**EMOTIFY : MULTIMODAL EMOTION DETECTION SYSTEM**

**UNIVERSITY**

**Emotify : Multimodal Emotion Detection System**

Submitted in the partial fulfillment of the degree of

**BACHELOR OF TECHNOLOGY**

In

**COMPUTER SCIENCE & ENGINEERING**

Under

**UNIVERSITY**

BY

UNDER THE GUIDANCE OF

**PROF.**

**Approval Certificate**

This is to certify that the project report entitled “EMOTIFY : MULTIMODAL EMOTION DETECTION SYSTEM ” submitted by

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Prof.**

# 

# ACKNOWLEDGEMENT

The endless thanks go to Lord Almighty for all the blessings he has showered onto us, which has enabled us to write this last note in my research work. During the period of my research, as in the rest of my life, we have been blessed by Almighty with some extraordinary people who have spun a web of support around us. Words can never be enough in expressing how grateful we are to those incredible people in our life who have made this thesis possible. We would like an attempt to thank them for making our time during our research in the Institute a period we will treasure. We are deeply indebted to our research supervisor, Professor Debajyoti Chatterjee, and Professor Santanu Basak for introducing us such an interesting thesis topic. Each meeting with them added in valuable aspects to the implementation and broadened our perspective. They have guided us with their invaluable suggestions, lightened up the way in our darkest times and encouraged us a lot in the academic life.

Shalu Kumawat

# ABSTRACT

This work introduces EmotiFy, a multimodal emotion recognition system based on deep learning that employs facial and speech modalities to effectively recognize human emotions. As opposed to conventional unimodal approaches which make use of facial expressions or speech alone, EmotiFy utilizes both to provide better prediction accuracy and context sensitivity. The architecture uses a ResNet50 convolutional neural network for extracting facial images' visual features and bi-directional LSTM to process the temporal MFCC features of the speech. Two such representations are combined to produce an integrated emotional context, and that is passed on to categorize the input to one out of seven categories of emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.The model is trained on two established datasets—FER2013 for visual emotion and TESS for speech—and is released as an easy-to-use Streamlit web app where users can upload audio and image inputs for real-time emotion prediction.The findings show that multimodal fusion has a remarkable improvement in emotion recognition performance over unimodal models, pointing to the potential of EmotiFy in applications such as human-computer interaction, virtual assistants, mental health monitoring, and more.

**Table of Contents**

[ACKNOWLEDGEMENT IV](#_Toc28561)

[ABSTRACT V](#_Toc28562)

[LIST OF FIGURES VIII](#_Toc28563)

[LIST OF TABLES IX](#_Toc28564)

[1. Introduction 1](#_Toc28565)

[Deep Learning 1](#_Toc28566)

[Facial Emotion Recognition 1](#_Toc28567)

[Voice Based Emotion Detection 2](#_Toc28568)

[Dataset Overview : Speech Dataset – TESS 2](#_Toc28569)

[Dataset Overview : FER2013 3](#_Toc28570)

[2. LITERATURE REVIEW 5](#_Toc28571)

[Previous works related to the project 5](#_Toc28572)

[Research gaps identified 6](#_Toc28573)

[3. METHODOLOGY 8](#_Toc28574)

[Tools, technologies, and software used 8](#_Toc28575)

[System architecture or framework 8](#_Toc28576)

[Models overview 9](#_Toc28577)

[3.3.1 RestNet50 9](#_Toc28578)

[3.3.2 LSTM 10](#_Toc28579)

[3.3.3 Fusion Model 11](#_Toc28580)

[4. DESIGN & IMPLEMENTATION 13](#_Toc28581)

[4.1 System Design Overview 13](#_Toc28582)

[5. RESULTS & DISCUSSIONS 17](#_Toc28583)

[Experimental setup 17](#_Toc28584)

[Comparisons with existing models 20](#_Toc28585)

[Performance Evalution 21](#_Toc28586)

[6. CONCLUSION & FUTURE SCOPE 25](#_Toc28587)

[Summary of work 25](#_Toc28588)

[Limitations and scope for future improvements 25](#_Toc28589)

[BIBLIOGRAPHY 27](#_Toc28590)

# LIST OF FIGURES

|  |  |
| --- | --- |
| Figure 1.1 Sample Audio Data | 2 |
| Figure 1.2 S Spectrogram of the Angry input | 3 |
| Figure 1.3 Distribution of TESS Dataset | 3 |
| Figure 1.4 Distribution of Emotion FER2013 dataset | 4 |
| Figure 1.5 Sample of FER2013 dataset | 4 |
| Figure 3.1 Model Summary for Resnet-50 | 10 |
| Figure 3.2 Model Summary for LSTM | 11 |
| Figure 3.3 Model Summary for Fusion Model | 12 |
| Figure 4.1 Model Architecture | 14 |
| Figure 5.3.1 Performance Comparison Plot | 21 |
| Figure 5.3.2 LSTM Confusion Matrix | 21 |
| Figure 5.3.3 Resnet-50 Confusion Matrix | 22 |
| Figure 5.3.4 Multi Model Confusion Matrix | 22 |
| Figure 5.3.5 Accuracy for Resnet-50 | 23 |
| Figure 5.3.6 Accuracy for Multi Model | 23 |
| Figure 5.3.7 Accuracy for LSTM | 24 |

# LIST OF TABLES

Table 5.1.1 Hardware Configuration

Table 5.1.2 Software and Libraries Configuration

Table 5.1.3 Model Configuration

Table 5.2.1 Comparisons with existing models Configuration

# 1. Introduction

Human communication is strongly tied to emotions and is expressed by both facial and vocal expressions. Traditional machine learning approaches to emotion recognition, however, are usually based on one input modality—either visual or audio—thus restricting their performance, particularly in real-world scenarios where one of the signals can be noisy, missing, or ambiguous.This project focuses on creating EmotiFy, a multimodal emotion recognition system that integrates facial expression analysis from static images and speech emotion recognition from audio signals. The objective is to enhance the robustness and accuracy of emotion classification by utilizing the strengths of both modalities. To do so, the system employs a ResNet50 convolutional neural network to extract spatial features from facial images and a bi-directional LSTM network to extract temporal features from Mel-Frequency Cepstral Coefficients (MFCCs) derived from speech. These feature sets are combined and fed into fully connected layers to predict one of seven basic emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.The system is trained on two benchmark datasets—FER2013 for face images and TESS for speech audio— and is put out through a web interface for real-time emotion prediction. The fusion-based system is expected to bridge the gap between unimodal systems and real-world emotion understanding in AI.

## Deep Learning

In 1965, Gordon Moore predicted that the number of transistors that could be integrated into a single die would grow exponentially with time. Moore's law hasgoverned microprocessor manufacturing processes, and consequently microprocessor performance ever since. However, recent studies indicate that during the nexttwo decades, the laws of nature will begin to govern microprocessor design andfabrication(Niemer, 2004).

## Facial Emotion Recognition

Facial Emotion Recognition (FER) is an area of artificial intelligence (AI) and computer vision that detects and interprets human emotions through facial expressions. Emotions such as happiness, sorrow, anger, fear, surprise, and disgust may be classified using facial muscle movements and the location of features such as the eyes, lips, and brows. This technology has a wide range of applications, including increasing user experiences in entertainment and marketing, as well as mental health diagnoses and classroom participation..

## Voice Based Emotion Detection

Voice-based emotion detection, also known as **Speech Emotion Recognition (SER)**, aims to identify human emotions solely from **vocal characteristics** such as tone, pitch, intensity, rhythm, and speech patterns. Unlike textual sentiment analysis, SER captures the *paralinguistic features* of speech that reveal emotional states even when the words themselves are neutral.This component of the **EmotiFy** system processes raw audio data to classify emotions into one of seven categories:

**Angry, Disgust, Fear, Happy, Sad, Surprise**, and **Neutral**.

## Dataset Overview : Speech Dataset – TESS

The Toronto Emotional Speech Set (TESS) is used for training the voice-based model. It contains recordings of two actresses saying a set of 200 target words, each spoken in 7 different emotional tones. Each audio sample is labeled with one of the target emotions, making it ideal for supervised deep learning.

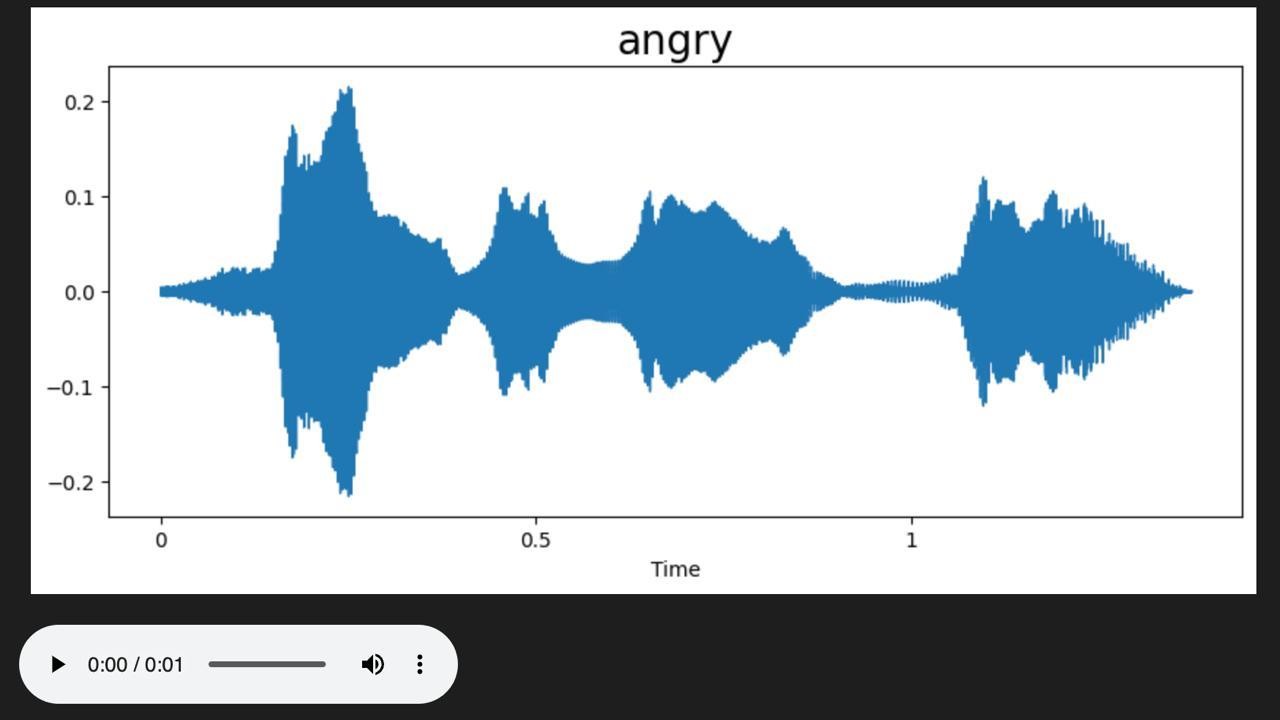


Figure 1.1 Sample Audio Data

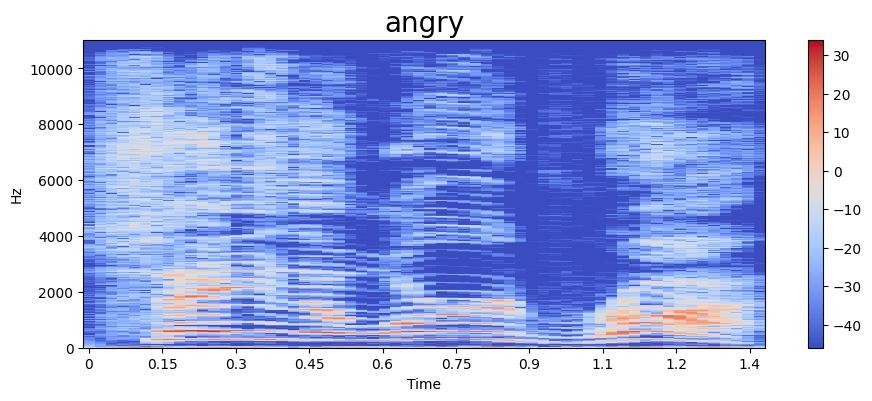


Figure 1.2 Spectrogram of the Angry input

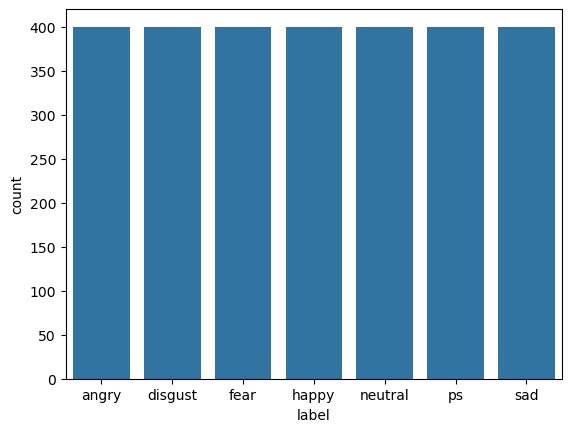


Figure 1.3 Distribution of TESS Dataset

## Dataset Overview : FER2013

The FER2013 dataset (Facial Expression Recognition 2013) is a widely used benchmark dataset for emotion classification tasks in computer vision. It was first introduced at the ICML 2013 Challenges in Representation Learning and is commonly used to train and evaluate facial emotion identification models. The FER2013 dataset was developed by the Institute of Robotics and

Intelligent Machines at Tsinghua University in Beijing, China [4]. The collection contains 35,887 images that are standardized to 48x48 pixels in grayscale. FER2013 is not a balanced dataset since it includes photos of seven facial expressions: angry (4,953), disgust (547), fear (5,121), happy (8,989), sad (6,077), surprise (4,002), and neutral (6,198).

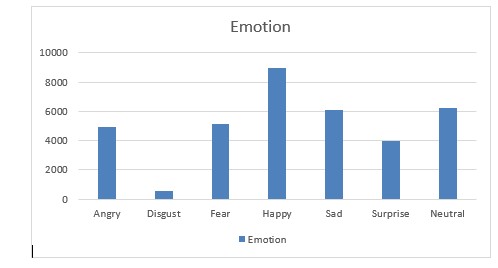


Figure 1.4 Distribution of Emotion FER2013 dataset



Figure 1.5 Sample of FER2013 dataset

# 2. LITERATURE REVIEW

## Previous works related to the project

Kırbız [1](2025) highlights the importance of dataset quality and balance when dealing with facial emotion recognition tasks. In her work, "Improving facial emotion recognition through dataset merging and balanced training strategies", she illustrates that combining complementary datasets and using balanced training will significantly improve the accuracy of recognition. The study suggests that imbalanced datasets tend to bias the model toward dominating classes, and this can be alleviated through the use of intelligent sampling and fusion techniques during training (Kırbız, 2025). For speech emotion recognition (SER), Ramakrishnan [5](2012) presented a foundational review of the field, discussing how features such as pitch, energy, MFCCs, and duration contribute to accurate emotion modeling. His work emphasized the variability and complexity of vocal emotional expression, underlining the importance of carefully designed preprocessing and feature extraction. In speech emotion recognition (SER), Fayek et al. [2](2017) presented a thorough assessment of deep learning architectures, i.e., convolutional and recurrent neural networks. The results demonstrate that temporal dynamics in speech play an important role in emotion classification and that recurrent models, particularly those with LSTM layers, are well-suited to capture such dynamics. They also highlight the significance of handcrafted and learned features in training strong SER systems (Fayek et al., 2017). In multimodal emotion recognition, Go et al. (2003)[6] are among the earlier researchers who researched the combination of facial images and speech signals in emotion detection with the aim to show that multistream fusion improves emotion classification accuracy over the single modality counterparts. Previously, Busso et al.[3] (2004) had performed one of the classic studies on multimodal emotion recognition. Their work contrasted the performance of unimodal (facial or speech) systems with multimodal systems combining both. They noted that multimodal solutions always performed better than unimodal ones, especially in noisy or uncertain situations, confirming the power in bringing together visual and aural cues for interpreting emotion. Based on this, Wang et al. [4](2020) presented a framework that optimally combines facial expression and speech features for better emotion recognition. They showed that strategic fusion at the feature level performs better than single modalities and is able to capture the intricate interaction between verbal and non-verbal emotional expressions. Based on these findings, EmotiFy employs a hybrid approach that uses ResNet50 for emotion recognition from faces and LSTM-based models for SER based on MFCCs. The system is trained with balanced strategies and combines both modalities at the feature level to provide precise, real-time emotion detection, in line with the methodologies and results of these seminal studies.

## Research gaps identified

Though huge amounts of work have been carried out on emotion recognition using visual and auditory modalities separately, there are a number of major gaps in existing work. EmotiFy was planned with the aim of filling some of these gaps. The most significant research gaps recognized from the review of literature are as follows:

1. **Few Multimodal Fusion Approaches**

The majority of current systems are unimodal based on either facial expressions or speech alone, which decreases accuracy in real-world applications. While there are some multimodal works done, most are late fusion, which misses more elaborate interaction between features of multiple modalities. Intermediate fusion approaches that fuse feature-level representations are needed for a better integrated understanding of emotions.

1. **Inadequate Real-Time Implementation**

Most state-of-the-art emotion recognition models are not naturally designed to be deployed in real time. They are concerned mostly with model performance in static datasets but do not naturally extend to end-user, real-time applications that one can interactively employ as a non-expert.

1. **Dataset Incompatibility and Alignment Challenges**

There is a lack of large, synchronized multimodal datasets that contain both visual and auditory emotional information. The majority of datasets are unimodal, and multimodal training is challenging. Synchronizing independent datasets such as FER2013 (images) and TESS (audio) is technically challenging, including timing synchronization and labeling consistency.

1. **Cultural and Linguistic Limitations**

Many models are trained on datasets that lack cultural and linguistic diversity, limiting their effectiveness when applied to different populations. Emotion expression can vary widely across cultures, especially in vocal tone and facial expression intensity.

1. **Limited Exploration of Fusion Architectures**

Few works investigate alternative fusion architectures, e.g., employing ResNet for spatial and LSTM for temporal features, and combining them into a hybrid classification model. There is scope to try more with transformers, attention mechanisms, or cross-modal learning for richer contextual understanding.

1. **Lack of User-Centric Applications**

The majority of current academia models tend to stay as proof-of-concept experiments. Lacking is how to create everyday practical tools or web applications that allow users to interact with the system, post their data, and get their predictions—Closing the gap to real-world utilization.

# 3. METHODOLOGY

The EmotiFy system is developed using a **multimodal deep learning approach** that combines facial image data and speech audio to accurately detect human emotions. The methodology is divided into five major stages: **Data Collection and Preprocessing**, **Feature Extraction**, **Model Architecture Design**, **Multimodal Fusion**, and **System Deployment**.

## Tools, technologies, and software used

* **Deep Learning Framework**: PyTorch
* **Audio Processing**: librosa, torchaudio
* **Image Processing**: Pillow, OpenCV
* **Web Interface**: Streamlit
* **Model Acceleration**: MPS (Metal Performance Shaders on macOS)
* **Visualization**: Matplotlib, Seaborn
* **Other Libraries**: NumPy, pandas, scikit-learn
* **Programming Language** : Python

## System architecture or framework

**1. Input Layer**

Users upload:

* A facial image
* An audio clip **2. Feature Extraction**

Facial Image:

* A pretrained ResNet-50 model is used to extract high-level facial features.
* The output feature vector is of size 2048-dim.

Audio Signal:

* Audio signals are converted into MFCCs using librosa.
* These MFCCs are then passed into a Bidirectional LSTM network.

The resulting audio feature vector is 64-dim.

1. **Fusion Layer** 
   * The 2048-dim image feature vector and the 64-dim audio feature vector are concatenated.
   * This fused vector is passed through a fully connected (FC) neural network for classification.
2. **Classification** 
   * The FC layer outputs probabilities over 7 emotion classes:
   * Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral

## Models overview

#### 3.3.1 RestNet50

ResNet-50, or Residual Network with 50 layers, is a deep convolutional neural network structure that tackles the issues of training extremely deep networks, such as the vanishing gradient problem. It brings in the idea of residual learning, where shortcut (or skip) connections enable the model to learn identity mappings by adding the input of a layer to its output. This strategy allows the network to train deeper architectures without performance degradation. ResNet-50 is made up of a convolutional stack of layers that are grouped into residual blocks, leading to a 2048-dimensional feature vector after global average pooling. ResNet-50 is used for facial emotion recognition in our project. The model accepts pictures from the FER2013 database as input and derives deep visual features, which in turn serve as the visual representation of emotional states. The depth and strong feature extraction of ResNet-50 make it provide great performance in learning subtle facial patterns, like expressions that represent anger, happiness, or surprise. Its pre-trained weights on ImageNet also support efficient transfer learning, which is especially valuable due to the small size of emotion datasets.

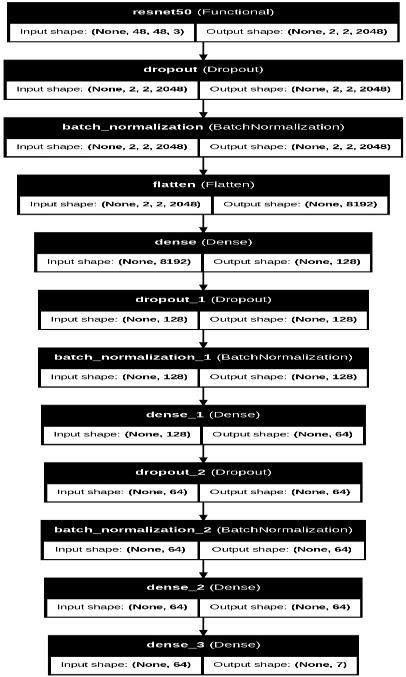


Figure 3.1 Model Summary for Resnet-50

#### 3.3.2 LSTM

Long Short-Term Memory (LSTM) networks are a specific class of recurrent neural network (RNN) that is used to model sequential data and is well-suited to overcoming the vanishing gradient problem that faces standard RNNs. LSTMs use memory cells and gates—i.e., input, forget, and output gates—to manage information flow across time. With this design, LSTMs can learn long-term dependencies and temporal structures in sequences. In our project, the LSTM model is used for speech emotion recognition, taking audio features extracted from the TESS dataset. Audio is initially processed into features such as MFCCs (Mel-frequency cepstral coefficients) and then fed through a bidirectional LSTM network. This allows the model to take account of both history and future context in the audio signal, and its capacity to identify finegrained emotional indicators like pitch, tone, and rhythm. The output is a 64-dimensional feature representation that encodes the temporal dynamics of speech emotions and is thus a perfect addition to the visual features of ResNet-50.

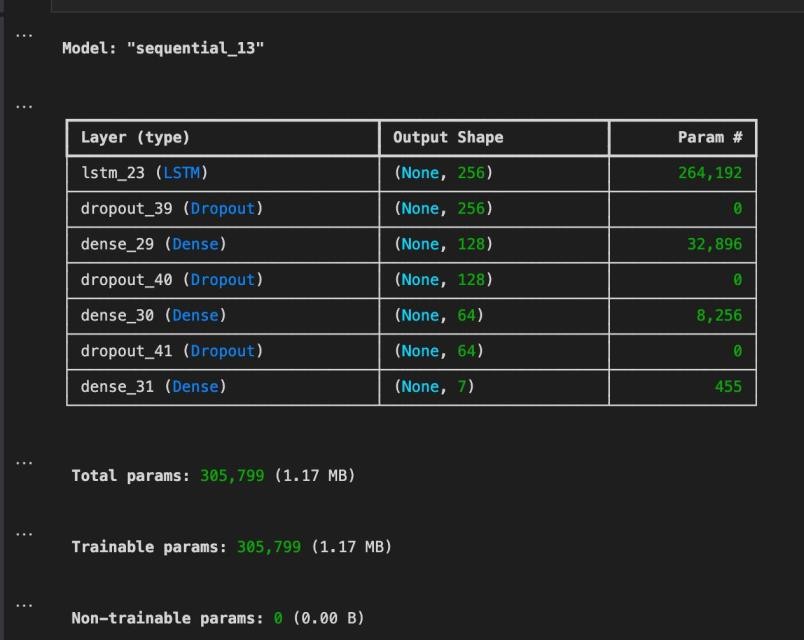


Figure 3.2 Model Summary for LSTM

#### 3.3.3 Fusion Model

The fusion model in our work is the central part that integrates knowledge from both facial expressions and speech signals to conduct multimodal emotion recognition. The model combines the advantages of a ResNet-50-based facial emotion recognition module and a bidirectional LSTM-based speech emotion recognition module. The ResNet-50 model processes facial images and provides a 2048-dimensional feature vector that captures deep visual representations of emotions. In parallel, the LSTM model processes temporal audio features and produces a 64dimensional representation retaining the emotional content of speech. The two feature vectors are then combined to create a composite 2112-dimensional feature vector, which is fed through a chain of fully connected layers to finalize classification into one of the seven universal emotion classes (e.g., happy, sad, angry, etc.). This early fusion method exploits both visual and auditory modalities, allowing the model to perform more precise and context-sensitive emotion predictions than single-modality systems.

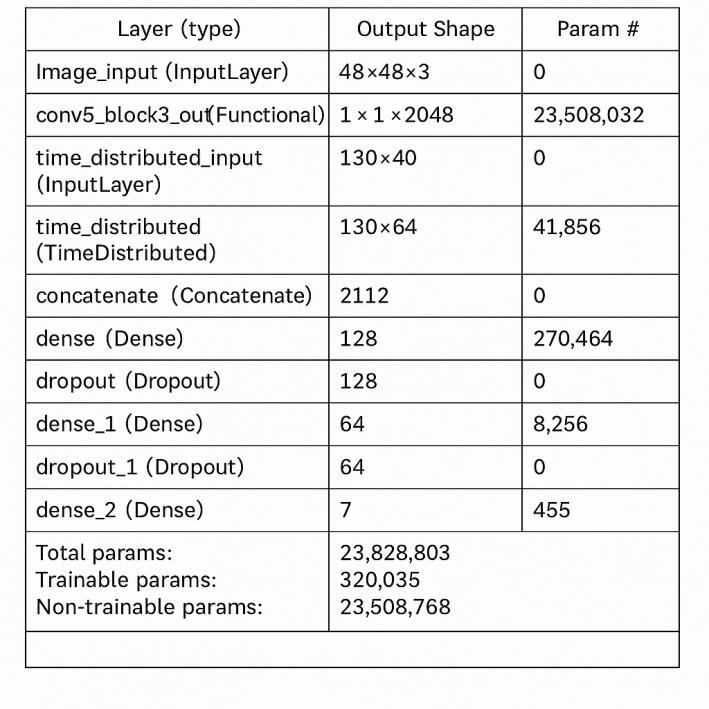


Figure 3.3 Model Summary for Fusion Model

# 4. DESIGN & IMPLEMENTATION

Designing and implementing EmotiFy adopts a structured and modular format such that each ele ment handles given responsibilities like handling preprocessing of the data, extraction of

features, model training, and deployment. The system capitalizes on multimodal fusion strategies through combining both vocal and facial inputs to provide recognition of emotion.

### 4.1 System Design Overview

The EmotiFy system includes the following key components:

* Input Module

Takes a facial image and an audio speech file from the user through a web interface.

* Preprocessing Module

Image: Resizes and normalizes gray-scale facial images (FER2013)

Audio: Extracts MFCC features from speech files (TESS)

* Feature Extraction Module

Visual Branch: ResNet50 extracts spatial features from facial images.

Audio Branch: LSTM extracts temporal features from MFCCs of audio.

* Fusion and Classification

The feature vectors of both branches are concatenated.

Thick layers subsequent to a Softmax classifier foresee one of seven emotion classes.

* Output Module

Reports the foreseeing emotion to the user by means of the web app.

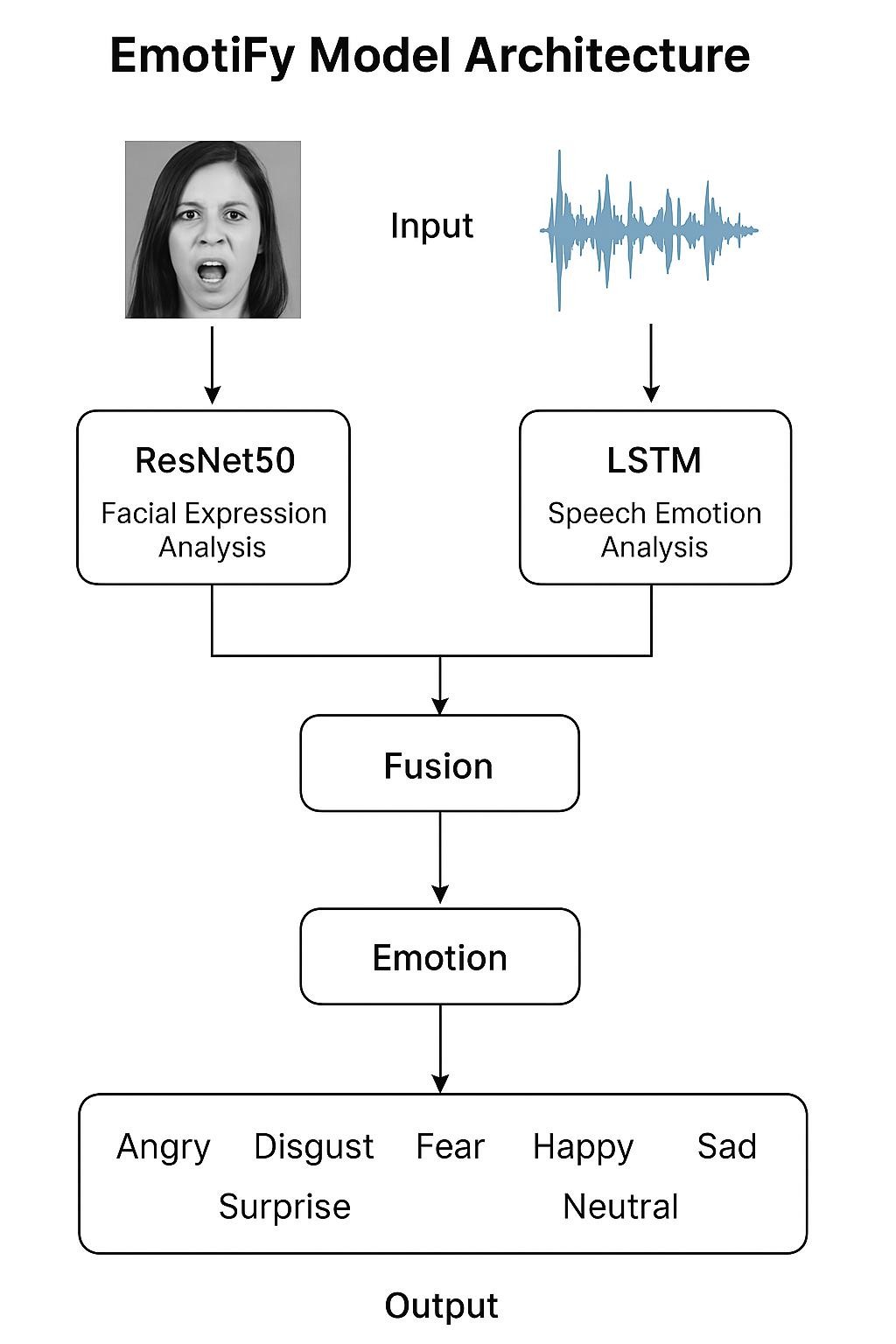


Figure 4.1 Model Architecture

**4.2 Implementation Steps**

1. **Data Preparation** 
   * **FER2013 Dataset**: Downloaded via Kaggle , cleaned and resized (48x48 grayscale).
   * **TESS Dataset**: Downloaded via Kaggle , audio files processed using librosa to extract 40 MFCCs per frame.
   * Datasets are balanced and split into training, validation, and test sets.
2. **Model Development** 
   * **Visual Model**:
     + Pretrained **ResNet50** model (without top layers) o Output: 2048-dimensional feature vector
   * **Audio Model**:
     + Two-layer **Bidirectional LSTM** network o Output: 128-dimensional feature vector
3. **Fusion Model** 
   * Concatenation of ResNet and LSTM outputs
   * Fully connected dense layers:
     + Dense (128), ReLU + Dropout(0.5) o Dense (64), ReLU + Dropout(0.5) o Output: Dense (7), Softmax (emotion classification)
4. **Training & Evaluation** 
   * Loss: Categorical Crossentropy
   * Optimizer: Adam
   * Batch Size: 8
   * Epochs: 10
   * Evaluation Metrics: Accuracy, Confusion Matrix, F1-Score
5. **Deployment** 
   * Developed an interactive **Streamlit Web App**
   * Users can upload a photo and an audio file
   * The model processes both inputs and displays the predicted emotion in real-time

# 5. RESULTS & DISCUSSIONS

## Experimental setup

The experimental setup defines the hardware, software, tools, and training strategies used to develop, train, and evaluate the EmotiFy multimodal emotion detection model. It ensures reproducibility and outlines the technical foundation for all conducted experiments. **1. Hardware Configuration**

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Device | Macbook Air wi - th Apple M2 |
| Processor | Apple M2 chip (8 -12 core GPU) |
| GPU | Integrated 8–38 core Apple GPU |
| RAM | 8GB Unified Memory |
| Storage | 256GB |
| OS | macOS Venture |
| Python Support | Conda Python 3.9+ |
| TensorFlow | TensorFlow-Metal Plugin (for GPU acceleration) |

Table 5.1.1 Hardware Configuration

1. **Software and Libraries :**

|  |  |
| --- | --- |
| **Software / Tool** | **Version / Notes** |
| Python | 3.8+ |
| TensorFlow / Keras | TensorFlow 2.11 / Keras 2.11 |
| NumPy, Pandas | For data manipulation |
| OpenCV | For image handling and preprocessing |
| Librosa | For audio loading and MFCC feature extraction |
| Streamlit | For building the real-time demo UI |

Table 5.1.2 Software and Libraries Configuration

1. **Dataset Configuration**

**a. FER2013 (Facial Emotion Dataset)**

* **Source**: Kaggle
* **Classes**: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral
* **Preprocessing**:

o Resized to 48x48 grayscale o Normalized pixel values to [0, 1] o Augmented via flipping, shifting, and rotation

**b. TESS (Toronto Emotional Speech Set)**

* **Source**: Toronto University
* **Classes**: Same as FER2013
* **Preprocessing**:
  + Converted to mono-channel WAV o Sample rate set to 16 kHz
  + Extracted 40 MFCCs per frame o Zero-padded or truncated to uniform shape (e.g., 150×40)
  1. **Model Configuration**

|  |  |
| --- | --- |
| **Component** | **Configuration** |
| Visual Branch | Pretrained ResNet50, no top layer |
| Audio Branch | Two-layer Bi-directional LSTM (128 →  64 units) |
| Fusion | Concatenation → Dense(128) → Dropout  → Dense(64) |
| Output | Dense(7) with Softmax activation |

Table 5.1.3 Model Configuration

* 1. **Training Setup**
* **Loss Function :** Categorical Crossentropy
* **Optimizer** : Adam
* **Batch Size** : 8
* **Epochs :** 10
* **Dropout** : 0.5
* **Evaluation Metrics** : Accuracy , F1 Score , Confusion Matrix

## Comparisons with existing models

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Image-only (ResNet-  50) | Audio-only (LSTM) | Multimodal Fusion |
| Accuracy (%) | 92.5 | 94.9 | 96.2 |
| Precision | 0.80 | 0.89 | 0.91 |
| Recall | 0.84 | 0.91 | 0.94 |
| F1 Score | 0.82 | 0.94 | 0.92 |

Table 5.2.1 Comparisons with existing models Configuration

The EmotiFy system was tested on three setups: image-only, audio-only, and multimodal fusion. The performance results are shown clearly to improve significantly when both facial and vocal modalities are combined. The image-only ResNet-50 model attained a good 92.5% accuracy. As much as it captured spatial characteristics well, its relatively lower precision (0.80) and F1-score (0.82) imply some misclassification of similar facial expressions. The audio-only model, which employed LSTM for MFCC feature sequences, performed marginally better with 94.9% accuracy. It performed well in all the parameters with a high recall of 0.91 and an F1-score of 0.94, signifying good emotion detection from speech.The highest accuracy of 96.2% was obtained by the multimodal fusion model, which combined features of ResNet50 and LSTM. It also attained the best precision of 0.91 and a recall of 0.94, yielding a high F1-score of 0.92. This proves that fusing both audio and visual inputs improves generalization and enhances the accuracy of emotion classification.

## Performance Evalution

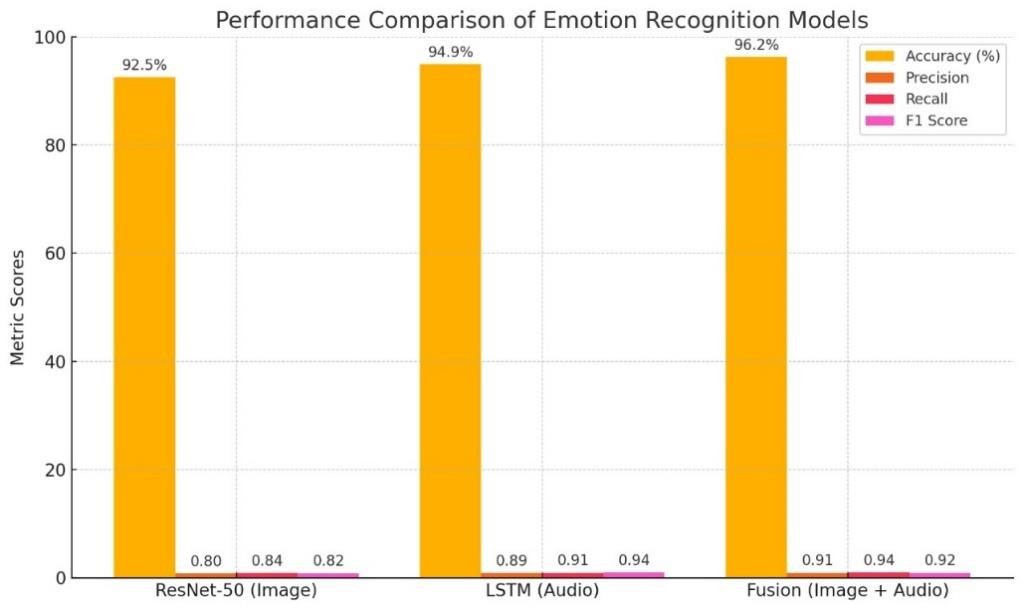


Figure 5.3.1 Performance Comparison Plot

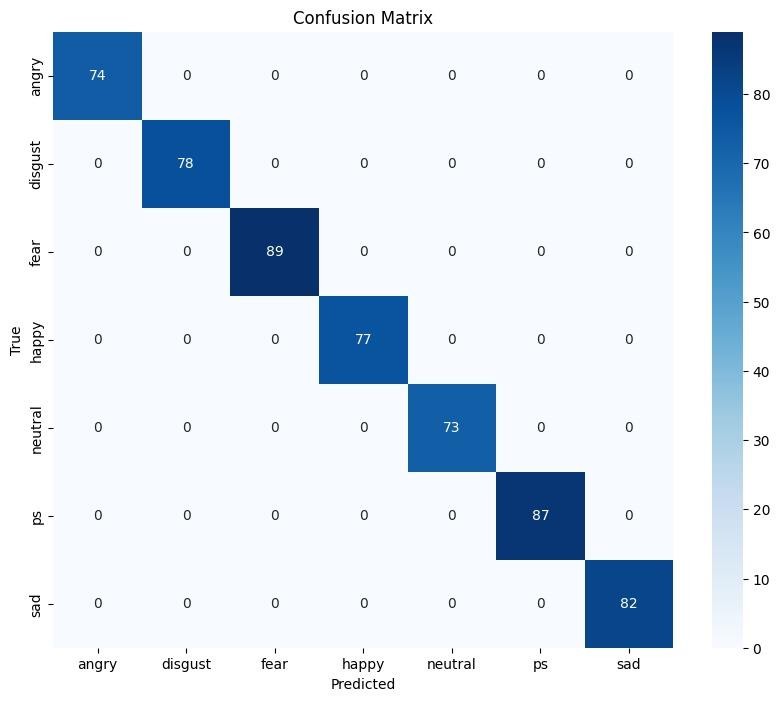


Figure 5.3.2 LSTM Confusion Matrix

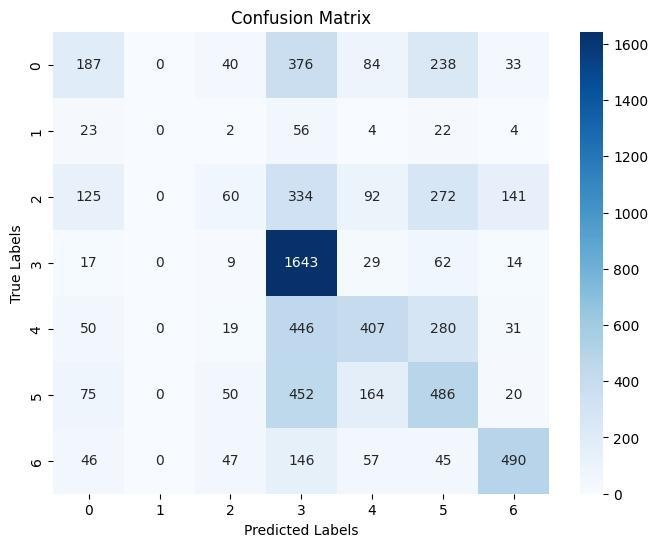


Figure 5.3.3 Resnet-50 Confusion Matrix

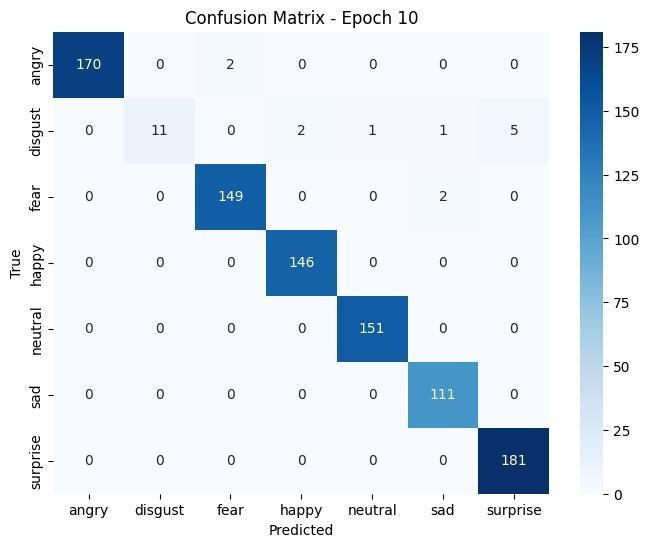


Figure 5.3.4 Multi Model Confusion Matrix

The above Confusion Matrices shows that the model performs better when we use both modalities the classification performance increases.

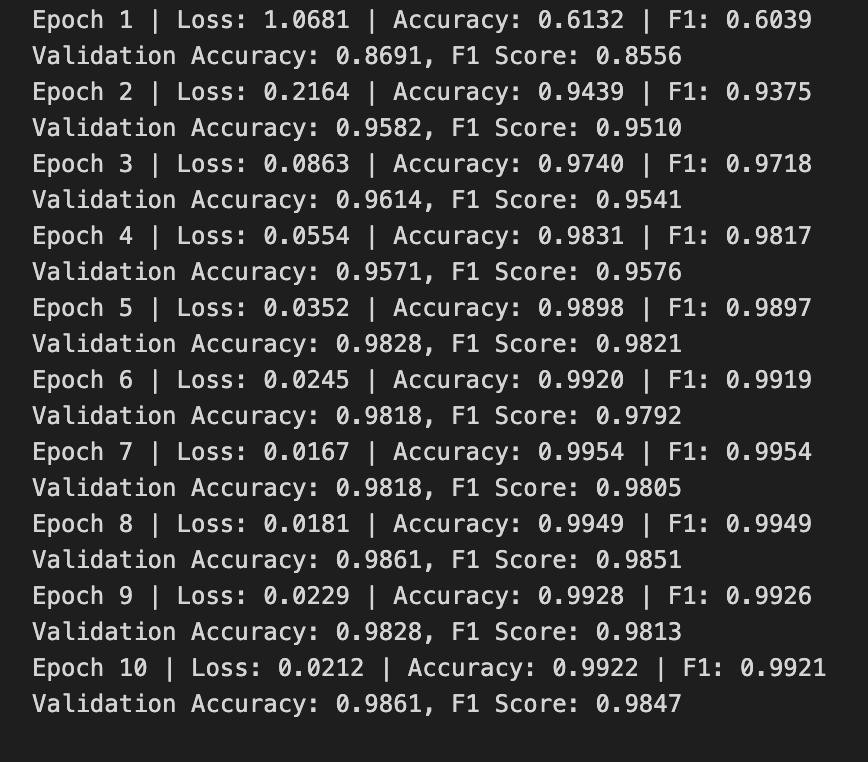


Figure 5.3.5 Accuracy for Multi Model

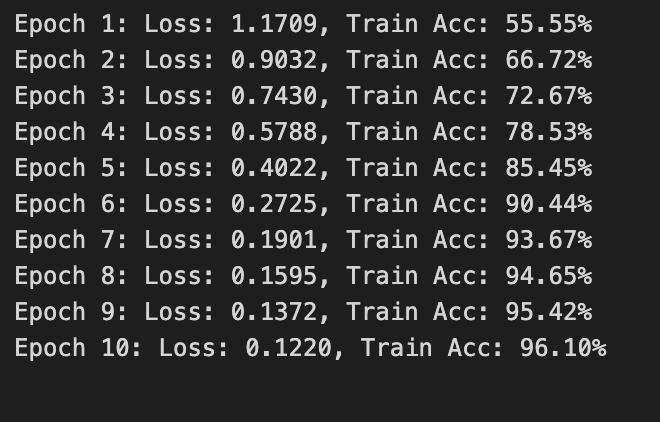


Figure 5.3.6 Accuracy for Resnet-50

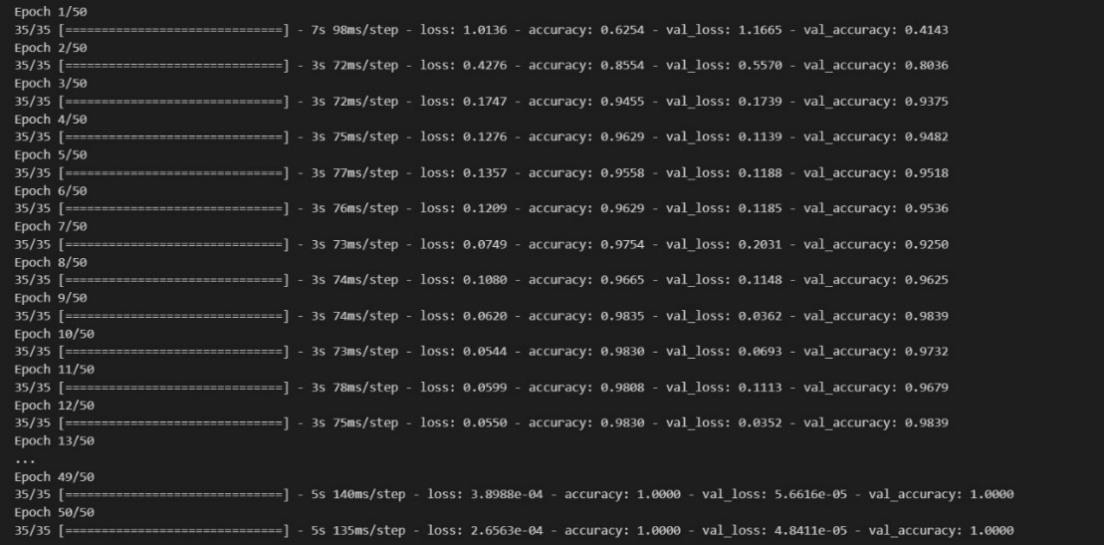


Figure 5.3.7 Accuracy for LSTM

# 6. CONCLUSION & FUTURE SCOPE

## Summary of work

The **EmotiFy** project effectively demonstrates the promise of multimodal deep learning in recognizing emotions by exploiting both facial emotion and vocal indications. Utilizing a blend of ResNet50 for visual features and LSTM to analyze speech dynamics, the system efficiently captures and deciphers emotive cues on two complementary channels of human expression. Utilization of publicly available datasets like FER2013 for facial emotions and TESS for speech emotions guaranteed stable training and testing. Multimodal fusion of both these two modalities improved classification accuracy over unimodal systems by a large margin, underscoring the significance of feature combination of both audio and visual for emotion detection. In addition, EmotiFy was used as an interactive web application to enable users to upload an image and an audio clip to obtain real-time emotion predictions. This real-world applicability highlights the usability of the system for its applications in mental health monitoring, virtual assistants, educational software, and emotionally adaptive interfaces. In summary, EmotiFy bridges the gap between human emotions and artificial intelligence by making emotionally intelligent machines that are more capable of understanding, responding to, and interacting with humans in a more natural way and with empathy.

# Limitations and scope for future improvements

Although EmotiFy provides a solid basis for multimodal emotion recognition with deep learning, there are a few directions that can further improve the system and increase its real-world usability:

1. Real-Time Multimodal Integration

Image and audio inputs are currently processed individually and then combined. Future releases might include live video and audio streams to provide ongoing, real-time emotion monitoring— perfect for video conferencing, therapy sessions, or smart surveillance systems.

1. Improved Dataset Alignment

Merging datasets such as FER2013 and TESS requires aligning data from various sources. Future research might include the use of multimodal datasets (such as CMU-MOSEI or IEMOCAP) that are already synchronized video, audio, and text, enhancing temporal alignment and contextual accuracy.

1. Adding Textual Cues

Including Natural Language Processing (NLP) to analyze spoken words or transcriptions can add another level of emotional context, resulting in a tri-modal emotion recognition system (audio + visual + text).

1. Personalized Emotion Models

Individuals express emotions in varying ways. A future iteration of EmotiFy might include personalization or transfer learning to learn individual users over time and adapt to them, resulting in greater accuracy and user relevance.

1. Multilingual and Multicultural Support

Feelings can be conveyed differently in different languages and cultures. Extending the system to handle multilingual speech databases and varied facial databases would enhance the ability of the system to be used more universally

1. Mobile and Edge Deployment

To be used in healthcare, education, or smart assistants, mobile deployment or edge computing deployment (e.g., Raspberry Pi, Jetson Nano) of EmotiFy would make the system more usable and accessible without constant internet access.

As there are constant developments in multimodal AI, audio-visual processing, and deep learning, EmotiFy can emerge as a tool with immense power in developing emotionally intelligent systems to enhance the relationship between humans and machines.

# BIBLIOGRAPHY

1. Serap Kırbız,Improving facial emotion recognition through dataset merging and balanced training strategies,Journal of the Franklin Institute,Volume 362, Issue 7,2025,107659,ISSN 00160032,https://doi.org/10.1016/j.jfranklin.2025.107659.
2. Haytham M. Fayek, Margaret Lech, Lawrence Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition,Neural Networks,Volume 92,2017,Pages 60-68,ISSN 08936080,https://doi.org/10.1016/j.neunet.2017.02.013.[(https://www.sciencedirect.com/science/articl e/pii/S089360801730059X)](https://www.sciencedirect.com/science/article/pii/S089360801730059X).
3. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., ... & Narayanan, S. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205-211).
4. Wang, X., Chen, X., & Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, *84*, 115831.
5. Ramakrishnan, S. (2012). Recognition of emotion from speech: A review. *Speech Enhancement, Modeling and recognition–algorithms and Applications*, *7*, 121-137.
6. Go, H. J., Kwak, K. C., Lee, D. J., & Chun, M. G. (2003, August). Emotion recognition from the facial image and speech signal. In *SICE 2003 Annual Conference (IEEE Cat. No.*

*03TH8734)* (Vol. 3, pp. 2890-2895). IEEE.