# Speech Emotion Recognition of Animal Vocals Using Deep Learning

**Aviral Pamecha**
Student of Bachelors of Technology,
SRM IST Delhi NCR Campus,Ghaziabad
E-mail: pamechaaviral@gmail.com

**Aditi Budholia**
Student of Bachelors of Technology,
SRM IST Delhi NCR Campus,Ghaziabad
E-mail: aditibudholia@gmail.com

**Saurabh Gupta**
Assistant Professor
Computer Science & Engineering Department
SRM IST Delhi NCR Campus,Ghaziabad
E-mail: saurabh256837@gmail.com

**Abstract***:* Emotion Detection is a crucial parameter in Communication. Sensing emotions correctly improves communication and helps us understand the context better. To have an effective communication between two people emotions are a must. Collaboration of sensory provocations providing information about the emotional state of others can be decoded using Emotion Detection. But this capability is not constrained only to Humans. Latest studies suggests that higher order social functions including emotions might be present in animal species also. Animal emotion detection can be very useful human-animal communication. When cats and dogs are captured in an animal shelter, they tend to show variety of emotions. This in turn can leave a long-term impact on them which can affect their emotional health. This Detecting animal emotions will help humans to detect pains in animals. So, a technology which can sense and detect animal emotions would be a boon. We aim to analyze audio data and speeches from animals. Machine Learning has advanced so much through Deep Neural Networks audios can be used to harness good amount of emotion-based information. When classifying the vocal abilities of the humans the machine learning has played an utmost important role.

*Keywords*: SER, CNN, RNN, DNN, MFCC, spectral audio features, speech emotion recognition, Deep Neural Network, feature extractions.

## 1. Introduction

Machine Learning has achieved milestones in the field of numerical and visual data. Research in audio recognition has traditionally focused on speech. The success of Machine Learning in predicting patterns in Visual Data is a great leap in the field of Artificial Intelligence. It has automated numerous sectors of the industry. Application of ML has also started on Audio data. Humans produce a lot of Audio data in form of speech, anger, love, surprise and many other emotions. This audio data can be very useful in synthesizing useful information from audio speeches. ML has played an utmost role in classifying vocal abilities of Humans. When we talk of Humans, we have enormous data. But when we talk about other species such as animals, we have very less data in context of Emotion detection. Detecting emotions in animals can be very useful in animal industry. We are thinking to apply Speech Emotion Recognition techniques to Animal Audio Speeches to decode the emotion underlying in their speeches. The advancement of ML in decoding human emotions made us go ahead with this idea.

Success of Emotion Detection in Animals through CNN (Convolutional Neural Network) is due to the CNN methods in [3]. In several studies, this method is widely used and results were commendable. The features extracted from Audio files can be represented in Images which can be fed into CNNs and classification can be done. Furthermore, there Supervised Learning methods which can also be applied to this topic like it is done in [5]. Application of Supervised Machine Learning Algorithms have shown some interesting results in [3], [5]. We can get to know the importance of Emotion Detection from [1]. There are various applications of Emotion Detection in animals. In the study of [1], we get to know that SER of animals has a great use in Education and Professional sector. Moreover, study of animal emotion can be of great use in Animal Industry. We can also get to know about importance of Emotion Detection from [1]. It is important in communication with animals. Through this we can also get to know level of pain intensity in animals. A significant study of Emotion recognition in Cats has been done in [2]. In this study, the emotion detection has been done on Cats and significant results were seen. The results from [2] showed that cats integrate acoustic and visual emotional signals of particular “hiss” and human “anger” and “happiness”. The combined study of Audio and Video for Emotion Recognition can be seen in [4]. In this research, the audio spectrograms were passed through various DNN like CNN, CNN+RNN & CNN+RNN. This research showed that CNN+RNN+3DNN performs better than other combinations of NN. Study of MPEG-7 and MFCC features is done in [6]. In this study, a comparison of 6 different methods has been done. From [4] we can get to know that animal emotions can be very useful in zoos, Animal Husbandry Industry, National Parks and wherever where communication with them is important. As humans, animals also feel pain and fear. In today’s world we sometimes are surrounded by animals but we are not sure about their emotion. SER in animals will be of great use to know the pain intensity, level of fear in animals.

A Speech Emotion Recognition system is defined as a collection of methodologies in which speech signals are processed and classified to detect the emotions. This system can be used in application areas like voice assistants or analysing a call between caller and agent. In this research paper we aim to study and apply audio analysing methods to animal speeches. A study in the field of SER in animals is done in [7]. In this research, we can see detailed information on how Deep Learning techniques can used in Speech Emotion recognition. DNN such as DBM, RNN, DBN, CNN, and AE have been the subject of much research in recent years. The large layer-wise internal architecture less efficiency for temporally varying input can be counted as limitations of these techniques. Research on Spectral Audio Features for Speech Emotion Recognition is done in [11]. From this research we can note that MFCC features has greatest accuracy in terms of classifying different classes. The results show that MFCC features can be mostly relied upon for the task of Emotion Recognition.

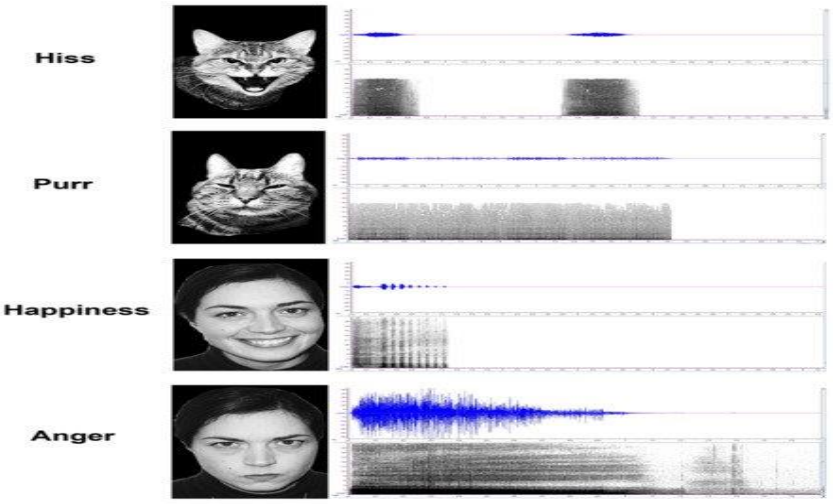


**Figure 1:** **Examples of faces and their corresponding vocalizations used in the cross-modal paradigm**

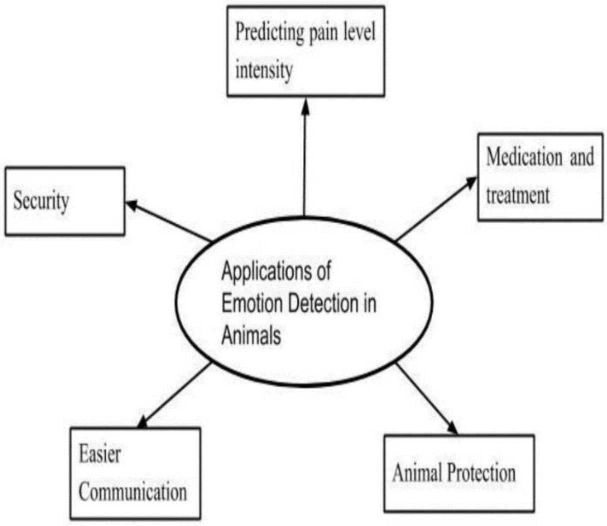

Predicting pain level intensity
Security
Medication and treatment
Applications of Emotion Detection in Animals
Easier Communication
Animal Protection

**Figure 2:** **Applications of Emotion Detection in Animals Medication and Treatment**

## 2. Related Work

Emotions play a very important role in order to identify the feelings of a living creature. It has been an easier job to predict the emotions of human beings than animals due to the availability of enormous and vast data sets. With the help of Deep Learning, it was possible to recognize the emotions through audio + video in 2 stages [4] by first creating a neural network by taking audio spectrograms as input in CNN+RNN architecture in order to identify the emotion and then in second stage recognizing the emotion in 3DCNN by using video frames.

In the same way the emotions can also be recognized in animals. Through video streaming [1] the face was detected by which different facial features were extracted like the shape of ears, eyes etc. which eventually helped to identify the emotion the animal wants to portray. In [3] for processing the visualization the Mel-frequency Cepstral coefficient (MFCC) for audio files were found which in turn was plotted using Librosa

Model to obtain the images in the same order as that of the audio to recognize the emotions. In [2] the comparison of cat's emotions was shown with respect to that of a human using both audio + video. The cat's "purr" and "hiss" were compared to that of the "happiness" and "angry" of humans and these results show that cats have some particular skill which allows cats to identify the emotion signals of humans.

The ANN Deep Learning methods were applied in [5] to predict emotions and the results obtained provided a contrast with the findings of Music Emotion Recognition (MER). The analysis of the features which were extracted from IADS set resulted in mixed set of findings. When the features were extracted from IADS dataset the results and findings were mixed. Only MFCC feature showed a major role in case of Arousal. But at the same time various spectral, MFCC and Chroma features were important to predict the valence.

In order to get a better approach towards Speech Emotion Recognition, in [12] a deep learning model was created in such a way that it has convolutional, pooling and fully connected layers which directly takes raw input data signals which are short voice recording that are further split into 20 milliseconds. The silent segments of data were deleted using Voice Activity Detection which then divided that the data into 3 segments i.e Training, Validation and Testing. It uses complex multi-layer models that represents data with multiple level of abstraction in order to achieve the desired accuracy.

[9] states that sometimes the process of speech emotion recognition becomes difficult due to the gap between human emotions and vocal characteristics as every human being has their own pitch and emotions. Therefore, for recognizing the emotion using speech the CNN algorithm/model is used. The raw input passed in the model undergoes normalization to train the CNN. As a result of this process the collection of weights is obtained as an outcome. The system with pitch and energy is selected by the dataset and with the help of the final weights obtained the emotion is recognized.

## 3. Dataset and Features

### 3.1. Dataset

Human Audio Data can be found in abundance but Animal Audio data is not in much quantity and Quality. We aim to use the Cats and Dogs dataset available on Kaggle. We also aim to create our dataset by recording audios of animals from street and zoos.

### 3.2. Features

The collected audio files will be in .wav format. The most important numerical feature which can be extracted from Audio speeches is the Mel-frequency Cepstral coefficient for the audio. We can Visualize these coefficients using Python's Library called 'Librosa'. We are extracting raw Chromagram Features of all the audios. Chromagram features are 12 different audio notes which are present in an audio. These all features are transformed into a Dataframe which now can be input to a Deep Neural Network Model.

In order to pre-process the audio files to extract numerical features from .wav audio files, we have used various python libraries. In python there is a library named as "Librosa" which is basically developed and designed for audio analysis. The major feature provided by this library is that we can create systems which can take music information.  We have also used ‘pyAudioAnalysis’ library for extracting raw numerical features from audio files. Using this library, we have extracted the chromagram features.

There are 12 different pitch classes in any audio. These pitches can also be called as Chroma features or Chromagram. They can also be pronounced as “pitch class profiles”. They can act as a powerful tool to take out voice-based information from audio speeches. We have used raw chromagram features to extract the note frequency and develop histogram. After data processing we transform the 12 frequency notes into a 12-column dataset. These 12 features columns are for every audio sample.

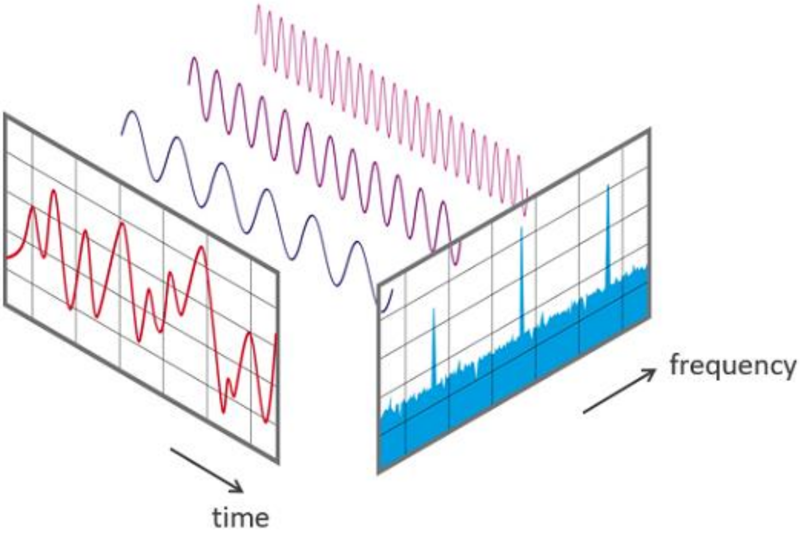


**Figure 3: Graphs representing frequency of Audio with Time**

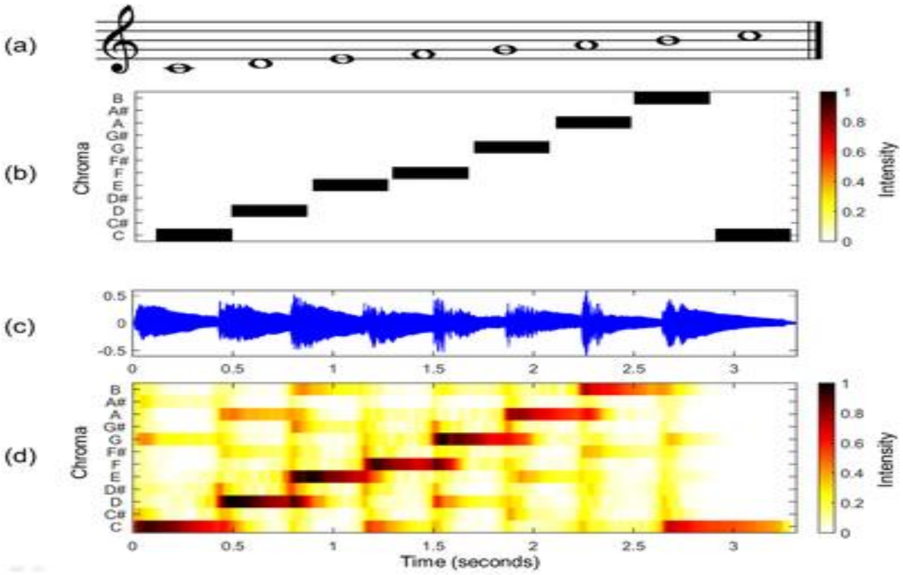


**Figure 4: All 12 different Notes of an Audio**

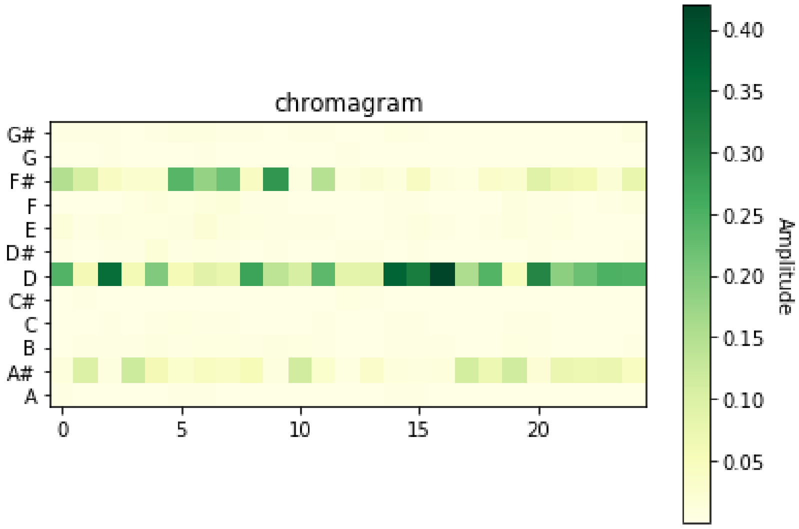


**Figure 5: A Heatmap representing all 12 notes of a single Audio**

## 4. Feature Engineering

As we are having 12 features, it would be a good idea to reduce or to increase number of features. We have reduced the number of features to 10. Number of features is something we can experiment with. Reducing number of features to very low quantity can speed up the training process but important data can be lost. We have decided to use 10 features out of 12 features. We will reduce the number of features by using Principal Component Analysis (PCA).

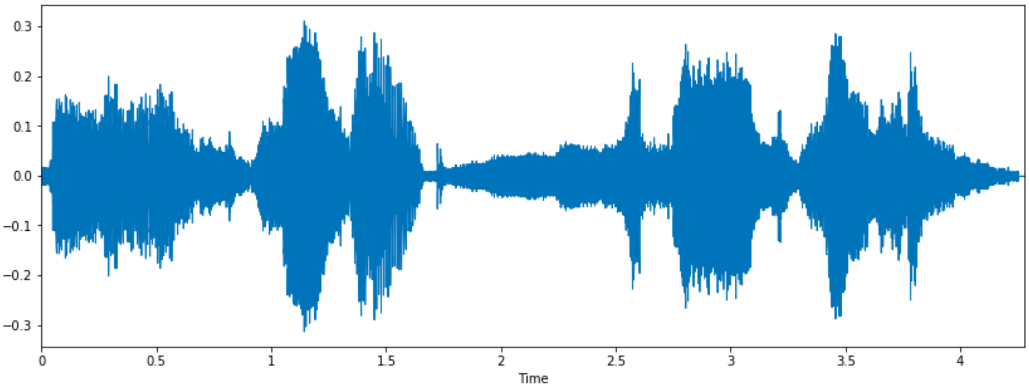


**Figure: 6 Image to wave Mapping**

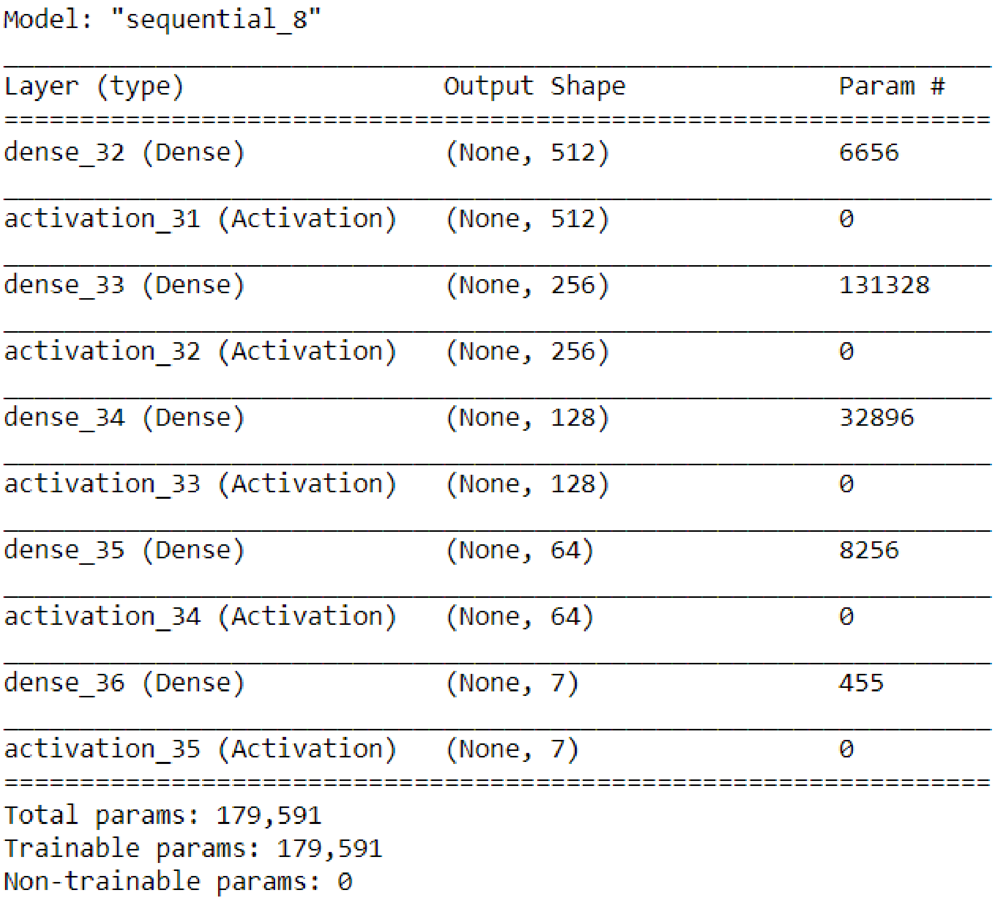
## 5. Methodology

In this paper, we consider acoustic features – Chromagram and MFCC. These two features provide a DNN (Deep Neural Network) with required and necessary low-level features. These low-level features are responsible for distinguishing of emotions. Along with these low-level features, Chromagram and MFFCC features also provide a semantic relationship so that DNN can accurately classify different emotion classes. There are 12 Chromagram features of an Audio file. These 12 features represent 12 important Audio notes of an audio. These 12 features can be incorporated into a dataframe. This dataframe can be used in a Deep Neural Network for Training purpose.

## 6. Model Building

The 12-features dataframe which is created above will be used in a Deep Neural Network. A 5 layered sequential network is built in which 3 layers are hidden. The first layer is known as the Dense layer. Seeing the complexity of the data we have kept 512 neurons in the first layer. Second layer is same layer just with a smaller number of neurons.

To tackle overfitting, we have used 'leaky_relu' as activation function with an alpha of 0.2. Third layer is also same with a smaller number of neurons. The last layer is Dense layer with 7 neurons as we have 7 different emotions. The activation function is kept as 'softmax' as this is the case of categorical classification. The model is then compiled using an Adam Optimizer. The loss in the network can be found with categorical cross entropy.

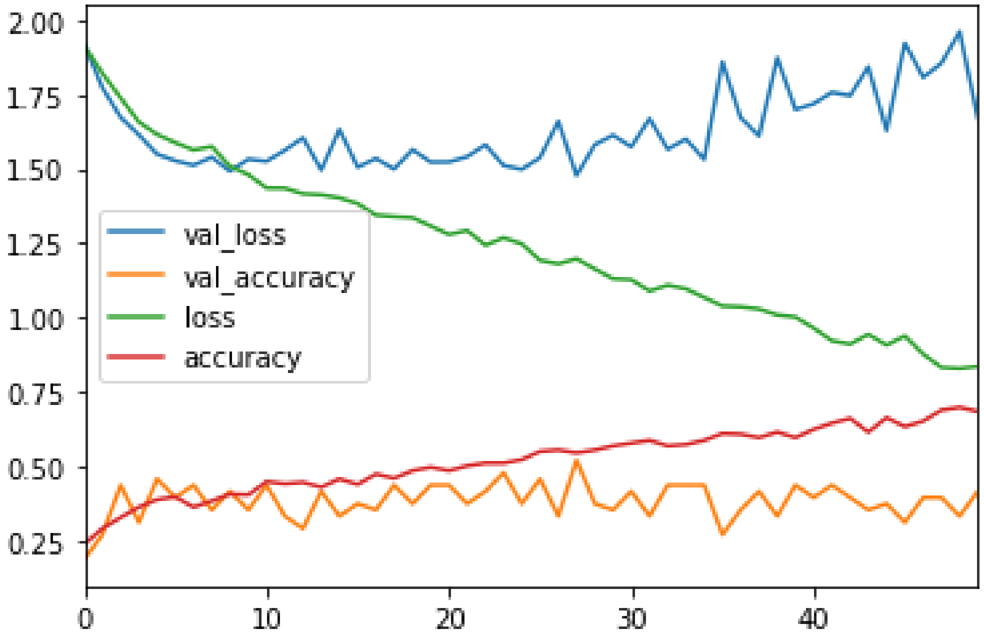


```
Model: "sequential_8"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_32 (Dense)             (None, 512)               6656
_________________________________________________________________
activation_31 (Activation)   (None, 512)               0
_________________________________________________________________
dense_33 (Dense)             (None, 256)               131328
_________________________________________________________________
activation_32 (Activation)   (None, 256)               0
_________________________________________________________________
dense_34 (Dense)             (None, 128)               32896
_________________________________________________________________
activation_33 (Activation)   (None, 128)               0
_________________________________________________________________
dense_35 (Dense)             (None, 64)                8256
_________________________________________________________________
activation_34 (Activation)   (None, 64)                0
_________________________________________________________________
dense_36 (Dense)             (None, 7)                 455
_________________________________________________________________
activation_35 (Activation)   (None, 7)                 0
=================================================================
Total params: 179,591
Trainable params: 179,591
Non-trainable params: 0
```

## 7. Results

| Epoch | val_loss | val_accuracy | loss | accuracy |
|---|---|---|---|---|
| 0 | 1.825973 | 0.333333 | 1.924034 | 0.206019 |
| 1 | 1.718995 | 0.333333 | 1.838121 | 0.261574 |
| 2 | 1.662616 | 0.395833 | 1.744466 | 0.328704 |
| 3 | 1.662637 | 0.4375 | 1.691121 | 0.349537 |
| 4 | 1.591711 | 0.458333 | 1.612831 | 0.372685 |
| 5 | 1.648664 | 0.458333 | 1.585886 | 0.37037 |
| 6 | 1.599826 | 0.479167 | 1.564551 | 0.375 |
| 7 | 1.759542 | 0.375 | 1.532655 | 0.381944 |
| 8 | 1.68576 | 0.354167 | 1.509513 | 0.418981 |
| 9 | 1.630127 | 0.4375 | 1.530943 | 0.398148 |
| 10 | 1.634473 | 0.416667 | 1.442551 | 0.409722 |
| 11 | 1.762719 | 0.416667 | 1.454837 | 0.405093 |
| 12 | 1.780549 | 0.354167 | 1.433592 | 0.407407 |
| 13 | 1.67442 | 0.416667 | 1.38033 | 0.465278 |
| 14 | 1.666486 | 0.395833 | 1.363796 | 0.456019 |
| 15 | 1.666699 | 0.4375 | 1.336924 | 0.453704 |
| 16 | 1.701909 | 0.3125 | 1.303796 | 0.502315 |
| 17 | 1.707368 | 0.416667 | 1.361756 | 0.456019 |
| 18 | 1.63315 | 0.416667 | 1.34051 | 0.497685 |
| 19 | 1.684071 | 0.4375 | 1.265692 | 0.516204 |
| 20 | 1.716381 | 0.4375 | 1.257053 | 0.511574 |
| 21 | 1.728432 | 0.458333 | 1.230584 | 0.516204 |
| 22 | 1.743204 | 0.4375 | 1.237318 | 0.541667 |
| 23 | 1.777185 | 0.395833 | 1.250397 | 0.516204 |
| 24 | 1.871462 | 0.354167 | 1.237502 | 0.525463 |
| 25 | 1.739593 | 0.4375 | 1.205062 | 0.546296 |
| 26 | 1.825222 | 0.333333 | 1.179822 | 0.555556 |
| 27 | 1.761825 | 0.416667 | 1.154968 | 0.555556 |
| 28 | 1.819383 | 0.416667 | 1.142445 | 0.550926 |
| 29 | 1.865758 | 0.395833 | 1.108519 | 0.592593 |
| 30 | 1.892333 | 0.395833 | 1.094702 | 0.587963 |
| 31 | 1.832382 | 0.375 | 1.088996 | 0.576389 |
| 32 | 1.910915 | 0.375 | 1.04037 | 0.608796 |
| 33 | 1.953571 | 0.333333 | 1.038572 | 0.618056 |
| 34 | 1.876336 | 0.416667 | 1.053931 | 0.581019 |
| 35 | 2.089959 | 0.270833 | 0.996775 | 0.618056 |
| 36 | 2.048887 | 0.395833 | 1.041399 | 0.613426 |
| 37 | 2.007513 | 0.416667 | 0.974735 | 0.622685 |
| 38 | 2.063566 | 0.354167 | 0.993674 | 0.62037 |
| 39 | 2.138384 | 0.291667 | 0.982775 | 0.641204 |
| 40 | 2.150193 | 0.395833 | 0.924058 | 0.638889 |
| 41 | 2.154113 | 0.333333 | 0.94278 | 0.638889 |
| 42 | 1.976458 | 0.395833 | 0.904761 | 0.664352 |
| 43 | 2.065085 | 0.375 | 0.898871 | 0.645833 |
| 44 | 2.093139 | 0.3125 | 0.864531 | 0.662037 |
| 45 | 2.178331 | 0.416667 | 0.845786 | 0.694444 |
| 46 | 2.211607 | 0.375 | 0.817919 | 0.673611 |

| | | | | |
|---|---|---|---|---|
| 47 | 2.242846 | 0.375 | 0.818409 | 0.6875 |
| 48 | 2.269721 | 0.395833 | 0.847504 | 0.671296 |
| 49 | 2.428202 | 0.354167 | 0.807558 | 0.668981 |



After training 50 epochs we are getting an Accuracy of 69% on Training Data and Validation Accuracy of 35%.

## 8. Conclusions

In this research paper, a Deep Learning model was created by extracting the features from the audio files of animals which were in .wav format. From this the chromogram features were extracted which were successful in determining various emotion which an animal is going though. For future scope, along with audio the video files can also be included which in turn will increase the accuracy to determine the emotion because of various facial expressions.

## 9. Acknowledgment

We would like to express our appreciation to Mr. Suarabh Gupta who have supported us during our research and study in SRM Institute of Science and Technology.

## 10.References

[1] B. K. Singh, D.P. Verma, A. Adane (2020): Animal Emotion Detection and Application. https://www.researchgate.net/publication/346733135

[2] A. Quaranta, S. d'Ingeo, R. Amoruso, M. Siniscalchi (2020): Emotion Recognition in Cats. https://doi.org/10.3390/ani10071107

[3] V. Totakura, M.I. Thariq Hussan (2020): Prediction of Animal Vocal Emotions using Convolutional Neural Network. https://www.researchgate.net/publication/339687484

[4] M. Singh, Y. Fang (2020): Emotion Recognition in Audio and Video Using Deep Neural Networks. https://arxiv.org/pdf/2006.08129.pdf

[5] S. Cunningham, H. Ridley, J. Weinel & R. Picking (2020): Supervised machine learning for audio emotion recognition. https://doi.org/10.1007/s00779-020-01389-0

[6] R. Radhakrishnan (2003): Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1221332&isnumber=27434

[7] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar & T. Alhussain (2019): Speech Emotion Recognition Using Deep Learning Techniques: A Review. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8805181&isnumber=8600701

[8] P. Shi (2018): Speech emotion recognition based on deep belief network. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8361376&isnumber=8361215

[9] H. Murugan (2020): Speech Emotion Recognition Using CNN. International Journal of Psychosocial Rehabilitation. https://www.researchgate.net/publication/342231090_Speech_Emotion_Recognition_Using_CNN

[10] P. Raguraman, R. Mohan & M. Vijayan (2019): LibROSA Based Assessment Tool for Music Information Retrieval Systems. https://www.researchgate.net/publication/332678924_LibROSA_Based_Assessment_Tool_for_Music_Information_Retrieval_Systems

[11]A. Shoiynbek, K. Kozhakhmet, N. Sultanova & R. Zhumaliyeva (2019): The Robust Spectral Audio Features for Speech Emotion Recognition http://www.naturalspublishing.com/files/published/9t24q1809wgyib.pdf

[12]P. Harár, R. Burget & M. K. Dutta (2017): Speech emotion recognition with deep learning. https://www.researchgate.net/publication/320089581_Speech_emotion_recognition_with_deep_learning

[13]H. Purwins, B. Li, T. Virtanen, J. Schlüter, S. Chang & T. Sainath (2019): Deep Learning for Audio Signal Processing. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8678825&isnumber=8717740

[14]P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar & J. Vepa (2018): Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1811.pdf

[15]S. Lalitha, A. Madhavan, B. Bhushan & S. Saketh (2014): Speech emotion recognition. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7002390&isnumber=7002373