Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 4 (2026) 878 - 887 e-ISSN: 2580-0760

Lightweight Convolutional Neural Network for Robust and Label-Efficient Animal Sound Classification

Ivan Saputra[1], Rikman Roedavan[2*]
[1,2]Department of Multimedia Engineering, School of Applied Sciences, Telkom University, Bandung, Indonesia
[*]*rroedavan@gmail.com*

Abstract

Automated bioacoustic monitoring is increasingly important for biodiversity observation, yet practical deployment is often limited by computational constraints and the scarcity of annotated audio data. This study introduces MelCNN, a custom compact end-to-end architecture designed to mitigate over-parameterization and negative transfer in animal sound classification under small-data conditions. A balanced dataset of 600 audio clips from three animal classes (Cat, Dog, and Cow) was standardized to 4.8 seconds, converted into Log-Mel spectrograms, and evaluated using 5-fold stratified cross-validation. The proposed Lightweight CNN with Global Average Pooling was compared against YAMNet-based transfer learning baselines and three additional modern lightweight CNN baselines, namely MobileNetV3-Small, ShuffleNetV2-1.0x, and EfficientNet-Lite0-style. The proposed model achieved 97.00% ± 2.01% Accuracy and 97.01% ± 2.00% Macro-F1 while retaining a compact parameter size of approximately 0.39 million, outperforming both the strongest transfer learning baselines and the added modern lightweight baselines in the present setting. The model also maintained performance above 90% at 12 kHz and preserved approximately 90.5% Macro-F1 when trained with only 30% of the available labeled data. These findings indicate that a domain-specific lightweight architecture can provide a favorable accuracy-efficiency trade-off for controlled small-data animal sound classification in resource-constrained monitoring scenarios.

Keywords: animal sound classification; bioacoustic monitoring; label efficiency; lightweight CNN; Log-Mel spectrogram

How to Cite: I. Saputra and R. A. Rudawan, "Lightweight Convolutional Neural Network for Robust and Label-Efficient Animal Sound Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 4, pp. 878 - 887, Jul. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i4.7366

Received: December 1, 2025
Accepted: June 21, 2026
Available Online: July 19, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

Environmental monitoring through acoustic surveillance has emerged as a cornerstone in biodiversity conservation and ecological assessment. The automated classification of animal sounds facilitates critical tasks ranging from early warning systems for poaching to the forensic analysis of wildlife presence in protected areas [1], [2]. As the volume of acoustic data grows exponentially, the reliance on manual annotation becomes untenable, necessitating robust automated systems driven by Deep Learning (DL) algorithms [3]. Recent studies have demonstrated the efficacy of Convolutional Neural Networks (CNNs) in capturing intricate bioacoustic patterns, often outperforming traditional signal processing techniques [4]. However, the transition from experimental prototypes to real-world deployment on edge devices, such as Raspberry Pi or remote sensors, is severely impeded by critical bottlenecks stemming from extreme computational constraints and the acute scarcity of high-quality annotated data in specific ecological domains [5].

The prevailing paradigm in audio classification heavily relies on Transfer Learning (TL) using large-scale pre-trained models such as VGGish, YAMNet, or complex Transformers trained on massive datasets like AudioSet or ImageNet [6]. While these models achieve high accuracy, they are fundamentally over-parameterized for specific bioacoustic tasks involving a limited number of classes [7]. The computational footprint of these heavy architectures often precludes their use in real-time monitoring scenarios where energy efficiency is paramount [8], [9]. Furthermore, blind application of transfer learning can lead to negative transfer when the source domain (e.g., human speech or music) significantly diverges from the target domain of environmental animal sounds, suggesting that custom-

designed architectures may offer superior efficiency without sacrificing accuracy [10], [11]. Consequently, there is an urgent need for lightweight architectures that are natively designed to minimize floating-point operations while maintaining discriminative power [5].

Beyond computational weight, the "data hunger" of deep learning poses a critical barrier. Collecting and annotating field recordings is labor-intensive and requires domain expertise, often resulting in small, imbalanced datasets [12], [13]. Although frameworks utilizing Knowledge Distillation [14], Few-Shot Learning [15], and cross-modal Self-Supervised Learning [3] address data scarcity, they introduce high computational overhead and architectural complexity [16], [17]. Similarly, alternative networks like BMNet are fundamentally optimized for broader domains such as speech or music rather than the highly non-stationary properties of environmental animal vocalizations [18], [19]. Consequently, these methods remain sub-optimal for edge deployment due to their intensive training requirements and lack of tuning for specific time-frequency acoustic representations.

To address these gaps, this study proposes MelCNN, a specialized lightweight architecture. The structural contribution lies in the optimization of a restricted parameter space paired with a Global Average Pooling (GAP) layer to prevent overfitting natively, bypassing the need for complex distillation pipelines or large pre-trained models. We rigorously benchmark this proposed lightweight model against both standard transfer learning baselines and additional modern lightweight baselines to examine whether a compact domain-specific architecture can remain competitive under controlled small-data conditions. Furthermore, this research provides a critical analysis of label efficiency by systematically evaluating the performance stability of the model under drastic data reduction scenarios, ranging from the full training set down to only 10% of available data. In addition to efficiency, we assess the robustness of the architecture against environmental perturbations, specifically analyzing its resilience to additive noise and signal resampling to ensure reliability in real-world deployment.

This study evaluates whether a compact architecture can maintain structural stability and accuracy under constrained data environments. This approach is necessitated by the high operational costs and inconsistency associated with collecting large-scale, annotated bioacoustic datasets [12], [13], [20]. Rather than asserting universal generalizability, this work establishes a baseline for classification efficiency using localized datasets. Accordingly, the present work emphasizes balanced class composition, stratified cross-validation, robustness analysis, and label-efficiency assessment as complementary forms of validation, thereby allowing the proposed model to be examined not only in terms of peak predictive accuracy, but also in terms of stability under constrained training conditions and suitability for lightweight deployment [5], [9]. From this perspective, the contribution of the study lies in demonstrating that rigorous evaluation on a carefully controlled dataset can still yield meaningful evidence regarding architectural efficiency, robustness, and annotation sensitivity, which are central requirements for practical bioacoustic systems operating in resource-constrained environments [5], [20].

2. Methods

This study establishes a rigorous research workflow designed to ensure reproducibility and robustness, aligned with recent standards in efficient deep learning deployment [20], [21]. The methodology comprises dataset curation, acoustic feature extraction using Log-Mel Spectrograms, and the design of a custom Lightweight CNN architecture optimized for label efficiency.

2.1 Dataset Curation and Preprocessing

A balanced dataset of 600 environmental animal sound clips was curated in this study, covering three target classes, namely Cat, Dog, and Cow, with 200 samples allocated to each class. This balanced composition was intentionally maintained to avoid class dominance during training and to ensure that the reported Macro-F1 score reflects discriminative performance across all categories rather than bias toward a larger class. For a lightweight bioacoustic classification task with limited labeled data, such a balanced design provides a controlled and methodologically reliable setting for evaluating model behavior and for comparing the proposed architecture against both transfer learning and modern lightweight baseline families [5], [20].

The dataset combines publicly available audio from Kaggle repositories with self-recorded clips captured using smartphone devices in natural everyday environments. The self-recorded samples were obtained at close range to the target animals so that the vocalizations could be captured as clearly as possible while reducing irrelevant background interference. This recording strategy was important because animal vocal events are often brief, irregular, and difficult to collect in large quantities under consistent acoustic conditions. Therefore, the resulting corpus was designed as a controlled but acoustically diverse dataset that still reflects practical recording variability relevant to edge-based monitoring scenarios [20].

As summarized in Table 1, the curated corpus has a median clip duration of 4.77 seconds and a total duration of approximately 2,864 seconds. Based on this empirical distribution, all recordings were standardized to a fixed duration of 4.8 seconds. This choice was made for three reasons. First, 4.8 seconds closely matches the natural central tendency of the collected recordings, so the selected duration preserves the dominant temporal structure of the dataset rather than imposing an arbitrary cutoff. Second, because many target vocalizations in this corpus are short and distinct, a 4.8-second window

is sufficient to retain the main acoustic event while minimizing excessive silence or irrelevant contextual sound. Third, fixed-duration inputs are necessary to produce uniform spectrogram tensors and to eliminate input-length bias during batch training, thereby improving numerical consistency and optimization stability in CNN-based classification.

Before feature extraction, all audio files were resampled to 16 kHz to ensure a uniform sampling format across recordings. Clips shorter than 4.8 seconds were padded with silence, whereas clips longer than 4.8 seconds were trimmed to preserve a consistent temporal window for analysis. RMS normalization was then applied to reduce loudness variation across recordings so that the classifier learned spectral-temporal structure rather than absolute amplitude differences. In addition, controlled augmentation was introduced during training as a regularization strategy to improve generalization under limited-data conditions, consistent with prior findings that augmentation can support audio classification when applied carefully [22].

Overall, this preprocessing pipeline was designed not only to standardize heterogeneous recordings, but also to preserve the essential acoustic content of short animal vocalizations within a compact and reproducible representation suitable for lightweight deep learning.

Table 1. Dataset Distribution and Duration Statistics

Class	Samples (N)	Total Duration (s)	Median Duration (s)
Cat	200	954.7	4.77
Cow	200	954.7	4.77
Dog	200	954.7	4.77
Total	600	2,864.2	

2.2 Feature Extraction: Log-Mel Spectrogram

We utilize Log-Mel Spectrograms as the input representation, a choice supported by their proven efficacy in capturing semantic audio features for complex attention-based and convolutional models [23], [24]. The Short-Time Fourier Transform (STFT) is first computed using a Hann window of size 25 ms ($N_{fft} = 400$) and a hop length of 10 ms ($H = 160$). The power spectrum $|S(f,t)|^2$ is then mapped onto the Mel scale using a filter bank of $M = 128$ triangular filters.

The final Log-Mel feature X_{mel} is obtained by applying a logarithmic operation to compress the dynamic range, as defined in Equation 1.

$$X_{mel(m,t)} = \log\left(S_{mel(m,t)} + \epsilon\right) \quad (1)$$

S_{mel} is the Mel-scaled power spectrogram and $\epsilon = 10^{-6}$ ensures numerical stability.

Prior to the initiation of the training phase, a systematic data validation protocol was executed to ensure the reliability of the input pipeline. The cleaning process confirmed that 100% of the 600 audio files were successfully transformed into Mel-spectrograms, guaranteeing the structural and numerical integrity of the dataset with zero instances of file corruption or missing values. Verification of the tensor dimensions established strict uniformity across the corpus, where every sample adheres to an identical shape of (1, 128, 480). This consistency eliminates the computational overhead associated with dynamic padding or truncation during the learning process. To illustrate the discriminative quality of these features, Figure 1 provides a visual representation of the extracted Log-Mel Spectrograms, highlighting the distinct time-frequency spectral signatures inherent to each class that serve as the fundamental basis for classification.

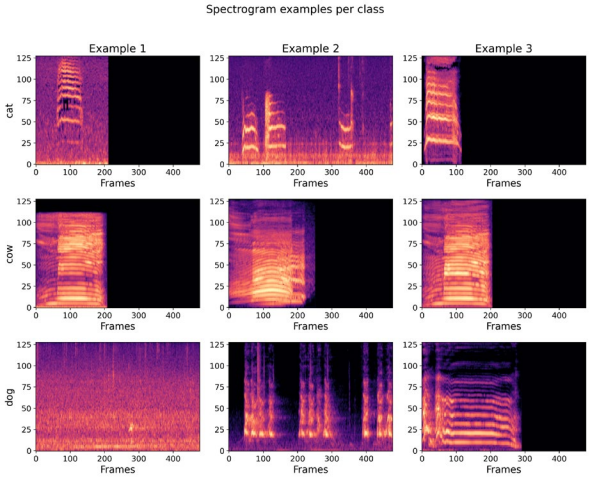


Figure 1. Visualization of Log-Mel Spectrogram features for (a) Cat, (b) Cow, and (c) Dog classes. The vertical axis represents frequency bins (Mel scale), and the horizontal axis represents time frames.

2.3 Proposed Lightweight CNN Architecture

To address the computational constraints of edge deployment, we propose a custom Lightweight CNN (MelCNN). This design philosophy mirrors the contemporary paradigm shift towards efficient architectures, exemplified by models such as EfficientNetV2 and squeezed CNNs [25], [26] which prioritize significant parameter reduction without compromising classification accuracy.

The backbone extracts hierarchical features through four sequential blocks. Each block i performs a convolution operation, Batch Normalization (BN), ReLU activation, and Max Pooling, as formulated in Equation 2.

$$y_i = \text{MaxPool}\left(\text{ReLU}\left(\text{BN}(W_i * x_{i-1})\right)\right) \quad (2)$$

W_i represents the kernel weights. The channel depth increases progressively $(32 \rightarrow 64 \rightarrow 128 \rightarrow 256)$ while the spatial dimensions are reduced by pooling layers to capture high-level abstractions.

A critical innovation in our design is the use of a Global Average Pooling (GAP) layer instead of traditional flattening. While diverse pooling strategies like Global Entropy Pooling have been explored [27], GAP remains the gold standard for minimizing overfitting in lightweight models by enforcing a direct correspondence between feature maps and categories.

GAP spatially averages the feature map $F \in \mathbb{R}^{C \times H \times W}$ into a feature vector $v \in \mathbb{R}^{C}$, drastically reducing the parameter count in the dense layer [28].

Table 2. Architecture Specification of Proposed Lightweight CNN

Layer	Output Shape	Parameters
Input	$(1, 128, 480)$	0
ConvBlock 1	$(32, 64, 240)$	320
ConvBlock 2	$(64, 32, 120)$	18,496
ConvBlock 3	$(128, 16, 60)$	73,856
ConvBlock 4	$(256, 8, 30)$	295,168
Global Avg Pool	$(256, 1, 1)$	0
Linear (Dense)	(3)	771
Total Params		~388,611

As detailed in Table 2, this architecture results in approximately 0.39 million parameters, which is substantially more compact than the transfer learning baselines and therefore more suitable for deployment on resource-constrained devices.

2.4 Experimental Setup

To ensure a reliable and unbiased evaluation, all experiments were conducted using 5-fold Stratified Cross-Validation, such that the class distribution remained consistent across the training and validation partitions in every fold. This protocol was selected to reduce the likelihood of favorable split effects and to provide a more stable estimate of model generalization under limited-data conditions. Model optimization was performed using the Adam optimizer and Cross-Entropy Loss, which are widely adopted for multi-class deep learning tasks due to their stable convergence behavior.

In addition to standard predictive evaluation, this study employed a controlled simulation-based stress-testing protocol to examine whether the proposed lightweight architecture remains reliable under practical constraints relevant to bioacoustic monitoring and edge-oriented deployment [5], [20]. In this paper, simulation testing does not refer to synthetic data generation. Rather, it refers to a structured set of controlled perturbation experiments designed to emulate three realistic operational challenges, namely signal degradation, reduced hardware fidelity, and limited label availability.

The first simulation scenario focused on acoustic degradation. To evaluate robustness against adverse environmental interference, the input signals were perturbed using additive noise at Signal-to-Noise Ratio (SNR) levels ranging from 0 dB to 20 dB. This test was designed to examine whether the classifier could preserve discriminative performance when the target vocalization was partially obscured by noise, which is a common condition in field recording and remote sensing applications [20].

The second simulation scenario focused on hardware-related fidelity constraints. Specifically, the input audio was resampled between 8 kHz and 12 kHz to mimic lower-quality acquisition conditions that may arise in bandwidth-limited or resource-constrained sensing devices. This resampling analysis was not intended to replace real-device benchmarking, but rather to provide a controlled proxy for assessing whether the proposed model remains functionally stable when high-frequency detail is partially lost, as often occurs in practical low-power audio systems [5], [9].

The third simulation scenario addressed annotation scarcity through a systematic Label Efficiency analysis. In this setting, the model was iteratively trained on stratified subsets representing 10%, 20%, 30%, 50%, and 100% of the available training data. This procedure was included to quantify the dependency of the proposed architecture on labeled data volume and to determine whether meaningful classification performance could still be maintained when annotation resources are severely constrained, which remains a major challenge in bioacoustic studies [20].

To further strengthen the evaluation under small-data conditions, augmentation policies were also examined as part of the experimental design. These augmentation settings were treated as controlled regularization strategies rather than as substitutes for real data. Their inclusion was motivated by prior evidence showing that augmentation can improve generalization in voice and audio classification when applied appropriately [22]. The effect of each policy was then analyzed separately in the ablation study to determine whether added acoustic variability improved or degraded recognition performance in short animal vocalizations.

Overall, the experimental setup was intentionally structured to evaluate the proposed model from four complementary perspectives: statistical stability across folds, comparative accuracy across transfer learning and modern lightweight baselines, robustness under controlled signal and hardware perturbations, and resilience to label scarcity. This design provides a more rigorous assessment of practical usability than a single train-test performance report and directly supports the central objective of developing a compact yet reliable animal sound classifier for resource-constrained environments [5], [20].

2.5 Comparative Baseline Configuration

To strengthen the comparative validity of the study, the evaluation protocol was extended to include both transfer learning baselines and additional modern lightweight convolutional baselines under a unified experimental setting. The transfer learning baselines consisted of YAMNet embeddings followed by MLP, Logistic Regression, and SVM classifiers, whereas the modern lightweight baselines consisted of MobileNetV3-Small, ShuffleNetV2-1.0x, and EfficientNet-Lite0-style. This expanded configuration was introduced to ensure that the proposed MelCNN was not assessed only against generic pre-trained embedding pipelines, but also against recent compact architectures that represent contemporary efficiency-oriented design trends in sound classification and resource-constrained inference [4], [5], [9], [10].

To preserve methodological fairness, all compared models were evaluated using the same curated dataset, the same standardized Log-Mel spectrogram input representation, and the same 5-fold stratified cross-validation protocol described in the preceding subsection. Accordingly, the comparison was designed to isolate architectural differences as much as possible, rather than allowing performance variation to be driven by inconsistent preprocessing or data partitioning. For the end-to-end lightweight CNN baselines, the models were trained directly on the same spectrogram inputs used by the proposed architecture, whereas the transfer learning baselines operated on the corresponding YAMNet feature representation. Comparative performance was then examined primarily in terms of Accuracy, Macro-F1, and model compactness, so that predictive effectiveness and computational efficiency could be considered jointly within the same analytical framework. From this perspective, the purpose of the expanded benchmark was not merely to identify the highest-scoring model, but to determine whether the proposed domain-specific lightweight architecture could maintain a superior accuracy-efficiency trade-off when challenged by both established transfer learning pipelines and more recent lightweight CNN families relevant to edge-oriented bioacoustic monitoring [5], [20].

3. Results and Discussions

This section presents a comprehensive evaluation of the proposed Lightweight CNN (MelCNN) by systematically examining the model's performance across four critical analytical dimensions. The investigation begins with an assessment of training stability to ensure convergence reliability and proceeds to a rigorous comparative benchmarking against both transfer learning and modern lightweight baselines. Subsequently, we conduct a granular class-wise error profiling to identify specific failure modes and conclude with an extensive stress-testing phase that quantifies both robustness against signal degradation and label efficiency under data scarcity.

3.1 Data Distribution and Training Stability

Before evaluating classification performance, it is necessary to examine both the empirical characteristics of the curated audio corpus and the stability of the training process. Figure 2 presents the distribution of raw audio clip durations in the curated dataset, with a median duration of 4.77 seconds. This empirical duration profile motivated the subsequent standardization of all recordings to a fixed length of 4.8 seconds, thereby eliminating input-length bias during batch processing and ensuring a uniform spectrogram representation for model training.

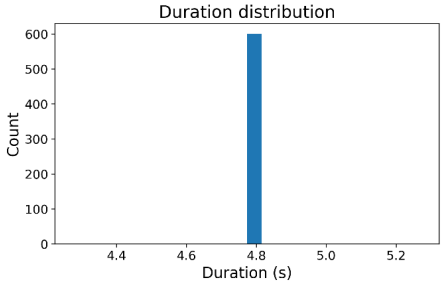


Figure 2. Distribution of raw audio clip durations in the curated dataset. The concentration of durations around the median provides the empirical basis for standardizing all recordings to a fixed length of 4.8 seconds.

The training dynamics of the proposed MelCNN are illustrated in Figure 3. Although the maximum training budget was set to 50 epochs, an early stopping mechanism terminated the optimization process between 16 and 22 epochs across all folds upon reaching convergence to prevent overfitting.

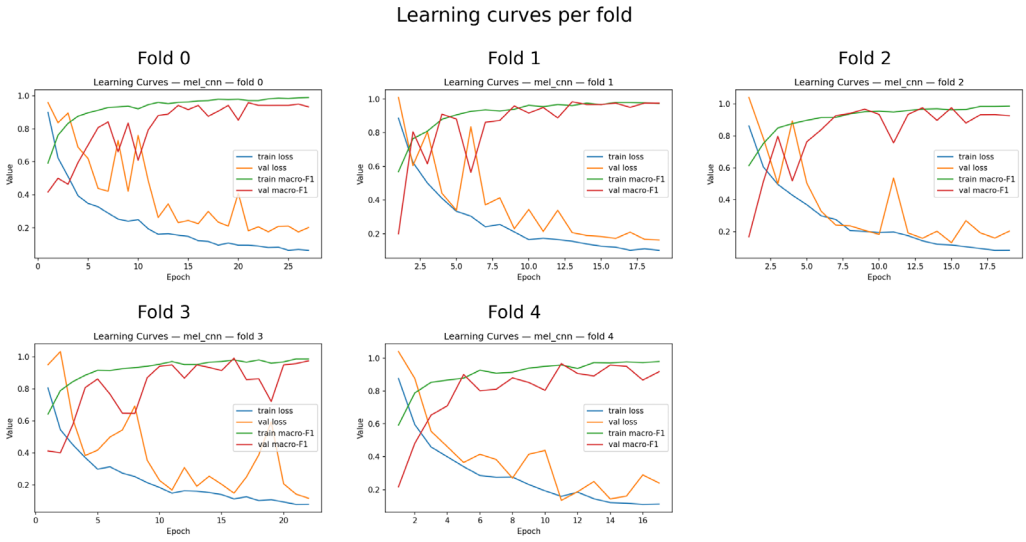


Figure 3. Learning curves of the proposed MelCNN across a maximum of 50 epochs with early stopping termination, showing training and validation loss on the left and accuracy on the right.

Examination of the learning curves reveals a stable convergence profile characterized by the synchronous decline of both training and validation losses. The negligible divergence observed between the training and validation accuracy trajectories provides empirical evidence that the model maintains robust generalization capabilities throughout the optimization phase. Although training Convolutional Neural Networks on restricted datasets of only 600 samples typically introduces a high risk of memorization, our architecture effectively mitigates overfitting. This stability is directly attributed to the constrained parameter space of 388k weights and the regularization effects inherent in the design, ensuring that the model learns discriminative acoustic features rather than fitting noise within the training data.

To rigorously substantiate the generalization capability of the model, Table 3 presents the performance metrics obtained through a comprehensive 5-fold cross-validation procedure. The empirical results show generally consistent performance across the folds, with the proposed model achieving a mean Macro-F1 of 97.01%, a mean mAP of 98.03%, and a mean Accuracy of 97.00%. The corresponding standard deviations of 2.00%, 1.32%, and 2.01% indicate that, although some variation across data partitions remains expected under a limited-data setting, the overall predictive performance of the architecture remains consistently high. These findings support the interpretation that the proposed MelCNN learns robust and discriminative acoustic representations, rather than benefiting from a favorable split or isolated partition effect.

Table 3. 5-Fold Cross-Validation Performance

Fold	Macro-F1	mAP	Accuracy
0	94.16%	96.08%	94.17%
1	98.34%	97.42%	98.33%
2	97.50%	99.31%	97.50%
3	99.17%	98.25%	99.17%
4	95.89%	99.10%	95.83%
Mean	97.01%	98.03%	97.00%
Std. Dev.	2.00%	1.32%	2.01%

3.2 Comparative Analysis with Baselines

We extended the comparative evaluation beyond the original transfer learning benchmark by additionally including three modern lightweight convolutional architectures, namely MobileNetV3-Small, ShuffleNetV2-1.0x, and EfficientNet-Lite0-style, all assessed under the same preprocessing pipeline, 5-fold stratified cross-validation protocol, and evaluation metrics. As presented in Table 4, the proposed MelCNN achieved the strongest overall classification performance in terms of Accuracy and Macro-F1 while simultaneously retaining the smallest parameter count among all compared models. Specifically, the proposed model reached 97.00% ± 2.01% Accuracy and 97.01% ± 2.00% Macro-F1 with approximately 0.39 million parameters. In contrast, the strongest transfer learning baselines, YAMNet + LogReg and YAMNet + SVM, both reached 96.50% Accuracy and 96.49% Macro-F1 with approximately 3.70 million parameters, whereas the modern lightweight baselines remained lower, with MobileNetV3-Small yielding 93.83% Accuracy and 93.85% Macro-F1, ShuffleNetV2-1.0x yielding 95.00% Accuracy and 94.97% Macro-F1, and EfficientNet-Lite0-style yielding 95.00% Accuracy and 94.98% Macro-F1. These results indicate that the proposed architecture not only outperforms generic transfer-learning pipelines, but also remains superior when compared against recent compact CNN families that were explicitly designed for efficient inference.

Table 4. Extended Performance Comparison of Transfer Learning Baselines, Modern Lightweight CNN Baselines, and the Proposed Model

Model Variant / Architecture	Accuracy	Macro-F1	Parameters
YAMNet + MLP	94.00% ± 2.60%	93.96% ± 2.59%	~3.70 M
YAMNet + LogReg	96.50% ± 1.09%	96.49% ± 1.09%	~3.70 M
YAMNet + SVM	96.50% ± 0.70%	96.49% ± 0.70%	~3.70 M
MobileNetV3-Small	93.83% ± 0.46%	93.85% ± 0.46%	~1.52 M
ShuffleNetV2-1.0x	95.00% ± 2.04%	94.97% ± 2.08%	~1.26 M
EfficientNet-Lite0-style	95.00% ± 2.04%	94.98% ± 2.07%	~3.37 M
Proposed Model Lightweight CNN	97.00% ± 2.01%	97.01% ± 2.00%	~0.39 M

Values are reported as mean ± standard deviation across 5-fold stratified cross-validation.

These findings are consistent with the broader literature showing that lightweight and task-specific architectures can offer a more favorable balance between predictive performance and computational efficiency than over-parameterized transfer learning models in environmental sound recognition [5], [9], [10]. At the same time, the present results should not be interpreted as evidence that transfer learning is ineffective for bioacoustic tasks, since the YAMNet-based baselines remained highly competitive and prior work has shown that transfer learning can be highly useful in acoustic species classification [4]. Rather, the expanded benchmark suggests a more specific conclusion: in a controlled three-class animal vocalization task with standardized Log-Mel inputs and limited labeled data, model suitability depends less on architectural popularity or pre-training scale alone and more on the degree of alignment between the network design and the acoustic structure of the target problem. From this perspective, the proposed MelCNN provides the most favorable accuracy-efficiency trade-off in the present setting, supporting the argument that a compact domain-specific architecture can be more scientifically appropriate and practically relevant than both generic transfer-learning embeddings and more general-purpose lightweight CNN backbones for resource-constrained bioacoustic monitoring [5], [20].

3.3 Class-wise Performance and Error Profiling

To identify specific weaknesses in the model, we conducted a granular class-wise analysis. Table 5 details the Precision, Recall, and F1-Score for each

animal class. The model achieves the highest recall for Cow (98.5%), likely due to the distinct low-frequency harmonic structure of cow moos.

Table 5. Class-wise Performance Metrics

Class	Precision (Mean)	Recall (Mean)	F1-Score (Mean)
Cat	96.05%	97.00%	96.51%
Cow	97.60%	98.50%	98.01%
Dog	98.97%	97.00%	97.97%

However, an analysis of the Confusion Matrix (Figure 4) combined with Table 5 reveals a slight performance drop in the Cat class (F1 96.51%). The matrix indicates that the primary source of error is Dog samples being misclassified as Cat. This misclassification can be attributed to the spectral similarity between high-pitched barks (small dogs) and aggressive cat meows, which share overlapping frequency bands in the 1-4 kHz range.

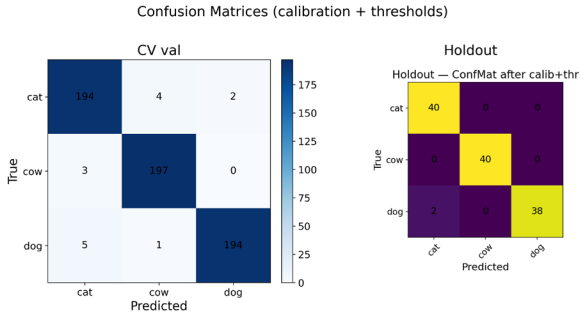


Figure 4. Confusion Matrix of the proposed model. While Cow predictions are nearly perfect, minor confusion exists between Dog and Cat classes due to spectral overlap.

3.4 Ablation Study and Robustness Analysis

A counterintuitive finding emerged from our ablation study on augmentation strategies, summarized in Table 6. While augmentation is typically recommended for small datasets [22], our experiments reveal that aggressive strategies like SpecAugment (time/frequency masking) significantly degrade performance (-5.1% F1-Score). This suggests that for short, distinct animal sounds, masking parts of the spectrogram destroys critical temporal-spectral cues.

Table 6. Ablation Study on Augmentation Policies

Augmentation Policy	Macro-F1	Accuracy	Observation
No Augmentation	97.83%	97.83%	Best Performance
RARPM	94.13%	94.17%	Moderate Degradation
SpecAugment	92.77%	92.83%	Significant Degradation

In the final phase of the evaluation, we applied a controlled stress-testing protocol to examine how the proposed model behaves under signal degradation, reduced audio fidelity, and limited training data, all of which are relevant to practical edge-based bioacoustic monitoring. Figure 5 illustrates the performance trajectory under varying sampling rates, providing a proxy for determining hardware compatibility. Although a performance attenuation is observed when downsampling to 8 kHz due to the inevitable loss of high-frequency spectral fidelity, the architecture exhibits strong stability by sustaining an F1-score above 90% at a 12 kHz sampling rate. This resilience provides meaningful support for the practical suitability of the model in edge-oriented monitoring scenarios where high-fidelity audio capture is often constrained by limited bandwidth and processing power.

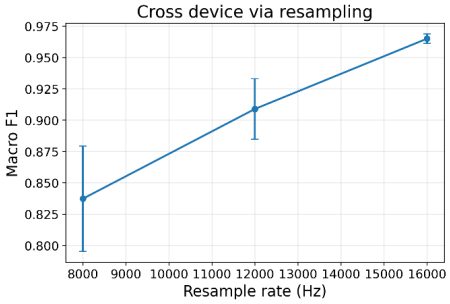


Figure 5. Robustness analysis showing performance degradation under signal resampling. The model maintains >90% F1-score even at 12 kHz sampling rate.

To provide a conclusive validation of the proposed model's practical utility, Figure 6 delineates the Label Efficiency curve, offering a granular analysis of performance degradation relative to training set availability.

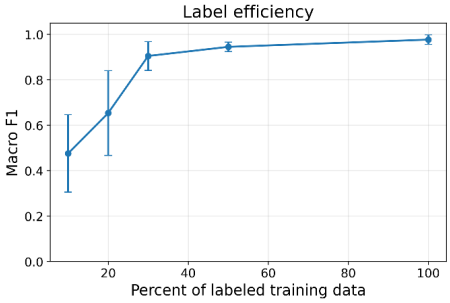


Figure 6. Label Efficiency Curve. The model demonstrates exceptional data efficiency, achieving over 90% F1-score with only 30% of the available training data.

The empirical trajectory reveals a highly favorable trade-off between annotation volume and classification accuracy. Most notably, the model exhibits exceptional resilience to data scarcity by sustaining a Macro-F1 score exceeding the 90% threshold (specifically achieving approximately 90.5%) even when the training corpus is drastically truncated to utilize a mere 30% of the original labeled samples. This finding serves as a robust confirmation that the proposed lightweight architecture, underpinned by the Global Average Pooling mechanism and Log-Mel feature extraction, successfully captures invariant acoustic signatures without necessitating the vast datasets typically required by over-parameterized transfer learning models. Consequently, this research establishes the proposed MelCNN as a pragmatic and scalable paradigm for biodiversity monitoring, as it effectively addresses the bottleneck of expert-level annotation and proves that high-performance bioacoustic surveillance

is achievable in resource-constrained environments where labeled data is both prohibitively expensive and scarce.

From a deployment perspective, these results provide meaningful evidence that the proposed architecture is practically oriented toward edge-based bioacoustic monitoring, even though direct hardware benchmarking was not performed in the present study. The compact parameter scale of approximately 0.39 million, combined with sustained performance above 90% under 12 kHz resampling and strong label efficiency at only 30% of the training data, indicates that the model is not merely accurate under controlled evaluation, but also structurally aligned with the operational constraints typically encountered in low-power sensing systems [5], [9], [20]. In particular, the ability to preserve discriminative performance under reduced audio fidelity suggests tolerance to bandwidth-limited acquisition, while the low parameter burden supports the feasibility of memory-efficient inference relative to heavier transfer learning alternatives. At the same time, it is important to state that computational practicality in this study is inferred from architectural compactness and robustness behavior rather than from direct latency, throughput, or energy measurements on embedded hardware. Therefore, although the present findings strongly support the suitability of the proposed MelCNN for resource-constrained deployment, future work should include empirical profiling on platforms such as Raspberry Pi or ESP32 to quantify real-time inference latency, memory usage, and energy consumption under field-operational conditions [11], [20].

3.5 Dataset Bias and External Validity

Although the proposed MelCNN demonstrated strong predictive performance under the present evaluation protocol, the external validity of the findings must be interpreted within the boundaries of the curated dataset. The corpus used in this study consists of 600 audio clips distributed across only three animal classes, namely Cat, Dog, and Cow, with each class intentionally balanced to support controlled comparative evaluation. This design is methodologically advantageous for assessing architecture behavior under limited-data conditions, but it also narrows the representational scope of the study. In particular, the current dataset does not capture the broader taxonomic, behavioral, and ecological diversity that characterizes real-world bioacoustic monitoring scenarios, where vocalizations may vary substantially across species, recording distance, habitat, season, and background acoustic context. Therefore, the present results should be interpreted as evidence of strong performance within a controlled three-class setting rather than as proof of universal generalization across heterogeneous bioacoustic environments [12], [13], [20].

A second source of potential bias arises from dataset composition and acquisition conditions. As described in Section 2.1, the dataset combines publicly available audio from Kaggle repositories with self-recorded smartphone audio collected at close range to the target animals. The public audio was taken from Kaggle repositories relevant to the three target classes and then selected and standardized according to the duration and quality-control criteria applied in this study. This strategy was appropriate for constructing a compact yet acoustically meaningful corpus, but it may also introduce source-dependent regularities related to microphone characteristics, recording distance, environmental context, and signal cleanliness. Close-range recordings, for example, may over-represent clearer vocal events and under-represent the more degraded or partially occluded signals that commonly occur in practical field deployment. Similarly, mixing clips from heterogeneous sources can unintentionally embed acquisition-specific artifacts that are not strictly related to the underlying animal vocalization itself. For this reason, the reported results should be understood as reflecting performance under a carefully standardized and balanced evaluation regime rather than under fully uncontrolled real-world soundscape conditions [20].

At the same time, these constraints do not invalidate the contribution of the present study. Rather, they clarify its intended scope. The objective of this work is not to claim exhaustive ecological generalization, but to determine whether a compact domain-specific architecture can remain accurate, stable, and label-efficient under a realistic small-data regime. From that perspective, the use of balanced class composition, standardized preprocessing, stratified cross-validation, robustness analysis, and label-efficiency evaluation provides meaningful methodological support for the internal validity of the findings, even though broader external validation remains necessary. Accordingly, the present study should be interpreted as a controlled and scientifically grounded step toward low-complexity animal sound classification, while future work should extend the evaluation to larger and more acoustically diverse public datasets in order to test the robustness of these conclusions under wider deployment conditions [5], [9], [20].

3.6 Limitations and Future Work

Several limitations of the present study should be acknowledged in order to position its contribution accurately. First, the generalizability of the proposed model is constrained by the three-class evaluation corpus. While the architecture demonstrates high internal stability, validating its performance against unconstrained and larger public bioacoustic repositories is required to confirm external validity. This cross-dataset evaluation is recognized as a necessary step and is designated for future research. In other words, the current results are best understood as establishing internal validity and practical promise within a restricted evaluation regime rather than as exhausting the full range of real-world ecological variability [12], [13], [20].

Second, the practical suitability of the proposed model for edge-oriented deployment is supported in this study through indirect but meaningful indicators, including the compact parameter scale of approximately 0.39 million, sustained performance above 90% under 12 kHz resampling, and strong label-efficiency behavior at reduced training-set sizes. These findings are important because they demonstrate that the architecture remains structurally compatible with resource-constrained monitoring scenarios. However, computational practicality in the present work is inferred from architectural compactness and controlled robustness behavior rather than verified through direct on-device measurement. Therefore, no explicit claims are made here regarding real-time inference latency, peak memory usage, throughput, or energy consumption on embedded platforms. This boundary is important, because empirical deployment readiness ultimately requires hardware-specific profiling rather than architectural reasoning alone [5], [9], [11], [20].

Third, although the comparative study has been strengthened by the inclusion of three additional modern lightweight CNN baselines, the current benchmark still represents a controlled comparison within one problem formulation, namely short-duration three-class animal sound classification using standardized Log-Mel inputs. Accordingly, the present findings should be interpreted as evidence that the proposed MelCNN offers a superior accuracy-efficiency trade-off within this setting, rather than as a universal ranking of lightweight architectures for all audio classification tasks. Future work should therefore proceed along two complementary directions. On the one hand, broader external validation is needed using larger and more acoustically diverse public bioacoustic datasets. On the other hand, direct embedded-device benchmarking should be conducted on platforms such as Raspberry Pi or ESP32 in order to quantify latency, memory usage, and energy efficiency under realistic operating conditions. These extensions would allow the practical implications suggested by the present results to be tested more directly and under substantially wider deployment scenarios [11], [20].

4. Conclusions

This study demonstrates that a purpose-built Lightweight Convolutional Neural Network can provide a strong and practically meaningful balance between predictive performance, parameter efficiency, and robustness for animal sound classification under limited-data conditions. Using a balanced dataset of 600 clips from three target classes and Log-Mel spectrogram inputs, the proposed model achieved 97.00% ± 2.01% Accuracy and 97.01% ± 2.00% Macro-F1 across 5-fold stratified cross-validation while retaining a compact parameter scale of approximately 0.39 million. In the expanded comparative evaluation, the proposed MelCNN remained superior not only to the original transfer learning baselines based on YAMNet embeddings, but also to three additional modern lightweight CNN baselines, namely MobileNetV3-Small, ShuffleNetV2-1.0x, and EfficientNet-Lite0-style, indicating that for this controlled three-class bioacoustic task, a domain-specific compact architecture can offer a more favorable accuracy-efficiency trade-off than both generic pre-trained embedding pipelines and more general-purpose lightweight backbones. Beyond classification accuracy, the model also showed meaningful resilience under constrained conditions, maintaining performance above 90% at 12 kHz and preserving approximately 90.5% Macro-F1 when trained with only 30% of the available labeled data. These findings directly support the central objective of the study, namely to determine whether a lightweight end-to-end architecture can remain accurate, stable, and label-efficient in a realistic small-data setting relevant to resource-constrained monitoring applications. At the same time, the contribution of this work should be interpreted within the scope of a controlled evaluation regime rather than as a claim of universal bioacoustic generalization or direct embedded-device validation. Accordingly, future work should extend the analysis to larger and more acoustically diverse public datasets and complement the present results with direct profiling on embedded platforms to quantify latency, memory usage, and energy efficiency under practical deployment conditions.

Acknowledgements

The authors declare no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] C. A. Quinn *et al.*, "Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data," *Ecol. Indic.*, vol. 138, p. 108831, May 2022, doi: https://doi.org/10.1016/j.ecolind.2022.108831

[2] F. Sessa *et al.*, "Advancing Diagnostic Tools in Forensic Science: The Role of Artificial Intelligence in Gunshot Wound Investigation—A Systematic Review," *Forensic Sciences*, vol. 5, no. 3, p. 30, Sep. 2025, doi: https://doi.org/10.3390/forensicsci5030030

[3] J. Zurn, W. Burgard, and A. Valada, "Self-Supervised Visual Terrain Classification from Unsupervised Acoustic Feature Learning," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 466–481, Apr. 2021, doi: https://doi.org/10.1109/TRO.2020.3031214

[4] B. Swaminathan, M. Jagadeesh, and S. Vairavasundaram, "Multi-label classification for acoustic bird species detection using transfer learning approach," *Ecol. Inform.*, vol. 80, p. 102471, May 2024, doi: https://doi.org/10.1016/j.ecoinf.2024.102471

[5] H. Xu, Y. Tian, H. Ren, and X. Liu, "A Lightweight Channel and Time Attention Enhanced 1D CNN Model for Environmental Sound Classification," *Expert Syst. Appl.*, vol. 249, p. 123768, Sep. 2024, doi: https://doi.org/10.1016/j.eswa.2024.123768

[6] M. K. H. Kanchon, M. Sadman, K. F. Nabila, R. Tarannum, and R. Khan, "Enhancing personalized learning: AI-driven identification of learning styles and content modification strategies," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 269–278, Jan. 2024, doi: https://doi.org/10.1016/j.ijcce.2024.06.002

[7] B. Tiple and M. Patwardhan, "Multi-label emotion recognition from Indian classical music using gradient descent SNN model," *Multimed. Tools Appl.*, vol. 81, no. 6, pp. 8853–8870, Mar. 2022, doi: https://doi.org/10.1007/s11042-022-11975-4

[8] Y. Chen *et al.*, "Effective Audio Classification Network Based on Paired Inverse Pyramid Structure and Dense MLP Block," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, 2023, pp. 70–84, doi: https://doi.org/10.1007/978-981-99-4742-3_6

[9] Z. Fang, B. Yin, Z. Du, and X. Huang, "Fast environmental sound classification based on resource adaptive convolutional neural network," *Sci. Rep.*, vol. 12, no. 1, p. 10382, Dec. 2022, doi: https://doi.org/10.1038/s41598-022-10382-x

[10] J. Park, T. Yoo, S. Lee, and T. Kim, "Urban Noise Analysis and Emergency Detection System using Lightweight End-to-End Convolutional Neural Network," *International Journal of Computers, Communications and Control*, vol. 18, no. 5, p. 5814, Oct. 2023, doi: https://doi.org/10.15837/ijccc.2023.5.5814

[11] A. H. Z. Ilahi, A. Irwansyah, and H. Oktavianto, "Comparative Study of CNN Architectures for Real-Time Audio-Based Car Accident Detection on Edge Devices," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 3, p. 1310, May 2025, doi: https://doi.org/10.62527/joiv.9.3.2985

[12] A. Khodadadi, S. Molaei, M. Teimouri, and H. Zare, "Classification of audio codecs with variable bit-rates using deep-learning methods," *Digital Signal Processing: A Review Journal*, vol. 110, p. 102952, Mar. 2021, doi: https://doi.org/10.1016/j.dsp.2020.102952

[13] S. Anand, N. K. Devulapally, S. Das Bhattacharjee, and J. Yuan, "Multi-label Emotion Analysis in Conversation via Multimodal Knowledge Distillation," in *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2023, pp. 6090–6100, doi: https://doi.org/10.1145/3581783.3612517

[14] A. M. Tripathi and O. J. Pandey, "Divide and Distill: New Outlooks on Knowledge Distillation for Environmental Sound Classification," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1100–1113, 2023, doi: https://doi.org/10.1109/TASLP.2023.3244507

[15] C. Fan, J. Hu, and J. Huang, "Few-Shot Multi-Agent Perception," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 1712–1720, doi: https://doi.org/10.1145/3474085.3475315

[16] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "slimIPL: Language-model-free iterative pseudo-labeling," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2021, pp. 2118–2122, doi: https://doi.org/10.21437/Interspeech.2021-740

[17] J. Liang *et al.*, "Adapting Language-Audio Models as Few-Shot Audio Learners," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2023, pp. 276–280, doi: https://doi.org/10.21437/Interspeech.2023-1082

[18] K. M. Hasib, A. Tanzim, J. Shin, K. O. Faruk, J. Al Mahmud, and M. F. Mridha, "BMNet-5: A Novel Approach of Neural Network to Classify the Genre of Bengali Music Based on Audio Features," *IEEE Access*, vol. 10, pp. 108545–108563, 2022, doi: https://doi.org/10.1109/ACCESS.2022.3213818

[19] Y. Wang, C. Liu, H. Wu, Q. Sui, C. Yang, and W. Gui, "Revolutionizing Flotation Process Working Condition Identification Based on Froth Audio," *IEEE Trans. Instrum. Meas.*, vol. 72, p. 3315388, 2023, doi: https://doi.org/10.1109/TIM.2023.3315388

[20] J. Segura-Garcia, S. Sturley, M. Arevalillo-Herraez, J. M. Alcaraz-Calero, S. Felici-Castell, and E. A. Navarro-Camba, "5G AI-IoT System for Bird Species Monitoring and Song Classification," *Sensors*, vol. 24, no. 11, p. 3687, Jun. 2024, doi: https://doi.org/10.3390/s24113687

[21] K. Shaheed *et al.*, "EfficientRMT-Net—An Efficient ResNet-50 and Vision Transformers Approach for Classifying Potato Plant Leaf Diseases," *Sensors*, vol. 23, no. 23, p. 9516, Dec. 2023, doi: https://doi.org/10.3390/s23239516

[22] H. Bakır, A. N. Çayır, and T. S. Navruz, "A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification," *Multimed. Tools Appl.*, vol. 83, no. 6, pp. 17601–17628, Feb. 2024, doi: https://doi.org/10.1007/s11042-023-16200-4

[23] N. Zacarias-Morales, J. A. Hernández-Nolasco, and P. Pancardo, "Full single-type deep learning models with multihead attention for speech enhancement," *Applied Intelligence*, vol. 53, no. 17, pp. 20561–20576, Sep. 2023, doi: https://doi.org/10.1007/s10489-023-04571-y

[24] W. Wang, Q. Li, J. Xie, N. Hu, Z. Wang, and N. Zhang, "Research on emotional semantic retrieval of attention mechanism oriented to audio-visual synesthesia," *Neurocomputing*, vol. 519, pp. 194–204, Jan. 2023, doi: https://doi.org/10.1016/j.neucom.2022.11.036

[25] Z. Zhao, E. B. A. Bakar, N. B. A. Razak, and M. N. Akhtar, "Corrosion image classification method based on EfficientNetV2," *Heliyon*, vol. 10, no. 17, p. e36754, Sep. 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e36754

[26] V. Naresh, G. Yogeswararao, B. Nageswararao Naik, R. Malmathanraj, and P. Palanisamy, "Empirical Analysis of Squeeze and Excitation-Based Densely Connected CNN for Chili Leaf Disease Identification," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 4, pp. 1681–1692, Apr. 2024, doi: https://doi.org/10.1109/TAI.2024.3364126

[27] K. Filus and J. Domańska, "Global Entropy Pooling layer for Convolutional Neural Networks," *Neurocomputing*, vol. 555, p. 126615, Oct. 2023, doi: https://doi.org/10.1016/j.neucom.2023.126615

[28] T. B. Sridharan and D. P. S. Sharagaharajan Akilashri, "Multimodal learning analytics for students behavior prediction using multi-scale dilated deep temporal convolution network with improved chameleon Swarm algorithm," *Expert Syst. Appl.*, vol. 286, p. 128113, Aug. 2025, doi: https://doi.org/10.1016/j.eswa.2025.128113