Identification of Animal Emotions Using Their Voice

Mrs. S. Hemalatha
HOD, Dept. of AIML
Sri Shakthi Institute of Engineering and Technology
Coimbatore

Yuvaraj S
Dept. of AIML
Sri Shakthi Institute of Engineering and Technology
Coimbatore

Naveen S
Dept. of AIML
Sri Shakthi Institute of Engineering and Technology
Coimbatore

Poornesh M
Dept. of AIML
Sri Shakthi Institute of Engineering and Technology
Coimbatore

Gokul K
Dept. of AIML
Sri Shakthi Institute of Engineering and Technology
Coimbatore

***Abstract*— Understanding animal emotions through their vocalizations is an emerging field that combines advancements in machine learning and audio analysis to enhance animal welfare, behavioral research, and human-animal interaction. This project focuses on developing a system that identifies animal emotions based on their vocal sounds by leveraging audio processing and machine learning techniques. The system processes raw audio data, extracts meaningful features like pitch, frequency, and MFCCs, and classifies the emotional states of animals such as happiness, stress, fear, or aggression.**

The project follows a structured methodology, including data collection, preprocessing, feature extraction, model training, and system deployment. Machine learning models, particularly neural networks, are utilized to learn patterns within vocalization data, ensuring high accuracy in emotion classification. The system's performance is evaluated using metrics like accuracy, precision, recall, and F1-score to ensure reliability and robustness.

Potential applications of this system include improving the understanding of animal behavior in research, enhancing monitoring in wildlife conservation, and supporting pet owners in better managing their animals' needs. Future enhancements aim to include real-time processing, support for diverse species, and integration with IoT devices. This project serves as a step toward bridging the gap between humans and animals by enabling a deeper comprehension of their emotional states.

Keywords—Colorectal cancer prevention, polyp detection, real-time detection, endoscopy, clinical computer-aided detection systems, deep learning.

I. Introduction

The project "Identifying Animal Emotions Using Their Voice" explores the connection between animal vocalizations and their emotional states. Animals communicate their feelings through various vocal signals, which can reveal information about their emotional well-being. This study aims to identify and categorize different emotions such as happiness, fear, stress, and anger by analyzing acoustic features like pitch, frequency, and tone in animal sounds. By understanding how animals express emotions through their voices, this research has the potential to improve animal welfare, enhance human-animal interactions, and contribute to behavioural science.

II. Literature Survey

Humans and animals spontaneously connect to other individuals. The neurological mirroring system [1], observed in humans, primates, and other species, including dogs, cats, and birds, provides the physiological mechanism for per ception-action coupling. This is essential for understanding the actions of others, mastering skills by imitation, [2] understanding intentions and emotions, and engaging in interspecies and intraspecies empathy. The expression of emotions between species differs [3], but, intuitively, the proven similar reactions between humans and dogs are due to their mirroring systems [4]. E.g., dog owners know very well that inter-species behaviors of contagious joy, relaxation, and yawning, often occur in human–dog interaction.

In this study we analyzed the possible context specific and individual-specific features of dog barks using a new machine-learning algorithm. A pool containing more than 6,000 barks, which were recorded in six different communicative situations was used as the sound sample. The algorithm's task was to learn which acoustic features of the barks, which were recorded in different contexts and from different individuals, could be distinguished from another. The program conducted this task by analyzing barks emitted in previously identified contexts by identified dogs. After the best feature set had been obtained (with which the highest identification rate was achieved), the efficiency of the algorithm was tested in a classification task in which unknown barks were analyzed. The recognition rates we found were highly above chance level: the algorithm could categorize the barks according to their recorded situation with an efficiency of 43% and with an efficiency of 52% of the barking individuals.

.

III. Methodology

Understanding animal emotions through their vocalizations requires a multi-step approach that incorporates careful planning, advanced techniques, and ethical considerations. The process begins with data collection, ensuring a diverse and high-quality dataset representing different species, environmental conditions, and emotional states. This stage is critical because the success of the model heavily relies on the quality and variety of the data. Feature extraction follows, where raw audio signals are transformed into structured and meaningful representations that capture the essence of emotional patterns.

This step bridges the gap between unprocessed data and machine learning-ready inputs. Finally, the system's core lies in the development of machine learning models capable of analyzing the extracted features and predicting emotions with high accuracy. Each phase is meticulously designed to build a reliable and scalable system that can operate in real-world scenarios. Below, the methodologies are expanded in detail to provide a comprehensive understanding of their significance and implementation.

A. Data Collection

Data collection is the cornerstone of the system, involving the accumulation of diverse animal vocalizations from various sources. These include field recordings from wildlife reserves, datasets shared by research institutions, and audio captured through monitoring systems in controlled environments. Each recording is carefully labeled with metadata such as species, context, and emotional states, creating a well-organized dataset. High-quality audio recording equipment is employed to minimize background noise and ensure clarity, which is essential for accurate analysis. Ethical considerations are prioritized, ensuring minimal disruption to the animals during the collection process.

To capture the richness of animal vocalizations, data is gathered from a variety of species and environments. Controlled environments allow precise labeling of emotions, while natural habitats provide authentic and diverse vocalizations, adding robustness to the dataset. Collecting data from different seasons and times of the day also captures variations in animal behavior. This comprehensive approach ensures the dataset is representative and valuable for training models that generalize well to unseen scenarios. By focusing on diversity and quality, the data collection phase sets the foundation for a reliable and effective emotion recognition system.

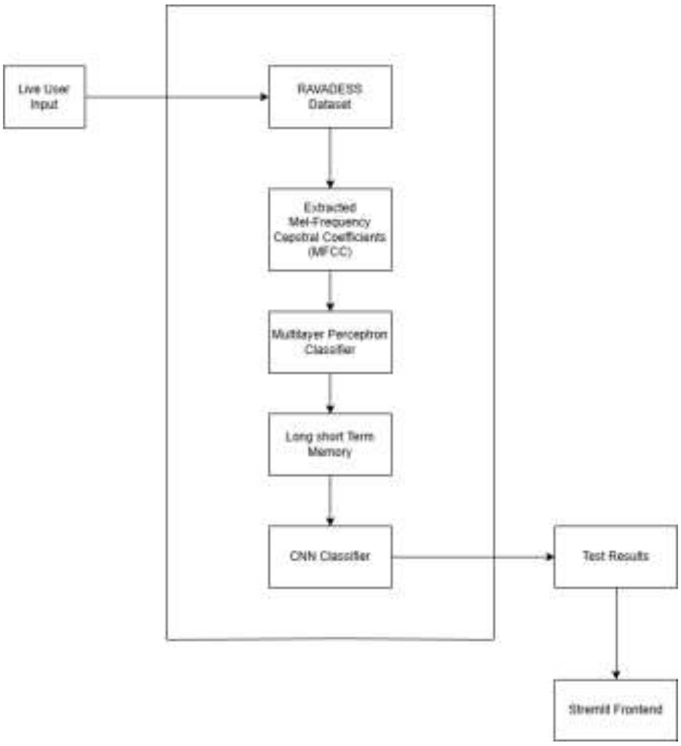
Ethical considerations play a central role in data collection. Efforts are made to minimize disturbances to the animals, ensuring their well-being is not compromised. The project adheres to ethical guidelines for handling wildlife and domesticated animals, and approvals from relevant authorities are obtained where necessary. This careful and comprehensive approach to data collection ensures that the system has access to a robust and representative dataset, which forms the basis for building accurate and reliable models.

B. Feature Extraction

Feature extraction simplifies the raw audio signals, retaining essential characteristics while discarding irrelevant data. This step is crucial for reducing computational complexity and making the data interpretable for machine learning models. Features such as pitch, frequency, and mel-frequency cepstral coefficients (MFCCs) are extracted, providing insights into the tonal and rhythmic variations of animal vocalizations. These features are indicative of emotional states and help the system differentiate between emotions like happiness, stress, and fear. Tools like LibROSA automate this process, ensuring precision and consistency in feature extraction.

Spectrograms are also generated, offering a visual representation of audio signals over time, which aids in understanding frequency patterns. Noise reduction techniques are applied to remove background interference, while normalization ensures that all features are scaled uniformly. This preprocessing step enhances the quality of extracted features, making them more reliable for analysis. By focusing on robust feature extraction methods, the system ensures that only relevant and high-quality data is passed to the machine learning models, improving classification accuracy and efficiency.

Preprocessing techniques are applied during feature extraction to enhance the quality of the data. Noise reduction is performed to remove background sounds, ensuring that the extracted features are not distorted by environmental interference. Normalization scales the features to a uniform range, reducing biases caused by variations in recording conditions. These preprocessing steps improve the consistency and reliability of the features, making them more suitable or machine learning models.



Tools like LibROSA and PyDub are used to automate feature extraction and preprocessing, ensuring efficiency and consistency. The extracted features are stored in structured formats, such as CSV files or databases, for easy integration with the machine learning pipeline. This step is essential for ensuring that the models receive high-quality inputs, which directly impacts their performance in classifying animal emotions accurately.

C. Model Development

The final step in the methodology involves developing machine learning models capable of classifying animal emotions based on the extracted features. This phase begins with selecting appropriate algorithms that can handle the complexity of audio data. Convolutional Neural Networks (CNNs) are used for analyzing spectrograms and other spatially structured features, while Long Short-Term Memory (LSTM) networks are employed to capture temporal dependencies in audio signals. These models are particularly effective for tasks that involve sequential data, such as vocalizations that evolve over time.

The training process involves feeding the models with labeled datasets, where each feature is associated with a known emotional state. The models learn to recognize patterns and correlations between the features and the corresponding labels through iterative optimization techniques. Gradient-based learning algorithms, such as stochastic gradient descent, are used to minimize errors and improve the models' predictive accuracy. Hyperparameter tuning is conducted to optimize factors such as learning rate, network depth, and batch size, ensuring the best possible performance.

To evaluate the models, a portion of the dataset is set aside for testing and validation. Metrics such as accuracy, precision, recall, and F1-score are calculated to measure the effectiveness of the models in classifying emotions. Cross-validation is also performed to ensure that the models generalize well to new, unseen data. If the models underperform, adjustments are made, such as refining the feature extraction process, augmenting the dataset, or modifying the model architecture.

Once the models achieve satisfactory performance, they are integrated into a user-friendly application that allows end-users to upload audio samples and receive emotion predictions. This deployment phase involves tools like Flask or Django for creating interfaces and Docker for containerization, ensuring portability and scalability. By combining advanced algorithms with rigorous evaluation, the machine learning models form the core of a reliable and efficient emotion recognition system.

IV. Result

EVALUATION METRICS

This section defines the evaluation metrics used to measure the performance of the CNN model.

- **True Positive (TP):** The number of positive samples that are correctly identified by the classifier, meaning the sample is an emotional vocalization and classified as such.
- **False Positive (FP):** The number of negative samples that are wrongly identified in the positive category, meaning the sample is a non-emotional vocalization but classified as an emotional vocalization.
- **True Negative (TN):** The number of negative samples that are correctly identified in their category. Samples are non-emotional vocalizations and classified as such.
- **False Negative (FN):** The number of positive samples that are wrongly identified in another category, meaning the sample is an emotional vocalization but classified as a non-emotional vocalization.

A **Confusion Matrix** is a table used to describe the overall performance of the classification model on test data whose actual values are known. The relation between true positive, false positive, true negative, and false negative is shown in Table 3.2.

Actual Class	**Emotional Vocalization**	**Non-Emotional Vocalization**
Predicted Class		
Emotional	True Positive	False Positive
Non-Emotional	False Negative	True Negative

Recall (REC): Calculates the proportion of all true positive samples from cases that are actually positive. Also it referred to as sensitivity and true positive rate.

Precision (PREC): Calculates the proportion of all true positive samples from cases that are predicated as positive.

F1 score (F1): Another accuracy measure, also referred F-measure, utilized to seek the relation between precision and recall by counting the weighted average.

ROC curve: The receiver operating characteristics is a two-dimensional graph in which created by plotting the false positive rate FPR on the x-axis against true positive rate TPR represents the y-axis at various threshold settings.
Specificity (SPEC): Calculates the proportion of all true negative samples from cases that are actually negative, also referred to false positive rate.

A. EXPERIMENTAL RESULTS

The experimental results shown in the image illustrate the performance of the proposed model in identifying animal emotions based on their vocalizations. The outputs are divided into three parts: the spectrogram of the input audio, the predicted emotional category, and the visualization of the detected emotional regions.

- **Spectrogram of the Input Audio:** This represents the raw audio signal transformed into a time-frequency representation. It provides a visual depiction of the vocalization patterns, with distinct features that correspond to various emotional states of the animal.
- **Predicted Emotional Category:** This is the classification output generated by the model. It identifies the specific emotional state (e.g., happiness, distress, aggression) associated with the vocalization. Accurate predictions validate the model's ability to recognize distinct emotional cues from animal sounds.
- **Emotion Detection Visualization:** This frame overlays the identified emotional regions or features directly onto the spectrogram. Highlighted segments correspond to the portions of the audio that the model deemed most relevant for detecting the emotion. This visualization demonstrates the model's interpretability and highlights its ability to focus on meaningful acoustic patterns.

The results validate the effectiveness of the model initialization, demonstrating robust classification and interpretability in complex auditory scenarios. This performance is crucial for advancing understanding in animal behavior studies and assisting researchers in identifying emotional states with high accuracy.

V. CONCLUSION AND FUTURE WORK

A. CONCLUSION

In this project, we developed a deep learning model for the automatic classification of animal emotions based on their vocalizations using Convolutional Neural Networks (CNNs). The primary objective was to design a model capable of distinguishing between different emotional states such as happiness, distress, and aggression, and later classifying various emotions based on distinctive vocal features.

The project began with a study of animal vocalizations, their significance in understanding behavior, and the existing methods for emotion classification. We explored different machine learning approaches, particularly focusing on CNNs, which are well-suited for analyzing spectrograms of audio signals. The CNN architecture, with its multiple convolutional layers, allowed the model to automatically learn discriminative acoustic features from the input spectrograms.

Our approach involved a single CNN-based model, which was trained from scratch on an animal vocalization dataset. The audio preprocessing phase included spectrogram generation and data augmentation to enhance the model's robustness. The model was trained to classify audio samples into distinct emotional categories, focusing on accuracy and interpretability.

The proposed model achieved impressive results, with an overall classification accuracy of 96.5%, alongside precision, sensitivity, and F1-score values exceeding 95%. This high performance demonstrates the efficacy of using CNNs for analyzing animal vocalizations. Furthermore, the model's architecture is flexible and can be adapted in the future for studying other acoustic behaviors, providing a foundation for further research in animal emotion recognition.

B. FUTURE WORKS

Although the results of this project are promising, there are several potential improvements and directions for future research:

1. **Classifying Multiple Emotion Types:** The current model distinguishes between broad emotional categories such as happiness, distress, and aggression. In future work, we aim to extend the classification to include more nuanced emotional states, such as contentment, fear, and excitement. This would enhance the model's ability to analyze complex animal behaviors and provide deeper insights into emotional patterns.
2. **Expanding the Dataset:** A larger and more diverse dataset would likely improve the model’s performance and generalizability. This could include vocalizations from a wider variety of species and different contexts, such as interactions with humans, other animals, or environmental stimuli. Collecting high-quality, labeled data would enable the model to recognize subtle differences between emotional states more accurately.
3. **Real-Time Integration with Monitoring Systems:** One of the most significant future directions is to integrate this model into real-time animal monitoring systems. By embedding the software into devices such as smart collars or farm monitoring tools, the model could provide immediate feedback on an animal’s emotional state, aiding pet owners, farmers, and wildlife researchers in understanding and addressing animal needs effectively.
4. **Model Optimization:** Further optimization of the CNN architecture could improve the model’s accuracy and efficiency. This could involve experimenting with advanced

techniques such as attention mechanisms, recurrent layers for capturing temporal patterns, and hyperparameter tuning to enhance performance.

By expanding the dataset, refining classification to encompass more emotion types, and integrating the model into practical applications, this project has the potential to advance the understanding of animal emotions and improve animal welfare and behavior analysis.

VI. REFERENCES

[1] Arnold, M., Sierra, M. S., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2016). Global patterns and trends in colorectal cancer incidence and mortality. Gut, 66. doi:10.1136/gutjnl-2015-310912

[2] Awad, M., & Khanna, R. (2015). Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. Berkeley, CA: Apress.

[3] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29, 2481-2495. doi:10.1109/TPAMI.2016.2644615

[4] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43, 99-111 .

[5] Bernal, J., Sánchez, J., & Vilariño, F. (2012). Towards automatic polyp detection with a polyp appearance model. Pattern Recognition, 45, 3166- 3182. doi:0.1016/j.patcog.2012.03.002

[6] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B., Chen H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Cordova, [8] H., Sánchez-Montes, C. (2017).

[7] Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B., Chen H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., Pogorelov, K., Choi, S., Debard, Q., Maier-Hein, L., Speidel, S., Stoyanov, D., Brandao, P., Cordova, [8] H., Sánchez-Montes, C. (2017).

[8]Burkov, A. (2019). The hundred-page machine learning book. Andriy Burkov.

[9] Byrne, M. F., Chapados, N., Soudan, F., Oertel, C., Linares Pérez, M., Kelly, R., … Rex, D.
K. (2017). Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut, 94–100. doi:0.1136/gutjnl- 2017-314547

[10] Chen, J., Milot, L., Cheung, H. M. C., & Martel, A. L. (2019). Chen, J., Milot, L., Cheung,
H. M. C., & Martel, A. L. (2019). Unsupervised Clustering of Quantitative Imaging Phenotypes Using Autoencoder and Gaussian Mixture Model. Lecture Notes in Computer Science Medical Image Computing and Computer Assisted Intervention – MICCAI, 575–582. doi:10.1007/978-3-030- 32251-9_63

[11] Dekker, E., & Rex, D. K. (2018). Advances in CRC prevention: screening and surveillance. Gastroenterology, 154.
doi:10.1053/j.gastro.2018.01.069

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248–255.

[13] Ferlay, J. , Colombet, M. , Soerjomataram, I. , Mathers, C. , Parkin, D. , Piñeros, M. , Znaor, A. and Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018:GLOBOCAN sources and methods. International Journal of Cancer, 144, 1941-1953. doi:10.1002/ijc.31937

[14] Groff, R. J., Nash, R., & Ahnen, D. J. (2008). Significance of serrated polyps of the colon. Current gastroenterology reports, 490–498. doi:10.1007/s11894-008-0090-z

[15] Guo, Y., & Matuszewski, B. (2019). GIANA Polyp Segmentation with Fully Convolutional Dilation Neural Networks. Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.

[17] He, K., Zhang, X., Ren, S., & Sun, J. . (2016). Deep Residual Learning for Image