World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

A Multimodal Deep Learning Model for Emotion Recognition from Video, Audio and Text Data

Ogochukwu Patience Okechukwu [1,*] and Godson Nnaeto Okechukwu [2]

[1] *Department of Computer Science, Faculty of Physical Science, Nnamdi Azikiwe University, Awka, Nigeria.*
[2] *Department of Electronic/Computer Engineering, Faculty of Engineering, Nnamdi Azikiwe University, Awka, Nigeria.*

World Journal of Advanced Engineering Technology and Sciences, 2026, 19(02), 121-135

Publication history: Received on 02 April 2026; revised on 11 May 2026; accepted on 13 May 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.19.2.0257

Abstract

This study presents the development of a multimodal deep learning model for emotion recognition using video, audio, and text data extracted from a video data. The proposed system employs a structured pipeline that begins with video input, from which audio is extracted and transcribed into text using the Whisper model. Emotion recognition is performed across three modalities: facial expressions analysed using DeepFace, vocal features classified using a Long Short-Term Memory (LSTM) network, and textual features processed using a Feedforward Neural Network (FNN). Results are visualized in bar chart plotted using Seaborn and integrated into an interactive user interface built with Flet. This multimodal approach enhances accuracy and robustness over unimodal systems by leveraging complementary information from each data stream. The model has broad applications in affective computing, human-computer interaction, surveillance, and sentiment-aware technologies.

Keywords: Multimodal; DeepFace; LSTM; FNN; Deep Learning; Emotion

1. Introduction

(Bejjagam & Chakradhara, 2022) observed that for every behavioural and physiological changes that occurred in the human body, emotions are generated showing the body's mental reactions to that change. Emotion recognition enables the identification of basic human emotions such as anger, fear, disgust, sadness, happiness, and surprise based on some input data. (Nouman, Iqbal, & Saleem, 2023) recognized that human emotion recognition is a rapidly growing field that has gained significant attention in recent years due to its potential applications in various areas, including psychology, healthcare, education, crime investigation, and entertainment. Furthermore, that emotions are complex and subjective experiences crucial to communication, decision-making, and well-being. Essentially, for effective human-robot interaction, personalized mental health interventions, and many other applications, there is need to understand emotions.

Accurate emotion recognition is critical for developing “emotionally intelligent” systems that can interact naturally with humans (Filali, et al., 2025). Recognizing human affect has become increasingly important in many domains, from human–computer interaction and social media analysis to healthcare and marketing (Filali, et al., 2025). Emotions are expressed through a rich combination of cues including facial expressions, vocal intonations, and linguistic content. Interpreting them reliably can greatly enhance user experience and decision-making (Filali, et al., 2025). For example, emotion recognition can inform adaptive tutoring systems, mental health monitoring tools, and responsive robotics. In large-scale applications with “big data” such as social media streams, integrating affective analysis helps to uncover population-level trends and contextual sentiments that single-channel systems miss.

* Corresponding author: Ogochukwu Patience Okechukwu.

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0.

Traditional emotion classification systems have typically been *unimodal*, processing only one signal type (such as speech, text, or facial video) in isolation. However, these unimodal approaches often prove brittle. As (Filali, et al., 2025) noted, focusing on a single modality "often fails to capture the full spectrum of human emotions, leading to information loss". For instance, speech emotion recognition can be undermined by background noise or ambiguous intonation, while text alone may miss nonverbal cues. Similarly, visual-only systems can be sensitive to lighting or occlusion. In practice, a person's emotional state is conveyed through multiple channels simultaneously, so relying on one cue can result in degraded performance and missed subtleties. Large reviews of affective computing emphasize that unimodal systems are susceptible to noise, bias, and lack contextual robustness (Lian, et al., 2023); (Filali, et al., 2025). These limitations motivate a shift toward multimodal recognition, which combines diverse data sources to capture complementary information and improve accuracy.

The research objective of this paper is to design and validate a unified deep-learning framework that inputs raw video and outputs emotion labels by jointly leveraging its audio, text content as well as the visual information.

By fusing predictions from text, audio, and visual channels, this multimodal system aims to produce a more reliable emotion estimate than any unimodal method alone. The model was trained and evaluated on publicly available datasets sourced from Kaggle, demonstrating scalability to large, real-world data.

2. Literature Review

(Chutia & Baruah, 2024) recognized that emotion recognition has become a crucial area of study in affective computing, particularly with the advancement of deep learning and the increasing availability of multimodal data. Traditional emotion recognition systems were limited to unimodal analysis, such as text, audio, or facial expressions, often resulting in incomplete understanding due to the loss of contextual cues. In recent years, researchers have explored the fusion of multiple methods to improve emotion recognition accuracy and robustness.

2.1. Text-based Emotion Recognition

Text-based emotion recognition is one of the most extensively researched modalities due to the abundance of textual data available from social media, chatbots, and transcripts. Feedforward Neural Networks (FNNs) and transformer-based architectures have shown significant promise in this domain. For instance, (Machová, Szabóova, Paralič, & Mičko, 2023) used multimodal datasets in their model and clearly showed that simple neural networks could still achieve competitive performance when the training data was carefully pre-processed and annotated. Incorporating semantic context and word embeddings has also enhanced model performance, especially when dealing with emotion-laden texts.

2.2. Audio-based Emotion Recognition

Audio emotion recognition focuses on extracting emotional states from voice signals by analysing prosodic features such as pitch, tone, and rhythm. Long Short-Term Memory (LSTM) networks have emerged as the dominant approach due to their ability to preserve temporal dependencies across audio frames. (Alam, Nigar, Erler, & Banerjee, 2023) observed that in recognizing complex emotions such as frustration and sarcasm, LSTM models do far better than the traditional approaches like Support Vector Machines (SVMs). They further demonstrated the viability of simpler architectures such as Feedforward Neural Networks (FNNs) for speech emotion recognition, offering reduced computational complexity for real-time applications.

2.3. Face-based Emotion Recognition

Facial expressions are highly expressive of human emotion and have been widely adopted in emotion recognition tasks. Tools such as DeepFace enable automated facial feature extraction and emotion inference using deep convolutional neural networks. (Fan, Jing, & Wang, 2025) noted that for more accurate emotion classification when combined with vocal input, visual cues, such as eye movement and lip positioning, significantly in that classification. Their study revealed that multimodal fusion of facial and speech features achieved higher classification accuracy than either modality used in isolation.

2.4. Multimodal Emotion Recognition

Multimodal systems leverage the complementary strengths of multiple modalities, text, audio, and video, to create a more holistic emotion recognition model. (Fan, Jing, & Wang, 2025) proposed a multi-stream transformer-based architecture that integrates features from all three modalities and showed a significant performance boost over unimodal models. In the same vein, (Chutia & Baruah, 2024) delved into deep learning techniques used in multimodal

systems and determined that models incorporating attention mechanisms and feature fusion layers tend to outperform traditional concatenation-based methods.

3. Methodology

The proposed system will be able to take big data (video Data) as input, extract the audio from this data, use the extracted audio, transcribe it to text, giving out three data from the input data, which includes video, audio, and text. Deep learning models will be used to estimate emotions on the video data, estimate emotions on the audio data, as well as the text data.

The DeepFace model will be used to estimate emotions on the video data. The DeepFace model is selected because of its extensiveness, accuracy, and it was trained on a massive 4 million data. A deep learning model will be trained to estimate emotions from the audio and the text. After training each of the models and evaluating the accuracy of the DeepFace model, a robust integration will be done to enable the models to work together. A UI will be developed to enable the use of the models, which will be working in the backend.

The Architecture of the Multimodal system is shown figure 1.

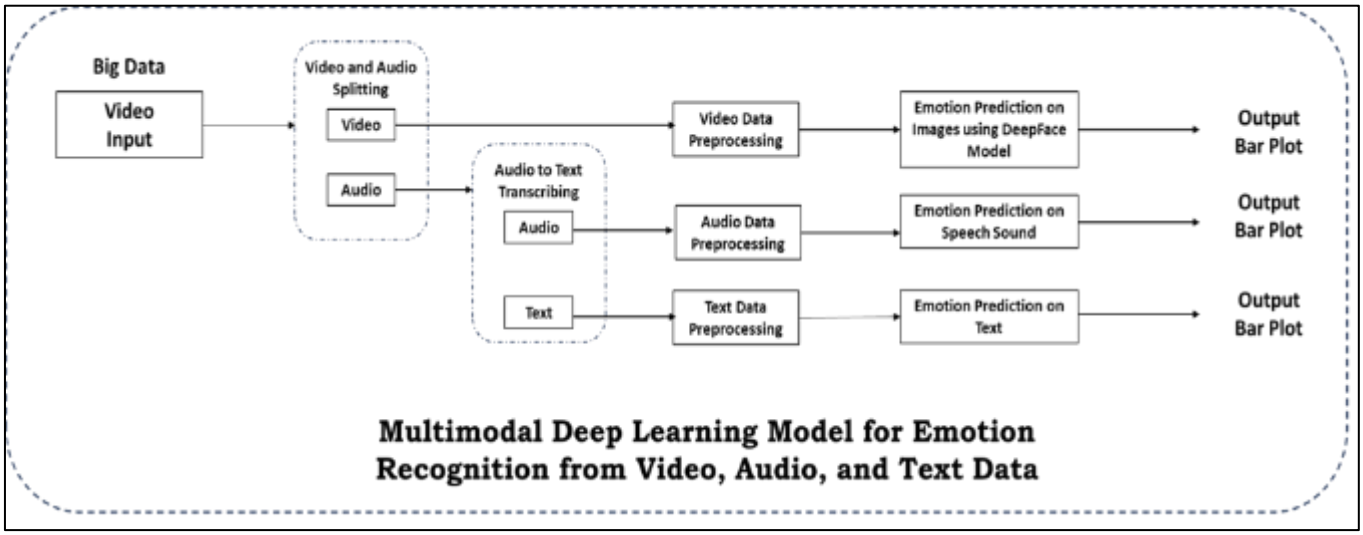


Figure 1 System Architecture of the Multimodal Deep Learning Model for Emotion Recognition from Video, Audio, and Text Data

The methodology adopted for the development of this system is the multi-phase methodology. This methodology is most suitable for the development of a complex system. The methodology is a structured approach that divides the system development into phases of different milestones, each with a specific objective to achieve. This methodology ensures that the complex task of designing and implementing the multimodal model is handled systematically. Each phase focuses on a specific aspect of the system, starting from data acquisition to preprocessing, model development, integration, and visualization. The methodology facilitates the seamless combination of multiple modalities (video, audio, and text) to ensure robust and accurate emotion recognition, leveraging tools like DeepFace, TensorFlow, pandas, numpy, Librosa, nltk, etc. The phases in the multiphase methodology include:

- System planning and design
- Data collection and pre-processing
- Feature extraction from big data
- Model selection, development, and training
- Model accuracy evaluation
- System implementation
- Comprehensive system testing, result visualization, and documentation

4. Model Selection and Training

Three models will be working on the system, which includes: Video model, for classifying emotions on the video file, Audio model for classifying emotions on the audio data, and Text model for classifying emotions on the text data.

4.1. Video Model

The video data comprises frames of images. The prediction that will be done on the image data will be done by predicting emotions on individual frames and totalling the entire result of emotion scores of each image. For the implementation of this system, a pre-trained model will be used to estimate emotions in the image. The selected model is DeepFace. The DeepFace model is a DNN model for face emotion recognition, facial expression recognition, etc. It is one of the pioneers of DNN models for facial expression recognition. This model was selected because of its high accuracy of over 97.35%. It was trained on a massive data set of over 4 million labelled images.

The DeepFace model was evaluated on different faces of different emotions. This was to evaluate the performance of the model, its performance, and its accuracy.

The different text done with different images of different facial expressions with DeepFace is shown in figure 2.

Figure 2 Different Face Predictions from DeepFace

4.2. Audio Model Training

The model for emotion classification was trained on a large 5600-labeled audio dataset. The tools being used include: Anaconda, Jupyter Notebook, Librosa, tensorflow, matplotlib, seaborn, pandas, numpy. etc. The dataset for training the model was the Toronto Emotional Speech set data. The dataset is a well-known and extensive dataset for speech-emotional analysis. The dataset is comprised of 5600 labeled speech sounds for different emotions which include: Anger, disgust, fear, happy, neutral, surprise, and sad. Each of the emotions has 800 labelled data making the dataset a total of 5600. The dataset was downloaded from Kaggle.

Each of the emotions are audio file in respective folders according to their label. The directory where they are stored was read using a Python code to extract the file name of each file and its label. These data were stored in a panda's series for easy access. After this, the bar chart of the data count was plotted using Seaborn. This is to confirm the size of the dataset before further processing. The bar chart is shown figure 3.

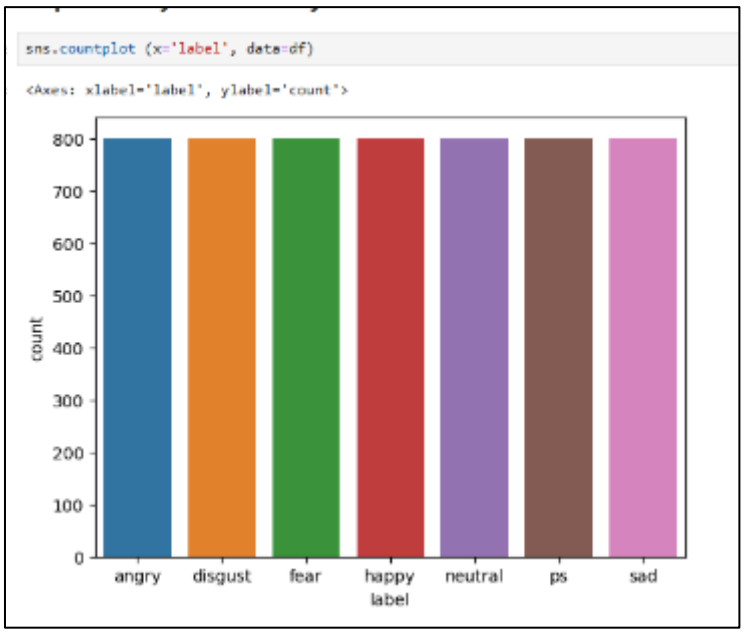


Figure 3 Bar chart of audio data count

4.2.1. Audio feature extraction

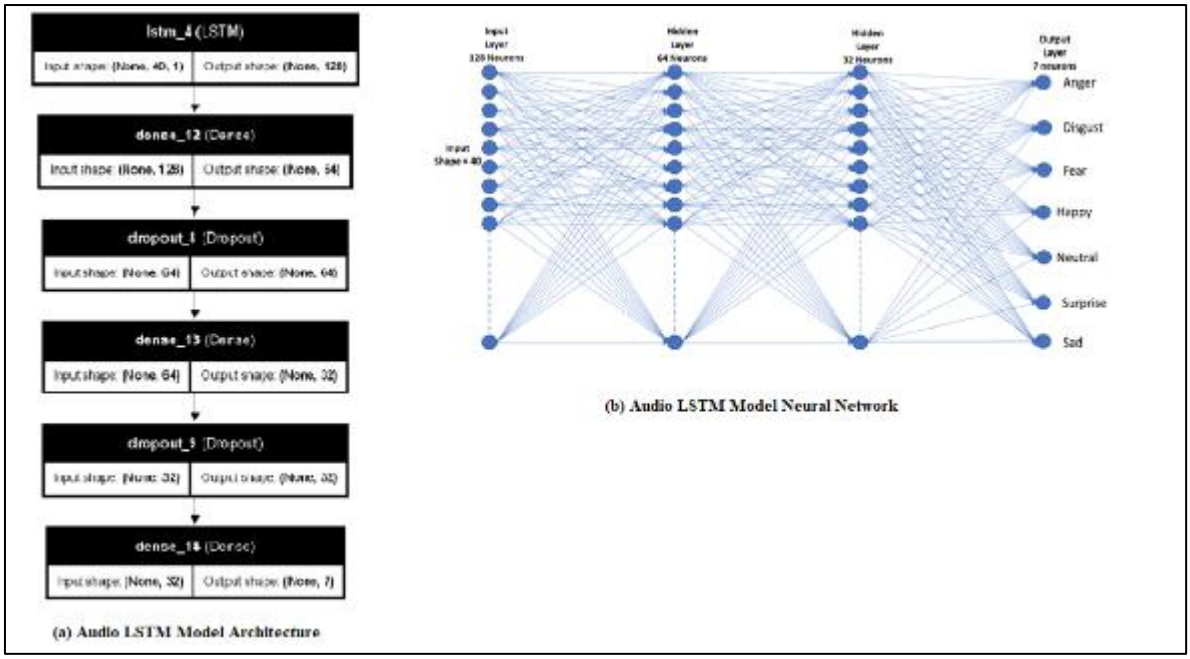
At this stage, different features were extracted from the audio sound and visualized. The different features extracted are the waveform and spectrogram. The waveform shows the amplitude of the audio signal over time, showing the intensity of the audio at each point, while the spectrogram is a visual representation of the audio's frequency spectrum as it varies over time. **Short-Time Fourier Transform (STFT)** was used to generate the spectrogram**.** Both the waveform and spectrogram were computed using the Librosa Python library and plotted using matplotlib.

The waveform and spectrogram of different emotions are shown below:

The main feature needed from the audio files is the mel-frequency cepstal coefficient (mfcc). The first 40 mfcc were extracted and compiled as a numpy array. This is the feature fed into the neural network for training. After this extraction, the data was slice into train data and text data. The train data was used to train the model while the text data was used to evaluate the model performance after the training.

The LSTM was selected to be used for the audio model. LSTM is a variant of RNN. LSTM is suitable for time series which audio data fall under.

The input shape of the model will be 40 because the MFCC extracted from the audio data is 40, and the output is 7 because we have 7 emotion datasets. The architecture of the LSTM model and its neural logic is shown in figure 4



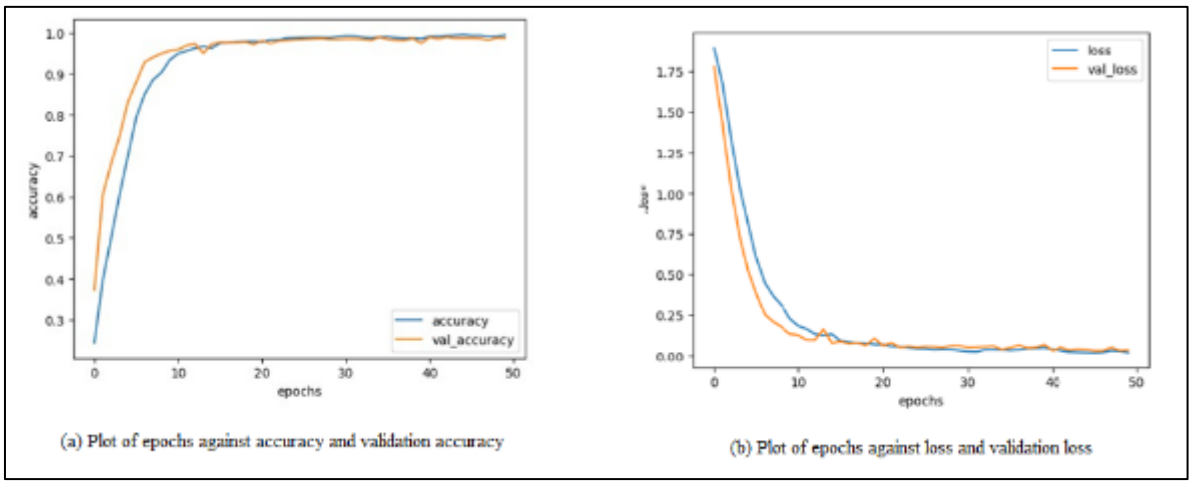
(a) Audio LSTM Model Architecture

(b) Audio LSTM Model Neural Network

Figure 4 Audio LSTM Model Architecture and Neural Network

4.2.2. Audio Model Training

The model was configured to train on the training dataset while using the testing dataset for validation. It underwent training for 50 epochs, during which the accuracy and loss metrics were monitored and recorded. The last validation accuracy was 0.9866, and the previous validation loss was 0.0345. This means that the model has an accuracy of 98% and a loss of 3%. The progress of these metrics is visualized in the plots in figures 5.



(a) Plot of epochs against accuracy and validation accuracy

(b) Plot of epochs against loss and validation loss

Figure 5 Plot of epochs against accuracy and validation accuracy; and loss and validation loss

After the training, the test data were used to make predictions about the model. Using these predicted labels and the labels of the test data, the accuracy of the model was evaluated.

4.2.3. Audio Model evaluation

The model evaluation showing the accuracy and loss, and the confusion matrix is shown figure 6.

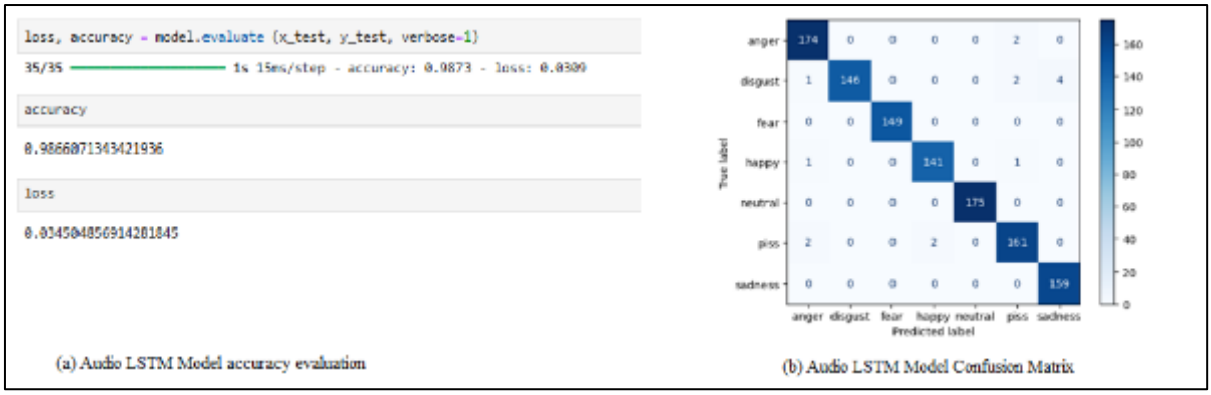


Figure 6 Audio LSTM Model accuracy evaluation and its Confusion Matrix

4.3. Text Model Training

The model for classifying emotions on the text data was trained on a vast 34,000 labelled text data. The tools being used include Anaconda, Jupyter Notebook, NLTK, TensorFlow, matplotlib, seaborn, pandas, and NumPy.

The dataset used for this model is the emotion dataset for NLP downloaded from Kaggle. The dataset comprises labelled emotional text. The different emotions in the dataset include sad, anger, love, joy, fear, and surprise. The dataset has a total of 34,000 labelled data.

The Emotion dataset for Natural Language Processing (NLP) is provided in text files, each containing raw textual data corresponding to emotion labels. To prepare this data for analysis and model training, several critical pre-processing steps were performed:

- **Reading and Transferring Data to a Pandas DataFrame**: The text data and associated emotion labels were read from the dataset and loaded into a Pandas DataFrame for efficient manipulation and analysis. This step provided a structured format, enabling easy access to the textual content and their labels.
- **Handling Missing Values**: The dataset was carefully inspected for null or missing values, which can negatively impact the performance and reliability of the model. All rows with null values were identified and removed to ensure that the dataset was complete and consistent. This cleaning process is crucial as it prevents errors during feature extraction and model training phases.
- **Ensuring Data Integrity**: After removing null values, the dataset was re-examined to confirm its integrity. Ensuring that no unintended data points were lost during the cleaning process is a vital step in maintaining the quality of the dataset.
- **Rationale for Cleaning**: The removal of null values is an essential pre-processing task, as incomplete data can lead to inaccuracy during training, reducing the overall performance of the model. Additionally, cleaning the dataset eliminates potential noise, allowing the models to focus on learning meaningful patterns from the available data.

By structuring the data into a clean and well-prepared data frame, it was ready for the subsequent steps of feature extraction and model training, ensuring a strong foundation for building a robust emotion recognition model. A screenshot of the dataframe is shown figure 7.

```
[3]: df = pd.concat ([train, test, val])
     df.columns = ['text', 'label']
     df.reset_index(drop=True)
     df.head()
```

[3]:

	text	label
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger

Figure 7 Text Model Pandas Data after Pre-processing

Feature extraction from text data is essential in Natural Language Processing (NLP) which transforms raw text into a structured format suitable for model training. In this project, feature extraction from the text data was carried out in different stages.

- **Removing Punctuation:** Punctuations usage have important meaning to humans but insignificant to AI models. They introduce noise to the data and increase the complexity of processing data. All punctuation was systematically removed from the data frame, leaving only raw text data.

- **Tokenization:** Tokenization is the process of breaking down text into individual units; which can be words, phrases, or even characters. Each sentence was split into a list of words allowing a better handling and analysis.

- **Stop words removal**: Stop words are commonly used words such as 'the', 'is', 'in', etc. These words occur in sentences but contribute little contextual understanding of the data. These words were identified and removed using the stop word function in nltk.

The dataset after the feature extraction was split into training data and testing data. The training data was used to train the model while the testing data was used to evaluate the model after training. A screenshot of the data after the feature extraction is shown figure 8.

[13]:

	text	label	features
0	i didnt feel humiliated	sadness	didnt feel humili
1	i can go from feeling so hopeless to so damned...	sadness	go feel hopeless damn hope around someon care ...
2	im grabbing a minute to post i feel greedy wrong	anger	im grab minut post feel greedi wrong
3	i am ever feeling nostalgic about the fireplac...	love	ever feel nostalg fireplac know still properti
4	i am feeling grouchy	anger	feel grouchi
...	...	...	...
1995	im having ssa examination tomorrow in the morn...	sadness	im ssa examin tomorrow morn im quit well prepa...
1996	i constantly worry about their fight against n...	joy	constantli worri fight natur push limit inner ...
1997	i feel its important to share this info for th...	joy	feel import share info experi thing
1998	i truly feel that if you are passionate enough...	joy	truli feel passion enough someth stay true suc...
1999	i feel like i just wanna buy any cute make up ...	joy	feel like wanna buy cute make see onlin even one

34000 rows × 3 columns

Figure 8 Text Model data after Feature Extraction

The Feedforward Neural Network is selected for building the text model. The input shape of the model is 5000, and the output is 6. The architecture of the FNN model and its Neural Network is shown figure 9.

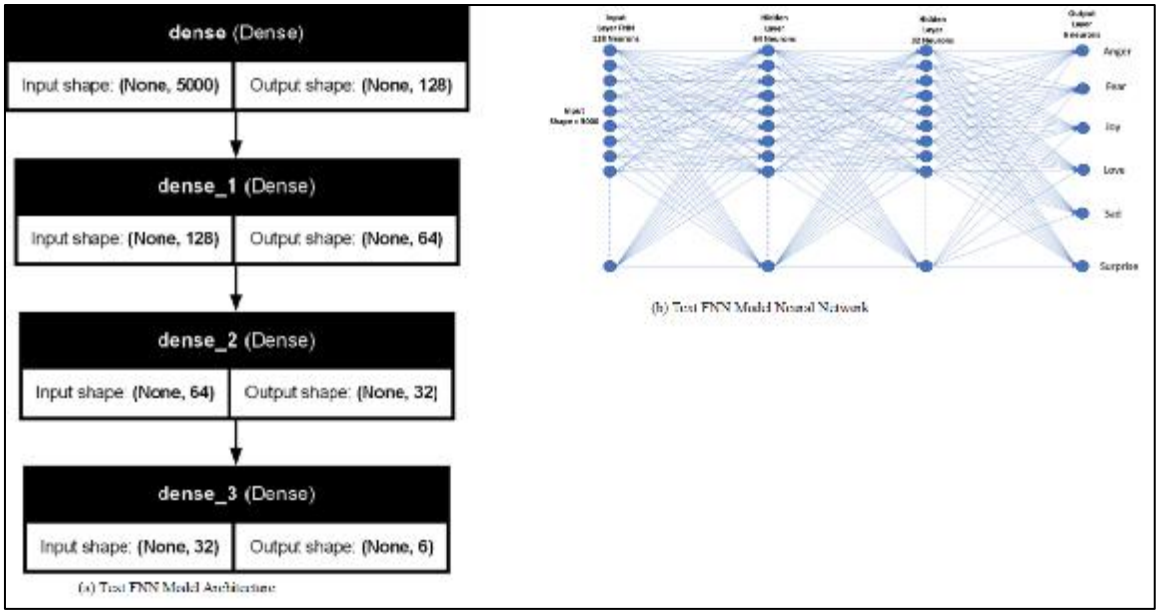
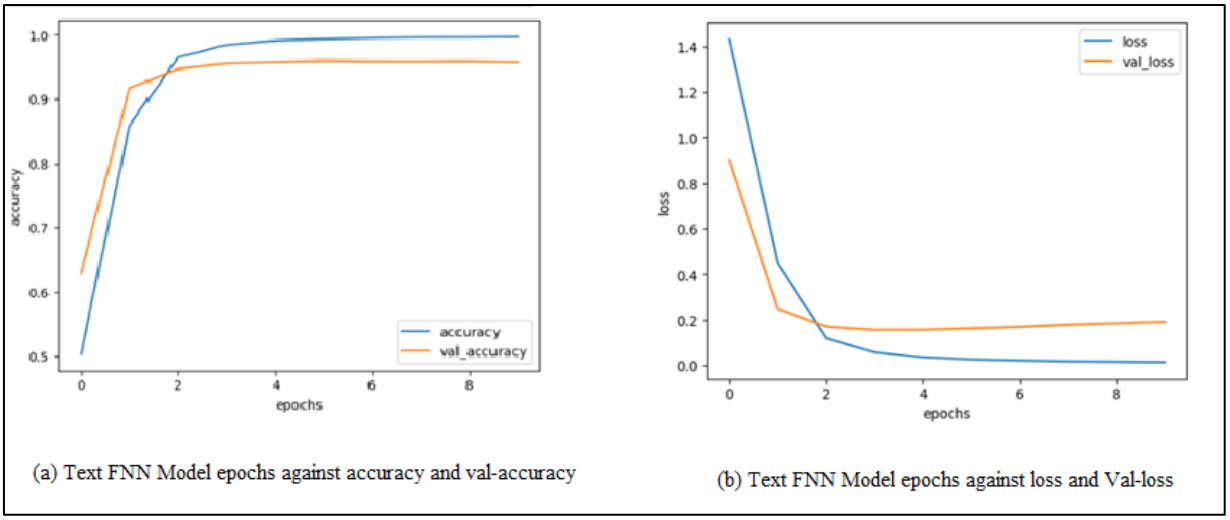


Figure 9 Text FNN Model Architecture and Neural Network

The model was configured to train on the training dataset while using the testing dataset for validation. It underwent training for 10 epochs, during which the accuracy and loss metrics were monitored and recorded. The last validation accuracy was 0.9565 and the previous validation loss was 0.1901. This means that the model has an accuracy of 95% and a loss of 19%. The progress of these metrics is visualized in the plots in figure 10.



(a) Text FNN Model epochs against accuracy and val-accuracy

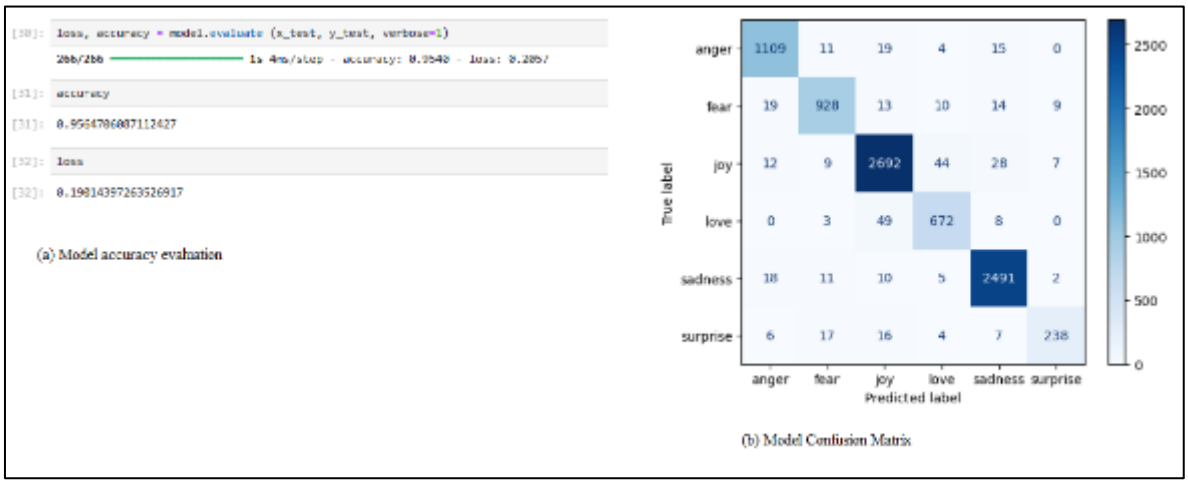
(b) Text FNN Model epochs against loss and Val-loss

Figure 10 Text FNN Model epochs against accuracy and val-accuracy, loss and Val-loss

4.3.1. Text Model Evaluation

After the training, the test data was used to make predictions on the model. Using these predicted labels and the labels of the test data, the accuracy of the model was evaluated.

The model evaluation and confusion matrix are shown figure 11.



(a) Model accuracy evaluation

(b) Model Confusion Matrix

Figure 11 Model accuracy evaluation and Confusion Matrix

5. Implementation

An Interactive UI was built for the implementation of the system. The System as its input, accepts video file. The file can be of the format MP4, AVI, etc. The system takes this video file, extracts the audio, then further transcribes the audio to text and run emotion classification on the video, audio and text using DeepFace, LSTM model and FNN model respectively.

The prediction done by the model comprises the emotion score from the input data. This score was plotted in a bar chart for easy visualization and shown on the system UI.

5.1. System Algorithm

The algorithm of the system operation is shown below:

- Click the Load File button

- Select a video file
- Open the video file
- Extract audio from the video
- Extract text from the audio
- Extract frames from the video
- Predict on the frames
- Extract features from the audio
- Extract features from the text
- Make predictions on the audio
- Make predictions on the text
- Plot a bar plot of the output
- Display the plot.

5.2. System Flowchart

The flowchart of the system operation is shown in figure 12.

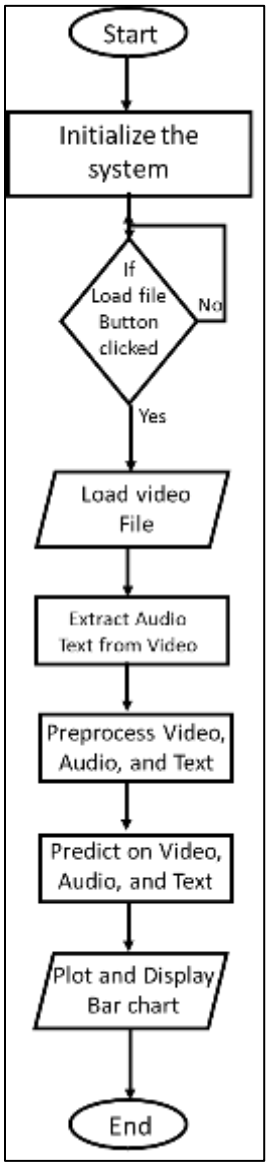


Figure 12 System Flowchart

6. Result

After the different models were trained and evaluated, they were combined to achieve the multimodal system. This enables the system to handle the different features of each specific model's different data types.

A simple user interface was built to test the models. This enables the user to feed data into the model and visualize the output result from the model. The user interface is shown in the figure 13.

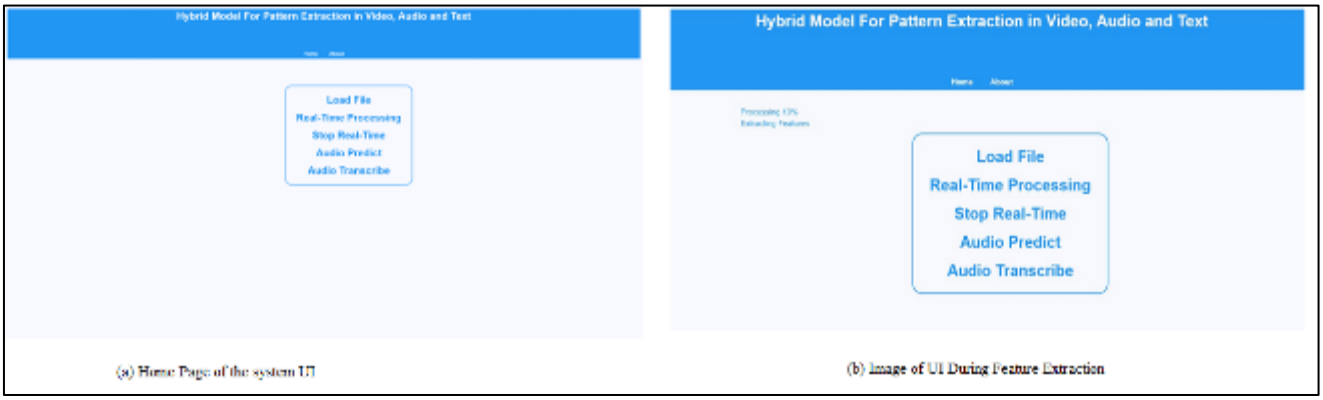


Figure 13 Home Page of the system UI and its Image During Feature Extraction

When the load button was clicked, the system opened the file picker for the user to select a video file. When a video file is selected, the file picker returns the file path to the selected file. This file path is used to open the video file to extract various features from it. A screenshot of this from the UI is shown figure 13.

The feature extractions include extracting the audio from the video. When the system first extracts the audio from the video, it goes ahead to transcribe the audio into text. After the feature extraction, the system has three types of data to work with which include video data, audio data, and text data.

After extracting the features, the system goes on to start predicting on the data. First, it predicts on the video by processing the video frame by frame and feeding each frame to the deepface model for emotion classification. After the prediction on the video, the system starts predicting on the audio. The extracted audio is fed to the LSTM model for emotion classification. Lastly the extracted text will be fed to the FNN model for emotion classification. The screenshot of the following processes is shown figure 14.

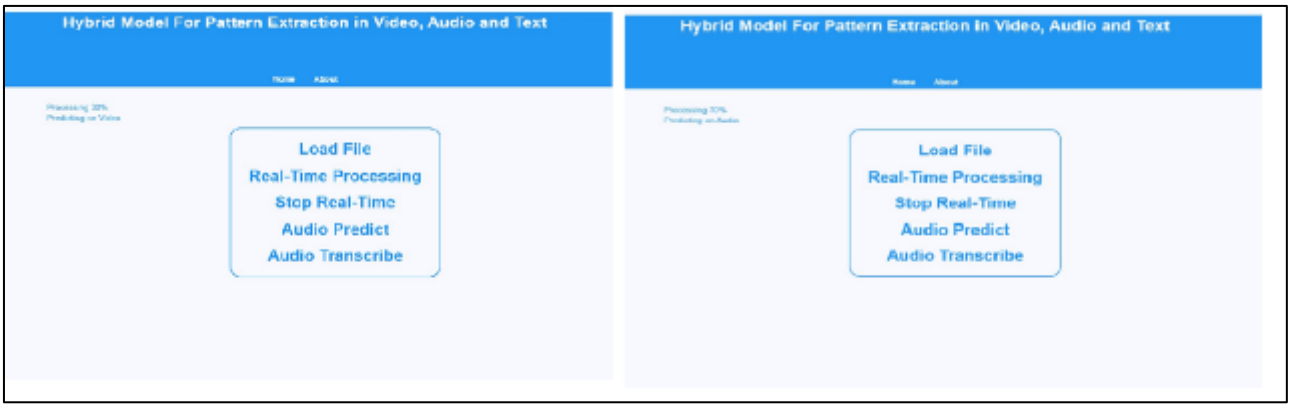


Figure 14 Image of UI during Prediction on Video and Prediction on Audio

After the prediction, the system now processes the entire data and plots a bar plot of the different predictions using seaborn and matplotlib. After plotting the barplot, the system displays the plot on the UI for further comparison and analysis. The bar plots from this testing are shown figure 15.

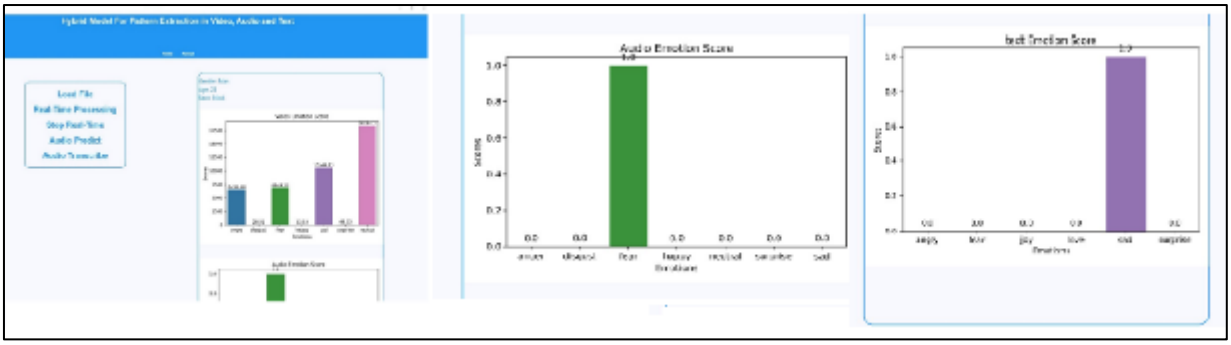


Figure 15 Result from Pre-Recorded Video Processing

From the result above, the facial expressions from the video are more neutral, the speech sound shows the person is afraid, and the text shows the person is sad.

7. Conclusion

This research demonstrates the effectiveness of a multimodal deep learning model in recognizing human emotions from synchronized video, audio, and text inputs. By integrating DeepFace for facial emotion and demographic recognition, an LSTM model for audio-based emotion detection, and an FNN model for text emotion analysis, the system delivers a more holistic and accurate interpretation of emotional states. Experimental results affirm that combining modalities significantly outperforms unimodal models, especially in complex or ambiguous contexts where a single data form may be insufficient. Overall, this model contributes to the growing body of affective computing research by providing a scalable and adaptable solution for emotion recognition in real-world applications.

Recommendation

Future studies should consider using datasets that include many different languages for both text and speech. At the moment, many emotion recognition models are trained mostly on English language data, which makes it harder for them to work well in other languages. By adding support for more languages, including fewer common ones, the system can better recognize emotions in people from different cultures and backgrounds. This would make the model more useful in real-world applications like global customer service, multilingual chatbots, and international healthcare support.

Another area to improve is how well the model works on small devices like mobile phones, smartwatches, or other edge devices. Right now, the system may need a lot of computing power. In the future, the model can be made smaller and faster using techniques like model compression or pruning. This would help the system give quick emotion predictions without needing a powerful computer. Real-time emotion recognition on edge devices could be useful in areas like live online learning, therapy apps, smart assistants, and even security systems.

Compliance with ethical standards

Statement of ethical approval

The authors declare that all procedures performed in this Study were in accordance with ethical standards and relevant guidelines.

Disclosure of conflict of interest

The authors declare that there is no conflict of interest.

References

[1] Alam, K., Nigar, N., Erler, H., & Banerjee, A. (2023). Speech Emotion Recognition from Audio Files Using Feedforward Neural Network. *International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-6). Chittagong: IEEE.

[2] Bejjagam, L., & Chakradhara, R. (2022). Facial Emotion Recognition using Convolutional Neural Network with Multiclass Classification and Bayesian Optimization for Hyper Parameter Tuning. Karlskrona, Sweden: Faculty of Computing, Blekinge Institute of Technology. Retrieved from https://urn.kb.se/resolve?urn=urn:nbnse:bth-23359

[3] Chutia, T., & Baruah, N. (2024). A Review on Emotion Detection by Using Deep Learning Techniques. *Artificial Intelligence Review, 57*(203). doi:10.1007/s10462-024-10831-1

[4] Fan, S., Jing, J., & Wang, C. (2025). Audio-Visual Learning for Multimodal Emotion Recognition. *Symmetry, 17*(418), 1-21. doi:10.3390/sym17030418

[5] Filali, H., Boulealam, C., Tairi, H., El Fazazy, K., Mahraz, A. M., & Riffi, J. (2025). Comparative study of Unimodal and Multimodal Systems Based on MNN. *Journal of Information Systems Engineering and Management, 10*(18s), 355–363.

[6] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face. *Entropy, 25*(1440), 1-33. doi:https://doi.org/10.3390/e25101440

[7] Machová, K., Szabóova, M., Paralič, J., & Mičko, J. (2023). Detection of Emotion by Text Analysis Using Machine Learning. *Frontiers in Psychology, 14*(1190326), 1-14. doi:10.3389/fpsyg.2023.1190326

[8] Nouman, A., Iqbal, H., & Saleem, M. K. (2023). An Emotion Recognition System: Bridging the Gap Between Human-Machine Interaction. *IAES International Journal of Robotics and Automation (IJRA), 12*(4), 315-324. doi:10.11591/ijra.v12i4.pp315-324