Deep Learning-Based Multi-Class Forest Sound Classification Using a CNN–LSTM–Attention Framework

Maddhigalla Lakshumaiah and D. S. Rao *
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, India
Email: lakshumaiah@klh.edu.in (M.L.); dsrao@klh.edu.in (D.S.R.)

Manuscript received February 22, 2026; revised March 23, 2026; revised again April 23, 2026; accepted May 15, 2026
*Corresponding author

***Abstract*—Acoustic monitoring is commonly used in forest studies since sound recordings can capture information related to animal activity, human presence, and environmental changes over time. However, traditional monitoring methods based on manual observation or remote sensing are limited in detecting short and irregular acoustic events and often require continuous human effort. This work presents a deep learning-based approach, termed CLASED (CNN–LSTM–Attention-based Sound Event Detection). The framework combines convolutional neural networks for extracting spectral features with long short-term memory networks to model temporal variations in audio signals. An attention mechanism is incorporated to improve the contribution of relevant time segments during the classification process. A dataset of 3,000 forest-related audio recordings was prepared and processed through resampling, augmentation, and feature normalization before training. The developed model classifies audio signals into three categories: animal calls, human activity, and natural environmental sounds. Experimental results show an overall classification accuracy of 79%, with better performance observed for animal vocalizations and consistent performance across the remaining classes. The study demonstrates that integrating temporal modeling with attention can support reliable automated analysis of forest sound data and reduce dependence on manual monitoring.**

***Index Terms*—acoustic signal analysis, convolutional neural networks, deep learning for audio recognition, environmental sound classification, forest acoustic analysis, long short-term memory networks**

I. Introduction

The forest is a complex environment that carries significant information about nature as well as human activities through sound. Birds, insects, and mammals produce acoustic signals that, together with wind, rain, and anthropogenic noise, shape the overall forest soundscape. It has been discovered that the acoustics of the forest relate to its biodiversity and overall forest health [1]. Recent ecological research further validates that acoustic measurements can predict species richness and ecosystem health metrics [2]. These findings highlight that forest soundscapes contain rich ecological information reflecting biodiversity dynamics, requiring effective multisource sound classification techniques [3].

Conventional approaches to monitoring, such as ground-based observations, camera traps, and satellite imagery, also have limitations. Although satellites can monitor broad areas, they often fail to detect small and sudden events such as illegal logging, wildlife disturbance, or early-stage incidents [4]. In dense forest habitats, thick canopy cover obscures undergrowth activities, and acquiring frequent imagery in remote regions is often impractical and costly. Acoustic monitoring offers a practical alternative: it works continuously, even at night or in low-visibility conditions, and can quickly detect both natural and human-caused events [5, 6].

Advances in low-power audio recorders and automated analysis now enable processing of long recordings to identify different sound types [3]. Recent developments in embedded and edge-based deep learning systems further support efficient environmental sound recognition. Similar data-driven approaches have also been explored in related environmental monitoring domains, such as agriculture and geospatial analysis, where machine learning and satellite-based methods are used to model environmental variability and predict ecosystem-related outcomes [7]. Yet forest sound classification is not easy. Natural soundscapes are often noisy, with overlapping sounds and wide frequency ranges, while datasets are usually small and imbalanced, making it hard to train models that generalize well [8, 9]. Recent reviews further highlight the growing adoption of semi-supervised and unsupervised deep learning approaches to address data scarcity and labeling challenges in ecoacoustic tasks [10].

Early machine learning methods that rely on handcrafted features often struggle in such environments [5]. Deep learning models have improved performance; however, convolutional neural networks (CNNs) alone fail to capture temporal patterns, while recurrent networks alone lack strong spectral feature extraction. Many existing studies also use generic datasets that fail to reflect the variety of real forest sounds, limiting practical application [8, 11].

This study addresses these challenges with a hybrid convolutional–LSTM (long short-term memory) model featuring an attention mechanism for multi-class forest sound classification. The model learns both spectral and temporal features and focuses on the most informative

parts of the audio. Its performance is tested on curated forest sound datasets and selected subsets of the ESC-50 benchmark, keeping in mind practical use in resource-limited monitoring systems.

II. Literature Review

Studies for environmental sound classification have evolved over the years from conventional machine learning algorithms to state-of-the-art deep learning techniques. The initial work relied on classifier algorithms such as the Gaussian mixture model, support vector machine, and k-nearest neighbors, along with engineered features such as mel-frequency cepstral coefficients and elementary spectrum feature descriptors. Many such techniques are easy to code with low computational requirements, thus apt for initial development work with limited datasets [12]. Nonetheless, their performance relies heavily on good feature extraction, with adverse effects because of background noise and overlapping signals. In natural environments, for example, in a forest where wind, rainfall, insects, and human interactions coexist, the above constraints become more evident. Although certain gains have been achieved through the combination of features and adjustments to the classifiers, this kind of work has faced difficulties in sustaining a strong performance level within complex acoustic environments [13, 14].

However, with the increased use of deep learning techniques, the CNN has become the most popular technique used for the classification of environmental sounds. The CNN can learn features directly from the spectrogram image without any predefined features to focus on local patterns within the time-frequency domain. The CNN has demonstrated a marked improvement over traditional approaches for sound classification on standard benchmark datasets like ESC-50 and UrbanSound8K. However, the CNN has limited capability to focus on long-range dependencies within sound patterns for environment-related sounds like animal calls and human-related events that are often larger in duration compared to the local patterns emphasized by CNNs. This problem has been solved by integrating CNNs with recurrent networks like LSTMs and GRUs to develop convolutional recurrent neural networks that demonstrated an improved level of robustness for sound classification, albeit at the cost of increased computational requirements [8, 15].

Recently, attention models were proposed to further enhance the performance of sound classification in difficult environments. The primary motivation behind the attention mechanism is to encourage the network to pay more attention to the most informative time-frequency areas, thus discarding the noise areas. This is a significant advantage, especially in recorded environments, where the event of interest might consist of a small portion of the recorded signal. The attention models were successful in differentiating sounds in difficult environments, such as noise and overlap, by giving more weight to the relevant audio areas. This further resulted in more robust predictions, especially with multiple sources or a low signal-to-noise ratio [8, 16].

In spite of these developments, however, little work has been done on studies concentrating on forest soundscapes. Most methods have been tested on generic environmental audio samples and do not completely cover the acoustic diversity that is present in forests. Issues like model size and computational power have also been ignored. More recently, some studies have demonstrated that hybrid models combining CNN and LSTM layers and assisted by transfer learning methods could be very helpful in improving robustness to unseen data. These works indicate that combining learning methods is effective and useful for handling complicated acoustic scenes [17]. Transformer models and dynamic convolutions have also been introduced. While these models have been able to exhibit powerful representational ability, concerns have been growing about computational costs relative to long-term measurement applications in forests [18, 19]. Table I summarizes the above related work discussed above.

Despite these advancements, a clear research gap remains. Existing approaches are typically evaluated on general-purpose environmental datasets and do not adequately capture the acoustic diversity and noise characteristics of real forest environments. Furthermore, many models focus on either spectral or temporal features in isolation, or introduce attention mechanisms without explicitly addressing the combined challenges of overlapping sounds, class imbalance, and deployment constraints in resource-limited forest monitoring systems.

Therefore, there is a need for a unified framework that effectively integrates spectral feature extraction, temporal modeling, and attention mechanisms while remaining computationally practical for real-world ecoacoustic applications. This gap directly motivates the development of the proposed CLASED model.

Most existing studies highlight the need to balance model performance with practical deployability. In this context, forest soundscapes require a unified approach that integrates spectral learning, temporal modeling, and attention mechanisms. This requirement forms the foundation for the CNN–LSTM–Attention framework adopted in the present work.

Table I: Related Work Summary

Study	Method	Key Features	Strengths	Weaknesses
[13]	Traditional ML (GMM, SVM, kNN)	Handcrafted features (MFCCs, etc	Simple, low compute	Feature-dependent, noise sensitive
[8]	CNN	Spectrogram learning	Strong spectral learning	Poor temporal modeling alone
[15]	CRNN (CNN+RNN)	Sequential modeling	Captures temporal & spectral	Higher compute cost
[8]	Attention-CNN	Temporal/frequency attention	Highlights informative regions	Complexity
[17]	Hybrid CNN-LSTM	Integrated spectral & temporal	Balanced feature learning	Demands design choices
[18, 19]	Dynamic/Transformer models	Self-attention, adaptive filters	Global dependencies	Larger models

III. Dataset Design and Preprocessing

Most of the performance of sound classification depends on the data it has been trained on. This is even more robust in forest monitoring, considering the wide variations in acoustic conditions across locations, seasons, and weather. Most publicly available audio datasets are targeted for urban environments or general environmental sounds, and do not represent forest soundscapes properly. Due to this limitation, a dataset was prepared for the current study in which forest-relevant audio samples were collected and filtered from existing public datasets.

The dataset is centered on three classes of sounds that often exist in the forest environment. These classes of sound include animal sounds, sounds produced by human activities, and environmental sounds. The classification was tailored to suit the nature of forest monitoring.

A. Dataset Composition

The data was generated based on audio clip extraction from free online environmental and natural sounds databases. Only clips labeled for outdoor and forest-like conditions were used for analysis. The sounds of interest include those of birds and insects, as well as mammals common to forest ecosystems, recorded when vocalizing.

The audio clips used for data analysis were obtained from two free online databases, namely the 'iNaturalist Sounds' database [15] and 'Xeno-canto' database [16], both of which contain audio clips recorded in natural field conditions. Soundscapes, including sounds of human activity in the forest, such as chainsaw cuts and forestry equipment, were collected from selected classes within the ESC-50 dataset [17]. Despite being a general-purpose dataset for environmental sounds, some relevant sound classes, such as "forest" soundscapes, can be filtered from the soundscapes within the ESC-50 dataset.

Nature sounds like wind, thunder, and crackling fire were also selected from the suitable categories in the ESC50 database. Additional audio examples were taken from the Forest Sound Classification Dataset (FSC22) [18] to improve the quality of the sound dataset. After applying filtering and balancing techniques, a total of 3,000 audio samples were obtained (Table II). Every audio clip is between 5 and 10 seconds long and is stored in mono audio format with a sampling rate of 44.1 kHz. Audio samples were selected based on relevance to forest environments, clarity of the dominant sound source, and absence of excessive background distortion. To improve label reliability, audio samples were manually reviewed to ensure consistency with their assigned class labels and to remove ambiguous or mislabeled recordings. Class balancing was achieved by limiting the number of samples per class to comparable ranges during dataset construction and further equalizing representation through augmentation applied only to the training set.

To further reduce potential dataset bias, care was taken to include audio samples from multiple sources and recording conditions, ensuring diversity in background noise, recording devices, and environmental contexts. Additionally, augmentation techniques were applied only to the training set to prevent information leakage into validation and test sets. The use of stratified sampling during dataset splitting also ensured that class distributions remained consistent across all subsets, improving the generalization capability of the model.

Table II: Dataset Composition

Class	Sound Types Included	Source Datasets	No. of Samples
Wild Animals	Bird calls, insect sounds, mammal vocalizations	iNatSounds, Xeno-canto	1,200
Anthropogenic Activities	Chainsaws, logging vehicles, machinery	ESC-50	1,000
Natural Environmental Events	Wind, thunder, fire crackling, storm sounds	ESC-50, FSC22	800

B. Audio Preprocessing and Augmentation

Before the extraction of features, each audio file was standardized. This entailed resampling all audio files to a frequency rate of 44.1 kHz, as well as making all audio files mono and normalizing all audio files in terms of amplitude.

For enhanced resilience during training, data augmentation was employed. Time-stretching was employed to introduce slight variations in audio duration. Pitch-shifting was employed to incorporate variations in frequency response across records. Background mixing as well as noise injection were employed for simulating occlusion of sounds from the forest, as well as emulation of environmental and sensing noises. After data augmentation, the size of the training set was ensured to get close to 9,000 samples.

C. Feature Extraction

The raw audio signals were not used directly for model training; instead, time–frequency features were extracted. MFCCs were computed using 40 coefficients per frame, with a window of 25 ms and a hop size of 10 ms. Log-Mel spectrograms with 128 Mel bands were also calculated to capture the temporal and spectral patterns of speech. Besides, chroma features, spectral contrast, and zero-crossing rate were also extracted for complementary information. The extracted features (MFCC, log-Mel spectrogram, chroma, spectral contrast, and zero-crossing rate) are concatenated along the channel dimension to form a multi-channel input tensor. Each feature type is treated as a separate channel, analogous to color channels in image data, enabling the CNN to learn complementary representations across different acoustic feature spaces.

D. Dataset Splitting

Stratified sampling was used to divide the dataset into training, validation, and testing sets so that the class distribution can be preserved. For this, 70% of the samples were taken for training, 15% for validation, and 15% for testing. This split was maintained throughout all the experiments.

E. Preprocessing Pipeline

The entire preprocessing flow, ranging from raw audio files to feature extraction, is depicted in Fig. 1. This flow involves resampling, normalization, augmentation, feature extraction, and splitting datasets.

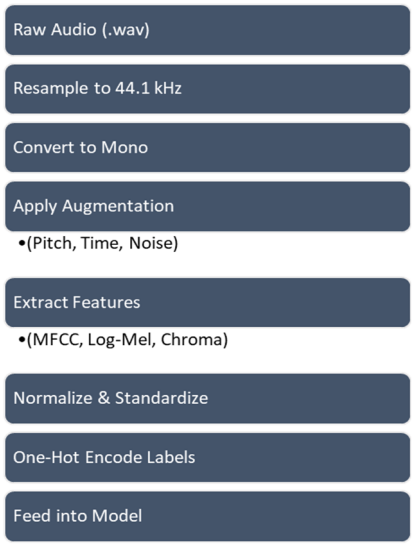

Fig. 1. Audio preprocessing pipeline used for forest sound classification.

IV. Proposed Methodology

The technique presented here describes a hybrid CNN-LSTM-Attention-based deep learning model called CLASED. Proposed for the sound classification task in a forest environment, the model is designed considering three major challenges observed in natural soundscapes: the presence of overlapping acoustic events, wide frequency variations, and the presence of persistent environmental noise. The proposed technique puts forward the importance of the combination of spectral feature extraction, temporal sequence modeling, and the attention mechanism to improve robustness and class discrimination under realistic field conditions.

The preprocessed time-frequency representations are the raw audio signals upon which the CLASED framework operates. The convolutional layers capture local spectral patterns, while longer temporal dependencies across frames are modeled by recurrent layers. Meanwhile, an attentional layer is added to enhance the informative parts of the sequence in audio and suppress irrelevant or noisy regions. This architecture lets the model learn these complementary representations, which could be hard to capture using single-architecture approaches.

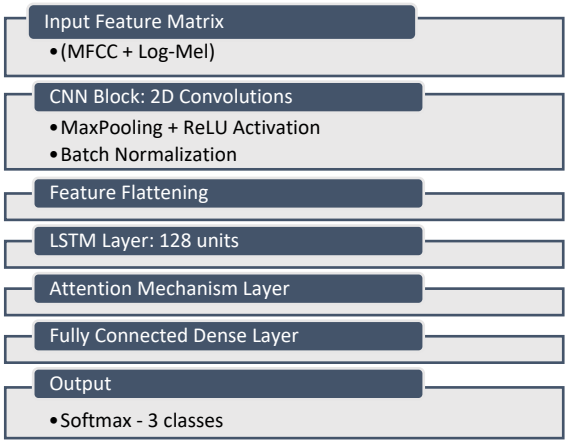

Fig. 2. Architecture of the proposed CLASED CNN–LSTM–Attention model.

A. Architecture Overview

The overall architecture of the CLASED model is composed of five main components: an input layer, a convolutional feature extraction block, a temporal modeling block based on LSTM units, an attention mechanism, and a dense output layer. The input layer receives multi-channel acoustic features, including MFCCs and log-Mel spectrograms, extracted during preprocessing. These features are passed sequentially through the network to produce final class probabilities corresponding to different forest sound event categories. Fig. 2 illustrates the high-level structure of the proposed model and the flow of data through each component.

B. Convolutional Block for Spectral Feature Extraction

The convolutional block is responsible for learning discriminative spectral patterns from the input feature maps. The input to this block is a two-dimensional representation of size $F \times T$, where F denotes the number of frequency coefficients, and T represents the number of time frames.

Two convolutional layers are employed with kernel sizes of 3×3, using 32 and 64 filters, respectively. Each convolution operation is followed by a ReLU activation function to introduce nonlinearity. Max-pooling layers with a 2×2 stride are applied to reduce spatial dimensions and control model complexity. Batch normalization is used to stabilize training and improve convergence.

The convolution operation for the k-th filter can be expressed as:

$$\mathbf{y}_{i,j}^{(k)} = \mathrm{ReLU}\left(\sum_{m,n} \mathbf{x}_{i+m,j+n} \cdot \mathbf{w}_{m,n}^{(k)} + \mathbf{b}^{(k)}\right) \quad (1)$$

where $\mathbf{y}_{i,j}^{(k)}$ denotes the output value at spatial location (i, j) of the feature map produced by the kth convolutional filter, $\mathbf{x}$ is the input feature map, $\mathbf{x}_{i+m,j+n}$ represents the input feature map value at location $(i+m, j+n)$, i and j index the spatial location of the output feature map, m and n index the spatial dimensions of the convolutional kernel. $\mathbf{w}_{m,n}^{(k)}$ denotes the weight at position (m, n) of the kth convolutional kernel, $\mathbf{b}^{(k)}$ denotes the bias of the k-th filter, k denotes the filter index, the operator "·" denotes multiplication between corresponding elements during the convolution operation, and $\mathrm{ReLU}(\cdot)$ is the rectified linear unit activation function.

C. LSTM Block for Temporal Modeling

The convolutional outputs are then reshaped into a sequence and passed to an LSTM layer for analyzing the temporal relationships between consecutive audio frames. For forest sound events, analyzing the temporal relationships is crucial since the information might be embedded in the form of call durations, repetitions, or intensity changes over time. The architecture of this LSTM layer comprises 128 hidden units and a dropout value of 0.3. These operations allow the network to retain the relevant temporal context while removing less informative data.

D. Attention Mechanism

To further improve the ability to discriminate between

noisy and overlapping speech, an attention mechanism is applied to the sequence of output values produced by the LSTMs. The purpose of this layer is to assign higher importance to time steps that contain salient acoustic information related to the target sound event.

The attention weights are computed as:

$$\mathbf{e}_t = \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a) \quad (2)$$

$$\alpha_t = \frac{\exp(\mathbf{e}_t)}{\sum_t \exp(\mathbf{e}_t)} \quad (3)$$

$$\mathbf{z} = \sum_t \alpha_t \mathbf{h}_t \quad (4)$$

where $\mathbf{h}_t$ denotes the hidden state output of the LSTM at time step t, $\mathbf{e}_t$ represents the unnormalized attention score corresponding to time step t, α_t denotes the normalized attention weight assigned to the hidden state at time step t, $\mathbf{z}$ is the context vector obtained as the weighted sum of the LSTM hidden states, $\mathbf{W}_a$ and $\mathbf{b}_a$ are the weight matrix and bias vector of the attention layer, respectively, The summation index t spans all time steps in the input sequence, and exp (·) denotes the exponential function.

E. Classification Layer and Training Strategy

The output vector from the attention layer is fed into a fully connected dense layer with softmax activation to generate class probabilities for the three sound event categories: wild animals, machinery and tools, and natural disasters.

The model is trained using categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.001. Training is performed for up to 100 epochs with a batch size of 32. Dropout and L2 regularization are applied to reduce overfitting. Early stopping and model checkpointing based on validation loss are used to ensure stable generalization.

The selected hyperparameters were chosen based on commonly adopted configurations in environmental sound classification literature and preliminary validation experiments. The learning rate of 0.001 with the Adam optimizer provided stable convergence, while a batch size of 32 ensured a balance between computational efficiency and gradient stability. The number of LSTM units and dropout rate were selected to capture sufficient temporal dependencies while reducing overfitting, as observed from validation performance.

F. Algorithm Description

Algorithm 1 outlines the complete pipeline for CLASED, including preprocessing of audio, feature extraction, model construction, training, and inference. This procedural description complements the architectural explanation and provides clarity on the end-to-end workflow implemented in this study.

Algorithm 1: CLASED – CNN–LSTM–attention-based sound classification

Input:
$D = \{x_1, x_2, ..., x_n\}$ // audio signals
$Y = \{y_1, y_2, ..., y_n\}$ // class labels
Output:
Trained CLASED model for forest sound event classification

Step 1: Audio Preprocessing
for each audio sample $x \in D$ do
Resample x to 44.1 kHz
Convert x to mono channel
Apply data augmentation (time stretching, pitch shifting, noise injection)
Extract MFCC and log-Mel spectrogram features
Normalize extracted features
Store processed features in feature set F
end for

Step 2: Model Construction
Initialize CNN layers for spectral feature extraction
Apply convolution, activation, batch normalization, and max pooling
Reshape CNN output into sequential format
Add LSTM layer with 128 hidden units for temporal modeling
Apply attention mechanism over LSTM outputs
Add fully connected dense layer with softmax activation

Step 3: Model Training
Split F and Y into training, validation, and test sets
Compile model using categorical cross-entropy loss and Adam optimizer
Train model for 100 epochs with batch size of 32
Apply early stopping and model checkpointing based on validation loss

Step 4: Inference
For an unseen audio sample:
Apply preprocessing and feature extraction
Feed features into trained CLASED model
Output predicted sound event class
Return trained CLASED model

V. Results

The CLASED model was evaluated using the forest-oriented dataset described in Section III, which includes three sound categories: wild animals, anthropogenic activities, and natural environmental events. The dataset was split into training, validation, and test sets in a 70:15:15 ratio using stratified sampling. All results reported here are based on the test set.

A. Training and Validation Performance

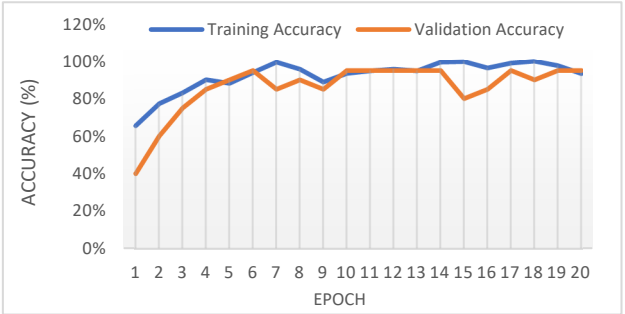

Fig. 3. Training and validation accuracy showing steady convergence and minimal overfitting.

The model was trained for up to 100 epochs; however, early stopping resulted in convergence after 20 epochs, and the final results correspond to this optimal stopping point. Training accuracy increased steadily from around 0.65 in

the initial epochs and reached values above 0.95 after approximately 10 epochs. Validation accuracy followed a similar pattern, increasing from roughly 0.40 in the early stages and stabilizing between 0.90 and 0.95 toward the end of training (Fig. 3).

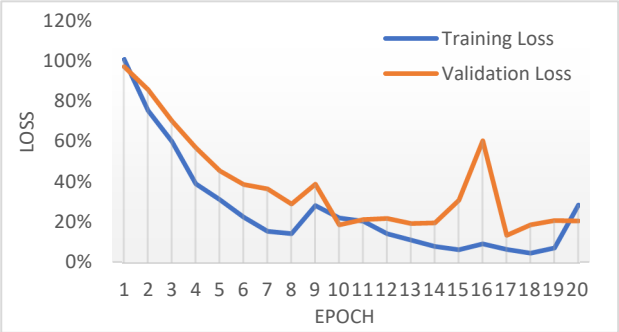

Fig. 4. Training and validation loss demonstrating stable optimization and convergence behavior.

Training loss decreased from about 0.97 at the start to below 0.20 by the final epochs. Validation loss also showed a downward trend and remained mostly between 0.20 and 0.30 after convergence (Fig. 4). The gap between training and validation curves remained small throughout training, indicating that the model did not suffer from severe overfitting.

B. Classification Metrics

The class-wise classification performance of the proposed model is summarized in Table III. Although the full test set consists of approximately 450 samples (15% of the dataset), Table III presents results on a representative subset of 24 samples for concise illustration of class-wise metrics. The wild animal class achieved strong performance, with a precision of 0.857, a recall of 1.000, and an F1-score of 0.923, indicating reliable detection of animal sounds. In contrast, anthropogenic activity sounds (chainsaw) obtained a precision of 0.600, a recall of 0.750, and an F1-score of 0.667, reflecting moderate classification accuracy for this class.

For natural environmental events (rain), the model achieved a precision of 0.800, while recall was comparatively lower at 0.500, resulting in an F1-score of 0.615. The overall classification accuracy across all classes was 79.17%. The macro-averaged precision, recall, and F1-score were 0.752, 0.750, and 0.735, respectively, while the weighted F1-score reached 0.778, indicating balanced performance when accounting for class sample sizes.

Support values reported in Table III indicate the number of test samples per class and are provided for contextual interpretation of the evaluation results rather than as performance metrics. For illustration purposes, representative labels are shown for each class in Table III: "Dog" corresponds to the wild animal category, "Chainsaw" represents anthropogenic activity, and "Rain" represents natural environmental events.

TABLE III: CLASS-WISE PERFORMANCE OF THE MODEL

Class/Metrics	Precision	Recall	F1-Score	Support
Chainsaw	0.600000	0.750000	0.666667	4.000000
Dog	0.857143	1.000000	0.923077	12.000000
Rain	0.800000	0.500000	0.615385	8.000000
Accuracy	0.791667	0.791667	0.791667	0.791667
Macro Avg	0.752381	0.750000	0.735043	24.000000
Weighted Avg	0.795238	0.791667	0.777778	24.000000

C. Metric Distributions and 2D Analysis

The metric distribution plots (Fig. 5) illustrate the spread of precision, recall, and F1-score across classes. Precision values mostly fall between 0.75 and 0.85, while recall values range from 0.50 to 1.00. The F1-scores lie between approximately 0.62 and 0.92.

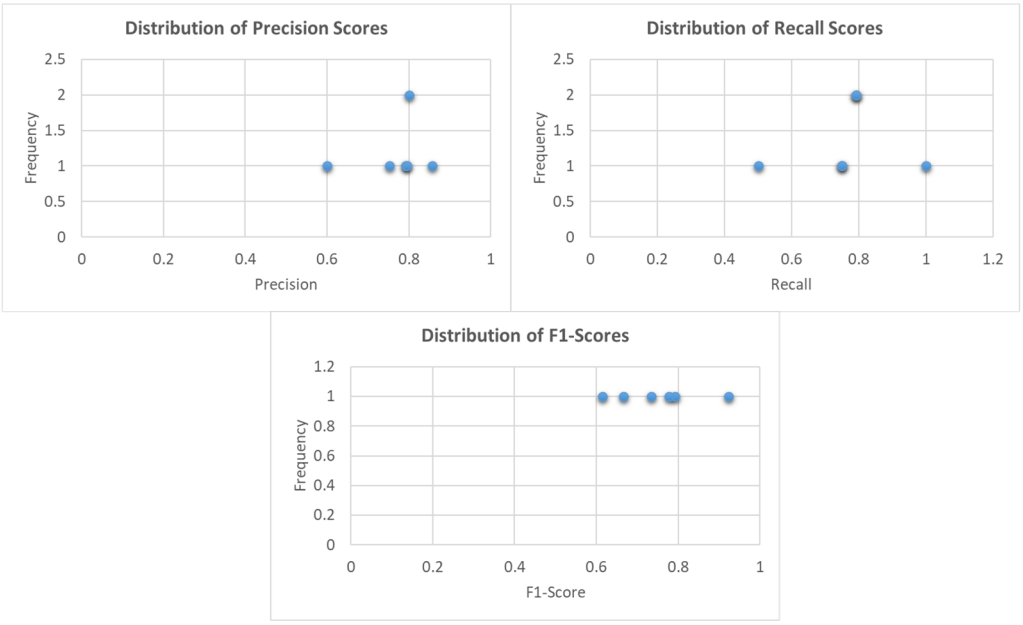

Fig. 5. Distribution of precision, recall, and F1-score across classes, highlighting class-wise performance variability.

The two-dimensional metric plots show clear relationships between recall and F1-score. Samples with

higher recall tend to achieve higher F1-scores. The lower-performing points mainly correspond to the natural environmental class.

VI. Discussion and Conclusion

The results show that the proposed CLASED model can effectively classify forest sounds, particularly wild animal calls, anthropogenic activity, and natural environmental events. The overall accuracy of nearly 79% indicates that the model can reliably distinguish these categories, even in challenging acoustic conditions with overlapping sounds and background noise.

Wild animal sounds were the easiest for the model to identify, with perfect recall and a high F1-score. These sounds tend to have distinct frequency patterns and repeated temporal structures, which the CNN–LSTM–Attention architecture can capture well. Anthropogenic sounds, such as chainsaws or machinery, were moderately well classified, but some misclassifications occurred, likely due to similarities with other environmental noises and the smaller number of samples. Natural environmental noises, such as wind, rain, and fire-crackling sounds, were the toughest to classify. A high level of overlap among these and other sounds resulted in lower recall values, implying that more samples or better features are needed for classification among these sounds.

From the training and validation curves, it is evident that the model learned comparatively without any signs of overfitting, indicating that the integration of the spectral, temporal, or attention model is effective for these types of audio signals. Despite the fact that the data is partially imbalanced, the model performed relatively well, which is indicative that this method is useful for actual forest monitoring.

Although this study focuses on the proposed hybrid architecture, a comparative evaluation with baseline models such as CNN-only and CNN–LSTM without attention could further quantify the contribution of each component. This remains an important direction for future work, along with systematic hyperparameter tuning to optimize model performance.

In the end, the CLASED framework offers a practical and flexible means for monitoring forest sounds. It works best when the sounds have some sort of pattern, but it still struggles with highly variable or overlapping environmental sounds. The future effort can be done by expanding the dataset to cover more regions and sound types, exploring better attention or feature extraction methods, or optimizing the model to be deployed on low-power devices. In general, this approach demonstrates the idea that spectral, temporal, and attention-based learning in combination can support automated real-time monitoring of forest ecosystems that can provide valuable insights for biodiversity assessment and environmental management.

Dataset Availability Statement

All datasets used in this study are publicly accessible. The ESC-50 dataset is available at https://github.com/karolpiczak/ESC-50 [20]. Wildlife audio recordings were obtained from the iNaturalist Sounds dataset (https://cvl-umass.github.io/iNatSounds/) [21] and the Xeno-canto bird sound repository (https://www.xeno-canto.org/) [22]. Forest-specific recordings from the FSC22 dataset are available at https://dx.doi.org/10.21227/40ds-0z76 [11].

Conflict Of Interest

The authors declare no conflict of interest.

Author Contributions

Maddhigalla Lakshumaiah conducted the research, including dataset preparation, audio preprocessing, feature extraction, model design and implementation, experimentation, and performance evaluation, and wrote the original draft of the manuscript. D. S. Rao supervised the study, contributed to the research design and methodological validation, guided the analysis and interpretation of results, and reviewed and edited the manuscript. All authors discussed the results and approved the final version of the manuscript.

References

[1] C. Sánchez-Giraldo, C. C. Ayram, and J. M. Daza, "Environmental sound as a mirror of landscape ecological integrity in monitoring programs," *Perspect. Ecol. Conserv.*, vol. 19, no. 3, pp. 319–328, Jul. 2021.

[2] A. Hutschenreiter, F. Aureli, A. Torres-Araneda, and E. Andresen, "Performance of four ecoacoustic indices as proxies of bird species richness in a neotropical forest," *Biodivers. Conserv.*, vol. 35, no. 1, Dec. 2025. doi: 10.1007/s10531-025-03208-5

[3] C. Yang, X. Liu, Y. Li, and X. Yu, "Deep learning-based multi-label classification for forest soundscape analysis: A case study in Shennongjia National Park," *Forests*, vol. 16, no. 6, May 2025. doi: 10.3390/f16060899

[4] S. Newey, P. Davidson, S. Nazir, G. Fairhurst, F. Verdicchio, R. J. Irvine, and R. van der Wal, "Limitations of recreational camera traps for wildlife management and conservation research: A practitioner's perspective," *Ambio*, vol. 44, no. 4, pp. 624–635, Oct. 2015.

[5] D. Meedeniya, I. Ariyarathne, M. Bandara, R. Jayasundara, and C. Perera, "A survey on deep learning based forest environment sound classification at the edge," *ACM Comput. Surv.*, vol. 56, no. 3, Mar. 2024. doi: 10.1145/3618104.

[6] D. Teixeira, P. Roe, B. J. van Rensburg, S. Linke, P. G. McDonald, D. Tucker, S. Fuller, "Effective ecological monitoring using passive acoustic sensors: Recommendations for conservation practitioners," *Conserv. Sci. Pract.*, vol. 6, no. 6, art. no. e13132, Jun. 2024.

[7] N. Mahankale, S. Gore, D. Jadhav, G. S. P. S. Dhindsa, P. Kulkarni and K. G. Kulkarni, "AI-based spatial analysis of crop yield and its relationship with weather variables using satellite agrometeorology," in *Proc. 2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)*, Mumbai, India, 2023, pp. 1–7. doi: 10.1109/ICACTA58201.2023.10392944.

[8] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, Nov. 2021. doi: 10.1038/s41598-021-01045-4

[9] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intell. Syst. with Appl.*, vol. 16, Nov. 2022. doi: 10.1016/J.ISWA.2022.200115

[10] D. A. Nieto-Mora, S. Rodríguez-Buritica, P. Rodríguez-Marín, J. D. Martínez-Vargaz, and C. Isaza-Narváez, "Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring," *Heliyon*, vol. 9, no. 10, Oct. 2023. doi: 10.1016/j.heliyon.2023.e20275

[11] M. Bandara, R. Jayasundara, I. Ariyarathne, D. Meedeniya, and C. Perera, "Forest sound classification dataset: FSC22," *Sensors*, vol. 23, no. 4, Feb. 2023. doi: 10.3390/S23042032
[12] J. Sharma, O. Granmo, and M. Goodwin, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network centre for artificial intelligence research department of information and communication technology," in *Proc. Interspeech 2020*, 2020, pp. 1186–1190. doi: 10.21437/Interspeech.2020-1303
[13] A. Ashurov, Y. Zhou, L. Shi, Y. Zhao, and H. Liu, "Environmental sound classification based on transfer-learning techniques with multiple optimizers," *Electronics*, vol. 11, no. 15, Jul. 2022. doi: 10.3390/ELECTRONICS11152279
[14] A. Kaçar, D. Avci, and İ. Türkoğlu, "Enhancing environmental sound classification performance through data fusion: A comparative machine learning analysis," Research Square, Nov. 2025. doi: 10.21203/RS.3.RS-7898377/V1
[15] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, pp. 896–903, Sep. 2021. doi: 10.1016/j.neucom.2020.08.069
[16] J. Guo, C. Li, Z. Sun, J. Li, and P. Wang, "A deep attention model for environmental sound classification from multi-feature data," *Appl. Sci.*, vol. 12, no. 12, Jun. 2022. doi: 10.3390/APP12125988
[17] R. Akter, M. R. Islam, S. K. Debnath, P. K. Sarker, and M. K. Uddin, "A hybrid CNN-LSTM model for environmental sound classification: Leveraging feature engineering and transfer learning," *Digit. Signal Process.*, vol. 163, 105234, Aug. 2025. doi: 10.1016/J.DSP.2025.105234
[18] K. Zaman, K. Li, M. Sah, C. Direkoglu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digit. Signal Process.*, vol. 158, 104956, Mar. 2025. doi: 10.1016/J.DSP.2024.104956
[19] T. Song and W. Zhang, "Frequency-aware convolution for sound event detection," *Lect. Notes Comput. Sci.*, vol. 15520, pp. 415–426, 2025. doi: 10.1007/978-981-96-2054-8_31
[20] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 2015 ACM Multimed. Conf.*, Oct. 2015, pp. 1015–1018.
[21] M. Chasmai, A. Shepard, S. Maji, and G. Van Horn, "The Naturalist Sounds Dataset," *Adv. Neural Inf. Process. Syst.*, 2024. doi: 10.52202/079017-4213
[22] Xeno Canto Repository. Accessed: Jan. 08, 2026. [Online]. Available: https://xeno-canto.org/

Maddhigalla Lakshumaiah received his B.Tech. degree from Sri Venkateswara University (SVU), Tirupati, India, in 2005, and his M.Tech. degree from Jawaharlal Nehru Technological University (JNTUA), Anantapur, India, in 2011. He has 15 years of teaching experience in engineering education. He is currently working as an Assistant Professor at KL University, India. His academic interests include teaching, curriculum development, and mentoring undergraduate and postgraduate students. He has been actively involved in academic activities and contributes to quality technical education.

D. S. Rao received his Ph.D. degree from the National Institute of Technology (NIT), Rourkela, and is currently a Professor in the Department of Computer Science and Engineering at the KLEF (KL Deemed to be University), Hyderabad Campus. He is an associate member of the Institute of Noise Control Engineering (INCE), USA. He has published more than 20 peer-reviewed research articles indexed in SCI and Scopus databases, along with 7 patents and 7 book chapters. His research contributions span soft computing, noise prediction, noise-induced hearing loss, and lung cancer prediction. He also serves as a reviewer for three SCI-indexed journals and is a member of the editorial boards of two academic journals. In recognition of his contributions, Dr. Rao has received several awards from various organizations, including those in the field of noise pollution and healthcare.