International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(1):1280-1298

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i01.01882

Bridging AI and Ecology: CILNN and XAI for Acoustic Based Prediction of Dangerous Wild Animals

Govindaprabhu GB[1]*, Sumathi M[2], Sharan Neyvasagam[3], Naveen Ananda Kumar J[4]

[1]Madurai Kamaraj University (MKU), Madurai, Tamilnadu, India, [2]Department of Computer Science, Sri Meenakshi Govt. Arts College for Women, Madurai, Tamilnadu, India, [3]Life point Health, TN, USA, [4]Tekinvaderz LLC, Florida, USA. *Corresponding Author's Email: prabhupri.pp@gmail.com

Abstract

In habitats that are encroaching on humans, human-wildlife conflict is an increasing global challenge. There is a significant risk of human injury and retaliatory action being taken if humans encounter dangerous animals. This work presents a novel approach to automated detection and classification of dangerous animals using audio signals, with a focus on model interpretability. This work introduces the Convolutional Interconnected Layer Neural Network (CILNN), a deep learning architecture designed to effectively process and classify animal vocalizations. Our method leverages a comprehensive set of audio features, including Mel-frequency cepstral coefficients (MFCCs) and spectral characteristics, optimized through SHAP-based feature selection. The CILNN incorporates interconnected layers and attention mechanisms to enhance feature extraction and model performance. It evaluates proposed approach on a diverse dataset of vocalizations from five dangerous animal species: bears, bison, cheetahs, elephants, and wild boars. Experimental results demonstrate that the CILNN outperforms traditional machine learning models such as Random Forests and Decision Trees in classification accuracy and robustness. Crucially, it employs Explainable AI (XAI) techniques, including SHAP values and decision tree visualizations, to interpret the decision-making processes of both our CILNN (90.6% accuracy) and other models. This interpretability analysis provides insights into feature importance and model behavior, enhancing trust and understanding in the classification process. Our work contributes to wildlife monitoring and human-wildlife conflict mitigation by offering an efficient, accurate, and interpretable method for acoustic-based animal detection.

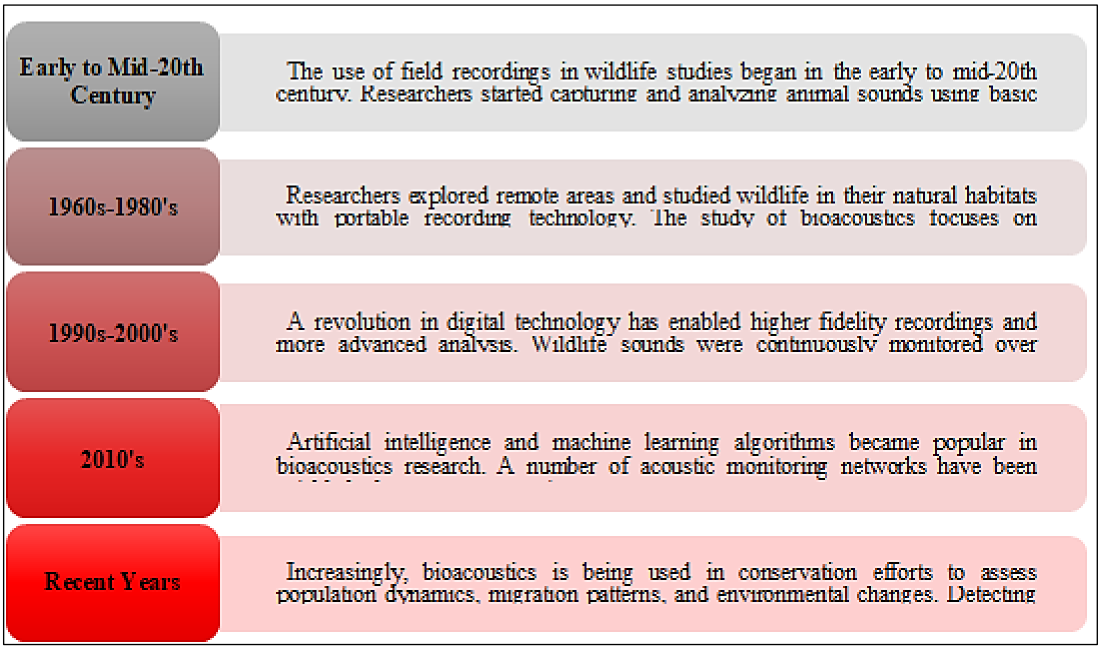
Keywords: Acoustic of Animals, CILNN, Dangerous Wild Animals, Explainable AI (XAI), MFCC, SHAP.

Introduction

A growing global challenge is human-wildlife conflict, especially in habitats that are encroaching on humans. Human encounters with dangerous animals pose significant risks to human safety and can also lead to retaliatory actions that threaten wildlife conservation. Natur e uses vocalizations to communicate, mating calls, to defend territory, and to warn. Voices can convey a wealth of information about an animal's state, intentions, and environment. Through acoustic monitoring, animal populations, behaviors, and movements can be studied and tracked without direct human intervention. Bioacoustics analyzes animal sounds based on biology and acoustics. The methods of modern acoustic monitoring typically involve recording animal vocalizations using specialized microphones. Analyzing these recordings using machine learning algorithms and signal processing techniques allows species to be identified, individuals counted, and patterns detected. Humans have used bioacoustics for situational awareness of dangerous animals since ancient times. During the early 20th century, bioacoustics emerged as a scientific field from basic observations by naturalists. Practical applications in conservation and safety replaced purely scientific study. In the late 20th and early 21st centuries, digital technology, machine learning, and artificial intelligence led to automated monitoring systems. Today's systems are highly effective for reducing wildlife conflict, preventing poaching, and enhancing tourism safety. Human-wildlife conflicts are intensifying because of habitat loss, population growth, and land use changes. Therefore, species that once lived in undisturbed habitats now navigate landscapes transformed by humans. Both wildlife

(Received 02nd August 2024; Accepted 30th January 2025; Published 31st January 2025)

conservation and human safety depend on understanding and addressing these conflicts. Conflicts between humans and wildlife pose a significant threat to communities and wildlife. Conflicts between humans and wildlife are becoming more frequent as human populations expand and encroach upon natural habitats. Figure 1shows the Time line of Animal Acoustic Detection. These conflicts can often be mitigated using physical barriers or human patrols, but these methods tend to be costly, labor-intensive, and ineffective. A scalable solution that provides early warnings of potential dangerous animal presence, thereby reducing the likelihood of harmful encounters, is urgently needed. There is a pressing need for innovative solutions to proactively prevent or mitigate conflict situations. Automated early warning systems stand out as a promising solution, providing timely alerts to communities and authorities, reducing the risk of encounters with potentially dangerous animals. Using artificial intelligence, such as deep learning algorithms, enables the system to recognize specific animal vocalizations and identify potentially dangerous situations. Automated methods ensure a faster and more accurate response to emerging conflict scenarios, allowing for timely intervention and risk mitigation.



The use of field recordings in wildlife studies began in the early to mid-20th century. Researchers started capturing and analyzing animal sounds using basic

Researchers explored remote areas and studied wildlife in their natural habitats with portable recording technology. The study of bioacoustics focuses on

A revolution in digital technology has enabled higher fidelity recordings and more advanced analysis. Wildlife sounds were continuously monitored over

Artificial intelligence and machine learning algorithms became popular in bioacoustics research. A number of acoustic monitoring networks have been

Increasingly, bioacoustics is being used in conservation efforts to assess population dynamics, migration patterns, and environmental changes. Detecting

Figure 1: Time line of Animal Acoustic Detection

Artificial intelligence and audio signal processing advances enable machine listening techniques to address this challenge. By identifying acoustic signatures, these techniques provide a cost-effective, continuous monitoring solution that is noninvasive and noninvasive. Real-world applications of these techniques face several challenges:

- Complexity of natural soundscapes: As wildlife vocalizations occur in a cacophony of environmental sounds, robust algorithms are needed to isolate and identify species.
- Variability in animal vocalizations: Variations within species caused by factors with precise, adaptable classification models.
- Limited labeled data: It is challenging to obtain large, accurately labelled datasets of dangerous animal vocalizations, especially for rare or elusive species.
- Interpretability of AI models: The decision-making process of these models must be transparent and explainable to allow them to be deployed practically and gain user trust.

An automated acoustic detection and classification of dangerous animals is presented in this study to overcome these challenges. Our key contributions are:

- Development of the Convolutional Interconnected Layer Neural Network (CILNN): New deep learning architecture for processing and classifying animal vocalizations.
- Comprehensive feature engineering: In our approach, SHAP-based feature selection is used to select Mel-frequency cepstral

coefficients (MFCCs) and spectral characteristics from audio signals.

- Integration of Explainable AI (XAI) techniques: By incorporating SHAP values and decision tree visualizations, it enhances the transparency of the classification process using both our CILNNs and traditional machine learning approaches.
- Comparative analysis: The CILNN is compared to traditional machine learning models (Random Forests and Decision Trees) on a diverse dataset of vocalizations from five dangerous animal species: bears, bison, cheetahs, elephants, and wild boars.
- Real-world applicability: Our system in wildlife monitoring and human-wildlife conflict mitigation is discussed in terms of potential deployment scenarios and challenges.

Considering these aspects will help develop practical, interpretable tools for monitoring and management of wildlife by advancing the field of bioacoustics. Through non-invasive monitoring techniques, the proposed approach could significantly enhance human safety and support conservation efforts in wildlife-rich areas. The field of bioacoustics and AI has significantly transformed wildlife monitoring. Animal vocalizations can now be precisely classified using sophisticated DL models, such as CILNN. The use of Explainable AI (XAI) techniques has increased model transparency. Enhanced feature extraction techniques and attention mechanisms allow models to adapt to species-specific variability. Using these innovations, humans and wildlife can coexist efficiently and non-invasively. The research work combined sound analysis with linear prediction coding and artificial neural networks to detect stress vocalizations in noisy pig units with few recognition errors (<5%) (1). STREMODO (Stress monitor and documentation unit) is insensitive to noise, human speech, and pig vocalizations other than screams. Various farming environments can use it routinely as an objective, non-invasive measure of acute stress. A study proposed that African elephant vocalizations reflect emotional intensity (2). Researchers examined four adult female rumbles in different social contexts at Disney's Animal Kingdom to determine whether they varied between negative (dominance interactions) and neutral (minimal social activity) situations. It appears that negative social contexts elicit higher intensity vocalizations with specific acoustic features, while positive contexts show similar but less pronounced effects. The review work discusses current trends in sound analysis. This is followed by a description of three important farm livestock species: chickens (Gallus gallus domesticus), pigs (Sus scrofa domesticus), and cattle (Bos taurus) (3). This method has the potential for developing automated methods for large-scale farming to monitor animal welfare. The novel study was to investigate whether wild boar calls (i.e., grunts, screams, or squeals) change depending on their emotional state. Positive and negative situations resulted in different types of calls (4). As emotions changed, their acoustic structure changed as well. Generally, positive calls are shorter and lower frequency than negative calls. Thus, wild boars seem to express their emotions through their vocalizations. Various attempts have been made to decipher the meaning of farm animal vocalizations in recent years (5). A review of the current state-of-the-art is given in this discipline focusing on important farm animal species (pigs, cattle, and poultry) as well as current problems and future developments. Modern sound analysis techniques can discriminate, analyze, and classify specific vocalizations. AudioMoth is a low-cost, small, full-spectrum alternative described by the work (6). The device consists of a printed circuit board, a microcontroller, and a microphone. By combining its small size and a simple mechanism, this device can be retrofitted into many low-cost ruggedized enclosures for deployment in remote locations; 1. Long-term monitoring with low-power operation; 2. Modular expansion with easy access general purpose input and output pins; and acoustic detection with onboard processing power. An audio processing workflow combining automated detection and human review was described (7). By using this workflow, it reduces human effort by more than 99%. The Shiny package provides a user-friendly way to run the neural network through RStudio, creating a portable and portable workflow for field biologists. This work focused animals can identify their species or breed by the sound they produce during vocalizations, even if the sound is similar to unaided human ears (8). In order to test this hypothesis, three artificial neural networks (ANNs) were developed to automatically identify 13 bird species, eight dog breeds, and 11 frog species using bioacoustics properties. By tenfold cross-validation, the converted values of the vocalizations and breed or species identifications were used to train the ANNs. The respective ANNs correctly identify 71.43% of birds, 94.44% of dogs, and 90.91% of frogs. The recent work presents an end-to-end feedforward convolutional neural network that reliably classifies source and type of animal calls (9). It two streams of audio data in a noisy environment with imperfect labels and modest datasets. Several cages of captive marmoset monkeys were nearby, with their audio recordings. Using audio-specific feature extraction techniques and machine learning models present a multi-purpose livestock vocalization classification algorithm (10). As part of testing the algorithm's multi-purpose nature, three separate data sets were created targeting sheep, cattle, and Maremma sheepdogs. Several continuous recordings were conducted at three different operational farming enterprises to reflect real-world conditions. All data sets were highly accurate (sheep: 99.29%, cattle: 95.78%, dogs: 99.67%). This work focused the effectiveness of the UOZ-1, which emits the natural warning calls of animals, in protecting animals living near railway tracks (11). Two study sites along the E-20 line where UOZ-1 devices had been installed were investigated between 2008 and 2012. Rail transport was not observed in 76% of observations. Most wild mammals escaped when a train approached and acoustic signals were emitted (93–85% of cases).In the study investigated factors that may affect microchipping and neutering decisions among dogs and cats (12). An analysis of 1047 valid responses was conducted using the non-parametric Chi-Square test, among companion animal guardians in Portugal. The latest work proposed sensor-based IoT system using Arduino and sensors. A Short Message Service (SMS) can also be sent to the farm owner through GSM (13). It also allows the farmer to control the entry of the animal into the farm field automatically or manually from home. Through the developed system, the farm owner can receive live pictures of the animals via telegram bot, so they can protect them. This work detected the animal's presence, identify the animal, and divert the identified animal away from the field (14). An infrared passive sensor detects the presence of animals and a sound analysis system identifies them based on their sound. Using a bright light-emitting diode (LED) that works only in dark environments, specific ultrasonic sounds will be generated to irritate the identified animal. By using machine learning, these recent works solved challenges animal intrusion detection and reach the objectives (15). Regularly taken pictures of the entire farm were taken in their study. Using the Watershed technique, it analyzed the photos. 2D Gabor filter banks are used to retrieve training set features. The Support Vector Machines method is used for classification.

Research Gap

There are several gaps in existing knowledge and methods for acoustic-based wildlife monitoring. A major challenge is identifying specific animal vocalizations within complex soundscapes where multiple overlapping sounds overlap. In noisy environments, existing models struggle with such variability. By incorporating advanced feature extraction and attention mechanisms, the CILNN enhances the model's robustness to diverse acoustic conditions. Traditional models suffer from a lack of labeled data for endangered species, which hampers their performance. By using SHAP-based feature selection, this work mitigates this issue. The model performs well even with smaller datasets. AI models lack interpretability, which limits their adoption in wildlife conservation. With Explainable AI (XAI), this work provides transparency into decision-making.

Handling Complex Soundscapes: Many current systems struggle in noisy environments or with multiple species vocalizing simultaneously. It is critical to improve AI's ability to separate and identify species in complex acoustic environments.

Rare and Endangered Species: The training data for rare or endangered species is often limited. This gap needs to be filled by developing techniques that accurately identify these species.

Handling Variations within Species: There are many regional dialects and individual variations among species. It is important to improve AI's ability to handle intra-species variability.

Explainable AI: Ecologists and conservation biologists would be more inclined to trust and adopt models that explain their decision-making process.

Methodology

A novel XAICILNN based framework for acoustic detection and classification of dangerous animals is proposed, addressing the challenges of complex soundscapes and the need for interpretable AI in wildlife monitoring. Our approach relies on Convolutional Interconnected Layer Neural Networks (CILNN), a deep learning architecture specifically designed for processing audio signals of animal vocalizations. Figure 2 shows the framework of the Proposed Work.

There are several key components to the methodology:

- Audio Dataset Preparation: This work collects vocalizations from five dangerous animals: bears, bison, cheetahs, elephants, and wild boars.
- Spectrogram Analysis: Spectrogram analysis is used to visualize and understand each species' acoustic signature.
- Feature Extraction: This phase capture the nuanced aspects of animal vocalization, 30 audio features are extracted, including Mel-frequency cepstral coefficients (MFCCs).
- SHAP-based Feature Selection: SHAP (SHapley Additive Explanations) values are used for feature selection to optimize model performance.
- CILNN Architecture: This work incorporates convolutional layers and attention mechanisms to enhance feature extraction and improve classification accuracy,
- Explainable AI Integration: The model's decision-making process is analyzed using XAI techniques, including SHAP value analysis and decision tree visualizations.

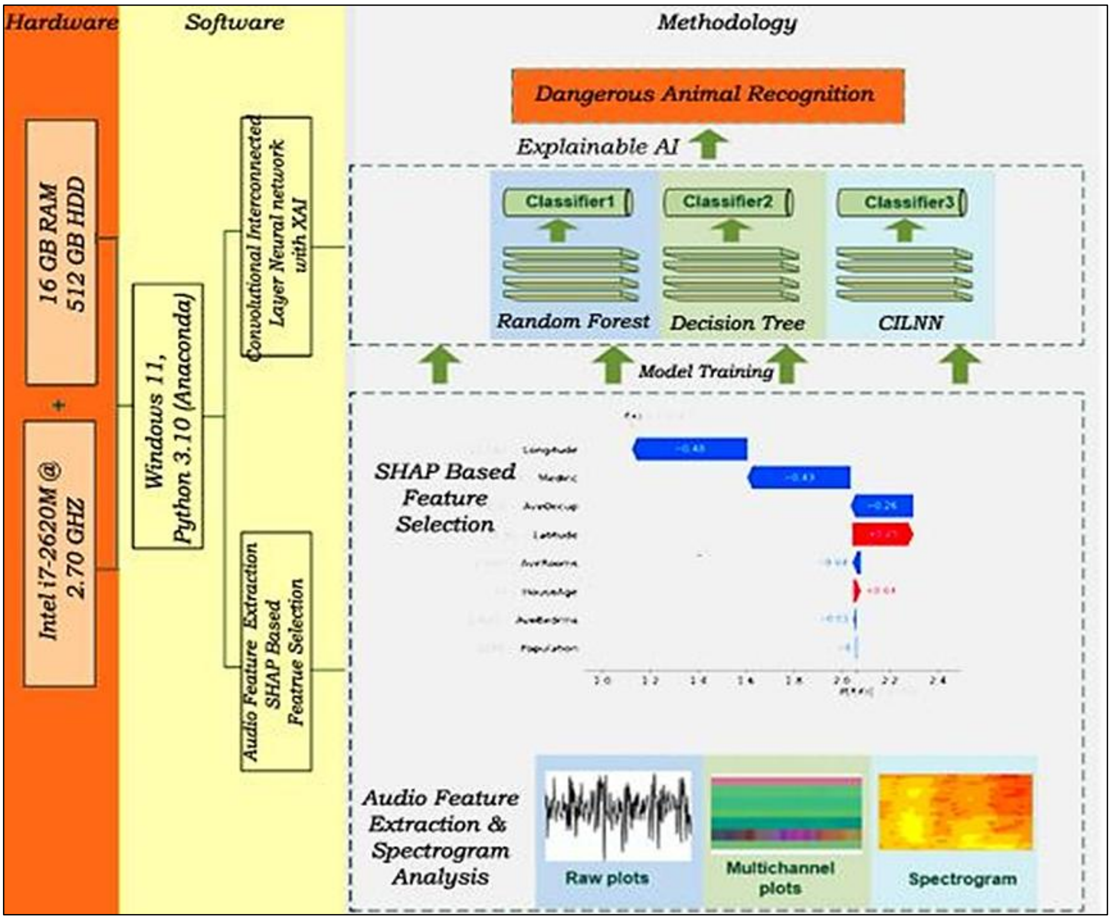
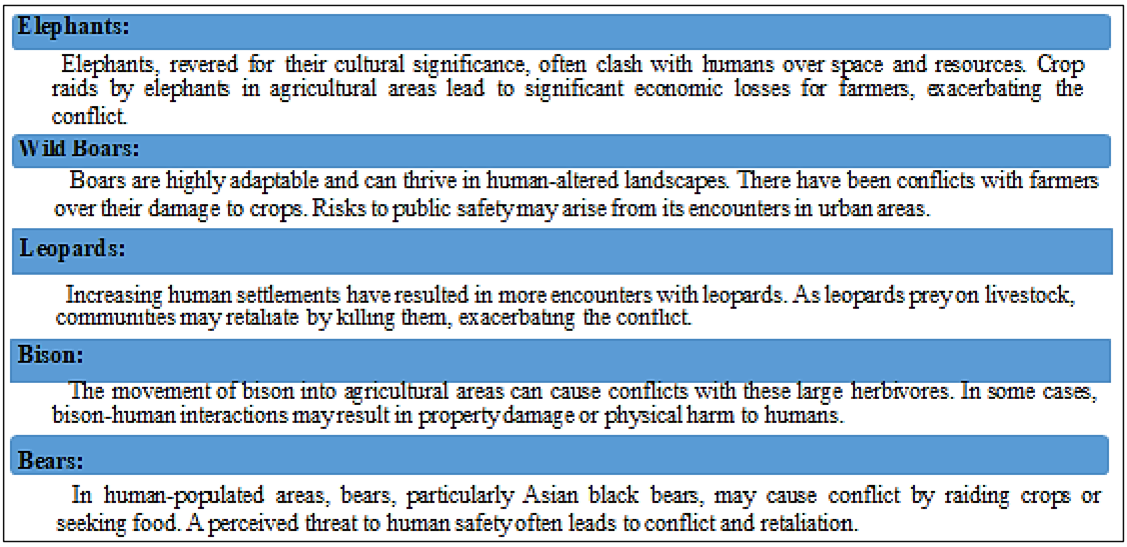


Figure 2: XAICILNN Based Framework of the Proposed Work

Dangerous Animal Audio Dataset

In order to collect the audio dataset, five dangerous animals were used: bears, bison, cheetahs, elephants, and wild boars. It includes open platforms like YouTube and publicly available datasets (16). Selection of audio samples was based on capturing unique vocalizations of each species. During preprocessing, the duration of the audio files was standardized to maintain uniform temporal resolution. Clear vocalization, minimal background noise, and representativeness of the species' typical sounds were criteria for selection. By using this approach, we were able to provide a diverse but high-quality dataset for accurate training and testing. Figure 3 shows the various dangerous animals list.



Elephants:

Elephants, revered for their cultural significance, often clash with humans over space and resources. Crop raids by elephants in agricultural areas lead to significant economic losses for farmers, exacerbating the conflict.

Wild Boars:

Boars are highly adaptable and can thrive in human-altered landscapes. There have been conflicts with farmers over their damage to crops. Risks to public safety may arise from its encounters in urban areas.

Leopards:

Increasing human settlements have resulted in more encounters with leopards. As leopards prey on livestock, communities may retaliate by killing them, exacerbating the conflict.

Bison:

The movement of bison into agricultural areas can cause conflicts with these large herbivores. In some cases, bison-human interactions may result in property damage or physical harm to humans.

Bears:

In human-populated areas, bears, particularly Asian black bears, may cause conflict by raiding crops or seeking food. A perceived threat to human safety often leads to conflict and retaliation.

Figure 3: Dangerous Animals List

Spectrogram Analysis and Observation

This work involves spectrogram analysis to visualize and interpret the acoustic signatures of bears, bisons, cheetahs, elephants, and wild boars. The spectrum provides a visual representation of the frequency content of audio signals over time, allowing us to observe unique patterns in animal vocalizations. The Short-Time Fourier Transform (STFT) is used to generate spectrograms for each species.

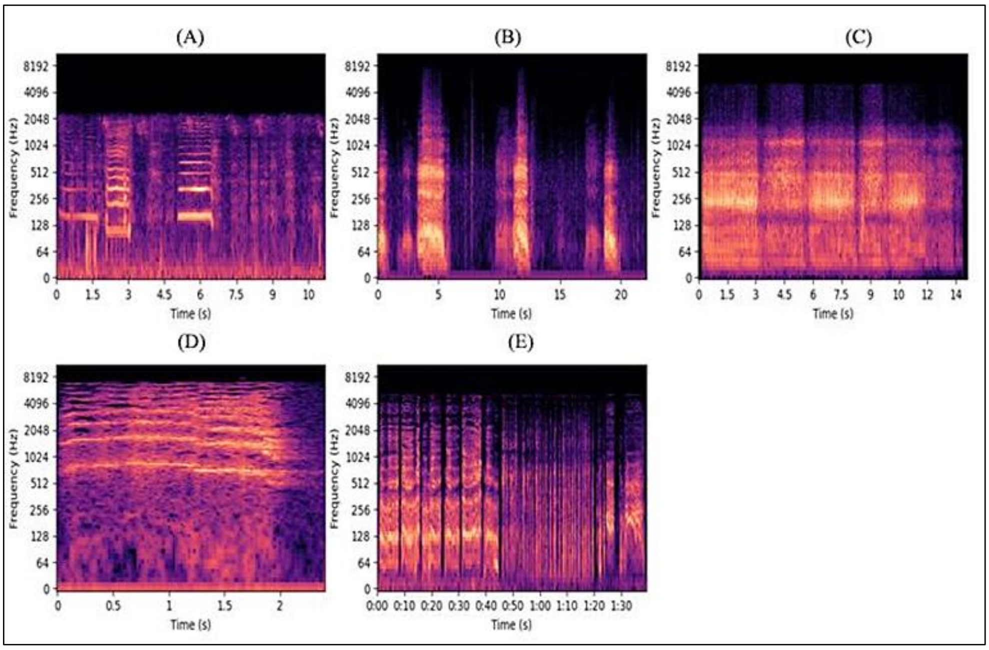


Figure 4: Various Animals Acoustic Frequencies: (A) Bear (B) Bison (C) Cheetah (D) Elephant (E) Wild Boar

In Figure 4, spectrograms of five different animal's vocalizations are shown: bear, bison, cheetah, elephant, and wild boar. Each spectrogram shows the frequency content of an animal's sound over time. Color represents the intensity of the sound at each time and frequency, with yellow and red indicating higher intensities and purple and black indicating lower intensities. Bears' vocalizations show vertical streaks, suggesting short, sharp sounds in their spectrograms. Longer vocalizations are indicated by the bison's spectrogram. Cheetahs' sounds are horizontal streaks, possibly chirps or calls. Elephant spectrograms show horizontal bands, possibly corresponding to low-frequency rumbles. There is a series of grunts and snorts in the wild boar's vocalization. The visualizations illustrate the diversity of animal vocalizations across different species by analyzing and comparing their acoustic characteristics. Table 1 shows the Frequency, Loudness, Energy and observation of the animal acoustic.

Table 1: Animal sound Frequency and Observation

Animal	Dominant Frequency	Loudness (Mean Amplitude)	Energy	Observation
Bear	703.12 Hz	0.14	999021.19	The energy level is moderate, typical for bear vocalizations. The sound is relatively quiet, possibly indicating a distant bear vocalization.
Bison	193.80 Hz	0.14	1932787.25	The energy level is moderate, typical for Indian bison vocalizations. The sound is relatively quiet, possibly indicating a distant Indian bison vocalization.
Cheetah	515.62 Hz	0.44	4412070.50	The energy level is high, suggesting sharp and intense cheetah vocalizations. The sound has a moderate loudness, consistent with typical cheetah vocalizations.
Elephant	575.62 Hz	1.66	2799742.75	The energy level is moderate, consistent with typical elephant vocalizations. The sound is relatively loud, which aligns with the known powerful vocalizations of elephants.
Wild Boar	281.25 Hz	0.53	4908632.00	The energy level is high, suggesting intense and potentially close wild boar vocalizations. The sound has a moderate loudness, consistent with typical wild boar vocalizations.

Audio Feature Extraction

There are three main steps to preparing audio files for analysis. A library like pydub decodes MP3 files into uncompressed audio samples, which are then converted to a raw audio format like WAV. After that, the sampling rates are standardized (e.g., 22050 Hz), ensuring uniform temporal resolution across the dataset using librosa. With libraries like librosa, audio features are extracted. Each of these 30 features captures an aspect of the audio, such as MFCCs, chroma, and spectral centroid. Every audio file in the dataset is processed this way, resulting in a consistent set of features that can be used for further analysis.

MFCCs (Mel-Frequency Cepstral Coefficients): MFCCs are derived by taking the Fourier transform and mapping the powers of the spectrum onto the Mel scale. Mel scales are based on human pitch perception, making MFCCs (13 features) particularly useful for speech and music analysis. Energy is represented by the first coefficient, while spectral details are represented by the higher coefficients.

Chroma Feature: Chroma features are calculated by: a) Computing the spectrogram b) Mapping the frequencies to the 12 pitch classes c) Summarizing the energy in each pitch class.

Spectrogram: The spectrogram is computed with the Short-Time Fourier Transform (STFT). An overlapping signal is segmented and windowed (usually with Hann windows). For each windowed segment, the FFT is computed, and the magnitude is squared to get the power spectrum. The spectrogram shows how the signal changes in frequency over time.

Spectral Bandwidth: It is calculated by weighting the standard deviations around the spectral centroid. It measures a sound's "spread" and can indicate the sound's noise or harmonic content.

Harmonic-to-Noise Ratio (HNR): In order to calculate HNR, this work separates the signal into harmonic and noise components. HNR is calculated based on the energy of each component and the ratio between harmonic energy and noise energy. A higher HNR indicates a more tonal sound, while a lower HNR indicates a noisier sound.

Tonnetz Features: The Tonnetz (tone network) represents tonal space geometrically. This space is derived from chroma features, which can reveal harmonic relationships that aren't apparent in raw chroma information.

Pitch: Analysis of autocorrelations or cepstrums is often involved in pitch extraction. A probabilistic algorithm is used by the librosa track function, which is more robust to noise than a simple autocorrelation.

Constant-Q Transform (CQT): Unlike the STFT, which has linearly spaced frequency bins, the CQT uses logarithmically spaced bins. This corresponds to musical scales, which double in frequency with every octave. All bins have the same Q-factor (ratio of center frequency to bandwidth).

STFT (Short-Time Fourier Transform): The STFT is calculated similarly to the spectrogram, but without taking into account the magnitude. For tasks like phase coding or source separation, it retains phase information.

Spectral Centroid: This is calculated by weighing the magnitudes of the frequencies present in the signal. A sound's perceived brightness is correlated with it.

Spectral Bandwidth: Calculated as the weighted standard deviation of frequencies around the spectral centroid. Information about the spectrum's shape is provided by it.

Spectral Flux: Normalized spectra are typically calculated using the Euclidean distance between them. Because it detects sudden changes in the spectrum, it is useful for detecting onsets.

Spectral Rolloff: In this case, a certain percentage (usually 85%) of the spectral energy lies below a certain frequency. The skewness of the spectral shape can be indicated by it.

Spectral Flatness: This is calculated as the ratio of the geometric mean to the arithmetic mean of the spectrum. A high flatness (close to 1) indicates a noisy signal, while a low flatness indicates a tonal signal.

Table 2: Sample Features

Chroma_ Feature	Stft	Spectral_ Rolloff	Spectral_ Flatness	hnr	Harmonic	Percussive	Class
-418.786	1.164482	-5.74144	-0.15534	-7.80E-10	-0.09633	-5.83481	Bear
-253.413	4.499015	-3.61844	-4.78478	-5.03E-07	-2.65346	-5.56889	Bear
-366.548	3.333036	-5.03716	-8.52473	-0.00057	3.637412	-3.20641	Bear
-234.422	-26.0454	-20.7128	4.981143	-7.02E-07	-5.02228	-8.60392	Bear

Table 2 shows the sample values of the customer generated features. Animal vocalizations can be classified using features like MFCCs, chroma features, and spectral characteristics. MFCCs are highly effective for analyzing sound frequency components because they closely mimic human auditory perception. As MFCCs represent both energy and spectral details of audio signals, they can distinguish subtle variations in animal vocalizations. The chroma features of audio signals were used to summarize the energy distribution across the 12 semitone classes. When vocalizations exhibit tonal patterns, such as specific calls or chirps, they can assist in species identification. Sound spectrum characteristics, such as spectral centroid, bandwidth, roll-off, and flatness, provide insight into the overall shape and distribution of sound. A robust classification relies on identifying features like brightness, harmonic structure, and noisiness, which vary widely between species. As a result, the model understands audio signals in both frequency and temporal domains. It improves its ability to classify complex animal vocalizations.

SHAP Based Feature Selection

Feature selection significantly reduces the training time in the dataset when the model is applied to it. In this sense, the training time without feature selection and that with feature selection were compared to understand the impact of feature selection. SHAP (SHapley Additive Explanations) values identify important features in a dataset by quantifying their contribution to model predictions (17). Based on game theory, SHAP values assign "credit" for a prediction among the input features. For animal sound classification, it helps to identify which audio features, such as MFCCs, chroma features, and spectral characteristics and influence species differentiation. It provides insight into how different features interact and affect model predictions by calculating the marginal contribution of each feature. In this process, features are selected to enhance model accuracy. The dataset is reduced to the most relevant and impactful features. Using essential features improves the model's generalization and reduces overfitting. It improves the model's interpretability and aid in further refining the model by visualizing feature importance. The SHAP tool allows understanding machine learning models using game theory. Shapley values from game theory and their extensions are used to connect optimal credit allocation with local explanations. As a model agnostic XAI, SHAP can be applied to any model post-training. In SHAP values, each prediction is explained by the features of the dataset contributing to the model's output. Accordingly, SHAP approximates Shapley values, a concept from game theory that solves the problem of computing the contribution of each subset of features to a model's prediction given a dataset with m features. Despite the exponential nature of the problem, SHAP approximates the Shapley value solution using weighted linear regression for all models or ensemble tree models with different assumptions about feature dependence. In linear regression models, the coefficients used to weight the features are used to explain all predictions, but they don't account for individual data points' heterogeneity. The effect of a feature on a data point may differ from that on another data point. This is consistent with local explanations being more accurate than global ones. Similarly, non-linear dimensionality reduction methods estimate global similarities through local similarities. In SHAP, local explainability is explored and used to build surrogate models for black boxes. Then, SHAP tests the change in prediction by slightly changing the input value for a feature, and if the prediction doesn't change much, the feature for that data point may not be an important predictor. Figure 5 shows the generation of the SHAP Features.

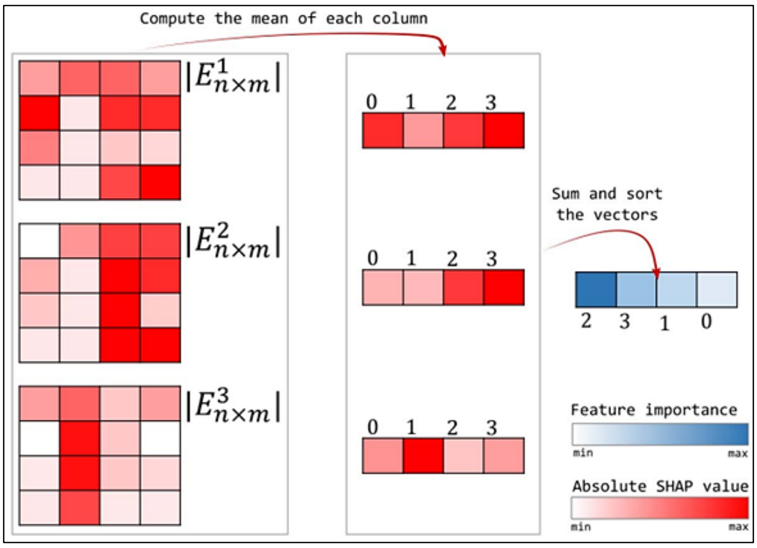


Figure 5: Generation of SHAP Features

The feature information shown in Figure 5, this work evaluates SHAP as a feature selection approach. To begin, a SHAP values matrix is generated for each class (c) in the dataset, which encodes the features that contribute to each data point, and then the mean of the columns of each matrix is calculated. Each class's mean SHAP values are summed and ordered in decreasing order. First, the most important feature appears in the first position, the second feature appears in the second position (13).

```
Input: Selected Feature Set X, Class Set C
Output: Ranked Feature Set R, Predicted Output (f_x)
Algorithm: SHAP Based Feature Selection
Train a base model M on the full feature set X
Set number of iterations K for SHAP value estimation
Initialize empty list R for ranked features
For each feature j in X:
        Initialize S_j = 0 (SHAP value for feature j)
For k = 1 to K:
Randomly select a subset of features $S \subseteq X \setminus \{j\}$
Predict $f\_x(S \cup \{j\})$ using model M
Predict f_x(S) using model M
Calculate marginal contribution: $MC\_k = f\_x(S \cup \{j\}) - f\_x(S)$
Update S_j: S_j += MC_k
$\varphi\_j = S\_j / K$
R = sort (|φ_j| for j in X, descending)
Select top N features from R based on a threshold or desired number of features
Retrain model M' on selected features
For each new input x':
f_x = M'(x')
Ranked feature set R and SHAP values φ for interpretation
End Algorithm
```

Algorithm 1: SHAP Based Feature Selection

Algorithm 1 shows the SHAP Based Feature Selection and Figure 6 shows the SHAP Based Selected Features. HNR, MFCC, Chroma Features are the important features in the selection.

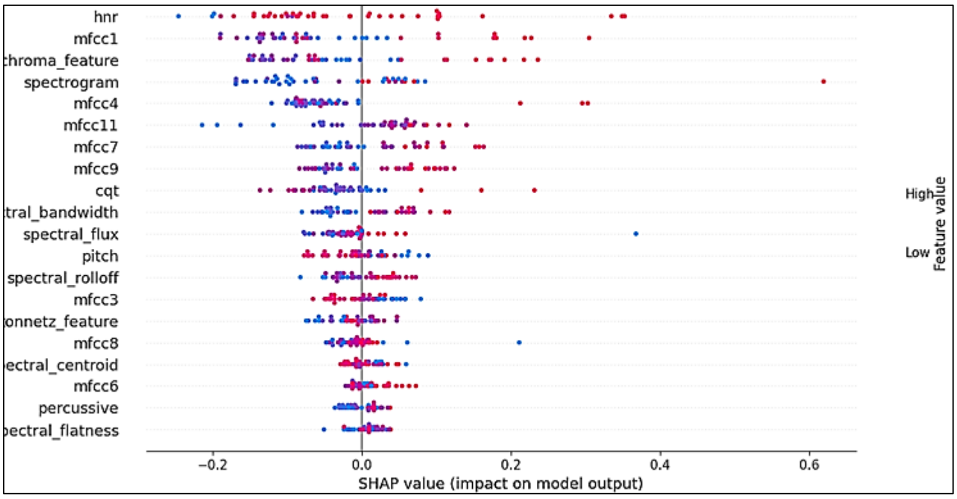


Figure 6: SHAP Based Selected Feature

CILNN - Convolutional Interconnected Layer Neural Network

Numerous applications depend on audio classification, ranging from speech recognition to music genre classification. Traditional CNNs have been successful in this domain, learning hierarchical feature representations from raw audio data. Nevertheless, these models face challenges in capturing complex patterns and long-range dependencies. To overcome such limitations, this work proposes the Convolutional Interconnected Layer Neural Network (CILNN), which incorporates convolutional layers and attention mechanisms. CILNNs begin with a series of convolutional layers that extract low-level audio features. Layers are designed to capture local patterns such as frequency components and temporal dynamics. A max-pooling layer follows each convolutional layer to reduce spatial dimensions. In CILNN, low-level features are extracted from audio input through a series of convolutional layers. In these layers, frequency components and temporal dynamics are captured. In order to reduce the spatial dimensions and retain the most salient features, each convolutional layer is followed by a max-pooling layer.

$$Y[i,j] = \sum_{m=0}^{K-1} \sum_{n=0}^{k-1} X[i+m, j+n].W[m,n] + b$$

$$Y = Output\ feature\ map, W = Filter, b = Bias, K = Filter\ size$$

Through the use of interconnected layers, parallel convolutional paths can be introduced to enhance feature extraction. Using these paths, the model learns complementary features while applying different convolutional filters, capturing a broader range of patterns. Feature representation is enhanced by concatenating these parallel paths. X is the input feature map of shape H×W×C, where H and W are the height and width of the feature map, and C is the number of channels. Two parallel convolutional layers are applied to the input feature map X. The convolutional filters of the two parallel paths are f1 and f2. Y1 and Y are respectively the outputs of these convolutional operations.

$$Y1 = f1(X)\ and\ \ Y2 = f2(X)$$

Where f1and f2represent convolution operations defined as:

$$Y_1[i,j,k] = \sum_{m=1}^{H\prime} \sum_{n=1}^{W\prime} \sum_{C=1}^{C} X[i+m, j+n, c].W_1(m,n,c,k)$$

$$Y_2[i,j,k] = \sum_{m=1}^{H\prime} \sum_{n=1}^{W\prime} \sum_{C=1}^{C} X[i+m, j+n, c].W_2(m,n,c,k)$$

Where W1and W2are the weights of the convolutional filters f1 and f2, respectively, and H ′and W′ are the height and width of the convolutional filters. The outputs Y1 and Y2 are concatenated along the channel dimension to form the combined feature map Y

$$Y = Concatenate(Y1, Y2)$$

The CILNN incorporates attention mechanisms to focus on the most relevant parts of the feature maps. Attention is focused on the outputs of the parallel convolutional paths in the interconnected layers. Using this method, the model can emphasize important features and reduce the impact of irrelevant information. By generating attention weights, the attention mechanism scales feature maps and highlights significant areas. Convolution and interconnected layers are followed by flattening of the feature maps and passing them through fully connected layers. With the help of these layers, high-level reasoning and classification are carried out based on the extracted features. The dropout technique is applied to these layers to prevent over fitting and improve generalization. Using softmax activation, the final output layer generates class probabilities based on the input data. Final output layers produce class probabilities using soft-max activation functions. A target class in the audio classification task corresponds to the number of units in this layer. This CILNN architecture has been improved in the following ways:

- Interconnected layers: The model incorporates interconnected layers that connect the outputs of different convolutional layers.
- Attention mechanism: A second interconnected layer includes an attention mechanism that aids in classification by focusing on the most relevant features.
- Global average pooling: This method reduces parameter numbers and enables spatial information to be aggregated.
- Fully connected layers with dropout: Adding fully connected layers with dropout prevent over fitting and helps learn higher-level representations.

Input: Selected Feature Set X, Class Set C
Output: Predicted Output (f_x)
Algorithm: Convolutional Interconnected Layer Neural Network (CILNN)

```
Define input layer $X \in R^{\wedge}(H \times W \times C)$
Initialize convolutional layers with filters $W\_l \in R^{\wedge}(K \times K \times C\_in \times C\_out)$
Set up interconnected layers with parallel paths
Initialize attention mechanism weights $A \in R^{\wedge}(H \times W \times C)$
Define fully connected layers with weights W_fc and biases b_fc
Set up output layer for |C| classes
For each convolutional layer l:
    $$Y[i,j] = \sum_{m=0}^{K-1} \sum_{n=0}^{k-1} X[i+m, j+n].W[m,n] + b$$
Apply activation: Y_l = ReLU(Y_l)
Apply max pooling: Y_l = MaxPool(Y_l)
For interconnected layers:
    $$Y_1[i,j,k] = \sum_{m=1}^{H\prime} \sum_{n=1}^{W\prime} \sum_{C=1}^{C} X[i+m, j+n, c].W_1(m,n,c,k)$$
    $$Y_2[i,j,k] = \sum_{m=1}^{H\prime} \sum_{n=1}^{W\prime} \sum_{C=1}^{C} X[i+m, j+n, c].W_2(m,n,c,k)$$
Y = $Concatenate(Y1, Y2)$
A = $\sigma\ (W\_a \cdot Y + b\_a)$
Y_att = A ⨀ Y
Flatten: Y_flat = Flatten(Y_att)
For each fully connected layer:
Z = $W\_fc \cdot Y\_flat + b\_fc$
Y_fc = ReLU(Z)
Apply dropout: Y_fc = Dropout (Y_fc, p)
P = softmax (W_out · Y_fc + b_out)
Compute cross-entropy loss: $L = -\sum\_(i = 1)^{\wedge}|C|\ y\_i\ log(P\_i)$
Calculate gradients: $\nabla W = \partial L / \partial W$
Update weights: W = W - η∇W (using optimizer, e.g., Adam)
Perform forward pass on new input X_new
P = argmax_c(P_c)
Apply SHAP values: $\varphi\_j = \sum\_(S \subseteq F\setminus\{j\})\ (|S|!\ (|F| - |S| - 1)!)/(|F|!)\ (f\_x(S \cup \{j\}) - f\_x(S))$
where F is the set of all features, and **f_x** is the model output for input x
End Algorithm
```

Algorithm 2: Convolutional Interconnected Layer Neural Network (CILNN)

Compared to other state-of-the-art deep learning models for audio classification, CILNN's performance would be more comprehensively evaluated. CNNs, LSTMs, and hybrid CNN-LSTM architectures are commonly used in audio classification tasks and could serve as benchmarks. They capture spatial and temporal patterns in audio signals. For example, CNNs excel at extracting features from spectrograms, while LSTMs model sequential dependencies in time-series data, making them suitable for processing dynamic vocalizations. A comparison of CILNN with these models might reveal areas where it might fall short, such as handling complex temporal dependencies (Algorithm 2, Table 3).

Table 3: CILNN Parameters

Layer Type (Parameters)	**Parameter Values**
Input Layer	Defined by the input shape of the model
Conv2D (First Block)	32 filters, (3, 3) kernel, ReLU
MaxPooling2D (First Block)	(2, 2)
Conv2D (Second Block)	64 filters, (3, 3) kernel, ReLU
MaxPooling2D (Second Block)	(2, 2)

Conv2D (Parallel Block 1)	64 filters, (3, 3) kernel, ReLU
Concatenate (Block 1)	Original output + Parallel Conv1 layer
Conv2D (Third Block)	128 filters, (3, 3) kernel, ReLU
MaxPooling2D (Third Block)	(2, 2)
Conv2D (Parallel Block 2)	128 filters, (3, 3) kernel, ReLU
Conv2D (Attention)	1 filter, (1, 1) kernel, Sigmoid (for attention mechanism)
Multiply (Attention)	Parallel Conv2 features + Attention layer
Concatenate (Block 2)	Original output + Attended Features
Conv2D (Fourth Block)	256 filters, (3, 3) kernel, ReLU
GlobalAveragePooling2D	Average pooling over all spatial dimensions (height and width)
Dense (Fully Connected 1)	256 units, ReLU
Dropout	0.5
Dense (Output Layer)	num_classes (Speices), Softmax

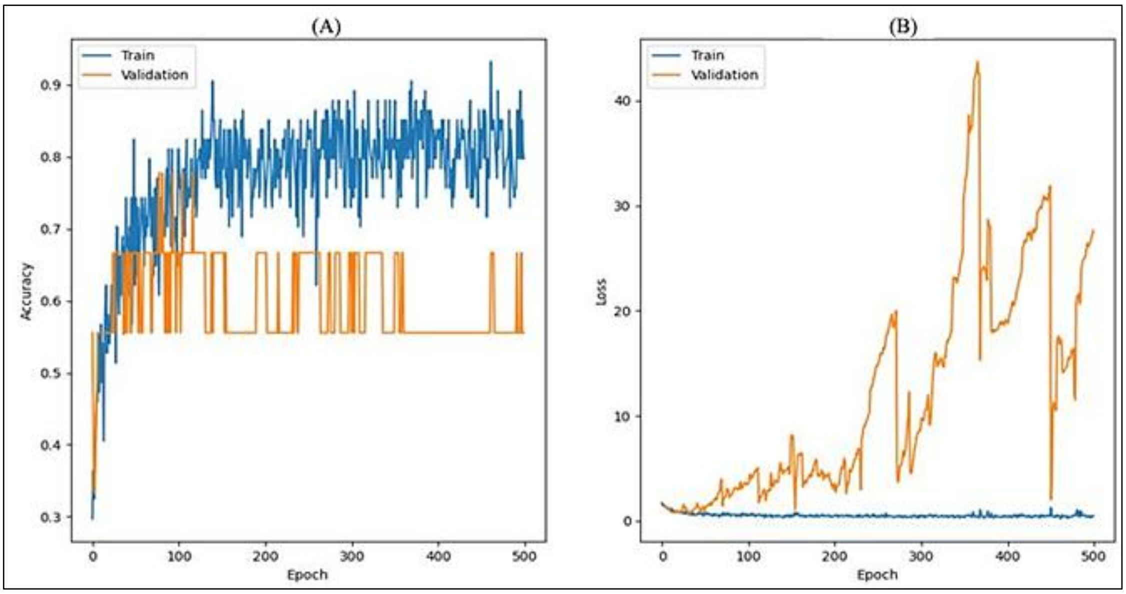


Figure 7: CILNN Accuracy and Loss – (A) Model Accuracy, (B) Model Loss

Figure 7 illustrates two graphs tracking the performance of a machine learning model over time. Model accuracy is displayed on the left, while model loss is displayed on the right. Training accuracy improves and fluctuates at a high level, while validation accuracy remains lower and more stable. The loss graph shows consistently low training loss, but large spikes in validation loss over time. A model that performs well on training data but struggles to generalize to new, unknown data strongly suggests over fitting. A growing gap between training and validation performance indicates that the model memorizes the training set rather than learning generalizable patterns.

Results and Discussion

In this research work, Python was used as the primary programing language and various Python libraries were used. To manage the dependencies efficiently, MiniConda, a lightweight package manager for Python, was used. A 2.70 GHz Intel Core i7-2620M processor was used for processing, providing substantial power. With 16 GB of RAM, the system can handle large datasets and complex computations with ease. A 64-bit Windows 7 operating system provides a stable and familiar platform for conducting research. Accuracy measures correctly classified data, but it does not distinguish foreground errors from background errors (18).

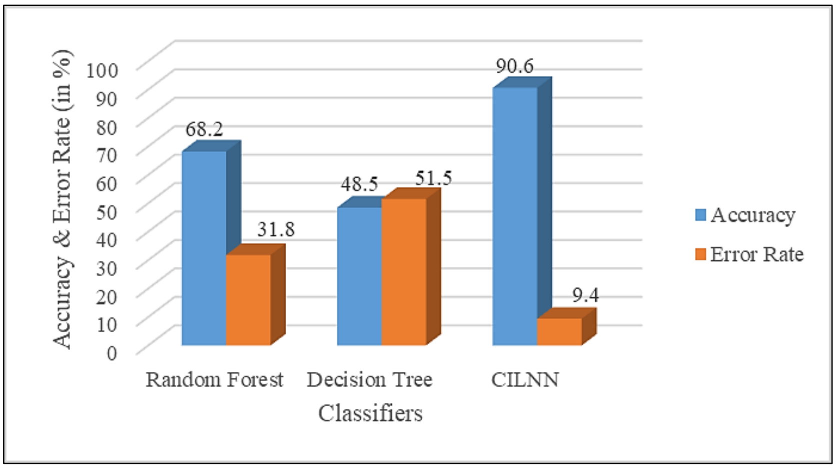


Figure 8: Accuracy and Error Rate

Figure 8 compares the performance of three machine learning classifiers: Random Forest, Decision Tree, and CILNN. It displays accuracy (blue) and error rate (orange) for each model. CILNN demonstrates superior performance with 90.6% accuracy and 9.4% error rate, significantly outperforming the others. Random Forest follows with 68.2% accuracy and 31.8% error rate, while Decision Tree shows the weakest performance at 48.5% accuracy and 51.5% error rate. The complementary nature of accuracy and error rate is evident, as they sum to 100% for each classifier. This visual representation effectively highlights the substantial performance gap between CILNN and the other two classifiers for the given task or dataset.

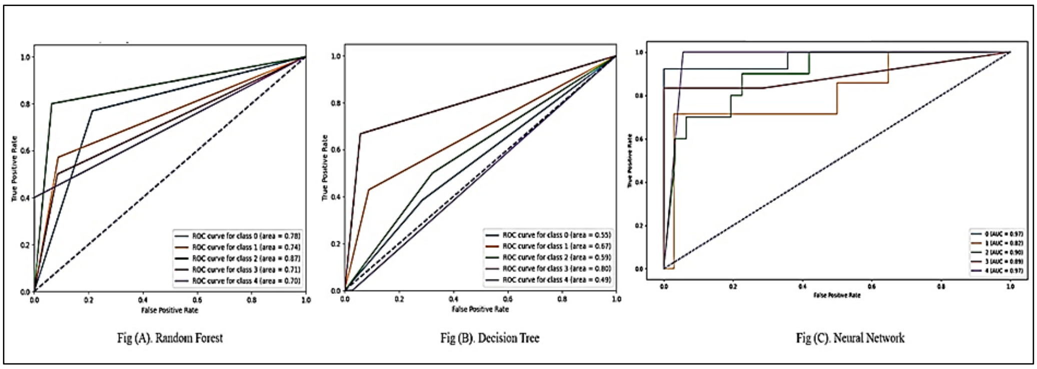


Fig (A). Random Forest

Fig (B). Decision Tree

Fig (C). Neural Network

Figure 9: Precision-Recall

Figure 9 shows the Receiver Operating Characteristic (ROC) curves for Random Forest, Decision Tree, and CILNN. At various thresholds, dotted diagonal lines represent random classifier performance. Multi-curves consistently above the diagonal indicate good performance across various configurations (19). Decision Tree graphs display similar multiple curves, but they perform slightly worse. For some configurations, the CILNN graph shows fewer but more distinct curves, with some reaching high into the top-left corner. AUC values in each graph indicate the performance of the overall model, with larger values indicating better performance. Figure 10 shows the confusion matrix.

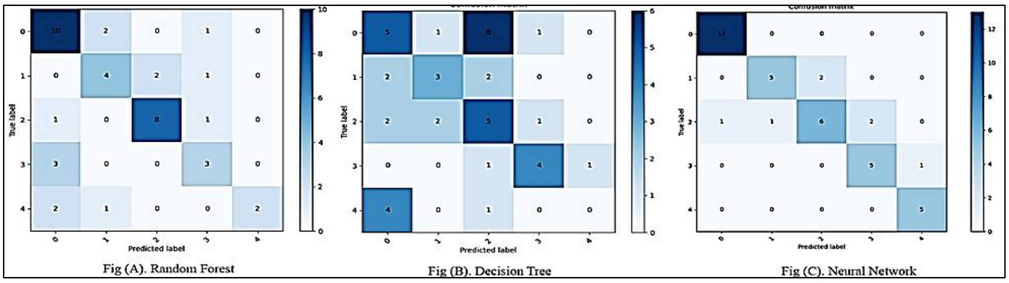


Fig (A). Random Forest

Fig (B). Decision Tree

Fig (C). Neural Network

Figure 10: Confusion Matrix

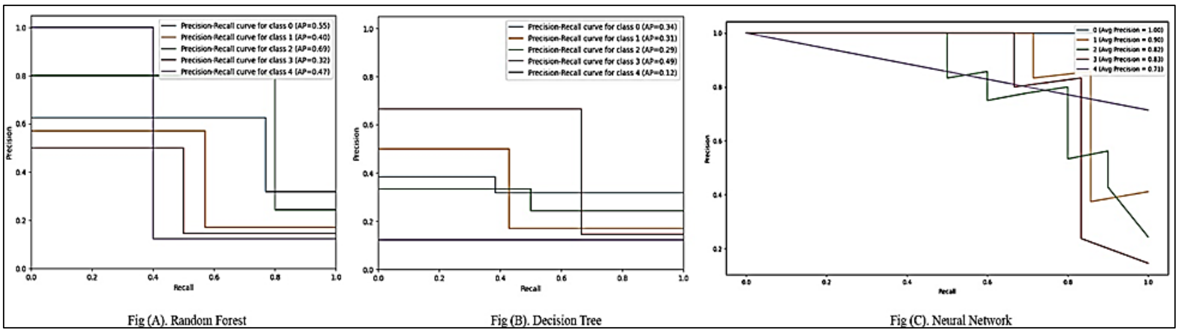


Fig (A). Random Forest

Fig (B). Decision Tree

Fig (C). Neural Network

Figure 11: ROC Graph

Three different machine learning models are compared in terms of their performance across different recall thresholds in figure 11. Graphs plot precision versus recall for different classes, likely representing different animal species. This model shows consistent performance across recall values, with clear separation between classes. As recall varies, the Decision Tree model shows less nuanced changes in precision. Multiple intersecting lines and variable precision-recall trade-offs are revealed in the Neural Network graph. There are areas of high precision for some classes, but also areas of greater volatility. Each model has unique strengths and weaknesses when it comes to balancing precision and recall. Model selection and tuning can be influenced by whether the priority is overall consistency, simplicity, or maximizing performance for specific classes.

Discussion

An audio correlation heat map shows how audio features relate in figure 12. These features show complex interdependencies, with many MFCCs showing moderate to strong correlations, both positive and negative. There could be overlapping information about the audio signal's spectral properties represented by these coefficients. There is an inverse relationship between spectral bandwidth and several MFCCs, indicating that this overall measure of frequency spread is negatively correlated. Audio signals demonstrate interesting correlations between frequency-weighted energy (FWR), harmonics, and percussive features, highlighting their multifaceted nature. For audio analysis tasks like classification, speech recognition, understanding these relationships is crucial, as they can guide feature selection processes, identify redundant information, and ultimately improve performance.

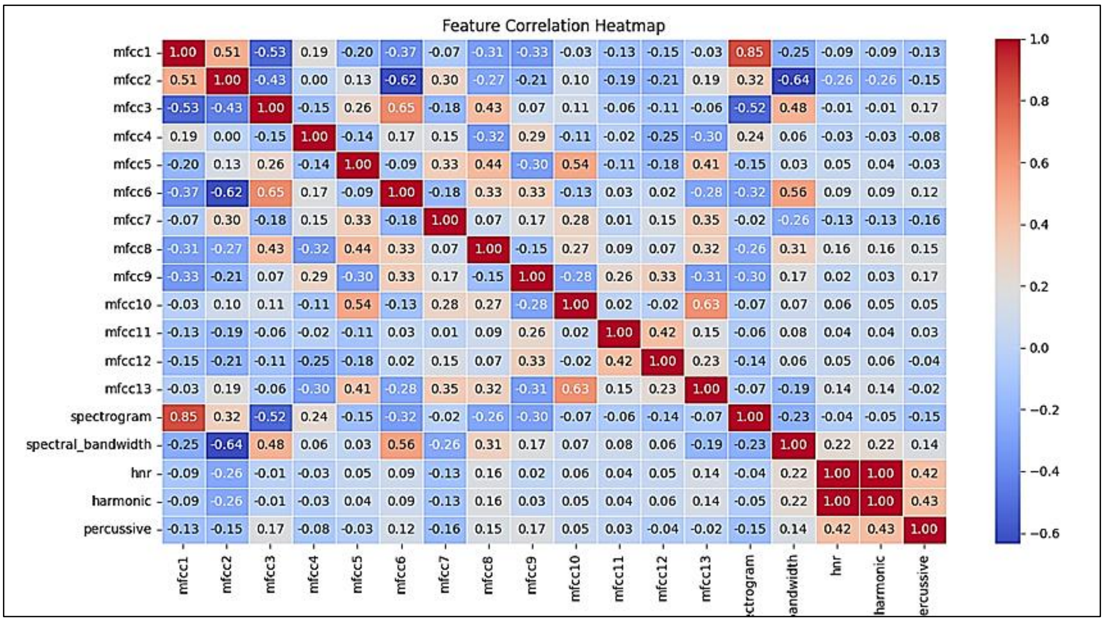


Figure 12: Correlation Map

Acoustic prediction using AI classifiers is enhanced by Explainable AI (XAI) to provide transparency and interpretability. It helps make the algorithm's black-box nature more transparent by explaining how it makes acoustic predictions. By using XAI techniques, it can explain individual predictions based on acoustic features. It is possible to validate the model's validity by understanding the reasoning behind predictions. New relationships or patterns may be discovered in acoustic data using XAI.

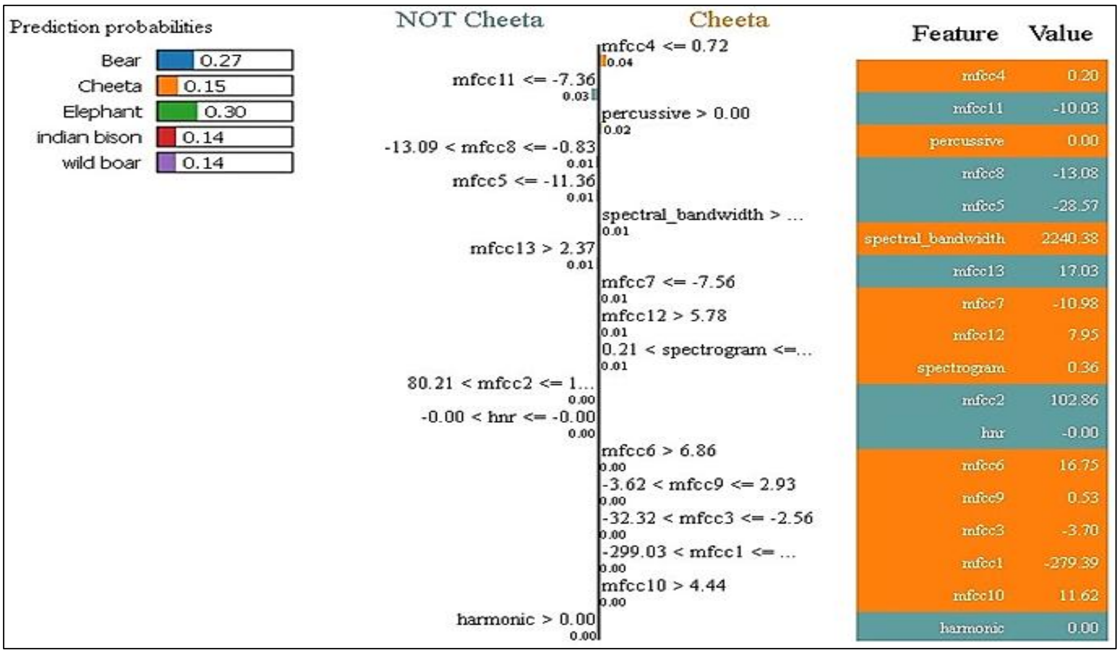


Figure 13: Explainable Tree for Random Forest

Figure 13 shows a visualization of the results of a Random Forest model for identifying different animals with the goal of distinguishing cheetahs from other species. Various animal classes are predicted by the model. At 0.27, "Bear" has the highest probability, followed by "Elephant" at 0.20. The probability for "Cheeta" (likely a misspelling) is 0.15. The model's decision-making process is divided into two branches, "NOT Cheeta" and "Cheeta". An input is classified using various features (such as MFCC11, MFCC4, etc.). The right table shows the values for each feature. This model includes audio-related measurements (MFCCC: Mel-frequency cepstral coefficients, spectral bandwidth, etc.). This is likely an audio-based classification model using animal vocalizations. Different audio features are used to make the final classification.

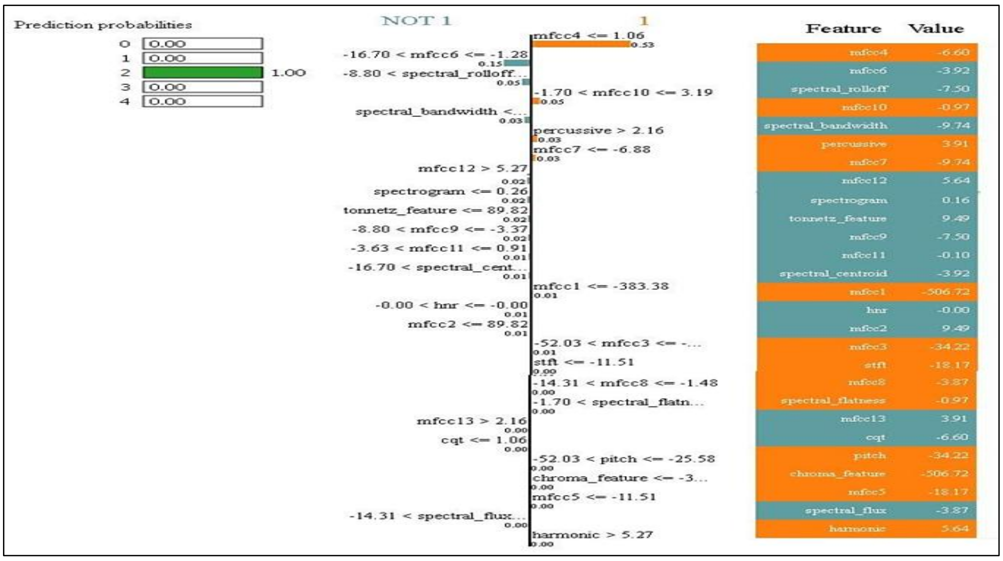
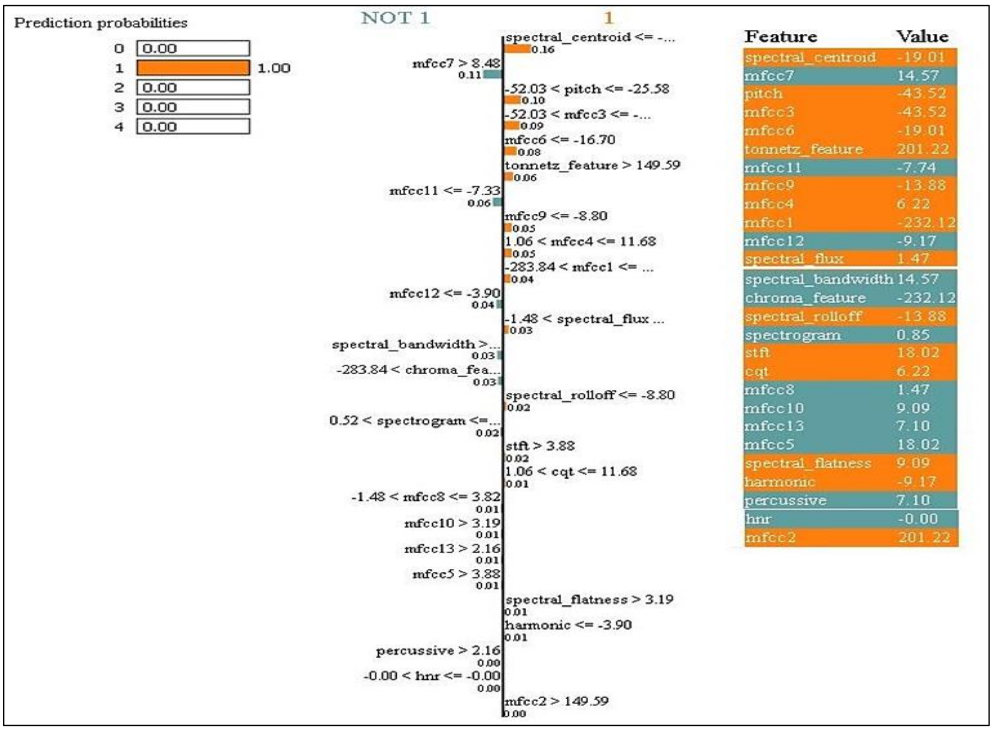


Figure 14: Explainable Tree for Decision Tree

Figure 14 shows the output of a decision tree. Here are the prediction probabilities (0 to 4) for different classes. This instance predicts class 2 with a probability of 1.00. It shows a simplified decision tree or rules used to make the classification. There are a number of decision nodes or leaves, each with a condition, such as "mfcc4 <= 1.06" or "spectral_rolloff <= 7.50". The right side of the table shows various features and their corresponding values. MFCCs, spectral roll offs, spectral bandwidth, etc., appear to be related to audio analysis. In this model, audio samples appear to be classified based on their genre or event, possibly for the purpose of audio classification. MFCCs, spectral properties, and other acoustic measurements are among the features used. In the decision tree, each node represents a decision based on a specific feature value threshold that is used to navigate to a final classification.



Feature	Value
spectral_centroid	-19.01
mfcc7	14.57
pitch	-43.52
mfcc3	-43.52
mfcc6	-19.01
tonnetz_feature	201.22
mfcc11	-7.74
mfcc9	-13.88
mfcc4	6.22
mfcc1	-232.12
mfcc12	-9.17
spectral_flux	1.47
spectral_bandwidth	14.57
chroma_feature	-232.12
spectral_rolloff	-13.88
spectrogram	0.85
stft	18.02
cqt	6.22
mfcc8	1.47
mfcc10	9.09
mfcc13	7.10
mfcc5	18.02
spectral_flatness	9.09
harmonic	-9.17
percussive	7.10
hnr	-0.00
mfcc2	201.22

Figure 15: Explainable Tree for CILNN

Figure 15 shows the to be a visualization of a CLINN. Prediction probabilities are shown on the left (0, 1, 2, 3). The probability of Class 1 is 1.00, suggesting it is the predicted class. In the middle, it seems a simplified decision tree. Different decision nodes are displayed with features and thresholds. One of the decision points is "mfcc7 > 8.48". Feature values are shown on the right. The model uses these features as inputs. A positive or negative influence might be indicated by the colors (orange and blue). It shows the some notable features include: spectral centroid: 19.11, mfcc7: 14.57, tonnetz_feature: 201.22 and chroma_feature: -232.12. A tree structure in the middle shows how the model predicts. By visualizing the classification process, it can better understand how the model makes predictions. Developing the system for real-world environments is one significant limitation. Real-world applications of the CILNN model show strong performance. The use of labeled datasets is another limitation, which can be scarce or incomplete. For broader application, the dataset needs to include more species and environmental conditions. Furthermore, with its attention mechanisms and interconnected layers, the CILNN model may require significant hardware resources, limiting its scalability for low-resource settings.

Conclusion

The CILNN (Convolutional Interconnected Layer Neural Network) was implemented to detect dangerous animals using audio signals, with an emphasis on XAI model interpretability. A new neural network architecture combines advanced feature extraction, SHAP-based feature selection, and advanced feature extraction to achieve high classification accuracy. CILNN outperforms traditional machine learning models, demonstrating its potential for wildlife monitoring and management applications. With the help of XAI methods, including SHAP values and decision tree visualizations, it gained valuable insight into the decision-making processes of our CILNN and traditional models. Analyzing the interpretability of various audio features shed light on the models' classification strategies and revealed their relative importance. This transparency improves trust in the models and offers avenues for further refinement and optimization. Despite promising results, there are several directions for future research. This includes expanding the dataset to include more species and environmental conditions, investigating real-time processing capabilities for field deployment, and integrating the audio-based system with other sensors for more comprehensive wildlife monitoring. Furthermore, XAI techniques tailored for audio classification tasks could provide even deeper insights into model behavior. Bioacoustics and wildlife conservation have benefited greatly from interpretable deep learning.

Abbreviations

XAI: Explainable AI, CILNN: Convolutional Interconnected Layer Neural Network, SHAP: SHapley Additive explanation, MFCC: Mel-Frequency Cepstral Coefficients.

Acknowledgement

None.

Author Contributions

The corresponding (first) author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Conflict of Interest

There is no conflict of interest.

Ethics Approval

Not applicable.

Funding

No Funding.

References

1. Schön PC, Puppe B, Manteuffel G. Automated Recording of Stress Vocalizations as A Tool to Document Impaired Welfare in Pigs. Animal Welfare. 2004; 131(2):105–110.
2. Soltis J, Blowers TE, Savage A. Measuring positive and negative affect in the voiced sounds of African elephants Loxodonta Africana. J Acoust Soc Am. 2011; 129(2):1059-1066.
3. Mcloughlin MP, Stewart R, McElligott. Automated bioacoustics: Methods in Ecology and Conservation and Their Potential for Animal Welfare Monitoring. J R Soc Interface. 2019; 161(55):1-10. https://doi.org/10.1098/rsif.2019.0225.
4. Maigrot AL, Hillmann E, Briefer EF. Encoding of emotional valence in wild boar Sus Scrofa calls. Animals. 2018; 86(1):1-10. https://doi.org/10.3390/ani8060085.
5. Manteuffel G, Puppe B, Schön PC. Vocalization of farm animals as a measure of welfare. Appl Anim Behav Sci. 2004;88(1-2):163-82. https://doi.org/10.1016/j.applanim.2004.02.012.
6. Hill AP, Prince P, Snaddon JL, Doncaster CP. AudioMoth: A low-cost acoustic device for monitoring biodiversity and the environment. Hardware-X. 2019; 61:1-10. https://doi.org/10.1016/j.ohx.2019.e00073
7. Ruff ZJ., Lesmeister DB., Appel CL, Sullivan CM. Workflow and Convolutional Neural Network for Automated Identification of Animal Sounds. Ecol Indic. 2021; 1241:1-10. https://doi.org/10.1016/j.ecolind.2021.107419
8. Pabico JP, Gonzales AMV, Villanueva MJS, Mendoza AA. Automatic Identification of Animal Breeds and Species Using Bioacoustics and Artificial Neural Networks. arXiv preprint. 2015; 1(2):1-10. https://arxiv.org/pdf/1507.05546.pdf
9. Oikarinen T, Srinivasan K, Meisner O, Hyman JB. Deep Convolutional Network for Animal Sound Classification and Source Attribution Using Dual Audio Recordings. J Acoust Soc Am. 2019; 145(2):654-662. https://doi.org/10.1121/1.5087827
10. Bishop JC, Falzon G, Trotter M, Kwan P, Meek PD. Livestock vocalization Classification in Farm Soundscapes. Comput Electron Agric. 2019; 162(1):531-542. https://doi.org/10.1016/j.compag.2019.04.02.
11. Joanna Babińska-Werka, Dagny Krauze-Gryz, Michał Wasilewski, Karolina J. Effectiveness of an acoustic wildlife warning device using natural calls to reduce the risk of train collisions with animals. Transp Res Part D Transp Environ. 2015; 381(2):6-14. https://doi.org/10.1016/j.trd.2015.04.021.
12. Sandra Cardoso, Ceres Faraco, Gonçalo Pereira, Harry Eckman. A survey of acquisition and animal-related factors leading to microchipping and neutering of dogs and cats in Portugal. J Vet Behav. 2023; 641(1): 9-15. https://doi.org/10.1016/j.jveb.2023.06.005.
13. Manikandan T, Kumaresan SJ, Muruganandham A. IoT Based Animal Detection and Alert System for Farm Fields. Proceedings of the International Conf. on Commu., Computing and IoT (IC3IoT). 2024; 1-3. https://doi.org/10.1109/IC3IoT60841.2024.10550250.
14. Dissanayake CM, Tennakoon P. Sound Analysis-Based Animal Recognition and Repel System for Agriculture. Moratuwa Engineering Research

Conference (MERCon), Sri Lanka. 2023; 328-333. https://doi.org/ 10.1109/MERCon60487.2023.10355383.

15. Chandralekha E, Thiruselvan P, Ravikumar S, Vijay K. Animal Intrusion Detection System: Protected Crops and Promoted Safety using Machine Learning. Proceedings of the International Conf. on Research Methodologies in Knowledge Mgmt, AI and Telecommunication Engineering. 2023; 1-5. doi: 10.1109/RMKMATE59243.2023.10368943.
16. Dataset: Cornell Lab of Ornithology. Macaulay Library. Cornell University.; 2024; Available from:https://search.macaulaylibrary.org/catalog?taxonCode=t-11036143searchField=animals.
17. Pimentel da Silva M, Correia da Silva N. Feature selection using SHAP: An explainable AI approach. Brasília, DF. 2021; 1-10.
18. Govindaprabhu GB, Sumathi M. Ethno medicine of Indigenous Communities: Tamil Traditional Medicinal Plants Leaf detection using Deep Learning Models. Procedia Comp Sci. 2024; 235(1):1135-1144.
19. Govindaprabhu GB, Sumathi M. Safeguarding Humans from Attacks Using AI-Enabled (DQN)Wild Animal Identification System. Int Res J Multi-discip Scope (IRJMS), 2024; 5(3): 285-302.