB.

5.5. Stochastic negative mini-batch undersampling

Analysis of the challenge dataset revealed that whale vocalisations are rare, with a prevalence of approximately 5 %. Therefore, we propose a technique where, during each epoch of finetuning, we sample a different subset of *negative* segments (containing no calls) while ensuring that there are approximately as many negative as positive segments per mini-batch. This ratio was determined though early informal experimentation After each training epoch, a new subset of negative segments is sampled. The set of positive calls remains consistent for each epoch during finetuning.

Table 2: Summary of Whale-VAD model layer configuration. The kernel size (K), stride (S), number of input channels (C_{in}) and output channels (C_{out}), are shown.

Layer	K	S	C_{in}	C_{out}
Filterbank	(7, 1)	(3, 1)	1	64
Feature extractor				
└ Conv2D	(5, 5)	(3, 1)	64	128
└ Max pool	(5, 1)	(1, 1)	–	–
└ Conv2D	(3, 3)	(2, 1)	128	128
└ Max pool	(3, 1)	(1, 1)	–	–
Bottleneck network				
└ Conv2d	(1, 1)	(1, 1)	128	64
└ Conv2d	(3, 3)	(1, 1)	64	64
└ Conv2d	(1, 1)	(1, 1)	64	128
Depth. Conv2d	(3, 3)	(1, 1)	128	128

5.6. Loss function

In our experiments, we considered weighted binary cross-entropy (BCE) and *focal loss* as loss functions. We found that, when computing the class weighting, rather than normalising by the duration of each class, it was better to normalise by the number of segments belonging to each class. When considering weighted BCE, we compute the weighting w_c for each class c as follows:

$$w_c = \frac{N}{P_c}$$

where N denotes the total number of negative (no-call) segments and P_c the number of positive (call) segments for class c.

In addition to weighted BCE, we evaluated the use of *focal loss* [22], a modified cross-entropy designed to focus training on hard-to-classify examples by reducing the contribution of easy examples. In our experiments, we set the class imbalance term to 0.25 and *focus* term to 2, following the recommendations in the original paper.

For all experiments, we rely on AdamW [23] as the numerical optimiser. Unless otherwise stated, the optimiser was configured with an initial learning rate of 1×10^{-5}, momentum terms of 0.9 and 0.999, and a weight decay factor of 0.001.

5.7. Multi-objective regression

The challenge dataset contained not only annotations in time, but also in frequency (bounding box). The best-performing baseline YOLO model, provided by the organisers, uses these box-level annotations. In addition to our Whale-VAD system (Section 5.4), we therefore evaluated a bounding box regression network. The network uses the same latent features as the classification model, with the addition of an adaptive pooling layer in order to reduce the time dimension. These reduced latent features are presented to two independent multi-layer perceptron (MLP) networks, each consisting of three layers, with GELU activation and dropout after each hidden layer. Each of the regression networks is applied to the 64 channels of the adaptive pooling layer, which is the maximum number of anchors (bounding boxes) the model can produce per input segment. The first network has a four dimensional output, corresponding to the bottom left and top right corners of the bounding box. The second network produces a confidence score, corresponding to the presence of the bounding box. Figure 2 illustrates the additional bounding box regression model. The regression model is trained using smoothed L1 loss [24]. Note that the regression model forms part of a multi-objective training regime, where the regression and classification loss are jointly optimised. The regression outputs (bounding boxes) are not used for final model evaluation. We postulated that training to jointly optimise both tasks might lead to improved classification performance.

5.8. Postprocessing

After training, the best model was chosen based on the lowest BCE development loss. The model outputs were smoothed using a median filter with a 500 ms kernel. The classification thresholds θ_c were selected per class c to maximise the development F1-score. The resulting per call threshold θ_c was applied to the posterior call probabilities computed by the model to obtain the binary classification result. Finally, the 7-class classifier output is collapsed int the 3-class variant posed by the challenge organisers.

The resulting binary labels were used to generate start and end boundaries relative to the start of the recording. These annotations were refined by merging overlapping calls of the same type, eliminating duplicates, and joining calls separated by less than 500 ms. Finally, calls longer than 30 s or shorter than 500 ms were discarded.

6. RESULTS

Table 3 presents the top performing models proposed in this work and the baselines provided by the organisers of the BioDCASE challenge. All reported figures have been measured on the development set.

We performed a series of model development experiments that lead to notable changes to the original CNN-BiLSTM architecture (AVA-VAD). The addition of residual bottleneck layers and depthwise convolutions, in particular, lead to substantial improvements in model recall, at the loss of some precision. However, incorporating phase information improved both recall and precision, and resulted in a 30 % improvement in the F1-score. The introduction of focal loss further improved the recall, precision and F1-score. Finally, training the model on collapsed labels (three-class problem) yielded an additional 15.2 % improvement, resulting in an F1-score of 0.440. The addition of multi-task bounding box regression, as well as the inclusion of input noise perturbation and spectral augmentations, were found to be counterproductive. When considering the top performing baseline model (YOLOv11), we see that our models exhibit superior recall, while the baseline achieves greater precision.

Table 4 and Fig. 3 presents classification performance of our best performing model for each call and each development set. We observe consistent performance across both kerguelen sites, but poor performance for casey2017. We also observe that our models produce a high number of false positive (FP) classifications for the minority call types (d and bp), which leads to poor precision. Manual inspection revealed these FPs usually to be other marine vocalisations.

7. CONCLUSION

We have presented a CNN-BiLSTM architecture and shown it to be viable for whale call event detection. Incorporating residual bottleneck and depthwise convolutional layers lead to substantial improvements in recall. While most prior research disregards the phase information of the STFT, we have demonstrated that incorporating it for whale call classification can improve model precision substantially, compared to models trained solely on magnitude features. This finding opens avenues for future research in bioacoustics to reconsider the role of phase in time-frequency representations, particularly in the design of feature extractors and model architectures that can more effectively exploit both magnitude and phase components.

Table 3: Development set results for the official ResNet18 and YOLOv11 baselines, and our MFCC, AVA-VAD and Whale-VAD models. Scores are averages across call types and validation and test sets. Challenge results shown, where available.

Experiment	Development			Test		
	Recall	Precision	F1-score	Recall	Precision	F1-score
ResNet18 (Baseline)	0.36	0.29	0.32	—	—	—
YOLOv11 (Baseline)	0.32	**0.67**	0.43	0.331	**0.480**	**0.392**
MFCCs + BiLSTM	0.409	0.226	0.291	—	—	—
AVA-VAD [5]	0.310	0.219	0.245	—	—	—
Whale-VAD	0.424	0.207	0.278	—	—	—
+ Phase	0.461	0.316	0.375	—	—	—
└ + Focal loss	**0.484**	0.348	0.405	**0.414**	0.302	0.349
└ + Three class	0.461	0.420	**0.440**	0.382	0.336	0.357

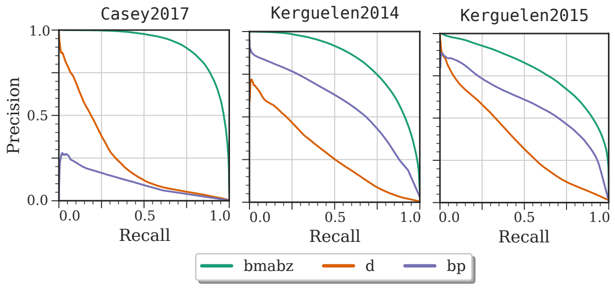

Fig. 3: Separate precision-recall curves for the three-call problem on each development set using the top-performing Whale-VAD model.

Table 4: Detailed development set results for the top-performing Whale-VAD model.

Dataset	Label	TP	FP	FN	Recall	Precision	F1
casey2017	bmabz	1984	1956	434	0.821	0.504	0.624
casey2017	d	179	5928	374	0.324	0.029	0.054
casey2017	bp	5	101	287	0.017	0.047	0.025
kerguelen2014	bmabz	2739	1120	1558	0.637	0.710	0.672
kerguelen2014	d	229	2248	550	0.294	0.092	0.141
kerguelen2014	bp	1391	663	2355	0.371	0.677	0.480
kerguelen2015	bmabz	2137	2676	611	0.778	0.444	0.565
kerguelen2015	d	366	2545	1158	0.240	0.126	0.165
kerguelen2015	bp	665	355	605	0.524	0.652	0.581

While the models we present exhibit performance improvements for whale call event detection, much room for advancement remains. The achieved recall and precision of 48.4 % and 42 %, respectively, means that most human annotated calls are not identified by the model, and among the calls identified, most are false positives.

8. ACKNOWLEDGMENT

The authors gratefully acknowledges Telkom (South Africa) for their financial support, and the Stellenbosch Rhasatsha high performance computing (HPC1) facility for the compute time provided to the research presented in this paper.

REFERENCES

[1] K. A. Kowarski and H. Moors-Murphy, "A review of big data analysis methods for baleen whale passive acoustic monitoring," *Marine Mammal Science*, vol. 37, no. 2, pp. 652–673, 2021.

[2] J. G. Cooke, "Balaenoptera musculus," *The IUCN Red List of Threatened Species*, 2018, erratum published in 2019.

[3] ——, "Balaenoptera physalus," *The IUCN Red List of Threatened Species*, 2018.

[4] B. S. Miller, The IWC-SORP/SOOS Acoustic Trends Working Group, K. M. Stafford, I. Van Opzeeland, D. Harris, F. Samaran, A. Širović, S. Buchan, K. Findlay, N. Balcazar, S. Nieukirk, E. C. Leroy, M. Aulich, F. W. Shabangu, R. P. Dziak, W. S. Lee, and J. K. Hong, "An open access dataset for developing automated detectors of Antarctic baleen whale sounds and performance evaluation of two commonly used detectors," *Scientific Reports*, vol. 11, no. 1, p. 806, 2021.

[5] N. Wilkinson and T. Niesler, "A Hybrid CNN-BiLSTM Voice Activity Detector," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, Canada: IEEE, 2021, pp. 6803–6807.

[6] S. Anderson, A. S. Dave, and D. Margoliash, "Template-based automatic recognition of birdsong syllables from continuous recordings." *The Journal of the Acoustical Society of America*, vol. 100 2 Pt 1, pp. 1209–1219, 1996.

[7] N. Vaughan, G. Jones, and S. Harris, "Identification of British bat species by multivariate analysis of echolocation parameters," *Bioacoustics*, vol. 7, no. 3, pp. 189–207, 1997.

[8] R. H. Campbell, S. K. Martin, I. Schneider, and W. R. Michalson, "Analysis of mosquito wing beat sound," *Journal of the Acoustical Society of America*, vol. 100, pp. 2710–2710, 1996.

[9] D. K. Mellinger and C. Clark, "Recognizing transient low-frequency whale sounds by spectrogram correlation." *The Journal of the Acoustical Society of America*, vol. 107 6, pp. 3518–29, 2000.

[10] J. A. Ward, M. Fitzpatrick, N. A. Dimarzio, D. J. Moretti, and R. P. Morrissey, "New algorithms for open ocean marine mammal monitoring," *OCEANS 2000 MTS/IEEE Conference and Exhibition. Conference Proceedings (Cat. No.00CH37158)*, vol. 3, pp. 1749–1752 vol.3, 2000.

[11] D. Gillespie and O. Chappell, "An automatic system for detecting and classifying the vocalisations of harbour porpoises," *Bioacoustics*, vol. 13, pp. 37 – 61, 2002.

[12] D. K. Mellinger and C. W. Clark, "Methods for automatic detection of mysticete sounds," *Marine and Freshwater Behaviour and Physiology*, vol. 29, pp. 163–181, 1997.

[13] P. J. Dugan, A. N. Rice, I. R. Urazghildiiev, and C. W. Clark, "North Atlantic right whale acoustic signal processing: Part II. improved decision architecture for auto-detection using multi-classifier combination methodology," *2010 IEEE Long Island Systems, Applications and Technology Conference*, pp. 1–6, 2010.

[14] M. Pourhomayoun, P. J. Dugan, M. Popescu, and C. W. Clark, "Bioacoustic Signal Classification Based on Continuous Region Processing, Grid Masking and Artificial Neural Network," 2013, arXiv:1305.3635.

[15] D. Stowell, "Computational bioacoustics with deep learning: a review and roadmap," *PeerJ*, vol. 10, p. e13152, 2022.

[16] S. Kahl, C. M. Wood, M. Eibl, and H. Klinck, "BirdNET: A deep learning solution for avian diversity monitoring," *Ecol. Informatics*, vol. 61, p. 101236, 2021.

[17] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. K. Maier, "ORCA-SPOT: An Automatic Killer Whale Sound Detection Toolkit Using Deep Learning," *Scientific Reports*, vol. 9, 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[19] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," 2017, arXiv:1605.07146.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, 2019, pp. 2613–2617.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," 2018, arXiv:1708.02002.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[24] R. Girshick, "Fast R-CNN," 2015, arXiv:1504.08083.