IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

VOCAL EMOTION CLASSIFIER FOR PARROTS USING DEEP LEARNING

[1] Selvakumar G, [2] Srinivasan S A, [3] Kamalesh S
[1] Assistant Professor (Sr. G.),
Department of Artificial Intelligence and Data Science
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.

Abstract: People are fascinated by parrots because of their capacity to communicate. They are well-known for their mimicry and colorful personalities. Yet, deciphering their emotions remains a fascinating, multifaceted challenge. Knowing when someone is happy, scared, stressed, or excited even if it is a pet bird, it might help us to adjust our care to suit their emotional state. Their well-being can be greatly enhanced by offering stimulating surroundings, managing stress, and engaging in suitable interaction. The proposed paper investigates the potential of deep learning models to analyze parrot vocalizations and classify their emotions as "happy" or "sad." By adopting a carefully curated and enhanced dataset, the aim is to unlock the secrets hidden within their squawks and screeches. With the annotated parrot vocalizations, a well-established pre-processing pipeline is employed. Each audio recording will be converted into a Mel-spectrogram, preserving the temporal and spectral information crucial for emotional cues. This spectrogram representation then serves as the input for our deep learning model. Comprehending their affective reactions facilitates the development of more robust and significant connections. Being sensitive to their emotions builds affection and trust, which improves our lives as well as theirs.

Keywords— Emotion Detection, Classification, Deep Learning, Convolutional Neural Networks.

I. INTRODUCTION

The use of machine learning to ensure animal welfare has significantly increased as concerns about animal welfare in the legal and social realms have grown. Animal expressions are complex tapestries made of behavioral, psychological in nature and physiological elements that serve as essential emotional gauges [1]. One of the most popular uses of machine learning in the human race is emotion detection, which is applied to software, websites, games, education, and healthcare. The aim of this work is to comprehend and identify the ways in which emotions are connected and seen in animals, with the purpose of concentrating on potential disparities in communication styles and research vantage points [2]. Beyond moral and societal issues, machine learning can yield financial gains by reliably identifying animal emotions. Early diagnosis of stress or sickness in animals can result in preventive interventions, lower mortality rates, and eventually higher production in industries like agriculture.

Emotion is a powerful sensation produced by one's situation or the environment. It is an innate or intuitive feeling that is different from logic or knowledge. Not only humans but also animals and birds have emotions [3]. While parrots are popular birds to keep as pets, little is known about the relationship between them and their owners and about the frequency of the problematic behaviors they display. Even though birds have grown in popularity as pets in recent years—they are even thought to be among the best-known animal companions in the USA and Europe. The quality of life of owners is frequently enhanced overall by their relationships with their pets [4]. The extent of connection and the strong emotional bonds that owners may have with pets may help to prevent the owner from being physically or mentally down. Pet's welfare should be taken into account seriously and steps should be taken to ensure it. Figuring out a pet parrot's emotion would help the pet owners to analyze and make the parrot comfortable and happier.

Utilizing deep learning to address a range of identification challenges, machine learning has seen a sharp increase in usage over the recent years. Deep learning techniques are being used for a variety of recognition applications, including speech, image, and music identification [5]. The primary benefit of deep learning is that it requires very little manual engineering and can take use of the current improvements in data volumes and computing capacity. It is very beneficial to employ spectral and prosodic features that are taken from unprocessed audio data before the actual recognition process. Pitch, formants, energy, linear prediction cepstrum coefficient (LPCC), perceptual linear prediction cepstrum coefficient (PLP), Mel frequency cepstrum coefficient (MFCC), Mel energy spectrum dynamic coefficient (MEDC), and so on are typically used when we utilize audio format of inputs [6]. Convolutional Neural Networks (CNN) can be a clear winner in using it for audio classification as it can efficiently classify data and learn high-level representation from the raw features.

The proposed paper is structured as follows: Section 2 provides a quick overview of the related works. Details regarding the suggested architecture for audio emotion recognition are given in Section 3. Results of an experiment to assess the suggested system's performance are shown in Section 4. Section 5 shows the future works and lastly, Section 6 presents the conclusion.

II. Literature Survey

Anchan et. al. studied emotion detection utilized MLP, SVM, CNN, and DNN with LSTM layer, trained on CREMA-D, TESS, SAVEE, and RAVDESS datasets. The study achieved high accuracy rates post processing, introduced a Real-time Speech Detection System with Al for emotion classification based on gender [7]. Kania et. al. worked with automation of rodent vocalization analysis by combining k-means & SVM and creating a hybrid model. The machine learning models in MoUSE achieve high accuracy in classifying and characterizing rodent vocalizations, providing reliable data for research studies as well as configuring MoUSE for optimal performance can be challenging, requiring some technical expertise and understanding of the underlying algorithms [8].

Nur Korkmaz et. al. developed a vanilla CNN and transfer learning with a VGG16 architecture, to automate the detection of dolphin whistles in underwater audio recordings. The VGG architecture, demonstrate superior accuracy (92.3%) compared to a baseline method (PamGuard, 66.4%), showcasing their effectiveness in detecting dolphin vocalizations. The study acknowledges the need for further research to validate these methods in diverse environments and with different species [9]. Kumar et. al. proposed an Emotion detection system employed SVM classification on MFCC and delta coefficients extracted from male speech audio files in the RAVDESS dataset. The SVM model achieved an accuracy of 85% on male audio files and 83% on combined (male and female) audio files for classifying eight emotions but the study utilized only one feature set, and future work could explore additional features and apply deep learning models for improved emotion classification [10].

Vryzas et. al. proposed Speech Emotion Recognition (SER) model utilizes CNN architecture, trained on the AESDD with data augmentation techniques. Data augmentation really helped in improving the accuracy of the model [11]. Franzoni et. al. implemented a methodology adopts a transfer learning approach, fine-tuning a pre-trained AlexNet CNN on a dataset of emotionally labeled dog images. This involves replacing the last fully-connected layer for specific dog emotion recognition and retraining only that layer, balancing performance, and computational cost. It leverages pre-trained CNNs enables efficient transfer of low-level visual features, reducing the need for extensive training on emotion recognition tasks but pre-trained CNNs, like AlexNet, may not have been trained on emotional classes relevant to animal emotions, requiring additional training for optimal performance in dog emotion recognition [12].

Gupta et. al. developed a 5-layered sequential DNN with 12 Chromagram features, 'leaky-relu' activation, Adam optimizer, and categorical crossentropy loss for emotion classification, emphasizing the integration of audio features effectively. The chromagram and MFCC features enhance a Deep Neural Network's emotion classification by providing both low-level acoustic details and semantic relationships. But the model's limited improvement in validation accuracy (35%) after 50 epochs raises concerns about overfitting [13]. Hantke et. al. employs established research about affective computing-based features (EGEMAPS, COMPARE) and Bag-of-Audio-Words (BoAW) representations, conducting comprehensive experiments with UAR and RMSE metrics, indicating a systematic approach. It addresses a research gap by applying affective computing-based features to analyze the emotional aspects of dog bark sequences, contributing to the understanding of non-human mammalian emotions [14].

Jain et. al. used a sequential model, involving a rotating process for hyperparameter configuration testing and model performance evaluation, provides a systematic approach to fine-tune the algorithm and achieve a 70% accuracy in classifying bird types based on audio data. The utilization of Mel Frequency Cepstral Coefficients (MFCCs) through Librosa enhances the model's ability to capture essential audio features, contributing to accurate bird type classification [15]. Schoneveld et. al. used the knowledge distillation, specifically self-distillation, in training the facial expression embedding network contributes to improved performance. This approach involves training a teacher model (fine-tuned FaceNet) and a student model (facial expression embedding network) to mimic the teacher's predictions. The proposed architecture effectively integrates visual and audio modalities, leveraging a deep CNN for facial expression recognition and a modified VGGish backbone for audio-based emotion recognition. This allows the model to capture and combine information from both visual and audio cues, enhancing the overall emotion recognition performance [16].

III. METHODOLOGY

3.1 Feature Extraction:

A spectrogram is a graphic representation of the frequency that comprises a signal across time. The STFT (Short Term Fourier Transform) of a windowed audio stream which produces a spectrogram [17]. To obtain the mel-spectrogram, the estimated magnitude spectrogram is transformed to the mel-scale. Mel-frequency scale, which is comparable to the perceptual capabilities of human ears, emphasizes low frequency over high frequency [18]. The following eq.1 relates the Mel scale to Hertz, where mel stands for Mels and f for Hertz:

$$mel = 2595\ log10\ (1+\ f\ 700) \qquad (1)$$

3.2 CNN:

Convolutional Neural Networks are a perfect fit for classification when using mel-spectrograms and audio data. In any case, 2-D spectrograms differ from natural images as they don't have the time and spatial information that natural images provide. CNN models were designed for natural images. But because spectrograms include a temporal component, they are sequential data [19].

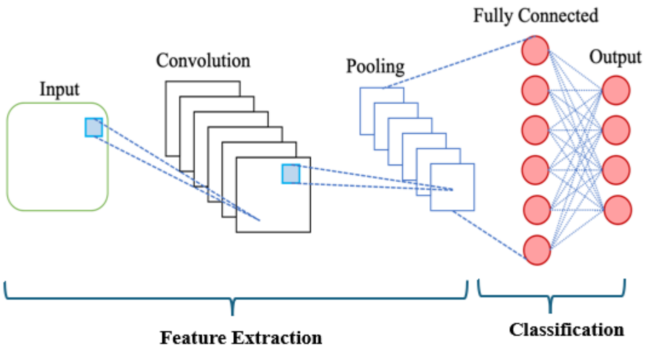
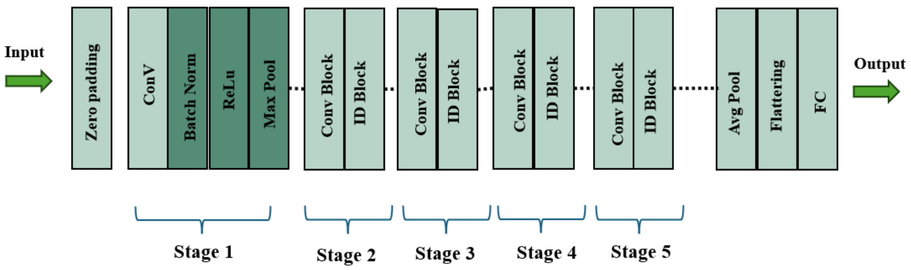


Fig 1.Architecture of CNN

3.3 ResNet:

ResNet is made up of many residual blocks layered on top of one another as it has two 3x3 layers of convolution with an equal number of output channels are present in the residual block. By employing skip connections, residual networks, or ResNets, enable the training of very deep CNNs.

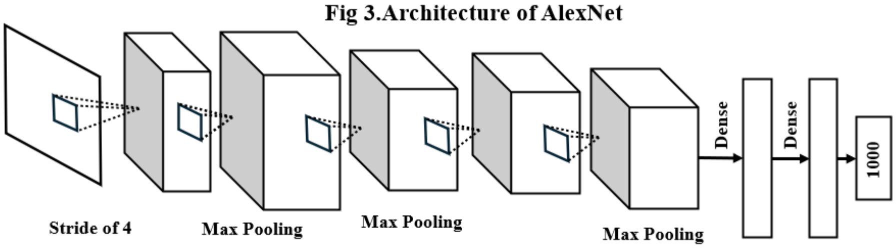
Fig 2.ResNet50 Model Architecture



3.4 AlexNet:

AlexNet is a key function of pooling layers, like max pooling, is to facilitate translation invariance, lower computational complexity, and downsample feature maps. These layers compile local data, keeping important elements and eliminating unnecessary information. Fully connected layers, sometimes referred to as dense layers, are usually used toward the end of the network architecture. By understanding intricate relationships between features and target labels, they help with classification or regression tasks by consolidating high-level features that were extracted by earlier layers. AlexNet's remarkable success in image classification tasks can be attributed to its strategic integration of dense layers and pooling, which paves the way for future deep learning architectures.

Fig 3.Architecture of AlexNet



IV. Implementation

The audio of pets especially birds is hard to find. So, web scrapping was a better solution to this problem. Only Google is more popular worldwide than YouTube.

Every month, nearly 2 billion users that are signed in to YouTube watch over a billion hours of video and produce billions of views every day [20]. In 2020, YouTube saw more than 500 hours of video uploaded every minute which is a tremendous amount of information generated. Data from YouTube and other online multimedia platforms were scrapped and converted to audio file for multiple emotions of the parrot. The data consists of two classes of emotions, happy and angry as it was accessible in abundance. Each class was segregated based on the emotions shown by the bird in the media. The lengthy audio utterances were sliced to a shorter length as it is believed that the frequency changes that capture the emotion content would remain present throughout the audio and would not be lost if the long audios were clipped. So, all the audio files are trimmed to 5 seconds in duration. By using this dataset, the proposed model was established.

The "librosa" Python module was utilised to calculate the mel-spectrogram. The fresh audio waveform is converted into a numpy array format and shown as a spectrogram, a two-dimensional visual representation.

In this research, a novel online application enhanced with a frontend module designed to efficiently fetch audio files from local storage is shown. This novel functionality is made possible by combining state-of-the-art technologies with strong programming languages. Our frontend interface, which makes use of HTML5, CSS3, and JavaScript, offers consumers a responsive and user-friendly platform for interacting with audio material. In addition, we leverage the power of contemporary web frameworks like React.js to enable dynamic user interfaces and effective data management. We guarantee smooth integration between the frontend and backend components by using rigorous coding methods and thorough attention to detail, which makes it possible for audio file retrieval operations to run smoothly. Our technological implementation demonstrates how user-centric design and software engineering concepts work together to provide a powerful solution.

V. Experimental Results

The ResNet model worked better than other models in terms of accuracy. Accuracy means how often the model provides expected result The ResNet model worked better than other models in terms of accuracy. Accuracy means how often the model provides expected results.

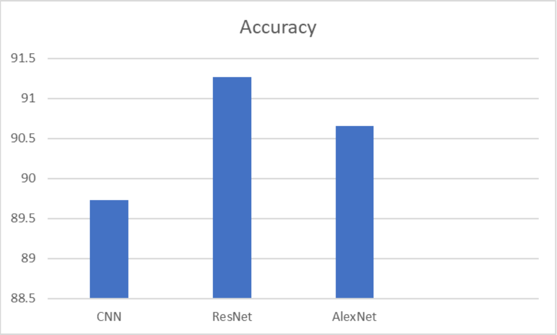


Table 1 Accuracy of the models

Model	Accuracy
CNN	89.73
ResNet	91.26
AlexNet	90.65

In terms of precision, ResNet leads the metrics with improved percentage. Precision is also one of the important evaluation metrics which plays a major role when comes to evaluating the results of a classification model.

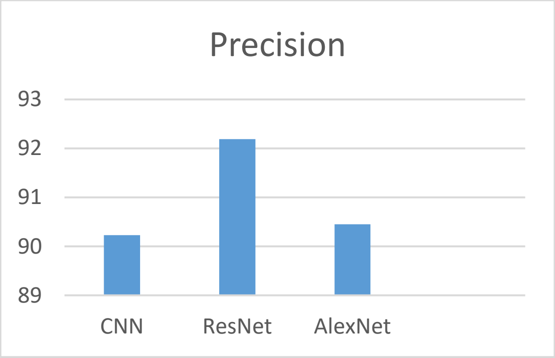


Table 2 Precision of the models

Model	Precision
CNN	90.23
ResNet	92.19
AlexNet	90.45

VI. Future Scope

While the research lays the groundwork for understanding parrot emotions, the future holds exciting possibilities for further exploration. Expanding the emotional spectrum beyond "happy" and "sad" to encompass a wider range of states like fear, anger, and contentment will provide a more nuanced understanding of their emotional landscapes. Recognizing individual differences by tailoring emotion classification to specific parrots based on their unique vocal signatures can personalize care and communication further. Developing real-time emotion detection systems can enable immediate responsiveness to their emotional state, fostering deeper connections and well-being. Insights gained from studying parrot emotions can even be applied to understand the emotional lives of other bird species, contributing to broader advancements in animal welfare and conservation efforts.

VII. Conclusion

Because their owners don't understand their emotional needs, a lot of birds wind up in shelters. By being aware of these demands, one can encourage responsible ownership and stop emotional abuse or neglect. The proposed Because their owners don't understand their emotional needs, a lot of birds wind up in shelters. By being aware of these demands, one can encourage responsible ownership and stop emotional abuse or neglect. The proposed paper explored the potential of deep learning to figure the emotions hidden within parrot vocalizations, focusing on classifying "happy" and "sad" states. By creating a dataset of annotated parrot vocalizations using web scraping, converting them into mel-spectrograms, and employing a deep learning model, the model provides an efficient classifier. The proposed research holds immense promise for improving parrot well-being. Communication is not limited to words. Understanding their mental and emotional language enables us to "speak their language" and interact with them more successfully, improving their quality of life and lowering dissatisfaction.

REFERENCES

[1] Nie, Lili, Bugao Li, Yihan Du, Fan Jiao, Xinyue Song, and Zhenyu Liu. "Deep learning strategies with CReToNeXt-YOLOv5 for advanced pig face emotion detection." *Scientific Reports* 14, no. 1 (2024): 1679.

[2] Singh, Bhupesh Kumar, Tanu Dua, Durga Prasad Sharma, and Abel Adane Changare. "Animal Emotion Detection and Application." In *Data Driven Approach Towards Disruptive Technologies: Proceedings of MIDAS 2020*, pp. 449-460. Springer Singapore, 2021.

[3] Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." *In 2018 9th International Conference on Computing, Communication and Networking Technologies* (ICCCNT), pp. 1-5. IEEE, 2018.

[4] Tygesen, Anne, and Björn Forkman. "The Parrot–Owner Relationship and Problem Behaviors in Parrots." *Anthrozoös* 36, no. 6 (2023): 985-997.

[5] Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In 2016 *Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pp. 1-4. IEEE, 2016.

[6] Harár, Pavol, Radim Burget, and Malay Kishore Dutta. "Speech emotion recognition with deep learning." In *2017 4th International conference on signal processing and integrated networks (SPIN)*, pp. 137-140. IEEE, 2017.

[7] Anchan, A., Manasa, G. R., & Pinto, J. P. (2024). Gender based Real Time Vocal Emotion Detection. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 282-289.

[8] Kania, A., Ormaniec, W., Zhylko, D., Grzanka, L., Piotrowska, D., & Siódmok, A. (2024). Joseph the MoUSE—Mouse Ultrasonic Sound Explorer. *SoftwareX*, *25*, 101606.

[9] Nur Korkmaz, B., Diamant, R., Danino, G., & Testolin, A. (2023). Automated detection of dolphin whistles with convolutional networks and transfer learning. Frontiers in Artificial Intelligence, 6, 1099022.

[10] Kumar, R., & Punhani, A. (2021). Emotion Detection from Audio Using SVM. In *Proceedings of International Conference on Big Data, Machine Learning and their Applications: ICBMA 2019* (pp. 257-265). Springer Singapore.

[11] N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas and G. Kalliris, "Continuous PAPERS Speech Emotion Recognition with Convolutional Neural Networks" J. Audio Eng. Soc., vol. 68, no. 1/2, pp. 14–24, (2020 January/February.).

[12] Franzoni, Valentina, Alfredo Milani, Giulio Biondi, and Francesco Micheli. "A preliminary work on dog emotion recognition." In *IEEE/WIC/ACM International Conference on Web Intelligence-Companion Volume*, pp. 91-96. 2019.

[13] Gupta, Saurabh, Amrapali S. Chavan, A. Deepak, Anil Kumar, Sumit Pundir, Ram Bajaj, and Anurag Shrivastava. "Speech Emotion Recognition of Animal Vocals Using Deep Learning." *International Journal of Intelligent Systems and Applications in Engineering* 12, no. 13s (2024): 129-136.

[14] Hantke, Simone, Nicholas Cummins, and Bjorn Schuller. "What is my dog trying to tell me? The automatic recognition of the context and perceived emotion of dog barks." In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5134-5138. IEEE, 2018.

[15] Jain, Niyati, Medini Kamble, Amruta Kanojiya, and Chaitanya Jage. "Implementation of Bird Species Detection Algorithm using Deep Learning." In *ITM Web of Conferences*, vol. 44, p. 03042. EDP Sciences, 2022.

[16] Schoneveld, Liam, Alice Othmani, and Hazem Abdelkawy. "Leveraging recent advances in deep learning for audio-visual emotion recognition." *Pattern Recognition Letters* 146 (2021): 1-7.

[17] Yenigalla, Promod, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding." In Interspeech, vol. 2018, pp. 3688-3692. 2018.

[18] Thornton, B. Z. J. L. S. "Audio recognition using mel spectrograms and convolution neural networks." (2019).

[19] Palanisamy, Kamalesh, Dipika Singhania, and Angela Yao. "Rethinking CNN models for audio classification." arXiv preprint arXiv:2007.11154 (2020).

[20] Buf, Diana-Maria, and Oana Ștefăniță. "Uses and gratifications of YouTube: A comparative analysis of users and content creators." *Romanian Journal of Communication and Public Relations 22*, no. 2 (2020): 75-89.