Listening to Nature: Automated Bird Species Identification for Biodiversity Monitoring

1st Vencel Bódi
Faculty of Electronics Engineering and Informatics
Budapest University of Technology and Economics
Budapest, Hungary
bodi.vencel@edu.bme.hu

2nd Márk Mitrenga
Systems and Control Laboratory
HUN-REN Institute for Computer Science and Control (SZTAKI)
Budapest, Hungary
mitrenga.mark@sztaki.hun-ren.hu

3rd Bálint Kővári
Department of Control for Transportation and Vehicle Systems
Budapest University of Technology and Economics
Budapest, Hungary
kovari.balint@kjk.bme.hu

4th Szilárd Aradi
Department of Control for Transportation and Vehicle Systems
Budapest University of Technology and Economics
Budapest, Hungary
aradi.szilard@kjk.bme.hu

***Abstract*—Bird populations are important bioindicators, as the diversity of their habitats and their outstanding mobility make them sensitive to ecosystem changes. They also contribute to ecosystems by providing valuable services such as pest control, seed dispersal, and pollination, which are vital for maintaining ecological balance. Hence monitoring the composition of bird populations is essential for assessing ecosystem health and guiding conservation efforts. However, traditional observer-based surveys are expensive, time-consuming, and impractical for large-scale and hardly accessible applications. Our study proposes an approach by integrating automated acoustic monitoring with machine learning applications to develop a system that achieves high identification accuracy while keeping the pipeline simple. Our model design emphasizes a balance between efficiency and scalability, enabling deployment in resource-constrained environments. We provide a tool for low-complexity ecological monitoring by prioritizing lightweight and simple architectures. This work bridges the gap between ecological research and practical, scalable conservation technologies.**

I. Introduction

Bird populations are essential contributors to innovative and sustainable agriculture [1], offering natural pest control solutions [2] and aiding seed dispersal, which lessens the need for chemical interventions and supports biodiversity. Their ecological roles extend beyond agricultural systems, influencing forest regeneration, pollination, and nutrient cycling. The presence of different bird species may come with various ecological contribution - be it advantageous or not [3]. Monitoring and controlling the composition of bird populations is therefore essential [4] to ensure that their ecological roles are maintained, while mitigating potential negative impacts. By adopting effective monitoring techniques, we can better understand the dynamics of bird populations and their interactions with ecosystems, which is crucial for sustaining biodiversity and ecosystem services.

Understanding these ecological contributions is crucial not only for conservation efforts but also for improving ecosystem services that directly benefit our society. However, as bird populations face increasing threats from habitat loss, climate change, and industrial activities, it is urgent to monitor and protect these species effectively. Traditional monitoring methods, such as field surveys, are resource-intensive and limited in scalability.

A. Motivation

By creating autonomous monitoring systems, we can achieve continuous, large-scale, and cost-effective data collection, surpassing the limitations of traditional field surveys [5]. These systems, often leveraging technologies like passive acoustic monitoring combined with machine learning, enable precise identification of bird species to provide real-time insights into population dynamics [6]. Autonomous monitoring not only reduces human effort but also ensures consistent data quality, making it an invaluable tool for ecological research, conservation planning, and the management of sustainable natural resources. Furthermore, such systems can be deployed in remote or resource-constrained environments, expanding their applicability and impact. Such systems provide effective means to monitor endangered or invasive species [7] [8], allowing for effective management while minimizing human interference with their natural habitats [9] [10].

B. Related Work

Previous studies have explored passive acoustic monitoring for bird species. In most cases, these techniques were used to determine bird species density, generally utilizing Autonomous Recording Units (ARUs) [11].

A more effective solution can be achieved by converting audio files into spectrograms. Different spectrograms can be generated using various techniques. One approach is to create magnitude spectrograms [12]. The spectrogram is a time-frequency representation that visualizes the frequency components of an audio signal over time. The dark and light bands show which frequencies are active at a given moment. In the case of bird calls, different species produce characteristic frequency patterns (e.g. pitch, trills, calls) that can be visually distinguished in spectrograms. However, this alone does not provide sufficient information about the sounds, making it impossible to achieve the desired classification results.

Another technique, closer to our approach, is the conversion of bird calls into Mel spectrograms and their classification using different Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) models [13]. In addition to the Mel spectrogram, the effectiveness of the Mel Frequency Cepstral Coefficient (MFCC) was also utilized for sound representation, followed by bird classification based on audio using various models. As a result of model comparisons, the CNN proved to be the optimal choice for implementing bird classification based on sound.

To achieve a more robust model, selective fusion of CNN-based networks may provide a solution [14]. By utilizing three time-frequency representations (TFRs) — Mel spectrogram, Harmonic-component based spectrogram, and Percussive-component based spectrogram — they capture different acoustic patterns from the same audio file. These representations are then used in combination with a Visual Geometry Group (VGG) based neural network, a SubSpectralNet model, and an additional class-based late fusion method to enhance the accuracy of classification.

However, this solution is only capable of distinguishing between 43 different bird species. If we aim to classify a larger number of species, the structure of the methodology makes further fine-tuning of the model a time-consuming, resource-intensive, and costly task.

C. Contribution

Our proposed machine learning solution integrates passive acoustic observation with advanced classification algorithms to develop an autonomous system that can identify bird species with high accuracy.

In contrast to previous approaches, which relied on complex deep learning models or a fusion of these, our approach emphasizes the use of lightweight and scalable architectures, ensuring its applicability in resource-constrained environments, thus reducing both the cost of ownership and the cost of maintenance.

By combining innovative augmentative signal processing techniques with machine learning models, we aim to create a robust framework that not only supports ecological monitoring, but the system also can facilitate large-scale, automated surveys of bird populations, aiding in the early detection of invasive species and the tracking of biodiversity trends over time. By monitoring the density and location of bird populations on a large scale, we can plan trajectories of autonomous agricultural vehicles using other deep learning algorithms [15] [16] and other real-time applications [17] that do not intervene with the habitat of avian species.

II. Objective and Key Challenges

Our objective was to train a deep neural network that is capable of classifying birds based on their vocalization recorded through passive acoustic monitoring, while keeping low the model's complexity. This work builds upon the BirdCLEF 2024's [18] dataset provided in the Kaggle competition [19]. BirdCLEF 2024 is part of the LifeCLEF competition series [20]. The dataset consists of 24,459 labelled and 8,444 unlabelled audio samples, complemented by a wide range of metadata. The labelled dataset includes annotations for 182 bird species, geographic locations, and recording quality, making it suitable for training and validating machine learning models. The unlabelled samples present opportunities for semi-supervised or unsupervised learning approaches, enabling the exploration of methods that utilize unannotated data effectively.

The dataset exhibits significant disparities in the number of recordings per class, with some bird species being underrepresented. This class imbalance poses a significant challenge during model training, as the network tends to focus on classes with abundant examples while potentially ignoring underrepresented species. This can lead to biased predictions, low recall values, and inaccurate species detection, which is particularly problematic in biodiversity monitoring. Therefore, balancing techniques such as data augmentation are necessary.

Additionally, recording lengths vary drastically, ranging from a few seconds to several minutes, or even hours, complicating data standardization. Moreover, length differences raise other issues as well. Very short samples may lack characteristic sound patterns, which can reduce the model's performance. Additionally, processing long recordings requires significant computational capacity and time, making efficient resource management a key consideration in model design. These challenges must be addressed through robust preprocessing steps and carefully crafted training strategies.

The combination of class imbalance and varying audio durations presents unique obstacles that must be addressed both in preprocessing and model training. Our solution addresses these challenges by implementing a robust preprocessing pipeline and a tailored deep learning approach.

III. Methodology

A. Data Segmentation

A key aspect of our approach is data segmentation. Since many recordings exceed the desired input size for our neural network, we divide them into fixed-length segments. Unlike

conventional methods that discard excess audio, our approach retains all segments to maximize dataset utilization. Shorter recordings are padded with zeros or mirror the edge values. It was found that zero padding gave greater accuracy, so this method was used to obtain the final result. Table I presents the percentile distribution of audio lengths, Fig. 1 illustrates their overall variation. The presence of extreme outliers further underscores the need for effective preprocessing.

B. Short Time Fourier Transform

There are multiple conventional approaches regarding audio processing [21] [22]. One of the widely spread approaches is based on the spectral decomposition of audio waves using the short time Fourier transform (STFT).

The STFT provides a time-frequency representation of an audio signal by dividing it into overlapping segments and applying the Fourier Transform to each segment [23]. Mathematically, the STFT of a signal $x(t)$ is defined as in eq. 1:

$$\text{STFT}(x(t))(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j\omega\tau}d\tau \tag{1}$$

where $w(\tau - t)$ is a sliding window function that localizes the signal in time. We used the Hann window function, which is decomposition allows us to convert raw audio into a spectrogram, which is a 2D representation of frequency over time. [24]

TABLE I
PERCENTILE DISTRIBUTION OF AUDIO LENGTHS.

Percentile (%)	Audio Length (s)
10	6.72
25	11.21
50	22.39
75	44.67
90	82.25
95	122.47
99	300.13
100	5964.23

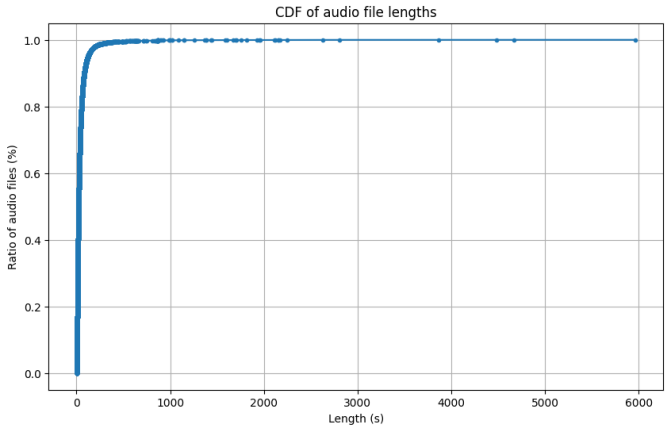

Fig. 1. Cumulative distribution of audio lengths

C. Mel Spectrogram Transformation

While spectrograms provide a detailed frequency representation, they are not directly aligned with human auditory perception. To address this, we transformed the spectrograms into Mel spectrograms using the mel scale, which is a perceptually motivated scale of pitches [25]. This highly contributes to achieving better model performance [26].

The mel frequency f_{mel} corresponding to a linear frequency f (in Hz) is given by eq. 2:

$$f_{\text{mel}} = 2595 \cdot \log_{10}(1 + \frac{f}{700}). \tag{2}$$

Visualization of the mel frequencies can be seen of Fig. 2. Using this scale, the frequency bins of the spectrogram are mapped to the mel scale. The resulting Mel spectrogram $M(t,m)$ is computed as seen in eq. 3:

$$M(t,m) = \sum_{k} |S(t,f_k)|^2 \cdot H_m(f_k) \tag{3}$$

where $S(t,f_k)$ is the STFT of the signal at frequency f_k, and $H_m(f_k)$ is the triangular filter corresponding to the m-th mel band. The power $|S(t,f_k)|^2$ represents the energy at frequency f_k and time t. Such a spectrogram made from an audio signal can be seen on Fig. 3.

D. Amplitude Scaling

The raw Mel spectrogram values are further converted to a logarithmic amplitude scale, commonly referred to as the decibel (dB) scale. This step enhances the representation of lower energy signals, making it more suitable for neural network processing. The conversion to the dB scale is given by the eq. 4:

$$M_{\text{dB}}(t,m) = 10 \cdot \log_{10}(\max(M(t,m), \epsilon)) \tag{4}$$

where ϵ is a small constant to avoid taking the logarithm of zero.

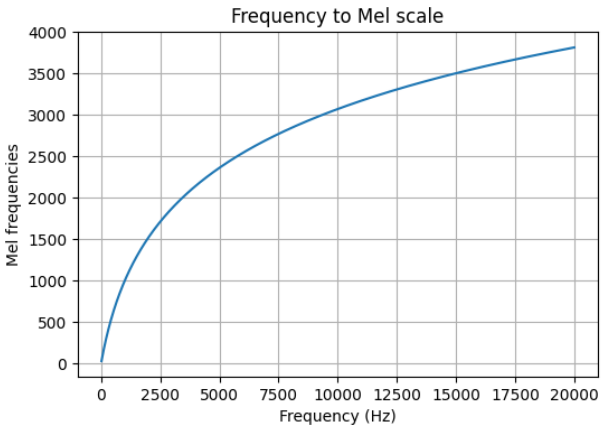

Fig. 2. The mel frequency f_{mel} corresponding to a linear frequency f (in Hz).

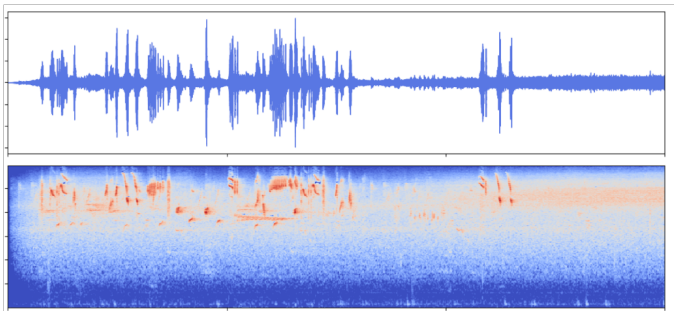

Fig. 3. Waveform (top) and corresponding Mel spectrogram (bottom) of a bird vocalization. [27] The spectrogram highlights frequency components over time, revealing distinct patterns characteristic of the species' call.

E. Data Augmentation

To further enhance model performance, we applied data augmentation techniques to increase the diversity of the training data and help the model generalize better to unseen samples [28]. Examples for these augmentations can be seen on Fig. 4.

F. Spectrogram Classification Using CNNs

Once the spectrograms were generated, we treated them as image-like inputs for classification. A Convolutional Neural Network (CNN) was employed to extract spatial and temporal features from the spectrograms. CNNs are particularly well-suited for this task as they can identify intricate patterns in the time-frequency domain, such as unique frequency modulations or harmonic structures characteristic of bird species.

Our model architecture was based on pre-trained EfficientNet [29] variants, which were fine-tuned for the task of birdsong classification. Using weights of pre-trained models can be highly beneficial in audio recognitions and can highly improve the convergence of the models [30]. These models provide a balance between computational efficiency and high accuracy, making them ideal for handling the challenges posed by the dataset, such as class imbalance and noisy samples.

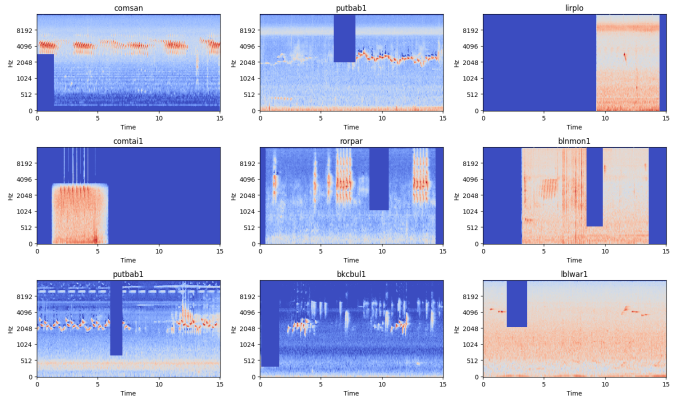

Fig. 4. Examples of data augmentation applied to bird vocalization spectrograms [27]. The transformations include frequency masking, time masking, and additive Gaussian noise.

IV. Experimental Setup

While conducting our research, we tried numerous convolutional neural networks. We have found that the EfficientNetV2-M architecture strikes the best balance between classification performance and computational efficiency, and by fine-tuning it, we have obtained a model that is suitable for implementing classification [31].

Ensuring the robustness of the model heavily depends on the appropriate selection of hyperparameters and parameters. Therefore, we implemented and tested multiple approaches to achieve the best possible results. Since the model and its hyperparameters were predefined, we did not modify them; however, we conducted several experiments with the optimizer, learning rate schedule, and loss function to identify the optimal settings.

The best results were achieved using the Adam optimizer, which provided fast convergence and a stable learning process. One of the main advantages of Adam is its adaptive learning rate adjustment for each parameter, making weight fine-tuning more efficient and reducing the likelihood of getting stuck in local minima.

To ensure successful training, the learning rate was dynamically adjusted using a Warmup Cosine Annealing schedule, which can be seen on Fig. 5. In the initial phase, gradually increasing the learning rate helped stabilize the weight updates, while the cosine-based decay ensured smooth convergence toward the optimal solution. This scheduling not only facilitated faster convergence but also reduced the risk of overfitting, as weight updates became smaller in the final phase of training.

The loss function plays a fundamental role in optimizing the network's weights, for this problem, we applied the focal loss function, which is a dynamically scaled cross-entropy loss designed to effectively handle class imbalances.

Hyperparameters used for the final results can be seen on Table II.

V. Results and Discussion

In terms of results, we have found a good practice to solve the problem. The method we propose achieves better results

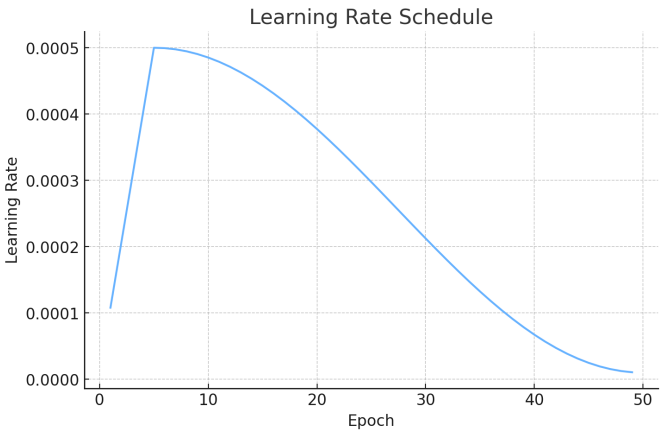

Fig. 5. Warmup Cosine Annealing learning rate schedule.

TABLE II
HYPERPARAMETERS USED IN THE TRAINING OF THE FINAL MODEL AND DATA PROCESSING.

Training Parameters	
Batch Size	8
Epochs	50
Warmup Epochs	5
Starting Learning Rate	1e-5
Warmup Learning Rate	5e-4
Final Learning Rate	2e-5
Mel Spectrogram Parameters	
Audio Length (s)	15
Sample Rate	16000
Number of Mel Frequencies	224
Window Length	512
Maximum Frequency (Hz)	16000
Minimum Frequency (Hz)	20

in terms of ROC-AUC than the best solutions tested on the available database published on BirdCLEF.

The proposed solution can be considered lightweight and highly scalable, as although the EfficientNetV2-M architecture is itself a compact and efficient model, the novelty of our work lies primarily in the applied preprocessing and data processing methods. These innovative steps — including the segmentation strategy, handling of noisy and short recordings, and targeted data augmentation — significantly contributed to improving the model's performance without the need to apply more complex or computationally demanding networks. Thus, our approach stands out precisely because we achieved computational efficiency not by complicating the model, but by optimizing data processing.

We trained our model with both the conventional way, using only a 15 second long segment of the audio files and by the segmentation method we proposed. The results of the two methods we applied are shown in Fig. 6. The gray curve represents the validation accuracy during training with the conventional approach, while the red curve corresponds to our approach. It is clearly visible that the method proposed by us approach achieves better results in terms of the evaluated accuracy.

The final ROC-AUC scores of our proposed methods and the best publicly available solution from the BirdCLEF 2024 competition are summarized in Table III. It is evident that the method we proposed outperforms the best publicly available solution. This improvement is largely due to our enhanced data processing approach, which effectively leverages the available dataset.

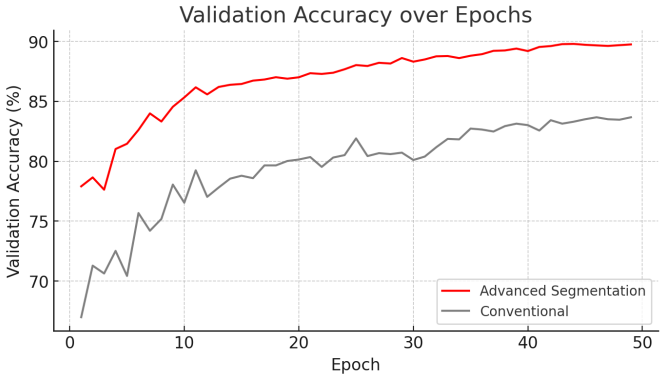

Fig. 6. Validation accuracy

VI. CONCLUSION

In this study, we investigated a method that improves the performance of classification for audio files through data processing and augmentation. We used a single convolutional neural network to do classification between 182 different bird species based on their songs.

By dividing the data into uniform segments and supplementing segments shorter than the desired length, it is possible to enhance the model's accuracy on the same dataset even for classes with fewer samples. This type of preprocessing improves the model's performance, as it is trained with a larger number of samples than in the original dataset. Consequently, with a larger dataset, the model is more likely to learn the unique characteristics of the different classes, allowing it to solve the classification task with excellent results.

With the method proposed by us the complete monitoring of endemic and epidemic bird species can be achieved with a lower cost associated with the machine learning solution. By improving the data usage efficiency and using advanced augmentation techniques even a single, much smaller neural network can achieve outstanding results, thus the need for highly complex models and ensemble techniques is no longer needed.

Furthermore, the methodology outlined in this study holds promise for several future applications. First, it can be adapted to other domains involving time-series data, such as speech recognition, music genre classification. The proposed segmentation techniques can also benefit tasks where data imbalance is a critical issue, such as rare event detection.

VII. ACKNOWLEDGMENTS

We would like to thank the organizers of the BirdCLEF 2024 competition for providing the dataset, which served as the foundation for this work.
This research was supported by the European Union within the framework of the National Laboratory for Autonomous Systems. (RRF-2.3.1-21-2022-00002).
Bálint Kővári is supported by project no. 2024-2.1.1-EKÖP-2024-00003 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the EKÖP-24-4-I-BME-150 funding scheme.

TABLE III
BEST RESULTS ACHIEVED WITH THE METHODS.

	Public Best	Conventional	Segmentation
Final metrics (AUC)	98.77	98.6303	99.4243

REFERENCES

[1] J. A. Jedlicka, R. Greenberg, and D. K. Letourneau, "Avian conservation practices strengthen ecosystem services in california vineyards," *PLOS ONE*, vol. 6, no. 11, pp. 1–8, 11 2011. [Online]. Available: https://doi.org/10.1371/journal.pone.0027347

[2] M. Johnson, J. Kellermann, and A. Stercho, "Pest reduction services by birds in shade and sun coffee in jamaica," *Animal conservation*, vol. 13, no. 2, pp. 140–147, 2010.

[3] C. J. Whelan, H. Şekercioğlu, and D. G. Wenny, "Why birds matter: from economic ornithology to ecosystem services," *Journal of Ornithology*, vol. 156, no. 1, pp. 227–238, 2015. [Online]. Available: https://doi.org/10.1007/s10336-015-1229-y

[4] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, no. S1, pp. S163–S173, 2008.

[5] B. J. Furnas and R. L. Callas, "Using automated recorders and occupancy models to monitor common forest birds across a large geographic region," *The Journal of Wildlife Management*, vol. 79, no. 2, pp. 325–337, 2015.

[6] M. Depraetere, S. Pavoine, F. Jiguet, A. Gasc, S. Duvail, and J. Sueur, "Monitoring animal diversity using acoustic indices: Implementation in a temperate woodland," *Ecological Indicators*, vol. 13, no. 1, pp. 46–54, 2012.

[7] M. C. Cerqueira and M. Aide, "Improving distribution data of threatened species by combining acoustic monitoring and occupancy modeling," *Methods in Ecology and Evolution*, vol. 7, no. 11, pp. 1340–1348, 2016.

[8] A. Gasc, J. Anso, J. Sueur, H. Jourdan, and L. Desutter-Grandcolas, "Cricket calling communities as an indicator of the invasive ant wasmannia auropunctata in an insular biodiversity hotspot," *Biological Invasions*, vol. 20, pp. 1099–1111, 2018.

[9] R. Gibb, E. Browning, P. Glover-Kapfer, and K. E. Jones, "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring," *Methods in Ecology and Evolution*, vol. 10, no. 2, pp. 169–185, 2019.

[10] D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, J. L. Deppe, A. H. Krakauer, C. Clark, K. A. Cortopassi *et al.*, "Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospectus," *Journal of Applied Ecology*, vol. 48, no. 3, pp. 758–767, 2011.

[11] C. Pérez-Granados and J. Traba, "Estimating bird density using passive acoustic monitoring: a review of methods and suggestions for further research," *Ibis*, vol. 163, no. 3, pp. 765–783, 2021.

[12] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks." *CLEF (working notes)*, vol. 1866, 2017.

[13] S. Carvalho and E. F. Gomes, "Automatic classification of bird sounds: using mfcc and mel spectrogram features with deep learning," *Vietnam Journal of Computer Science*, vol. 10, no. 01, pp. 39–54, 2023.

[14] J. Xie, K. Hu, M. Zhu, J. Yu, and Q. Zhu, "Investigation of different cnn-based models for improved bird sound classification," *IEEE Access*, vol. 7, pp. 175 353–175 361, 2019.

[15] K. Balint, A. B. Gergo, and B. Tamas, "Deep reinforcement learning combined with rrt for trajectory tracking of autonomous vehicles." *Transportation Research Procedia*, vol. 78, pp. 246–253, 2024, 25th Euro Working Group on Transportation Meeting. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352146524000851

[16] B. Kővári, F. Hegedüs, and T. Bécsi, "Design of a reinforcement learning-based lane keeping planning agent for automated vehicles," *Applied Sciences*, vol. 10, no. 20, p. 7171, 2020.

[17] H. Sándor, K. Bálint, K. Máté, G. Tamás, V. Dániel, R. József, and B. György, "Area scanning with reinforcement learning and mcts in smart city applications," *Repüléstudományi Közlemények*, vol. 32, no. 2, pp. 137–153, 2020.

[18] S. Kahl, T. Denton, H. Klinck, V. Ramesh, V. Joshi, M. Srivathsa, A. Anand, C. Arvind, H. Cp, S. Sawant, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of BirdCLEF 2024: Acoustic Identification of Under-studied Bird Species in the Western Ghats," in *CLEF 2024 Working Notes - 25th Conference and Labs of the Evaluation Forum*, ser. CEUR workshop proceedings, vol. 3740, Grenoble, France, Sep. 2024, pp. 1948–1957. [Online]. Available: https://hal.inrae.fr/hal-04719578

[19] H. Klinck, Maggie, S. Dane, S. Kahl, T. Denton, and V. Ramesh, "Birdclef 2024," https://kaggle.com/competitions/birdclef-2024, 2024, kaggle.

[20] A. Joly, L. Picek, S. Kahl, H. Goëau, V. Espitalier, C. Botella, B. Deneu, D. Marcos, J. Estopinan, C. Leblanc, T. Larcher, M. Sulc, M. Hruz, M. Servajean, J. Matas, H. Glotin, R. Planqué, W.-P. Vellinga, H. Klinck, and H. Müller, *LifeCLEF 2024 Teaser: Challenges on Species Distribution Prediction and Identification*, 03 2024, pp. 19–27.

[21] M. Raghuram, N. R. Chavan, R. Belur, and S. G. Koolagudi, "Bird classification based on their sound patterns," *International journal of speech technology*, vol. 19, pp. 791–804, 2016.

[22] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," *arXiv preprint arXiv:1706.07156*, 2017.

[23] M. AI, "Short time fourier transform and spectrograms," *Audio Analysis Techniques*, 2024, accessed: 2024-12-07. [Online]. Available: https://books.mercity.ai/books/Audio-Analysis-and-Synthesis---Introduction-to-Audio-Signal-Processing/audio_analysis_techniques/01_ShortTime_Fourier_Transform_and_Spectrograms

[24] A. Zhao, K. Subramani, and P. Smaragdis, "Optimizing short-time fourier transform parameters via gradient descent," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 736–740.

[25] U. N. Lab, "Audio recognition using mel spectrograms and convolution neural networks," *ECE228 2019 Reports*, 2019, accessed: 2024-12-07. [Online]. Available: https://noiselab.ucsd.edu/ECE228_2019/Reports/Report38.pdf

[26] Y. Hwang, H. Cho, H. Yang, I. Oh, and S. Lee, "Mel-spectrogram augmentation for sequence to sequence voice conversion," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2001.01401

[27] M. A. Rahman, L. G. Martin, M. Görner, F. Chollet, and P. Culliton, "Birdclef 2024 - kerascv starter train," Kaggle Notebook, 2024, accessed: 2025-01-29. [Online]. Available: https://www.kaggle.com/code/awsaf49/birdclef24-kerascv-starter-train

[28] 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 *Interspeech 2019*. ISCA, 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[29] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020. [Online]. Available: https://arxiv.org/abs/1905.11946

[30] E. Tsalera, A. Papadakis, and M. Samarakou, "Comparison of pre-trained cnns for audio classification using transfer learning," *Journal of Sensor and Actuator Networks*, vol. 10, no. 4, 2021. [Online]. Available: https://www.mdpi.com/2224-2708/10/4/72

[31] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.