Crowd Emotion Analysis with Deep learning using Convolutional Neural Networks

Nandini khandelwal ,
Shourya Sharma
Guided by Prof. Sameer Sayyad

Robotics and Automation Department
Symbiosis Institute of Technology, Pune

Abstract

With the use of their words, a group of people can convey their feelings or a collective sense of emotion. Humans communicate with one another in order to share their emotions. These emotions may include grief, rage, stress, and other emotions. Whistles, shouts, claps, and other noises made by a crowd are also included. Crowd noises convey more than just a few words; they also reveal people's mental and emotional states. The term "crowd roar" wonderfully captures the idea of a group's collective passion being expressed through sound by the group as a whole.Crowd Emotion Recognition (CER) is a notion that identifies an individual's or a group's emotion based on their vocalizations. These sounds, which are frequently created by the crowd, can include cheering, booing, and screaming. Labels are frequently employed in tasks that deal with identifying the emotions in sounds originating from a group of people. These designations serve as a signal of the crowd's acceptance, neutrality, or rejection of the particular activity at hand. Our project's primary goal is to determine the emotions of a crowd. The information that is now accessible makes it quite obvious that employing artificial intelligence to recognise crowd emotions is not an easy undertaking.Basic machine learning algorithms may complete the task, but their accuracy levels fall short of what would be considered optimal performance. The CNN model is useful in this situation

Keywords : Speech Recognition ,Convolutional Neural Networks, Mel-spectrograms , Deep Learning ,Machine Learning ,

1.Introduction

Over the years, numerous studies have been done on crowd emotion recognition. It could be feasible to identify particular emotional states or forecast crowd behavior by looking at the acoustic characteristics of crowd emotional sound, which can be helpful for crowd safety and security. It could be feasible to identify particular emotional states or forecast crowd behavior by looking at the acoustic characteristics of crowd emotional

sound, which can be helpful for crowd safety and security. Overall, studying crowd emotional sound may provide important information about a group's overall emotional state and be helpful for a variety of applications in disciplines including psychology, sociology, and event planning.
Many individuals have a very vested interest in this field. A major question which comes to our mind while studying at these researches is what is the problem statement which is related to this field?How can we use crowd emotional sounds to accomplish emotion recognition using audio and how can we use CNN to turn crowd voices into comprehensible and interpretable visuals? This is the issue statement for this study.Another question which comes to mind is whether the emotional content of crowd sounds can be recognised using frequency-amplitude features and analytic methods which are analogous to those used on individual voices. While using the traditional methods for crowd emotion recognition, Low precision performance, Expensive computing and lengthy identification of a variety of emotive scenarios are seen as the results. These questions have raised an interesting curiosity among researchers which has resulted in many bold and informative researches and development in this field. Crowd emotion recognition, even with so much development at this very moment, has got many other problems to be solved and advancements to be unfolded.

1.1. Literature review

This topic of crowd emotion recognition has been a major subject for research and studies. Over the years many researches have been conducted. Some of the prominent studies have been described in the following table. These researches have involved usage of algorithms such as SVM, KNN, Decision Tree and many more. These approaches have resulted in varying accuracies with the majority of them being between 80-90 percent. These accuracy scores although being very impressive, Our CNN model has been able to achieve a better accuracy percentage for the dataset.

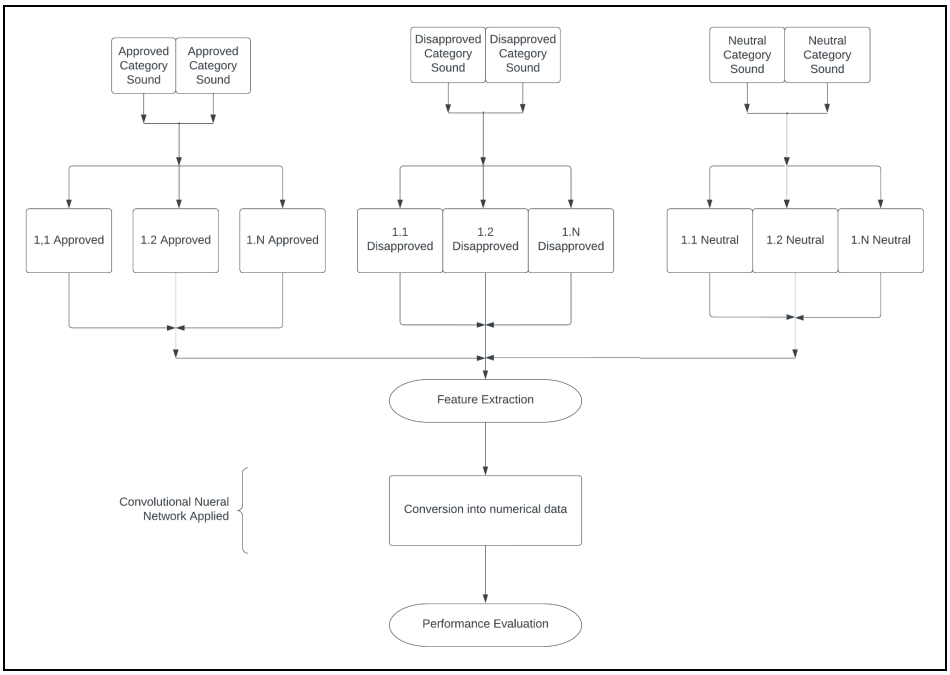
Reference Number	Author Name	Description	Accuracy
[1]	Valentina Franzoni, Guilio Biond, Alfredo Milani	A research was conducted on the final trained convolutional neural network to test its ability to characterize the emotion of a crowd.	80-90 %

[2]	Chung-Hsien Wu and Wei-Bin Liang	AP & SL technique is used to identify emotional states using Gaussian mixture, Support vector machine, multilayer perceptron and decision trees.	85.79%, 80.92% and 83.55% respectively.
[3]	T.Mary and T.Maya	An experiment was conducted to produce effective and trustworthy results when classifying emotions from voice cues using multiclass SVM, KNN & Discriminant analysis.	83.08% for Berlin, 82.67% for eNTERFACE, 81.79% for RAVDESS, 82.99% for EMOVO, 84% for Urdu and 83.75% for SAVEE

2. Research Gap

Multiple researches and experiments have been conducted on emotion recognition based on the sound of a crowd or an individual. There are different approaches which are being used in these researches and a dissimilarity has been seen in the accuracy results of these approaches. With this trend of different accuracy scores between these experiments, it is necessary to create a model which tackles all the possible inconsistencies and problems which can possibly occur while conducting an experiment in this field. We did not use any other model because they did not achieve the accuracy scores which our model did by the end of our experiment. And in comparison to other models, ours takes less time and is also more accurate.
Our model is more time efficient in comparison with other models because we used feature based spectrograms and converted the .wav file into numerical data instead of using frequency based spectrograms.

3. Methodology:



(Fig : System Flowchart)

Our system's overall flowchart is shown above. The approach taken for this dataset includes the following steps :

- Data Acquisition
- Data Preprocessing
- Scaling and Filtering
- Feature Extraction
- Model Training
- Result analysis and Visualization

3.1.Data Acquisition:

the collection of information derived from authentic audio snippets captured during large-scale events when the crowd behaves as a single entity collectively , the given audio files were normalized priorly . For accurate spectrogram production as in audio signal processing and analysis,Mel spectrogram is frequently employed. It is a representation of the frequency content of an audio stream over time using a Mel scale as opposed to a linear scale for the frequency axis.

A perceptual measure of pitch, the Mel scale is based on the characteristics of human hearing. Given the sensitivity of human hearing, it is non-linear and has a better

resolution at lower frequencies.and to fine-tune the domain-specific characteristics, the original sound samples are filtered and normalized in amplitude. The normalized audio files were categorized into 3 types. The types of these audio files are as follows:

1. Approved- 39 .wav files
2. Disapproved- 14 .wav files
3. Neutral- 15 .wav files

The normalization of the dataset is set to −23 Loudness Units, following the EBU R128 standard. We filtered the sound in the 20–20,000 Hz range[1]; Because it encompasses the whole spectrum of human hearing, we filter sound in the 20–20,000 Hz range. The human ear is sensitive to noises in this frequency range, with the band between 2,000 and 5,000 Hz being the most sensitive. We can identify and examine the key frequencies for human perception and hearing by filtering the sound in this range. Overall, filtering sound between this range is a feasible and efficient technique to record and examine the key frequencies for human perception and hearing.

3.1.1.Data Preprocessing

The first step in our project is importing all of the libraries that will be used to carry out the various operations. These are the libraries that were imported:

1. Pydub Library - Pydub is a python library which is used to process audio files(.WAV or MP3 files). In projects like audio processing and analysis, speech recognition, and music creation, the Pydub library can be extremely useful.It offers a user-friendly interface for interacting with several audio formats, like WAV and MP3. With Pydub, you may quickly edit audio files by splitting, concatenating, and exporting them in various formats.
2. AudioSegment - We import AudioSegment from the Pydub library, which is used to extract the audio files' attributes (channels, sample width, frame rate, frame count, and length in milliseconds).
3. Glob Module - Glob module is a robust Python utility for pattern-based file searching and retrieval.
4. Librosa Library - for feature extraction of .WAV files. It offers a variety of tools for working with audio data, such as ways to extract features from audio signals, and ways to visualize audio data.
5. Pandas Library - for creating the data frame.
6. Numpy Library - to create an array for the cnn module and the “np.mean” function is used for taking out the mean of the array of the extracted numeric value from the .wav file

3.1.2. Data Filtering & Data Scaling

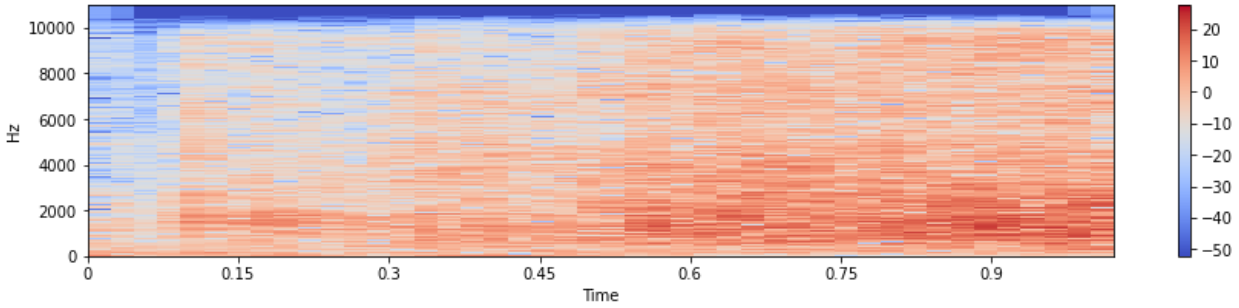
We obtained data like frame count, intensity, and duration of audio files in order to examine the content of the file we are dealing with. We noticed discrepancies in the

duration of the audio files as a result. We decided to divide these files into units with a window size of 1 sec each in order to address this problem. These divided files are kept in a directory. From the original audio recordings, we currently have 1778 approval chunks, 388 disapproval chunks, and 7340 neutral chunks [1]. Using the shifting window method, we separated the audio recordings into segments. As part of the shifting window technique, the audio stream is split into short, overlapping chunks or frames with specified durations. The size of the window or the duration of each segment is frequently determined by the frequency content of the signal and the specific analysis being performed. with 0.25 seconds of moving window and 0.75 seconds of overlap for each block of one second.

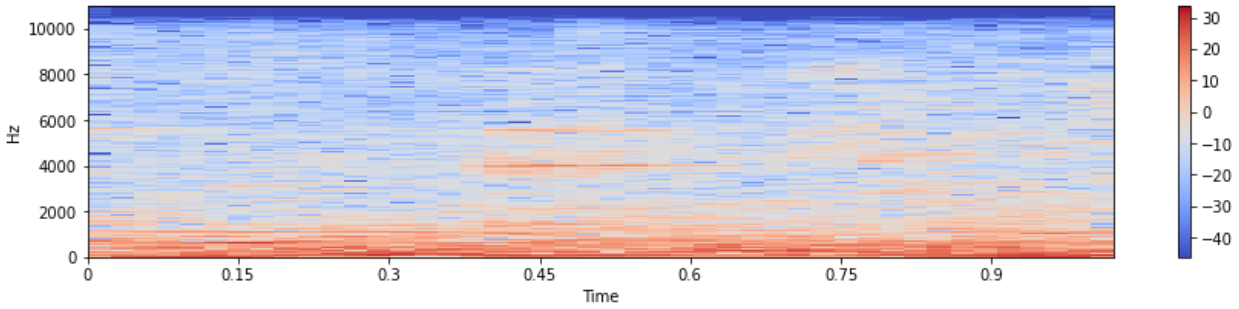
In order to transform the noises of the crowd into accessible and comprehensible images, we employed a technique for creating a spectrogram.

The frequency content of an audio stream is shown visually in a spectrogram over time. The time-varying acoustic characteristics of the crowd voices may be transformed into a 2D picture that can be studied and understood using a spectrogram , after converting them into 1s blocks.

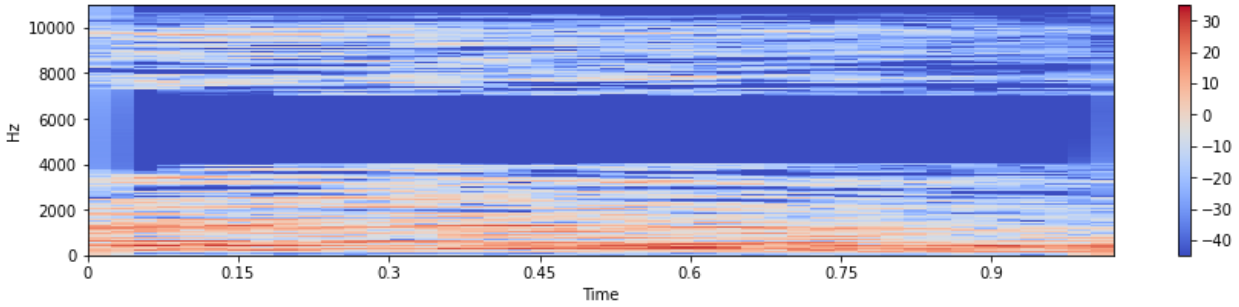
The frequency time spectrograms for each category we have obtained are as follows:



(Fig 1 : Approval Spectrogram)



(Fig 2 : Neutral Spectrogram)



(Fig 3: Disapproval Spectrogram)

3.2. Feature Extraction

In many audio processing applications, including emotion recognition, music analysis, feature extraction is a crucial step. The methods we used for extracting features from an audio file were

Mel-Frequency Cepstral Coefficients (MFCC) is a well-liked method for extracting features from audio signals. Spectral Features - The short-term Fourier Transform (STFT) of the audio stream yields spectral characteristics. The power spectrum, spectral centroid, spectral flatness, and spectral roll-off are examples of spectral characteristics. These characteristics can be used to describe an audio signal's frequency content.Chroma features are obtained from an audio signal's pitch class profile. They depict how different pitch classes are distributed in a musical composition.

After the audio files split, we retrieved characteristics such as:

3.2.1.Mel-frequency

A common feature extraction method used in audio signal processing and speech recognition . It entails examining the spectrum of a sound wave and organizing the frequency bands into mel frequency bins, which are spaced in accordance with how pitch is perceived by the human auditory system.

The logarithm of the energy contained in each mel frequency bin is then used to create MEL coefficients, which are then applied to the resultant mel spectrum using a Discrete Cosine Transform (DCT). For tasks like voice recognition, speaker identification, and music genre classification, machine learning algorithms can utilize these coefficients to describe the spectral envelope of a sound source.

3.2.2.Mel Frequency Cepstral Coefficients (MFCC)

The MFCC method produces a set of coefficients that succinctly and accurately depict the spectral envelope of the sound stream. These coefficients may be applied to a number of tasks, including speaker identification, emotion recognition, and speech recognition.The steps of the MFCC algorithm are as follows:

- Pre-emphasis: At this stage, a high-pass filter is used to enhance the high-frequency components of the audio stream.

- Frame blocking: The audio stream is broken up into brief frames, usually lasting 20 to 30 milliseconds.
- Windowing: To lessen spectral leakage and increase the precision of the spectral analysis, a window function is applied to each frame.
- Each frame is subjected to the Fourier Transform to produce the signal's frequency domain representation.
- Mel the filterbank. The filterbank's focus is on the frequency elements that are most crucial for human perception.
- Logarithm: The filterbank's output logarithm is used to reduce the signal's dynamic range.

The MFCC method produces a set of coefficients that succinctly and accurately depict the spectral envelope of the sound stream.

3.2.3 Complex Short-Time Fourier Transform(C STFT)

It is an adaptation of the STFT(Short-Time Fourier Transforms) algorithm, which is used to examine audio signals. The primary distinction between STFT and C STFT is that the latter preserves the fact that the Fourier Transform coefficients have complex values.
In order to determine the frequency content of the audio signal, the STFT algorithm divides the audio signal into tiny frames and uses the Fourier Transform on each frame. The Fourier Transform's output, on the other hand, is a collection of complex-valued coefficients that are both magnitude and phase-informed.
For spectrum analysis, the magnitude of the coefficients is frequently employed, whilst the phase information is frequently ignored.
The Fourier Transform coefficients' complex-valued character is preserved by the C STFT method, in contrast. Although phase information may also be applied to many applications, such as audio signal processing and synthesis, it enables a more thorough description of the signal.
The time-frequency representation of the audio stream containing magnitude and phase information is the product of the C STFT algorithm. Applications for this format include audio signal processing, synthesis, and analysis.
Along with this function, we compute a chromagram from a waveform or power spectrogram[2].

3.2.4.Complex Constant-Q Transform.(C CQT)

It is an adaptation of the CQT algorithm, which is used to analyze audio signals. The key distinction between CQT and C CQT is that the latter preserves the fact that the Fourier Transform coefficients have complex values.
The CQT method divides the audio stream into a series of narrow frequency bands, each band's bandwidth growing exponentially with frequency. The Fourier Transform

coefficients' complex-valued character is preserved by the C CQT method, in contrast. Although phase information may also be applied to many applications, such as audio signal processing and synthesis, it enables a more thorough description of the signal. The C CQT technique produces a time-frequency representation of the audio signal that includes information on the amplitude and phase.

3.2.5 Complex Cepstral Envelope Normalization Spectrum (C CENS)

It is referred to as C CENS. It is a feature extraction approach used in audio signal processing for a variety of applications, including speech recognition, emotion identification, and music information retrieval.

The C CENS method produces a set of coefficients that concisely and accurately depict the spectral envelope of the audio stream. The following actions are part of the C CENS algorithm:

- Short-Time Fourier Transform (STFT): This technique divides the audio stream into brief frames, usually lasting 20–30 milliseconds. Each frame is given a window function, and the Fourier Transform is calculated to determine the signal's frequency content.
- Mel Filterbank: To produce a collection of Mel frequency bands, the Mel filterbank is used.
- Logarithm: The logarithm used to condense the signal's dynamic range.
- Inverse Fourier Transform: To extract the complex coefficients, the Inverse Fourier Transform is done to the log-Mel spectrum.
- Then, the envelope is normalized to eliminate the impact of fluctuating signal amplitudes and noise.
- To produce the C CENS coefficients, the normalized envelope is smoothed using a Gaussian filter.

The characteristics of the retrieved audio files (.WAV files) were converted into numerical data. The numerical data shows that some of the retrieved characteristics were in array form and others were not. We see that some features are in array form and some are not, therefore we decide to calculate the mean of these characteristics and put them into a data frame. All three categories were combined into a single, sizable data frame, which was then saved as a.CSV file.

3.3 Model Training

From here we start the implementation of CNN (Convolutional Neural Network). A very powerful algorithm for deep learning is CNN (Convolutional Neural Network).The capacity of CNNs to recognise spatial and temporal patterns in data allows us to use

them to analyze numerical data, which improves accuracy and performance in comparison to conventional machine learning models. Because of its propensity to recognise patterns and apply previously acquired knowledge, it is a very popular neural network that is typically used for audio analysis. The implementation of CNN is divided into multiple steps which are as follows:

1. Preprocessing involves importing the data, dividing it into training, validation, and testing sets, and normalizing the data as the initial step.
2. Determining CNN's architecture is the next stage. The common input form for numerical data is a 1D array with the same number of features (or input channels) as variables in the dataset. Similar to an image-based CNN, the design can have convolutional layers, pooling layers, and fully linked layers.
3. Then, in order to prepare the data for an algorithm and obtain a better forecast, we employ the One Hot Encoder. With one-hot, we create a new category column for each categorical value and give it a binary value of 1 or 0.
4. Model compilation involves us defining the optimizer which was Adam, the loss function as categorical crossentropy , and the evaluation metric as accuracy for the model once the architecture has been established.
5. The total parameters and the trainable parameters are found out to be 4611.
6. The next step is testing these parameters with different epoch values. 50 is the best epoch value which is found for maintaining accuracy .

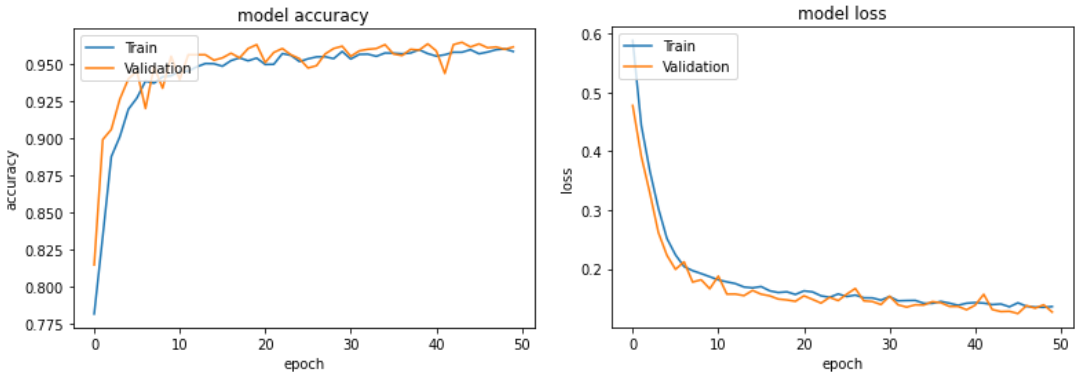
We also tried to implement the LSTM model on our current data.An LSTM model is a sort of recurrent neural network (RNN) that is often used for sequential data analysis, such as natural language processing and speech recognition. The primary benefit of LSTM models over typical RNNs is their ability to capture long-term relationships in input data. Language modeling, machine translation, speech recognition, and other applications have proven that LSTM models are quite successful.The LSTM model was able to perform its operations very efficiently and it achieved an accuracy percentage of 94% , it was lesser than the Convolutional Neural Network predicted accuracy which was implemented on this model.Being the model with the highest accuracy percentage value, Convolutional Neural Network was finally used for the research.

4.Result and Discussion

4.1.Result analysis:

The Convolutional Neural Network, when implemented on the current data, achieved an accuracy of 96.53%, it took us 1 minutes 3 seconds to run each , with the epochs of 50.

4.2.Visualization



(Fig 4: training & loss graph)

```
60/60 [==============================] - 0s 3ms/step - loss: 0.1208 - accuracy: 0.9653
Accuracy: 96.53%
```

(Fig 5: accuracy)

5.Conclusion:

Crowd Emotion Recognition (CER) is a notion that identifies an individual's or a group's emotion based on their vocalizations. This study aims to use artificial intelligence to determine the emotions of a crowd, using a CNN model to convert MP3 files into WAV files. The Pydub library is used to process audio files, and AudioSegment is used to extract the audio files' attributes in milliseconds.

Mel spectrograms are used to represent the frequency content of an audio stream over time and filter sound in the 20-20,000 Hz range, which is sensitive to noises in this frequency range, making them a feasible and efficient technique for recording and examining key frequencies for human perception and hearing. We separated audio samples into units with 1 sec window sizes and used a shifting window technique to build a spectrogram to convert crowd noises into accessible and comprehensible pictures. Mel-Frequency Cepstral Coefficients (MFCC) is a popular approach for extracting spectral characteristics, chroma details, and frequency domain representation from audio sources. The Complex Short-Time Fourier Transform (C STFT) and Complex Constant-Q Transform (C CQT) are two methods used to analyze audio signals while keeping the complex-valued nature of the Fourier Transform coefficients. C CENS is a feature extraction method used in audio signal processing to recognise speech, identify emotions, and retrieve music information.

CNN is a powerful algorithm for deep learning that can be used to analyze image and numerical data, improving accuracy and performance. LSTM models are successful for sequential data analysis due to their ability to capture long-term relationships.

This project uses artificial intelligence to determine the emotions of a crowd using a CNN model, Pydub, AudioSegment, Glob, and librosa libraries. The pandas library and numpy library were used to create the data frame and array for the cnn module and the "np.mean" function. The audio files extracted were transformed into numerical data and stored as a CSV file. The implementation of CNN (Convolutional Neural Network) was able to perform its operations efficiently and achieved an accuracy percentage of 96.5%, while the LSTM model was able to perform its operations efficiently and achieved an accuracy of 94%.

References:

1. Franzoni, V., Biondi, G. & Milani, A. Emotional sounds of crowds: spectrogram-based analysis using deep learning. Multimed Tools Appl 79, 36063–36075 (2020). https://doi.org/10.1007/s11042-020-09428-x[1]

2. Y. Zhou, X. Liang, Y. Gu, Y. Yin and L. Yao, "Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 695-705, 2022, doi: 10.1109/TASLP.2022.3145287. [2]

3. C. -H. Wu and W. -B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10-21, Jan.-June 2011, doi: 10.1109/T-AFFC.2010.1 [3]
4. Andry Chowanda, Irene Anindaputri Iswanto Esther Widhi Andangsari (2022) ; “Exploring deep learning algorithm to model emotions recognition from speech”,doi:10.1016/j.procs.2022.12.187[4]

5. Valentina Franzoni, Giulio Biondi, Alfredo Milani, February 25, 2021, "Emotional Crowd Sound", IEEE Dataport, doi: https://dx.doi.org/10.21227/xxwy-5869.