



DEEP LEARNING BASED DOG BEHAVIORAL ANALYSIS FROM THE AUDIO DATA USING MFCC FEATURE EXTRACTION

Kalaavathi Bhuvaneshwaran ^{*1}, Sujaykumar Reddy M ² and Arivoli A ³

¹Department of IoT, SCOPE, VIT University, Vellore, Tamilnadu, India.

²B. Tech. Student, SCOPE, VIT University, Vellore, Tamilnadu, India.

³Department of Computational Intelligence, SCOPE, VIT University, Vellore, Tamilnadu, India.

¹<http://orcid.org/0000-0003-4027-3137>, ²<http://orcid.org/0009-0000-4604-2490>, ³<http://orcid.org/0000-0002-8954-6346>

E-mail: kalaavathi.b@vit.ac.in, sujaykumarreddy.m2020@vitstudent.ac.in, arivoli.a@vit.ac.in

ARTICLE INFO

Article History

Received: March 05, 2026

Reviewed: April 05, 2026

Accepted: May 15, 2026

Published: September 30, 2026

Keywords:

Audio Analysis,
Convolutional Neural Networks
(Cnn),
Dog Behavior Analysis,
Emotional Intelligence,
Stacked Cnn-Bilstm.

ABSTRACT

Understanding the behavior and emotions of dogs through their vocalizations is vital for various applications that includes emotional intelligence analysis and remote monitoring services. In this study, the vocal behavior of Dachshund dogs has been examined by using a novel feature extraction technique based on Mel Frequency Cepstrum Coefficients (MFCC) and advanced Deep Learning architectures. Classifying the vocalizations of pets accurately, particularly when they are left alone, is a critical challenge. Hence, an approach is offered in this paper that integrates MFCC-based feature extraction with multiple neural network architectures like CNN, LSTM, Feedforward Networks (FFN), and hybrid CNN-LSTM models. In comparison to the other architectures, the Stacked CNN-BiLSTM model outperforms better at detecting Pet Sound Events. This study enhances pet dog vocalization analysis, enabling better communication in the field.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Human-Pet Interaction (HPI) is a developing research topic that looks into the complex relationship between humans and their pets like cats and dogs. Creating communication channels between dogs and their owners is one of the main goals of HPI, especially when they are away from home. Research indicates that dogs possess emotional intelligence and can detect changes in their owners' behavior [1]. This emotional intelligence extends to their ability to sense physical or cognitive impairments in humans, evidenced by widened eyes and shallow breathing [2]. PAD technology advancements now allow devices to interpret canine emotional states by recording behavioral cues through sensor systems which deliver information about pet physical activity [3]. But it's crucial to do more than just detect activity if you want to evaluate a dog's general health and wellbeing. Vezzosi et al. [4] investigated the viability of using smartphone ECG sensors to monitor dogs' heart rates and rhythms at home, highlighting non-invasive health monitoring. Using the AKC registration data [5] from the past five years, the study examines the results for the 10 most popular breeds. The ranking shows that breeds such as French Bulldogs, Labrador Retrievers, Golden Retrievers, German Shepherds, Poodles, and Bulldogs have always been persistently owned breeds in society. By focusing on breeds commonly found in average households, the study aims to investigate genetic diversity within a representative sample of canine companions. Retrievers and Bulldogs, consistently ranked among the most popular breeds by the AKC, exemplify this approach, ensuring that the study encompasses a broad spectrum of canine genetic variation. Deep Learning (DL) has transformed various fields, including image recognition or NLP, even extending to animal emotion detection. This study explores the AI's ability to accurately predict dog emotions using selected datasets. The datasets was selected based on accuracy and labelling, breed types, emotional states, and quality. This would provide a foundation for developing and evaluating AI models in dog emotion recognition. The proposed study is organized as: The overview of methods needed by the research has been presented in Section 2.

In the next Section 3, various processes like collecting data, evaluating it, and classifying them are discussed. Section 4 summarizes the data categorization problem and the mathematical underpinnings of the proposed solution. Finally, Section 5 offers a synopsis of the closing thoughts and suggestions for additional study.

II. BACKGROUND

II.1 DOG BARKING CLASSIFICATION

Dog bark categorization is an important part of emotion classification techniques that help us understand how dogs behave and communicate. According to a study by Peter et al. [6], people may infer basic interior states of dogs from their audio signals even at a young age. This result defies past research that claimed young children frequently misinterpreted visual cues from dogs. Larrañaga et al. [7] classified Mudi dogs by age, sex, and circumstance (e.g., Alone, Ball, Fight, Food, Play, Stranger, & Walk). used a ML technique. They successfully classified these characteristics by using a Wrapper technique with the K-Nearest Neighbors (KNN) algorithm for optimal feature selection. Advanced signal processing methods, including spectral energy, formant structure, and pitch, were used by Hantke et al. [8] to determine the emotional intensity and context of dog bark sequences. These methods assisted in capturing the subtle differences in canine vocalizations that represent distinct meanings and emotional states. The behavioural patterns of Indian dogs are studied by Mohandas et al. [9] by particularly emphasizing the Hound Breed with the help of reinforcement learning like Q-Learning algorithm. In this method, five different actions were assessed and results proved the study obtained 85% accuracy and highlighted implementation of reinforcement learning in the study of canine behaviour. Nikolett et al. [10] used the statistical analysis and demonstrated the novel method for classifying the human reactions based on the canine vocalizations.

For studying how particular audio sequences are attracting human reactions, the authors have utilized 153 participants. These participants are divided into three different age groups for evaluating 12 different dog barking sequences and the results of this method showed that there is a possibility for successful classification of human sentiments by using the various barking features of Canine. The creation of a Dog Behavioral Monitoring System has been advanced with the implementation of pattern recognition algorithms by Boteju et al. in [11]. In this study, various behavioural type such as recognition of dog breeds, their walking patterns, their resting patterns, and their barking patterns has been analysed. Also, the various dog breeds like the German Shepherd and Pomeranian has been examined to create an effective framework for evaluating and monitoring the behavior of canine thereby advancing both behavior management and animal welfare. The various studies conducted on analysing dog behaviour thereby classifying the dog barks has paved pay for better understanding of human-canine interactions provided with valuable insights for enhancing canine welfare thereby improving relationship between human and dog. With the various innovative methodologies, the area of HPI continues to evolve. This also fosters a more precise and objective assessments of dogs' physical and mental states.

II.2 SENSOR-BASED CLASSIFICATION

Sensor-based classification is used to identify dog behaviors and monitor their well-being with the help of multi-nodal data from sensors. Using a movement sensor with six degrees of freedom, Vehkaoja et al. [12] collected data from 17 dog breeds which paved the way for authors to predict seven distinct behaviors and activities. Using object detection techniques like YOLOv3, R-CNN, and YOLOv4, Jinah et al. [13] identified dog behaviors across seven classes by using the videos and sensor data and this research has demonstrated the efficiency of multi-modal data analysis in differentiating the multiple behaviours of the dogs. In the research, Electroencephalography (EEG) has been used as it is capable of offering valuable insights into neural activities in animals. Using EEG signals in dogs, Kujala et al. [14] researched to identify dogs' ability to identify multiple facial expressions, different faces, and scrambled objects. The results of the study indicated the specific response of brain to various types thus giving way for category-specific processing the canine brain.

II.3 BEHAVIORAL PATTERN ANALYSIS

With the help of behavioral pattern analysis, various dogs are studied and categorized depending upon how the dogs are behaving and interacting with people and their surroundings. Riggio et al. [15] conducted a study to find how the dogs are interacting with their owners by using a classification system that has been developed for humans especially the new born. This study extended the research for analyzing the behavioral variations between dogs in a detailed manner especially for the dogs that are categorized as secured and avoidant. This offered the vital insights in finding the complexities of the relationship between dog and its and how it influences canine behavior. Özen et al. [16] conducted a study to investigate the various ML methodologies for classifying dog walks. This study aims to develop a methodology for classifying walking habit of dogs by utilizing the information collected from dogs of various types and sizes. The study also contributed for a better understanding of canine mobility with a newly introduced metric for forecasting the difficulties faced by the dogs while they are walking. For improving not only the animal welfare but also the relationship between dogs and humans, sensor-based classification was implemented with behavioral pattern analysis with an extended understanding and capacity to monitor the behavior of canines.

III. PROPOSED METHODOLOGY

The study employs a custom DL architecture intended especially for the classification of canine behaviors from audio inputs, departing from conventional optimization considerations for pre-trained neural networks like DenseNet and ResNet. Our methodology entails a complete analysis of CNNs, LSTMs, a subclass of RNNs, and hybrid (CNN-LSTM) modeling, as illustrated in figure 1.

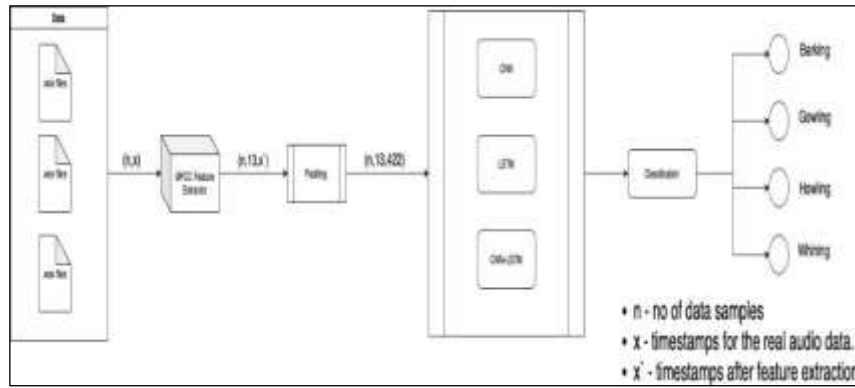


Figure 1: Proposed Methodology for Dog Behavioral Analysis.

Source: Authors, (2026).

The model utilizes CNNs for feature extraction (FE) and LSTMs for temporal sequence modeling to enhance classification accuracy. Barking, Growling, Howling, and Whining are classified based on MFCC features from .wav files arranged in (n, 13, 422). CNNs and RNNs combination allows improved analysis of canine vocalizations while optimizing classification efficiency.

III.1 DATA ACQUISITION AND PREPARATION

This work makes use of the dataset that Kim et al. [17] provided, which consists of four different classes of waveform-formatted audio data that represent dog barks. After this preprocessing stage, MFCC are used to extract statistical features from the audio data. Numerous research in the sector have confirmed that MFCC stands out for its remarkable recognition rates and constant performance. Vimal et al. [18] successfully showed that MFCC can identify eight different emotions, ranging from neutral and calm to happy and angry. Leveraging the demonstrated efficacy of MFCCs in extracting discriminative features from audio data, the incorporation into our methodology underscores the significance of MFCCs as a cornerstone in audio FE techniques, aligning with established methods and empirical findings. Data Structuring involves arranging these features into a sequential format (n, 13, 422), where each dimension holds significance. The first dimension 'n' represents the number of segments or time frames, the second dimension '13' corresponds to the number of statistical features extracted at each timestamp, and the third dimension '422' signifies the number of timestamps or frames considered. This structured format facilitates further processing and analysis of the extracted feature

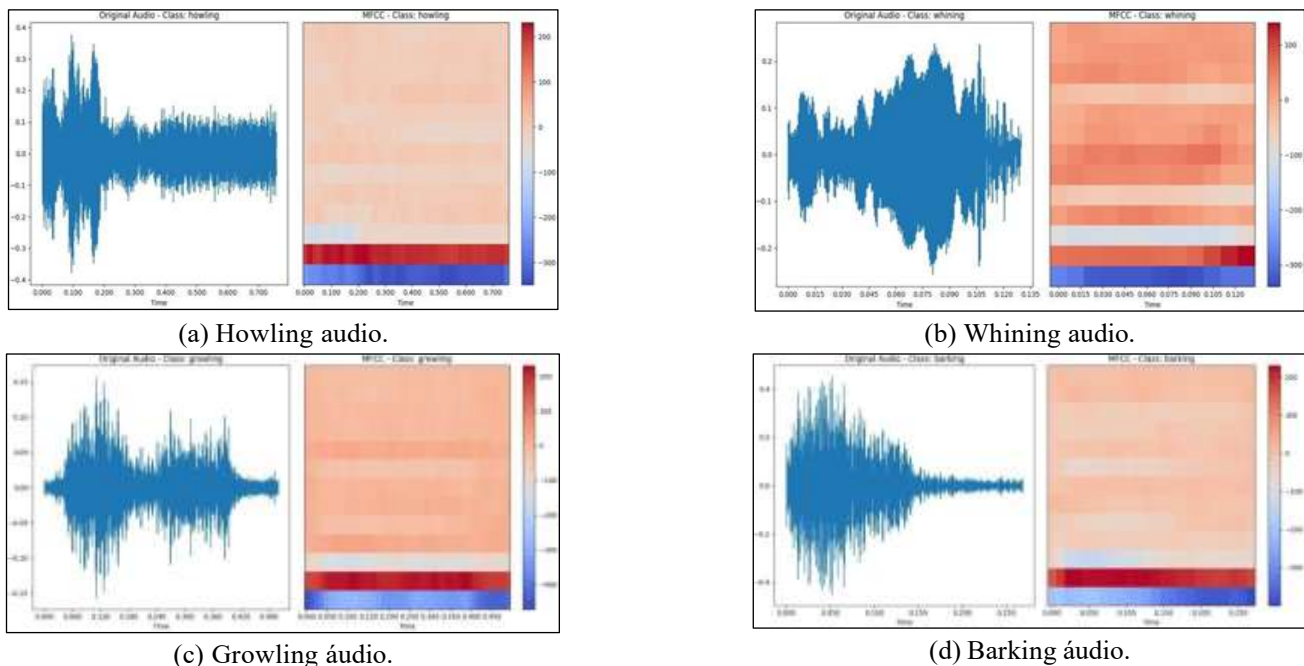


Figure 2: Sample heatmaps of MFCC-extracted data for each class.

Source: Authors, (2026).

The MFCC plot often looks like a heatmap, with x-axis denoting time, y-axis indicating frequency bands, and color intensity the magnitude of each MFCC coefficient. Figure 2 displays the sample audio figure. This visualization helps in capturing audio signals' spectral characteristics. This is vital for few tasks like speech recognition and sound classification. Figure 3 demonstrates the MFCC FE procedure and sheds light on its underlying ideas. Our methodology's incorporation of MFCC allows for reliable FE from the audio data, which establishes the groundwork for further classification tasks.

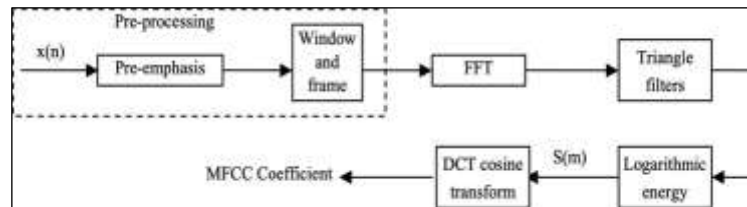


Figure 3: Standard MFCC Feature Extractor used by Xie et al.
Source: [19].

III.2 MODEL ARCHITECTURES

In outlining the DL models used, we focus on their configuration and the rationale behind their selection, employing both Base Architectures and Customized Sequential Models. The analysis begins with extracting MFCC essential for transforming dog vocalizations into analyzable data and revealing sound properties crucial for distinguishing vocal types, such as pitch and frequency. The base architecture models utilize 1D-CNNs for automated feature detection from sequential audio data and Feed Forward Neural Networks (FFNNs) for feature mapping, thereby optimizing both extraction and recognition processes, the same goes for LSTM models as well. CNNs and LSTM variations are used in customized sequential models to extract sequence patterns and temporal dynamics from audio data. The Stacked CNN-BiLSTM architecture, the apex of our model structure, greatly improves classification accuracy by combining CNNs for in-depth FE with BiLSTMs for examining temporal relationships. In order to ensure the creation of strong, flexible models suited for efficient real-time vocalization analysis, training uses gradient-based optimization with rigorous cross-validation.

III.2.1 Feature Extraction Using MFCC

Since human auditory perception does not change quickly over time, MFCCs are a feature set that is frequently used in speech and audio processing. The high frequencies, which usually contain more specific information about the dog's vocalization patterns, are amplified in each frame by a pre-emphasis filter before a Fourier transform transforms signal from time domain to frequency domain & computes power spectrum to determine magnitude of each frequency component. After that, spectrum is plotted onto the Mel scale, an auditory scale that simulates how the human ear reacts to various frequencies. This gives the frequencies that are perceptually significant greater detail and the remainder less. The log of the powers at each Mel frequency is calculated since the logarithmic scale is used by humans to perceive the loudness of sound. Lastly, to obtain the MFCCs, the log Mel powers are subjected to DCT. The characteristics that are most important for distinguishing between barks, howls, growls, and whines are highlighted in the MFCCs, which are a succinct representation of the power spectrum of a sound and of dog vocalizations in our study. Recent advancements in audio signal processing have significantly benefited from the adoption of 1DCNNs. Originally developed for image processing tasks, these networks have been adeptly adapted to manage sequential data, demonstrating their effectiveness in capturing temporal patterns within audio signals. Utilizing an input configuration described by (Batch size, 13, 422), where each batch contains 13 features recorded at 422 discrete time steps, 1D CNNs navigate through the temporal domain to identify complex patterns crucial for comprehensive audio analysis.

The convolutional operation involves sliding kernels across the spectrogram's temporal axis, capturing essential local temporal elements such as rhythm, timbre, and pitch. As the network stack multiple convolutional layers, each with progressively larger receptive fields, it extracts increasingly complex features at deeper levels, enhancing its FE capabilities. Pooling layers, which combine notable features and lower processing requirements, are used to down sample the feature maps after the convolutional layers. Max pooling, in particular, is used to retain the maximum value within each pooling window, preserving significant temporal characteristics. The features are extracted as a result of the above processes and they are fed as input to the fully connected layers where the synthesis and classification of the audio signals are done by the network. The audio signals are classified into distinct categories. To minimize the predefined loss function throughout the training phase, various optimization techniques like stochastic gradient descent (SGD) or Adam will be implemented thereby optimizing the performance of network classification. The 1D CNN mode has been used as a baseline model as shown in figure 4 and is applied in the process of audio signal processing. Using 1D CNN in audio signal processing delivers numerous advantages like robustness in signal length, computational efficiency, and the potential to capture both local & global temporal patterns. For types of audio processing including voice recognition, music genre categorization, and environmental sound analysis, these characteristics make 1D CNNs essential tools. CNNs are used in audio analysis to automatically and hierarchically extract temporal & spatial information from input data from MFCC. The convolutional layers serve as feature detectors from the audio signals' spectrograms, identifying important characteristics in the audio frames and learning filters that capture local dependencies.

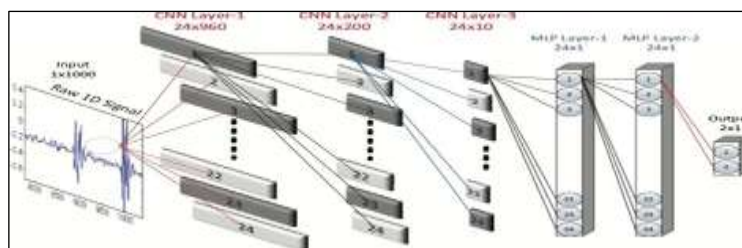


Figure 4: 1DCNN Architecture by Serkan et al.
Source: [20].

III.2.2 Recurrent Neural Networks (RNN)

In order to simulate cognitive processes, neural networks such as ANNs or FFN are inspired by interconnected structure of human neurons. ANNs consist of numerous interconnected processing elements, or neurons, where information is transmitted from one layer to another to capture non-linear relationships between input and output. RNNs are used to improve this skill by supplementing the input's channels or features, which enables better representation and comprehension of intricate temporal patterns. Unlike traditional FFNN, RNNs introduce recurrent connections between neurons, allowing them to learn from past information and utilize it for future predictions. By incorporating recurrent connections, RNNs can capture complex non-linear relationships and dependencies within the dog audio signals, leading to improved performance for the classification task. RNNs are useful in sequential processing, and they stand out from more conventional neural networks like FFN and CNN when used to categorize audio signals, a multi-to-one classification problem. The disappearing and expanding gradient problems that come with sequential processing, however, are a problem for RNNs. LSTM cells are specifically made to combat these problems. In order to control information flow, LSTMs implement three gating mechanisms namely Input, Output, and Forget gates as shown in Figure 5. While forget gate handles vanishing gradient issue by controlling information transmitted across time, input gate controls storage of new information in cell state. The forget gate lessens difficulties caused by vanishing gradients by choosing whatever information from prior cell state to keep or discard and allocating values between 0 and 1. At the same time, the output gate uses updated cell state to shape next hidden state. Together, these gating mechanisms regulate information flow into, out of, and within LSTM cell, addressing the subtleties of sequential processing. Because LSTMs are skilled at modeling time series data, they can accurately analyze temporal data by learning which properties to retain or forget over lengthy durations.

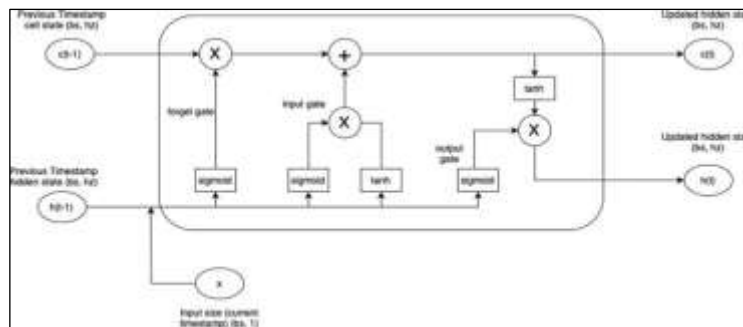


Figure 5: LSTM cell Structure.

Source: Authors, (2026).

III.2.3 Hybrid Cnn-Lstm Model

Utilizing hybrid CNN-LSTM model enables analysis of complex dog vocalization data, facilitating insights into their behavior and emotional states. For instance, different types of barks can be classified to determine whether a dog is in distress, anxious, or simply alerting their owner to a visitor. This level of analysis could be invaluable for remote monitoring services and emotional intelligence analysis, where understanding the nuanced differences in dog vocalizations can lead to better welfare and care for the animals. The CNN component acts as an automatic feature extractor. In this hybrid model, convolutional layers are applied to the MFCCs to identify and learn salient features within the vocalization data. This can include detecting edges, shapes, and textures in the spectrogram that are characteristic of different dog vocalizations, like the tonal quality of a bark versus a growl. Convolution layers are followed by activation functions like ReLU to produce non-linearity and enable model to learn more complex patterns. To summarize features identified by the convolutional filters while keeping the most crucial information, pooling layers typically max pooling reduce data dimensionality. The LSTM layers receive the features that were extracted from the CNN layers. LSTMs are designed to handle sequential data and are adept at learning from the temporal structure within the audio data. They can maintain information over long sequences, which is essential for capturing the context within the vocalization patterns that might be indicative of specific behaviors or emotional states. LSTM layers analyze the features over time, learning which characteristics of the sound are important for identifying the different types of vocalizations and when those characteristics appear in the time sequence. The hybrid model capitalizes on CNN's ability to reduce a complex input into a simpler representation and LSTM's ability to remember sequences over time. This combination is particularly potent for audio signal processing where both the content of the signal (which sounds are present) and the context of the signal (the order and duration of sounds) are important for accurate classification.

IV. RESULTS AND DISCUSSION

This study utilized three distinct DL architectures, namely CNN, LSTMs, and a hybrid CNN-LSTM model, to assess their efficacy in the designated task. It was implemented on Macbook M1 platform using TensorFlow 2.8, with Keras facilitating the execution, being a prevalent Python library. To mitigate overfitting and ensure optimal model performance, a uniform training duration of 25 epochs was employed across all architectures. Subsequent sections will provide detailed insights into the specific findings garnered from each model, elucidating their respective contributions and performance metrics. This study explored the effectiveness of 1D CNNs and LSTM networks for sequential modeling tasks. By varying key hyperparameters like No.of convolutional layers for CNN models & the number of LSTM units for LSTM models, several designs were investigated. The CNN models prioritized the assessment of the impact of network depth by maintaining the number of filters and kernel size at 64 and 3, respectively. Hybrid models, amalgamating CNN and LSTM components, utilized the most effective hyper parameter configurations identified in the individual models.

These methodological considerations were instrumental in comprehensively assessing the capabilities and limitations of each architectural variant. Kim et al. [18] achieved an accuracy of 83% utilizing a fine-tuned model proposed by Karim et al. [21]. In contrast, this paper develops a novel architecture for Dog audio classification problem. A 75:25 split ratio separates the training and validation datasets. Next, a model is constructed with Categorical Cross Entropy as loss function and Adam as the optimizer, which is applied uniformly across all tests. To avoid overfitting, the Fully Connected Networks (FCNs) hyperparameters are restricted to values between 1 and 6. Interestingly, the number of LSTM units is intentionally set greater due to the sequences' increased dimensionality during MFCC extraction. This tactical modification aims to enhance model's capacity for depicting data's intricate temporal relationships.

Table 1: Training Metrics at the last epoch for Different Models.

Model	Loss	Accuracy	Precision	Recall
FCN 1	0.033199	0.996875	0.996875	0.996875
FCN 2	0.001539	1.000000	1.000000	1.000000
FCN 3	0.002764	1.000000	1.000000	1.000000
FCN 4	0.008908	0.996875	0.996875	0.996875
FCN 5	0.006699	0.997917	0.997917	0.997917
FCN 6	0.001291	1.000000	1.000000	1.000000
LSTM 64	0.201519	0.939583	0.949947	0.929167
LSTM 124	0.193672	0.939583	0.944444	0.920833
LSTM 212	0.157053	0.959375	0.961175	0.954167
LSTM 300	0.145349	0.951042	0.953927	0.948958
BiLSTM-64	0.000768	1.000000	1.000000	1.000000
BiLSTM-124	0.000266	1.000000	1.000000	1.000000
BiLSTM-212	0.000148	1.000000	1.000000	1.000000
BiLSTM-300	0.000077	1.000000	1.000000	1.000000
CNN LSTM 4 64	0.017736	0.997917	0.997917	0.997917
CNN LSTM 4 212	0.001193	1.000000	1.000000	1.000000
CNN LSTM 4 300	0.011763	0.998958	0.998958	0.998958
CNN BiLSTM 4 64	0.000412	1.000000	1.000000	1.000000
CNN BiLSTM 5 212	0.000137	1.000000	1.000000	1.000000

Source: Authors, (2026).

Table 2: Validation Metrics after Training for Different Models.

Model	Loss	Accuracy	Precision	Recall
FCN 1	0.232302	0.929167	0.929167	0.929167
FCN 2	0.056682	0.979167	0.979167	0.979167
FCN 3	0.149895	0.954167	0.953975	0.950000
FCN 4	0.201651	0.937500	0.937239	0.933333
FCN 5	0.098520	0.975000	0.975000	0.975000
FCN 6	0.082193	0.970833	0.970833	0.970833
LSTM 64	0.233288	0.950000	0.948718	0.925000
LSTM 124	0.150943	0.950000	0.961538	0.937500
LSTM 212	0.222228	0.920833	0.924370	0.916667
LSTM 300	0.209840	0.954167	0.953587	0.941667
BiLSTM-64	0.136918	0.970833	0.970833	0.970833
BiLSTM-124	0.138462	0.970833	0.974895	0.970833
BiLSTM-212	0.133047	0.970833	0.974895	0.970833
BiLSTM-300	0.117315	0.970833	0.970833	0.970833
CNN LSTM 4 64	0.260246	0.945833	0.949580	0.941667
CNN LSTM 4 212	0.131032	0.966667	0.966667	0.966667
CNN LSTM 4 300	0.120291	0.966667	0.966667	0.966667
CNN BiLSTM 4 64	0.100486	0.975000	0.975000	0.975000
CNN BiLSTM 5 212	0.091694	0.970833	0.970833	0.970833

Source: Authors, (2026).

Table 3: F1 Scores for Training data for Different Models.

Model	Barking	Growling	Howling	Whining
FCN 1	0.993684	1.000000	1.000000	0.993815
FCN 2	1.000000	1.000000	1.000000	1.000000
FCN 3	1.000000	1.000000	1.000000	1.000000

FCN_4	0.993711	0.995781	1.000000	0.997930
FCN_5	0.997904	0.997895	0.997938	0.997930
FCN_6	1.000000	1.000000	1.000000	1.000000
LSTM_64	0.941650	0.948454	0.976939	0.889371
LSTM_124	0.962810	0.924051	0.977226	0.893528
LSTM_212	0.962810	0.962656	0.974895	0.936975
LSTM_300	0.972632	0.942387	0.972973	0.916318
BiLSTM-64	1.000000	1.000000	1.000000	1.000000
BiLSTM-124	1.000000	1.000000	1.000000	1.000000
BiLSTM-212	1.000000	1.000000	1.000000	1.000000
BiLSTM-300	1.000000	1.000000	1.000000	1.000000
CNN_LSTM_4_64	0.995798	0.995781	1.000000	1.000000
CNN_LSTM_4_212	1.000000	1.000000	1.000000	1.000000
CNN_LSTM_4_300	0.997904	0.997886	1.000000	1.000000
CNN_BiLSTM_4_64	1.000000	1.000000	1.000000	1.000000
CNN_BiLSTM_5_212	1.000000	1.000000	1.000000	1.000000

Source: Authors, (2026).

Table 4: F1 Scores for Validation data of Different Models.

Model	Barking	Growling	Howling	Whining
FCN_1	0.984127	0.887097	0.904348	0.939130
FCN_2	1.000000	0.967742	0.982759	0.965517
FCN_3	0.992000	0.919355	0.948276	0.956522
FCN_4	0.966667	0.918033	0.949153	0.916667
FCN_5	0.991870	0.968254	0.982759	0.956522
FCN_6	1.000000	0.951613	0.965517	0.965517
LSTM_64	0.967742	0.934426	0.949153	0.948276
LSTM_124	0.976378	0.960000	0.938053	0.921739
LSTM_212	0.939394	0.925620	0.931034	0.882883
LSTM_300	0.983871	0.924370	0.941176	0.966102
BiLSTM-64	0.983871	0.967213	0.957983	0.973913
BiLSTM-124	1.000000	0.950000	0.950000	0.982759
BiLSTM-212	1.000000	0.950000	0.957983	0.974359
BiLSTM-300	0.991870	0.958678	0.957983	0.974359
CNN_LSTM_4_64	0.976000	0.898305	0.941176	0.966102
CNN_LSTM_4_212	0.992000	0.950820	0.966102	0.956522
CNN_LSTM_4_300	0.992000	0.950820	0.966102	0.956522
CNN_BiLSTM_4_64	1.000000	0.958678	0.957983	0.982759
CNN_BiLSTM_5_212	1.000000	0.950000	0.957983	0.974359

Source: Authors, (2026).

The training measures, such as the model's performance at the most recent epoch, are shown in Tables 1 and 2, respectively, along with the model's post-training assessment using the validation data. In Table 3 and 4, the F1 scores for each class are reported, allowing for the examination of potential biases in the trained model. Notably, the FCN-2 and FCN-5 models yield accuracies equivalent to the maximum accuracy achieved by the CNN_BiLSTM_4_64 model. However, these FCN models exhibit less smooth convergence compared to the CNN_BiLSTM_4_64 model, as evidenced by fluctuations during feature learning, as depicted in Figure 6. Bi-LSTM models demonstrate consistent accuracy across various hyper parameter configurations.

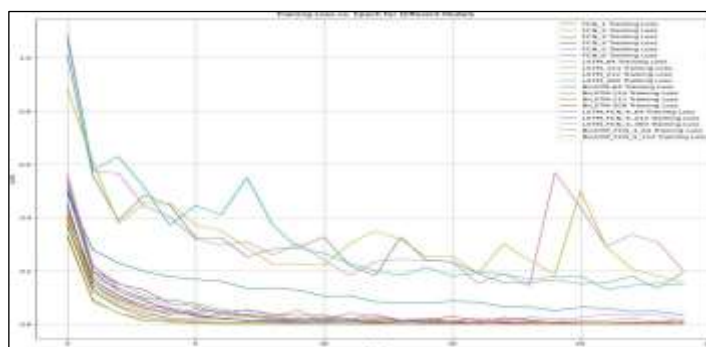


Figure 6: Loss Plots for all the models.

Source: Authors, (2026).

A thorough depiction of the CNN+Bi-LSTM model's classification performance across several classes is provided by Figure 7, which displays the confusion matrix. The Confusion Matrix study reveals significant results about the model's functionality. In particular, it labels "barking" vocalizations with remarkable accuracy and shows no false positives or false negatives. However, there is a discernible level of confusion observed between certain categories, particularly between 'growling' and 'howling', as well as 'howling' and 'whining', indicating inherent acoustic similarities that present challenges for the model's distinction.

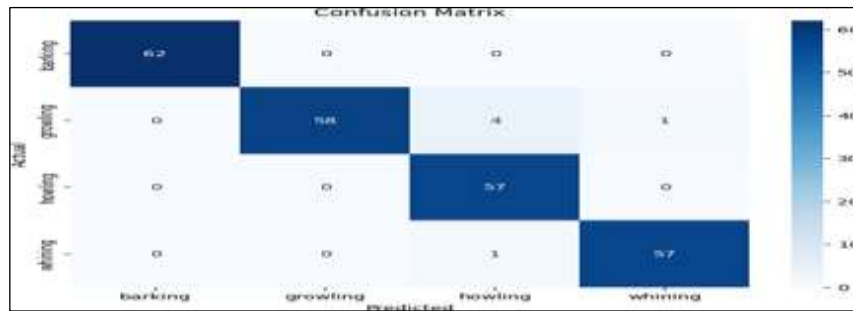


Figure 7: Confusion Matrix.

Source: Authors, (2026).

The model demonstrates robust performance overall, accurately classifying vocalizations with a high rate of true positives and few false classifications. However, targeted improvement is needed, particularly in distinguishing between 'growling' and 'howling', as well as 'howling' and 'whining'. Additional training data or refinement of FE techniques could enhance the model's capabilities in these areas.

V. CONCLUSIONS

The study classifies pet dog vocalization using sound sensor which aim to enhance the understanding canine behaviour and emotions, where dogs are unsupervised. It improves the communication interfaces between dogs and humans through Remote Monitoring Services. It helps in early detection of stress, anxiety .Focusing on the Dachshund breed, our study offers a framework extendable to other breeds. A novel FE technique utilizing MFCC, incorporated into DL architectures. Stacked CNN-BiLSTM model in particular performs exceptionally well in identifying and categorizing Pet Sound Events. Future work will expand dataset to include more breeds, emotional states, and behavioral contexts to enhance accuracy.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Methodology: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Investigation: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Discussion of results: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Writing – Original Draft: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Writing – Review and Editing: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Resources: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Supervision: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

Approval of the final text: Kalaavathi Bhuvaneshwaran , Sujaykumar Reddy M And Arivoli A.

VII. REFERENCES

- [1] Androscoggin Animal Hospital, "2024 Best animal hospital in Topsham, ME," Androscoggin Animal Hospital Website. [Online]. Available: <https://www.androscogginanimalhospital.com>. Accessed: Feb. 20, 2025.
- [2] T. J. Howell, S. Toukhsati, R. Conduit, and P. Bennett, "The perceptions of dog intelligence and cognitive skills (PoDIaCS) survey," *Journal of Veterinary Behavior*, vol. 8, no. 6, pp. 418–424, 2013.
- [3] A. Hussain, S. Ali, Abdullah, and H.-C. Kim, "Activity detection for the wellbeing of dogs using wearable sensors based on deep learning," *IEEE Access*, vol. 10, pp. 53153–53163, 2022.
- [4] T. Vezzosi et al., "Home monitoring of heart rate and heart rhythm with a smartphone-based ECG in dogs," *Veterinary Record*, vol. 184, no. 3, p. 96, 2019.
- [5] R. Kriss, "Most popular dog breeds," AKC Website. [Online]. Available: <https://www.akc.org/most-popular-breeds/>. Accessed: Feb. 20, 2025.
- [6] P. Pongrácz, C. Molnár, A. Dóka, and Á. Miklósi, "Do children understand man's best friend? Classification of dog barks by pre-adolescents and adults," *Applied Animal Behaviour Science*, vol. 135, nos. 1–2, pp. 95–102, 2011.
- [7] A. Larrañaga et al., "Comparing supervised learning methods for classifying sex, age, context and individual Mudi dogs from barking," *Animal Cognition*, vol. 18, no. 2, pp. 405–421, 2015.
- [8] S. Hantke, N. Cummins, and B. Schuller, "What is my dog trying to tell me? The automatic recognition of the context and perceived emotion of dog barks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

- [9] P. Mohandas et al., “Machine learning approach regarding classification and prediction of dog sounds: A case study of south Indian breeds,” *Applied Sciences*, vol. 12, no. 20, p. 10653, 2022.
- [10] N. Jégh-Czinege, T. Faragó, and P. Pongrácz, “A bark of its own kind: The acoustics of ‘annoying’ dog barks suggests a specific attention-evoking effect for humans,” *Bioacoustics*, vol. 29, no. 2, pp. 210–225, 2020.
- [11] W. J. M. Boteju et al., “Deep-learning-based dog behavioural monitoring system,” in *Proc. 3rd Int. Conf. Intelligent Sustainable Systems (ICISS)*, 2020, pp. 82–87.
- [12] A. Vehkaoja et al., “Description of movement sensor dataset for dog behavior classification,” *Data in Brief*, vol. 40, Art. no. 107822, 2022.
- [13] J. Kim and N. Moon, “Dog behavior recognition based on multimodal data from a camera and wearable device,” *Applied Sciences*, vol. 12, no. 6, p. 3199, 2022.
- [14] M. V. Kujala et al., “Time-resolved classification of dog brain signals reveals early processing of faces, species and emotion,” *Scientific Reports*, vol. 10, no. 1, p. 19846, 2020.
- [15] G. Riggio et al., “Quantitative behavioral analysis and qualitative classification of attachment styles in domestic dogs: Are dogs with a secure and an insecure-avoidant attachment different?” *Animals*, vol. 11, no. 1, p. 14, 2020.
- [16] G. Ozen, B. Karan, and T. Cakar, “Predicting animal behaviours: Physical and behavioural classification of dog walking levels,” in *Proc. 30th Signal Processing and Communications Applications Conf. (SIU)*, 2022, pp. 1–4.
- [17] Y. Kim, J. Sa, Y. Chung, D. Park, and S. Lee, “Resource-efficient pet dog sound events classification using LSTM-FCN based on time-series data,” *Sensors*, vol. 18, no. 11, p. 4019, 2018.
- [18] B. Vimal et al., “MFCC-based audio classification using machine learning,” in *Proc. 12th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2021.
- [19] C. Xie, X. Cao, and L. He, “Algorithm of abnormal audio recognition based on improved MFCC,” *Procedia Engineering*, vol. 29, pp. 731–737, 2012.
- [20] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, “1D CNN and applications: A survey,” *Mechanical Systems and Signal Processing*, vol. 151, Art. no. 107398, 2021.
- [21] F. Karim, S. Majumdar, H. Darabi, and S. Chen, “LSTM fully convolutional networks for time series classification,” *IEEE Access*, vol. 6, pp. 1662–1669, 2018.