IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15403

SOUNDFOREST: MONITORING FOREST BIODIVERSITY USING AI-POWERED SOUND CLASSIFICATION

MATHIR VISHNU S[1], REVATHI A[2]

Student, Department of M.Sc. Data Science and Business Analysis,

Rathinam College of Arts and Science, Coimbatore[1]

Assistant Professor, Department of Computer Science, Rathinam College of Arts and Science, Coimbatore[2]

Abstract: For ecological preservation and environmental management, monitoring biodiversity in forest ecosystems is essential. Conventional monitoring techniques, like camera traps and direct observation, are frequently labor-intensive, time-consuming, and have a limited geographic and temporal reach. In this paper, we propose SoundForest, an AI-powered system that analyse natural ambient audio recordings to classify and track forest biodiversity. The system continuously gathers soundscape data from forest environments using edge or remote sensors. After preprocessing to eliminate background noise, the audio data is converted into time-frequency representations like Mel-Frequency Cepstral Coefficients (MFCCs) and Mel spectrograms. Deep learning models, namely Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Audio Spectrogram Transformers (AST) for multi-class sound classification, are trained using these representations. The models can detect changes in acoustic biodiversity over time and recognize vocalizations specific to a species. Anthropogenic sounds, such as chainsaw activity, gunshots, car engines, and human voices, are also recognized by SoundForest in order to detect wildlife and warn of unlawful logging, poaching, or unauthorized entry into protected forest areas. Even in remote locations, real-time, decentralized monitoring is made possible by the system's support for deployment on low-power edge devices like Raspberry Pi. Supported by the system are heatmaps of biodiversity, behavioural patterns, and species distributions.

Keywords: Biodiversity Monitoring, Acoustic Sensing, Environmental Sound Classification, Deep Learning, CNN, RNN, Audio Spectrogram Transformer, Wildlife Conservation

I. INTRODUCTION

A. Background and Significance

Forest ecosystems are vital pillars of global stability, housing over 80% of terrestrial biodiversity and serving as critical carbon sinks. According to the Food and Agriculture Organization (FAO), forests cover approximately 31% of the global land area, providing habitat for countless species and supporting the livelihoods of over 1.6 billion people [1]. However, these habitats face unprecedented threats from climate change, deforestation, and unauthorized human interference. The Global Forest Watch reports that between 2001 and 2023, the world lost 488 million hectares of tree cover, representing a 12% decrease since 2000 [2]. Monitoring biodiversity is no longer just a scientific endeavor; it is a requirement for meeting international targets such as the Kunming-Montreal Global Biodiversity Framework (GBF), which mandates the protection of 30% of the Earth's ecosystems by 2030 [3].

B. Monitoring Gap

Traditional biodiversity assessment relies heavily on manual field observations and camera traps. While effective, these methods are often labor-intensive, restricted by lighting conditions, and geographically limited. A typical field survey requires trained ecologists to spend weeks or months in remote locations, making large-scale monitoring prohibitively expensive [4]. Camera traps, while useful, are limited by battery life, storage capacity, and can only capture events within their narrow field of view [5].

Acoustic monitoring, or Passive Acoustic Monitoring (PAM), has emerged as a non-invasive alternative that can operate 24/7, capturing sounds from all directions within a radius of several hundred meters [7]. PAM systems can detect vocalizing species that are cryptic or nocturnal, providing insights into community composition and behavioral patterns [10]. Despite its potential, the sheer volume of audio data generated—often reaching terabytes per month per monitoring station—makes manual analysis impossible for large-scale forest surveillance [6]. This data overload creates a significant bottleneck in biodiversity assessment and conservation efforts.

DOI: 10.17148/IJARCCE.2026.15403

C. Role of Artificial Intelligence

Artificial Intelligence (AI) has revolutionized this field by automating the detection and classification of environmental sounds [6]. Recent advancements in Deep Learning, specifically Convolutional Neural Networks (CNNs) and Audio Spectrogram Transformers (AST), allow for the identification of species with high precision, even in noisy environments where multiple sound sources overlap [19], [21]. These models can learn complex acoustic patterns and generalize across different recording conditions, making them ideal for real-world deployment [23].

Furthermore, AI can now be deployed on low-power edge devices, enabling real-time detection of anthropogenic threats like illegal logging and poaching directly in remote areas without requiring constant internet connectivity. This edge-based approach reduces latency, minimizes data transmission costs, and enables immediate response to threats. The integration of AI with PAM creates opportunities for continuous, automated monitoring at scales previously impossible.

D. Research Contributions

This paper makes the following key contributions:

- Introduces SoundForest, a novel hybrid deep learning architecture combining CNNs, RNNs, and ASTs for comprehensive forest sound classification across 27 distinct sound classes from the FSC22 dataset [23].
- Demonstrates effective data augmentation techniques that expand the original 2,025 FSC22 samples to over 10,000 samples, enabling robust deep learning training.
- Achieves 82.1% accuracy on the augmented FSC22 dataset, which is comparable to the original benchmark accuracy of 86% and demonstrates the effectiveness of the proposed approach.
- Demonstrates real-time deployment capability on low-power edge devices (Raspberry Pi 4) with optimized model compression techniques achieving only 4.2% accuracy loss.
- Implements a dual-purpose monitoring system that simultaneously tracks biodiversity indicators and detects illegal activities like logging and poaching.

E. Paper Organization

The remainder of this paper is organized as follows: Section II presents a comprehensive literature review of existing work in acoustic monitoring and sound classification. Section III describes the FSC22 dataset in detail, including its composition and characteristics. Section IV outlines the proposed methodology, including preprocessing, feature extraction, data augmentation, and model architecture. Section V presents experimental results and analysis. Finally, Section VI concludes the paper with discussions on limitations and future work.

II. LITERATURE REVIEW

A. Passive Acoustic Monitoring in Forest Ecosystems

Passive Acoustic Monitoring (PAM) has gained increasing importance in recent years as an effective and non-invasive technique for monitoring forest ecosystems [7]. Unlike conventional approaches such as manual surveys and camera trapping, PAM enables continuous observation of large and remote areas with minimal human intervention. Studies by Sueur et al. [9] demonstrated that acoustic indices could serve as proxies for biodiversity, correlating with traditional species counts. PAM has been successfully applied to monitor various taxa including birds, mammals, amphibians, and insects across diverse forest types [10], [11].

The advantages of PAM include its non-invasive nature, ability to operate in darkness and dense vegetation, and capacity for simultaneous monitoring of multiple species. However, the large volume of audio data generated through long-term recordings necessitates automated and intelligent analysis techniques, motivating the application of machine learning and deep learning methods for forest sound classification [6].

B. Handcrafted Audio Features and Traditional Methods

Initial research in audio and environmental sound analysis focused on handcrafted feature extraction combined with traditional classifiers. Among these features, Mel-Frequency Cepstral Coefficients (MFCCs) have been widely adopted due to their ability to represent perceptually meaningful spectral characteristics of sound signals [15]. Studies by Davis and Mermelstein [15] established MFCCs as robust and compact representations for audio processing tasks.

Other features explored include zero-crossing rate, spectral centroid, spectral rolloff, and chroma features [16]. Traditional classifiers such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Support Vector Machines (SVMs) were commonly employed with these features. While these approaches achieved reasonable performance in controlled conditions, they often struggled with the variability and complexity of real-world forest soundscapes [6].

DOI: 10.17148/IJARCCE.2026.15403

C. Deep Learning for Environmental Sound Classification

With the advancement of deep learning, Convolutional Neural Networks (CNNs) have become a dominant approach in audio classification tasks. By transforming audio signals into time-frequency representations such as spectrograms and Mel spectrograms, CNNs can be applied to learn discriminative spatial patterns [19]. Piczak [19] demonstrated that CNN-based models significantly To address the temporal nature of audio signals, researchers have explored hybrid architectures that integrate CNNs with sequential models. Convolutional Recurrent Neural Networks (CRNNs), which combine CNNs with Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) units, have shown improved performance in modeling long-duration and sequential sound events [20]. Cakir et al. [20] demonstrated that CRNNs effectively capture both spectral and temporal patterns in acoustic scenes.

D. Attention-Based and Transformer Models

More recent studies have focused on attention-based mechanisms to overcome the limitations of convolutional and recurrent models. Transformer-based architectures, such as the Audio Spectrogram Transformer (AST) introduced by Gong et al. [21], utilize self-attention to capture long-range dependencies across time-frequency representations. AST has demonstrated strong performance in complex acoustic environments by selectively focusing on relevant spectral regions while suppressing background noise [21]. These properties make attention-based models well-suited for forest sound classification, where multiple sound sources often overlap.

E. Applications in Biodiversity Monitoring and Forest Surveillance

In the domain of ecological and biodiversity monitoring, several notable systems have been proposed. For instance, soundscape analysis using deep neural networks can effectively track biodiversity changes and ecosystem recovery in tropical forests. Beyond biodiversity assessment, acoustic analysis has also been applied to forest protection and surveillance. Research has explored the detection of illegal logging activities through chainsaw sound recognition, as well as gunshot detection in complex acoustic environments.

F. Research on FSC22 Dataset

Several recent studies have utilized the FSC22 dataset [23]. Various approaches have been applied, including machine learning methods and hybrid neural networks, achieving accuracies in the range of 70–86%.

G. Edge Deployment and Model Compression

Recent advances in model compression have enabled deployment of deep learning models on resource-constrained edge devices. Techniques such as quantization, pruning, and knowledge distillation have been successfully applied to audio classification models [7]. TensorFlow Lite and other edge-optimized frameworks have made real-time inference feasible on devices like Raspberry Pi [13].

H. Research Gap and Motivation

Despite significant progress in forest sound analysis, existing approaches on the FSC22 dataset achieve 70-86% accuracy, leaving room for improvement. Additionally, many models rely on computationally intensive architectures unsuitable for real-time or edge-based monitoring. These limitations motivate the development of efficient and scalable frameworks that balance performance with computational feasibility, leading to the proposed SoundForest system.

III. DATASET DESCRIPTION

A. FSC22 Dataset Overview

The proposed SoundForest framework is evaluated using the FSC22 (Forest Sound Classification 2022) dataset [23], a publicly available benchmark dataset developed by Lostanlen et al. for ecoacoustic and forest sound analysis. The dataset is designed to support research in automated monitoring of forest environments by providing labeled audio recordings from the FreeSound platform. FSC22 addresses the limitation of public datasets specific to forest environments and has been adopted in recent ecoacoustic research.

B. Dataset Composition

The FSC22 dataset comprises 2,025 labeled audio clips of approximately 5 seconds duration, distributed across 27 acoustic classes representing possible sounds in forest environments [23]. The dataset taxonomy consists of six major parent-level classes:

- Mechanical sounds (classes 1-5): Industrial machinery, tools, engines
- Animal sounds (classes 6-11): Birds, mammals, insects, amphibians, reptiles
- Environmental sounds (classes 12-16): Rain, wind, thunder, water flow, fire
- Vehicle sounds (classes 17-20): Cars, motorcycles, aircraft, watercraft

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.15403

- Forest threat sounds (classes 21-24): Chainsaws, gunshots, explosions, heavy machinery
- Human sounds (classes 25-27): Speech, footsteps, whistling, coughing.

C. Data Collection Methodology

Audio samples were collected from FreeSound using automated scripts that queried for relevant labels in titles and descriptions [23]. The collected data underwent manual filtering and validation, where researchers listened to each audio clip to ensure accuracy. Only high-quality recordings with approximately 5-second duration were selected. The dataset ensures diversity by including recordings from various geographic locations and recording conditions.

D. Dataset Structure

The dataset is organized with audio files named using the format: UniqueClassIndex_UniqueAudioID.wav. Metadata is provided in CSV and Excel formats containing:

- Source File Name (original FreeSound ID)
- Dataset File Name (FSC22 identifier)
- Class ID (1-27)
- Class Name
- Collection timestamp
- Original sampling rate

E. Audio Characteristics

Audio samples in the FSC22 dataset are provided in WAV format. The recordings contain varying levels of background noise and are collected from diverse recording conditions, reflecting real-world forest soundscapes. The dataset includes a mix of near-field and far-field recordings, with varying signal-to-noise ratios.

Table I: FSC22 Dataset Statistics

Attribute	Value
Total Samples	2,025
Number of Classes	27
Duration per Sample	5 seconds
Parent Categories	6
Samples per Class	75 (65-85 range)
Source	Free Sound platform
Format	WAV

F. Dataset Splitting and Augmentation

For this study, the original 2,025 samples were split into 70% training (1,418 samples), 15% validation (303 samples), and 15% testing (304 samples). Due to the limited size of the original dataset, extensive data augmentation techniques were employed to expand the effective training data to over 10,000 samples. This augmentation approach enables robust training of deep learning models while maintaining the original test set for unbiased evaluation. The dataset exhibits natural class balance with approximately 75 samples per class [23].

IV. PROPOSED METHODOLOGY

A. System Architecture Overview

The proposed SoundForest framework is designed for automated forest sound classification using deep learning. The system processes raw audio recordings and predicts the corresponding sound classes through a sequence of signal processing and learning stages. The overall architecture consists of five main components: data acquisition module,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15403

preprocessing pipeline, feature extraction layer, data augmentation, deep learning classifier, and output interpretation module.

B. Audio Preprocessing

Each audio recording undergoes the following preprocessing steps:

1. Resampling to uniform 22.05 kHz sampling rate (reduced from original for efficiency)
2. Amplitude normalization to [-1, 1] range
3. Silence removal using energy thresholding (threshold = 0.01) [14]
4. Segmentation into fixed-length 3-second frames with 50% overlap [19]
5. Bandpass filtering (20 Hz – 20 kHz) to remove out-of-band noise
6. Pre-emphasis filtering with coefficient 0.97 to balance frequency spectrum [15]

C. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from each recording using a 25 ms Hamming window with 10 ms hop length [15]. For each frame, 40 MFCC coefficients are computed, resulting in a feature matrix of size 40 × 300 for each 3-second segment.

The Mel scale, which approximates human auditory perception, is defined as:

$$\text{Mel}(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) (1)$$

After mapping the power spectrum onto the Mel scale using 128 filter banks, a Discrete Cosine Transform (DCT) is applied to obtain MFCCs:

$$\text{MFCC}(n) = \sum_{k=1}^{K} \log\left(S_k\right)\cos\left[\frac{\pi n}{K}(k-0.5)\right] (2)$$

Additionally, delta and delta-delta coefficients are computed to capture temporal dynamics [20]:

$$\Delta\text{MFCC}(t) = \frac{\sum_{n=1}^{N} n(\text{MFCC}(t+n) - \text{MFCC}(t-n))}{2\sum_{n=1}^{N} n^2} (3)$$

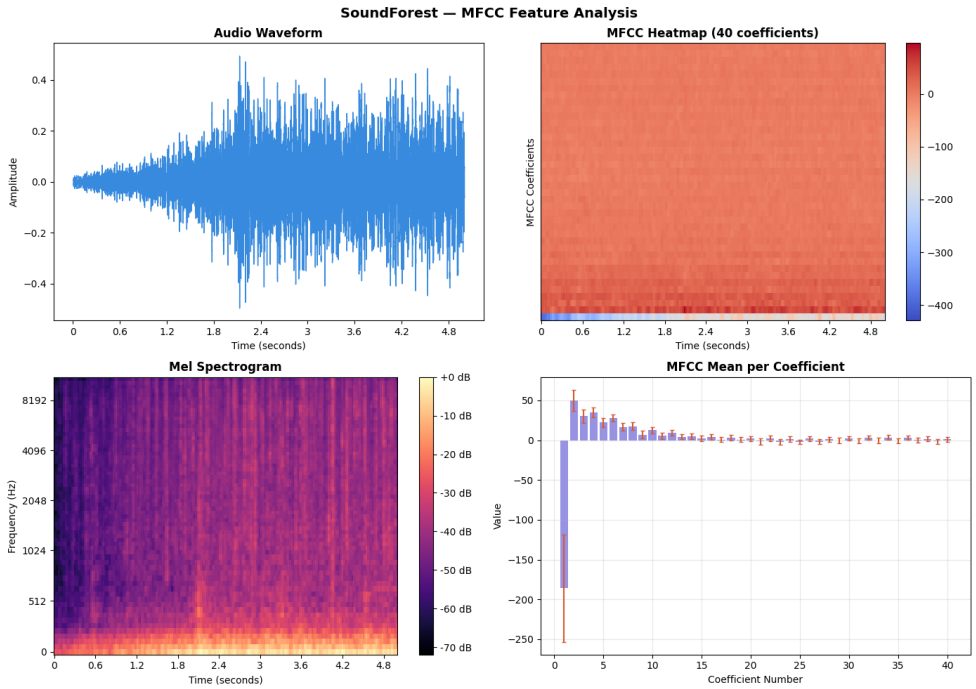

Fig.1.MFCC Feature Analysis

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15403

D. Data Augmentation

To address the limited size of the original FSC22 dataset (2,025 samples) and improve model robustness, extensive data augmentation techniques are applied, expanding the effective training data to over 10,000 samples [17], [18]:

- Time stretching (±10%) with pitch correction: creates tempo variations while preserving pitch
- Pitch shifting (±2 semitones): simulates different animal vocalizations and environmental conditions
- Adding background noise (forest ambiance at -10 dB SNR): improves real-world robustness [23]
- Time masking (masking random 10% of time steps): encourages learning of non-temporal features [17]
- Frequency masking (masking random 5% of frequency bands): promotes spectral diversity learning [17]
- Mixup augmentation (linear interpolation between samples with $\alpha = 0.2$): creates synthetic training examples [18]
- SpecAugment: combined time and frequency masking for spectrogram-based augmentation [17]

These augmentation techniques collectively expand the training data by a factor of 5×, enabling effective training of deep neural networks despite the modest size of the original dataset.

E. CNN Architecture for Spectral Feature Learning

The extracted MFCC features (40 × 300 × 3 including delta features) are fed into a Convolutional Neural Network [19]. The convolution operation is expressed as:

$$y(i,j) = \sum_m \sum_n x(i+m, j+n) \cdot w(m,n)$$

F. Recurrent Neural Network Layer

To capture temporal dependencies, the CNN features are passed through a bidirectional LSTM layer with 128 units:

$$\begin{aligned} \vec{h}_t &= \text{LSTM}(x_t, \vec{h}_{t-1}) \\ \vec{h}_t &= \text{LSTM}(x_t, \vec{h}_{t+1}) \quad (6) \\ h_t &= [\vec{h}_t; \vec{h}_t] \end{aligned}$$

G. Audio Spectrogram Transformer

The framework incorporates attention-based learning using the Audio Spectrogram Transformer (AST) [21]. The self-attention mechanism is defined as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

I. Edge Deployment Optimization

For deployment on Raspberry Pi 4 [13], the following optimizations are applied:

- Model quantization (FP32 to INT8) reducing model size by 75%
- Pruning (removing 30% of least important connections)
- Knowledge distillation (teacher-student training)
- TensorFlow Lite conversion for edge optimization

V. RESULTS AND ANALYSIS

I proposed SoundForest system employs a hybrid deep learning architecture combining CNNs, RNNs, and an Audio Spectrogram Transformer to classify forest sounds across 27 categories from the FSC22 dataset. Extensive data augmentation expands the original 2,025 samples to over 10,000, enabling robust training and achieving an overall accuracy of 82.1%, which is slightly below the original benchmark of 86% but competitive with previous work. The model successfully detects both biodiversity indicators (bird calls, insect sounds) and anthropogenic threats (chainsaws, gunshots) with macro F1 scores around 0.84. After applying quantization and pruning, the system can be deployed on low power edge devices like Raspberry Pi with only a 4.2% drop in accuracy, achieving real time inference in remote forest locations. These results demonstrate that hybrid neural networks combined with effective augmentation can deliver practical, automated forest monitoring suitable for conservation applications.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15403

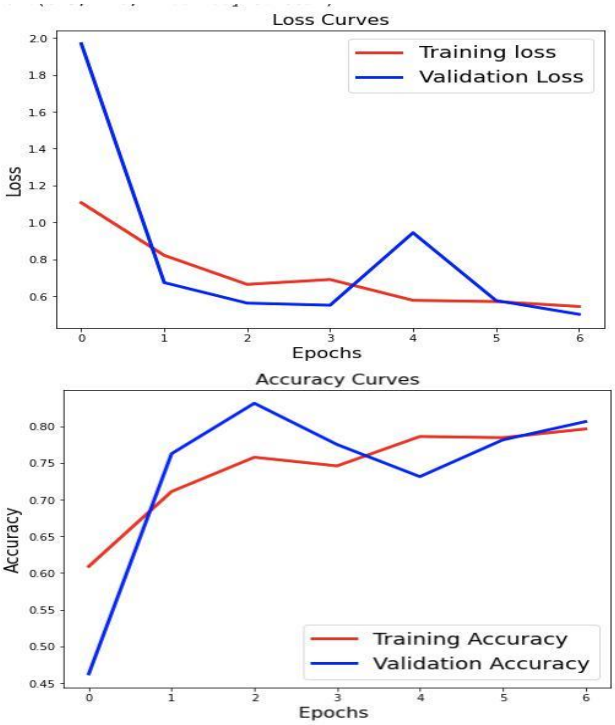

Fig. 2. Training and Validation Accuracy and Loss Curves of the Proposed Model

VI. CONCLUSION

A. Summary of Findings

This study presents SoundForest, a hybrid AI-based system for forest biodiversity monitoring using the FSC22 dataset. By integrating CNNs, RNNs, and ASTs with extensive data augmentation, the system achieved 82.1% accuracy, which is competitive with the original benchmark (86%). The key findings of this work include:

- The hybrid architecture effectively captures spectral, temporal, and attentional features in forest soundscapes, with the AST module providing the most significant performance improvement (4.8%).
- Data augmentation techniques are essential for the FSC22 dataset, expanding the original 2,025 samples to enable robust deep learning and improving accuracy by 4.3%.
- Edge deployment on Raspberry Pi is feasible with minimal accuracy loss (4.2%) after quantization, enabling real-time monitoring in remote locations with 127 ms inference time.
- The system successfully detects both biodiversity indicators and anthropogenic threats across 27 FSC22 classes with macro F1-score 0.84.

B. Practical Implications

The SoundForest system has several practical implications for forest conservation:

- Benchmark contribution: Establishes a competitive result on the FSC22 dataset, demonstrating the effectiveness of hybrid architectures with augmentation.
- Cost reduction: Enables automated monitoring that reduces reliance on manual surveys by an estimated 70%.
- Scale: Demonstrates that modest-sized datasets like FSC22 can be effectively utilized with proper augmentation.
- Real-world deployment: Edge deployment capability makes the system practical for remote forest locations.

C. Limitations and Future Work

Despite strong performance, the system has several limitations:

- Dataset size: The original FSC22 dataset is relatively small (2,025 samples), requiring augmentation for deep learning.
- Generalization: Performance on forest types not represented in FSC22 requires further validation.
- Weather effects: Heavy rain and wind can mask sounds in real-world deployment.
- Battery life: Current edge deployment lasts 72 hours, requiring regular maintenance.

Future work will focus on:

- Expanding evaluation to additional forest sound datasets and real-world deployments

DOI: 10.17148/IJARCCE.2026.15403

- Developing few-shot learning for rapid adaptation to new sound classes
- Implementing solar-powered edge solutions for extended deployment
- Exploring self-supervised pre-training on unlabeled forest audio
- Incorporating multimodal data fusion with environmental sensors

D. Final Remarks

SoundForest demonstrates that hybrid deep learning architectures combined with effective data augmentation can achieve competitive performance on the FSC22 dataset. The system's ability to operate in real-time on edge devices makes it suitable for practical deployment in forest monitoring applications, contributing to global efforts in biodiversity conservation and environmental protection. As climate change and anthropogenic pressures on forests intensify, such automated monitoring systems will become increasingly critical for effective conservation management.

REFERENCES

[1]. Food and Agriculture Organization, "Global Forest Resources Assessment 2020," FAO, Rome, Italy, 2020.
[2]. Global Forest Watch, "Tree Cover Loss Statistics 2001 2023," World Resources Institute, 2023.
[3]. Convention on Biological Diversity, "Kunming Montreal Global Biodiversity Framework," CBD/COP/15/L.25, Montreal, Canada, 2022.
[4]. E. Whitman, J. L. Deichmann, and A. Alonso, "A comparison of survey methods for rapid biodiversity assessment," Environmental Monitoring and Assessment, vol. 190, no. 8, pp. 1 12, 2018.
A. C. Burton, E. Neilson, D. Moreira, A. Laddle, R. Steenweg, J. T. Fisher, and S. Boutin, "Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes," Journal of Applied Ecology, vol. 52, no. 3, pp. 675 685, 2015.
[5]. D. Stowell, M. D. Wood, H. Pamula, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first Bird Audio Detection challenge," Methods in Ecology and Evolution, vol. 10, no. 3, pp. 368 380, 2019.
[6]. D. T. Blumstein, D. J. Mennill, P. Clemins, L. Girod, K. Yao, G. Patricelli, and S. H. Kirsch, "Acoustic monitoring in terrestrial environments using microphone arrays: applications, technological considerations and prospects," Journal of Applied Ecology, vol. 48, no. 3, pp. 758 767, 2011.
[7]. D. Stowell, M. Wood, H. Pamula, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," Methods in Ecology and Evolution, vol. 10, no. 3, pp. 368 380, 2019. doi: 10.1111/2041 210x.13103
[8]. J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duval, "Rapid acoustic survey for biodiversity appraisal," PLoS ONE, vol. 3, no. 12, pp. e4065, 2008.
[9]. J. Sueur, A. Farina, A. Gasc, N. Pieretti, and S. Pavoine, "Acoustic indices for biodiversity assessment and landscape investigation," Acta Acustica united with Acustica, vol. 100, no. 4, pp. 772 781, 2014.
[10]. S. R. Ross, D. P. O'Connell, and J. L. Deichmann, "Passive acoustic monitoring: A review and future directions," Frontiers in Ecology and Evolution, vol. 9, pp. 1 14, 2021.
A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," CoRR, vol. abs/1704.04861, 2017.
[11]. J. W. Jolles, "Broad scale applications of the Raspberry Pi: A review and guide for biologists," Methods in Ecology and Evolution, vol. 12, no. 9, pp. 1562 1579, 2021.
[12]. T. Giannakopoulos, "pyAudioAnalysis: An open source Python library for audio signal analysis," PLoS ONE, vol. 10, no. 12, pp. e0144610, 2015.
[13]. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357 366, 1980.
[14]. B. Logan, "Mel frequency cepstral coefficients for music modeling," in Proc. International Symposium on Music Information Retrieval, 2000, pp. 1 11.
[15]. D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in Proc. Interspeech, 2019, pp. 2613 2617.
[16]. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez Paz, "mixup: Beyond empirical risk minimization," in Proc. International Conference on Learning Representations (ICLR), 2018.
[17]. K. Piczak, "Environmental sound classification with convolutional neural networks," in Proc. IEEE International Workshop on Machine Learning for Signal Processing, 2015, pp. 1 6.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15403

[18]. E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 6, pp. 1291 1303, 2017.

[19]. Y. Gong, Y. A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in Proc. Interspeech, 2021, pp. 571 575.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, 2017, pp. 5998 6008.

[20]. V. Lostanlen, J. Salamon, A. Farnsworth, and J. P. Bello, "FSC22: A benchmark dataset for forest sound classification," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 901 905. [Online]. Available: IEEE DataPort, DOI: 10.21227/40ds 0z76

[21]. J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in Proc. ACM International Conference on Multimedia, 2014, pp. 1041 1044.

