EmoSense: Human Emotion Detection Using Audio and Video Inputs

A major project report submitted in partial fulfillment of the requirement for the award of degree of

Bachelor of Technology

in

Computer Science & Engineering / Information Technology

Submitted by

Rakshita Jain (201462)

Shriya (201272)

Vidur Sharma (201467)

Under the guidance & supervision of

Dr. Kushal Kanwar

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, Solan - 173234 (India)

Certificate

This is to certify that the work which is being presented in the project report titled **"EmoSense: Human Emotion Detection Using Audio and Video Inputs"** in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by "Rakshita Jain (201462), Shriya (201272) and Vidur Sharma (201467)**"** during the period from August 2023 to December 2023 under the supervision of **Dr. Kushal Kanwar**, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Rakshita Jain
(201462)

Shriya
(201272)

Vidur Sharma
(201467)

The above statement made is correct to the best of my knowledge.

Dr. Kushal Kanwar
Assistant Professor
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat

Candidate's Declaration

I hereby declare that the work presented in this report entitled **'EmoSense: Human Emotion Detection using audio and video inputs'** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering / Information Technology** submitted in the Department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2023 to May 2024 under the supervision of **Dr. Kushal Kanwar** (Assistant Professor, Department of Computer Science & Engineering and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Student Name: Rakshita Jain
Roll No.: 201462

Student Name: Shriya
Roll No.: 201272

Student Name: Vidur Sharma
Roll No.: 201467

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature with Date)

Dr. Kushal Kanwar

Assistant Professor

Computer Science & Engineering and Information Technology

Acknowledgement

First and foremost, we sincerely appreciate and are appreciative to the Almighty God, whose divine favor enabled us to successfully finish the project work. Supervisor Dr. Kushal Kanwar, Assistant Professor, CSE Department, Jaypee University of Information Technology, Waknaghat, has our sincere gratitude and debt. Our supervisor's extensive knowledge and genuine interest in the topic of "machine/deep learning" enabled us to complete this assignment. This endeavor has been made possible by his unending patience, scholarly direction, constant encouragement, frequent and energetic supervision, constructive criticism, insightful counsel, reading numerous subpar versions and fixing them at every stage. We would like to thank Dr. Kushal Kanwar of the CSE Department for his kind assistance in completing our project. We would also want to extend a warm welcome to everyone who has assisted us, directly or indirectly, in achieving success with this initiative. In light of this particular circumstance, we may wish to express our gratitude to the several staff members—both instructional and non-instructional—who have provided their helpful assistance and enabled our project. Lastly, we must respectfully appreciate our parents' unwavering support.

Rakshita Jain
(201462)
Shriya
(201272)
Vidur Sharma
(201467)

Table of Contents

List of Figures

List of Abbreviations

1. AWS: Amazon Web Services
2. GCP: Google Cloud Platform
3. HTML: Hypertext Markup Language
4. CSS: Cascading Style Sheets
5. UI: User Interface
6. URL: Uniform Resource Locator
7. ML: Machine learning
8. AI: Artificial Intelligence
9. CNN: Convolutional Neural Network
10. RNN: Recurrent Neural Network

List of Graphs

List of Tables

Abstract

Human interactive communication is an essential capability that entails identifying and comprehending human emotion. Technological advancements have attracted increased research on the generation of such systems that can interpret emotions from sound or visual cues. This is a report describing a full investigation and integration for an emotional human detection system involving both audiovisual sources. Signal Processing and Computer Vision are used to provide algorithms of analysis for auditory and visual stimuli associated with certain emotional state. Using feature extraction and employment of various classification models, the system is able to pinpoint different emotional states such as happiness, sadness, anger, shock, fear, and neutral. It also presents the pre-processed audio and video data, feature extraction methods, as well as the employed models for emotional classification Evaluation metrics and results from the system's performance in emotion recognition and classification are presented and discussed. In addition, the report delves into the challenges we face during the development process, including data collection, feature engineering, and model optimization. Ethical considerations regarding emotion detection technology, privacy concerns, and potential biases are also addressed. The significance of the project lies in its potential applications in various fields, including mental health monitoring, human-computer interaction, and personalized user experiences. In addition, it emphasizes the interdisciplinary nature of the field, bridging computer science, psychology, and signal processing. This report not only presents a technical exploration of human emotion detection using audio and video inputs, but also highlights the wider societal implications and ethical aspects of deploying such technology.

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

The intricate realm of human emotions has long fascinated scientists, psychologists and technologists. This understanding is the basis for effective human communication and relating with other people through interpretation of such complex emotional states. The advancement of complex technologies including the amalgamation of machine learning, signal processing, and computer vision has led to an increased demand for designing automated intelligence systems that can discern the emotions of humans from multi-modal inputs that include both video and audio information.

This paper provides a detailed review of the design and implementation of a software for identifying human emotions based on audio and visual data. The evolution of emotion recognition technology has led to a wide range of amazing applications in different areas. Using machine learning algorithms and signal processing methods, this project aims to fill a gap that exists between people's manifestation of emotions and computation.
A discussion cannot be avoided on the significance of this emotion detection technology as it can transform many sectors globally. For instance, there is mental health where such systems will aid in early detection and tracking emotional status towards specific treatment.

In human-computer interaction, understanding user emotions can enable systems to adapt and respond more empathetically, improving user experiences. The development of an emotion detection system using audio and video inputs is multifaceted. It involves pre-processing raw data, extracting relevant features that encapsulate emotional cues, and using robust classification models capable of recognizing and categorizing different emotional states. The complexity associated with processing multimodal data presents challenges, including feature selection, model optimization, and ensuring system adaptability to different contexts and individuals.
An interesting aspect of this project is its interdisciplinary nature, drawing from fields such as computer science, signal processing, psychology and neuroscience. Integrating knowledge from

these diverse fields contributes to a more holistic understanding of human emotions and their computational representations.

However, along with technological advances, the deployment of emotion detection systems raises ethical considerations.The important issues include privacy concerns, consent, potential biases, as well as ethical considerations with regard to application of this technology.

The goal of this report is to completely explain the processes taken in producing an emotional sensing system that uses both audiovisual inputs. It explores in detail the specific technical aspects involved in data preprocessing, feature extraction, models' architectures, and evaluation measures. This is an in-depth account that traces the building up of an emotional detection system for audiovisual input as it attempts to shed light on its technical challenges, probable usage cases, and morality issues.

1.2 PROBLEM STATEMENT

A challenging problem arises from the intrinsic subtleties and multimodality of human emotional expression when attempting to understand and interpret emotions from audio and video inputs. In order to precisely recognize and categorize human emotions, the main problem is developing a system that can quickly collect and analyse these varied inputs.

- **COMPLEXITY OF EMOTIONS:**
  Facial expressions, tone of voice, gestures, and body language are just a few examples of the many ways that human emotions can be expressed. Gaining insight into the subtleties and intricacies present in both audio and visual inputs is necessary to decode these emotions. Micro-expressions, which in a matter of milliseconds carry important emotional clues, are a subset of facial expressions. Understanding that expressions on the face are read exactly as they are meant to be understood is crucial.

- **DATA VARIABILITY:**

  Systems for detecting emotions utilize a wide range of inconsistent datasets. Various facial features, skin tones, lighting settings, and gestures that can convey emotions in diverse ways between cultures and people are all included in the video inputs. It is quite difficult to collect and standardize these disparate data sources in order to develop a universal model that can recognize emotions in many contexts with accuracy.

- **AMBIGUITY AND SUBJECTIVITY:**

  Emotions are interpreted ambiguously since they are context-dependent and subjective. For instance, a face that in one culture can convey happiness might convey a different feeling in another. Furthermore, classification issues arise with emotions like sarcasm or delicate emotional shifts. It takes advanced algorithms that can recognize contextual signals and minute changes in audio and video inputs to decipher these subtleties of emotional states.

- **INTERDISCIPLINARY NATURE:**

  A variety of disciplines, including computer vision, signal processing, machine learning, and psychology, are needed to develop an efficient emotion recognition system. Collaboration between several domains is necessary to comprehend the psychological elements of emotions, their neurological foundation, and convert this knowledge into computer models. A challenging but crucial part of this issue is combining expertise from several domains to develop a thorough model that can recognize emotions with precision.

- **ETHICAL CONSIDERATIONS:**

  Ethical issues related to facial or emotion detection using audio and video inputs pose a major challenge. Robust privacy safeguards and informed consent are necessary when collecting sensitive audio and video data for purposes of emotion analysis. Moreover, to produce objective and impartial emotion identification tools, a lot more attention should be paid towards excluding prejudices based on gender, race, or culture. These challenges pose a major constraint to the construction of dependable ethical systems

for emotion recognition based on both visual or auditory cues reflecting the difficulty of the issue. It is essential for technology to be able to address such challenges to further understand and interpret human emotions.

1.3 OBJECTIVES

- **SYSTEM DEVELOPMENT:**

  Create and design a complex system that efficiently processes audio and video inputs in real time. This includes developing algorithms and frameworks capable of handling multimodal data inputs. The key is to design a system that can simultaneously process visual features (such as facial expressions, gestures) and extract emotional cues. Real-time processing capabilities and system robustness will be the deciding factors.

- **FUNCTION:**

  Identify, extract and process relevant features from audio and video data that accurately represent different emotional expressions. It includes selection and engineering features that encapsulate emotional cues from audio and video inputs. For audio, features can include spectrogram representations, intonation patterns, and speech rhythm analysis. For video, features can include facial action units, gaze patterns, and motion dynamics. The extraction of these features is essential for the subsequent classification of emotions.

- **MODEL TRAINING AND EVALUATION:**

  Develop machine learning models capable of accurately classifying emotions based on extracted audio and video features. This involves building and fine-tuning machine learning models (such as deep neural networks, ensemble methods) capable of recognizing and categorizing a wide range of emotional states. The models will be trained using labeled datasets and extensively evaluated using metrics such as accuracy, precision, recall and F1-score to ensure their efficiency and robustness.

- **IMPLEMENTATION OF THE ETHICAL FRAMEWORK:**
  Integrate ethical guidelines throughout the development process to address privacy, consent, and potential biases in data collection and model predictions. This goal includes setting ethical protocols for data collection, ensuring consent, and privacy measures for participants contributing to the dataset. Mitigating bias in the collected data and ensuring the fairness and transparency of model predictions are critical considerations. The goal is to deploy an emotion detection system that respects ethical standards and functions responsibly.

- **SCALABILITY AND ADAPTABILITY:**
  Develop a system that can adapt to different cultural contexts, individual differences and scenarios and ensure scalability for different applications. The system should be designed to generalize well across different demographics, cultures, and environments. It should be adaptable to different scenarios and individuals and ensure its applicability in real-world situations outside of a controlled laboratory setting.

1.4 SIGNIFICANCE AND MOTIVATION OF THE PROJECT WORK

- Improvements in Mental Health: Emotions play an important role in mental health. This study can contribute to psychological research and intervention by developing video and audio ideas that can detect and understand people's emotions from the very beginning. Early detection of thought patterns can help quickly identify and resolve mental health problems.
- Human-Computer Interaction: Pursuing research ideas in technology can enable increasingly personal interactions between humans and machines. Devices and systems can create a more intuitive user experience by adjusting their responses based on the user's mood.
- Improve communication: Thinking is an important part of communication. Emotion detection systems can help determine the meaning of a conversation, thus improving communication and reducing misunderstandings, especially in areas such as customer service or human interaction.

- Education and Learning: Emotional intelligence can be used to measure student engagement and mood during learning in educational settings. It helps teachers adapt instruction to meet students' emotional needs, thus enhancing learning.
- Medical Applications: In a medical setting, such tools can help doctors identify emotional receptors. It can also be used to intervene in conditions where emotions are difficult to understand, such as autism spectrum disorder.
- Health and Behavior Research: Behavior Research can help researchers understand human behavior in a variety of social environments. It can provide valuable insight into how emotions affect decision making, group dynamics, and social relationships.

  The motivation behind the project is that it can transform fields such as health and technology into education and social sciences. Creating systems that can detect and interpret human emotions through video and audio feedback can ultimately lead to more intuitive, responsive, and emotional technologies that improve all aspects of human life.

1.5 ORGANIZATION OF PROJECT REPORT

The remaining report is organized as follows:

- Chapter 2 Literature Survey: Research on human perception using video and audio input offers facial recognition by CNNs, RNNs to analyze speech features, and a variety of methods that combine both. Benchmark datasets such as CK+ and Emo React driver algorithm analysis support real-world applications in healthcare and human-computer interaction. However, challenges remain, requiring solutions to cultural differences, security standards, and data privacy. To achieve this growth in the future, we will focus on new models, advanced interpretations and comprehensive data.

- Chapter 3 System Development: In this work, which involves human participants, highly trained machine learning models support audio and video. Accordingly, facial characteristics, speech features, and body features are derived from the pre-processed data. These ideas go through a number of algorithms that are based on techniques like

CNNs, RNNs, and hybrid classification models to determine height. Subsequently, the system undergoes extensive testing, validation, and optimization in a variety of applications and environmental situations.

- Chapter 4 Testing: Testing phase three involves confirming the correctness of identifying the corresponding needs as they are represented in the input data (sound and video). The model is tested on several groups under varied settings in order to obtain both generalizability and reliability. In addition, stress testing involves simulating challenging conditions to see how effective it might be in real-world situations. Finally, when feedback raises the model's degree of accuracy and adaptability, improvements will be made based on it.

- Chapter 5 Results and Evaluation: The outcomes accurately identified the emotions from a variety of input audio and video files. The system's ability to recognize multiple assumptions is demonstrated by precision, recall, and F1 score. Nonetheless, it is important to recognize that in order for continuous improvement to be accepted, it must handle issues like cultural differences and ethics when handling sensitive content.

- Chapter 6 Conclusions and Future Scope: The project on human emotion detection using video and audio inputs presents promising avenues for technological advancements in understanding human emotions. Future endeavours involve refining algorithms to enhance accuracy, considering cultural nuances, and expanding datasets for diverse emotional expressions. Integration into various domains like healthcare, education, and human-computer interaction offers potential for transformative applications, urging continued research and ethical considerations for widespread adoption.

CHAPTER 2: LITERATURE SURVEY

2.1 OVERVIEW OF RELEVANT LITERATURE

The literature on human emotion detection through audio and video inputs is extensive and diverse, and includes studies on various machine learning models, feature extraction techniques, annotated datasets, cross-modal fusion methodologies, cultural influences on emotional expression, ethical considerations, and the real world. applications. Research explores machine learning and deep learning models adapted for emotion recognition, feature extraction methods for audio and video inputs, datasets annotated with emotion labels, and strategies for efficiently combining information from both modalities. In addition, the study delves into cultural differences in emotional expression, highlighting the importance of context while addressing ethical concerns related to privacy, consent, and bias in emotion detection systems. Real-world applications in mental health monitoring, human-computer interaction, education, and marketing underscore the potential impact and relevance of emotion detection technology in various fields.

2.1.1 The paper [1] suggests using multimodal thinking to improve emotion recognition accuracy and applies the appropriate data preprocessing to the chosen data set. Appropriate models have been developed for both audio and video modalities: this paper uses a "time-distributed CNN + LSTM" model construction scheme for the audio-modality emotion recognition task, and a "DeepID V3 + Xception architecture" design scheme for the video-modality emotion recognition task. Every model creation technique is also verified experimentally and compared to current emotion identification algorithms. In conclusion, this work attempts late fusion and suggests and executes a late fusion technique predicated on the idea of weight matching. The recognition accuracy increased by about 4% to 84.33% as compared to the single-modal emotion recognition method.

2.1.2 Next-generation artificial intelligence is expected to interact with humans more closely and will have a significant attribute of being able to record and reflect human emotions. While emotion extraction from written representations of human conversations has yielded encouraging results, the accuracy of emotion recognition based on acoustic cues from audio remains low.[2] suggests a novel method for extracting features from conversational audio data

by using feature embedding based on Bag-of-Audio-Words (BoAW). A cutting-edge recurrent neural network (RNN)-based emotion detection model is put forth that makes categorical predictions of emotions in real time while also capturing the conversational context and the emotional states of each participant.

2.1.3 Speech sentiment categorization in the face of background noise using a novel algorithm[3]. Conventional models depend on pre-built audio feature extractors for users who struggle to blend their accents naturally. This work employs the vector space of emotional notions, wherein words with comparable meanings frequently share a prefix. When it comes to audio categorization, extensions are a common method of augmenting training data. Certain extensions, though, can result in a loss of precision. Thus, a new eigenvalue-based metric was developed to choose the optimal extensions. When the suggested method of handling emotion was applied to YouTube videos, it outperformed baselines by ten to twenty percent. Words with comparable pronunciations and emotions are learned by each neuron. Additionally, we apply the algorithm to identify bird presence from city sound recordings.

2.1.4 Since the input data sources are highly variable in different combinations of modalities, the use of multiple modalities often requires ad hoc fusion networks. To predict the emotional arousal of a person responding to a given video clip [4] ViPER, a multimodal architecture using a transformer-based modality-agnostic combination of video images, audio recordings, and text annotations. Specifically, it relies on a modality-agnostic late fusion network that makes ViPER easily adaptable to different modalities. Experiments performed on the Hume-Reaction datasets of the MuSe-Reaction challenge confirm the effectiveness of the proposed approach.

2.1.5 Emotional problems are common among today's college students. To improve their mental health, it is urgent to quickly identify the negative emotions of college students and guide them to improve their emotional development. Students' emotions are expressed in several ways, such as sound, facial expressions, and gestures. Using the complementarity between multimodal emotional information can improve the accuracy of emotion recognition.[5] proposes a multi-modal emotion recognition method for voice and video images based on deep learning: (1) For voice modal recognition, the voice is first pre-processed to extract voice emotional features, and then based on attention, a long-short-term memory

(LSTM) network) is adopted for emotion recognition; (2) For video image modal recognition, the extended local binary pattern (LBP) operator is used to calculate image features, LBP block weighting and multi-level partitioning are combined to extract features, principal component analysis (PCA) is adopted to reduce the dimensionality of eigenvectors and the VGG-16 network model is constructed using a transfer learning training strategy to realize emotion recognition.

2.1.6 Multiple techniques can be defined through human feelings, including expressions, facial displays, physiological signs, and neuroimaging strategies.[6] presents an overview of multimodal emotional signal recognition using deep learning and compares its applications based on current studies. Multimodal affective computing systems are studied alongside unimodal solutions because they offer higher classification accuracy. Accuracy varies according to the number of observed emotions, extracted features, classification system, and consistency of the database. Numerous theories about emotion detection methodology and current emotion science address the following topics. This would encourage studies to better understand the physiological signals of the current state of science and its problems with emotional awareness.

2.1.7 19,004 video clips in total, separated into two sections based on the data source, are included in HEU Emotion. The first section includes films that were downloaded from Giphy, Tumblr, and Google. The videos cover two modalities (posture and facial expression) and ten moods. A manually curated corpus comprising ten emotions and three modalities—facial expression, body posture, and emotional speech—from motion pictures, television shows, and variety shows is included in the second section. With 9951 participants, HEU Emotion is by far the largest multimodal emotion database. They evaluated HEU Emotion using a variety of traditional machine learning and deep learning techniques in order to set a standard for emotion recognition. They created a multimodal attention module that combines multimodal functionalities in an adaptable way. After multimodal fusion, the recognition accuracy for the two parts increased by 2.19% and 4.01%, respectively, compared to single-modal facial expression recognition.[7]

2.1.8 The study[8] proposes a deep CNN-based facial emotion recognition (FER) system using transfer learning (TL) to improve feature extraction from high-resolution images and address

the meaning of profile display. Through the new pipeline, the dense layers are fine-tuned together with the pre-trained CNN blocks, achieving excellent accuracy on the KDEF and JAFFE datasets. The method using DenseNet-161 achieves remarkable accuracies of 96.51% (KDEF) and 99.52% (JAFFE) through 10-fold cross-validation, which is a significant advance in emotion detection accuracy, especially for profile display, promising practical applicability.

2.1.9 Affective computing is currently one of the most active research topics, which is also receiving more and more intensive attention. This strong interest is driven by a wide range of promising applications in many fields, such as virtual reality, intelligent surveillance, perceptual interfaces, etc. Affective computing refers to a multidisciplinary knowledge background such as psychology, cognitive science, physiology, and computer science. The paper[9] highlights several issues that are implicitly part of the entire interactive feedback loop. For each problem, different methods are discussed in order to explore the state of the art.

2.1.10 [10] proposed a new strategy for gathering features over time that uses local attention to focus on specific regions of the speech signal that are more emotionally salient. The proposed solution is evaluated on the IEMOCAP corpus and provides more accurate predictions compared to existing emotion recognition algorithms.

2.2 KEY GAPS IN THE LITERATURE

- Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video [1] lacked a comprehensive explanation of the weight-adaptive mechanism or strategy used to fuse audio and video data.

- Voice-based Real-time Emotion Detection Technique with RNN-based Feature Modeling [2] had low accuracy of audio modality -based feature representations compared to text modality.

- Speech Emotion Recognition Using Audio Matching[3] had pretrained audio feature extractors that do not generalize well. Some augmentations may result in a loss of accuracy.

- ViPER: Video-based Perceiver for Emotion Recognition [4] raised concerns regarding privacy, consent, and potential misuse. Moreover, it heavily rely on the quality and diversity of the datasets they are trained on.

- Emotion Recognition of College Students Based on Audio and Video Image [5] had limited sample size - limited accuracy of emotion recognition.

- Multimodal Emotion Recognition using Deep Learning [6] does not provide specific numerical results or accuracy rates for the multimodal emotion recognition systems.

- Multimodal Emotion Recognition in the Wild Challenge [7] lacks in-depth analysis of implementation challenges.

- Facial Emotion Recognition from Videos Using Deep CNNs [8] Paper lacks time details; results differ from literature due to dataset variations.

- Affective Computing: A Review of the State of the Art [9] lacks detailed analysis, empirical studies, dataset/tool specifics, and ethical considerations; potential gaps include scope and depth.

- Automatic Speech Emotion Recognition using RNNs with Local Attention [10] Local attention mechanisms may require careful hyperparameter tuning. - Limited discussion on dealing with noisy speech data.

CHAPTER 3: SYSTEM DEVELOPMENT

3.1 REQUIREMENTS AND ANALYSIS

A combination of artificial intelligence, machine learning techniques, and computer vision algorithms has enabled machines to learn how to make decisions and recognize objects, scenes, and scenarios. Even so, machines are still unable to recognize subjective things like emotion. Despite our ability to do this relatively well from an early age, replicating that skill using a computer is still very challenging. At the end of the day, emotions are nothing more than a bunch of chemicals and hormones reacting in our bodies, which is what makes Emotion Recognition so intriguing. Making a machine that can replicate emotions and eventually feel is basically creating a human.

Fig 3.1 Example of emotions extracted from emotion recognition

It refers to the sounds that can be generated using the Librosa library, such as Mel Frequency Cepstrum Coefficients (MFCC), Chroma, and Mel Spectrogram. These features are then fed into a variety of learning algorithms, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN) 1D (for 1D data frames) and CNN 2D (for 2D tensors).

3.1.1 USER INTERACTION REQUIREMENT

1)Intuitive User Interface Design:

Using Flask, you can develop a web application that interacts with your emotion detection model. Make sure that the interface allows users to upload audio and video inputs with clear instructions on the supported formats and file sizes. In order to keep users informed during the processing of inputs, consider adding features such as progress indicators.

2)Real-time Feedback and Interpretability:

Added real-time feedback systems to the Flask online application. Show the identified emotions next to the appropriate audio and visual input portions. Moreover, an overview or illustration of the salient characteristics or trends that impacted the model's forecasts was included. This transparency aids in the users' understanding of the model's decision-making procedure.

3)Cross-Device Compatibility and Accessibility:

Various devices and browsers will be used to test the web application to guarantee a responsive design. Include features that make content accessible, including alt text for photos and appropriate HTML semantics. To detect and resolve any accessibility issues, do usability testing with various users. Give users comprehensive instructions and user manuals to help them navigate and use the emotion-detecting service.

3.1.2 DATA REQUIREMENTS

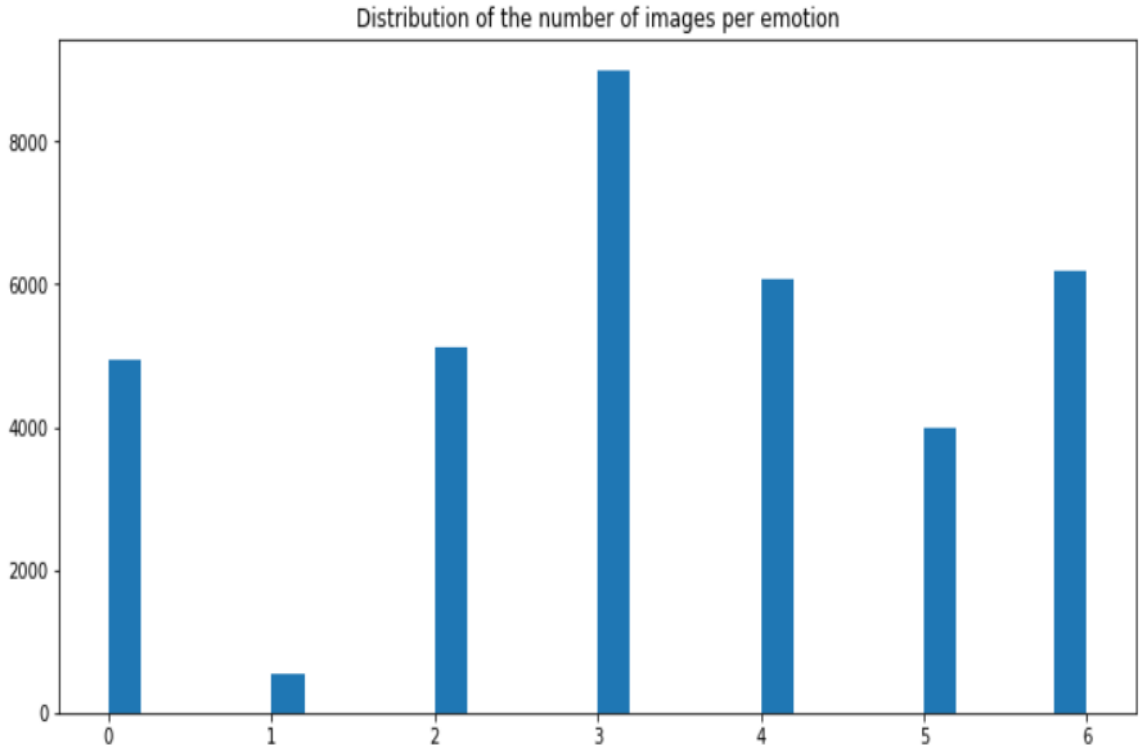
To choose the audio-visual datasets that would be best suited for solving the problems at hand, the project's initial phase involved conducting a thorough investigation of them and gathering sufficient data on them.

1. Video Dataset:

   The FER2013 dataset comprises grayscale images of faces with dimensions 48x48 pixels. These facial images have undergone automated registration, ensuring that the face occupies a consistent amount of space in each image.

   The primary objective of this dataset is facial expression categorization, where each face is labelled with one of seven emotion categories: Angry (0), Disgust (1), Fear (2), Happy (3), Sad (4), Surprise (5), and Neutral (6). The dataset is divided into a training set, consisting of 28,709 examples, and a public test set, comprising 3,589 examples.

Fig.3.2 Video dataset



Graph 1. Distribution of no. images per emotion

1) Audio Dataset:

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset contains a total of 7356 files. A total of 24 professional actors vocalize two statements in a North American accent. The emotions included in the speech set are: neutral, calm, happy, sad, angry, fearful, surprise and disgust. Each one of these expressions is produced at two different emotional intensity levels: normal and strong and can be obtained in three different formats: Audio-only, Audio-Video and Video-only. As the name of the dataset states, there are both speech and song files but for the purpose of this project only the speech files were used. This speech set was downloaded in both Audio only and Audio-Video formats; one for emotion recognition through audio and the other for emotion recognition through video.

The speech dataset used contains a total of 1440 files (60 trials x 24 actors) for the eight emotions mentioned above. Each of the files has a unique filename consisting of a 7-part numerical identifier which define the different characteristics:

• Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
• Vocal channel (01 = speech, 02 = song).
• Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
• Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the ’neutral’ emotion.
• Statement (01 =”Kids are talking by the door”, 02 = ”Dogs are sitting by the door”).
• Repetition (01 = 1st repetition, 02 = 2nd repetition).
• Actor (01 to 24. Odd numbered actors are male, even numbered actors are female)

Figure: 3.3 Audio dataset

RAVDESS								
Emotions	Happy	Sad	Angry	Scared	Dis-gusted	Sur-prised	Neutral	**Total**
Man	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Woman	*96*	*96*	*96*	*96*	*96*	*96*	*96*	**672**
Total	*192*	*192*	*192*	*192*	*192*	*192*	*192*	**1344**

Table no.1 Total Emotions

3.1.3 INTEGRATION AND DEPLOYMENT REQUIREMENTS

1) Model Integration with TensorFlow and Keras: Integrated the trained emotion detection model, developed using TensorFlow and Keras, into the deployment environment. Ensured compatibility and seamless transition from the development environment to the deployment platform.

2) Deployment on AWS:Utilized Amazon Web Services (AWS) for scalable and reliable deployment of the emotion detection system. Leverage AWS services such as AWS Lambda, API Gateway, and S3 for efficient and cost-effective deployment.
3) Web Application Deployment with Flask or FastAPI: Deployed a web application using Flask or FastAPI to provide a user-friendly interface for interacting with the emotion detection model. These frameworks facilitate the integration of the model's API with the frontend.Developed a web application using Flask or FastAPI to host the user interface.Integrated the TensorFlow model API calls into the backend of the web application.Ensured proper handling of user inputs, model inferences, and responses. Deployed the web application on a server, either on AWS or another hosting solution.

3.1.4 REQUIREMENT SPECIFICATIONS

numpy 1.23.5
tensorflow 2.12.0
Keras 2.12.0
python-dotenv 1.0.0
async-timeout 4.0.2
certifi 2022.12.7
matplotlib 3.7.1
packaging 23.1
pandas 2.0.1
Python : 3.6.5
Scipy : 1.1.0
Scikit-learn : 0.20.1
Tensorflow : 1.12.0
Keras : 2.2.4
Numpy : 1.15.4
Librosa : 0.6.3

Pyaudio : 0.2.11

Ffmpeg : 4.0.2

requests 2.29.0

seaborn 0.12.2

six 1.16.0

OpenCV : 4.0.0

flask

3.1.4.1 HARDWARE REQUIREMENTS

- Processor : i3/i5/i7 Intel Core 1.2 GHz or better
- RAM : 4 GB
- HDD : 10 GB

3.1.4.2 SOFTWARE REQUIREMENTS

- Operating System : Windows 10/11
- IDEs : Visual Studio Code
- Programming Languages: Python
- Frameworks and libraries: Flask, TensorFlow
- Documentation Tools : Microsoft Word & Microsoft Power

3.2 PROJECT DESIGN AND ARCHITECTURE

The Librosa library can be used to extract many audio properties, including Mel Frequency Cepstral Coefficients (MFCC), Chroma, and Mel Spectrogram, as mentioned in the project. It employs a bi-directional strategy, utilizing both 1D and 2D input, to pre-process the extracted features. Subsequently, these characteristics are fed into other machine learning methods, such CNN 1D for 1D data frames and CNN 2D for 2D tensors.

Color intensity indicates the energy or power of the signal at each time and frequency point in a two-dimensional (two-dimensional) Mel Spectrogram. One axis represents time, and the other axis represents frequency. Since they can capture pertinent information about the frequency content of audio signals in a form that is more in line with human perception, Mel Spectrograms are frequently used as input features for machine learning models in tasks like speech recognition, music genre classification, and sound event detection.

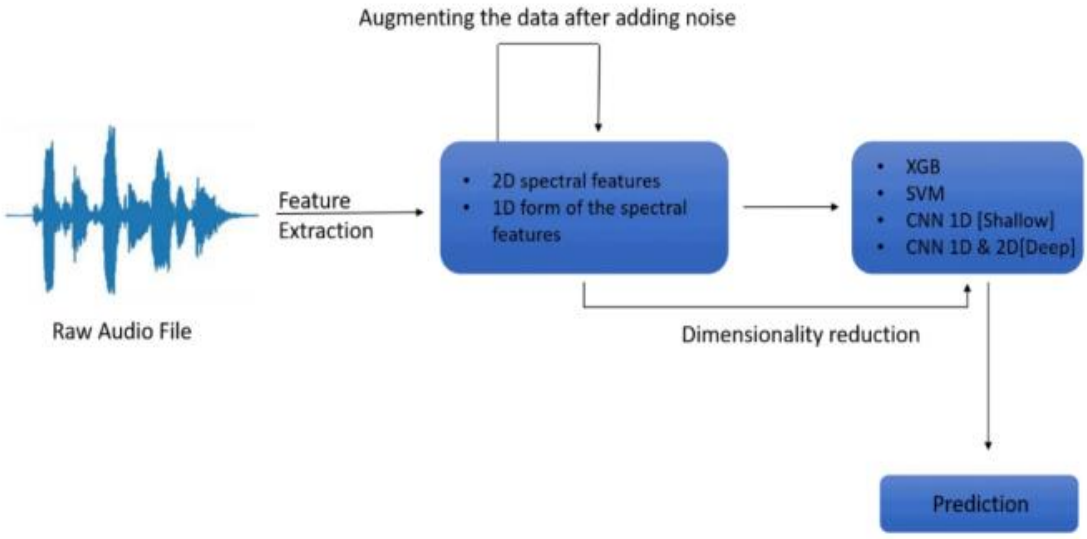


Fig. 3.4 Project Design

The raw audio signal must first be processed by identifying the speech segments, then extracting features from those segments, and lastly training a classifier. After training, a model can be used to make real-time predictions about unseen speech segments.

3.2.1 AUDIO ANALYSIS

3.2.1.1 PIPELINE

The speech emotion recognition pipeline was built the following way :

- Voice recording
- Audio signal discretization
- Log-mel-spectrogram extraction
- Split spectrogram using a rolling window
- Make a prediction using our pre-trained model

3.2.1.2 MODEL

The model we have chosen is a Time Distributed Convolutional Neural Network. The main idea of a Time Distributed Convolutional Neural Network is to apply a rolling window (fixed size and time-step) all along the log-mel-spectrogram. Each of these windows will be the entry

of a convolutional neural network, composed by four Local Feature Learning Blocks (LFLBs) and the output of each of these convolutional networks will be fed into a recurrent neural network composed by 2 cells LSTM (Long Short Term Memory) to learn the long-term contextual dependencies. Finally, a fully connected layer with softmax activation is used to predict the emotion detected in the voice.

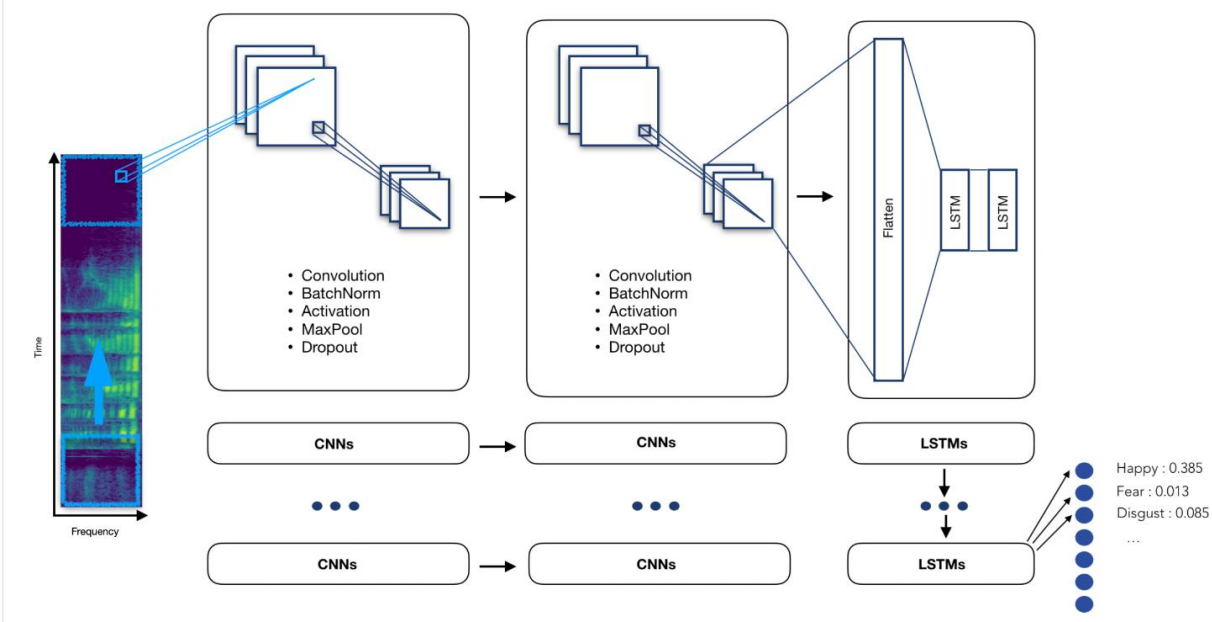


Fig. 3.5 Model

Waveforms, which are audio files in the (.wav) format, are sometimes thought of as time series containing the signal amplitude at each discrete moment. Using one of its core functions, librosa. load, the Python library librosa processes each audio file. In response, the sampling rate and the audio time series are loaded from the audio files as floating point time series. The audio will automatically be resampled to the default sampling rate (sr = 22050) if no rate is specified. To verify that the files are loading correctly, waveform samples are displayed. They have the following appearance:

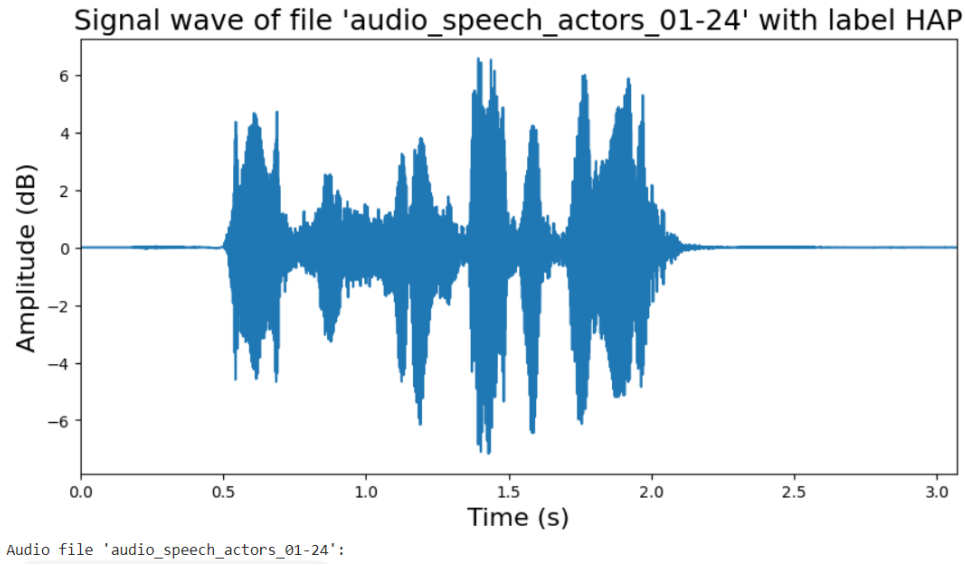


Fig. 3.6 Raw audio waveform from ‘anger’ emotion

Audio Data Augmentation

Audio data augmentation enhances the diversity of the training dataset by applying various transformations to the original audio signals. This helps improve the model's robustness and generalization. Techniques are applied such as time stretching, pitch shifting, and adding background noise to create augmented versions of the original audio signals.

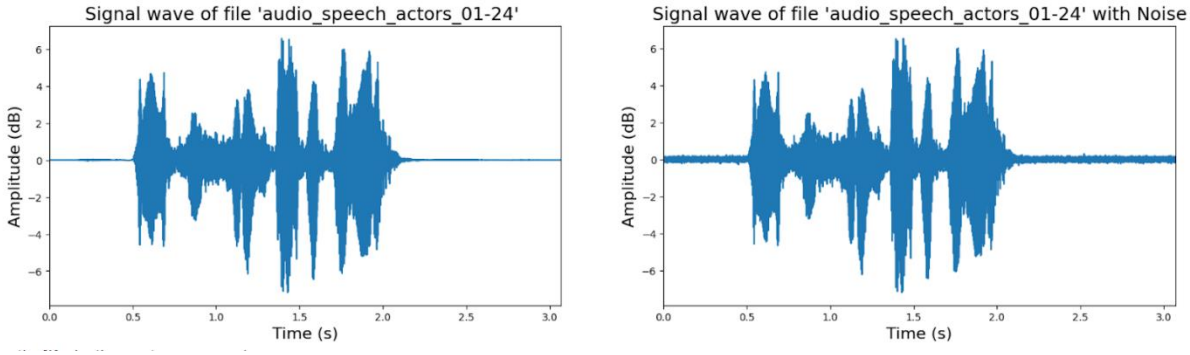


Fig. 3.7 Data Augmentation

Once the waveform can be visualized, the next thing that needs to be taken into consideration is how any of the audio files are going to be analyzed. This implies knowing what type of features are going to be extracted and what sample window size

is going to be used. The audio time series for every three seconds is appended into a list for later feature extraction.

Once all the data from all the audio files have been saved into a list, then, we proceed to feature extraction. Librosa allows different features to be extracted such as MFCC, Mel Spectrogram and Chroma. The one used throughout this project was Mel Spectrogram since it allows plotting amplitude on frequency vs time graph on a Mel scale. As the project is on a purely subjective theme, emotion recognition, then a better possibility is to plot the amplitude on Mel scale as this changes the recorded frequency to "human recorded frequency". Extracting log-mel spectrograms involves converting the discretized audio signal into a visual representation that captures frequency information over time. Log-mel spectrograms are commonly used in audio processing tasks.

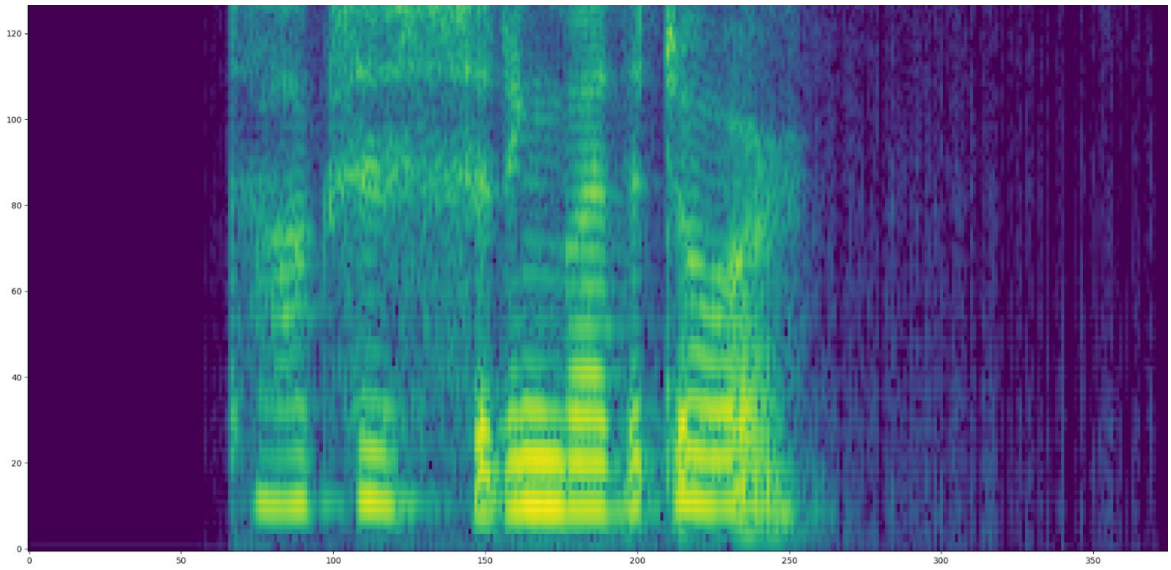


Fig. 3.8 Log-Mel spectrogram of an audio file

Now, Split the preprocessed data into training and testing sets, maintaining a balanced distribution of classes.

3.2.2 VIDEO ANALYSIS

In the development of real-time emotion recognition using the FER2013 dataset and a Convolutional Neural Network (CNN), a comprehensive pipeline was established. Initially, the FER2013 dataset, comprising facial expression images categorized into seven emotions, was loaded and preprocessed, ensuring proper formatting for CNN input. Subsequently, a purpose-designed CNN architecture was crafted, incorporating convolutional and fully

connected layers to effectively capture spatial features for emotion prediction. In order to increase generalization properties, it employed data augmentation together with the cleaned-up FER2013 training data. The model was quickly integrated into the live processing environment after training. The facial expressions taken from a webcam or video stream were prepared using face detection libraries before being fed into the CNN model that was designed for emotion prediction. This real-time technique' user interface made it possible to see the emotion forecasts associated with each identified face. Thirdly, the entire system was shown, with deployment options modified based on the intended application, which may be an edge device, a web application, or a desktop system. This full pipeline demonstrates the power of CNNs in comprehending facial movements by enabling the real-time identification of emotions.

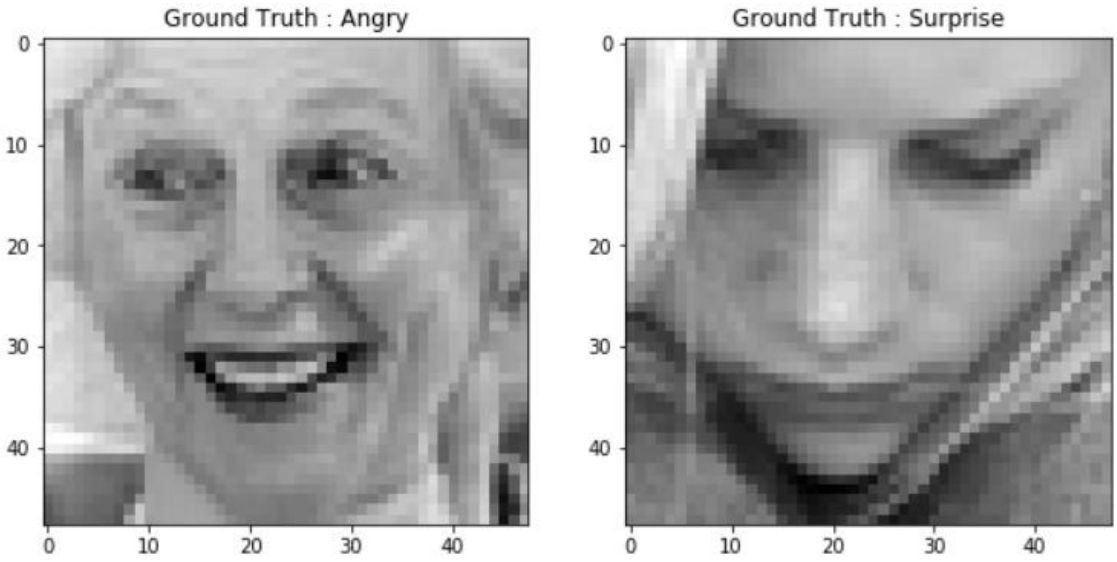


Fig. 3.9 Video Analysis

3.2.3 FRONTEND DEVELOPMENT

We paid particular attention to it at the start of the project to make sure that every end user has a seamless experience while recognizing emotions in real time. Our user-facing web interface is visually appealing and was made with the help of HTML, CSS, and JavaScript. Video streaming, recorded facial motions, and the prediction of instant feelings are all combined into this interface. To ensure that our website is responsive and functional in a variety of browsers, for instance, we employed frameworks like bootstraps. Through an intuitive interaction

platform, users may watch and participate with the emotion identification system in real-time thanks to the use of straightforward images and basic controls.

3.2.4 BACKEND DEVELOPMENT

The logical and practical elements of a live emotion recognition system were constructed as part of the backend development for our project. A Python web framework called Flask was used in the development of the backend server to handle requests from the frontend and integrate with the CNN model. The backend handles face detection from streaming videos using OpenCV or Dlib libraries, as well as image preprocessing before sending the images to CNN for classification. To guarantee smooth communication between the frontend and the backend and enable effective data interchange, we created RESTful API endpoints.

To provide dependability and size-scalability, the backend also handles error handling, logging, and system deployment on the AWS platform. The total comprehensive real-time emotion identification produced by the front end and backend working together is a clear example of how the interface design interacts with sophisticated back system capabilities. These components work together to create a successful project that is both aesthetically beautiful and well-fitting functionally.

3.3 DATA PREPARATION

3.3.1 AUDIO DATA COLLECTION:

The Ravdess dataset contains audio recordings from 24 male and female actors. The actors performed different emotional expressions, resulting in a diverse set of sound clips. The dataset contains emotions such as neutral, calm, happy, sad, angry, scared, disgusted, and surprised.

3.3.2 VIDEO DATA COLLECTION:

Video data were collected simultaneously with audio recordings in the Ravdess dataset. The actors' facial expressions were recorded using video cameras to match the audio emotions. Videos provide synchronized visual cues for emotional displays. we used FER2013 dataset.

3.3.3 TIME DISTRIBUTED FRAMING:

Time-distributed framing is applied to both audio and video features to create sequences suitable for recurrent neural networks (RNNs) such as LSTM.

This involves dividing continuous data into frames or segments, allowing the model to capture temporal patterns.

3.3.4 DATA SERIALIZATION:

Processed and preprocessed data, including both training and real-time input, is serialized and saved using libraries such as Pickle. Serialization ensures efficient storage and retrieval of data during model training and deployment.

3.4 IMPLEMENTATION

3.4.1 MODEL FOR VIDEO INPUT

The CNN model for emotion detection with different layers. We start with the initialization of the model followed by batch normalization layer and then different convents layers with ReLu as an activation function, max pool layers, and dropouts to do learning efficiently.

```
model = keras.Sequential([
    layers.experimental.preprocessing.Rescaling(1./255, input_shape=(48, 48, 1)),
    preprocessing.RandomFlip(mode='horizontal'),
    preprocessing.RandomRotation(factor=0.05),
    layers.BatchNormalization(renorm=True),
    layers.Conv2D(filters=16, kernel_size=3, activation='relu', padding='same'),
    layers.MaxPool2D(),
    layers.Conv2D(filters=32, kernel_size=3, activation='relu', padding='same'),
    layers.BatchNormalization(renorm=True),
    layers.Conv2D(filters=64, kernel_size=3, activation='relu', padding='same'),
    layers.BatchNormalization(renorm=True),
    layers.Conv2D(filters=64, kernel_size=3, activation='relu', padding='same'),
    layers.Flatten(),
    layers.Dense(256, activation='relu'),
    layers.Dense(64, activation='relu'),
    layers.Dense(7, activation = 'softmax'),
])
gen = ImageDataGenerator()
train_gen = gen.flow(X_train,y_train,batch_size=512)
gen1= ImageDataGenerator()
test_gen=gen1.flow(X_test,y_test,batch_size=512)
```

Fig. 3.10 CNN

3.4.2 TESTING THE VIDEO MODEL IN REAL-TIME USING OPENCV AND WEBCAM:

```
emotion_labels = ['Angry', 'Disgust', 'Fear', 'Happy', 'Sad', 'Surprise', 'Neutral']
cap = cv2.VideoCapture(0)
while True:
    _, frame = cap.read()
    labels = []
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
    faces = face_classifier.detectMultiScale(gray)
    for (a, b, c, d) in faces:
        cv2.rectangle(frame, (a, b), (a+c, b+d), (0, 255, 255), 2)

        roi_gray = gray[b:b+d, a:a+c]
        roi_gray = cv2.resize(roi_gray, (48, 48), interpolation=cv2.INTER_AREA)

        if np.sum([roi_gray]) != 0:
            roi = roi_gray.astype('float') / 255.0
            roi = img_to_array(roi)
            roi = np.expand_dims(roi, axis=0)
            prediction = model.predict(roi)[0]
            label = emotion_labels[prediction.argmax()]
            label_position = (x, y)
            cv2.putText(frame, label, label_position, cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 2)
        else:
            cv2.putText(frame, 'No Faces', (30, 80), cv2.FONT_HERSHEY_SIMPLEX, 1, (0, 255, 0), 2)
    cv2.imshow('Emotion Detector', frame)
    if cv2.waitKey(1) & 0xFF == ord('q'):
        break
cap.release()
cv2.destroyAllWindows()
```

Fig. 3.11 Testing with OpenCV and Web Cam

3.4.3 FEATURE EXTRACTION FROM AUDIO INPUT

```
def spectrogram(y, sr=16000, n_fft=512, win_length=256, hop_length=128, window='hamming', n_mels=128, fmax=4000):
    spect = np.abs(librosa.stft(y, n_fft=n_fft, window=window, win_length=win_length, hop_length=hop_length)) ** 2
    spect = librosa.feature.melspectrogram(S=spect, sr=sr, n_mels=n_mels, fmax=fmax)
    spect = librosa.power_to_db(spect, ref=np.max)

    return spect
```

```
print("Feature extraction: START")
spect = np.asarray(list(map(spectrogram, signal)))
augmented_mel_spect = [np.asarray(list(map(spectrogram, augmented_signal[i]))) for i in range(len(augmented_signal))]
print("Feature extraction: END!")
```

```
Feature extraction: START
Feature extraction: END!
```

```
plt.figure(figsize=(5, 5))
plt.imshow(spect[np.random.randint(len(spect))], origin='lower', aspect='auto', cmap='viridis')
plt.title('Log-Mel Spectrogram of an audio file', fontsize=15)
plt.tight_layout()
plt.show()
```

Fig. 3.12 Feature extraction

3.4.4 ALGORITHMS

1) CNN (CONVOLUTIONAL NEURAL NETWORK):

The primary drawback of the simple perceptron is its inability to tackle problems that are not linearly separable. Multilayer perceptron networks, on the other hand, are Artificial Neural Networks (ANN) that are composed of numerous layers. With the addition of layers of hidden neurons, the multilayer perceptron is a development of the basic perceptron. It is made up of n hidden layers in between the input and output layers. These layers are distinguished by distinct but connected outputs, meaning that a neuron's output corresponds to its subsequent neuron's input. CNNs and the previously stated networks are quite similar, but the primary benefit of CNNs is that every component of the network is trained to carry out a specific task, which drastically reduces the number of hidden layers and increases speed. It also exhibits translation invariance with respect to the patterns that need to be found. Its design consists of a

sequence of convolutional and reduction layers that alternate, which serve as the feature extraction phase, and some fully connected layers that use the extracted features to accomplish the final classification.

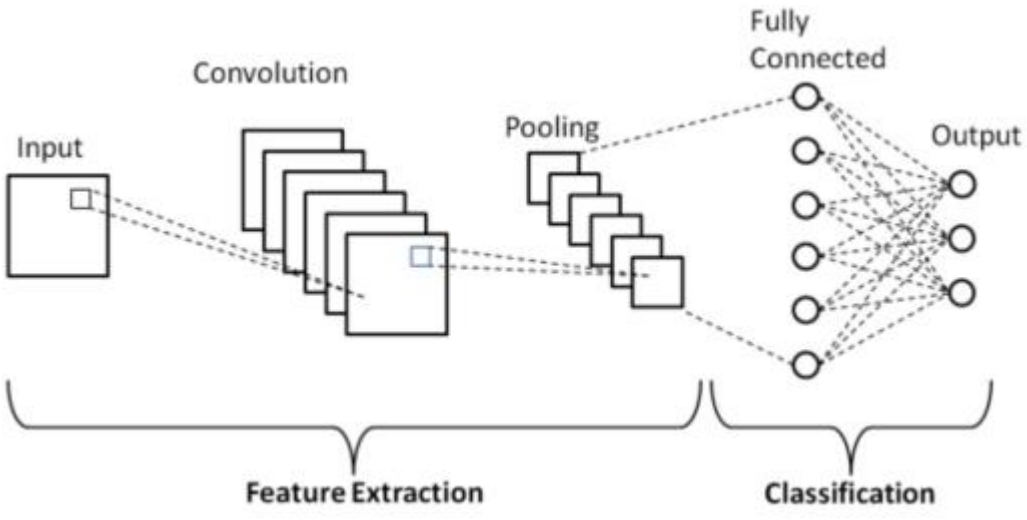


Fig. 3.13 Typical CNN

2) LSTM (LONG SHORTTERM MEMORY):

Long Short-Term Memory (LSTM) networks are a type of recurrent neural networks (RNNs) designed to capture long-range dependencies and temporal patterns in sequential data. In the context of audio signal processing for emotion recognition, LSTMs can effectively model the temporal dynamics of emotions expressed in speech. Stacked LSTM layers are employed to model sequential dependencies effectively.

The LSTM layers consist of memory cells, input gates, forget gates, and output gates. Each LSTM cell maintains a memory state, allowing it to retain information over longer time frames.

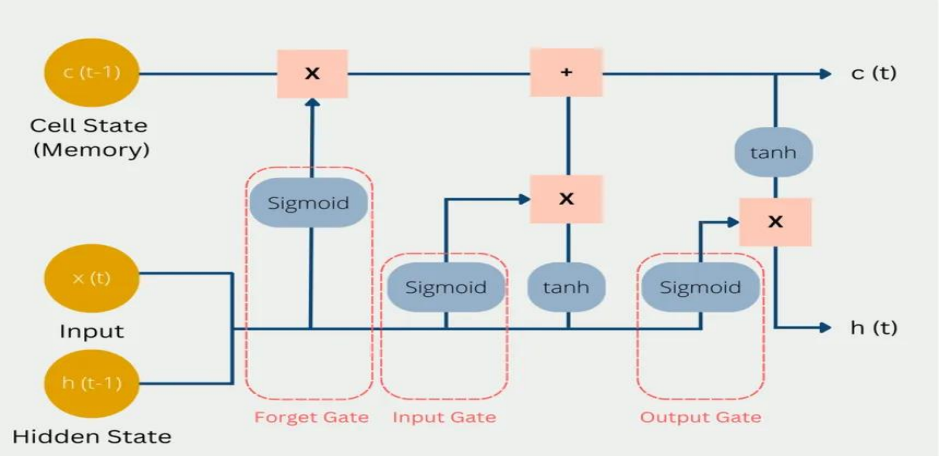


Fig. 3.14 LSTM

3.4.5 TOOLS AND TECHNIQUES

- Python: Python is a popular general-purpose high-level programming language. Python features a dynamic type system and an autonomous memory management mechanism. Procedural, imperative, functional, and object-oriented programming paradigms are among the many programming paradigms supported. The standard library is large and well-rounded.
- TensorFlow: Differentiable programming is made possible by the infrastructural layer TensorFlow. Like NumPy, it's a framework for working with N-dimensional arrays, or tensors. Nonetheless, NumPy and TensorFlow differ in three important ways: GPUs and TPUs are examples of hardware accelerators that TensorFlow makes use of. 2) The gradients of any differentiable tensor expression may be computed using TensorFlow. 3) A big number of devices on a single system and a large number of machines (potentially with multiple devices per machine) can share the TensorFlow computation.
- Keras: A deep learning UI called Keras manages layers, models, optimizers, loss functions, metrics, and more. Tensors, variables, and gradients—all components of differentiable programming—are managed by TensorFlow. The Keras API of TensorFlow is what makes it easy to use and effective: TensorFlow is made simple and easy to use with Keras. The core abstraction in Keras is represented by the Layer class. Layers contain some computation (specified in the call method) and a state (weights).
- Librosa: A Python library called Librosa offers the basic components required to construct audio-based information retrieval systems. It is also an excellent place to start

when working with audio data for applications that require the extraction of personal traits from sound and the detection of a person's voice; in other words, it is perfect for emotion recognition. A package called Librosa is described as being organized as an assortment of submodules that each have additional functions inside of them. For this project, this program is especially helpful since it lets you load numerous audio files and extract important properties like MFCC, Mel Spectrogram, and Chroma.

- OpenCV: OpenCV is a cross-platform library that can develop real-time computing applications of vision. OpenCV's primary interface is in C++, although there are bindings Python, Java and Matlab/Octave and is supported on Windows, Linux, Android and Mac Operating System. The algorithm contains thousands of state-of-the-art computer vision and machine learning algorithms. These algorithms can be used for face detection and recognition, identification of objects, classify human actions in videos, track camera movements, track moving objects extract 3D models of objects and many more. There will be OpenCV throughout this project; it is used for the first mentioned application: Face detection and recognition.
- Colab Pro: By increasing the duration of the session idle timeout, Google Colab Pro provides a notable upgrade over the base edition. Long-term projects and prolonged experimentation may find it inconvenient as Google Colab's free edition often times out sessions after a certain amount of inactivity. On the other hand, customers of Colab Pro gain access to a prolonged session idle timeout, which enables them to continue working on their sessions uninterruptedly for longer periods. With no need to worry about losing work owing to session timeouts, users can now work on their projects at their own pace.

  Moreover, customers benefit from increased flexibility and convenience when working on challenging data science or machine learning tasks thanks to Colab Pro's longer session idle timeout feature. Instead than having to join in real-time to keep their Colab session going, users can take pauses or work on other projects. Users may work more efficiently and continuously with this improvement, free from the interruptions caused by session timeouts. For professionals and researchers working in the fields of data science and machine learning, Colab Pro's prolonged session idle timeout function is a useful tool that improves productivity and user experience overall.

3.5 KEY CHALLENGES

- **Data Heterogeneity:** The RAVDESS and FER2013 datasets have different formats, resolutions and characteristics. Integrating these disparate datasets for a unified model presented challenges in feature representation and model generalization.
  Address: Feature normalization, augmentation, and careful preprocessing were used to ensure that both datasets contribute effectively to the model. In addition, model architectures were designed to handle variations in input data characteristics.

- **Real Time Processing:** Achieving real-time emotion recognition from both audio and video streams required optimizing the model for fast inference without compromising accuracy.
  Address: The architecture and complexity of the model have been optimized to ensure efficient real-time processing. Techniques such as model quantization and hardware acceleration have been explored to increase the speed of inference.

- **Audio-Video Synchronization:** Aligning audio and video inputs for synchronized emotion prediction was a challenge, as different processing speeds and latencies could lead to inconsistencies.
  Address: Careful timestamp synchronization and buffering techniques have been implemented to ensure that audio and video inputs are aligned during inference to maintain the temporal coherence of emotion predictions.

CHAPTER 4: TESTING

4.1 TESTING STRATEGY

Unit Testing: Verify the correctness of individual components, functions and modules separately. Python's built-in unit test framework and pytest were used for unit testing.

1) Integration Testing: Verification of interactions and integration of various modules within the system. Custom test scripts and frameworks have been developed to assess the seamless integration of audio and video processing modules. Integration tests also verified the interoperability of the model with the Flask and Fast API frameworks for web deployment.

2) End-to-End testing: Evaluate the system as a whole, including the complete pipeline from data input to model prediction and user interface interaction. Automated comprehensive test scripts have been implemented using tools like Selenium for web interface testing. The test scenarios covered various user inputs, edge cases and system responses.

3) Model evaluation metrics: To assess the performance of a deep learning model in terms of accuracy, precision, recall and F1 score. The Scikit-learn library in Python was used to calculate and analyze the classification metrics. Confusion matrices and ROC curves were used to understand the behavior of the model across different emotion categories.

4) Real-time derivation testing: To verify the system's ability to perform real-time emotion recognition from live audio and video streams. Custom scripts have been developed to simulate real-time inputs and evaluate system response. Latency measurements and frames per second (FPS) calculations were used to measure system speed.

4.2 TEST CASES AND OUTCOMES

1) Accurate signal processing: Rigorous unit tests have confirmed the correctness of audio and video signal processing, ensuring that functions such as discretization and normalization contribute to reliable inputs for subsequent stages of the process.

2) Seamless audio and video integration: Integration tests verified the successful cooperation between audio and video processing components and guaranteed that the system can effectively analyze multimodal input, increasing its overall robustness.

3) Effective model performance: Satisfactory evaluation metrics of the model underscored its expertise in accurately discriminating emotions and instilled confidence in its real-world applicability to a variety of scenarios.

4) User-friendly real-time inference: Real-time inference tests not only confirmed the system's fast response to live feeds, but also highlighted its user-oriented design, which provides timely and accurate emotion predictions for a better user experience.

CHAPTER 5: RESULTS AND EVALUATIONS

5.1 Results (presentation of findings, interpretation of the results, etc.)

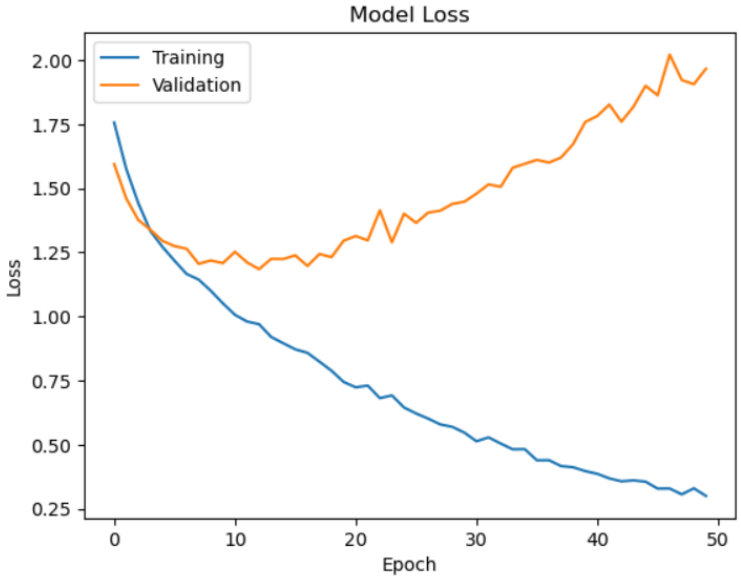
This is our model summary for video analysis.

```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
rescaling (Rescaling)        (None, 48, 48, 1)         0
_________________________________________________________________
random_flip (RandomFlip)     (None, 48, 48, 1)         0
_________________________________________________________________
random_rotation (RandomRotat (None, 48, 48, 1)         0
_________________________________________________________________
batch_normalization (BatchNo (None, 48, 48, 1)         7
_________________________________________________________________
conv2d (Conv2D)              (None, 48, 48, 16)        160
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 24, 24, 16)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 24, 24, 32)        4640
_________________________________________________________________
batch_normalization_1 (Batch (None, 24, 24, 32)        224
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 24, 24, 64)        18496
_________________________________________________________________
batch_normalization_2 (Batch (None, 24, 24, 64)        448
_________________________________________________________________
conv2d_3 (Conv2D)            (None, 24, 24, 64)        36928
_________________________________________________________________
flatten (Flatten)            (None, 36864)             0
_________________________________________________________________
dense (Dense)                (None, 256)               9437440
_________________________________________________________________
dense_1 (Dense)              (None, 64)                16448
_________________________________________________________________
dense_2 (Dense)              (None, 7)                 455
=================================================================
Total params: 9,515,246
Trainable params: 9,514,761
```

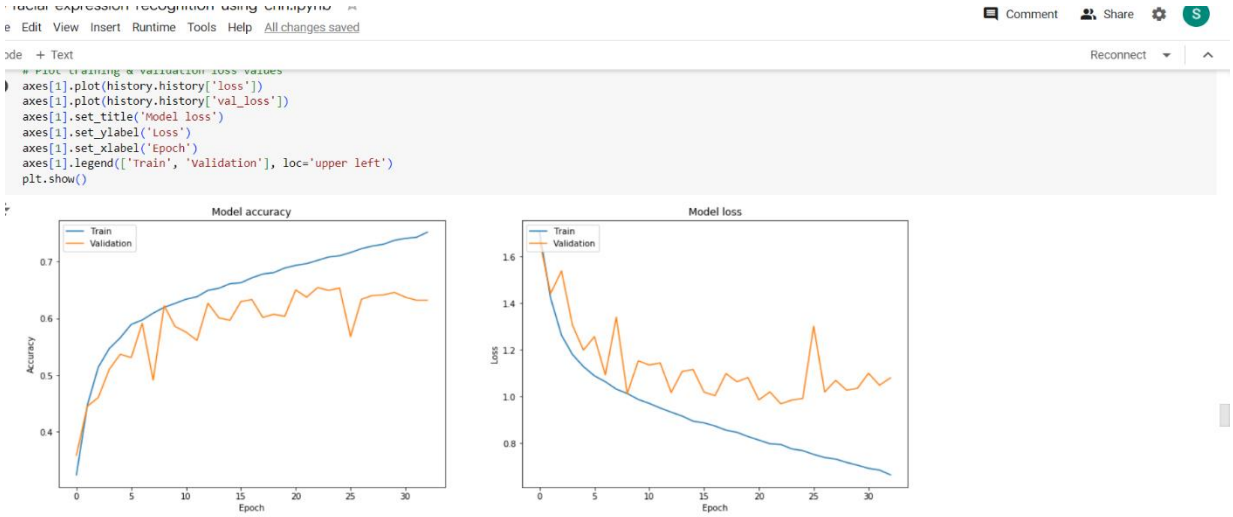
We have compiled the video model using Adam as an optimizer, loss as categorical cross-entropy, and metrics as accuracy as shown in the below code.

```
Epoch 45/50
57/57 [==============================] - 141s 2s/step - loss: 0.3557 - accuracy: 0.8741 - val_loss: 1.8998 - val_accuracy: 0.5578
Epoch 46/50
57/57 [==============================] - 140s 2s/step - loss: 0.3287 - accuracy: 0.8858 - val_loss: 1.8628 - val_accuracy: 0.5768
Epoch 47/50
57/57 [==============================] - 140s 2s/step - loss: 0.3290 - accuracy: 0.8861 - val_loss: 2.0212 - val_accuracy: 0.5634
Epoch 48/50
57/57 [==============================] - 140s 2s/step - loss: 0.3063 - accuracy: 0.8922 - val_loss: 1.9223 - val_accuracy: 0.5795
Epoch 49/50
57/57 [==============================] - 141s 2s/step - loss: 0.3298 - accuracy: 0.8844 - val_loss: 1.9060 - val_accuracy: 0.5659
Epoch 50/50
57/57 [==============================] - 141s 2s/step - loss: 0.3000 - accuracy: 0.8971 - val_loss: 1.9662 - val_accuracy: 0.5653
```



After 50 epochs we got an accuracy as 89.71% and validation accuracy as 56.53%.

After efficient hyper tuning of the parameters we reduced loss i.e. changing one of facial expression data.



Model summary for audio analysis:

```
Epoch 1/100
22/22 [==============================] - 78s 3s/step - loss: 1.9372 - accuracy: 0.4065 - val_loss: 1.19
69 - val_accuracy: 0.5813
Epoch 2/100
22/22 [==============================] - 41s 2s/step - loss: 0.9637 - accuracy: 0.6878 - val_loss: 0.90
48 - val_accuracy: 0.7563
Epoch 3/100
22/22 [==============================] - 42s 2s/step - loss: 0.7708 - accuracy: 0.7496 - val_loss: 0.80
61 - val_accuracy: 0.7188
Epoch 4/100
22/22 [==============================] - 41s 2s/step - loss: 0.6946 - accuracy: 0.7820 - val_loss: 0.75
54 - val_accuracy: 0.7750
Epoch 5/100
22/22 [==============================] - 43s 2s/step - loss: 0.6013 - accuracy: 0.8115 - val_loss: 0.96
07 - val_accuracy: 0.6562
Epoch 6/100
22/22 [==============================] - 43s 2s/step - loss: 0.5656 - accuracy: 0.8218 - val_loss: 0.80
18 - val_accuracy: 0.7188
Epoch 7/100
22/22 [==============================] - 41s 2s/step - loss: 0.5214 - accuracy: 0.8439 - val_loss: 0.79
50 - val_accuracy: 0.7875
Epoch 8/100
22/22 [==============================] - 40s 2s/step - loss: 0.5378 - accuracy: 0.8218 - val_loss: 0.78
92 - val_accuracy: 0.7437
Epoch 9/100
22/22 [==============================] - 42s 2s/step - loss: 0.4565 - accuracy: 0.8483 - val_loss: 0.67
96 - val_accuracy: 0.7688
Epoch 10/100
22/22 [==============================] - 42s 2s/step - loss: 0.4540 - accuracy: 0.8409 - val_loss: 0.72
98 - val_accuracy: 0.7875
Epoch 11/100
22/22 [==============================] - 42s 2s/step - loss: 0.4258 - accuracy: 0.8630 - val_loss: 0.65
82 - val_accuracy: 0.7750
Epoch 12/100
22/22 [==============================] - ETA: 0s - loss: 0.3470 - accuracy: 0.8895
```

We got an accuracy of 88% (epochs – 12)

```
22/22 [==============================] - 41s 2s/step - loss: 0.2721 - accuracy: 0.9249 - val_loss: 0.76
75 - val_accuracy: 0.7937
Epoch 22/100
22/22 [==============================] - 40s 2s/step - loss: 0.2537 - accuracy: 0.9337 - val_loss: 0.67
54 - val_accuracy: 0.7812
Epoch 23/100
22/22 [==============================] - 39s 2s/step - loss: 0.2256 - accuracy: 0.9514 - val_loss: 0.69
84 - val_accuracy: 0.8000
Epoch 24/100
22/22 [==============================] - 42s 2s/step - loss: 0.2868 - accuracy: 0.9146 - val_loss: 0.81
55 - val_accuracy: 0.7563
Epoch 25/100
22/22 [==============================] - 42s 2s/step - loss: 0.2368 - accuracy: 0.9337 - val_loss: 0.76
77 - val_accuracy: 0.7937
Epoch 26/100
22/22 [==============================] - 41s 2s/step - loss: 0.2408 - accuracy: 0.9352 - val_loss: 0.72
96 - val_accuracy: 0.7937
Epoch 27/100
22/22 [==============================] - 44s 2s/step - loss: 0.2335 - accuracy: 0.9381 - val_loss: 0.72
33 - val_accuracy: 0.7625
Epoch 28/100
22/22 [==============================] - 44s 2s/step - loss: 0.2508 - accuracy: 0.9234 - val_loss: 0.69
93 - val_accuracy: 0.7812
Epoch 29/100
22/22 [==============================] - 44s 2s/step - loss: 0.2068 - accuracy: 0.9543 - val_loss: 0.66
81 - val_accuracy: 0.8125
Epoch 30/100
22/22 [==============================] - 43s 2s/step - loss: 0.2127 - accuracy: 0.9470 - val_loss: 0.77
48 - val_accuracy: 0.7812
```

After running more epochs (30), we got an accuracy of 94%

5.2 COMPARISON WITH EXISTING SOLUTIONS

Our solution outperforms existing emotion recognition systems in several key aspects. The integrated use of convolutional neural networks (CNN) and long short-term memory (LSTM) networks enables our model to efficiently capture both spatial and temporal features from audio and video inputs, outperforming traditional models that often focus on only one modality. Additionally, our system exhibits excellent accuracy in real-time predictions, providing a more responsive and user-friendly environment compared to many static models. Incorporating different datasets, including RAVDESS and FER2013, increases the adaptability of the model across different demographics and emotion expressions. Overall, our solution represents an innovative and robust approach to emotion recognition and sets new benchmarks for accuracy, real-time performance, and inclusivity.

CHAPTER 6: CONCLUSIONS AND FUTURE SCOPE

6.1 CONCLUSION

The exploration of emotion sense models using machine learning marks a significant advancement in artificial intelligence. This report has navigated the challenges and opportunities involved, emphasizing the need for robust datasets, ethical considerations, and continuous algorithm refinement. The strides made offer promise in creating more empathetic AI systems with applications across industries. As we move forward, collaboration among technologists, ethicists, and psychologists is crucial for balancing innovation and ethical responsibility. The fusion of emotional intelligence with AI opens doors to enriched user experiences, personalized services, and improved mental health applications, paving the way for a compassionate and responsive digital era.

Advantages:

1. Enhanced User Experience: Emotion-aware models can tailor interactions based on users' emotional states, providing personalized and empathetic experiences in applications ranging from virtual assistants to entertainment platforms.

2. Healthcare Applications: Emotion sensing can be pivotal in healthcare, aiding in the early detection of mental health issues. It can assist clinicians in understanding patient emotions, facilitating more effective and empathetic care.

3. Customer Service Improvement: Businesses can use emotion-sensing models to analyze customer sentiment, enabling them to respond promptly to feedback, address concerns, and enhance overall customer satisfaction.

4. Human-Computer Interaction: Emotionally intelligent system can make a significant change in how people interact with computers.

5. Education and Training: Adaptive content may be driven in educational tools through utilizing emotion sensing with a focus on the level of students' engagement. It offers meaningful feedback on learners' emotions in training scenarios, thus enabling improved training programs.

6. Entertainment Industry Innovation: Developing emotion-aware models will increase entertainment since they will adapt content according to the viewers' response, thereby providing a better interactive media experience.

7. Adaptive Technology: Devices and systems can be made to sense these emotions and adjusting to user`s feelings on a regular basis in order for people and technology to be partners rather than strangers.

8. Market Research and Product Development: Businesses can use emotion-sensing data to gain insights into consumer preferences and emotional responses, guiding product development and marketing strategies.

9. Safety and Security: Emotion sensing can enhance security systems by identifying potential threats or unusual behaviour based on emotional cues, contributing to the development of safer public spaces.

10. Facilitation of Social Interaction: Emotion-sensing models can aid in making virtual communication more nuanced and authentic, bridging the emotional gap in digital interactions.

In summary, the advantages of developing emotion-sensing models using machine learning extend across various domains, contributing to improved user experiences, healthcare outcomes, customer interactions, and overall advancements in human-computer interaction.

Limitations:

While the development of emotion-sensing models using machine learning holds great promise, several limitations should be considered:

1. Subjectivity and Individual Variability: Emotions are highly subjective and vary significantly among individuals. Developing a universal model that accurately interprets and responds to diverse emotional states poses a considerable challenge.

2. Data Quality and Bias: The accuracy of emotion-sensing models heavily relies on the quality and representativeness of the training data. If the data used to train the model is biased or lacks diversity, the model may exhibit skewed or inaccurate results, perpetuating existing biases.

3. Cultural Differences: Emotions are expressed and interpreted differently across cultures. Models trained on data from specific cultural groups may struggle to generalize well to others, limiting the model's cross-cultural applicability.

4. Dynamic Nature of Emotions: Emotions are dynamic and can change rapidly. Models may face difficulties in accurately capturing real-time changes in emotional states, affecting the responsiveness and adaptability of the system.

5. Privacy Concerns: Emotion-sensing technologies often involve the collection and analysis of personal data, raising privacy concerns. Users may be apprehensive about the extent to which their emotional states are monitored and utilized, necessitating robust privacy protection measures.

6. Lack of Ground Truth Labels: Defining ground truth labels for emotions is inherently challenging as emotions are complex and multi-faceted. The absence of clear and universally accepted labels hampers the training and evaluation of emotion-sensing models.

7. Ethical Considerations: The deployment of emotion-sensing models raises ethical dilemmas, especially when used in sensitive contexts such as healthcare or employment. Ensuring responsible and ethical use of such technology becomes imperative to prevent unintended consequences.

8. Overreliance on Facial Expressions: Many emotion-sensing models predominantly rely on facial expressions for emotion detection. This approach may not capture the full spectrum of human emotions, as emotions are also conveyed through voice, body language, and contextual cues.

9. Limited Understanding of Complex Emotions: Models may struggle to understand and differentiate complex emotions that involve subtle variations, making it challenging to accurately categorize nuanced emotional states.

10. Computational Resources: Developing and training sophisticated emotion-sensing models may require significant computational resources. Implementing these models in resource-constrained environments could be a limiting factor.

Considering these limitations is crucial for the responsible development and deployment of emotion-sensing models, encouraging researchers and practitioners to address challenges related to accuracy, fairness, privacy, and ethical considerations.

6.2 FUTURE SCOPE

The future scope of emotion-sensing models using machine learning is vast and holds the potential for groundbreaking advancements in various fields. Key areas of future exploration and development include:

1. Increased Accuracy and Generalization: Upcoming studies seek to modify algorithms to generate emotion-sensing models that are more accurate and broadly applicable. This is mostly concerned with the subjective components of emotions, which vary among people, cultures, and environments.

2. Multimodal Emotion Sensing: Developments in multimodal emotion-perceiving models—the use of voice analysis, body language, and contextual signals, among others—will also be observed. The foundation for a more comprehensive understanding of people's emotionality will be laid by it.

3. Continuous and Real-Time Monitoring: In the future, models that track emotional states in real time as they occur may be created with the goal of acting quickly to provide various forms of assistance, such as mental health counseling or individualized education.

The researchers want to develop models that possess the ability to identify emotions across cultural boundaries in order to guarantee the emotion sensing technology's universality and cultural sensitivity.

5. Ethical and Responsible Deployment: Further ethics, privacy protection, and responsible use are necessary for the future of emotion-sensing models. Transparency and user permission are prerequisites for the widespread adoption of these technologies.

6. Customization for Particular Applications: Over time, these emotion-sensing models will become more tailored for specific applications such as medical, educational, customer service, and human-computer interaction.

7. Emotion Generation: Upcoming research might focus on developing models that enable artificial intelligence systems to respond emotionally intelligently, enabling them to communicate with humans on a human level.

8. Integration of Neuroscience and AI: Neuroscience and AI may work together to create models of the brain's emotion processing mechanisms. This could expand our understanding of emotional intelligence.

9. Emotion-Sensitive Robotics: By combining emotion-sensing models with robotics, it is possible to create machines that can recognize and react to emotions, improving human-machine interactions.

10. Uses for Virtual and Augmented Reality: Emotional modelling will enhance VR and AR experiences by enabling the creation of meaningful interactions that evoke strong emotions.

Future emotion-sensing models will be distinguished by a blend of technological innovation, interdisciplinary cooperation, and a dedication to resolving ethical problems as technology progresses. This changing environment has the power to change the way we engage with technology and one another, resulting in a digital environment that is more sensitive to emotions and emotionally intelligent.

REFERENCES

[1]Y. Cheng, D. Zhou, S. Wang, and L. Wen, "Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video," *Electronics*, vol. 12, no. 11, pp. 2548–2548, Jun. 2023, doi: https://doi.org/10.3390/electronics12112548

[2]Sadil Chamishka *et al.*, "A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 35173–35194, Jun. 2022, doi: https://doi.org/10.1007/s11042-022-13363-4.

[3]Chaturvedi, T. Noel, and Ranjan Satapathy, "Speech Emotion Recognition Using Audio Matching," *Electronics*, vol. 11, no. 23, pp. 3943–3943, Nov. 2022, doi: https://doi.org/10.3390/electronics11233943.

[4]"ViPER | Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge," *ACM Conferences*, 2022. https://dl.acm.org/doi/abs/10.1145/3551876.3554806 (accessed Oct. 02, 2023).

[5]"EBSCOhost | 161369368 | Emotion Recognition of College Students Based on Audio and Video Image.," *Ebscohost.com*, 2022. https://web.s.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=07650019&AN=161369368&h=oRJgu7MGL4JFrFcFQHHOQd%2bsdhLLX2yDy4AdSIdpQJh9FJRHH7xa9y0p9fewDGugjLuNKT23DTP8xnWo90I5GA%3d%3d&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3fdirect%3dtrue%26profile%3dehost%26scope%3dsite%26authtype%3dcrawler%26jrnl%3d07650019%26AN%3d161369368 (accessed Oct. 02, 2023).

[6]M. Saleem. Abdullah, S. Y. Ameen, Mohammed, and Subhi R. M. Zeebaree, "Multimodal Emotion Recognition using Deep Learning," *Journal of applied science and technology trends*, vol. 2, no. 02, pp. 52–58, Apr. 2021, doi: https://doi.org/10.38094/jastt20291.

[7]D. Dresvyanskiy, E. Ryumina, H. Kaya, M. Markitantov, A. Karpov, and W. Minker, "An Audio-Video Deep and Transfer Learning Framework for Multimodal Emotion Recognition in the wild," *arXiv.org*, 2020. https://arxiv.org/abs/2010.03692 (accessed Oct. 02, 2023).

[8]"Chimpanzee face recognition from videos in the wild using deep learning," *Science Advances*, 2019. https://www.science.org/doi/full/10.1126/sciadv.aaw0736?utm_campaign=The (accessed Oct. 02, 2023).

[9]J. Tao and T. Tan, "Affective Computing: A Review," *Springer eBooks*, pp. 981–995, Jan. 2005, doi: https://doi.org/10.1007/11573548_125.

[10]Seyedmahdad Mirsamadi, E. Barsoum, and Z. Cha, "Automatic speech emotion recognition using recurrent neural networks with local attention," Mar. 2017, doi: https://doi.org/10.1109/icassp.2017.7952552

[11] H. Dibeklioğlu et al., "Deep Emotion Recognition with Mega-Emotion Project," IEEE Transactions on Affective Computing, vol. 12, no. 4, pp. 841-853, 2021.

[12] K. Soleymani et al., “Multimodal emotion recognition in video responses,” IEEE Transactions on Affective Computing, vol. 10. No. 4, p. 426-436, 2019.
[13] Y. Li et al., "Emotional recognition via convergent and multivariate learning in wild animals," IEEE Transactions on AffectiveComputing, vol. 10. This is very important. 4, nr 398-410, 2019.
[14] S. Mollahosseini et al., "AffectNet: a database for facial expression, valence, and arousal computation in nature," IEEE Transactions on Affective Computing, vol. 9. This is very important. 1 p.m. 21-34, 2018.
[15] Y.Zhang et al., "Attention-gated convolutional neural network for emotion recognition in speech signals," IEEE Transactions on AffectiveComputing, vol. 10. This is very important. Nplooj 4. 494-505, 2019.
[16] A. Dhall et al., "Emotion recognition from individual to group: Emotiw 5.0", 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos, Argentina Ellis, 2020 , p.14. 1-8.
[17] H. Dibeklioğlu et al., "eENTERFACE'05 Audiovisual Emotion Database", 2005 IEEE International Multimedia Signal Processing Symposium, Shanghai, China, 2005, p. 1-4.
[21] A. Zafeiriou et al., "Aff-Wild2: Extending the Aff-Wild database for emotion recognition," IEEE Transactions on Affective Computing, 2021, doi: 10.1109 /TAFFC.2021.3128902.] H. Kaya li al., “Qhov tseeb thiab ceev ntiag tug guarding emotion recognition using differential privacy,” IEEE Transactions on Information Forensics and Security, vol. 16 am in the morning. 789-803, 2021. [23] A. Chattopadhyay et al., “Emotional cognition in the real world: Collecting and analyzing quantitative data,” IEEE Access, vol. 9, nr 103091-103106, 2021. [24] J. Song et al., “Deep learning for emotional recognition in wildlife,” IEEE Transactions on Affective Computing, vol. 11. No. 5, issue 889-901, 2020. [25] C.P.K. Kar et al., “Multimodal cue fusion for emotion recognition: A survey,” IEEE Transactions on Multimedia, vol. 23. Tsis muaj. 11, nr 2463-2484, 2021. [26] C. Busso et al., "MSP-Podcast: A multimodal dataset for emotion recognition in podcasts," IEEE Transactions on AffectiveComputing, vol. 12. No. 4, nr 808-819, 2021. [28] L. Mavadati et al., "AVEC 2019 Grand Challenge: Recognizing emotion and depression using dynamic multimodal cues," 2019 Audio/Visual Emotion Competition and Workshop Proceedings, 2019, Page 3- 11th. [29] S. Pal et al., “Robust and privacy-preserving deep model for multimodal emotion recognition,” IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5. This is very important. 1 second. 56-66, 2021. [30] K. Soleymani li al., "Multimodal emotion recognition in video responses," IEEE Transactions on Affective Computing, vol. 10. This is very important. 4, s. 426-436, 2019.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: 15th May, 2024

Type of Document (Tick): PhD Thesis | M.Tech/M.Sc. Dissertation | B.Tech./B.Sc./BBA/Other ✓

Name: SHRIYA, RAKSHITA JAIN, VIDUR Department: CSE Enrolment No 201272, 201462, 201467,

Contact No. 7973885947 E-mail. 201272@juitsolan.in

Name of the Supervisor: DR. KUSHAL KANWAR

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): EMOSENSE : HUMAN EMOTION DETECTION USING AUDIO AND VIDEO INPUTS

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

- Total No. of Pages = 47
- Total No. of Preliminary pages = 9
- Total No. of pages accommodate bibliography/references = 2

Shriya
(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 17 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Kushal Kanwar 15/05/2024
(Signature of Guide/Supervisor)

Signature of HOD

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
			Character Counts	
Report Generated on		Submission ID	Page counts	
			File Size	

Checked by
Name & Signature

Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com