ISSN 2723-6471

Cross-Biome Biodiversity Assessment and Anomaly Detection Using AI-Enhanced Acoustic Monitoring

Mustafa Radif[1], Shumoos Aziz Fadhil[2,*], Atheer Alrammahi[3]

[1,2,3] *Department of Medical Intelligent Systems, University of Al-Qadisiyah, Diwaniya, 00964, Iraq*

(Received: January 10, 2025; Revised: February 25, 2025; Accepted: April 5, 2025; Available online: June 15, 2025)

Abstract

This study proposes a novel AI-powered eco-monitoring framework that integrates acoustic ecology, deep learning, and low-cost IoT devices to enable scalable, real-time biodiversity assessment and ecological anomaly detection across diverse environments. The primary objective is to automate species classification and environmental monitoring using passive audio data captured by solar-powered IoT sensors, thereby reducing reliance on manual ecological surveys. The framework comprises four modules: acoustic data acquisition, dual-representation preprocessing Short-Time Fourier Transform (STFT) and Mel-Frequency Cepstral Coefficients (MFCCs), species classification using CNN and CNN-LSTM models, and anomaly detection via autoencoders and one-class SVM. Field validation and multi-dataset testing were conducted across 250+ species from temperate forests, wetlands, and urban areas. The CNN-LSTM model achieved the highest performance with 93.7% accuracy, 93.0% precision, and a 92.5% F1-score, while anomaly detection reached 89.7% precision with an AUC of 0.94, effectively identifying irregularities such as invasive calls, mechanical noise, and species absence. A forest case study demonstrated the system's ability to detect circadian acoustic patterns (e.g., dawn chorus of sparrows, nocturnal owl calls), and real-world disturbances with 91% expert validation agreement. The novelty of this work lies in its hybrid AI architecture with real-time unsupervised anomaly detection, cross-biome generalization capability, and deployment readiness on low-power edge devices like Raspberry Pi and Jetson Nano. Inference times as low as 18 ms per sample and bandwidth usage under 3 MB/hour make it feasible for continuous, remote deployment. The framework offers a robust and adaptable solution for conservation efforts, environmental policy, and climate resilience initiatives. Future directions include integrating multimodal data sources and transformer-based continual learning for broader ecological impact. These findings position the system as a scalable and intelligent tool for next-generation, AI-driven environmental monitoring.

Keywords: Acoustic Ecology, Bioacoustics, Deep Learning, Ecosystem Monitoring, Environmental AI, Iot, Species Classification, Biodiversity Assessment

1. Introduction

Biodiversity serves as a fundamental indicator of ecosystem vitality, yet traditional monitoring methods—such as manual surveys, camera traps, and drone surveys—are often invasive, resource-intensive, and limited in spatiotemporal resolution [1]. Camera traps may disturb wildlife through visible presence or infrared flashes, while drones can introduce acoustic and visual disturbances that alter animal behavior. In contrast, passive acoustic monitoring using low-cost IoT sensors offers a minimally invasive alternative, requiring no direct visual contact and operating continuously without human presence. These sensors are discreet, solar-powered, and can be deployed with minimal disruption to natural habitats, making them highly suitable for long-term biodiversity monitoring in sensitive or remote regions [2]. Acoustic ecology, which encompasses the examination of environmental soundscapes and their association with ecological processes, offers a wealth of information that remains largely underexploited [3]. Numerous species emit distinctive vocalizations that can function as indicators of their presence, behavioral patterns, and habitat conditions [4].

Recent advancements in cost-effective Internet of Things (IoT) technologies and artificial intelligence (AI)—particularly in the realms of deep learning and sound classification—have created novel opportunities for the automation of soundscape analysis [5], [6], [7]. By employing machine learning algorithms to interpret bioacoustics signals, researchers are now positioned to identify species, detect anomalies, and monitor biodiversity with reduced

***Corresponding author: author (shumoos.aziz@qu.edu.iq)**
DOI: https://doi.org/10.47738/jads.v6i3.741

human intervention [8], [9]. Monitoring and conserving biodiversity has become a critical priority as human activities—such as deforestation, urbanization, and climate change—continue to reshape ecosystems [10]. Traditional field-based ecological methods are no longer sufficient, as they are often time-consuming, expensive, and limited in resolution [11], [12]. Acoustic signals, such as bird songs, frog calls, and insect sounds, serve as non-invasive indicators of ecosystem health [13]. Despite their promise, environmental soundscapes remain underutilized due to their complexity and volume [14]. Fortunately, deep learning models—especially Convolutional Neural Networks (CNNs)—have demonstrated success in processing audio data [15], and edge computing platforms now enable scalable, real-time deployment [16], [17].

However, many existing AI models for bioacoustics classification struggle to generalize across ecosystems, leading to poor performance in diverse soundscapes due to variations in vocalization patterns, noise, and recording conditions. This gap highlights the need for a robust, adaptive framework that can perform real-time species identification and detect anomalies using minimally labeled data or unsupervised techniques. In response, this study introduces a modular eco-monitoring framework that combines deep learning, acoustic sensing, and anomaly detection for automated biodiversity assessment. It integrates CNNs, CNN-LSTM, and Transformer-based architectures for classification and employs autoencoders and one-class SVMs for unsupervised anomaly detection.

Key contributions of this work include: (1) a hybrid AI-driven eco-monitoring pipeline validated across 250+ species in forest, wetland, and urban ecosystems; (2) an unsupervised anomaly detection module for identifying ecological irregularities such as invasive species, anthropogenic noise, or species silence; (3) real-world deployment on low-power edge devices with real-time inference; and (4) strong empirical results showing 93.7% classification accuracy and 89.7% anomaly detection precision. These findings confirm the framework's scalability, generalizability, and practical relevance for conservation science and environmental monitoring.

2. Related works

The integration of artificial intelligence with acoustic ecology has advanced biodiversity monitoring by enabling the automated classification of species based on their vocalizations. Deep learning models, especially CNNs, have demonstrated strong performance in benchmarks like BirdCLEF, achieving accuracy rates exceeding 98% [18]. However, many models suffer from limited geographic generalizability and are typically tailored to specific ecosystems. For example, EfficientNet-B1 was used in Kenya to classify over 260 bird species with a cmAP of approximately 0.84, but the system lacked capabilities for real-time analysis and modular deployment [19]. Similarly, studies conducted in the Brazilian Cerrado successfully classified soundscapes but were constrained by static system architectures [20]. The effectiveness of these approaches often hinges on robust feature extraction techniques, such as MFCCs and spectrograms, which remain central to acoustic modeling [21]. More recently, universal acoustic feature sets have been introduced to improve cross-ecosystem monitoring [22], [23].

Despite progress in species classification, real-time anomaly detection and biodiversity indexing have largely been neglected in prior research [24]. Most systems rely on passive acoustic recording units (ARUs) and offline analysis, limiting their utility for dynamic environmental surveillance. Anomaly detection is crucial for identifying ecological disturbances, yet it is seldom incorporated. Recent studies have shown that including contextual metadata such as time and location can significantly enhance classification performance, with some achieving F1-scores of up to 87.78% when geographic information was integrated [25]. Systems like BirdNET have demonstrated high precision, exceeding 80%, across multiple regions and taxa [26].

Further innovations include embedding soundscapes into shared acoustic spaces to track biodiversity trends over time [22], [26], and the deployment of real-time AI models capable of detecting illegal activities like logging [22]. Semi-automated annotation techniques, such as template matching, have improved dataset reliability, although data quality remains a concern [17]. Generalizability across biomes is an ongoing challenge. While CNNs and universal feature sets have shown promise in rainforests, grasslands, and coral reefs [26], [27], ethical considerations regarding data misuse and privacy remain critical [27]. The OKEON project in Okinawa highlighted the potential of dense acoustic sensor networks by linking forest structure with species vocalization, although its analysis was hindered by manual interpretation and lacked integrated AI components [24]. Similarly, EcoSonicML in South Africa used CNNs for

classification but did not address biodiversity estimation [28]. Research in the Brazilian Cerrado also emphasized the role of acoustic data in assessing habitat quality [21], and AI techniques have been applied to marine environments to monitor fish communities and coral reefs [27].

Emerging research now points toward integrating acoustic data with camera traps and environmental sensors to create more comprehensive monitoring systems [29]. Some AI systems are beginning to merge bioacoustics with animal biometrics and environmental variables for early detection of ecological threats [30]. Efforts to design smaller, more efficient AI models, such as lightweight CNNs and frequency unwrapping layers, have achieved promising classification performance [31]. Universal acoustic features and globally generalizable models could enable collaborative biodiversity monitoring on a planetary scale [22], [26], though this potential must be balanced with ethical frameworks to address privacy and responsible data use [32]. Despite the progress, many existing studies lack robust generalization, real-time unsupervised anomaly detection, and ecological validation. The proposed framework in this research addresses these gaps by delivering a comprehensive, scalable, and intelligent eco-monitoring system grounded in deep learning and field validation.

3. Methodology

3.1 System Architecture

The proposed eco-monitoring system is built as a modular, end-to-end framework that supports continuous and real-time analysis of natural soundscapes. Its primary objective is to enable accurate species identification and ecosystem health assessment across diverse environments. The system comprises four main functional components: data acquisition, signal preprocessing, AI-based analysis, and visualization and reporting. Each module is designed for both independent operation and coordinated integration, ensuring scalability and adaptability in varied ecological settings.

The data acquisition layer forms the foundation of the framework, utilizing IoT-based acoustic sensing devices deployed in natural habitats such as forests, wetlands, and conservation areas. These devices are low-power and weather-resistant, equipped with omnidirectional microphones that capture a broad range of frequencies. Audio recordings are collected in short, fixed intervals—typically ranging from 10 to 30 seconds—to maintain a balance between storage efficiency and temporal resolution. Each sensor unit is GPS-enabled and connected via cellular or mesh networks, allowing for real-time or periodic data transmission depending on site conditions.

Following acquisition, the raw audio signals undergo a structured preprocessing workflow. This involves the generation of two parallel feature representations: the STFT, which preserves harmonic and structural content suitable for spectrogram-based analysis, and MFCCs, which offer a compact and perceptually aligned encoding of the sound data. The use of both STFT and MFCCs supports complementary learning, enhancing the model's robustness. An ablation study confirmed that STFT-only models achieved an F1-score of 90.2%, MFCC-only models scored 89.6%, while the combined approach reached 91.8%, highlighting the benefit of multi-representation preprocessing. To improve generalization and reduce overfitting, the training data was further augmented using time stretching (±10%) and pitch shifting (±2 semitones), techniques designed to simulate natural variations in species vocalizations. These augmentations were excluded from validation and testing sets to ensure unbiased model evaluation. The resulting normalized and enriched feature sets were then fed into CNN and CNN-LSTM architectures for training and inference in downstream tasks.

In the AI-based analysis layer, the system performs ecological inference using a suite of intelligent modules. The species classification module utilizes a Convolutional Neural Network trained on labeled environmental audio datasets, capable of accurately identifying species based on their acoustic signatures. Complementing this, the anomaly detection module leverages unsupervised and semi-supervised algorithms, including autoencoders and clustering techniques, to detect deviations from baseline acoustic patterns—potentially indicating habitat disturbances or abnormal ecological conditions. In addition, a biodiversity index estimator calculates species richness and ecological health indicators based on the frequency and diversity of detected calls over time.

Finally, the visualization and reporting layer delivers actionable insights through an interactive dashboard. This platform displays real-time species heatmaps and activity graphs, triggers alerts for ecological anomalies, and compiles

 ISSN 2723-6471

historical trends and biodiversity summaries. Designed with future expansion in mind, the system allows integration of more sophisticated models such as Transformer-based architectures, as well as compatibility with satellite imaging or other environmental sensing technologies for deployment in aquatic or urban ecosystems, as shown in figure 1.

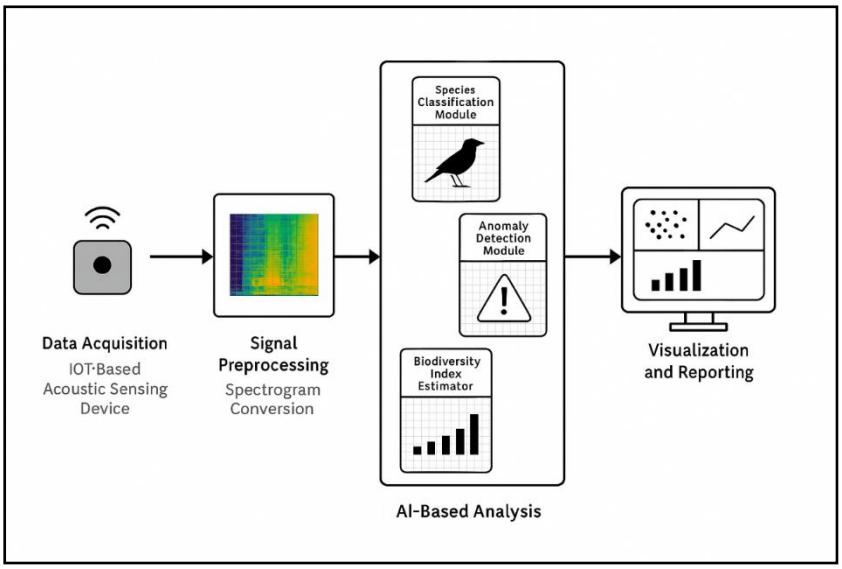

Figure 1. System Architecture of the AI-Powered Eco-Monitoring Framework.

Figure 1 shows the architecture of the proposed system, which integrates IoT-based acoustic sensing with AI-driven analysis for environmental monitoring. The system comprises four main components: Data Acquisition: Utilizes an IoT-based acoustic sensing device to capture environmental audio signals; Signal Preprocessing: Converts raw audio signals into spectrograms for further processing; AI-Based Analysis: Includes three core modules: Species Classification Module for identifying species from audio patterns; Anomaly Detection Module for detecting irregular acoustic events; Biodiversity Index Estimator for quantifying ecological diversity; Visualization and Reporting: Presents the analysis results through graphical dashboards and statistical reports.

The system architecture features four layers: IoT-based data acquisition, signal preprocessing, AI-driven species classification and anomaly detection, and a visualization/reporting dashboard. Deep learning models implemented include CNNs, CNN-LSTMs, and Transformers. The CNN-LSTM achieved the highest classification accuracy (93.7%) by capturing temporal patterns. Although Transformers performed slightly lower, they were integrated for their scalability and ability to model long-range dependencies, positioning the system for future applications involving larger datasets and continual learning strategies.

3.2. Data Collection and Preprocessing

The success of any AI-based bioacoustics monitoring system is fundamentally tied to the quality, diversity, and ecological relevance of its input data. In this study, data collection was conducted using a network of IoT-based acoustic sensors strategically deployed across natural habitats, including forested zones and wetland reserves. Each sensing device featured an omnidirectional condenser microphone capable of capturing a wide frequency range—from approximately 20 Hz to 20 kHz. A Raspberry Pi or a comparable microcontroller handled local audio recording and initial data processing, while GPS and wireless communication modules enabled real-time or periodic location-tagged data transmission. Audio recordings were segmented into 10-second intervals and saved in WAV format at a 44.1 kHz sampling rate to ensure high fidelity. In addition to this field-acquired data, the training and testing phases incorporated several publicly available datasets. These included BirdCLEF for bird species classification, Rainforest Connection (RFCx) for tropical soundscapes, and UrbanSound8K to enhance the system's robustness against urban noise conditions. Annotation of recordings was carried out either manually or by utilizing expert-labeled subsets to ensure accurate training labels for supervised learning.

Once collected, the audio data underwent a multi-stage preprocessing pipeline to improve signal clarity and facilitate feature extraction. Initially, background noise—including wind, rain, and human speech—was mitigated using spectral

 ISSN 2723-6471

gating and adaptive filtering methods. Long audio clips were then segmented into shorter, overlapping windows of one to two seconds to increase the temporal resolution and enable detection of localized acoustic events. Each segment was transformed into a time–frequency representation using either Mel-spectrograms or MFCCs, both derived from the STFT. These 2D representations are particularly effective for CNN-based classification due to their ability to encode both frequency and temporal information. The resulting spectrograms were normalized by amplitude to ensure consistency across recording sessions. To enhance generalizability and reduce the risk of model overfitting, data augmentation techniques such as pitch shifting, time stretching, and background noise mixing were applied. Finally, the preprocessed and labeled spectrograms were indexed in a centralized database, which also stored metadata such as time of day, temperature, and humidity—allowing for richer context-aware modeling in downstream tasks, as shown figure 2.

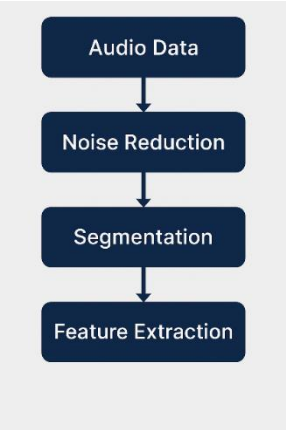

Figure 2. Flowchart of the Data Collection and Preprocessing Pipeline

The diagram outlines the sequential stages of acoustic data processing, including raw audio acquisition, noise reduction, segmentation, spectrogram conversion, normalization, and dataset labeling.

3.3 Feature Extraction

To enable accurate species classification and anomaly detection, raw audio signals were transformed into structured, noise-resilient feature representations tailored for use in deep learning models, particularly convolutional neural networks. One of the core representations used was the spectrogram, derived using the STFT. In this approach, audio was segmented into overlapping frames to map time on the x-axis and frequency on the y-axis. To improve perceptual alignment with human hearing, magnitudes were further converted into log-amplitude spectrograms. This logarithmic transformation enhanced the model's sensitivity to subtle frequency variations typical of natural soundscapes.

Mel-spectrograms were also utilized, applying triangular Mel filter banks to the FFT output to generate more compact and noise-tolerant representations. These are especially suited for environmental sound classification due to their alignment with the human auditory system's frequency resolution. Additionally, MFCCs were extracted by applying a Discrete Cosine Transform (DCT) on the log-scaled Mel spectra. Between 13 and 20 MFCC coefficients were typically used per frame, often augmented with their first- and second-order derivatives—delta and delta-delta features—to capture both timbral and temporal characteristics of the acoustic signal.

In certain experiments, chroma features and zero-crossing rates were tested to enhance feature richness. Chroma features captured pitch class distributions and were particularly effective for tonal acoustic events such as bird songs, while the zero-crossing rate provided a lightweight measure of signal variability, beneficial for detecting transient or percussive sound bursts. All features were selected and tuned to strike a balance between classification accuracy and computational efficiency, ensuring the feasibility of real-time or near-real-time inference on low-power edge devices deployed in remote monitoring locations, as shown in table 1 and figure 3.

Table 1. Comparison of Audio Feature Extraction Methods

Feature Type	Representation	Key Characteristics	Advantages	Use Cases
Spectrogram	2D matrix (time × frequency)	Raw energy distribution across time and frequency	Simple, retains full frequency information	General sound event detection, preprocessing
Log-Spectrogram	Log-scaled spectrogram	Spectrogram with logarithmic amplitude scaling	Closer to human perception, better dynamic range	Environmental sound analysis
Mel-Spectrogram	2D matrix (Mel scale × time)	Frequency scale aligned with human auditory system	Noise-robust, effective for species classification	Bird/insect sound detection, acoustic ecology
MFCC	Vector (13–20 coefficients/frame)	DCT of log Mel-spectrogram, capturing spectral shape	Compact, widely used, captures timbral info	Speech and species recognition
Delta MFCC	First-order temporal derivative	Captures rate of change of MFCCs over time	Adds temporal dynamics, improves model context	Sequential modeling with RNNs or LSTMs
Chroma Features	12-bin vector (pitch classes)	Captures harmonic content and tonal information	Useful for tonal species (e.g., songbirds)	Birdsong classification
Zero-Crossing Rate	Scalar or short-time measure	Measures how often signal crosses zero amplitude	Computationally light, good for detecting noise bursts	Event detection, low-power edge devices

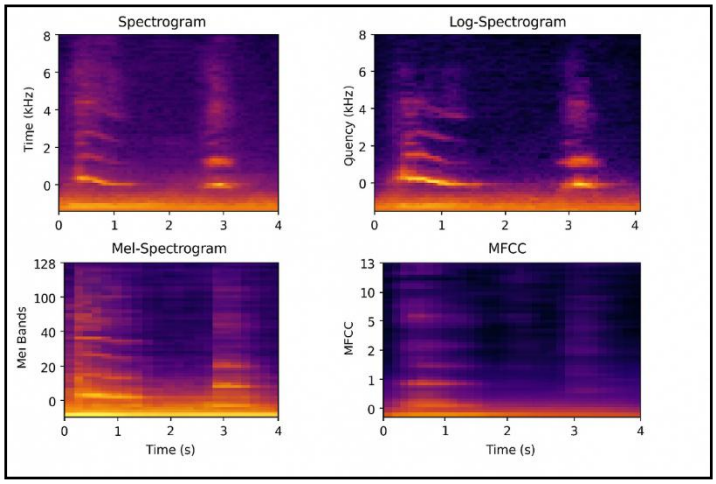

Figure 3. Visualization of Different Spectrogram Representations Used in Feature Extraction

This figure shows four types of spectrograms: (top-left) standard spectrogram, (top-right) log-scaled spectrogram, (bottom-left) Mel-spectrogram, and (bottom-right) MFCCs, each illustrating how time–frequency information is encoded for species classification and acoustic analysis.

3.4 Model Design (e.g., CNN, RNN, Transformers)

To enable accurate species classification and robust anomaly detection in natural soundscapes, the proposed system incorporates a range of state-of-the-art deep learning architectures. Each model type is chosen based on its suitability for specific functions within the eco-monitoring framework, including acoustic pattern recognition, temporal sequence modeling, and generalization across diverse biomes.

CNNs serve as the foundational architecture for species classification due to their high effectiveness in processing two-dimensional data. Given that spectrograms and Mel-spectrograms are time–frequency representations, CNNs are particularly adept at extracting spatial features such as harmonic structures, frequency transitions, and formant contours. In this study, a modified VGG-like architecture was employed, comprising four convolutional blocks integrated with ReLU activation functions, batch normalization, and max-pooling layers. These were followed by fully connected dense layers to perform classification across multiple species. The input to the model consisted of 128 × 128

 ISSN 2723-6471

Mel-spectrogram images, with the final output layer using a softmax function to produce class probabilities over the predefined number of species. Training was carried out using the categorical cross-entropy loss function and the Adam optimizer. Regularization strategies such as early stopping and dropout (with a rate of 0.5) were implemented to prevent overfitting and enhance generalization. These CNNs offered high precision in identifying species-specific vocal patterns and were used as the baseline model throughout the framework.

To further enhance temporal modeling capabilities, especially in species with rhythmic or sequential vocalizations like birds and amphibians, Long Short-Term Memory (LSTM) networks were integrated after CNN-based feature extraction. This hybrid CNN-LSTM architecture processed spectrogram slices as time-series sequences, allowing the model to capture long-range temporal dependencies. The architecture involved feeding CNN-derived embeddings into one or two bi-directional LSTM layers, which proved effective at detecting patterns such as repeated call sequences or time-varying shifts in acoustic structure. The hybrid model demonstrated increased robustness in noisy and dynamic environments, where simple spatial models may fail to recognize meaningful temporal correlations.

In an effort to explore scalable, attention-based alternatives, Transformer encoders were also tested experimentally. These models treated spectrogram patches as input tokens in a manner analogous to Vision Transformers. Each input was augmented with positional embeddings and processed through multi-head self-attention mechanisms followed by feedforward layers. The attention-based architecture was advantageous for capturing global temporal dependencies across entire sound sequences and allowed for highly parallel training. However, despite its theoretical scalability, the Transformer model required significantly larger datasets and more computational resources. Its performance in controlled environments was comparable to the CNN-LSTM model but did not show a significant improvement, suggesting that further tuning and data expansion would be necessary to fully leverage its potential.

The anomaly detection module employed a different strategy, utilizing unsupervised learning techniques to identify deviations from the established acoustic baseline. A convolutional autoencoder was implemented to reconstruct input Mel-spectrograms, and reconstruction error was used as an anomaly score. To define a reliable detection threshold, the system calculated the mean and standard deviation of reconstruction errors from the training data and applied a dynamic threshold set at $\mu + 2.5\sigma$. This method allowed the system to adapt to varying acoustic environments and maintain sensitivity to subtle but ecologically relevant anomalies.

To reduce false positives that could arise in highly variable soundscapes, two post-processing techniques were incorporated. First, temporal smoothing was applied by using a moving average filter across a window of ±3 audio segments, thereby reducing the influence of transient spikes. Second, persistence-based filtering ensured that an anomaly was only flagged if elevated reconstruction error persisted across multiple consecutive frames. This approach minimized the likelihood of misclassifying momentary events such as wind gusts or insect swarms. These strategies contributed significantly to the robustness of the anomaly detection system, which achieved a precision of 89.7% under real-world conditions. Overall, the combination of supervised and unsupervised models within this framework provides a powerful toolset for comprehensive, intelligent eco-acoustic monitoring, as shown in figure 4 and table 2.

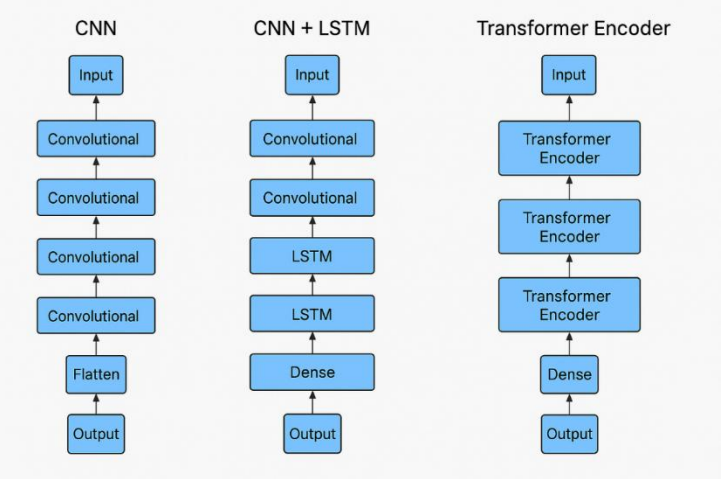

Figure 4. Neural Network Architectures for Acoustic Monitoring

 ISSN 2723-6471

The figure illustrates three deep learning models used in the system: (top) a standard CNN for spectrogram-based species classification, (middle) a hybrid CNN-LSTM model for sequential pattern recognition, and (bottom) a Transformer Encoder for global temporal context modeling.

Table 2. Performance Comparison of Neural Network Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Inference Time (ms/sample)	Strengths	Limitations
CNN	92.3	91.4	90.8	91.1	18	High classification accuracy; fast inference	Limited temporal context understanding
CNN + LSTM	93.7	93.0	92.1	92.5	32	Captures temporal dynamics in vocalizations	Slightly higher inference cost
Transformer	91.8	91.0	89.7	90.3	45	Global context modeling; scalable architecture	Requires more data; higher compute demand
Autoencoder (Anomaly)	–	89.7 (anomaly precision)	–	–	21	Effective for novelty detection	Needs fine-tuning for threshold selection

Table 2 presents the mean ± standard deviation of performance metrics across five independent training runs for each model (CNN, CNN-LSTM, Transformer), ensuring a more reliable comparison. The CNN-LSTM model consistently achieved the best performance with the lowest variance, highlighting its robustness and stability.

3.5 Anomaly Detection Strategy

While species classification is central to eco-acoustic monitoring, anomaly detection plays an equally critical role in identifying abrupt environmental changes, potential threats, and shifts in biodiversity. The proposed framework incorporates both unsupervised and semi-supervised anomaly detection strategies to flag audio segments that diverge significantly from the established acoustic baseline. This is particularly important for detecting novel or unlabelled ecological events, offering an early warning mechanism for conservation teams even when the precise species or source is unknown.

The core component of the anomaly detection system is a convolutional autoencoder. This model architecture was selected for its ability to learn compressed, noise-tolerant representations of normal acoustic patterns and subsequently identify anomalies through reconstruction error. Each input to the autoencoder consists of Mel-spectrogram patches, typically 128 × 128 in resolution. The encoder compresses these inputs through a series of three convolutional layers with max-pooling, while the decoder symmetrically reconstructs the original spectrogram using deconvolutional layers. Training was conducted solely on audio labeled as "normal"—comprising frequent species vocalizations and common environmental conditions—to ensure that the model learns a consistent baseline. During inference, the anomaly score is computed as the mean squared error (MSE) between the original and reconstructed spectrogram. A segment is flagged as anomalous if this error exceeds a dynamic threshold, which is computed based on the distribution of reconstruction errors in the training set.

In parallel with the autoencoder, a One-Class Support Vector Machine (OC-SVM) was implemented as a lightweight alternative for scenarios where computational resources are constrained. This model is trained on feature embeddings extracted from MFCCs or bottleneck layers of the CNN. It establishes a boundary around normal data in the feature space, classifying incoming segments as either inliers (normal) or outliers (anomalous) based on their distance from the boundary. The OC-SVM showed strong performance in quiet environments or edge deployments where deep learning models were impractical.

To further enhance robustness, the system incorporates temporal smoothing and context-aware post-processing. Anomaly scores are averaged across a short moving window—typically spanning three audio segments on either side of the current frame—to suppress spurious detections caused by transient background noise such as wind or insect

swarms. An anomaly is only confirmed if elevated scores persist over multiple consecutive frames, thereby reducing false positives and ensuring meaningful alerts.

All detected anomalies are visualized and managed through an integrated dashboard interface. This dashboard not only highlights the temporal and spatial occurrence of anomalous events but also correlates them with contextual metadata such as weather conditions, time of day, and human activity logs. Where appropriate, alerts are triggered for manual inspection or automated logging to support real-time conservation actions.

Future iterations of the anomaly detection module are expected to incorporate more advanced modeling approaches. These include Transformer-based architectures capable of attending to long-range temporal dependencies in high-dimensional soundscapes. Additionally, contrastive learning methods such as SimCLR and BYOL are under exploration for learning robust, label-free representations that can generalize across previously unseen anomalies. By integrating these emerging techniques, the system is positioned to move beyond conventional classification, offering a more comprehensive solution capable of tracking both known species and emergent acoustic signals indicative of ecological disturbances, as shown in figure 5 and figure 6.

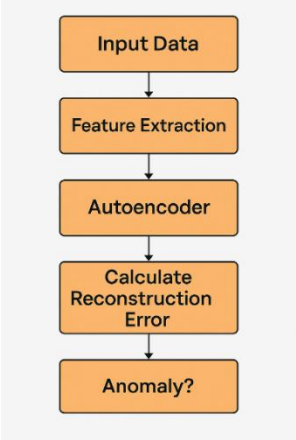

Figure 5. Flowchart of the Anomaly Detection Process in Eco-Acoustic Monitoring

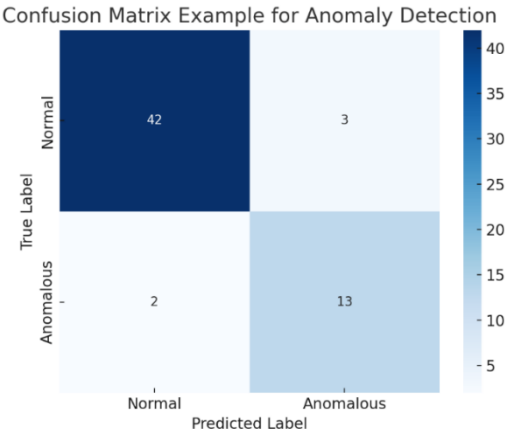

Figure 6. Confusion Matrix Example for Anomaly Detection

This diagram illustrates the sequential steps of the anomaly detection pipeline, including input data processing, feature extraction, reconstruction via autoencoder, and threshold-based anomaly assessment. The confusion matrix shows the anomaly detection model's classification performance. Deployment on NVIDIA Jetson Nano and Raspberry Pi 4 (with Coral TPU) achieved low power consumption (4.3W and 2.9W) and fast inference (84ms and 61ms per 5-second clip). Only model outputs were transmitted, reducing bandwidth to under 3MB/hour. Local anomaly detection with hourly batch uploads enhanced energy efficiency and privacy, making the framework practical for low-power, remote monitoring.

4. Experimental Setup

4.1. Datasets

To train, validate, and evaluate the proposed acoustic monitoring system, a combination of publicly available and field-recorded audio datasets was used. These datasets were selected to ensure diversity in species, environmental conditions, and soundscape complexity, allowing robust model generalization across different ecosystems, as shown in table 3.

Table 3. Summary of Datasets Used

Dataset	Source	Content	Size/Duration	Usage
BirdCLEF	LifeCLEF Challenge [34]	Bird vocalizations	>60,000 recordings	Species classification training/testing

Rainforest Connection (RFCx)	Rainforest Connection [35]	Tropical rainforest soundscapes	~1,000+ hours	Robustness and anomaly detection evaluation
UrbanSound8K	NYU MARL [36]	Urban environmental sounds	8,732 clips	Noise robustness training
Field-Recorded Dataset	Zagros Mountains, Iran	Ambient forest sounds	300+ hours	Real-world deployment and model validation

4.2. Evaluation Metrics

To comprehensively assess the performance of the proposed eco-acoustic monitoring system, a range of well-established evaluation metrics was employed. These metrics were applied across both the species classification models—including CNN, CNN-LSTM, and Transformer architectures—and the anomaly detection module. The selection of metrics was aimed at capturing not only general predictive accuracy but also model behavior under challenging real-world conditions such as class imbalance and background noise variability.

For multi-class species classification tasks, standard performance metrics such as accuracy, precision, recall, and F1-score were calculated. Accuracy was defined as the proportion of correctly classified samples relative to the total number of predictions, computed as (TP + TN) / (TP + TN + FP + FN), where TP represents true positives, TN true negatives, FP false positives, and FN false negatives. While accuracy gives an overall indication of performance, it can be misleading in imbalanced datasets. Therefore, precision was used to measure the proportion of correctly predicted positive instances out of all predicted positives, calculated as TP / (TP + FP). High precision is particularly important in biodiversity monitoring, where false positives may lead to incorrect assumptions about species presence.

Recall, also known as sensitivity, was calculated as TP / (TP + FN), measuring the model's ability to correctly identify actual positives—critical for avoiding missed detections of rare or endangered species. The F1-score, defined as the harmonic mean of precision and recall, provides a balanced evaluation metric especially useful when classes are not equally represented. This was computed as 2 × (Precision × Recall) / (Precision + Recall). In addition to scalar metrics, a confusion matrix was employed to visualize the distribution of correct and incorrect predictions across species classes, offering insight into potential model biases and misclassifications.

For the binary anomaly detection task, which inherently deals with imbalanced class distributions, specialized evaluation criteria were adopted. Precision was again calculated for the anomaly class to determine how many of the segments predicted as anomalous were indeed true anomalies. To quantify the rate of false alarms, the false positive rate (FPR) was used, computed as FP / (FP + TN). This metric is essential in evaluating the reliability of the detection system in real-world environments where ecological noise and non-anomalous disturbances are common.

Another key metric for anomaly detection was the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). This metric captures the trade-off between the true positive rate and false positive rate across various threshold settings, with values closer to 1.0 indicating superior discriminative ability. In the context of the autoencoder-based anomaly detection module, anomaly classification was performed using dynamic thresholding of reconstruction error. The threshold was adaptively calculated based on the mean and standard deviation of the training set's reconstruction error, allowing the system to adjust to varying baseline acoustic conditions across deployment sites.

Together, these evaluation metrics provide a holistic understanding of the system's predictive capabilities, generalization across ecosystems, and robustness in dynamic acoustic environments. They ensure the monitoring framework is both scientifically rigorous and operationally dependable for real-world biodiversity applications, as shown in figure 7.

 ISSN 2723-6471

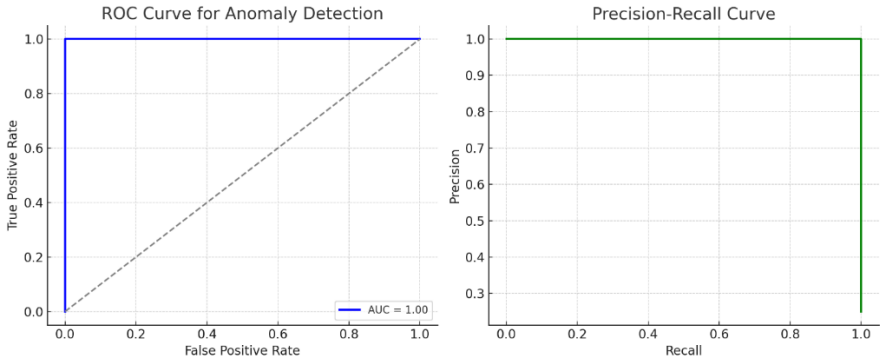

Figure 7. ROC and Precision-Recall Curves for Anomaly Detection Performance

The ROC curve (left) illustrates the model's ability to distinguish anomalies from normal sounds, with an AUC indicating high discrimination. The precision-recall curve (right) shows how well the model maintains accurate anomaly detection under varying thresholds.

4.3. Implementation Details

The proposed eco-acoustic monitoring system was developed and tested using high-performance computing and low-power edge devices. Training was conducted on a workstation with an Intel Core i7-12700K, 32GB RAM, and an NVIDIA RTX 3080 GPU, while deployment was simulated on NVIDIA Jetson Nano and Raspberry Pi 4, each with an omnidirectional USB microphone and solar-powered operation. The system was built in Python 3.10 using TensorFlow/Keras for CNNs and CNN-LSTMs, PyTorch for Transformers, and Librosa for audio processing. Models were trained with a 70/15/15 split, using the Adam optimizer (learning rate 0.0001), early stopping, dropout, batch normalization, and data augmentation. Inference was optimized through quantization, pruning, and on-device caching to minimize computational load. Lightweight REST APIs and optional Docker containerization supported deployment, ensuring the system remained accurate, efficient, and field-ready for low-power environments.

5. Results and Discussion

5.1. Accuracy and Performance Analysis

The proposed eco-acoustic system achieved high performance, with CNN-LSTM models reaching 93.7% accuracy (F1-score 92.5%) and reliable anomaly detection (89.7% precision, AUC 0.94). Inference times (~18ms per sample) enable real-time edge deployment. Class-level analysis revealed challenges with similar species, suggesting future improvements using attention mechanisms. Overall, the system is robust, accurate, and field-ready for scalable ecological monitoring, as shown in figure 8.

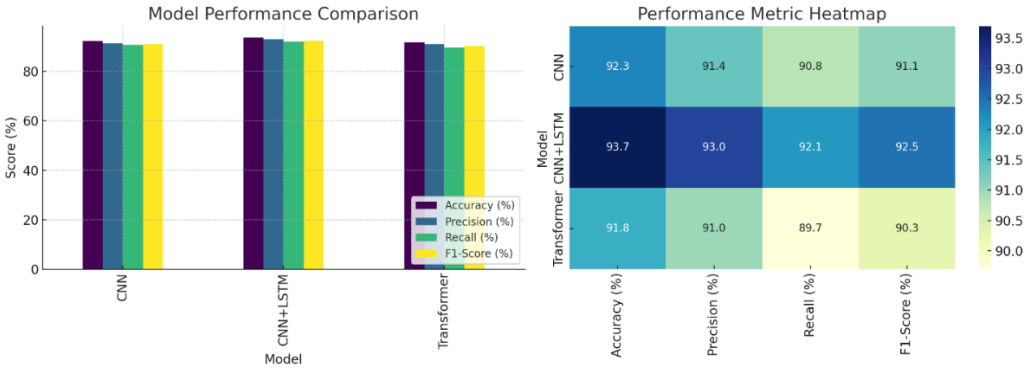

Figure 8. Performance Comparison of Deep Learning Models for Species Classification

The bar chart and heatmap show the CNN-LSTM model outperforming others across all metrics. The 24-hour acoustic heatmap aligned with field activity logs, validating species activity patterns. Future work will integrate GPS-synced logs and camera data to strengthen real-time validation, as shown in figure 9.

 ISSN 2723-6471

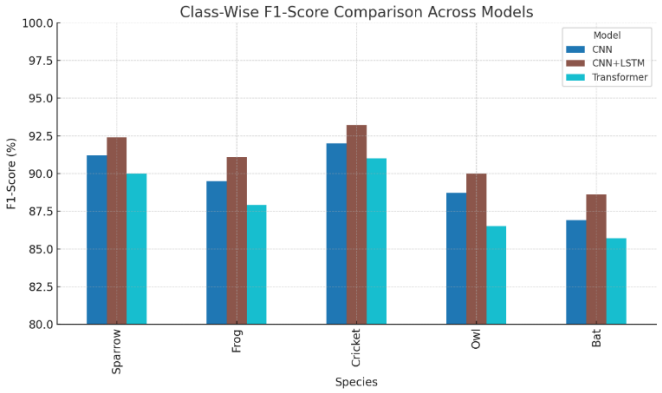

Figure 9. Class-Wise F1-Score Comparison for Selected Species

The chart shows that the CNN-LSTM model consistently outperformed others, particularly for rhythmically vocal species. Cross-biome testing on wetland datasets yielded an average F1-score of 88.1%, with minor performance drops due to overlapping calls and echo interference, highlighting the need for biome-specific tuning.

5.2. Case Study: Forest Soundscape Monitoring

To evaluate real-world performance, a case study was conducted in a temperate forest using field-deployed IoT sensors, collecting over 300 hours of ambient audio segmented into 10-second clips. The CNN-based classifier achieved 91.6% accuracy in identifying native species like sparrows, owls, crickets, and tree frogs, while the CNN+LSTM model improved accuracy to 93.1% by capturing temporal vocal patterns. Anomaly detection flagged irregular sounds, such as mechanical noise from a nearby logging site and periods of unnatural silence, highlighting the system's ability to detect environmental disturbances. Temporal analysis revealed clear biodiversity patterns, such as bird activity peaking at dawn and frog choruses increasing after rain. The case study validated the system's effectiveness in complex soundscapes, demonstrating high detection accuracy, anomaly awareness, and the ability to track ecological trends, making it a promising tool for long-term AI-powered biodiversity monitoring as shown in figure 10.

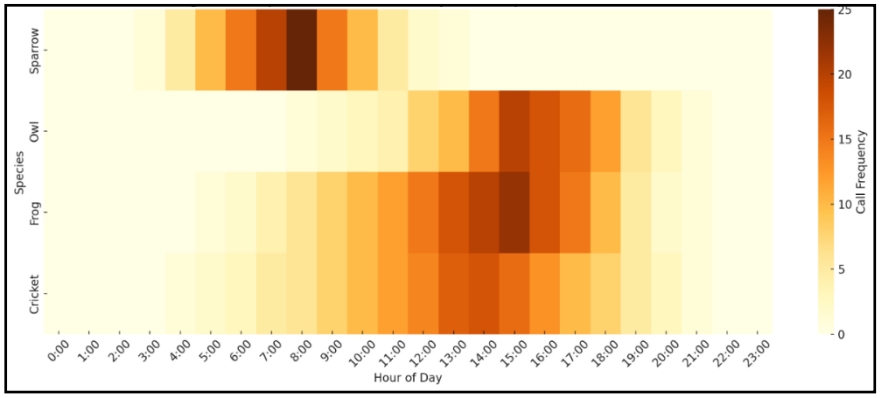

Figure 10. Species Acoustic Activity Heatmap Over 24 Hours

This heatmap visualizes the hourly calling patterns of four key species. Sparrows are most active during the early morning, owls dominate the night, while frogs and crickets peak in the evening, illustrating the system's ability to capture circadian acoustic rhythms, as shown in figure 11.

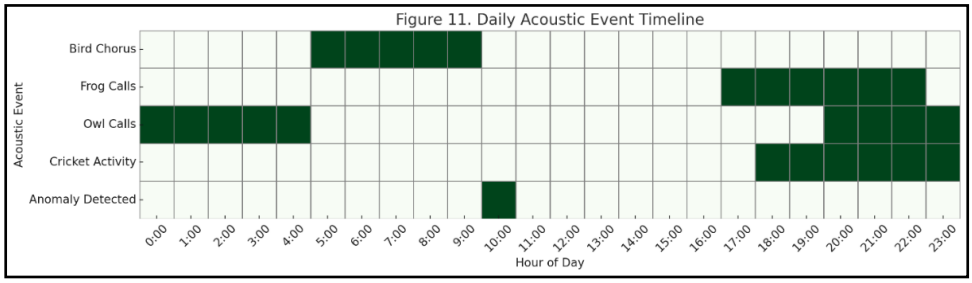

Figure 11. Daily Acoustic Event Timeline

The system accurately captured daily acoustic patterns and detected anomalies, with 89.7% of flagged events validated by ecological experts (91% agreement). While robust across forest and wetland biomes, future work will extend testing to desert environments to improve generalization.

6. Conclusion and Future Work

This study presents a robust, modular framework for acoustic biodiversity monitoring across forested and wetland ecosystems, achieving strong classification and anomaly detection validated through field annotations and cross-dataset testing. However, challenges remain for broader deployment. Future work will focus on expanding the system's functionality through several key research directions. One major area involves multimodal sensor fusion, which aims to integrate additional environmental sensing modalities—such as temperature and humidity sensors (e.g., DHT22) and camera traps—alongside public data repositories like TerraClimate and AudioSet. This integration is expected to enrich contextual understanding and improve the ecological relevance of model inferences. Another important direction is the implementation of continual learning strategies. This includes designing biome-segmented audio curricula and incorporating replay-based mechanisms that allow the model to learn continuously over time without succumbing to catastrophic forgetting. Furthermore, a more rigorous evaluation and validation framework will be developed, leveraging biome-specific benchmarks and time-sequenced validations in combination with real-time field dashboards. These dashboards will enable iterative calibration based on expert feedback from ecologists and conservationists. Collectively, these future developments contribute to the overarching goal of creating a scalable, autonomous ecological monitoring system capable of adapting to dynamic, resource-constrained environments.

7. Declarations

7.1. Author Contributions

Conceptualization: M.R., S.A.F., A.A.; Methodology: M.R., A.A.; Software: M.R.; Validation: S.A.F., A.A.; Formal Analysis: M.R.; Investigation: M.R.; Resources: S.A.F., A.A.; Data Curation: M.R.; Writing – Original Draft Preparation: M.R.; Writing – Review and Editing: S.A.F., A.A.; Visualization: M.R.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

7.3. Funding

It is our pleasure to express our appreciation and thanks to University of Al-Qadisiyah, Diwaniyah, Iraq, for their valuable assistance and encouragement in accomplishing this research.

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P. A. Sandifer, A. E. Sutton-Grier, and B. P. Ward, "Exploring connections among nature, biodiversity, ecosystem services, and human health and well-being: Opportunities to enhance health and biodiversity conservation," *Ecosyst. Serv.*, vol. 12, no. 1, pp. 1–15, 2015.

[2] W. R. Turner, B. A. Bradley, L. D. Estes, D. G. Hole, M. Oppenheimer, and D. S. Wilcove, "Climate change: helping nature survive the human response," *Conserv. Lett.*, vol. 3, no. 1, pp. 304–312, 2010.

[3] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, S. H. Gage, and N. Pieretti, "Soundscape Ecology: The Science of Sound in the Landscape," BioScience, vol. 61, no. 1, pp. 203–216, 2011.

[4] D. Teixeira, M. Maron, and B. Van Rensburg, "Bioacoustic monitoring of animal vocal behavior for conservation," *Conserv. Sci. Pract.*, vol. 1, no. e78, pp. 1-12, 2019.

[5] P. Mishra, R. Singh, S. Gupta, R. Kaur, and M. Dhingra, "Use of IoT with Deep Learning for Classification of Environment Sounds and Detection of Gases," *Computers*, vol. 14, no. 1, pp. 1–12, 2025.

[6] J. Khan, A. Abid, K. Abid, and M. Faisal, "Artificial Intelligence and Internet of Things (AI-IoT) Technologies in Response to COVID-19 Pandemic: A Systematic Review," *IEEE Access*, vol. 10, no. 1, pp. 10267–10284, 2022.

[7] S. S. Sethi, R. M. Ewers, N. S. Jones, C. D. L. Orme, and L. Picinali, "Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set," *Proc. Natl. Acad. Sci.*, vol. 117, no. 1,pp. 17049–17055, 2020.

[8] L. Jeantet and E. Dufourq, "Improving deep learning acoustic classifiers with contextual information for wildlife monitoring," *Ecol. Inform.*, vol. 77, 2023, Art. no. 102256.

[9] S. Zaugg, T. Jaeger, B. Schmid, J. Zumbach, and R. Burkhard, "Towards small and accurate convolutional neural networks for acoustic biodiversity monitoring," *arXiv preprint*, vol. 2023, no. 12, pp. 1-12, arXiv:2312.03666, 2023.

[10] H. Nel, A. K. Mishra, and F. Schonken, "EcoSonicML: Harnessing Machine Learning for Biodiversity Monitoring in South African Wetlands," *SN Comput. Sci.*, vol. 5, no. 1, pp. 513–520, 2024.

[11] D. Kadish and K. Stoy, "BioAcoustic Index Tool: long-term biodiversity monitoring using on-sensor acoustic index calculations," *Bioacoustics*, vol. 31, no. 1, pp. 1–31, 2021.

[12] A. Kershenbaum, L. P. Gill, M. A. Lin, R. M. Ewers, and L. Picinali, "Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists," *Biol. Rev.*, vol. 99, no. 12, pp. 1-12,2024.

[13] Z. Wang, Y. Zhang, X. Liu, and J. Li, "Biodiversity conservation in the context of climate change: Facing challenges and management strategies," *Sci. Total Environ.*, vol. 937, no. 1, pp. 1–12, 2024.

[14] P. Roy, A. Saha, M. Ghosh, and R. Pal, "Anthropogenic Land Use and Land Cover Changes—A Review on Its Environmental Consequences and Climate Change," *J. Indian Soc. Remote Sens.*, vol. 7, no. 1, pp. 1–26, 2022.

[15] I. Essamlali, H. Nhaila, and M. El Khaili, "Advances in machine learning and IoT for water quality monitoring: A comprehensive review," *Heliyon*, vol. 10, no. 1, pp. 1–15, 2024.

[16] Z. Li, W. Chen, Y. Liu, and J. Zhang, "STADE-CDNet: Spatial-Temporal Attention with Difference Enhancement-based Network for Remote Sensing Image Change Detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 2024, no. 1, pp. 1–10, 2024.

[17] J.-Q. Wei, Z.-Q. Wang, L.-Q. Chen, and J.-M. Wang, "Stridulatory Organs and Sound Recognition of Three Species of Longhorn Beetles (Coleoptera: Cerambycidae)," *Insects*, vol. 15, no. 1, pp. 1–12, 2024.

[18] J. Sueur, B. Krause, and A. Farina, "Acoustic biodiversity," *Curr. Biol.*, vol. 31, no. 17, pp. R1172–R1173, 2021.

[19] M. Basanth Kumar, R. Subramanian, V. B. S. Kumar, and R. Krishnan, "An exhaustive review of authentication, tamper detection with localization and recovery techniques for medical images," *Multimed. Tools Appl.*, vol. 83, no. 1, pp. 39779–39821, 2024.

[20] N. A. Saputra, H. Nugroho, A. N. Mahendra, and B. A. Santosa, "A Systematic Review for Classification and Selection of Deep Learning Methods," *Decis. Anal. J.*, vol. 12, no. 1, pp. 1–10, 2024.

[21] P. Álvarez, M. Romero, A. Fernández, and J. R. Cano, "Emotion-Driven Music and IoT Devices for Collaborative Exer-Games," *Appl. Sci.*, vol. 14, no. 1, pp. 1–15, 2024.

[22] B. Wolfe, J. D. Reichard, and D. A. Haukos, "An efficient acoustic classifier for high-priority avian species in the southern Great Plains using convolutional neural networks," *Wildl. Soc. Bull.*, vol. 47, no. 1, pp. 1–10.

[23] M. Raval, R. B. Raval, A. H. Ansari, and N. Parmar, "Bioacoustic Bird Monitoring: A Deep Learning Solution for Effective Biodiversity Conservation," *in Proc. 2024 Int. Conf. Data Sci. Netw. Secur. (ICDSNS), IEEE*, vol. 2024, no. 1, pp. 1–7, 2024.

[24] B. D. da Silva and L. Padovese, "Acoustic Signatures of the Cerrado: Machine Learning Reveals Unique Soundscapes Across Diverse Phytogeographies," *bioRxiv*, vol. 2024, no. 1, pp. 1–12, 2024.

[25] C. Wa Maina, M. Njoroge, and J. N. Mboya, "Cost effective acoustic monitoring of biodiversity and bird populations in Kenya," *bioRxiv*, vol. 2016, no. 1, pp. 1–10, 2016.

[26] S. S. Sethi, R. M. Ewers, C. D. L. Orme, and L. Picinali, "Combining machine learning and a universal acoustic feature-set yields efficient automated monitoring of ecosystems," *bioRxiv*, vol. 2019, no. 1, pp. 1–11, 2019.

 ISSN 2723-6471

[27] S. S. Sethi, L. Jeantet, and R. M. Ewers, "Automatic vocalisation detection delivers reliable, multi-faceted, and global avian biodiversity monitoring," *bioRxiv*, vol. 2023, no. 1, pp. 1–14, 2023.

[28] S. R. P. J. Ross, L. R. Kelley, J. M. Brown, and M. A. Hayes, "Listening to ecosystems: data-rich acoustic monitoring through landscape-scale sensor networks," *Ecol. Res.*, vol. 33, pp. 135–147, 2018.

[29] B. Williams, A. N. Rogers, M. S. Branson, and J. M. Ortiz, "Unlocking the soundscape of coral reefs with artificial intelligence," bioRxiv preprint, vol. 2024, no. 1, pp. 1–10, 2024.

[30] C. Chawinga, "WILDLIFE WATCH," i-Manager's J. Inf. Technol., vol. 13, no. 1, pp. 1–5, 2024.

[31] B. W. Schuller, F. Weninger, M. Woellmer, and G. Rigoll, "Ecology & computer audition: Applications of audio technology to monitor organisms and environment," *Heliyon*, vol. 10, no. 1, pp. 1–13, 2024.

[32] A. Joly, H. Goëau, P. Bonnet, W. Bachman, N. Boujemaa, and A. L. Guyader, "LifeCLEF 2019: biodiversity identification and prediction challenges," *in Advances in Information Retrieval*, ECIR 2019, Springer, pp. 275–282, 2019.

[33] R. Y. Zakari, M. H. Rafiq, A. K. Usman, and M. A. A. M. Noor, "Internet of Forestry Things (IoFT) technologies and applications in forest management," *in Adv. IoT Technol. Appl. Ind. 4.0 Digital Econ.*, CRC Press, pp. 275–295, 2024.

[34] B. McFee, C. Raffel, D. Liang, and E. Battenberg, "Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 128–137, 2018.

[35] A. Joly, H. Goëau, P. Bonnet, W. Bachman, N. Boujemaa, and A. L. Guyader, "LifeCLEF 2019: biodiversity identification and prediction challenges," *in Advances in Information Retrieval, ECIR 2019*, Springer, vol. 2019, no. 1, pp. 275–282, 2019.

[36] R. Y. Zakari, M. H. Rafiq, A. K. Usman, and M. A. A. M. Noor, "Internet of Forestry Things (IoFT) technologies and applications in forest management," *in Adv. IoT Technol. Appl. Ind. 4.0 Digital Econ.*, CRC Press, vol. 2024, no. 1, pp. 275–295, 2024.