DualStack: A Multi-Resolution Spectrogram Fusion Framework for Enhanced Bird Sound Classification

Chitrang Patel, Tanishka Gupta, Sapan H Mankad
Department of Computer Science and Engineering
Institute of Technology, Nirma University
Ahmedabad, India
Email: {22bcm014,22bcm061,sapanmankad}@nirmauni.ac.in

Abstract—**Bird sound classification is pivotal in bioacoustic monitoring, species identification, and ecological conservation. Recent studies have demonstrated that CNN-based approaches, which convert bird sounds into image spectrograms, provide higher classification accuracies. However, in bioacoustics, many researchers still rely on single-resolution spectrograms, which often struggle to capture the diverse temporal and spectral characteristics of avian vocalizations. To address this limitation, we introduce DualStack, a new multi-resolution fusion technique that vertically stacks high-resolution and low-resolution Mel spectrograms into a unified input for a CNN. For comparison, the Biparallel ResNet18 model is employed that simultaneously processes multi-resolution Mel spectrograms parallelly using two branches that handle different resolutions respectively. The results are tested on a dataset comprising 22 bird species with a total of 967 bird sounds, indicate that these two multi-resolution fusion models yield higher accuracies than models using single resolutions. DualStack achieved an accuracy of 86.63%, while the Biparallel ResNet18 model demonstrated an accuracy of 83.66%. In contrast, a single high-resolution model ResNet50 scored 82.18%, and a single low-resolution model ResNet50 achieved an accuracy of 75.74%. To our knowledge, this is the first application of vertical multi-resolution stacking in bioacoustics, offering an automated and scalable approach for real-time ecological monitoring applications.**

Index Terms—**Bird sound classification, multi-resolution fusion, spectrogram, CNN, bioacoustics**

I. Introduction

Bird sound classification is a cornerstone of bioacoustic research as it enables automated species identification, biodiversity, and ecological studies. However, with the increased availability of large-scale audio datasets from Xeno Canto, there is a need for new robust classification methods that handle the complexity of bird vocalizations [2]. Bird calls exhibit significant variability ranging from rapid tonal chirps to sustained broadband vocalizations. Thus, when single resolution-based approaches are applied for classification, they often fail because they have an inherent trade-off: high-resolution spectrograms capture fine temporal details but lose broader spectral context, while low-resolution spectrograms preserve frequency trends at the expense of temporal granularity.

Recently, deep learning, particularly Convolutional Neural Networks (CNNs), has revolutionized audio classification by learning hierarchical features directly from spectrogram representations. Mel spectrograms are used in models as they mimic the human auditory system by providing finer resolutions at lower frequencies [13]. However, single-resolution spectrograms fail to fully exploit the multi-scale nature of bird vocalizations, which limits classification accuracy, especially for species with overlapping frequency distributions or subtle temporal modulations.

Thus, to solve this problem, multi-resolution analysis provides an effective solution. Multi-resolution analysis is widely used in image processing and speech recognition. In speech processing, multi-resolution techniques have improved recognition accuracy by capturing both short-term phoneme transitions and long-term prosodic patterns [6]. However, its use in bioacoustics is still lacking, particularly in bird sound classification. Also, existing multi-resolution fusion approaches like parallel CNN branches or late-stage feature concatenation introduce computational overhead when larger models are used, and they often fail to fully leverage resolution synergies due to independent feature extraction.

To address this, we introduce DualStack, a new multi-resolution fusion approach that vertically stacks high-resolution and low-resolution Mel spectrograms into a single input image, processed by a unified CNN pipeline. Unlike parallel architectures, DualStack ensures spatial coherence between resolutions, allowing CNNs to jointly learn multi-scale features without additional complexity. We evaluated DualStack on a dataset of 967 recordings across 22 bird species [1] and compared it against single-resolution baselines and a BiParallel ResNet18 model. Our results show that DualStack achieves a validation accuracy of 86.63%, significantly outperforming single-resolution models (82.18% high-resolution, 75.74% low-resolution) and BiParallel ResNet18 (83.66%). This opens new gateways for future research using multi-resolution fusion approaches in other areas of bioacoustics.

II. Related Work

Bird sound classification has evolved significantly over the past few years, transitioning from traditional handcrafted features to deep learning-based approaches. However, the application of multi-resolution spectrogram fusion in bioacoustics remains a nascent area of research.

A. Traditional Approaches to Bird Sound Classification

The earlier methods relied on manually engineered features such as Mel-Frequency Cepstral Coefficients (MFCCs), wavelet transforms, and Linear Predictive Coding (LPC). These features were then paired with statistical classifiers like Support Vector Machines (SVMs), Gaussian Mixture Models (GMMs), and Hidden Markov Models (HMMs). However, they achieve moderate success on controlled datasets. However, these methods struggled with real-world challenges such as background noise, overlapping calls, and inter-species variability because they relied on predefined feature sets.

B. Deep Learning and Spectrogram-Based Methods

With the advent of deep learning, the paradigm quickly shifted toward leveraging CNN-based approaches for the classification of spectrogram representations. Various spectrograms other than Mel are used for classification, such as Log-Mel spectrograms and Constant-Q Transforms (CQT) [9] [3]. Log-Mel spectrograms have been explored for their ability to enhance low-energy frequencies, and Constant-Q Transforms (CQT) for their ability to provide better frequency resolution. Recent studies have applied CNNs to large-scale bird sound datasets, demonstrating significant improvements over traditional methods [7]. However, these approaches typically use single-resolution spectrograms which fail to capture the full spectrum of spectral-temporal features, therefore limiting robustness.

C. Multi-Spectrogram Fusion in Audio Classification

Multi-spectrogram fusion technique refers to combining different spectrogram types (e.g., Mel, Gammatone, CQT) to leverage feature representation. Lambamo (2022) fused cochleogram and mel spectrogram for speaker recognition, highlighting the benefits of spectral diversity [5]. However, these studies demonstrated the value of multi-spectrogram fusion but focused on combining different spectrogram types rather than varying resolutions of the same type, which helps capture multi-scale features in bioacoustics.

D. Multi-Resolution Spectrogram Fusion

Multi-resolution analysis has been successfully applied in speech and music processing. Toledano (2018) used multi-resolution speech representations to improve automatic speech recognition [4]. However, in computer vision, multi-resolution techniques like pyramid representations are standard for tasks that require multi-scale feature extraction, but their application in bioacoustics is still limited.

III. Methodology

This section provides a detailed description of the dataset, spectrogram generation, model architectures, training setup, and evaluation metrics.

A. Approach

Our dataset is a small subset from the Xeno Canto dataset. It consists of a total of **967 audio recordings** spanning **22 bird species** [1]. The original dataset comprises raw audio recordings, which we converted into Mel spectrograms to serve as input for CNN-based classification models.

B. Single-Resolution Models

1) High-Resolution Mel Spectrogram Model:

- **Dataset**: High-resolution spectrograms are generated using a 44.1 kHz sampling rate, 4096 FFT window, 1024 hop length, and 256 mel bands.
- **Architecture**: A ResNet50 CNN processes these spectrograms, extracting hierarchical feature representations.
- **Results**: This model achieves 82.18% accuracy, demonstrating that high-resolution spectrograms effectively capture fine-grained temporal details.

2) Low-Resolution Mel Spectrogram Model:

- **Dataset**: Low-resolution spectrograms are generated using a 16 kHz sampling rate, 1024 FFT window, 256 hop length, and 64 mel bands.
- **Architecture**: The same ResNet50 CNN architecture is employed.
- **Results**: The model achieves 75.74% accuracy, indicating that low-resolution spectrograms retain broader spectral trends but sacrifice finer temporal resolution.

C. Multi-Resolution Fusion Models

1) DualStack (Vertical Stacking of Spectrograms):

- **Dataset**: High-resolution and low-resolution spectrograms are concatenated vertically, forming a single stacked image.
- **Architecture**: The ResNet50 model is trained on these vertically stacked spectrograms, processing them as standard input.
- **Results**: Achieves 86.63% accuracy, confirming that combining different resolutions enhances classification performance.

2) BiParallel ResNet18 (Dual-Branch Network):

- **Dataset**: The same dataset previously generated for single-resolution models (high-resolution and low-resolution) was used and passed independently to a dual-branch CNN.
- **Architecture**: A BiParallel ResNet18 processes each spectrogram separately before concatenating the extracted features in the fully connected layer.
- **Results**: Achieves 83.66% accuracy, outperforming single-resolution models but slightly lower than DualStack, due to the lack of direct spatial fusion between resolutions.

D. Training Setup

The training setup was kept the same for all models for fair comparison:

- **Optimizer**: Adam with a learning rate of 0.0001 and weight decay of 1e-4 to prevent overfitting [8].
- **Batch Size**: 32.
- **Loss Function**: CrossEntropyLoss, suitable for multi-class classification.
- **Scheduler**: ReduceLROnPlateau with a factor of 0.1.
- **Hardware**: NVIDIA GTX 1650 GPU, 16GB RAM, Intel i5 CPU.

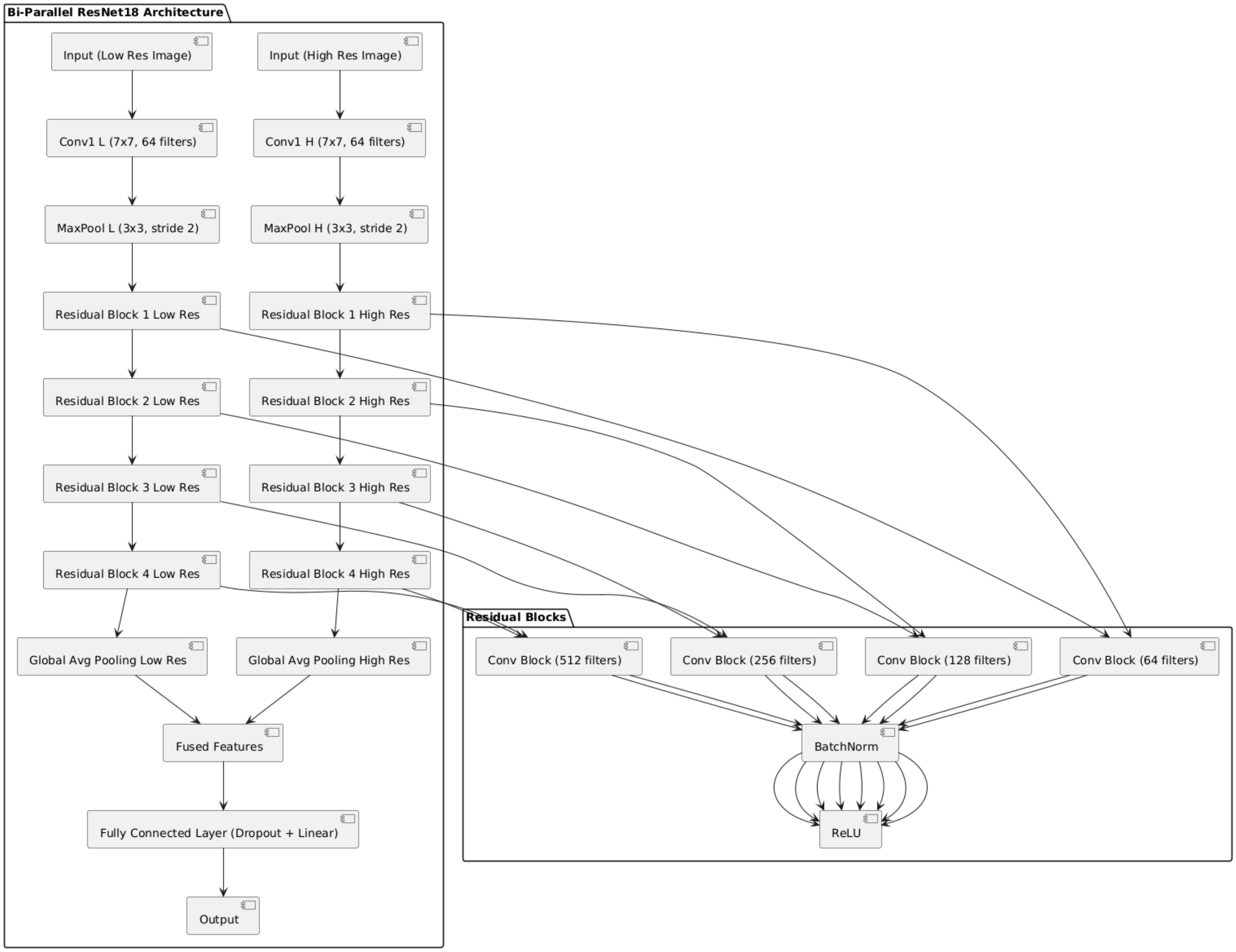
Bi-Parallel ResNet18 Architecture
Input (Low Res Image)
Input (High Res Image)
Conv1 L (7x7, 64 filters)
Conv1 H (7x7, 64 filters)
MaxPool L (3x3, stride 2)
MaxPool H (3x3, stride 2)
Residual Block 1 Low Res
Residual Block 1 High Res
Residual Block 2 Low Res
Residual Block 2 High Res
Residual Block 3 Low Res
Residual Block 3 High Res
Residual Block 4 Low Res
Residual Block 4 High Res
Global Avg Pooling Low Res
Global Avg Pooling High Res
Fused Features
Fully Connected Layer (Dropout + Linear)
Output
Residual Blocks
Conv Block (512 filters)
Conv Block (256 filters)
Conv Block (128 filters)
Conv Block (64 filters)
BatchNorm
ReLU

Fig. 1. BiParallel ResNet18 architecture

E. Evaluation Metrics

- **Validation Accuracy**: Percentage of correctly classified samples. It reflects general classification performance.
- **Macro F1-Score**: Harmonic mean of macro precision and recall. It is averaged across classes, balancing performance despite class imbalance.
- **Macro Precision**: Average precision across classes. It measures the proportion of correct positive predictions.
- **Macro Recall**: Average recall across classes. It assesses the proportion of actual positives correctly identified.
- **Top-1 Accuracy**: Same as overall accuracy. It indicates the percentage of correct top predictions.
- **Top-3 Accuracy**: Percentage of samples where the correct class is among the top three predictions. It shows ranking reliability.
- **Top-5 Accuracy**: Percentage of samples where the correct class is among the top five predictions. It evaluates broader ranking performance.
- **Average per Sample**: Average inference time per sample in milliseconds. It indicates computational efficiency.
- **FPS (Frames Per Second)**: Number of samples processed per second. It measures throughput for high-speed applications.

F. Evaluation Metrics Formulas

The following metrics are used to evaluate the performance of our models, where C is the number of classes (22 bird species), TP_i, FP_i, TN_i, and FN_i represent the true positives, false positives, true negatives, and false negatives for class i, respectively, and N is the total number of samples.

$$\text{Accuracy} = \frac{\sum_{i=1}^{C} TP_i}{N} \tag{1}$$

$$\text{Macro Precision} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i} \tag{2}$$

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i} \tag{3}$$

$$\text{Macro F1-Score} = \frac{1}{C}\sum_{i=1}^{C}\left(2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}\right) \quad (4)$$

where $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$ and $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$.

IV. Results and Analysis

TABLE I
Performance Comparison of Different Models

Metric	DualStack	BiParallel	High Res	Low Res
Validation Accuracy	86.63%	83.66%	82.18%	75.74%
Macro F1-Score	78.44%	68.35%	68.90%	61.70%
Macro Precision	85.59%	70.42%	73.10%	64.48%
Macro Recall	76.19%	67.96%	68.81%	62.65%
Top-1 Accuracy	86.63%	83.66%	82.18%	75.74%
Top-3 Accuracy	94.06%	95.05%	94.06%	90.59%
Top-5 Accuracy	97.52%	97.52%	97.03%	94.55%
Avg. Time (ms)	8.28	6.77	5.77	5.59
FPS	120.81	147.77	173.31	179.02

A. Low-Resolution ResNet50

Low-Resolution ResNet50 achieved an accuracy score of 75.74%, which is the lowest among all the models. It was trained on spectrograms with a 16 kHz sampling rate, a 1024 FFT window, a 256 hop length, and 64 Mel bands. Due to this configuration, the model prioritizes broader spectral trends over fine-grained temporal details. The macro F1-score of 61.70%, macro precision of 64.48%, and macro recall of 62.65% indicate challenges in balanced classification across the 22 classes. The higher precision compared to recall, suggests that the model is more confident in its positive predictions but misses some true positives, particularly for classes with subtle vocalizations. The 16 kHz sampling rate limits the frequency analysis to below 8 kHz, omitting critical high-frequency harmonics, while the longer hop length reduces temporal resolution, affecting the detection of rapid vocalizations.

The model excels in computational efficiency despite these limitations, with an inference time of 5.59 ms per sample and an FPS of 179.02, which is the highest among all models. Thus, the model can be useful in remote resource-constrained ecological monitoring environments.

B. High-Resolution ResNet50 Model

High-Resolution ResNet50 achieved an accuracy score of 82.18%, showing a 6.44% improvement over the Low-Resolution model (Table I). It was trained with spectrograms using a 44.1 kHz sampling rate, a 4096 FFT window, a 1024 hop length, and 256 Mel bands. This model captures high-frequency components up to 22.05 kHz, which is crucial for distinguishing classes with rapidly modulated calls. The macro F1-score of 68.90%, macro precision of 73.10%, and macro recall of 68.81% reflect improved performance, driven by finer temporal resolution that resolves short-duration frequency shifts.

However, the model struggles with inter-class frequency overlap. Its inference time (5.77 ms) and FPS (173.31) are slightly less efficient due to the larger image input size. The top-3 accuracy (94.06%) and top-5 accuracy (97.03%) indicate high prediction confidence, making it suitable for scenarios where ranking classes is valuable, such as biodiversity assessments. Overall, this model shows the importance of capturing fine details, but its limitations suggest that multi-resolution fusion is necessary for further improvements.

C. BiParallel ResNet18 Model

BiParallel ResNet18 achieved an accuracy score of 83.66%, processing high and low-resolution spectrograms through two independent ResNet18 branches before late-stage fusion. It outperforms single-resolution models based on validation accuracy (Table I). The macro F1-score of 68.35%, macro precision of 70.42%, and macro recall of 67.96% indicate that the model delivers balanced performance by leveraging complementary features from both resolutions. A key limitation is the lack of early cross-resolution interaction, as independent processing delays fusion, reducing the model's ability to learn joint patterns, thus contributing to a 16.34% error rate. Its inference time (6.77 ms) and FPS (147.77) reflect higher computational complexity compared to single-resolution models. However, the top-3 accuracy (95.05%) and top-5 accuracy (97.52%), the highest among models, thus making BiParallel architecture suitable for applications needing ranked predictions, such as bioacoustic monitoring. Despite these strengths, it still lags behind DualStack performance, indicating that early fusion is more effective for increasing accuracy.

D. DualStack (Vertical Stacking) Model

DualStack ResNet50 achieved the highest accuracy score of 86.63% (Table I). It vertically stacks high and low-resolution Mel spectrograms into a single input image. Its macro F1-score (78.44%), precision (85.59%), and recall (76.19%) demonstrate superior balanced performance, driven by early fusion that integrates spectral and temporal features from the input level. This approach outperforms BiParallel (83.66%) by 3% and High-Resolution (82.18%) by 4.45%. Top-3 (94.06%) and top-5 (97.52%) accuracies shows that it makes reliable predictions. However, the inference time (8.28 ms) and FPS (120.81) are the highest and lowest, respectively, mainly due to larger image input sizes. Also, a 13.37% error rate suggests challenges like fixed stacking order and intra-class variability exist. Overall, DualStack sets a new benchmark that is ideal for accuracy-focused bioacoustic applications despite its computational cost, highlighting the effectiveness of early fusion in multi-resolution learning.

V. Future Research Directions

The success of DualStack in achieving the highest accuracy among all models clearly shows that early multi-resolution

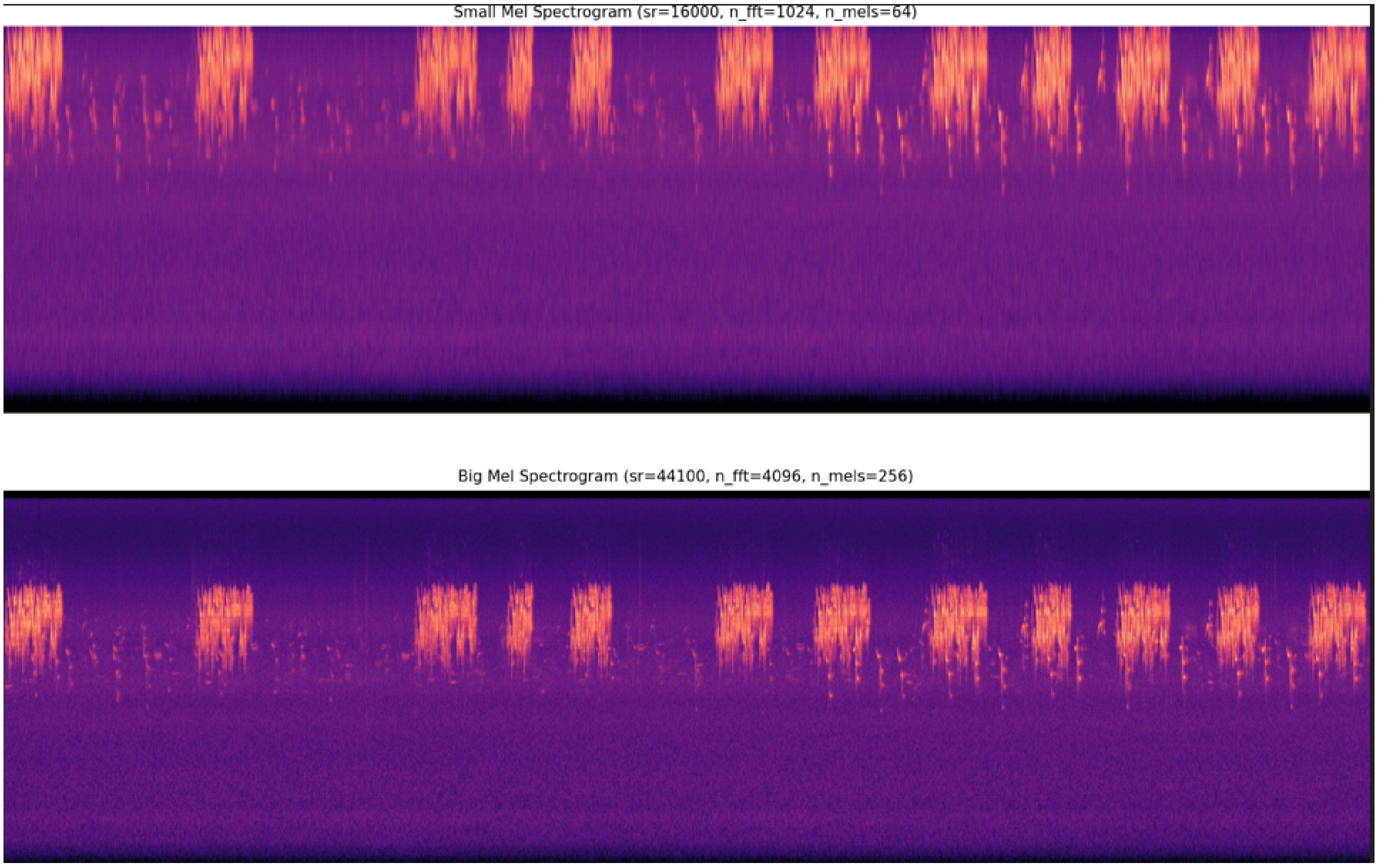

Fig. 2. Example of a single DualStack image, where the high-resolution and low-resolution spectrogram is vertically stacked within one image. This spectrogram is extracted from an MP3 audio recording of the bird Troglodytes pacificus.

fusion opens new avenues for future research in more fields of bioacoustics. One promising direction is the exploration of adaptive multi-resolution learning, where the model dynamically adjusts the resolution based on the characteristics of each class's vocalization, thus improving robustness in learning more diverse patterns. Also, incorporating self-supervised learning techniques could further enhance DualStack's performance by leveraging large-scale, unlabeled audio datasets, reducing reliance on labeled data and improving generalization to underrepresented classes or noisy environments. Future work should also aim to reduce the computational overhead of 8.28 ms by using lightweight models such as MobileNet to enable on-the-edge processing in remote monitoring systems [11]. Additionally, extending the use of the DualStack framework to other areas of the bioacoustics domain, such as amphibian or bat monitoring, and integrating uncertainty estimation methods could enhance its reliability for real-time ecological applications [12]. Finally, expanding the dataset to include more diverse environmental conditions and a broader range of classes could mitigate intra-class variability and frequency overlap challenges and other overall challenges, further improving classification performance [10].

VI. Conclusion

In this paper, we introduced DualStack, a new multi-resolution fusion technique for bird sound classification. Our experiments have shown that DualStack outperforms both single-resolution models and the BiParallel ResNet18 model, thus, achieving an accuracy of 86.63%. The success of DualStack also highlights the importance of early fusion in capturing multi-scale features from bird vocalizations. Future studies will now focus on reducing computational efficiency and extending this DualStack framework to other bioacoustic applications.

References

[1] L. Jeantet and E. Dufourq, “Manually labeled bird song dataset of 22 species from Xeno-canto to enhance deep learning acoustic classifiers with contextual information,” Zenodo, 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7828148

[2] S. J. Sober, M. J. Wohlgemuth, and M. S. Brainard, ”Central contributions to acoustic variation in birdsong,” J. Neurosci., vol. 28, no. 41, pp. 10370-10379, 2008.

[3] F. Wolf-Monheim, ”Spectral and rhythm features for audio classification with deep convolutional neural networks,” arXiv preprint arXiv:2410.06927, 2024.

[4] D. T. Toledano, M. P. Fernández-Gallego, and A. Lozano-Diez, ”Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT,” PLoS ONE, vol. 13, no. 10, p. e0205355, Oct. 2018, doi: 10.1371/journal.pone.0205355.

[5] W. Lambamo, R. Srinivasagan, and W. Jifara, ”Fusion of cochleogram and mel spectrogram features for deep learning based speaker recognition,” 2022.

[6] S. Harding and G. F. Meyer, ”Multi-resolution auditory scene analysis: robust speech recognition using pattern-matching from a noisy signal,” in Proc. INTERSPEECH, 2003, pp. 2109-2112.

[7] G. Gupta, M. Kshirsagar, M. Zhong, S. Gholami, and J. Lavista Ferres, ”Comparing recurrent convolutional neural networks for large scale bird species classification,” Sci. Rep., vol. 11, no. 1, p. 17085, 2021.

[8] D. P. Kingma and J. Ba, ”Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[9] P. Singh, G. Saha, and M. Sahidullah, ”Non-linear frequency warping using constant-Q transformation for speech emotion recognition,” in Proc. 2021 Int. Conf. Comput. Commun. Informatics (ICCCI), pp. 1-6, 2021.

[10] C. M. Wood, J. Champion, C. Brown, W. Brommelsiek, I. Laredo, R. Rogers, and P. Chaopricha, ”Challenges and opportunities for bioacoustics in the study of rare species in remote environments,” Conserv. Sci. Pract., vol. 5, no. 6, p. e12941, 2023.

[11] O. D. Incel and S. Ö. Bursa, ”On-device deep learning for mobile and wearable sensing applications: A review,” IEEE Sensors J., vol. 23, no. 6, pp. 5501-5512, 2023.

[12] A. Kendall and Y. Gal, ”What uncertainties do we need in Bayesian deep learning for computer vision?,” Advances in Neural Information Processing Systems, vol. 30, 2017.

[13] B. Z. J. L. S. Thornton, Audio recognition using mel spectrograms and convolutional neural networks, Academia, San Francisco, CA, USA, 2019.