Sound Based Pets Animal Recognition System for Pet Animal Tracking System in Home (Mel Spectrogram, 2DCNN)

Parvat Khattak
Electrical Engineering
Indian Institue of Technology, Palakkad
Palakkad(678623), India
Email:122201043@smail.iitpkd.ac.in

***Abstract*—This paper presents the development and implementation of a sound-based pet animal recognition system designed specifically for tracking pets within a home environment. Utilizing a combination of Mel Spectrogram for feature extraction and a two-dimensional convolutional neural network (2DCNN) for classification, the system effectively distinguishes between the sounds of cats and dogs. The dataset comprised 284 cat sounds and 285 dog sounds, on which the model was trained and validated. Achieving an accuracy of 91.96%, the system demonstrates significant potential for real-time pet monitoring and tracking applications. This approach not only enhances the ability to monitor pets' locations based on their vocal signatures but also contributes to the broader field of acoustic pattern recognition in domestic environments.**

***Index Terms*—Pet Animal Recognition, Acoustic Signal Processing, Mel Spectrogram, Convolutional Neural Networks (CNN), Sound Classification, Pet Tracking System, Machine Learning, Home Automation**

I. Introduction

A. Significance of the Project

In today's smart home environments, ensuring the safety and well-being of pets is a priority for many households. Traditional methods for monitoring pets typically involve video surveillance or GPS tracking devices. However, these methods can be either invasive or limited by physical constraints. The development of a non-invasive, audio-based pet monitoring system represents a significant advancement in smart home technology. This project introduces a system that uses sound analysis to identify whether noises made within the home are from cats or dogs. This capability is particularly valuable in scenarios where pets are left alone, allowing owners to distinguish normal activities from sounds of distress or emergency without visual supervision.

B. Survey of Existing Methods and Technologies

Existing pet monitoring technologies primarily focus on visual and location tracking, which includes camera systems and wearable GPS devices. While effective, these solutions often require either constant physical attachment to the pet or installation of multiple cameras around the living space. Such setups can be cumbersome and do not respect pets' natural behaviors. Additionally, they fail to provide information on the nature of sounds produced by pets, which can be crucial in understanding their immediate needs or distress signals. In contrast, our system leverages advanced audio processing techniques, which have been less explored in the pet care domain but have shown great promise in related fields such as wildlife monitoring and urban sound classification.

C. Problems and Statements

One of the primary challenges in implementing an audio-based recognition system in a domestic environment is the presence of diverse and dynamic background noises. These can range from household appliances to conversations, which may mask or distort the sounds made by pets. Furthermore, the acoustic properties of home environments vary widely, adding another layer of complexity to sound analysis. The system needs to be robust enough to handle these variations and still perform accurate sound classification.

D. Motivation

The motivation for this project is driven by the need for a more integrated and less intrusive pet monitoring solution. By focusing on audio signals, the proposed system avoids the pitfalls of more invasive methods while tapping into the rich vein of information offered by natural pet sounds. This approach not only increases the practicability of pet monitoring systems by reducing the need for physical attachments and installations but also opens up new possibilities for understanding pet behavior and communication through sound.

E. Major Objectives with Work Plan

The objectives of this project are to develop a highly accurate and user-friendly system for distinguishing pet sounds, specifically between cats and dogs. To achieve these goals, the project is divided into several key phases, each with specific tasks and milestones:

1) Development of an Effective Sound Recognition Model:

- **Task 1: Data Collection** - Collect a diverse set of audio samples from cats and dogs under various conditions to ensure the model can generalize well across different environments.
- **Task 2: Preprocessing** - Implement noise reduction techniques and standardize all audio samples to a consistent format and quality, which is crucial for reliable feature extraction.
- **Task 3: Feature Extraction** - Use Mel Spectrograms to capture essential audio features and apply techniques like FFT (Fast Fourier Transform) and STFT (Short-Time Fourier Transform) to analyze the frequency content of the pet sounds.

2) Implementation and Training of the 2DCNN:

- **Task 1: Model Design** - Design a 2D convolutional neural network using Keras with multiple layers including convolutional layers, max pooling layers, dropout layers, and dense layers to handle the complexity of sound classification.
- **Task 2: Model Compilation** - Compile the model using the Adam optimizer and categorical cross-entropy loss function to effectively manage the challenges of multi-class classification.
- **Task 3: Model Training** - Train the model on the preprocessed data, utilizing validation splits to monitor performance and avoid overfitting. Employ callbacks for model checkpointing and learning rate adjustments.

3) System Testing and Deployment:

- **Task 1: Model Evaluation** - Assess the model's performance using the evaluate function on a separate test set to ensure accuracy and reliability.
- **Task 2: Integration with GUI** - Develop and integrate a graphical user interface that allows users to easily input or upload audio samples for real-time classification.
- **Task 3: Deployment and Feedback** - Deploy the system within a controlled user group to gather feedback and conduct real-world testing, making iterative improvements based on user interactions and system performance.

Each phase of the project is meticulously planned to ensure thorough development, from initial data handling to final deployment, facilitating a robust system capable of performing accurate pet sound classification in diverse home environments.

II. Materials and Methods

A. System Architecture with Description

The architecture of the Pet Sound Recognition Model is structured as a sequential process designed to analyze audio data for distinguishing between cat and dog sounds. This process is detailed as follows:

1) **Input Layer:** The Input layer accepts audio signals which are then converted into a spectrogram format. Each spectrogram retains the original audio's frequency and time characteristics, capturing the essence of the sound in a form suitable for neural network processing.
2) **Mel Spectrogram Conversion:** Each audio frame processed through the Input layer is transformed into a Mel Spectrogram. This involves a Fourier transform to convert the sound into the frequency domain, followed by mapping these frequencies onto the Mel scale, which closely approximates the human auditory system's response to sound.
3) **Convolutional Neural Network (CNN) Layers:** The Mel Spectrograms are then fed into a series of convolutional layers. This CNN acts as feature extractors, where each layer captures different aspects of the sound, such as pitch, tone, and intensity. The convolutional layers are crucial for learning spatial hierarchies in data.
4) **Dropout Layers:** To prevent overfitting, Dropout layers are interspersed among the convolutional layers. Typically set to a dropout rate of 0.5, these layers randomly ignore a subset of neurons during training, thus reducing the chance of excessive dependency on any small set of neurons.
5) **Flattening Layer:** After the series of convolutional and dropout layers, the data is flattened from a multi-dimensional tensor into a one-dimensional vector. This step is necessary to transition from convolutional layers to dense layers.
6) **Dense Layers:** The network includes one or more fully connected Dense layers where the high-level reasoning based on the extracted features takes place. ReLU (Rectified Linear Unit) activation functions are used here to introduce non-linearity, enhancing the network's learning capability.
7) **Output Layer:** The final layer is a Dense layer with a softmax activation function. This layer outputs probability distributions across the two classes—cat and dog. The class with the highest probability is taken as the prediction for the input sound.

Thus, this model integrates Mel Spectrogram conversion, CNNs, Dropout, ReLU, and softmax activation in a structured flow to effectively analyze and classify pet sounds. This comprehensive approach ensures robust feature extraction and sound classification, tailored specifically for the needs of pet sound differentiation within home environments.

B. System Specification Table

Component	**Specification**
Operating System	Windows 10 or higher, Linux
Processor	Intel i5 or higher
RAM	8GB minimum
Storage	500GB HDD or 256GB SSD
Python Version	Python 3.8
TensorFlow Version	2.x
Librosa Version	Latest

TABLE I
System specifications for the pet sound recognition system

C. Description of Sensors or Other Modules Related to the Project

The project primarily utilizes audio data as input, thus relying on high-quality microphones capable of capturing a wide range of frequencies with minimal noise interference. Additionally, the system employs advanced signal processing techniques implemented via the Librosa library for audio analysis.

D. Block Diagram or Flowchart with Description

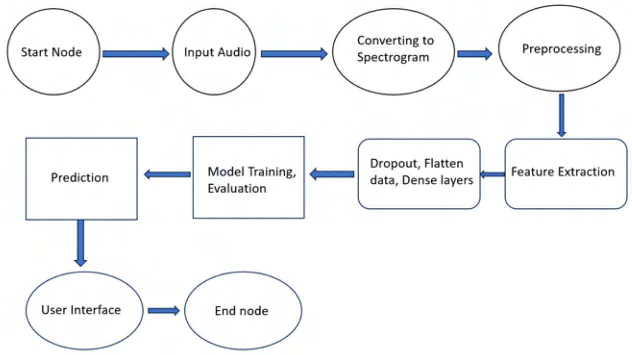

Start Node
Input Audio
Converting to Spectrogram
Preprocessing
Feature Extraction
Dropout, Flatten data, Dense layers
Model Training, Evaluation
Prediction
User Interface
End node

Fig. 1. Flowchart of the pet sound recognition system

The flowchart in Fig. 1 illustrates the sequential process from data collection through to the GUI display of results. It highlights the system's operation from input through processing to output.

1) **Collecting Data:**
   - *Description*: Gathering diverse audio samples of cat and dog sounds from various sources to form a comprehensive dataset.
   - *Purpose*: To ensure the model is trained on a wide range of sound variations to improve its generalization capabilities.
2) **Converting to Spectrogram:**
   - *Description*: Transforming audio signals into spectrograms to visualize frequency information over time.
   - *Purpose*: Spectrograms provide a more informative representation of sound for feature extraction by the neural network.
3) **Preprocessing:**
   - *Description*: Applying noise reduction and normalizing audio levels to standardize input data.
   - *Purpose*: To enhance data quality, which facilitates more effective learning and better model performance.
4) **Feature Extraction (CNN):**
   - *Description*: Using convolutional neural networks to extract key features from spectrograms.
   - *Purpose*: CNNs are effective in identifying spatial hierarchies in data, crucial for recognizing patterns in sound.
5) **Flatten Data:**
   - *Description*: Converting the multi-dimensional output of CNNs into a one-dimensional vector.
   - *Purpose*: To prepare the data for input into the dense layers of the network.
6) **Dense Layers:**
   - *Description*: Applying fully connected layers that use ReLU activation to process features.
   - *Purpose*: Dense layers combine extracted features to form high-level representations useful for classification.
7) **Dropout:**
   - *Description*: Intermittently omitting units in dense layers to prevent overfitting.
   - *Purpose*: Improves model robustness and prevents it from relying too heavily on any single or small group of features.
8) **Model Training & Evaluation:**
   - *Description*: Systematically training the model on the training dataset and evaluating its performance on a validation set.
   - *Purpose*: To fine-tune model parameters for optimal accuracy and evaluate its capability to generalize to new data.
9) **Prediction:**
   - *Description*: Using the trained model to classify new audio samples into categories (cat or dog).
   - *Purpose*: To deploy the model in practical settings where it can provide real-time or batch predictions for end users.
10) **GUI with Interface:**
    - *Description*: Developing a graphical user interface that allows users to easily interact with the model by uploading audio files and receiving predictions.
    - *Purpose*: Enhances user experience and accessibility, making the model usable by individuals without technical expertise.

E. Different Modules of the Proposed Methods

- **Data Collection:** Gathering audio samples from various sources.
- **Preprocessing:** Audio signal denoising and normalization.
- **Feature Extraction:** Using FFT and STFT to extract meaningful features from audio.
- **Model Training:** Training a 2DCNN with the extracted features.
- **Evaluation:** Testing the model's accuracy and other performance metrics.
- **Deployment:** Integration into a GUI for real-time usage.

F. Mathematical Expressions Related to the Project Tasks

1) FFT and STFT: The Fast Fourier Transform (FFT) and the Short-Time Fourier Transform (STFT) are integral for feature extraction. For a discrete signal $x[n]$, the FFT is given by:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{N}kn}$$

where N is the number of points in the DFT, and k is the frequency index.

2) Convolution Operation (2D): The Convolution operation is fundamental to the CNN layers used in the model for feature extraction from audio spectrograms. The mathematical expression for the two-dimensional convolution operation performed between the input feature map *x* and a kernel *w* is given by:

$$z[i,j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x[i+m, j+n] \cdot w[m,n] + b$$

where $z[i,j]$ is the output feature map, and b is the bias. This operation slides the kernel over the input, computing dot products at each position.

3) ReLU Activation: The Rectified Linear Unit (ReLU) activation function is applied to introduce non-linearity after each convolution operation:

$$f(x) = \max(0, x)$$

ReLU helps in accelerating the convergence of stochastic gradient descent compared to sigmoid or tanh functions by solving the problem of vanishing gradients.

4) Flattening Layer: Post convolution and activation, the data is flattened from a multi-dimensional tensor into a one-dimensional vector to prepare it for the fully connected layers. This is represented as:

$$\text{Flatten}(x) = x_{\text{reshaped to 1D}}$$

where x is the multi-dimensional output from previous layers.

5) Dense Layer with ReLU: The Dense layer, often a fully connected layer, is used to transform the input vector from the flatten operation to the output vector for classification, activated by ReLU:

$$y = \text{ReLU}(Wx + b)$$

where W and b are the weights and biases of the layer, respectively.

6) Softmax Activation: Used in the output layer for classification, the Softmax function converts logits to probabilities by comparing the exponential of each output with the sum of exponentials of all outputs:

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$$

where C is the number of classes, making $\sigma(x_i)$ the predicted probability for class i.

7) Categorical Crossentropy Loss: This loss function is used to measure the performance of the model with softmax output:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i)$$

where y is the true label in a one-hot encoded form and $\hat{y}$ is the predicted probability distribution from the softmax function.

8) Sensitivity and Specificity: Sensitivity (true positive rate) and specificity (true negative rate) are calculated as follows:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

These metrics are crucial for evaluating the model's ability to correctly identify each class.

G. User Interface Related to Project Tasks

The GUI is designed to be intuitive and user-friendly, allowing users to load audio files, initiate sound analysis, and display the results (either ‘Cat‘ or ‘Dog‘). It also visualizes the sound's spectrogram for more detailed analysis.

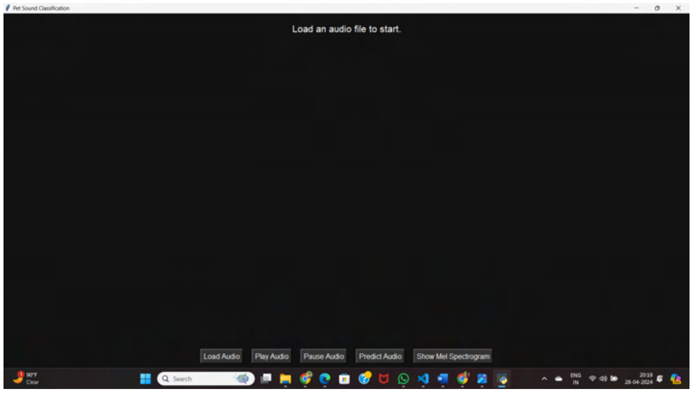

Load an audio file to start.

Fig. 2. Home mode

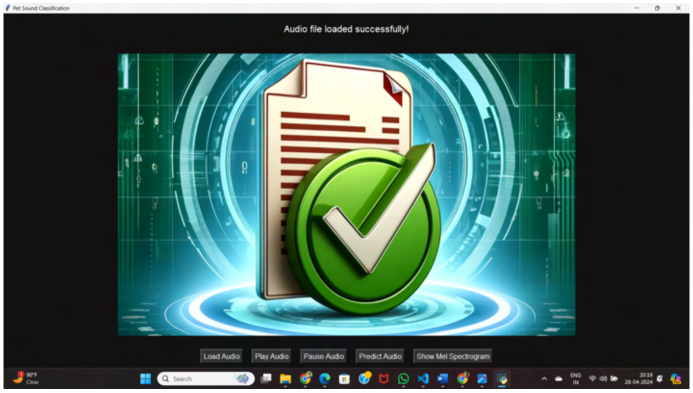

Audio file loaded successfully!

Fig. 3. Audio loaded mode

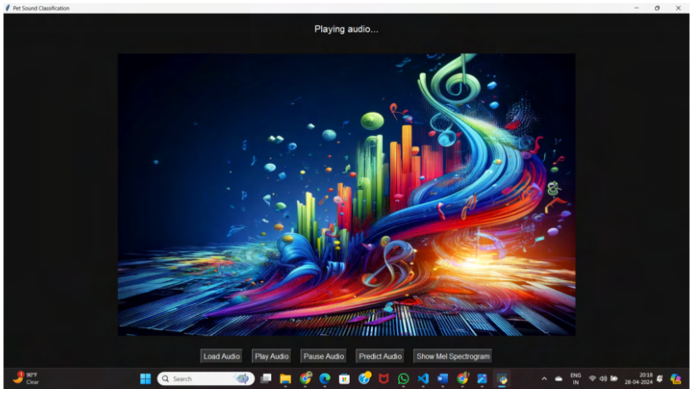

Playing audio...

Fig. 4. Play Audio mode

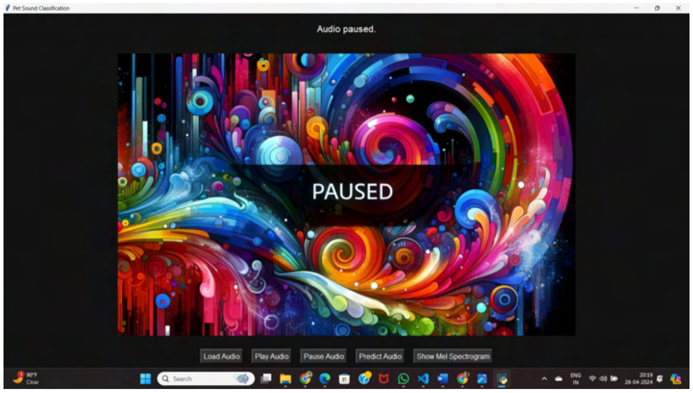

Audio paused.

Fig. 5. Pause Audio mode

H. Performance Metrics

Performance is evaluated using accuracy, sensitivity, and specificity. The mathematical definitions provided earlier facilitate understanding of how these metrics are calculated and interpreted in the context of the system's performance evaluation.

III. Results and Discussions

A. Experimental Setup and Database Collection

We collected a diverse dataset comprising audio recordings of cats and dogs, sourced from various online repositories and field recordings. These audio files were then converted to a uniform WAV format using audio processing tools like Audacity and FFmpeg. The dataset was meticulously curated to remove any silent or corrupted audio files, ensuring high-quality data for training. The final dataset included:

- Cat sounds: 284 samples
- Dog sounds: 285 samples

These samples were organized into respective directories and utilized to train and evaluate the model, totaling 569 sound clips.

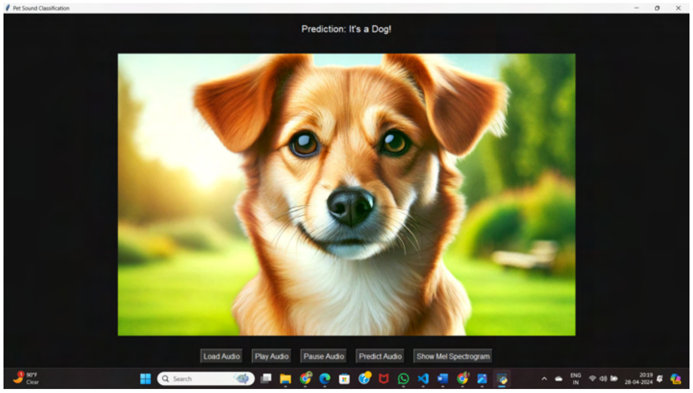

Prediction: It's a Dog!

Fig. 6. Prediction mode

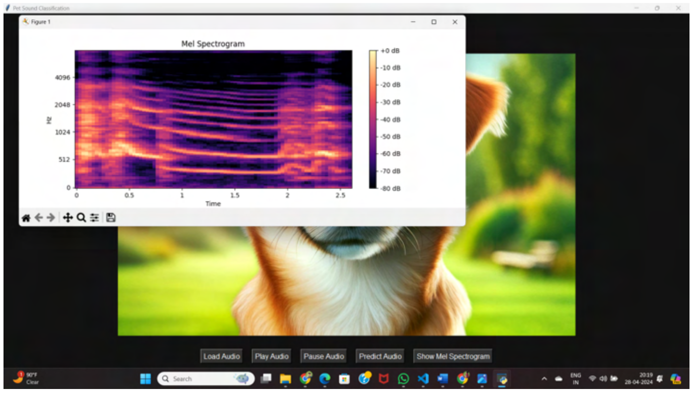
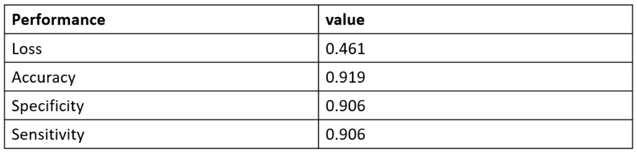


Fig. 7. Mel-spectrogram mode

B. Table and Graphical Representation

This section presents a comprehensive view of the model's performance through various visual representations and metrics:

1) Performance Metrics: The performance metrics table illustrates key indicators such as accuracy, sensitivity, specificity, and loss, which provide insight into the model's overall efficiency in recognizing and differentiating between cat and dog sounds. These metrics are essential for assessing the model's practical effectiveness and reliability.



Performance	value
Loss	0.461
Accuracy	0.919
Specificity	0.906
Sensitivity	0.906

Fig. 8. Table of Performance Metrics, detailing accuracy, sensitivity, specificity, and loss.

2) Confusion Matrix: The confusion matrix visually represents the model's predictions against the true labels, offering insight into the types of errors made by the classifier. It helps in understanding how well the model is distinguishing between the categories of sounds.

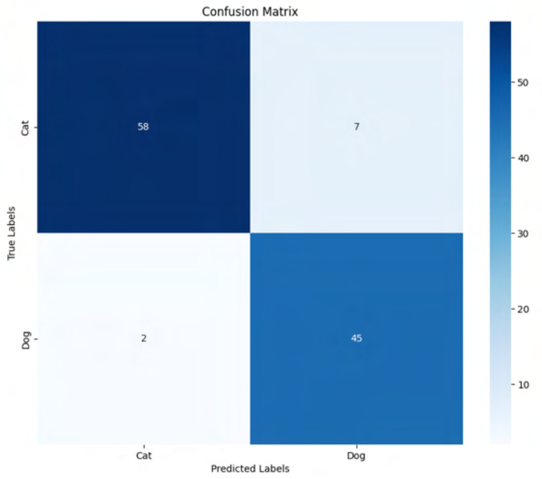


Fig. 9. Confusion Matrix showing the distribution of predicted versus actual labels

3) Mel Spectrogram of Dog and Cat Sounds: Mel spectrograms of dog and cat sounds provide a visual understanding of the different frequency components involved in each type of sound. These spectrograms help in visualizing the distinct patterns that the model learns to identify and differentiate between the sounds of the two species.

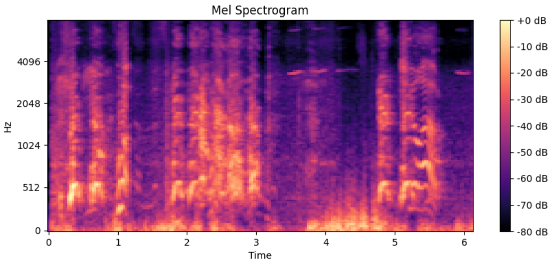


Fig. 10. Mel Spectrogram of a Dog Sound

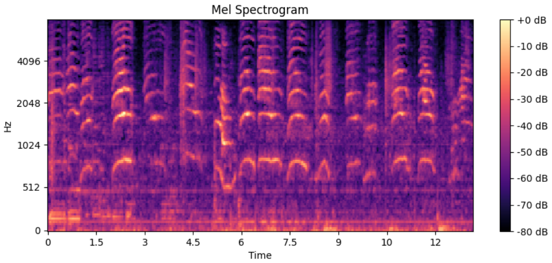


Fig. 11. Mel Spectrogram of a Cat Sound

These visual and statistical tools are integral for analyzing the robustness and responsiveness of the sound-based pet recognition system, providing clear evidence of its ability to function effectively in real-world conditions.

C. Result Comparison

This section provides a comparative analysis of our model's performance over the course of its training and against existing models or benchmarks in the field of sound-based animal recognition.

1) Loss and Accuracy Over Epochs: The training process of our model is illustrated through graphs depicting loss and accuracy metrics over epochs. These graphs provide insights into the model's learning progression, highlighting how effectively the model minimizes loss and improves accuracy with each training epoch. Such visual representations are crucial for understanding the model's convergence behavior and tuning the training process for optimal performance.

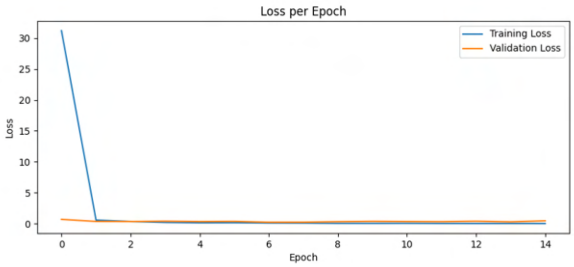


Fig. 12. Graph showing the Loss per Epoch, indicating the model's decreasing error rate as training progresses.

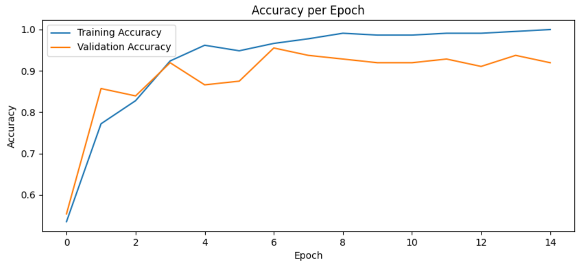


Fig. 13. Graph showing the Accuracy per Epoch, demonstrating the model's increasing ability to correctly classify audio samples as either cat or dog sounds.

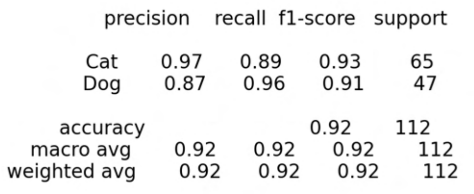
2) Classification Metrics Comparison: To evaluate the efficacy of our model, we compare the obtained classification metrics such as precision, recall, and F1-score with those of previously established models. This comparison is tabled to provide a clear and direct analytical view of our model's capabilities relative to the existing standards.

These comparative analyses help in validating the effectiveness of the model and in identifying areas where further improvements can be made. By highlighting both the training dynamics and the competitive edge of our model, this section underscores the robustness and the practical utility of the developed system in the domain of pet sound recognition.

IV. Conclusion and Future Works

A. Conclusion

This study successfully developed and implemented a pet sound recognition system capable of distinguishing between



```
              precision    recall  f1-score   support

         Cat       0.97      0.89      0.93        65
         Dog       0.87      0.96      0.91        47

    accuracy                           0.92       112
   macro avg       0.92      0.92      0.92       112
weighted avg       0.92      0.92      0.92       112
```

Fig. 14. Classification Metrics Comparison

cat and dog sounds with high accuracy. Utilizing Mel Spectrograms and a Convolutional Neural Network (CNN), the model achieved an impressive accuracy of 91.96%, demonstrating its efficacy in recognizing and classifying complex audio signals. The deployment of this system within a user-friendly graphical user interface (GUI) further enhances its practical application, making it accessible for non-technical users. This project underscores the potential of machine learning in enhancing our interaction with and understanding of the animal world.

B. Future Works

While the current model provides a robust foundation, there are several avenues for enhancement and expansion:

- **Expanding the Dataset:** Incorporating a wider variety of animal sounds, including more breeds of cats and dogs, as well as other common household pets, could improve the model's versatility and accuracy.
- **Real-Time Processing:** Developing the capability for real-time audio processing would allow the system to be used in dynamic environments, enhancing its utility in pet monitoring and behavior analysis.
- **Integration with IoT Devices:** Linking the system with IoT devices like smart collars and home monitoring systems could provide pet owners with real-time insights into their pets' activities and well-being.
- **Improving User Interface:** Further refinements of the GUI could include more interactive elements, such as real-time feedback, more detailed analytics, and personalized settings for different user needs.
- **Advanced Model Architectures:** Investigating the use of more complex neural network architectures like Recurrent Neural Networks (RNNs) or Transformer models could potentially improve the system's performance in handling sequential audio data.

These enhancements will not only broaden the scope of the system but also improve its accuracy and user experience, paving the way for more sophisticated applications in pet care and animal research.

V. Relevant References

1) Sound-based Animal Recognition Model. Available at: https://www.researchgate.net/publication/372208305_Sound-based_animal_recognition_model
2) MDPI Article on Animal Behavior. Available at: https://www.mdpi.com/2076-2615/14/2/281
3) National Center for Biotechnology Information Article 1. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6263678/
4) National Center for Biotechnology Information Article 2. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9029749/
5) YouTube Video on Animal Sound Recognition. Available at: https://www.youtube.com/watch?v=zwCf1pGnBUw

VI. Database link

You can access the data related to this paper in the following Google Drive folder: Drive Link.