www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

PREDICTION OF ANIMAL VOCAL EMOTIONS USING CONVOLUTIONAL NEURAL NETWORK

Dr. Thanveer Jahan[1*], Masampelly Sathvik[2], Varikollu Naveen[2], Sangani Saiteja[2], Sathuri Sakaleshwar[2]

[1]Associate Professor & Head, [2]UG Student, [1,2]Department of CSE(AI&ML)

[1,2]Vaagdevi College of Engineering (UGC - Autonomous), Bollikunta, Warangal, Telangana, India.

*Corresponding Email: Dr. Thanveer Jahan (thanvivcecse@gmail.com)

ABSTRACT

Animal vocalization has long intrigued researchers, particularly for its potential to reveal emotional states and enhance human-animal interaction. Understanding emotions through vocal cues can significantly improve animal welfare, inform behavioral studies, and foster better communication with pets. Traditionally, emotion recognition in animals relied on manual observation and acoustic analysis—focusing on features like pitch, frequency, and duration. While useful, these methods were limited by subjectivity, human error, and insufficient analytical depth to capture the complexity of animal emotions. Additionally, the variability in vocalization patterns across species and individuals posed challenges, leading to inconsistent results and limiting the reliability of such systems.

To overcome these limitations, researchers are now turning to machine learning, especially Convolutional Neural Networks (CNNs), for more accurate and scalable emotion recognition. CNNs excel at identifying intricate patterns in large datasets and are well-suited for analyzing spectrograms of animal sounds. The proposed system leverages CNNs to automatically classify emotions in animal vocalizations by training on a comprehensive dataset of labeled sound samples. This approach not only reduces human bias but also improves prediction accuracy by learning nuanced differences in vocal patterns.

By automating the emotion recognition process, this deep learning-based system offers a robust solution to the shortcomings of traditional techniques. Its potential applications span veterinary diagnostics, pet behavior monitoring, and the advancement of human-animal communication. Ultimately, such innovations promise to transform how we understand and care for animals by providing deeper insight into their emotional well-being.

Keywords: animal welfare, acoustic analysis, communication, machine learning, Convolutional Neural Networks (CNNs)

1. INTRODUCTION

Animal vocalization has long been studied by scientists and animal behaviorists as a vital means of understanding non-verbal emotional states in animals. This is particularly significant in countries like India, where over 192 million cattle and 105 million buffaloes are integral to agriculture and dairy farming. Recognizing emotions such as stress or discomfort through vocal cues can directly impact productivity, health, and welfare. Traditionally, animal emotion detection relied on manual observation and basic acoustic analysis, which were often inconsistent, subjective, and incapable of capturing complex, context-dependent vocal patterns. The diversity among species and individual animals further limited the effectiveness of these traditional methods, making real-time emotional analysis nearly impossible.

Driven by the growing demand for precision and automation, researchers are increasingly adopting deep learning techniques, particularly Convolutional Neural Networks (CNNs), to analyze animal sounds. CNNs are capable of learning subtle variations in pitch, frequency, and modulation, making them ideal for classifying emotional states with high accuracy. The motivation behind this research lies in overcoming the limitations of earlier approaches and harnessing AI to enhance veterinary diagnostics, animal training, and welfare monitoring. Real-time emotion detection holds immense value in livestock management, pet care, wildlife research, and animal rehabilitation. It enables early identification of distress or illness, supports effective training, and contributes to better living conditions in shelters and zoos. This study aims to develop a CNN-based system capable of recognizing and predicting animal emotions from vocalizations, offering a transformative tool for improving animal care and deepening human-animal understanding.

2. LITERATURE SURVEY

Speech is the most natural way of human communication. Affective computing systems based on speech play an important role in promoting human–computer interaction, and emotion recognition is the first step. Due to the lack of a precise definition of emotion and the inclusive and complex influence of emotion generation and expression, accurately recognizing speech emotions is still difficult. Speech emotion recognition (SER) is an important problem that is receiving increasing interest from researchers due to its numerous applications, such as e-learning [1], clinical trials [2], audio monitoring/surveillance, lie detection [3], entertainment, video games [4], and call centers [5]. Machine learning (ML) is a revolutionary method in which we feed a machine an adequate amount of data, and the machine will use the experience gained from the data to improve its own algorithm and process data better in the future [6]. One of the most significant approaches in machine learning is the use of neural networks (NNs). Neural networks are networks of interconnected nodes and are loosely modeled towards the way the human brain processes information. Neural networks store data, learn from it, and improve their abilities to sort new data. For example, a neural network with the task of identifying dogs can be fed a set of characteristic values extracted from various images of dogs tagged with the type of dog. Over time, it will learn what kind of image corresponds to what kind of dog. The machine therefore learns from experience and improves itself. Deep learning (DL) is a predominant machine learning approach, where neural networks are enabled at nodes to remember past values with many layers that are trained using massive amounts of data. In deep learning, the sprawling artificial neural network is fed representations of raw data (e.g., raw image representations) and not given any other instructions. This means that, in contrast to other machine learning approaches, it determines the important characteristics and purpose of the data itself while storing it as experience. In other words, according to studies, deep neural networks (DNNs) can solve the data representation problem through learning a series of task-specific transformations [7]. The network layers extract abstract representations and filter out the irrelevant information, which leads to a more accurate classification and better generalization. Temporal models were also proposed for modelling sequential data with mid- to long-term dependencies. Deep learning models are currently used to solve problems such as face recognition, voice recognition, image recognition, computational vision, and speech emotion recognition. One of the main advantages of deep learning techniques over other machine learning techniques is the automatic selection of features, which could, for example, be applied to important features inherent in audio files that have a special emotion in the task of recognizing speech emotions.

When it comes to recognizing emotion through speech, deep learning models, such as convolutional neural networks (CNN), deep neural networks (DNNs), deep belief networks (DBNs), etc., approach the detection of high-level features for better accuracy compared to hand-made low-level features. Furthermore, the use of deep neural networks enhances the computational complexity of the entire model. However, according to Mustaqeem and Kwon [8] there are still many challenges in recognizing emotion from speech, such as the fact that the current CNN architectures have not shown significant improvement in speech accuracy and complexity in speech signal processing, or the fact that the use of recurrent neural networks (RNNs) and long short-term memory neurons (LSTMs) is useful for training sequential data, but they are difficult to train effectively and are computationally more complex.

In the literature, there is a huge research interest, and several works attempt to perform emotion detection from speech [9,10]. Various works study the way that emotions can be automatically identified and accurately recognized in speech data [11,12]. In this regard, deep learning techniques have achieved breakthrough performance in recent years, and as a result, have been thoroughly examined by the research community [13,14]. Many existing studies in the literature have focused on improving and extending deep learning techniques [15].

In the work presented in [16], the authors present a new random deep belief network (RDBN) method for speech emotion recognition, which consists of a random subspace, DBN and SVM in the context of ensemble learning. It first extracts the low-level characteristics of the input speech signal and then applies them to the construction of many random sub-intervals. Second, it creates many different sub-intervals. In addition, the DBN continues to use the stochastic gradient descent method to optimize the parameters. To solve the problem, a random space is applied for the training of the basic classifiers as a whole, where the same classification method is used. The best accuracy achieved is 82.32% on the Emo-DB database, 48.5% on the CASIA database, 48.5% on the FAU database, and 53.60% on the SAVEE database.

In the work presented in [17], the authors introduce a method for identifying speech emotions using a spectrogram and convolutional neural network (CNN). The proposed model consists of three convolution layers and three fully connected layers, which extract distinctive features from spectrograph images and predictions for the seven emotions of the Emo-DB Database. Layer C1 has 120 cores (11 × 11) applied at a rate of four pixels. The ReLU acts as an activation function instead of the standard sigmoid functions that improve the efficiency of the educational process. Layer C2 has 256 cores of size 5 × 5 and is applied to the input with one step. Similarly, C3 has 384 cores of size 3 × 3. Each of these convolution layers are followed by ReLUs. Layer C3 is followed by three FC layers that have 2048, 2048, and 7 nodes, respectively. More than 3000 spectrograms were generated from all the audio files in the dataset. Overall, the proposed method achieved 84.3% accuracy.

In [18], the authors present two convolutional neural networks with a long-short memory network (CNN-LSTM), one one-dimensional (1D) and one two-dimensional (2D), stacking four designed local features learning blocks (LFBL). The 1D CNN-LSTM network is intended to recognize the feeling of speaking from raw audio clips, while the 2D CNN-LSTM network focuses on learning high-level capabilities from log-Mel spectrograms. The experimental study was conducted on the Berlin Emo-DB and IEMOCAP databases. The 1D CNN LSTM network achieved 92.34% and 86.73% recognition accuracy on the speaker-dependent and speaker-independent EmoDB databases, respectively, and also delivered 67.92% and 79.72% recognition accuracy on the IEMOCAP speaker-dependent and speaker-independent databases, respectively. The 2D CNN LSTM network achieved 95.33% and 95.89% recognition accuracy on the speaker-dependent and speaker-independent Emo-DB databases, respectively, and delivered 89.16% and 85.58% recognition accuracy on the IEMOCAP speaker-dependent and speaker-dependent experiment databases, respectively.

In the work presented in [19], the authors proposed a new approach to the multimodal recognition of emotions from simple speech and text data. The attention network implemented consists of three separate convolutional neural networks (CNNs), two for extracting features from speech spectrograms and word integration sequences and one for the emotion classifier. The CNN outputs from word integration and spectrograms are used to calculate an attention matrix to represent the correlation between word integration and spectrograms in relation to emotion signaling. To evaluate the model, they used audio and text data from the CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset. The dataset is organized by video IDs and corresponding segments with six emotion and sentiment labels. The video IDs are then further split into segments. The training set consisted of 3303 video IDs and 23,453 segments, while the validation set consisted of 300 non-overlapping video IDs and 1834 segments. The total accuracy of the proposed method was 83.11%.

In [20], the authors present three methods based on CNNs in combination with extensive features, a CNN + RNN and ResNet, respectively. The authors investigate different types of features as the end-to-end frame input, including primary wave data, the Q-transform constant spectrogram (CQT), and the Fourier transform short-term spectrogram (STFT). In this way, the authors create multiple data samples with slightly modified speed ratios, which helps them achieve significant improvements and handle the overfitting issue in the framework from end to end. For their experiments, they used the EmotAsS dataset. The CNN + RNN model achieved the best performance (45.12%) with data balancing; the CNN model in combination with features showed a performance of 34.33% with data balancing, while the ResNet model achieved a performance of 37.78%.

In the work presented in [21], the authors propose a new architecture called attention-based 3-dimensional convolutional recurrent neural networks (3-D ACRNN) for recognizing emotion from speech, combining CRNN with an attention mechanism, because they hypothesized that calculating delta and delta–deltas for individual functions not only retains effective emotional information but also reduces the effect of emotionally unrelated factors, leading to a reduction in misclassification. First, the CNN 3-D is applied to the entire logarithmic-Mel spectrogram, which has been compiled into a patch that contains only multiple frames. The attention layer then takes a sequence of high-level attributes as the input to generate expression-level attributes. The authors evaluated the model using the Berlin Emotional Speech Database (Emo-DB) and IEMOCAP database. From the ten speakers, for each evaluation, they selected eight as the training data, one as the validation data, and the rest as the test data. The method achieved an accuracy of 64.74% on IEMOCAP database and 82.82% on Emo-DB.

In the work presented in [22], the authors propose an attention-pooling representation learning method for recognizing emotions from speech (SER). Emotional representation is learned from end to end by applying a deep convolutional neural network (CNN) directly to speech spectrograms extracted from speech. Compared to existing aggregation methods, such as max pooling and average pooling, the proposed attention pooling can effectively integrate bottom–up class-agnostic attention maps and top–down class-specific attention maps. Given an expression, they segment it into 2 s sections for training and use an overlay of 1 s to allow them to receive more training data. Each section corresponds to the same tag with the corresponding expression. They used a 1 × 1 convolutional layer after Conv5 to create a top–down attention map and used another 1 × 1 convolutional layer to create bottom–up attention maps. The IEMOCAP improvised dataset was used, and the accuracy achieved by the proposed method was 71.75% for WA and 68.06% for UA.

In [23], the authors explore how to take full advantage of low-level and high-level audio features taken from different aspects and how to take full advantage of DNN's ability to merge multiple information to achieve better classification performance. For this reason, they proposed a hybrid platform consisting of three units, namely, a features extraction unit, a heterogeneous unification unit, and a fusion network unit. Besides low-level acoustic features, such as IS10, MFCCs, and eGemaps, that are extracted, high-level acoustic feature presentations named SoundNet bottleneck feature and VGGish bottleneck feature are considered for speech emotion recognition tasks. The heterogeneous integration unit is a Denoising

AutoEncoder (DAE), which is a multilayer feed-forward neural network and is introduced in order to convert the heterogeneous space of various features into a unified representation space by deploying this unsupervised feature learning technique. The fusion network module is utilized to capture the associations between those unified joint features for emotion recognition tasks and is constructed as a four-layer neural network, containing one input layer and three hidden layers. They evaluated the model using the IEMOCAP database, and the proposed method improved the recognition performance, reaching an accuracy of 64%.

In the work presented in [24], the authors propose a platform that, at the training layer, has three main stages, such as verbal/non-verbal audio segmentation, the integration of feature extraction, and the construction of an emotion model. The verbal sections were used to train the CNN-based emotion model to derive emotion features, while the non-verbal sections were used to train the CNN audio model to extract audio features. The CNN's combined features are used as the input to the LSTM-based sequence-to-sequence emotion recognition model. Here, the sequence-to-sequence model based on the LSTM with an attention mechanism was selected for emotion recognition. The LSTM and the attention mechanism for developing a sequence emotion recognition model contained a bidirectional LSTM (Bi-LSTM) as the coder for the attention mechanism and an unidirectional LSTM as the decoder for the emotional sequence output. They evaluated the model using the NTHU-NTUA Chinese interactive multimodal emotion corpus (NNIME); the proposed method achieved a 52.0% accuracy.

The work presented in [25] introduces a model that includes one-dimensional convolutional layers combined with dropout, batch-normalization, and activation layers. The first layer of their CNN receives 193 × 1 number arrays as the input data. The initial layer is composed of 256 filters with a kernel size of 5 × 5 and a stride of 1. After that, batch normalization is applied, and its output is activated by a rectifier linear units layer (ReLU). The next convolutional layer, consisting of 128 filters with the same kernel size and stride, receives the output of a previous input layer. The final convolutional layer, with the same parameters, is followed by the flattening layer and dropout layer, with a rate of 0.2. Their model was tested in the Berlin (EMO-DB), IEMOCAP, and RAVDESS databases and obtained 71.61% for the RAVDESS with eight classes, 86.1% for EMO-DB with 535 samples in seven classes, 95.71% for EMO-DB with 520 samples in seven classes, and 64.3% for IEMOCAP with four classes on speaker-independent audio classification tasks.

Attention-oriented parallel convolutional neural network encoders that capture the essential features required for emotion classification are introduced in [26]. The authors extracted and encoded features such as paralinguistic information and speech spectrogram data, and distinct CNN architectures were designed for each type of feature, and those encoded features were subsequently passed through attention mechanisms to enhance their representations before undergoing classification. Empirical evaluations were carried out on the EMO-DB and IEMOCAP open datasets, and the proposed model achieved a weighted accuracy (WA) of 71.8% and an unweighted accuracy (UA) of 70.9%. Furthermore, with the IEMOCAP dataset, the model yielded WA and UA recognition rates of 72.4% and 71.1%, respectively.

The authors in [27] present a work on enhancing the overall generalization performance and accuracy of SER with a balanced augmented sampling technique on spectrograms that aims to address the imbalance in sample distribution among emotional categories. A deep neural network is utilized, comprising the combination of a convolutional neural network (CNN) and an attention-based bidirectional long short-term memory network (ABLSTM) for feature extraction. Multitask learning is incorporated to enhance the deep neural network's performance. The methodology is assessed on the IEMOCAP and MSP-IMPROV databases, yielding a weighted average recall and unweighted average recall of 70.27% and 66.27% on the IEMOCAP database, respectively, while on the MSP-IMPROV database, the approach achieves 60.90% and 61.83%, respectively.

In the work presented in [28], the authors introduce a method to enhance SER performance by viewing Mel Frequency Cepstral Coefficients (MFCC), which accelerates the learning process while maintaining a high level of accuracy. The authors employ a supervised learning model, specifically a functional support vector machine (SVM), directly on the MFCCs represented as functional data. This enables the utilization of complete functional information, resulting in more precise emotion recognition. The authors' method demonstrates competitive results in terms of accuracy, underscoring its effectiveness in emotion recognition as well as reducing learning time, making it computationally efficient and practical for real-world applications.

A framework called Convolutional Auto-Encoder and Adversarial Domain Adaptation (CAEADA) for cross-corpus SER is introduced in [29]. The CAEADA framework starts by creating a one-dimensional convolutional auto-encoder (1D-CAE) for feature processing. This 1D-CAE is designed to capture correlations among adjacent one-dimensional statistical features, and the feature representation is enhanced through an encoder–decoder-style architecture. Following this, the adversarial domain adaptation (ADA) module works to reduce the differences in feature distributions between the source and target domains by confusing a domain discriminator. Specifically, it employs the maximum mean discrepancy (MMD) method to achieve effective feature transformation. The evaluation results demonstrate that the method performs quite satisfactorily on SER tasks.

In the work presented in [30], the authors present an attention-based dense long short-term memory (LSTM) approach for speech emotion recognition. The authors integrate LSTM networks, which are well-suited for handling time series data such as speech, with attention-based dense connections. This entails the incorporation of weight coefficients into skip connections for each layer, enabling the differentiation of emotional information across layers and preventing interference from redundant information in the lower layers with valuable information from upper layers. The experiments showcase an improvement in recognition performance by 12% and 7% on the eNTERFACE and IEMOCAP datasets, respectively.

3. PROPOSED METHODOLOGY

The proposed methodology utilizes Convolutional Neural Networks (CNNs) for emotion classification from animal vocalizations. First, a dataset of labeled animal sounds is collected and preprocessed into spectrograms, representing the frequency and amplitude of audio over time. These spectrograms serve as input features for the CNN model. The model is trained to recognize patterns associated with various emotional states such as stress, calmness, or excitement. Data augmentation techniques are applied to improve model generalization and robustness. The CNN architecture includes convolutional, pooling, and fully connected layers optimized through backpropagation. Model performance is evaluated using metrics like accuracy, precision, recall, and F1-score. Once trained, the model is integrated into a real-time application that receives animal audio input. The system analyzes incoming sounds and classifies emotions on the spot. This approach enables scalable, objective, and real-time monitoring of animal emotional well-being.

Step 1: Dataset

Collect a diverse dataset of animal vocalizations, including different emotional states such as angry, distressed, or content sounds. Ensure the dataset covers various species and environments for robust model generalization. Label each audio sample according to its emotional category to facilitate supervised learning. Preprocess the dataset by removing noise and standardizing the format (e.g., sampling rate and duration) to maintain consistency.

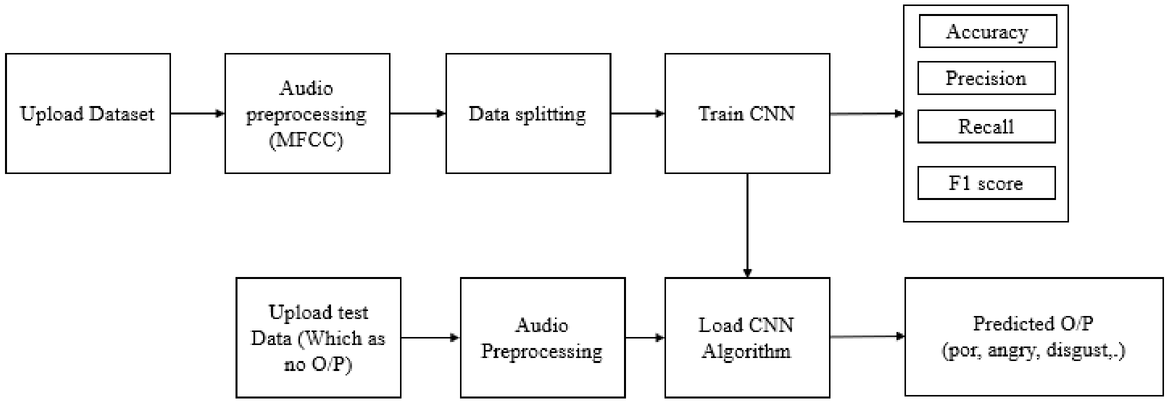

Fig. 1: Block Software

Step 2: Audio Preprocessing (MFCC Feature Extraction)

Convert raw audio signals into numerical representations using Mel Frequency Cepstral Coefficients (MFCCs), which effectively capture the frequency distribution of animal sounds. First, segment the audio into small frames to analyze short-time spectral properties. Apply windowing to minimize spectral leakage, then compute the power spectrum using the Fast Fourier Transform (FFT). Convert the spectrum into the Mel scale to mimic human auditory perception and extract cepstral coefficients. Normalize the extracted MFCC features to standardize input data, improving the CNN’s ability to learn relevant patterns and reducing variability caused by different recording conditions.

Step 3: Proposed Algorithm (CNN)

Design a Convolutional Neural Network (CNN) model to classify animal vocalizations based on MFCC features. Construct multiple convolutional layers to learn spatial hierarchies, followed by pooling layers to reduce dimensionality and computational complexity. Use batch normalization to stabilize training and dropout layers to prevent overfitting. Fully connected layers map extracted features to output categories, with a softmax activation function for classification. Train the model using a labeled dataset, employing an appropriate optimizer like Adam and categorical cross-entropy loss function. Fine-tune hyperparameters, including learning rate, batch size, and number of filters, to achieve optimal classification performance.

Step 4: Performance Evaluation

Assess the model's accuracy, precision, recall, and F1-score using a test dataset to ensure generalization. Utilize confusion matrices to analyze misclassification rates and identify areas for improvement. Conduct k-fold cross-validation to validate model stability across different subsets of data. Compare the CNN's performance with traditional machine learning classifiers like Support Vector Machines (SVM) or Random Forest to highlight deep learning advantages. Measure real-time inference speed for practical applications. If necessary, refine the model by augmenting the dataset, optimizing hyperparameters, or incorporating attention mechanisms to enhance the classification of complex and overlapping vocalizations.

3.2 Data Splitting & Audio Preprocessing

First, split the dataset into training, validation, and testing sets, typically in a 80:10:10 ratio, ensuring a balanced distribution of different vocalization classes. Use stratified sampling to maintain class proportions across splits. Next, preprocess audio by converting it into a uniform format (e.g., same sampling rate and duration). Apply noise reduction techniques and normalize amplitudes to enhance consistency. Extract Mel Frequency Cepstral Coefficients (MFCCs) as features, segmenting audio into short frames, applying windowing, and transforming signals into the Mel scale. Normalize extracted features before feeding them into the CNN model for robust classification.

3.3 CNN Model Building

To build a CNN model for audio classification, first, design a deep learning architecture that can process MFCC features as input. Use a sequential model with multiple convolutional layers (e.g., 2D Conv layers) to extract spatial patterns from the MFCC spectrograms. Apply batch normalization and ReLU activation to enhance feature learning. Add max-pooling layers to reduce dimensionality and prevent overfitting. Use dropout for regularization. Flatten the feature maps and pass them through dense layers. The final output layer uses a softmax activation function for classification. Compile the model with categorical cross-entropy loss and optimize using Adam. Train using labeled audio data.

3.3.1 Proposed Algorithm: Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a deep learning model designed for processing structured grid data, such as images and audio spectrograms. CNNs are particularly effective in pattern recognition tasks by automatically learning spatial hierarchies of features. Unlike traditional neural networks, CNNs reduce the need for manual feature extraction by learning features directly from raw input data. They use convolutional layers to detect patterns, pooling layers to reduce spatial dimensions, and fully connected layers for classification.

3.4 Flask

This project is a web-based application built using the **Flask framework** in Python. The goal is to allow users to upload an animal sound (in .wav format) and receive a prediction about the emotional state of the animal based on its vocalization. The backend of the application leverages a pre-trained **Convolutional Neural Network (CNN)** model which has been trained on a dataset of animal sounds categorized by emotions such as *anger, fear, happy, sad*, and so on. This makes the system capable of interpreting various emotional tones from vocal input, aiding in animal behavior studies or intelligent monitoring systems. The application begins by setting up the Flask environment and configuring the upload folder to store incoming audio files. It restricts uploads to only .wav format to ensure compatibility with the audio processing pipeline. The trained CNN model is loaded into memory using TensorFlow's load_model() function. Additionally, a LabelEncoder is initialized and fitted with the list of target emotion labels, which allows the system to map predicted numerical outputs back to human-readable emotion categories. To process the uploaded audio, the application uses the **Librosa** library, a popular tool for audio analysis in Python. Specifically, it extracts **MFCC (Mel-Frequency Cepstral Coefficients)** features from each audio file. MFCCs are a standard feature set used in audio and speech recognition tasks because they capture the timbral texture of sounds, which is important for distinguishing between emotional tones. The extracted features are padded or trimmed to maintain a uniform input size of 40 x 100 before being reshaped to a 4D tensor suitable for the CNN (1 x 40 x 100 x 1).Once the features are prepared, the model makes a prediction and the class with the highest probability is selected as the output. This predicted class index is then decoded back into the corresponding emotion using the label encoder. This entire process is encapsulated within a predict_emotion() function that returns the final emotion result as a string. The user interface consists of two simple HTML pages: index.html and result.html, placed within a templates/ folder. The index.html file provides a form for users to upload their .wav file. Upon submission, the form sends the file to the Flask backend using a POST request. The backend then processes the audio, predicts the emotion, and renders the result.html page, passing the emotion result dynamically using Flask's templating engine (Jinja2). This result page simply displays the predicted emotion and provides a link to go back and test another file.

4. RESULTS AND DISCUSSION

4.1 Dataset Description

The dataset comprises 400 WAV audio files capturing vocal expressions of animals across four distinct emotional states: Anger, Disgust, Fear, and Purr (representing Pleasure or Contentment), with each category containing 100 audio recordings. The Anger category includes aggressive vocalizations such as growls, roars, barks, or screeches, often reflecting defensive or territorial behavior and characterized by high intensity and irregular waveform patterns. Disgust features vocal reactions to unpleasant stimuli, including gagging, whining, or low growls, typically marked by abrupt and sharp tonal variations. Fear is represented by distress calls, whimpers, or shrieks that are usually high-pitched with rapid frequency modulation, often used to signal danger. The Purr category contains soft, rhythmic vibrations like gentle coos, hums, or murmurs, generally steady and low-frequency, associated with relaxed or content states. These files vary in duration from 1 to 10 seconds and are recorded in WAV format with a typical sampling rate of 44.1 kHz or 16 kHz, 16-bit PCM bit depth, and either mono or stereo channels depending on the source. The dataset is suitable for a range of applications, including training machine learning models for animal emotion recognition, conducting bioacoustic studies on animal communication, integrating into AI-driven animal welfare monitoring systems, supporting wildlife conservation through vocalization analysis, and contributing to emotion-based audio synthesis for entertainment or academic research.

4.2 Result analysis

```
Epoch 43/50
320/320 [==============================] - 1s 5ms/sample - loss: 0.1744 - acc: 0.9375 - val_loss: 0.1816 - val_acc: 0.9625
Epoch 44/50
320/320 [==============================] - 2s 5ms/sample - loss: 0.1581 - acc: 0.9375 - val_loss: 0.1853 - val_acc: 0.9625
Epoch 45/50
320/320 [==============================] - 1s 5ms/sample - loss: 0.1714 - acc: 0.9375 - val_loss: 0.1762 - val_acc: 0.9625
Epoch 46/50
320/320 [==============================] - 1s 5ms/sample - loss: 0.1650 - acc: 0.9375 - val_loss: 0.1845 - val_acc: 0.9625
Epoch 47/50
320/320 [==============================] - 2s 5ms/sample - loss: 0.1588 - acc: 0.9375 - val_loss: 0.1872 - val_acc: 0.9625
Epoch 48/50
320/320 [==============================] - 2s 5ms/sample - loss: 0.1611 - acc: 0.9375 - val_loss: 0.1875 - val_acc: 0.9625
Epoch 49/50
320/320 [==============================] - 2s 5ms/sample - loss: 0.1680 - acc: 0.9375 - val_loss: 0.2016 - val_acc: 0.9625
Epoch 50/50
320/320 [==============================] - 2s 5ms/sample - loss: 0.1629 - acc: 0.9375 - val_loss: 0.2055 - val_acc: 0.9625
New CNN model trained and saved as animal_emotion_model.h5!
```

Fig. 2: Epochs

Fig 2 shows that the training log shows the performance of a deep learning model over 50 epochs for classifying animal vocal emotions. The model trained on 320 samples, achieving a final training accuracy of **93.75%** and validation accuracy of **96.25%**, indicating good generalization. The loss values (which measure error) remained low, with the final training loss at **0.1629** and validation loss at **0.2055**, suggesting that the model effectively learned patterns without significant overfitting.

```
80/80 [==============================] - 0s 2ms/sample - loss: 0.2055 - acc: 0.9625
Model Test Accuracy = 96.25%
```

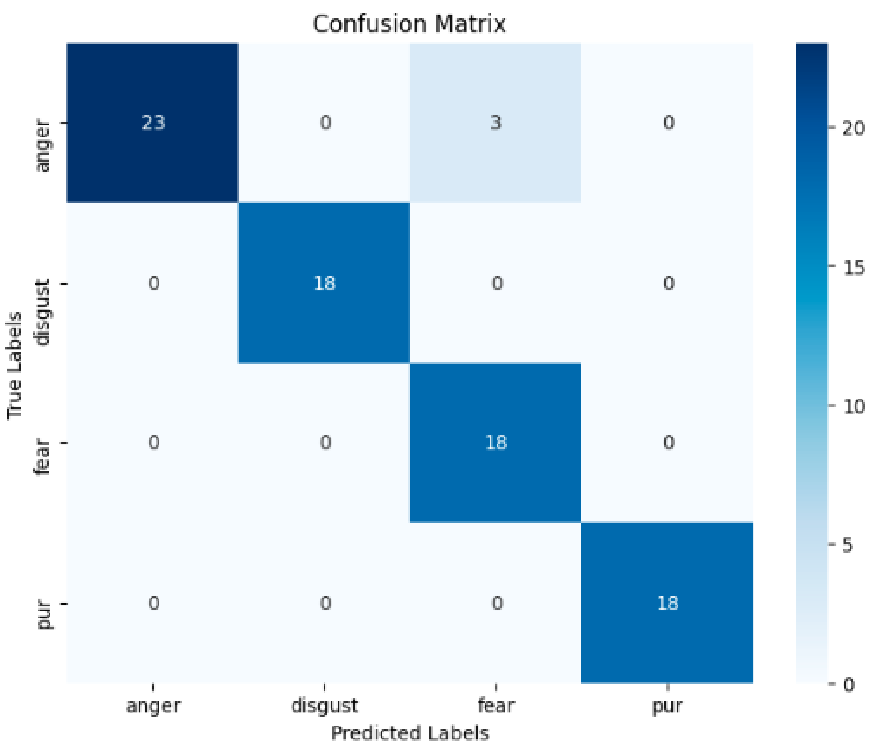

Fig. 3: Confusion matrix

Fig. 3 shows that the The confusion matrix evaluates the performance of a classification model on four vocal emotion categories: **anger, disgust, fear, and pur**. The rows represent the **true labels**, while the columns represent the **predicted labels**.

5. CONCLUSION

This study presents the effective implementation of deep learning techniques, specifically Convolutional Neural Networks (CNNs), for the classification of animal vocalizations into four distinct emotional categories: Anger, Disgust, Fear, and Purr. To address the issue of class imbalance within the dataset, comprehensive data augmentation strategies were employed, resulting in a more uniform distribution of audio samples across all emotion classes. Feature extraction was performed using Mel-Frequency Cepstral Coefficients (MFCCs), which provided a robust and consistent representation of audio signals for model training. The CNN-based model achieved a high validation accuracy of 96.25%, demonstrating strong generalization capability and the ability to differentiate nuanced variations in vocal expression. Furthermore, the model was integrated into a Flask-based web application, enabling real-time emotion prediction and enhancing its applicability in practical domains such as animal behavior analysis, welfare monitoring, and bioacoustics research. This approach underscores the potential of deep learning for advancing emotion recognition in non-human species, contributing to the broader field of affective computing and intelligent animal–machine interaction systems.

REFERENCE

[1] Wang, X.; Zhang, Y.; Yu, S.; Liu, X.; Yuan, Y.; Wang, F. E-learning recommendation framework based on deep learning. In Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 5–8 October 2017; pp. 455–460.

[2] Gligorijevic, J.; Gligorijevic, D.; Pavlovski, M.; Milkovits, E.; Glass, L.; Grier, K.; Vankireddy, P.; Obradovic, Z. Optimizing clinical trials recruitment via deep learning. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1195–1202.

[3] Davatzikos, C.; Ruparel, K.; Fan, Y.; Shen, D.; Acharyya, M.; Loughead, J.; Gur, R.; Langleben, D.D. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage* **2005**, *28*, 663–668.

[4] Justesen, N.; Bontrager, P.; Togelius, J.; Risi, S. Deep Learning for Video Game Playing. *IEEE Trans. Games* **2020**, *12*, 1–20.

[5] Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio Replay Attack Detection with Deep Learning Frameworks. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 82–86.

[6] Ajuzieogu, U. *The Role of AI in Modern Computing and Education*; Lulu Publisher: Morrisville, NC, USA, 2019; ISBN 978-0-359-72121-4.

[7] Jalal, M.A.; Loweimi, E.; Moore, R.K.; Hain, T. Learning Temporal Clusters Using Capsule Routing for Speech Emo-tion Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 1701–1705.

[8] Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* **2020**, *20*, 183.

[9] Singh, Y.B.; Goel, S. A systematic literature review of speech emotion recognition approaches. *Neurocomputing* **2022**, *492*, 245–263.

[10] Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814.

[11] Yadav, S.P.; Zaidi, S.; Mishra, A.; Yadav, V. Survey on Machine Learning in Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN). *Arch. Comput. Methods Eng.* **2022**, *29*, 1753–1770.

[12] Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A review on speech emotion recognition using deep learning and attention mechanism. *Electronics* **2021**, *10*, 1163.

[13] Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access* **2019**, *7*, 117327–117345.
[14] Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249.
[15] de Lope, J.; and Graña, M. An ongoing review of speech emotion recognition. *Neurocomputing* **2023**, *528*, 1–11.
[16] Wen, G.; Li, H.; Huang, J.; Li, D.; Xun, E. Random Deep Belief Networks for Recognizing Emotions from Speech Signals. *Comput. Intell. Neurosci.* **2017**, *2017*, 1945630.
[17] Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.
[18] Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **2019**, *47*, 312–323.
[19] Lee, C.; Song, K.Y.; Jeong, J.; Choi, W.Y. Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data. *arXiv* **2019**, arXiv:1805.06606.
[20] Tang, D.; Zeng, J.; Li, M. An End-to-End Deep Learning Framework for Speech Emotion Recognition of Atypical Individuals. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 162–166.
[21] Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition. *IEEE Signal Process. Lett.* **2018**, *25*, 1440–1444.
[22] Li, P.; Song, Y.; Mcloughlin, I.; Guo, W.; Dai, L. An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018.
[23] Jiang, W.; Wang, Z.; Jin, J.S.; Han, X.; Li, C. Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network. *Sensors* **2019**, *19*, 2730.
[24] Huang, K.; Wu, C.; Hong, Q.; Su, M.; Chen, Y. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 5866–5870.
[25] Issa, D.; Demirci, M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894.
[26] Makhmudov, F.; Kutlimuratov, A.; Akhmedov, F.; Abdallah, M.S.; Cho, Y.-I. Modeling Speech Emotion Recognition via Attention-Oriented Parallel CNN Encoders. *Electronics* **2022**, *11*, 4047.
[27] Liu, Z.-T.; Han, M.-T.; Wu, B.-H.; Rehman, A. Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning. *Appl. Acoust.* **2023**, *202*, 109178.
[28] Saumard, M. Enhancing Speech Emotions Recognition Using Multivariate Functional Data Analysis. *Big Data Cogn. Comput.* **2023**, *7*, 146.
[29] Wang, Y.; Fu, H.; Tao, H.; Yang, J.; Ge, H.; Xie, Y. Convolutional Auto-Encoder and Adversarial Domain Adaptation for Cross-Corpus Speech Emotion Recognition. *IEICE Trans. Inf. Syst.* **2022**, *105*, 1803–1806.

[30] Xie, Y.; Liang, R.; Liang, Z.; Zhao, L. Attention-Based Dense LSTM for Speech Emotion Recognition. *IEICE Trans. Inf. Syst.* **2019**, *102*, 1426–1429.