Emotion Recognition From 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

Deepankar Singla (201148)

Rohan Sood (201232)

Under the guidance & supervision of

Mr. Aayush Sharma

Ms. Seema Rani

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 “**Emotion Recognition From 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 / Information Technology** 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 “**Deepankar Singla(201148)**” and “**Rohan Sood(201232)**” during the period from August 2023 to May 2024 under the supervision of **Mr. Aayush Sharma ,** Assistant Professor, Department of Computer Science & Engineering and Information Technology and **Ms. Seema Rani** Assistant Professor, Department of Computer Science & Engineering and Information Technology.

Student Name: Deepankar Singla	Student Name: Rohan Sood
Roll No.: 201148	Roll No. :201232

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

Supervisor Name: Mr. Aayush Sharma
Designation: Assistant Professor
Department: Computer Science & Engineering and Information Technology.
Dated:

Supervisor Name: Ms. Seema Rani
Designation: Assistant Professor
Department: Computer Science & Engineering and Information Technology.
Dated:

CANDIDATE'S DECLARATION

We hereby declare that the work presented in this report entitled **'Emotion Recognition From 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 **Mr. Aayush Sharma ,** Assistant Professor, Department of Computer Science & Engineering and Information Technology and **Ms. Seema Rani** 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: Deepankar Singla
Roll No.: 201148

Student Name: Rohan Sood
Roll No.: 201232

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

Supervisor Name: Mr. Aayush Sharma
Designation: Assistant Professor
Department: Computer Science & Engineering and Information Technology.
Dated:

Supervisor Name: Ms. Seema Rani
Designation: Assistant Professor
Department: Computer Science & Engineering and Information Technology.
Dated:

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor **Mr. Aayush Sharma, Assistant Professor**, Department of CSE Jaypee University of Information Technology,Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of "**Data Science**" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to **Mr. Aayush Sharma and Ms. Seema Rani,** Department of CSE, for their kind help to finish my project

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

Deepankar Singla (201148)
Rohan Sood (201232)

TABLE OF CONTENT

LIST OF FIGURES

FIGURE **PAGE NO.**

LIST OF TABLES

TABLES **PAGE NO**

LIST OF ABBREVIATIONS

	ABBREVIATION	DEFINITION
1.	**CNN**	Convolutional Neural Network
2.	**LSTM**	Long Short-Term Memory
3.	**DeepID V3**	Deep Identity V3 (a type of deep learning model for face recognition)
4.	**IEMOCAP**	Interactive Emotional Dyadic Motion Capture (a database for emotion analysis)
5.	**MELD**	Multimodal EmotionLines Dataset (a dataset for emotion recognition)
6.	**BoAW**	Bag of Audio Words
7.	**RNN**	Recurrent Neural Network
8.	**ADFES-BIV**	Audio-Visual Facial Expression Signals with Bilingual and Multi-culture Valence and Arousal Labels (a dataset for facial expression analysis)
9.	**DCNN**	Deep Convolutional Neural Network

ABSTRACT

The project’s focus is on developing technology so that it can be able to understand and interpret emotional messages conveyed in audio and visual inputs. Today, as we live in a digital era, it is significant for machines to understand and react with human feelings so that relationships between users and technology become more friendly, compassionate or personal.

This project aims at improving and updating the technical systems for capturing complex hidden details in psychological facial images of emotions. An elaborate examination of the intricacies behind the multi-dimensionality of emotional sensibility, which manifests in terms of audio pitches, facial expressions, and linguistics within video clips.

Focus of the project is on reducing the perception difference between emotional human beings and technological understanding. The project aspires to create superior algorithms and machine learning structures to realize this. The emergence of such technologies is very helpful for unraveling the layers of emotional cues that lie beneath audiovisual text.

The project aims at making machines understand the various types of human emotions and respond in a more meaningful way depending on the user's feeling. Ultimately this will change how humans interact with future technological interfaces through the revolution of a new era in which technologies adapt to humans’ finest feelings.

This undertaking goes beyond technology improvement with the possibility of man-machines interactions that will be conducted at a higher level of comprehension and feedback creating more meaningful and enriched transactions.

CHAPTER 1 : INTRODUCTION

1.1 INTRODUCTION:

In today's super-techy world, understanding how we feel when we're using all these gadgets is pretty darn important. Whether we're having a chat with someone, chilling out watching stuff, or even checking in on our mental well-being, using tech to figure out our emotions has become a pretty big deal.

But here's the tricky part: when these devices try to read our feelings from our faces and what we say in videos, it's like they're facing a real challenge. IThat’s what it feels like – as if we were trying to catch a bunch of small things which make each of us so different emotionally. That smile has likely made them scratch their heads in disbelief. “Was that one of those ‘I’m sad but tired smiles?” or “What does it mean when one smiles from exhaustion?”.

These super intelligent systems are doing their best, but it’s as if our feelings are one giant riddle they cannot solve. While they are struggling to catch those small signals which we send, they know that it’s not easy. Therefore, despite the fact that everybody wants tech to see us well at present, they squint their eyes to understand our emotional expressions via video and sound. It is no longer a small mistake that occurs nowadays, where these devices do not completely hit upon our feelings. This is just a huge wave that causes a worldwide impact on all the technical aspects. That’s almost as if they are not able to comprehend who we are but deeply; that’s huge!

Sometimes their interpretation of how we feel and what we require is not right, which makes it impossible for them to genuinely bond with us and understand us. In this sense, for example, they may fail to identify indications that an individual is depressed or finding it tough psychologically.. And when they miss those emotional cues, it's tough for them to respond in a way that feels just right. So, taking all these important bits into account, this project is all about diving headfirst into the challenge of figuring out our feelings from our expressions and words in videos. It's like this big mission to untangle the knots of understanding human emotions through the stuff we say and how we look.

The idea? To help these machines bridge that gap between what we're feeling and how they understand and interact with us. It's like we're trying to build this smoother, more understanding link between technology and our emotional side. If we can help tech "get our feelings better, it could totally change the game in how we relate to all these gadgets in our lives.

These super-smart gadgets are really trying their best, but understanding our emotions is like handing them this massive, mind-boggling puzzle. It's as if they're scratching their digital heads, trying to catch those tiny signs we drop about how we're feeling, but it's just not easy for them. So, even though everyone's buzzing about making tech sharper at understanding us humans, it's like they're squinting at a blurry picture, trying to see the full story of how we express ourselves through sound and video.

And hey, when these devices miss the mark on picking up our feelings, it's not just a little glitch—it's like this massive ripple effect that messes with the whole tech scene. It's like they're missing out on really getting who we are. It's as if they're standing at the edge, peering into this deep well of our emotions, but they just can't dive in deep enough to truly grasp how we're feeling, and that's a pretty big deal. When they can't quite crack our emotions, it's hard for them to genuinely connect with us or to understand what we're really going through. For example, they might miss out on clues that someone is feeling down or struggling with their mental health. And when they can't catch these emotional vibes, it's tough for them to react in a way that feels natural or right.

So, considering all this, it's a big deal. That's why this project or whatever that's tackling the challenge of deciphering our emotions through the things we say and how we look in videos is such a big deal. It's like this major mission to untangle the mess and help these machines bridge that gap between what we're feeling and how they perceive and interact with us. It's like trying to build this smoother, more understanding link between technology and our emotional side. If they can really nail our feelings, it could totally change how we vibe with all this high-tech stuff in our lives.

Consider this: It is difficult for them to truly connect with us or provide for our needs when they are unable to fully understand our feelings. For example, they might overlook indications that someone is depressed or not doing well mentally. Furthermore, it's challenging for them to react in a way that feels appropriate when they're unable to fully capture the feelings we're expressing.

In light of all these significant variables, the project's primary objective is on diving deeply into the specifics of interpreting our facial expressions and words in films to determine our moods. It is akin to this grand endeavor to comprehend the intricacies involved in identifying and comprehending human emotions through audio and visual media.

Sort of sorting out the knotty aspects of identifying and understanding our feelings is the aim. As in, educating the robots to truly understand us in our joyful, sad, and everything in between states. The goal is to assist these machines in bridging the gap between our emotions and their perceptions of us and their responses.

It's similar to attempting to create a more seamless link between technology and our emotional self. Our interactions with all these devices and gizmos in our lives may change significantly if we can make technology "get" our sentiments better.

1.2 PROBLEM STATEMENT:

One significant challenge in today's technological world is teaching machines to recognize and understand human emotions from audio and visual inputs. In order to close the communication gap between technological interpretation and human expression, this challenge is essential.

Though the technology today is not simple at all, it’s always hard to represent what makes humans emotionally complex using these systems. It's like giving these systems a complicated jigsaw puzzle made up of different pieces: tone of voice, facial expressions and the cues embedded within video clips. Despite being highly sophisticated, such systems often do not sufficiently represent the whole spectrum of human feelings because of the complexity of expressing emotions.

This becomes an obstacle when our tech systems fail to sense and understand our emotions properly, which is not just a minor hitch. This lack helps them not really be able to relate with us and as such forms shallow relationships that do not last for long. Therefore they have no idea what human emotions even mean, so their genuine communication and understanding is incomplete.

This creates an obstacle in their ability to effectively respond appropriately and empathically towards our emotional states. They may not even see such indicative signals as signs of distress, and therefore lose a chance to help or react appropriately.

Consequently, the main task at hand is to improve these systems' ability to decipher and understand the complexities of human emotions as they are presented in audio and visual formats. To encourage a more sympathetic and accommodating relationship between technology and users, a deeper comprehension of our emotional expressions is required.

1.3 OBJECTIVE :

In today's rapidly changing digital world, the project has served as a beacon for understanding, or how enabling technology can understand emotions from audio and video inputs. The ability of machines to accurately detect and react to human emotions is the foundation for developing a more personal and intuitive relationship between people and machines. This project's goal encompasses a wide range of concerns about enhancing the systems that currently link people's emotional expression through diverse technologies.

This project essentially addresses the flaws in current approaches that arise when trying to extract the subtleties that are present in human-emotion displays. The suggested method tackles the complexity of emotional recognition by taking into account linguistic, facial, and audio cues. Through improved understanding and a more sophisticated definition, the project aims to further delve into the technology's capacity to comprehend all facets of human emotions.

With this project, we would like to fully read the emotional signals in audiovisuals by fine-tuning sophisticated algorithms and advanced machine learning. The various technological developments are what provide technology the ability to comprehend and interpret human emotions with far more detail. The goal of the project is to give technology emotional intelligence so that it can recognize human emotions and respond to them appropriately.

Furthermore, the goal of this project is to lay the groundwork for drastically altering the way users interact with interfaces. More than anything, what we want from technology is not just the ability to identify human emotions, but also the ability to respond in a way that is consistent with our expectations. The anticipated shift in HMI heralds the end of the status quo because new technologies enable machines to become more interactive and sensitive, forming strong relationships with their partners.

However, the project's effects span several domains and are not just confined to technological advancements. It predicts increasingly intricate, emotionally intelligent human-machine interactions where technology is able to identify and respond to a wide range of nuanced aspects of human emotions. Users' experiences will be drastically altered by this evolution, leading to more meaningful, intuitive, and humanistic interactions between humans and machines.

Thus, by stepping into an emotion-driven audio-visual culture that will influence future emotional interactions between people and technology, this project aims to deepen technology's understanding of human emotions.

1.4 SIGNIFICANCE AND MOTIVATION OF THE PROJECT WORK:

1.4.1 ENHANCING HUMAN-MACHINE INTERACTION:

Constructing new parameters for communication between humans and machines. The way that people and technology interact will fundamentally shift as machines learn to perceive emotions and respond accordingly.

Imagine, for example, if technology went beyond conventional responses to commands and advanced to recognize the subtleties that underlie human emotions. Imagine an intelligent assistant who is able to decipher not only what is explicitly asked of them, but also the emotion that lies behind those words. As technology advances, it will be able to react suitably to spoken words as well as the emotions that accompany them. modifying technology to better comprehend human emotions in order to enable it to respond to them in a suitable and sympathetic manner. Consequently, depending on the situation at hand, technology may be able to provide pleasurable experiences, guidance, or support through emotion-sensitive responses. This implies that the way we view many technological products, like interactive voice assistants and responsive entertainment platforms, may change as a result of this potentially evolving evolution.

Anticipated advancements ought to play a major role in fostering intimate connections that will put everyone at ease when using technology. It presents a world where technology is sensitive to human needs, impacting socially similar human-to-human communication.

Ultimately, the goal of this evolution is to improve the quality of communication between humans and technology by creating new avenues for more engaging, intelligent, and sentient interactions with it.

1.4.2 IMPROVING USER EXPERIENCE:

Refined emotions are one step ahead in the way humans interact with technology in that they can precisely discern users' demands in technological places. It's expected that technology will completely change user interactions in certain important ways once it can recognize and respond to human emotions.

Imagine a user interface that can recognize and process commands, as well as the nuanced emotions contained inside them. Consider, for instance, a voice-activated gadget that recognizes whether the user is happy or angry and responds to it with the right amount of enthusiasm and empathy. When someone possesses this level of emotional intelligence, interactions become the most personal and intimate experience, giving the impression that they are speaking with a real person rather than a computer.

The individualized approach is crucial to improving user-technology relationships. Using technology to effectively interpret and react to people's genuine intentions or emotions increases their level of active engagement with a given system or product. On the other hand, tailoring responses based on feelings may significantly raise user satisfaction and result in happier interactions all around.

This is accomplished by technology, which reacts to emotions by making adjustments that create dynamic dialogue and a relatable setting that animates the exchange. Because the technology can sense and react to their emotions, improving overall user experience, users will grow to appreciate it.

This refined approach has the potential to completely change antiquated interfaces in humanistic, perceptive, and intelligent settings. It is anticipated that this update would improve consumers' interaction and satisfaction levels, which will enhance their use of contemporary devices.

1.4.3 ADVANCING MENTAL HEALTH AND WELL-BEING:

This suggests using technology to identify emotions in order to improve mental health. Using this technique that uses audiovisual clues as a possible pathway, such cases can be detected early.

Consider how technology might be able to recognize subtle variations in people's voice tones, facial emotions, or physical movements that are presented as audio and visual data. These could be early warning indicators of emotional troubles or even suggestions of feelings like tension or anxiety.

It is essential for timely interventions based on early identification. Hence, with this capability, technology could provide aid or support right away to people who seem in need. When the technology detects emotional cues and initiates connections that lead the user to appropriate mental health resources or care services, it can be employed, for example, in video conversations and other types of communication.

Additionally, during examinations, mental health doctors can benefit from this technology. In order to help with the development of tailored interventions, experts may learn more about the psychological condition of their patients by examining the emotional cues displayed during consultation sessions or remote sessions.

This kind of technology has far-reaching effects on mental health. By giving required support as soon as it's needed, early detection and intervention can help avoid mental health problems from getting worse. It is anticipated that this technology will improve the lives of those who are experiencing mental health issues by strengthening and promoting these groups.

1.4.4 ENABLING COMMUNICATION WITH DIVERSE AUDIENCES:

People are more likely to engage in successful cross-cultural discussions if they understand the emotions that are expressed in different cultural contexts. As it stands, this capacity transcends language barriers to foster genuinely communicative interactions and recognize various affect presentation modalities.

Consider how technology could be able to decipher the emotional cues buried in cultural specifics. For instance, using closed hands to convey sadness or smiling at the same time as expressing other emotions can also be accomplished by other hand movements, tones, and facial expressions. Since this technology takes into account the various emotional expressions that go along with these subtleties, it can aid in closing communication gaps.

Its capacity to interpret diverse emotional cues in a range of cultural contexts makes communication more inclusive. This enables technology to modify its replies so that users can comprehend them and relate to them on a culturally diverse level. Emotional recognition and response require a very high degree of sensitivity and flexibility, which fosters an environment that is more inclusive.

It can also be used to prevent misunderstandings brought on by cultural differences. It facilitates the accurate interpretation of emotional cues, which results in the successful communication of intentions and feelings and reduces the likelihood of cross-cultural miscommunication.

In particular, the ability of technical instruments to identify various emotional states is the key to having universally successful cross-cultural communication. It improves relationships between individuals from different cultural backgrounds and fosters mutual understanding and acceptance, which unites people and enriches global interactions.

1.4.5 TECHNOLOGICAL ADVANCEMENT AND INNOVATION:

In AI and machine learning, this is a step forward since it requires building models that can interpret nuanced emotional cues from visual or auditory inputs. In five ways, the project—which aims to introduce new technological trends—can be seen as the initial step toward these changes.

Imagine a technological advancement that surpasses the mere processing of audio and visual data to comprehend the nuanced emotions that are inherent in it. It involves teaching robots to interpret minute variations in tone, gestures, facial expressions, and other non-verbal cues that convey emotion. By enabling technology to understand such intricate emotional distinctions, this initiative aims to elevate artificial intelligence and machine learning to new heights.

This is a major advancement in artificial intelligence and the emergence of a deep layer of emotion perception, which has proven extremely challenging for robots to do thus far. The aim of this work is to increase the sensitivity of machine learning algorithms for the analysis of voice and visual signals that contain emotional information.

Nevertheless, this technical innovation's effects are not confined to certain domains. These opportunities include more emotionally intelligent and humanistic user interfaces as well as advanced AI-focused systems that are able to intuitively understand human emotion. Modern artificial intelligence and machine learning have the potential to drastically alter a variety of industries, including customer service, entertainment, and health care.

This technological advancement also marks a significant advancement in one of AI's several challenges: comprehending and interpreting human emotions. This study advances the development of models capable of deciphering intricate emotional cues, improving the technology that can adjust to human emotions.

In summary, this project aimed to explain how complex emotions in audio or visual input could be recognized by technology. Like any other industry, this one is leading the way in breaking new ground and developing more emotional intelligence, sensitivity, and responsive technologies. These developments have far-reaching effects on how people engage with one another in social contexts as well as on how they deal with people they interact with through machines.

1.5 ORGANIZATION OF PROJECT REPORT :

Chapter 1: Introduction

1.1 Introduction

Discuss briefly the significance of Emotion detection using audio and video and why it is a key way through which to understand human emotions via multimedia signals.

1.2 Problem Statement

The challenges and limitations of traditional emotion recognition techniques necessitate integrated audio-video approach to fully understand emotions.

1.3 Objectives

Describe the particular goals guiding the formation of the Emotion Detection tool that aims to correct errors associated with conventional emotion analysis systems.

1. 4 Significance and motivation behind the project.

What motivated the development of an Integrated Emotion Detection system? How does technology help us understand emotional responses through multimedia input?

1.5 Project Report Organization

Give a general introduction to the structure of the report outlining the following chapters which together contribute in detail to Emotion Detection using audio and video

Chapter 2: Literature Survey

2. 1 Overview of Relevant Literature

A review of the literature that has been carried out in the last few years, on studies, papers, or research related to Emotion-detection using Audio and Video input and forming the basics of the project.

2.2 Two key gaps in the literature.

This will include identifying gaps, limitations or issues that have not been tackled in existing studies and provide direction for the proposals suggested by the Emotion Detection system.

Chapter 3: System Development

3.1 Requirements and Analysis

Explicitly outline the first step comprising requirements and an analysis which led to the development of Emotion Detection.

3. 2 Project Design and Architecture

Describe the Emotion Detection system based on technologies and architectures; highlight employed tools, frameworks and methods for audio-video integrations.

3.3 Data Preparation

Talk about pre-processing of data which are received as audios or videos through the integration process in order to create an efficient emotion recognition algorithm.

3.4 Implementation

Offer technical information on implementing the system, including algorithms, code fragments, as well as tools utilized for audio-video Emotion Detection.

3.5 Key Challenges

Provide insight into the challenges that arose during the development of systems as well possible ways applied in mitigating these difficulties.

Chapter 4: Testing

4.1 Testing Strategy

Discuss how the testing procedure and methods were employed to verify the dependability on Audio and Video

4. 2 Test Cases and Outcomes

Also, use test cases that make the system strong and quality assured.

Chapter 5: Results and Evaluation

5.1 Results

Showcase the outcomes, analysis and appraisal of the performance of the Emotion Detection System that accurately detects emotions via audio-video inputs or any other mode.

5. 2 Comparison with Existing Solutions

Optionally compare the Emotion detection system against the existing solutions pointing out the advances and advantageous performance of the solution itself.

Chapter 6: Conclusions and Future Scope

6.1 Conclusion

What were our key findings, limitations, and major contributions during the creation of the Emotion Detection model?

6.2 Future Scope

Investigate possible extensions, improvements, and future uses of the system to make it adjustable and competent, especially in the domain of multimedia-based emotion recognition.

CHAPTER 2 : LITERATURE SURVEY

2.1 OVERVIEW OF RELEVANT LITERATURE:

S. No.	Paper Title [Cite]	Journal/ Conference (Year)	Tools/ Techniques/ Dataset	Results	Limitations
1.	Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video	2023	Tools: time-distributed CNNs + LSTMs, DeepID V3 + Xception architecture - Dataset: selected dataset for emotion recognition	Accuracy of recognition increased by almost 4% - Recognition accuracy reached 84.33%	N/A
2.	Voice-based Real-time Emotion Detection Technique with RNN-based Feature Modeling	2022	datasets: IEMOCAP and MELD Big-of-Audio-Words (BoAW) Recurrent Neural Network (RNN)	A weighted accuracy of 60.87% and an unweighted accuracy of 60.97%	Low accuracy of audio modality -based feature representations compared to text modality
3.	Speech Emotion Recognition Using Audio Matching	2022	Algorithm for sentiment classification in speech - Eigenvalue-based metric for selecting audio augmentations	Outperformed baselines in emotions classification by 10-20% - Improved sentiment classification in presence of noise	Pretrained audio feature extractors do not generalize well. - Some augmentations may result in a loss of accuracy.

Table 1 Literature Review

S. No.	Paper Title [Cite]	Journal/ Conference (Year)	Tools/ Techniques/ Dataset	Results	Limitations
4.	ViPER: Video-based Perceiver for Emotion Recognition	2022	ViPER: multimodal architecture for emotion recognition - Hume-Reaction dataset used for experiments	This indicates that ViPER is able to better estimate the scores associated to some specific emotion classes whereas struggles to classify other ones	N/A
5.	Emotion Recognition of College Students Based on Audio and Video Image	2022	Tools: Audio and video processing tools - Techniques: Emotion recognition algorithms and machine learning - Dataset: College students' audio and video recordings	Emotion recognition accuracy of audio and video images - Comparison of different emotion recognition algorithms	Limited sample size - Limited accuracy of emotion recognition
6.	Multimodal Emotion Recognition using Deep Learning	2021	Deep learning algorithms, such as multi-layer LSTM architecture, are used for multimodal emotion recognition.	It review of emotional recognition of multimodal signals using deep learning and compares their applications based on current studies.	The paper does not provide specific numerical results or accuracy rates for the multimodal emotion recognition systems discussed
7.	Multimodal Emotion Recognition in the Wild Challenge	2020	Tools random forests and state-of-the-art deep neural networks	Approach details for Emotion Recognition in the Wild Challenge tracks.	Paper lacks in-depth analysis of implementation challenges

S. No.	Paper Title [Cite]	Journal/ Conference (Year)	Tools/ Techniques/ Dataset	Results	Limitations
8.	Facial Emotion Recognition from Videos Using Deep CNNs	2019	deep convolutional neural network ADFES-BIV dataset	Paper employs DCNN via TensorFlow for video-based emotion recognition, achieving 95.12% accuracy for ten emotions,	Paper lacks time details; results differ from literature due to dataset variations; limited dataset use; no audio;
9.	Affective Computing: A Review of the State of the Art	2018	Affective computing spans psychology, physiology, and computer science	Paper reviews affective computing's applications, technologies, challenges, and future directions.	Paper lacks detailed analysis, empirical studies, dataset/tool specifics, and ethical considerations; potential limitations include scope and depth.
10.	Automatic Speech Emotion Recognition using RNNs with Local Attention	2017	Deep learning Recurrent Neural Networks IEMOCAP	Deep recurrent neural networks with local attention improve speech emotion recognition compared to existing methods.	Local attention mechanisms may require careful hyperparameter tuning. - Limited discussion on dealing with noisy speech data.

Table 1 Literature Review

Yongjian Cheng et al [1] This paper presents a new emotion-recognition algorithm that utilizes multimodal approaches in order to surpass the shortcomings of monomodal recognition techniques. The single-mode recognition loses its ability both in terms of accuracy and comprehensiveness due to a large scale of data. Multimodal methods applied for increasing the recognition accuracy, and pre-processing of a data set. It builds the models in the context of audio and video, each through different build modes, compared to existing methods for validation purposes. Furthermore, it considers late fusion methods and introduces a weight adaptive late-fusion approach that improves accuracy by almost 4%, bringing up the total recognition accuracy to 84.33%.

Sadil Chamishka1 & et al[2] The approach uses a novel eigenvalues-based metric along with optimum augmentations to improve training data. As a result, there has been a ten to twenty percent improvement over conventional methods of evaluating emotions on YouTube. Neural cells in particular are capable of comprehending words with similar sounds and meanings. In addition to being adaptable for phrase analysis, the model can also distinguish birds in the recorded urban soundscape. The multidimensional technique discussed above exhibits great promise in terms of the possible impact magnitude for various scenarios within the auditory analysis domain, demonstrating its adaptability and versatility.

Iti Chaturvedi , et al [3] This paper outlines an effective way of sentiment classification in speech that is fit to TikTok or YouTube use as it helps buyers make wise purchases decisions. Tones of speech and their impact in the sentiment determination, in languages such as Spanish where the translation is very uncertain. To that end, it proposes a new sentiment classification approach in the context of noisy environments. The traditional approach based on pre-trained, handcrafted audio features have difficulty handling accented spoken speech. The classification of emotions is achieved by using an emotional concept vector which considers related words carrying the same prefix for the same meaning.By choosing the best augmentations, it enhances training data and suggests a new metric based on eigenvalues. Their impact is demonstrated by the 10–20% improvement in evaluations for YouTube emotions over baselines. Given that they can all comprehend words with similar sounds and feelings, the brain units are adaptable. Moreover, the model may be applied not only to phrase analysis but also to the recognition of birds from audio recordings made in cities. This illustrates the versatility of this multifaceted technique and suggests that it may have a substantial impact on many audio analysis scenarios.

Lorenzo Vaiani, et al. [4] Multiple inputs, including text, voice, and images, are used to perceive human emotions in videos. However, in order to fuse the data sources collectively, due to their disparate nature, particular fusion networks are required. to develop ViPER, a new transformer-based paradigm for assessing how people feel about various video clips in terms of their emotional states. VIPE is able to integrate all of the modalities—text, audio, and video—together thanks to laten fusion networks. Experiments conducted inside the MuSe-ReAction challenge using the HumeReAction datasets have confirmed the validity of this technique.

Chenjie Zhu , et al. [5] In order to help college students improve their mental health, it is critical to recognize emotional issues in them. This article presents a deep learning based multi-modal emotion identification technique using audio and video pictures. Lastly, the emotional elements for voice identification are extracted using an attention-based Long Short Term Memory network. In the second section, features are extracted using a modified local binary patterns operator and principal components analysis, and then a transfer learning process utilizing the VIFFN-VGG-16 network model is performed for video image recognition.5. Thirdly, there is a layer of decision-making that groups multi-modal emotions by combining emotional data from each single modality. Compared to single modal approaches, the Chinese emotion database Cheavd2.0 has a higher recognition rate. Therefore, this approach has the potential of correctly recognizing and handling negative emotions in a college student to boost emotional growth.

Sharmeen M.Saleem Abdullah, et al.[6] In order to create relationships between people and computer systems that are natural, recent research in human-computer interaction aims to take users' emotions into account. Things like health and education could be improved by this progress. The writers are investigating several methods of obtaining data from human emotions, such as facial expressions, bodily responses, and brain imaging. The study offers a thorough analysis of the literature on the topic of deep learning-based multimodal signal emotion recognition, along with a comparison of the results obtained thus far. It demonstrates that the multimodal affective system improves classification accuracy and functions more effectively than a single mode of emotion-based processing. This research states that the amount of observed emotions, extraction techniques, classifiers, and data consistency are among the aspects that determine accuracy.It also goes over how to identify emotions and the most recent advancements in the field of emotional psychology, highlighting the necessity of looking at consciousness and physiological indicators. In order to spur further research in this area that would offer more details on the topic at hand, the current study examines emotions in detail, including their identification and effects on various facets of life.

Shivam Srivastava ,et al, [7] The eight Emotion Recognition in the Wild Challenge (EmotiW 2020) has four courses, all of which they cover in this study. These tracks include the following: predicting involvement in natural environments, identifying group emotions, predicting driver gaze, and identifying emotions using physiological cues. Several approaches are used in our investigation, from state-of-the-art deep neural networks to traditional machine learning instruments like random forests. To find out how well these methods work in these different tasks, we also test fusion and ensemble-based approaches. They cover all four of the courses in the eight Emotion Recognition in the Wild Challenge (EmotiW 2020) in this study. Predicting engagement in natural situations, recognizing emotions in a group, anticipating driver gaze, and identifying emotions based on physiological indicators are some of the songs in this collection. Their study employs multiple methodologies, ranging from cutting-edge deep neural networks to conventional machine learning tools like random forests. They also test fusion and ensemble-based approaches to see how well these techniques perform in these various tasks.

Wisal Hashim Abdulsalam, et al [8] In this study, they explore how important it is to decipher facial expressions to understand human emotions, especially in the context of human-computer interaction (HCI). In this sector, facial cue-based emotion identification has long been a challenge. Their research focuses on using Google's TensorFlow machine-learning toolkit in tandem with a deep convolutional neural network (DCNN) [11]. They specifically investigate the use of this DCNN model for video data facial expression identification.Ten different emotions from the Amsterdam Dynamic Facial Expression Set-Bath Intensity Variations (ADFES-BIV) dataset were analyzed in the experiment using the TensorFlow library. Two different datasets were used for testing in order to confirm the model's efficacy. The purpose of this work is to clarify how well deep learning methods—in particular, DCNNs—perform when it comes to facial emotion recognition from video inputs.

Jianhua Tao & Tieniu Tan [9] This paper explores the field of affective computing, which is gaining significant attention because of its potential applications in a variety of fields, including smart surveillance, virtual reality, and perceptual interfaces. Numerous academic fields, including computer sciences, physiology, psychology, and cognitive science, are integrated by affective computing. Several implicit difficulties that are important to affective computing are addressed in this work, which emphasizes the interactive feedback loop. It provides a perceptive assessment of the state of the field today by closely examining the several approaches used to address each of these problems. Furthermore, the report identifies research challenges and suggests future directions in addition to exploring current approaches. It seeks to provide a thorough grasp of affective computing by going over these points and highlighting the urgent need for more study and advancement.

Seyedmahdad Mirsamadi, Emad Barsoum ,et al [10] This work explores the challenging field of automatic speech-to-emotion recognition, a laborious operation that depends on the effectiveness of speech features used for categorization. This work investigates the use of deep learning methods to automatically extract emotionally relevant information from speech signals. The study uses a deep recurrent neural network to show how the network can learn both short-term, frame-level auditory characteristics that are important for emotions and combine these into a succinct representation at the utterance level. Furthermore, a novel feature pooling strategy is presented that uses local attention to focus on particular speech signal segments with greater emotional weight. The IEMOCAP corpus is used to rigorously evaluate the suggested methodology, which shows better predicted accuracy than other emotion detection systems already in use. This strategy shows promise for developing the area by improving the accuracy of speech signals used for automatic emotion recognition.

2.2 KEY GAPS IN THE LITERATURE:

Within the research titled "Emotion Recognition using Audio and Video Inputs," a discernible void appears with regard to the multimodal emotion detection system's scalability and real-time applicability.

Real-time Applicability: A thorough examination of the system's applicability in real-time circumstances is lacking in this research. Instantaneous emotion recognition is essential for many applications, such as mental health monitoring and human-computer interaction. For the system to be applicable, it is essential to comprehend how quickly it detects emotions in continuous audio-visual streams—aspects such as processing speed, latency, and its capacity to comprehend quickly shifting emotional cues.

Scalability :The paper also fails to consider an important factor: the scalability of the system. How effectively does it function with big datasets and different levels of complexity? In order to handle a wide range of emotional expressions, deploy this technology across platforms, and handle future increases in data volume, scalability is essential. Deploying it in the real world requires an understanding of its adaptability to changing demands and input sources.

Despite concentrating on Spanish-language videos, the research on voice emotion recognition is not extensive enough to evaluate other languages. This discrepancy restricts our comprehension of the algorithm's performance outside of the Spanish language setting. Given the notable differences in emotional expressions between languages and cultures, assessing its efficacy in a variety of languages is essential. A comprehensive evaluation in several languages would provide information about the algorithm's flexibility and performance in various linguistic contexts. Understanding the algorithm's advantages, disadvantages, and difficulties in handling various linguistic subtleties requires this review. In the end, this strategy would guarantee the algorithm's wider application and efficacy across a variety of linguistic and cultural contexts.

There is a dearth of thorough investigation on ViPER's adaptability to a broad range of emotional circumstances in the study on multimodal emotion recognition. Furthermore, there is a lack of knowledge regarding ViPER's effectiveness while handling a variety of data sources. Comprehending ViPER's flexibility in diverse affective settings is essential. Emotions are quite complex and differ greatly depending on social, cultural, and personal factors. Further research on ViPER's response to and interpretation of this emotional variance would shed light on how resilient and reliable the system is at accurately recording a wide range of emotions. Moreover, there is still little research on how well ViPER works with fluctuating data sources. The identification of emotions may be impacted by the differences in characteristics and nuances between various data sources, including textual annotations, audio recordings, and video frames. ViPER's adaptability and possible drawbacks would be brought to light by assessing how well it works with these various data sources; this evaluation would help shape future developments and guarantee ViPER's effectiveness with a range of modalities.

Particularly when it comes to speech and video picture recognition, there is a dearth of thorough comparison metrics between single-modal and multi-modal techniques in the research on multi-modal emotion recognition utilizing deep learning techniques. It is essential to comprehend how much more effective multimodal approaches are than single-modal ones.

Although the study concentrates on using both voice and video image modalities for emotion recognition, it is unable to provide a strong framework for comparison. There are no metrics that specifically point out the advantages and disadvantages of single- and multi-modal approaches. An in-depth analysis would compare and contrast the performance metrics of voice or video pictures as individual modalities with their combined multi-modal recognition. Metrics like recall, accuracy, precision, F1-score, and confusion matrices would highlight the relative advantages and disadvantages of each strategy. This lack of information makes it more difficult to fully comprehend the advantages or disadvantages of using multi-modal approaches compared to single-modal ones. A more comprehensive comparative study would greatly develop emotion identification systems by providing insightful information about either method, single-modal or multimodal is better at extracting and identifying emotions from audio and visual picture input.

A complete investigation of the consistency and performance variances across different emotion recognition databases is lacking in the research on deep learning algorithms for emotional recognition across multimodal signals. A comprehensive knowledge of the effectiveness and consistency of the suggested deep learning approaches is hampered by the lack of a thorough investigation across several databases. Several databases are frequently used to assess emotion detection systems; each database has unique traits, intricacies, and subtleties in emotional expressions. The variability and consistency of the suggested deep learning approaches when used across these various databases are not sufficiently explored in this research.Assessing the performance measures for each database, including recall, accuracy, precision, and specificity, would be a comprehensive analysis spanning several databases. This evaluation would show how effectively the suggested deep learning techniques generalize and function consistently across various datasets of emotional expression. Comprehending the disparities in performance among diverse databases is imperative in evaluating the flexibility and resilience of the suggested methodologies in authentic situations. An in-depth assessment and validation of the suggested techniques would be possible with the use of this analysis, which would reveal information about the system's dependability and efficacy in identifying emotions in a variety of situations.

Gaps in the Emotion Recognition in the Wild Challenge (EmotiW 2020) are mostly related to the scant talks regarding the methodologies' usefulness in real-world circumstances and their generalizability across other tasks. A significant omission is the lack of thorough discussion of the flexibility and applicability of the suggested methods to various tasks within the challenge.

Group emotion identification, driver gaze prediction, engagement prediction, and physiological signal-based emotion recognition were among the tasks included in EmotiW 2020. It is possible that the research does not provide a thorough understanding of how the suggested approaches and algorithms could be easily expanded upon or modified to fit a variety of different jobs. To ensure that these approaches are strong and applicable to a wide range of scenarios, it is essential to comprehend how these approaches may be scaled and transferred across diverse activities.An additional deficiency is the scant investigation of probable obstacles encountered in practical situations. Although there appears to be little discussion of the difficulties and constraints in practical implementation, the research may yield encouraging results in controlled environments or with particular datasets. The efficacy of these methods may be impacted by variables like data heterogeneity, environmental fluctuations, and the dynamic character of real-world events. A thorough investigation of these issues would highlight the shortcomings and potential areas for progress, encouraging the creation of stronger and more workable solutions.

There are some gaps in the field of facial emotion recognition with deep convolutional neural networks that need to be investigated further, according to the research. A notable omission is the lack of detail in the discussion of the real-time application performance of the model. Although a deep convolutional neural network (DCNN) for facial emotion recognition is presented in the research, a thorough assessment of the model's performance in real-time scenarios is absent. Understanding the model's latency, responsiveness, and accuracy under real-time limitations are essential since real-time applications frequently require quick and accurate detection. Thoroughly examining the behavior of the model in real-time settings would yield significant insights into its practical usability and steer future enhancements aimed at optimizing its performance for real-time applications. One other noteworthy void is the scant attention paid to the model's cross-dataset resilience. The study may demonstrate the model's performance on a particular dataset in a controlled setting. Nevertheless, there isn't much information available regarding its effectiveness and generalizability over other datasets, differences in face expressions, lighting, or demographic characteristics. To confirm the model's applicability in many situations and guarantee its performance beyond specific scenarios, it is imperative to assess its robustness over a variety of datasets.

The review study highlights a gap in the field of affective computing that needs more investigation. A notable deficiency is the requirement for a more thorough examination of particular uses and real-world applications of affective computing in a variety of domains.

The review article recognizes the wide range of potential uses in fields such as smart surveillance, virtual reality, and perceptual interfaces, but it does not go far enough in examining and analyzing actual applications and their effects in these fields. Affective computing has enormous promise in a variety of domains, providing chances to transform human-computer interaction and propel a number of companies forward. Nevertheless, a thorough examination of the practical use of affective computing in these industries is absent from the article. Furthermore, the paper's capacity to demonstrate the revolutionary impact and tangible benefits that affective computing can offer to various industries is limited by the lack of particular examples or case studies showing the technology's practical application in real-world circumstances. A more thorough examination of successful applications, difficulties encountered, and solutions would offer insightful information about the viability and efficacy of incorporating affective computing into other fields.

For Automatic Emotion Recognition from Speech, a significant gap in the discussed methodology is identified. The void is in the scant discussion of how the suggested approach can be applied to a variety of languages and speech patterns outside of the particular IEMOCAP corpus that was used for the research. Although the study explores the nuances of emotion recognition from speech, the suggested method's scope and application in real-world scenarios including different languages, dialects, and speech patterns are limited by its primary focus on a single dataset. Particularly when it comes to speech cues, emotion identification is a complicated and diverse process that is frequently impacted by regional speech patterns, linguistic quirks, and cultural differences. Although the IEMOCAP corpus is an invaluable tool, its exclusive use in this study presents challenges for evaluating the generalizability of the technique in various linguistic and cultural situations. The research would be much more thorough if it included a larger variety of datasets that represented different languages, dialects, and speech patterns in its discussion. A more comprehensive analysis spanning several datasets with different languages and speech properties would provide information about how flexible and reliable the approach is in extracting emotions from speech outside of a particular dataset.

A significant vacuum exists in the field of Deep Learning-Based Emotion Recognition in Speech when it comes to a thorough assessment that takes into account a variety of emotional expressions and speech traits. The primary focus of the research is on using deep learning techniques for emotion recognition; nevertheless, the evaluation appears to be restricted to a small range of speech patterns and emotional expressions. Emotions are complex and multifaceted, differing from person to person and impacted by contextual, cultural, and individual factors. But the model's assessment seems constrained to a particular group of emotional expressions or limited speech features, which might miss the model's effectiveness over a wider range. For a more complete assessment of the model's performance, a larger range of emotional states and speech patterns should be included. The model's validation is restricted and its usefulness in real-world circumstances is hampered by the lack of different emotional manifestations, such as subtle subtleties, fluctuating intensities of emotions, or complicated emotional states. Additionally, evaluating its effectiveness in various cultural or demographic contexts would provide insightful information about its generalizability.

To gain a more comprehensive knowledge of the model's advantages and disadvantages, the evaluation could be expanded to encompass a wider range of emotional expressions and speech features. Its efficacy and dependability in a range of affective circumstances will be determined by this thorough evaluation, which will guarantee its application in real-world situations where emotions take on complex forms.

CHAPTER 3 : SYSTEM DEVELOPMENT

3.1 REQUIREMENT AND ANALYSIS :

3.1.1 SOFTWARE REQUIREMENTS:

1. **Kaggle Notebook:** One of Kaggle's most popular coding environments is Kernel, also known as interactive coding environment. Kaggle Notebooks builds its foundation on the popularly used Jupyter platform, where you can write and run code together with other team members which is more convenient and easier to share your codes.
2. **Jupyter Environment :** Kaggle provides a runtime in the form of a jupyter notebook environment, which supports languages like Python and R suitable for data analytics, visualization and machine learning exercises.
3. **Amazon Web Services (AWS) :** It is a virtual service provider. This implies that users can access computational resources, storage, and databases with minimal efforts. These allow companies and individuals to host websites, run applications, and carry out data storage at a low cost and provide reasonable options instead of expensive servers. AWS ensures that the level of complexity in the IT environment is manageable for small-sized users.
4. **Python :** Several others from this group also like to use Python, a commonly used programming language. Thanks to its syntax which is friendly to users and rich in libraries, it finds usefulness in machine learning and data science.

3.1.2 LIBRARY REQUIREMENTS:

1. **NumPy :** NumPy[12] is a crucial library in numerical computing. It allows for easy operations on large multidimensional array and matrix. NumPy makes many such jobs such as linear algebra, statistical analysis and even calculations involving mathematical functions. It is an indispensable tool for executing scientific and data tasks in python.
2. **Pandas :** Pandas[13] for Python is very powerful and convenient when it comes to dealing with databases and statistics. It has structured data objects like DataFrames. Pandas has clean, examine, and transform functions that are critical in data cleaning, data exploration, and transforming the data which are essential in data processing and performing analytical tasks.
3. **Matplotlib :** Matplotlib[14] is one of such python libraries often used in order to make it easier to construct static or dynamic, even interactive picture plots. Plotting enables an individual

to view data in both two and three-dimensional forms. The application of Matplotlib which uses an easy interface frequently is used for both exploration and presentation of data.

4. **Seaborn :** Data Visualization tools like Seaborn[15] , which was developed as a wrapper for Matplotlib and Python-based. This instrument is user-friendly because it assists in making presentable and significant bar graphs showing data statistics. In order to achieve this, sea born is equipped with heat maps and violin plots as a simple and clear aesthetic.
5. **TensorFlow :** Programming has been made powerful by the use of TensorFlow[16]. It provides a platform for data science application development. It offers an organized setting that makes it possible to develop complex models and applications in a timely manner. The users' tasks are made simpler by the availability of many functionalities and tools on this platform during the development processes.
6. **Keras :** Keras[17] is a Python based API, which is used to write neural networks. This is an entry point to the TensorFlow library where we can define and run our deep learning models. This is a straightforward, modular approach that facilitates rapid prototyping and suits well with the datascience folk. Keras has been designed to be a painless tool, through which individuals can readily build and train neural networks for several objectives.
7. **OpenCV :** OpenCV[18] is basically a tool with multiple uses that may include computer vision, image processing, etc. It is a set of operations that are useful in image and video analysis as well as feature extraction and machine learning. Using this platform, one can use a vast range of programming languages starting from very basic image modification to more sophisticated computer vision projects. The user-friendly interface and comprehensive handbook make this software widely used by programmers and scientists of all ranks.
8. **Sklearn :** The most user friendly machine learning library for python is sklearn[19] or scikit-learn. It helps in several phases of machine learning ranging from classification, regression, clustering and also dimensionality reduction. The sklearn interface is user friendly and supports the implementation of different algorithms. Several reasons make it preferred. Firstly, it's because it is easily understandable by newbies and, secondly , it offers proper documentation for professional users.

3.1.3 HARDWARE REQUIREMENTS:

1. **GPU P100 :** In machine learning, a GPU is the most essential hardware element. It is superior to other processors and especially with regard to parallel processing where it conducts multiple calculations simultaneously thereby supporting complicated ML algorithms. Training and model convergence on GPUs is necessary for deep neural networks

to obtain better results. They accelerate the training and deployment of the machine learning algorithms. NVIDIA's high-performing GPU is called the GPU P100 or NVIDIA Tesla P100. Its computational capacity is crucial and as such, it can undertake complex scientific simulations, deep learning, and artificial intelligence. It is capable of running in parallel mode which improves the speed and reliability at which the data intensive tasks are executed giving fast and dependable computational performance.

2. **Web Cam :** We utilize a webcam to record live video, this video is then processed and sent through the model for accurate predictions in real time.
3. **Microphone :** We utilize the computer microphone in order to record live user audio, this audio is then processed through the audio model to make accurate predictions.

3.2 PROJECT DESIGN AND ARCHITECTURE :

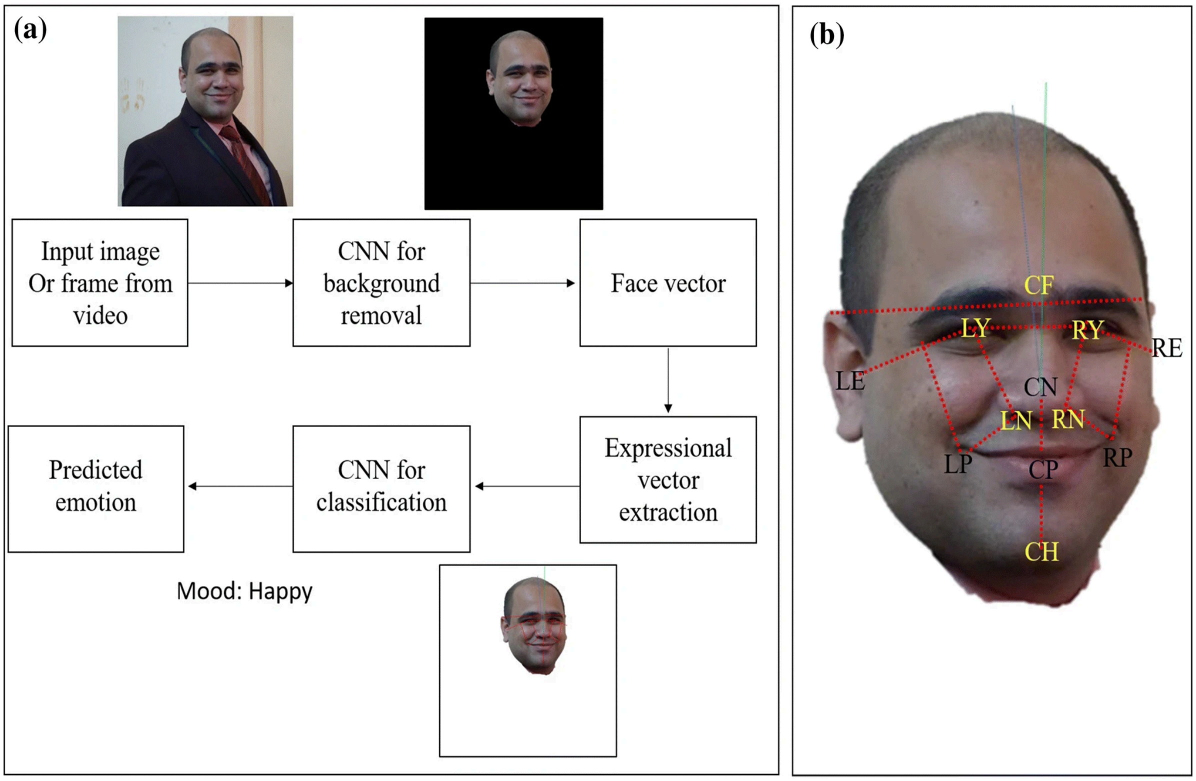


Fig. 1(a). Model architecture.

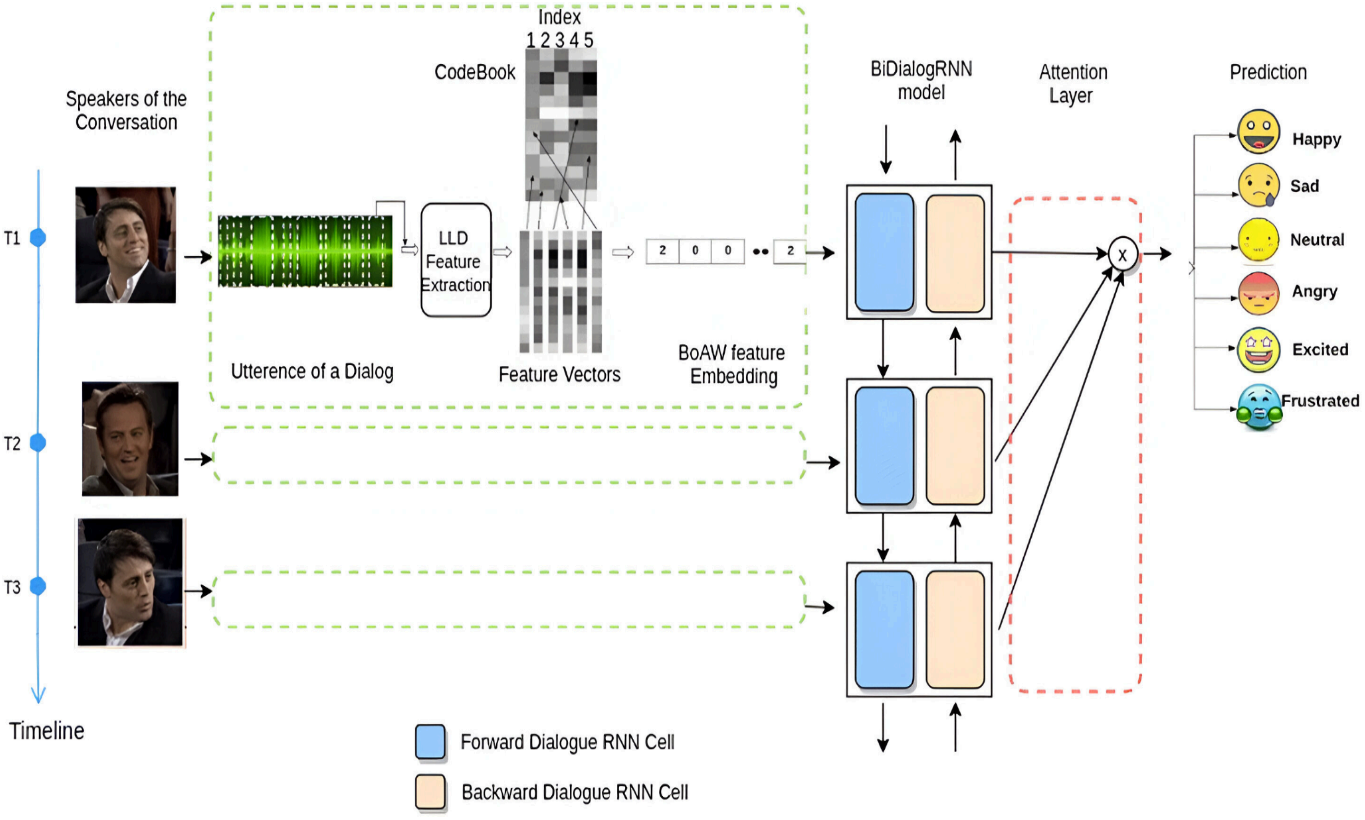


Fig. 1(b). Model architecture.

3.3 DATA PREPARATION :

1. **FER 2013 Dataset :** The fer2013[20] is a dataset of face images for training and testing emotion recognition models. It features grayscale images depicting faces expressing seven fundamental emotions: This includes emotions like anger, disgust, fear, happiness, sadness, surprise, and neutral. These images have their dimensions normalized to 48×48 pixels each so that they can be comparable for analysis. Originally created for the "Challenges in Representation Learning: For instance, the challenge known as “Facial Expression Recognition Challenge” that was once available on Kaggle webpage and later turned out to be very valuable in terms of machine learning studies.

The Fer2013 has the primary purpose of enhancing the algorithms that give accurate readings of human emotions from facial signs. The dataset that enables researchers and developers constructing models towards building intelligent artificial intelligence systems emotionally intelligent. Exposing the models to the facial expressions allows for the creation of more emotive AI applications and in turn this dataset helps the same. An easy to understand and straightforward model that could be explored by an individual interested in trying the emotion recognition machine learning model.

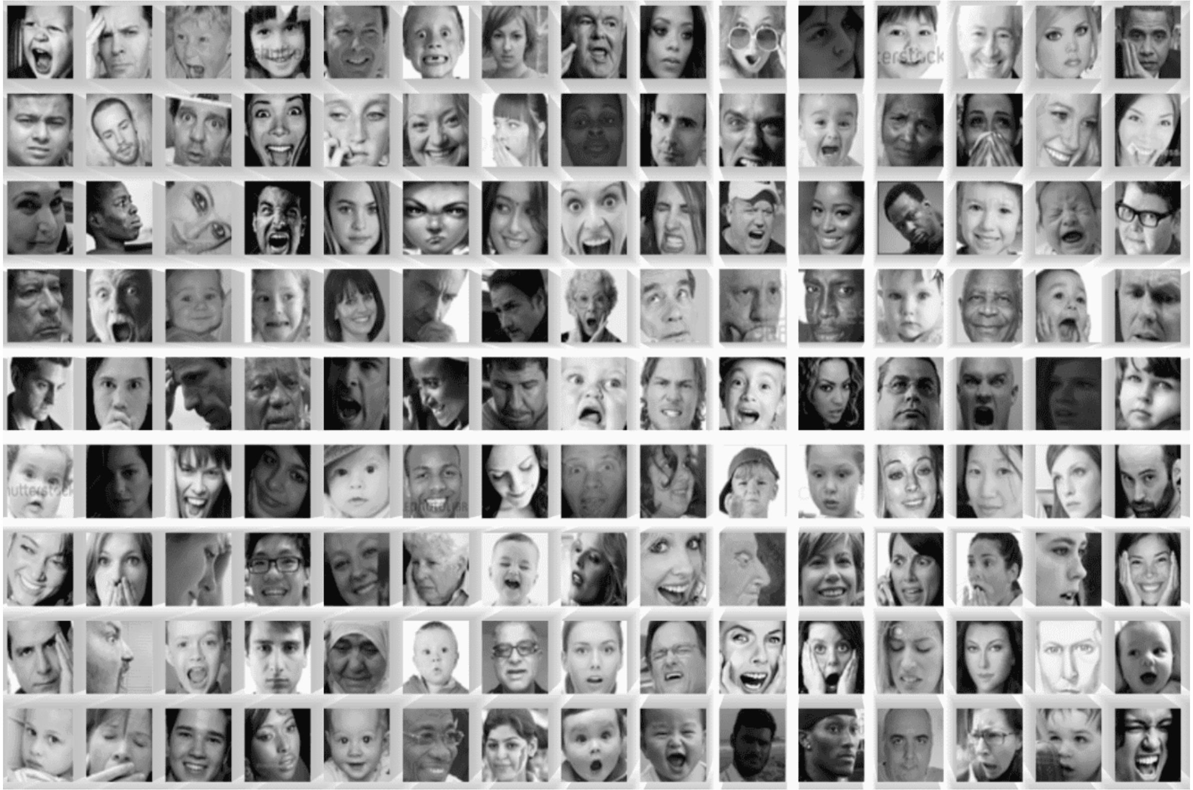


Fig. 2. Sample images from FER2013 dataset.

2. **RAVDESS Dataset :** The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[21] is made to study emotion recognition and comprises of audio-visual data. It includes acts of spoken speech and song including but not limited to neutral, calm, happiness, grief, anger, terror, surprise, revulsion, disgust. The recordings involve 24 professionals across RAVDESS in terms of age, gender, and ethnicity. It comprises each participant singing out in Canadian English with being a critical data for investigations on emotional expressions in different lingual and cultural settings. Data set is well laid out with simple file nomenclature. As such, researcher and practitioner friendly. RAVDESS is an open access resource for researching affective computing, human-computer interaction and machine learning. This tool is highly valuable in the development of artificial emotions, whereby it is possible to identify and respond to speech and song.

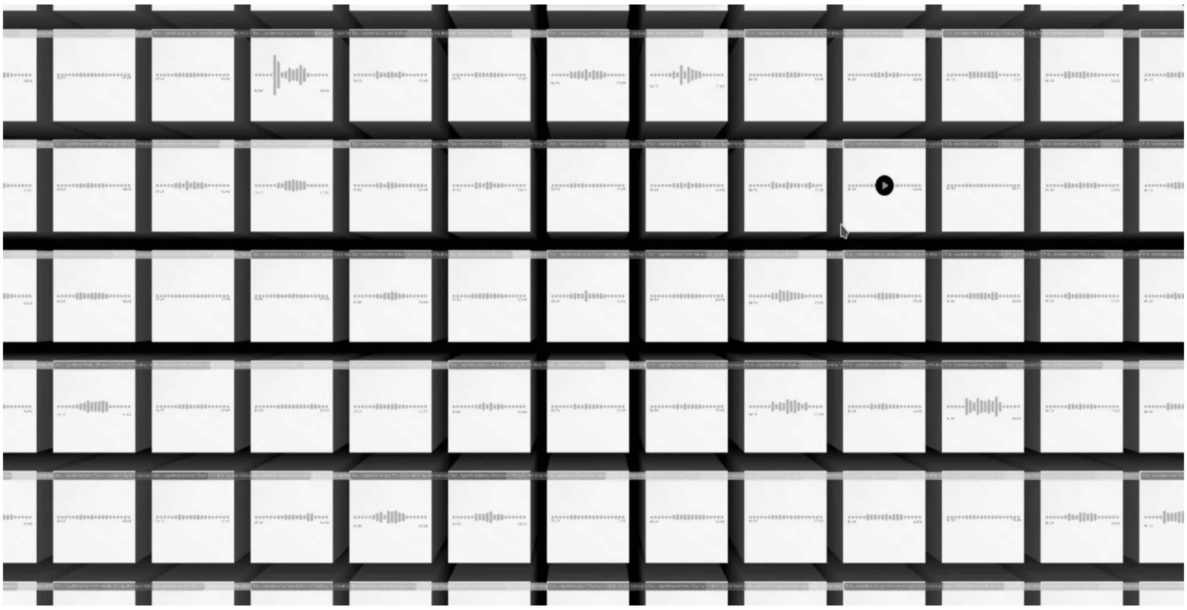


Fig. 3. Various audio samples from RAVDESS dataset.

3. **TESS :** Toronto Emotional Speech Set or TESS[22] is one of such datasets that seek the development of research on emotional speech analysis. TESS is comprised of a collection of vocalizations from 200 trained actors that cover different emotions including happiness, sadness, anger, fear, surprise, and disgust. Emotional expressions are enacted by each actor in a script, giving a sense of authenticity to different affective states. The TESS is designed to be easy to use as well as a significant support for emotion recognition, speech processing studies, and artificial intelligence. The simple organization of the dataset helps researchers and practitioners who can leverage it to make headway in the domain of emotion-aware technologies as well as machine learning models.

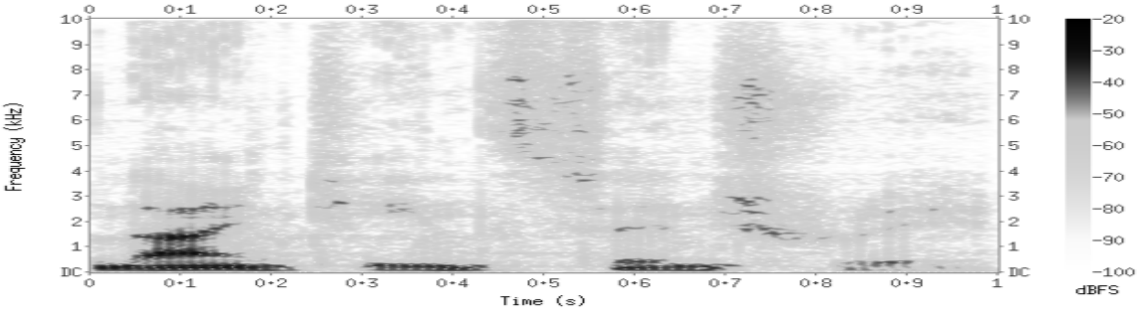


Fig. 4. Audio Spectrograph

4. **CREMA-D :** The CREMA-D[23] (CrowdEmotionMAnual Dataset) comprises of different emotional speech recordings meant for research and analysis. The dataset comprises more than 7,000 samples from over 90 professional actors covering various emotions, such as

happiness, sadness, anger, fear, and surprise. Every actor gives different shades of expression depending on the intensity of such expression. The distinguishing factor about CREMA-D is the manual annotation by human evaluators who indicate the emotional content and offer reliable ground truth data. CREMA-D helps researchers and practitioners in understanding complexities of emotional communication and is used to strengthen the studies on emotion detection, speech processing and machine learning applications because of its depth and authenticity. Enhancement of emotion technology is greatly influenced by the dataset's accessibility and richness.

5. **SAVEE :** SAVEE[24] is an important collection of audio-visual material for investigating emotional aspects of research. SAVEE is a collection of facial expressions of four male actors portraying seven basic human emotions which include sadness, happiness, surprise, disgust, fear, anger and a neutral facial expression. Despite these, this dataset is unique since the researchers are able to investigate interaction of facial expressions and vocal cues on emotional communications. SAVEE is a worthwhile tool in emotion's recognition research as well as the improvement and testing of algorithms in multimodal understanding of human emotions.

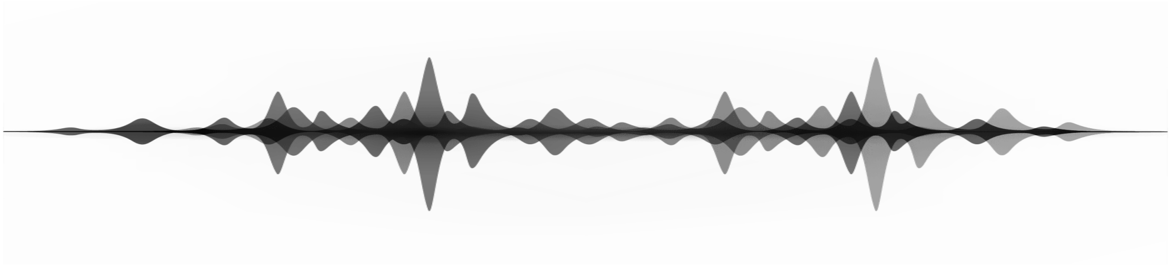
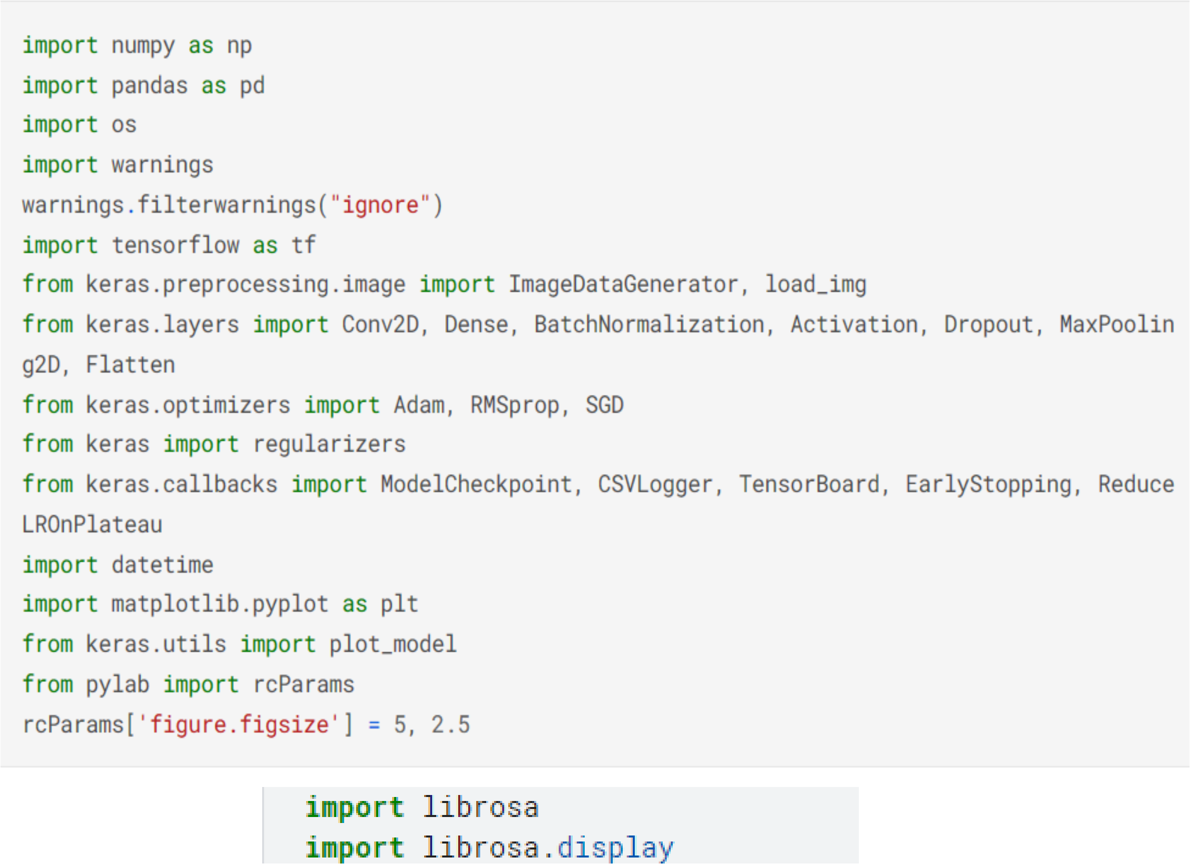


Fig. 5. Audio signal.

3.4 IMPLEMENTATION :

Below is the implementation of the video based recognition model:

Importing the necessary libraries:



```
import numpy as np
import pandas as pd
import os
import warnings
warnings.filterwarnings("ignore")
import tensorflow as tf
from keras.preprocessing.image import ImageDataGenerator, load_img
from keras.layers import Conv2D, Dense, BatchNormalization, Activation, Dropout, MaxPoolin
g2D, Flatten
from keras.optimizers import Adam, RMSprop, SGD
from keras import regularizers
from keras.callbacks import ModelCheckpoint, CSVLogger, TensorBoard, EarlyStopping, Reduce
LROnPlateau
import datetime
import matplotlib.pyplot as plt
from keras.utils import plot_model
from pylab import rcParams
rcParams['figure.figsize'] = 5, 2.5
```

```
import librosa
import librosa.display
```

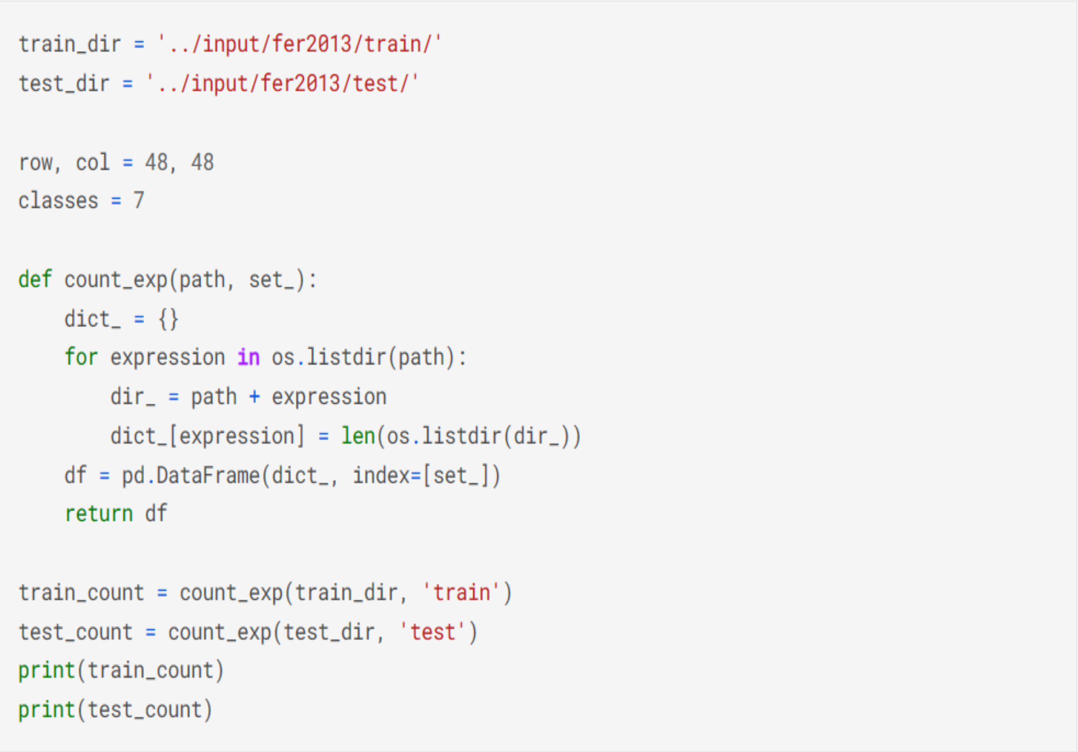
Fig. 6. Code for importing libraries.

1. import numpy as np: Imports NumPy library and assigns the alias 'np'.
2. import pandas as pd: Imports Pandas library and assigns the alias 'pd'.
3. import os: Imports the 'os' library for interacting with the operating system.
4. import warnings: Imports the 'warnings' module for handling warnings.
5. warnings.filterwarnings("ignore"): Ignores warning messages during code execution.
6. import tensorflow as tf: Imports TensorFlow library and assigns the alias 'tf'.
7. from keras.preprocessing.image import ImageDataGenerator, load_img: Imports specific modules from Keras for image data preprocessing.
8. from keras.layers import ...: Imports various layer types from Keras for building neural networks.
9. from keras.optimizers import Adam, RMSprop, SGD: Imports optimization algorithms for training neural networks.

10. from keras import regularizers: Imports regularization techniques from Keras.
11. from keras.callbacks import ...: Imports callback functions for model training monitoring.
12. import datetime: Imports the 'datetime' module for working with dates and times.
13. import matplotlib.pyplot as plt: Imports Matplotlib's 'pyplot' module and assigns the alias 'plt'.
14. from keras.utils import plot_model: Imports a Keras utility for plotting model architectures.
15. from pylab import rcParams: Imports a module to configure Matplotlib plot sizes.
16. rcParams['figure.figsize'] = 5, 2.5: Sets the default plot size to width=5 inches and height=2.5 inches.
17. import librosa:This is a Python package for music and audio analysis.

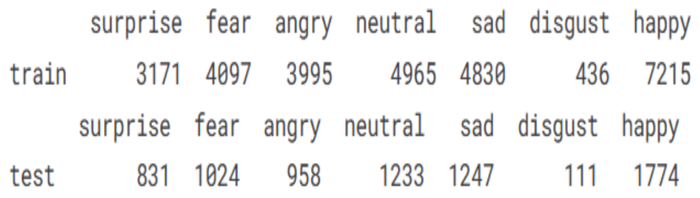
Understanding the Data Distribution :

We need to study the distribution of data across various classes for better understanding. The ‘count_exp’ function iterates over all the sub-directories belonging to the expression category and counts the number of images that depict each emotion. The results are saved in DataFrames train_count and test_count which shows what the data set is made up of. It is important because it shows how balanced and mixed up emotions are between train and test sets which determine whether a model can generalize over a variety of expressions.



```python
train_dir = '../input/fer2013/train/'
test_dir = '../input/fer2013/test/'

row, col = 48, 48
classes = 7

def count_exp(path, set_):
    dict_ = {}
    for expression in os.listdir(path):
        dir_ = path + expression
        dict_[expression] = len(os.listdir(dir_))
    df = pd.DataFrame(dict_, index=[set_])
    return df

train_count = count_exp(train_dir, 'train')
test_count = count_exp(test_dir, 'test')
print(train_count)
print(test_count)
```

Fig. 7. (a) Code for studying data distribution.

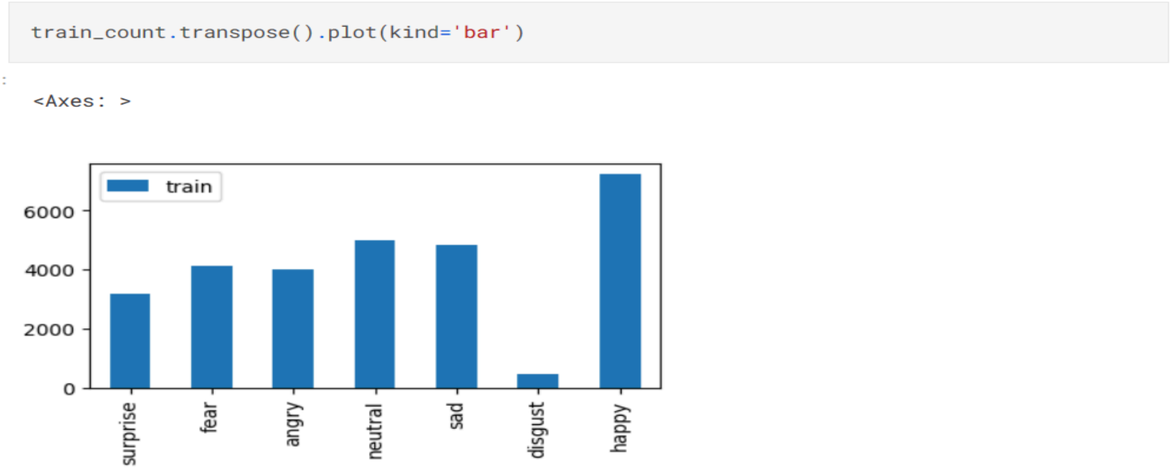


```
         surprise  fear  angry  neutral   sad  disgust  happy
train        3171  4097   3995     4965  4830      436   7215
         surprise  fear  angry  neutral   sad  disgust  happy
test          831  1024    958     1233  1247      111   1774
```

Fig. 7. (b) Distribution across various classes.

Basic Plots for the Data:

Visualizing data distribution across various classes for a better understanding.



```
train_count.transpose().plot(kind='bar')
```

```
<Axes: >
```

Fig 8 . Data distribution plot for Video

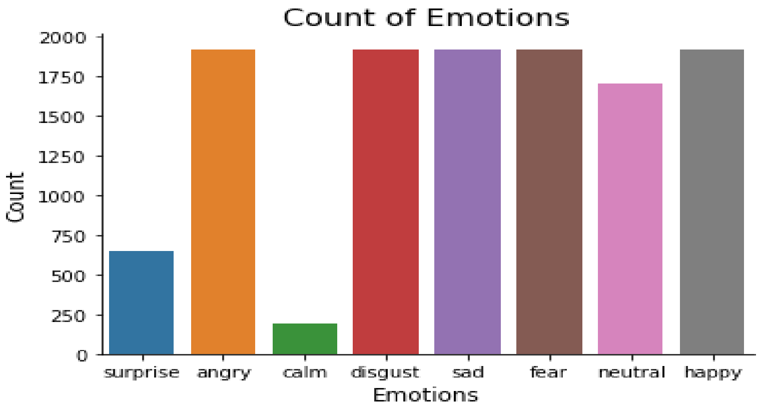
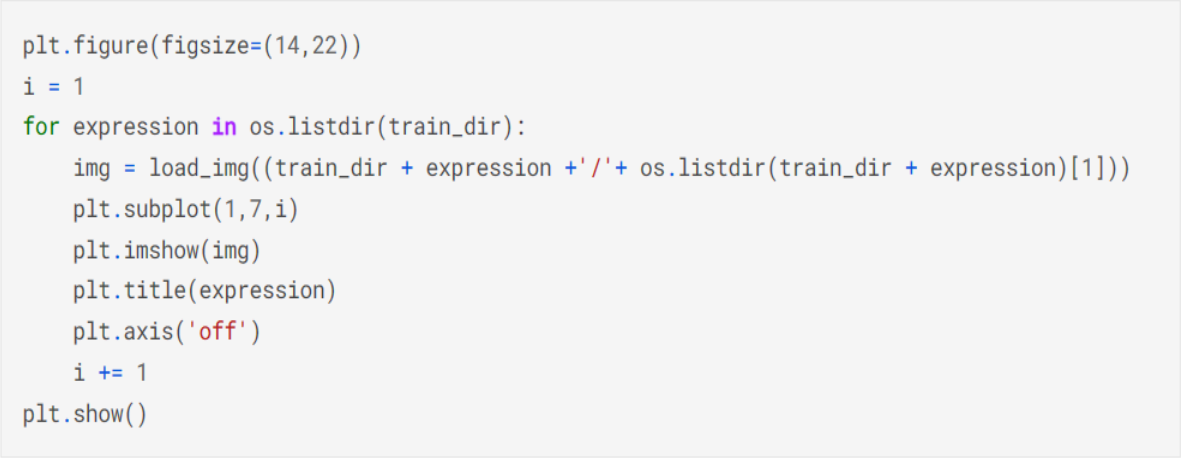


Fig. 9. Data distribution plot for Audio



```python
plt.figure(figsize=(14,22))
i = 1
for expression in os.listdir(train_dir):
    img = load_img((train_dir + expression +'/'+ os.listdir(train_dir + expression)[1]))
    plt.subplot(1,7,i)
    plt.imshow(img)
    plt.title(expression)
    plt.axis('off')
    i += 1
plt.show()
```

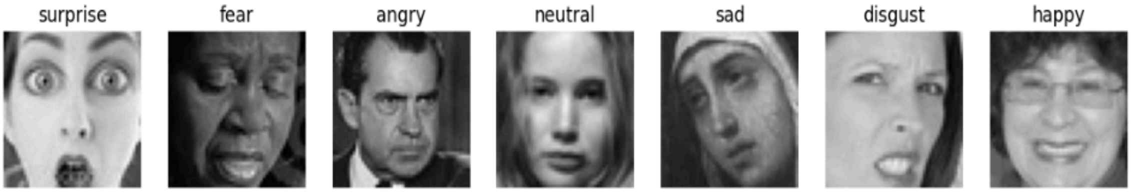
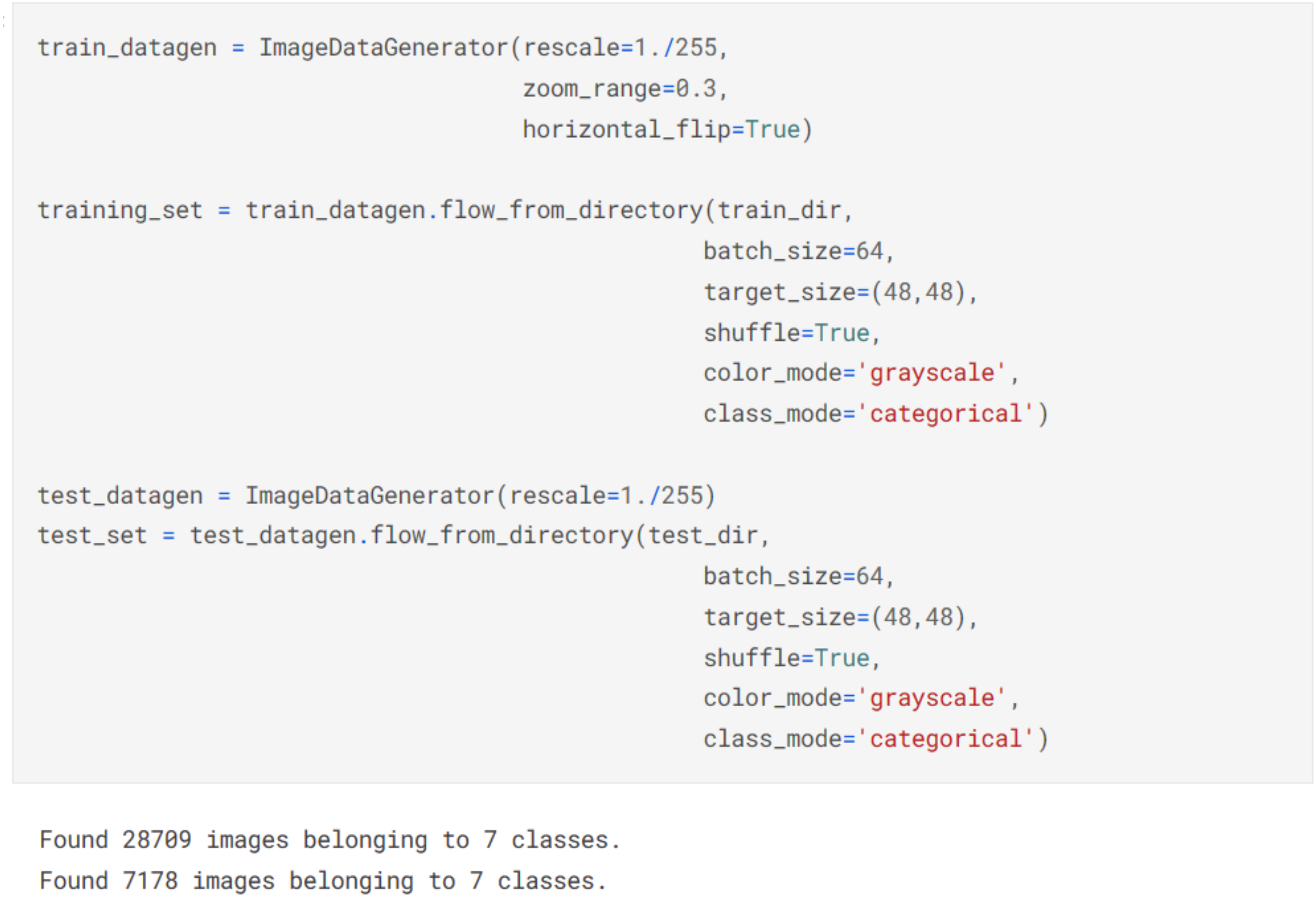


Fig. 10. Sample images from the dataset.

Creating Training and Testing Sets:

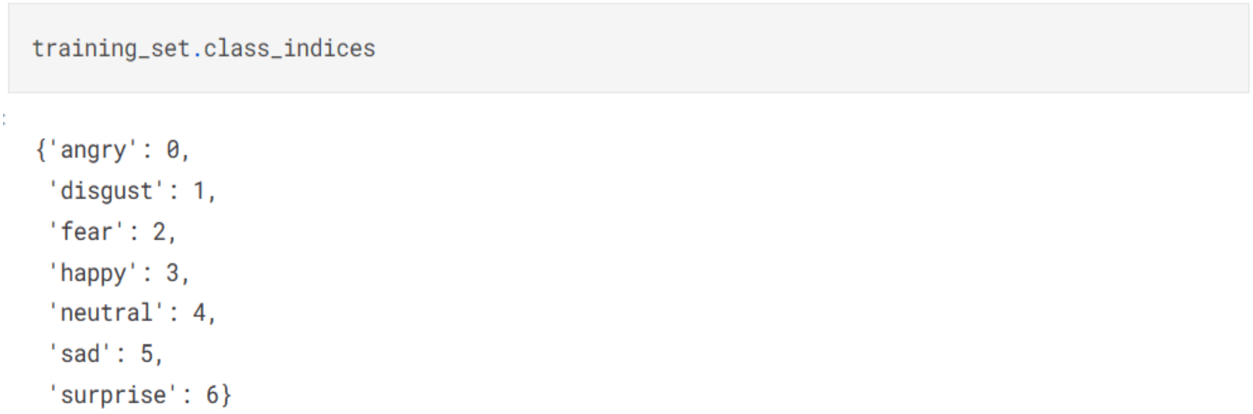
To create training and testing datasets here, we use an ImageDataGenerator to enrich and treat our facial expressions dataset for an emotion recognition model. It loads the training data from ‘train_dir’, it rescales using zooming option with horizontal flip and this gives a better chance to the model on generalization. All images are rescaled to a uniform (48,48) size, converted to grayscale and packed into batches of sixty-four for quick operations. ‘Categorical’ class mode is a setup with multi-class classification. ‘test_dir’s test dataset is also treated in a similar manner. Such data augmentation, pre-processing pipeline is very essential in the building of resilient and flexible emotion recognition model.



```
train_datagen = ImageDataGenerator(rescale=1./255,
                                   zoom_range=0.3,
                                   horizontal_flip=True)

training_set = train_datagen.flow_from_directory(train_dir,
                                                 batch_size=64,
                                                 target_size=(48,48),
                                                 shuffle=True,
                                                 color_mode='grayscale',
                                                 class_mode='categorical')

test_datagen = ImageDataGenerator(rescale=1./255)
test_set = test_datagen.flow_from_directory(test_dir,
                                                 batch_size=64,
                                                 target_size=(48,48),
                                                 shuffle=True,
                                                 color_mode='grayscale',
                                                 class_mode='categorical')
```

```
Found 28709 images belonging to 7 classes.
Found 7178 images belonging to 7 classes.
```

Fig. 11. Code for creating training and testing datasets.



```
training_set.class_indices
```

```
{'angry': 0,
 'disgust': 1,
 'fear': 2,
 'happy': 3,
 'neutral': 4,
 'sad': 5,
 'surprise': 6}
```

Fig. 12. Class labels

Define the Model:

We then describe a CNN architecture for an emotion recognition model in the Keras framework. Here's a breakdown of each layer:

1. Input Layer:

- Shape: `input_size` (48, 48, 1)

- Convolutional layer with 32 filters, kernel size (3, 3), ReLU activation.

2. Hidden Layer 1:

- A linear 64-layer convolution with (3, 3) for dimensions.

- Batch Normalization for normalization.

- 2×2 MaxPooling layer.

- A dropout layer at rate 0.25.

3. Hidden Layer 2:

- Layer 3: 128 convolutional filters, a 3×3 kernel size, ReLu activation function and L2 regularization.

- The network structure consists of convolutional layer with 256 filters, each with the shape of 3×3 and Rectified linear unit activation.

- Batch Normalization.

- two by two max-pooling layer.

- A dropout layer with dropout rate at 0.25.

4. Flatten Layers:

- Two Flatten layers for reshaping data before feeding it to the fully connected layers.

5. Fully Connected Layers:

- Contains 1024 units with ReLu activation function.

- A dropout layer using a 0.5 dropout rate.

- A classifying layer comprising of ‘classes’ dense units, utilizing softmax activation.

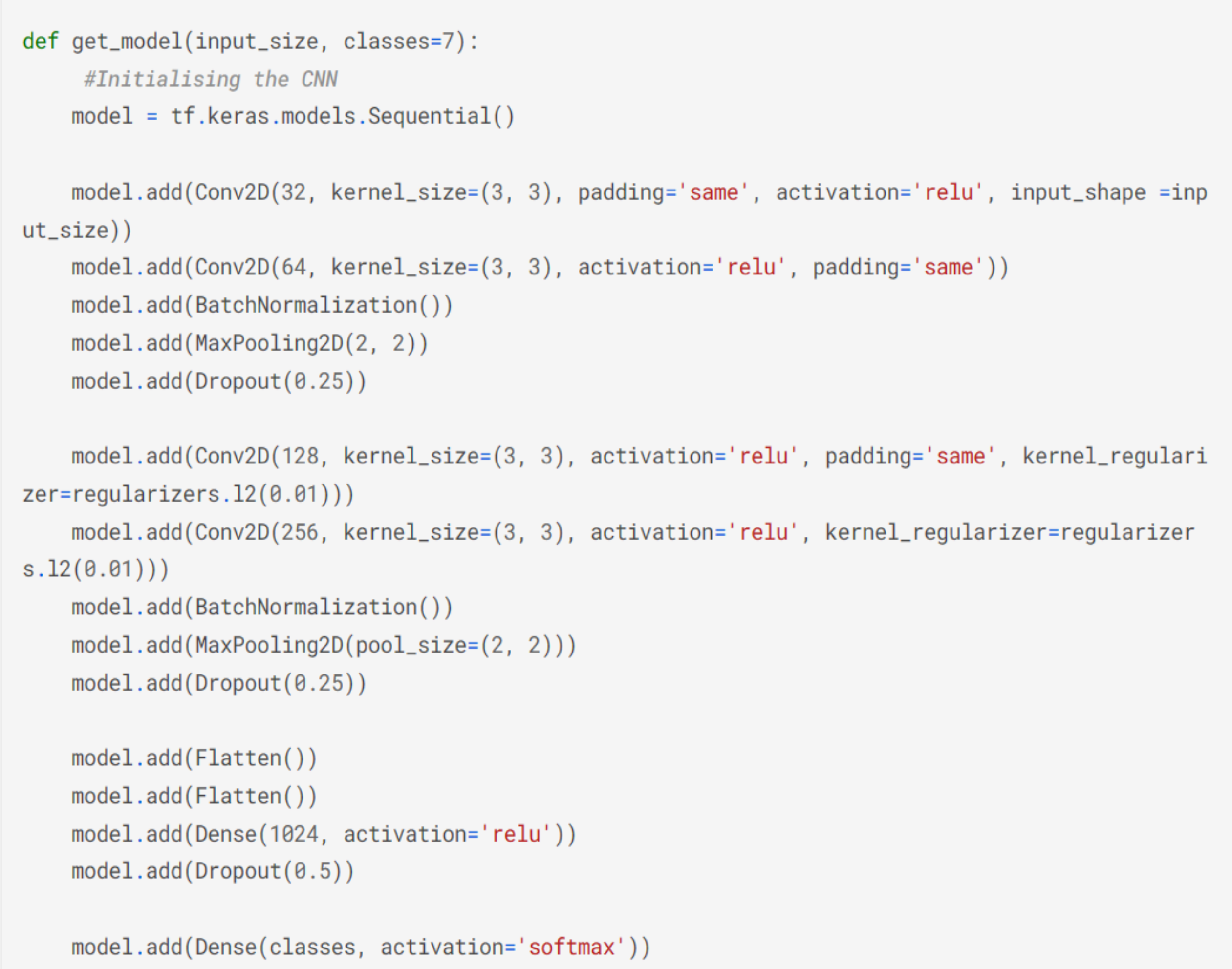
6. Compilation:

- Adam optimizer[25] (learning rate : 0.0001, weight_decay: 1e-6).

- Multi-class classifier loss, categorical cross entropy.

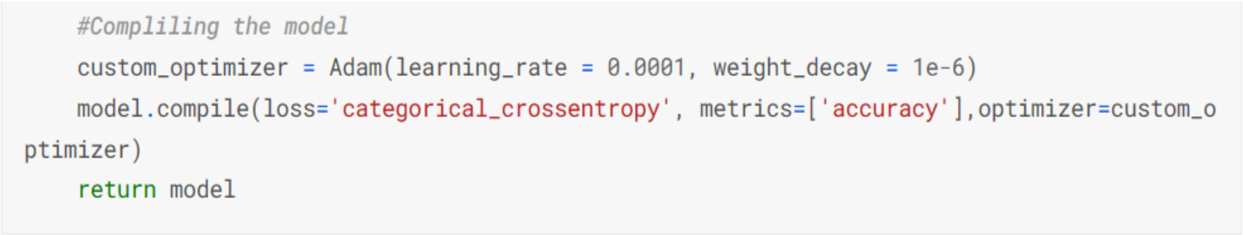
- Accuracy is the evaluation metric.

The model comprises of a series of convolutional layers designed to extract hierarchical features from input samples, followed by multiple fully connected layers meant for classifications. Some of the regularization methods like dropout and batch normalization are used for enhancing the generalization.



```
def get_model(input_size, classes=7):
     #Initialising the CNN
    model = tf.keras.models.Sequential()

    model.add(Conv2D(32, kernel_size=(3, 3), padding='same', activation='relu', input_shape =inp
ut_size))
    model.add(Conv2D(64, kernel_size=(3, 3), activation='relu', padding='same'))
    model.add(BatchNormalization())
    model.add(MaxPooling2D(2, 2))
    model.add(Dropout(0.25))

    model.add(Conv2D(128, kernel_size=(3, 3), activation='relu', padding='same', kernel_regulari
zer=regularizers.l2(0.01)))
    model.add(Conv2D(256, kernel_size=(3, 3), activation='relu', kernel_regularizer=regularizer
s.l2(0.01)))
    model.add(BatchNormalization())
    model.add(MaxPooling2D(pool_size=(2, 2)))
    model.add(Dropout(0.25))

    model.add(Flatten())
    model.add(Flatten())
    model.add(Dense(1024, activation='relu'))
    model.add(Dropout(0.5))

    model.add(Dense(classes, activation='softmax'))
```

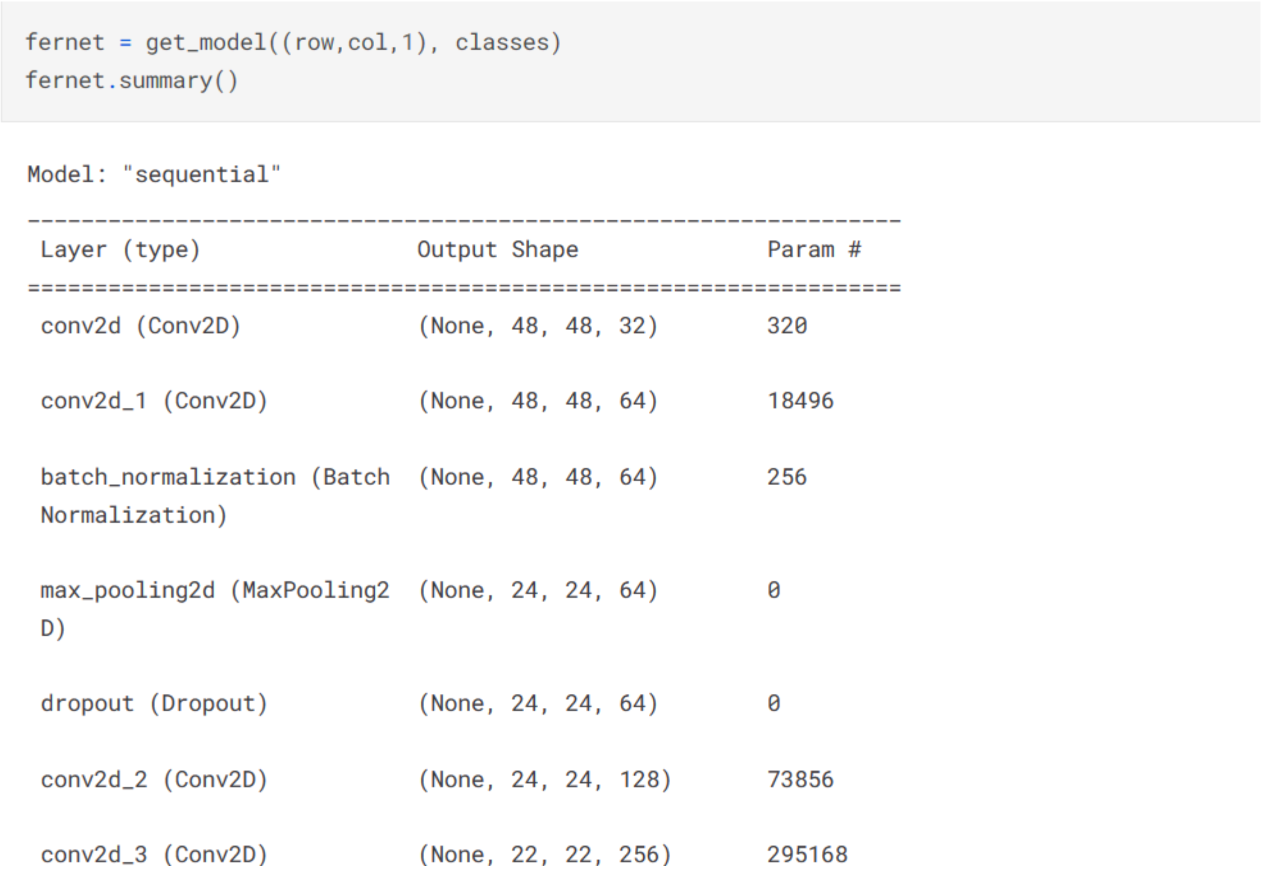
Fig. 13. Code defining the model.



```
    #Compliling the model
    custom_optimizer = Adam(learning_rate = 0.0001, weight_decay = 1e-6)
    model.compile(loss='categorical_crossentropy', metrics=['accuracy'],optimizer=custom_o
ptimizer)
    return model
```

Fig 14. Compiling the model.

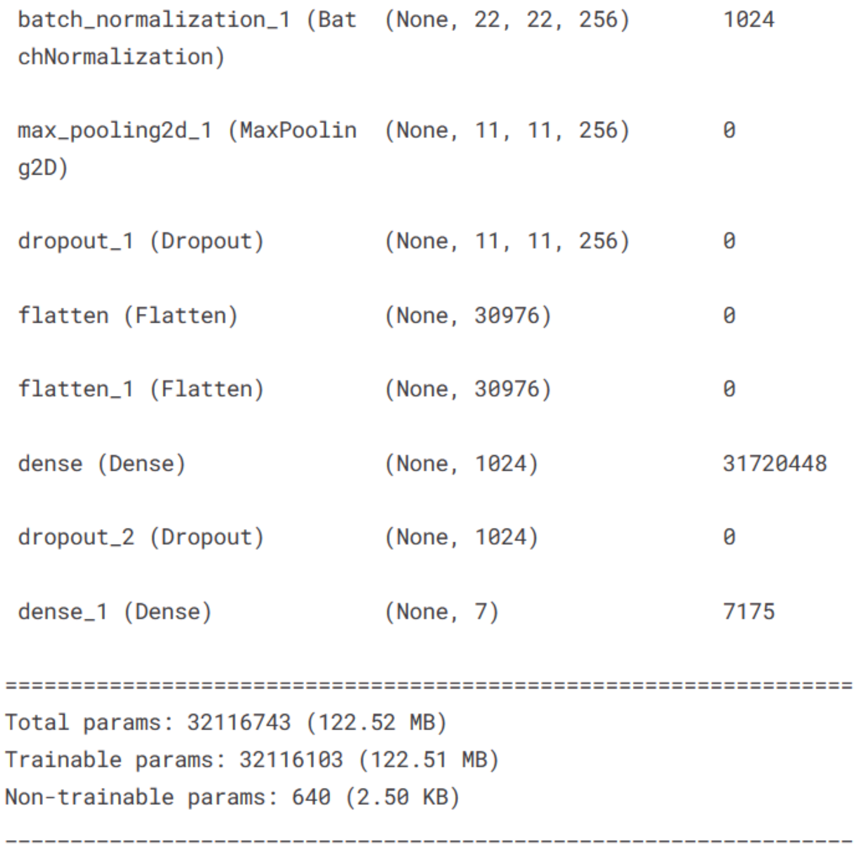
Model Summary :



```
fernet = get_model((row,col,1), classes)
fernet.summary()
```

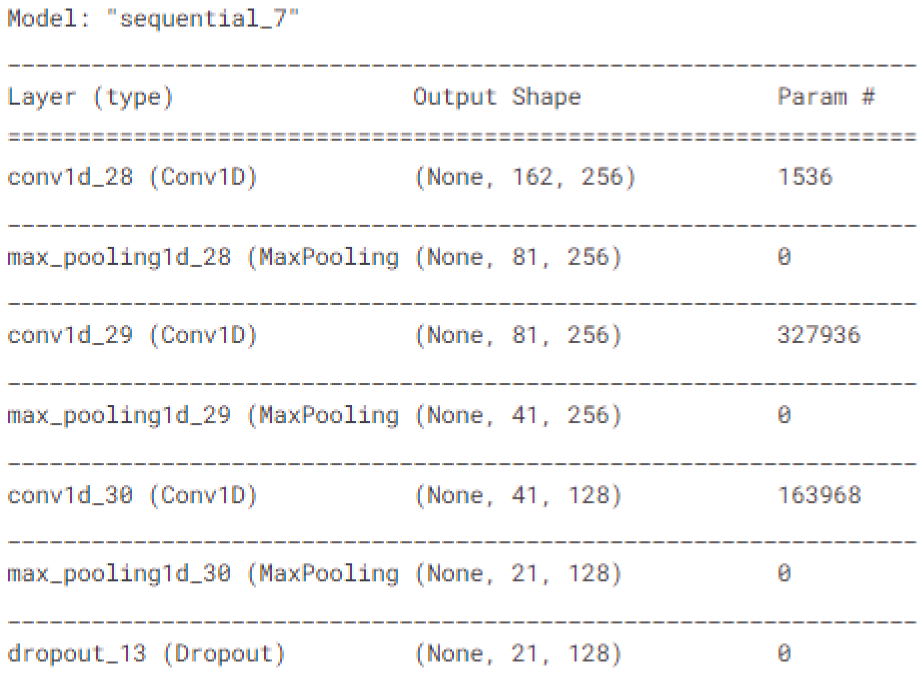
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 conv2d (Conv2D)             (None, 48, 48, 32)        320       
                                                                 
 conv2d_1 (Conv2D)           (None, 48, 48, 64)        18496     
                                                                 
 batch_normalization (Batch  (None, 48, 48, 64)        256       
 Normalization)                                                  
                                                                 
 max_pooling2d (MaxPooling2  (None, 24, 24, 64)        0         
 D)                                                              
                                                                 
 dropout (Dropout)           (None, 24, 24, 64)        0         
                                                                 
 conv2d_2 (Conv2D)           (None, 24, 24, 128)       73856     
                                                                 
 conv2d_3 (Conv2D)           (None, 22, 22, 256)       295168    
```

Fig. 15. (a) Code summarizing the Video model



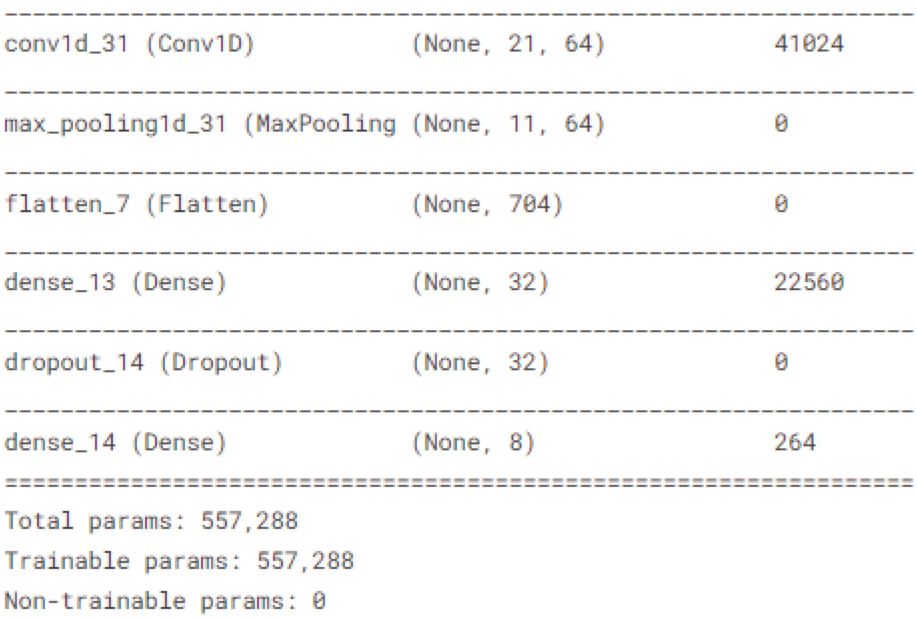
```
 batch_normalization_1 (Bat  (None, 22, 22, 256)       1024
 chNormalization)

 max_pooling2d_1 (MaxPoolin  (None, 11, 11, 256)       0
 g2D)

 dropout_1 (Dropout)         (None, 11, 11, 256)       0

 flatten (Flatten)           (None, 30976)             0

 flatten_1 (Flatten)         (None, 30976)             0

 dense (Dense)               (None, 1024)              31720448

 dropout_2 (Dropout)         (None, 1024)              0

 dense_1 (Dense)             (None, 7)                 7175

=================================================================
Total params: 32116743 (122.52 MB)
Trainable params: 32116103 (122.51 MB)
Non-trainable params: 640 (2.50 KB)
_________________________________________________________________
```

Fig. 15. (b) Code summarizing the Video model



```
Model: "sequential_7"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_28 (Conv1D)           (None, 162, 256)          1536
_________________________________________________________________
max_pooling1d_28 (MaxPooling (None, 81, 256)           0
_________________________________________________________________
conv1d_29 (Conv1D)           (None, 81, 256)           327936
_________________________________________________________________
max_pooling1d_29 (MaxPooling (None, 41, 256)           0
_________________________________________________________________
conv1d_30 (Conv1D)           (None, 41, 128)           163968
_________________________________________________________________
max_pooling1d_30 (MaxPooling (None, 21, 128)           0
_________________________________________________________________
dropout_13 (Dropout)         (None, 21, 128)           0
```

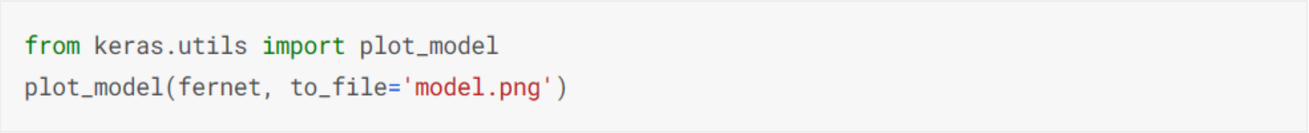
Fig. 15. (c) Code summarizing the Audio model



```
_________________________________________________________________
conv1d_31 (Conv1D)           (None, 21, 64)            41024
_________________________________________________________________
max_pooling1d_31 (MaxPooling (None, 11, 64)            0
_________________________________________________________________
flatten_7 (Flatten)          (None, 704)               0
_________________________________________________________________
dense_13 (Dense)             (None, 32)                22560
_________________________________________________________________
dropout_14 (Dropout)         (None, 32)                0
_________________________________________________________________
dense_14 (Dense)             (None, 8)                 264
=================================================================
Total params: 557,288
Trainable params: 557,288
Non-trainable params: 0
```

Fig. 15. (d) Code summarizing the Audio

Model Visualization :



```
from keras.utils import plot_model
plot_model(fernet, to_file='model.png')
```

Fig. 16. Code for model visualization.

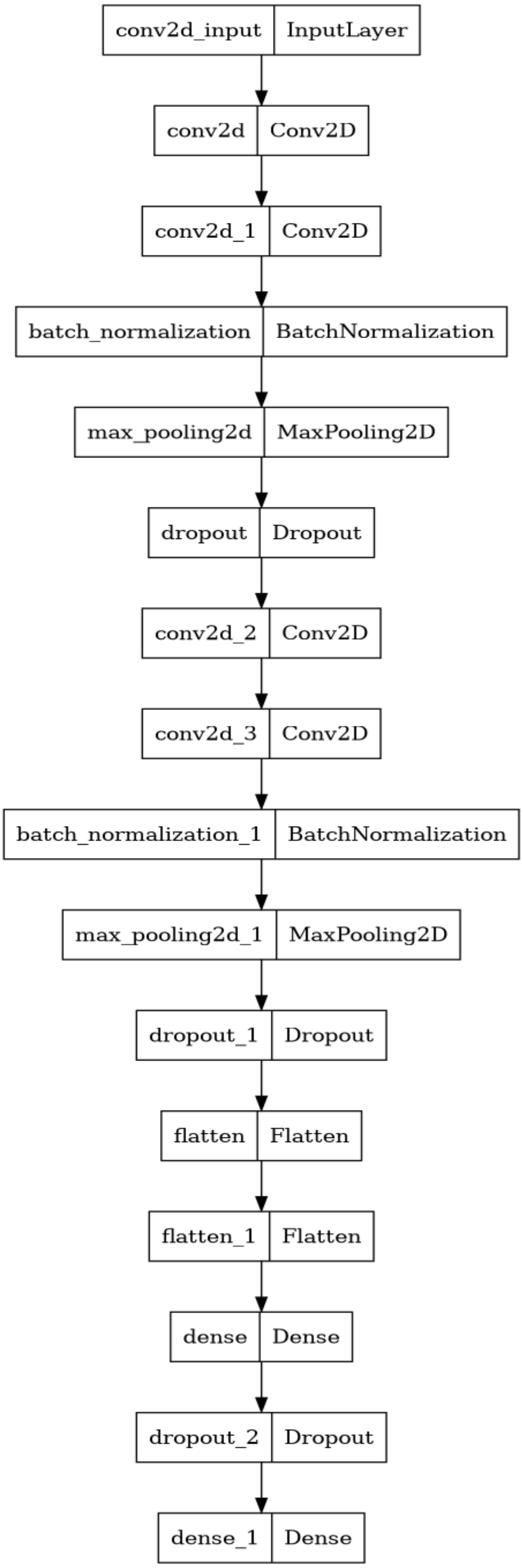


Fig. 17. Model overview

Callbacks Function:

Following this, we configure different callbacks[26] for model training and model checkpointing in the sentiment recognition project. Here's an explanation of each callback:

1. ModelCheckpoint:

- The model’s weights are saved at the end of every epoch, if it gives the best performance on the validation set.

- Filepath: 'ferNet.h5'.

- Mode: Minimum validation loss is monitored.

2. EarlyStopping:

- Validates loss and terminates training if performance does not improve for at least three epochs.

- Restores the best weights.

- Patience: 3 epochs.

3. ReduceLROnPlateau:

- Sets the learning rate when a plateau on validation loss is detected.

- Multiplies the learning rate by 0.2 when there is zero improvement in 6 epochs.

- Minimum delta: 0.0001.

4. TensorBoard:

- Creates logs for TensorBoard visualization.

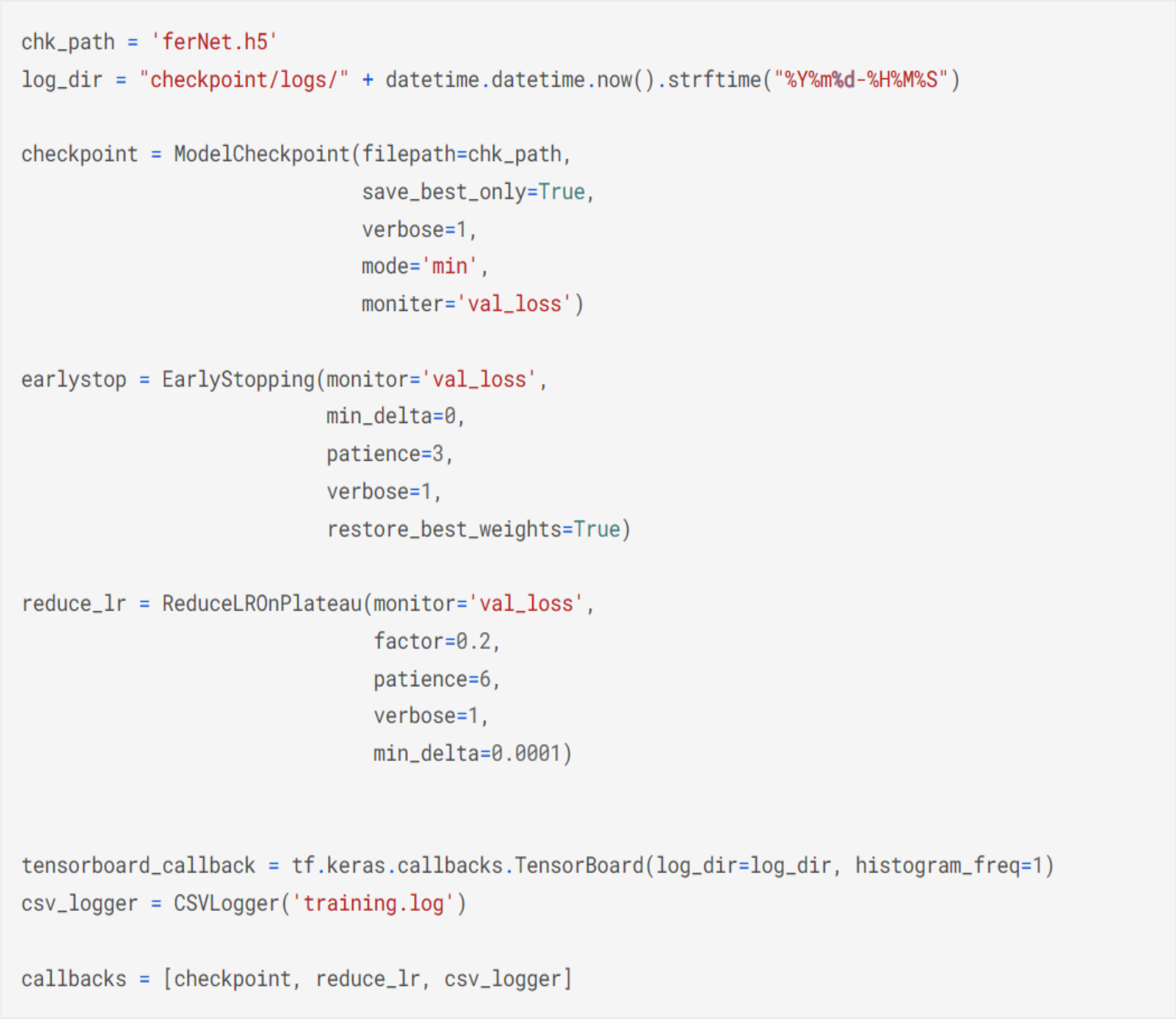
- Logs contain histograms of weights and biases.

- Log directory: "checkpoint/logs/" with a timestamp.

5. CSVLogger:

- It logs training details such as epoch, loss, and accuracy to a csv file named ‘training.log’.

Collectively, they improve the training process with model checkpointing, early stopping to avoid overfitting, dynamic learning rate adaptation, and visual insights of TensorBoard. CSVLogger keeps training data for future study.

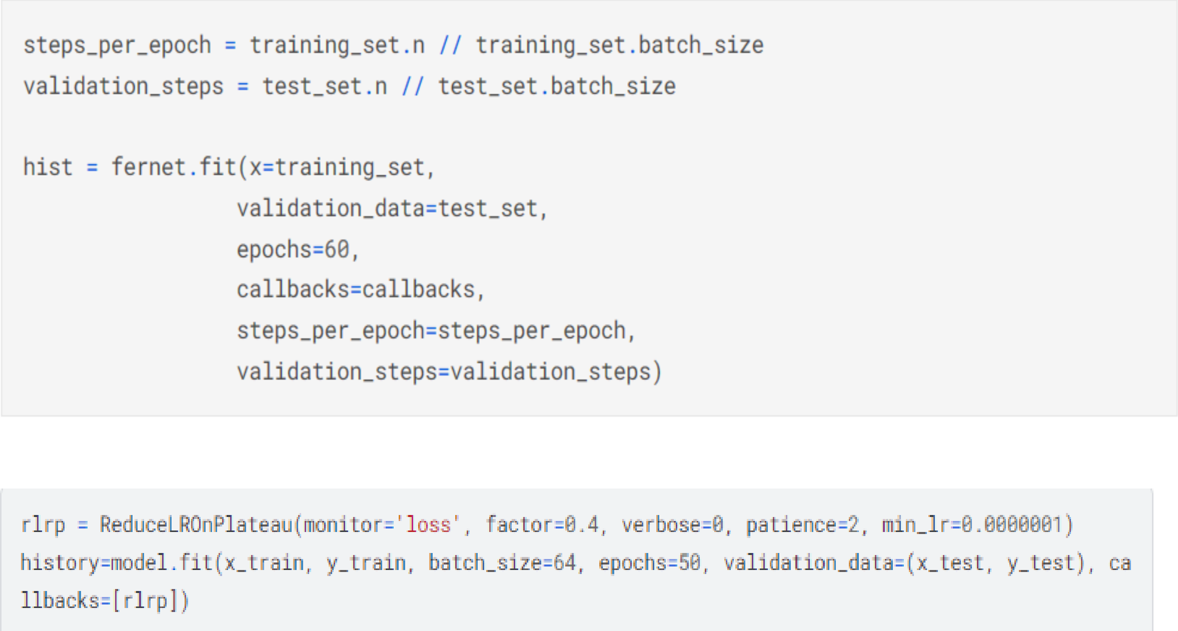


```
chk_path = 'ferNet.h5'
log_dir = "checkpoint/logs/" + datetime.datetime.now().strftime("%Y%m%d-%H%M%S")

checkpoint = ModelCheckpoint(filepath=chk_path,
                             save_best_only=True,
                             verbose=1,
                             mode='min',
                             moniter='val_loss')

earlystop = EarlyStopping(monitor='val_loss',
                          min_delta=0,
                          patience=3,
                          verbose=1,
                          restore_best_weights=True)

reduce_lr = ReduceLROnPlateau(monitor='val_loss',
                              factor=0.2,
                              patience=6,
                              verbose=1,
                              min_delta=0.0001)


tensorboard_callback = tf.keras.callbacks.TensorBoard(log_dir=log_dir, histogram_freq=1)
csv_logger = CSVLogger('training.log')

callbacks = [checkpoint, reduce_lr, csv_logger]
```

Fig. 18. Code for declaring callbacks.

Training Model :

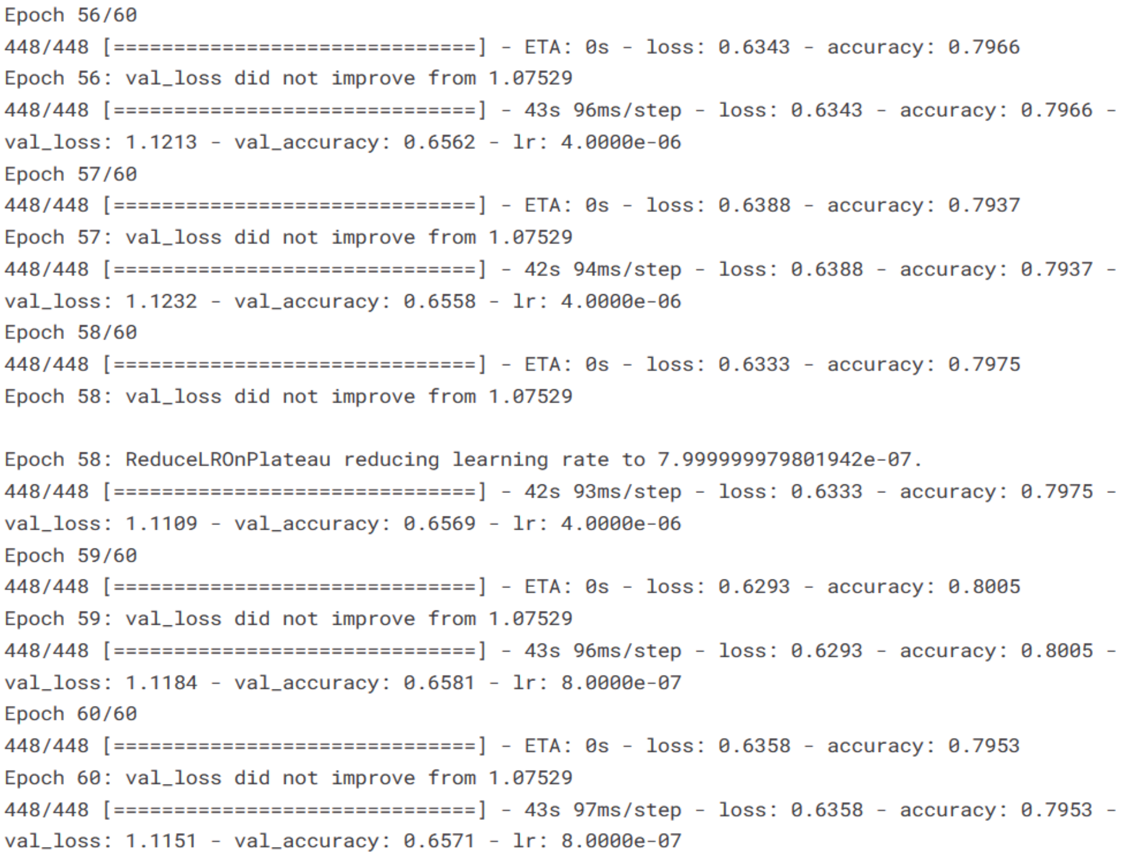
Now we move on to training the model that we created. We determine the number of batches processed in each training and testing epoch. We train our model for 60 epochs with callbacks like model checkpoints, early stopping, etc.



```
steps_per_epoch = training_set.n // training_set.batch_size
validation_steps = test_set.n // test_set.batch_size

hist = fernet.fit(x=training_set,
                 validation_data=test_set,
                 epochs=60,
                 callbacks=callbacks,
                 steps_per_epoch=steps_per_epoch,
                 validation_steps=validation_steps)
```

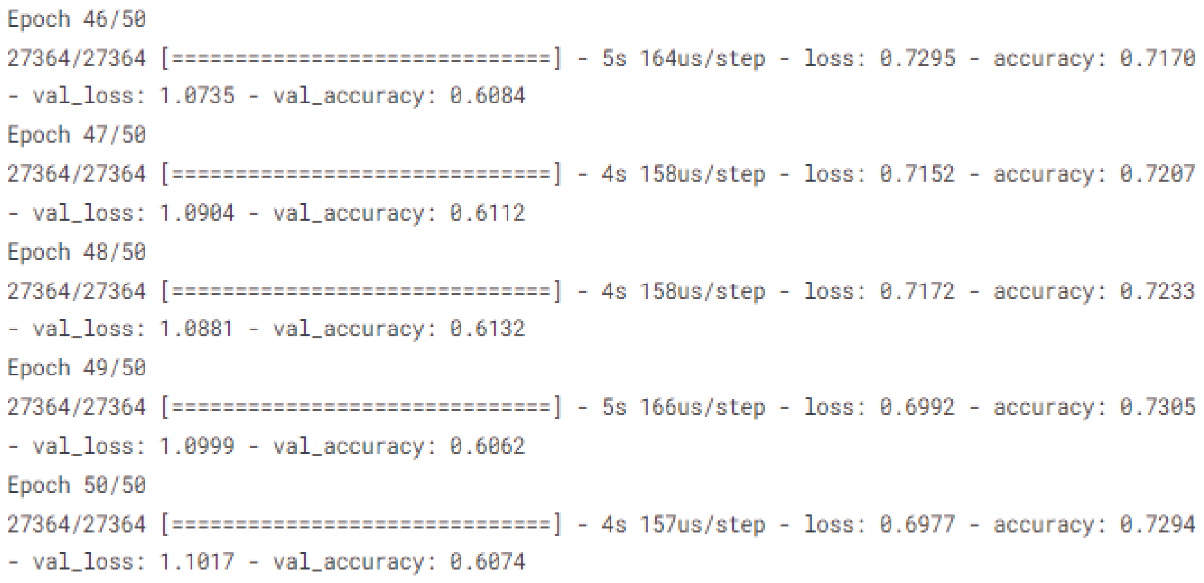
```
rlrp = ReduceLROnPlateau(monitor='loss', factor=0.4, verbose=0, patience=2, min_lr=0.0000001)
history=model.fit(x_train, y_train, batch_size=64, epochs=50, validation_data=(x_test, y_test), ca
llbacks=[rlrp])
```

Fig. 19. (a) Code for training the model.



```
Epoch 56/60
448/448 [==============================] - ETA: 0s - loss: 0.6343 - accuracy: 0.7966
Epoch 56: val_loss did not improve from 1.07529
448/448 [==============================] - 43s 96ms/step - loss: 0.6343 - accuracy: 0.7966 -
val_loss: 1.1213 - val_accuracy: 0.6562 - lr: 4.0000e-06
Epoch 57/60
448/448 [==============================] - ETA: 0s - loss: 0.6388 - accuracy: 0.7937
Epoch 57: val_loss did not improve from 1.07529
448/448 [==============================] - 42s 94ms/step - loss: 0.6388 - accuracy: 0.7937 -
val_loss: 1.1232 - val_accuracy: 0.6558 - lr: 4.0000e-06
Epoch 58/60
448/448 [==============================] - ETA: 0s - loss: 0.6333 - accuracy: 0.7975
Epoch 58: val_loss did not improve from 1.07529

Epoch 58: ReduceLROnPlateau reducing learning rate to 7.999999979801942e-07.
448/448 [==============================] - 42s 93ms/step - loss: 0.6333 - accuracy: 0.7975 -
val_loss: 1.1109 - val_accuracy: 0.6569 - lr: 4.0000e-06
Epoch 59/60
448/448 [==============================] - ETA: 0s - loss: 0.6293 - accuracy: 0.8005
Epoch 59: val_loss did not improve from 1.07529
448/448 [==============================] - 43s 96ms/step - loss: 0.6293 - accuracy: 0.8005 -
val_loss: 1.1184 - val_accuracy: 0.6581 - lr: 8.0000e-07
Epoch 60/60
448/448 [==============================] - ETA: 0s - loss: 0.6358 - accuracy: 0.7953
Epoch 60: val_loss did not improve from 1.07529
448/448 [==============================] - 43s 97ms/step - loss: 0.6358 - accuracy: 0.7953 -
val_loss: 1.1151 - val_accuracy: 0.6571 - lr: 8.0000e-07
```

Fig. 19. (b) Accuracy and loss after 60 epochs of video.

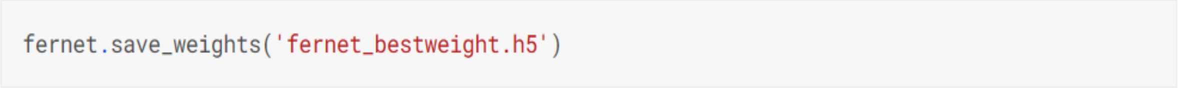


```
Epoch 46/50
27364/27364 [==============================] - 5s 164us/step - loss: 0.7295 - accuracy: 0.7170
- val_loss: 1.0735 - val_accuracy: 0.6084
Epoch 47/50
27364/27364 [==============================] - 4s 158us/step - loss: 0.7152 - accuracy: 0.7207
- val_loss: 1.0904 - val_accuracy: 0.6112
Epoch 48/50
27364/27364 [==============================] - 4s 158us/step - loss: 0.7172 - accuracy: 0.7233
- val_loss: 1.0881 - val_accuracy: 0.6132
Epoch 49/50
27364/27364 [==============================] - 5s 166us/step - loss: 0.6992 - accuracy: 0.7305
- val_loss: 1.0999 - val_accuracy: 0.6062
Epoch 50/50
27364/27364 [==============================] - 4s 157us/step - loss: 0.6977 - accuracy: 0.7294
- val_loss: 1.1017 - val_accuracy: 0.6074
```

Fig. 19. (c) Accuracy and loss after 50 epochs of audio.

Saving the Weights :

Finally we save the weights trained in a file so that they may be used for future predictions.



```
fernet.save_weights('fernet_bestweight.h5')
```

Fig. 20. Saving model weights to a .h5 file.

3.5 KEY CHALLENGES :

1. **Variability in Expressions :**

The biggest problem faced by most emotion recognition systems is a difference in expression. Because emotions are inherently personal and individualistic, they have many possible and even opposing expressions among people. It becomes more complex as different people can portray a single feeling individually hence it is difficult to discern. It is through this variability that our model becomes vital. This means that it should be able to understand and embrace this broad range of sentiments. In this respect, generalization becomes important; the model should comprehensively be trained on varied datasets to identify emotions regardless of how individually expressed they can be.

This calls for intensive training of our model using multicultural, multi-ethnic, cross-age datasets in order to avoid making generalization assumptions regarding language translation or its speakers and their intentions. The aim in this training is to acquaint the model with numerous emotional displays so that it can understand the different shades and subtleties of what it means to feel something. The model also encompasses resilience mechanisms. It is modifiable and compatible with various shapes of faces, hand signs, and speech patterns. Models become more robust towards such diversity by techniques such as data augmentation that introduce variations into expression during training.

2. **Data Quality :**

The success of any FER model depends on data quality. Quality of annotations in the dataset has a very important effect on learning process and evaluation tasks. False or inconsistent tagging of emotions would be detrimental for the work of this kind and make it hard to tell emotions correctly. Quality annotation is vital in training the FER model because it acts as the grounds' truth. These tell the model how to learn the connection between facial expressions and related feelings. Any inaccuracies in annotations could lead to noise and bias in the training data resulting in poor generalization of the model.

Accurate and consistent annotations of the dataset also involve different stages. It starts with extensive annotation carried out painstakingly by trained annotators or experts. The annotators should have good knowledge of emotional expressions so as to accurately label the dataset. Additionally, the whole quality control procedure including the validation of data ought to be employed for checking and confirming the annotations. The process consists of comparing labels between several annotators in order to detect and correct possible mismatches or mistakes. Uniformity of annotations through the entire dataset is required to avoid confusions while building a model.

3. **Limited Diversity :**

This helps to ensure that diversity exists within the Facial Emotion Recognition (FER) dataset to avoid biases and aid the model's capability to generalize crosswise over varied cultures, ethnicities, age groups, as well genders.The major issue related with FER datasets is that not all demographic group may have been represented.ICENSE: ## Parameter: The license is a crucial feature due to the significance of this contract. However, if the data set is not diversified enough, the model may not be accurate in predicting different facial expressions since it has not seen enough of them.

This can be addressed by selecting and sampling data from different demographic groups based on various factors, both individual and societal. This consists of obtaining facial pictures of people from various ethnicities, age clusters, male/female groups, and cultures. The addition of such diversified samples can give more scope to the model for learning different facial expressions which may help in recognizing emotions across different sections of society correctly.Secondly, the dataset has to be well-balanced across various demographic groups. Predictions can be biased when a group or individual, who are usually outnumbered by other groups for instance, over-dominates an imbalanced dataset. Hence, there should be a balanced collection of equal measures and distribution of facial images from different demographic groups.

4. **Computational Resources :**

The use of a 15-layer deep CNN in developing a face recognizer for recognition of emotions calls for huge computing capacity. Training such a complex model typically requires significant computational resources and hours.The computational challenge in addressing this can be solved by optimizing the model architecture or using other approaches that allow for balancing between performance and computationally efficiency. This can be done by looking into available optimization techniques which will lead to the trimming of the model without compromising on its ability to effectively identify the facial emotions.

Various approaches can help to reduce computation needs without affecting the model's effectiveness. For instance, there are model compression strategies like pruning and quantization that can be used to reduce the number of excess parameters in the network and make its representation efficient. During this optimization process, the objective is not to burden the computer with additional computation work but to enable it to distinguish emotions efficiently.In addition, investigation of transfer learning or feature extraction from pre-trained models will also be a good approach. This approach can help reduce the computation costs required for training the deeper layers of the network by borrowing already-prepared models for extracting facial features. Using such an approach, the model can work on those learned features from other pre-trained models which speeds up learning processes thus saving resources.

5. **Overfitting :**

The phenomenon of overfitting occurs quite often in the domain of deep learning, especially when one deals with small databases. The concept refers to a situation whereby the model overfits the training data such that it can't appropriately generalize to fresh or unseen data.Effective regularization techniques are necessary to overcome overfitting with the deep learning model that you are developing for facial emotion recognition. The use of these techniques help the model not to overfit the data and therefore allow it to generalize better on unseen cases.

A common method is using dropout which is one of the regularization techniques in which some randomly chosen neuron's are dropped out at each training step. Additionally, this ensures that the model does not get dependent on certain neurons or features. Instead, such a strategy leads to stronger learning of non-specific and more generalized representations.Batch normalization is another helpful way that the activations of each layer are normalized in order to smooth out the learning process and minimize the chance of overfitting. Batch normalization helps to stabilize the distribution of activations and therefore promotes better generalization of the model.

6. **Interpretable Features :**

It is important in deep learning to ensure the model generates interpretable and meaningful features so as one can understand how the model arrived at the conclusion. For instance, deep neural nets frequently operate as "black boxes" that can hardly explain why certain predictions are performed in facial emotion recognition tasks.Therefore, methods have to be used which make decisions of models easier for interpretation. This is where visualization methods can be employed to show where and/or what inputs give the greatest contribution to the decision making process of the model under inspection. Saliency maps, activation maximization and Grad-CAM are some of these methods that will enable one to obtain highlights of the important areas within the input images as pertains to this model's decisions.

Moreover, adopting methods which help understand the learned representations will improve interpretability. Visualization of various high-dimensional feature spaces using dimensionality reduction techniques like t-SNE and PCA gives an opportunity to have a better understanding of what features describe certain patterns in data and their mutual separability.Additionally, attention mechanisms contained within the model architecture would assist in comprehending which of the input data is key. Attention mechanisms help assign different weights on various portions of the input thereby allowing the model to concentrate on some areas when making predictions.

7. Model Evaluation :

Evaluation measures for a video model used for task specific applications such as facial emotion recognition need to capture the temporal nature of video data. Traditional metrics for the tasks of images remain very useful but might not fully characterize the model performance for video sequence.This necessitates the inclusion of more detailed assessment parameters based on temporality. Emotion metrics encapsulating the temporal dynamics of emotions across video frames will provide deeper insights into model success. In this respect, it is important to consider metrics such as temporal consistency, which measures how consistent predicted emotions are over time, or temporal alignment metrics focused on comparing predicted emotions and ground truth frames.

Moreover, the time needed to perceive each emotion differs across different videos, hence making measures taking into consideration the granularity of a temporal metric necessary. Time interval based metrics like frame level F1-score or temporal IoU could be used to estimate how accurately the model locates the emotions throughout the span of a video.Holistic metrics, which integrate performance throughout a complete video clip, are yet another crucial aspect. A measurement approach for whole video such as weighted F1 score or sequence level accuracy can be used to establish comprehensive metrics for assessing the model’s consistency in recognizing emotions across all frames.

CHAPTER 4 : TESTING

4.1 TESTING STRATEGY :

1. **Unit Testing :** During this initial phase of testing, single elements are examined separately. A sequence of tests are carried out on each module, which involve data preprocessing, feature extraction and model implementation to verify independence and accuracy.

2. **Integration Testing :** In this phase, the interactions and connections among diverse subsystems are authenticated. Therefore, it ensures smooth communication of the information between different units, as well as makes every unit fit into architectural requirements of the whole system.

3. **Data Quality Testing :** A validation of the quality of the dataset is made to ascertain if it's suitable for training and evaluation purposes. In this testing stage, it confirms that there is not any imbalance in data, the accuracy of the annotations, and consistency that helps create a reliable ground fact for model evaluation.

4. **Functionality Testing :** A thorough evaluation of its core functionality. The model's ability to correctly distinguish a series of different emotions in multiple video clips is examined. It evaluates whether the model adequately presents various forms of expressed emotions.

5. **Performance Testing :** Important measures for determining the model's performance include accuracy, precision, recall, and F1 score, as well as confusion matrices. The measure assesses the model's prediction capabilities on both training and out-of sample testing sets, guaranteeing that it generalizes well beyond the training set.

6. **Robustness Testing :** Measuring the model's response to such challenges as noise, different illumination levels, diverse face orientations, and occlusions is important. In this stage, the model is tested against difficult real life scenarios.

7. **Temporal Testing :** This implies that the model's ability to detect emotions in successive time frames within a video sequence is determined. They check if the model continues to predict emotions faithfully until the end of the video

8. **Cross-validation Testing :** Some techniques like k-fold cross-validation[27] are used for checking out of the model's accuracy over different subset of datasets. This assists in determining whether the model is generalizable and reduces the risk of overfitting.

9. **User Testing** : Feedback in respect of system's usability, accuracy and overall functioning is collected via user-testing. The information from this users feed will be essential in tuning the system with actual users' experience.

10. **Scalability Testing :** The efficiency of the model in processing different length, time, and size of videos will be tested as a part of testing the scalability of the model. Scalability is confirmed by measuring its performance for videos with various running times and resolutions.

4.2 TEST CASES AND OUTCOMES :

1. **Data Preprocessing :**

   Test Case: Validate the quality of loading, re-sizing, and standardizing processes for the data set.

   Expected Outcome: All data were successfully loaded and preprocessed with no errors confirmed. The dataset should be properly formatted, resized, and normalized for other applications.

2. **Feature Extraction :**

   Test Case: Test whether feature extraction techniques using facial landmark or motion features are valid.

   Expected Outcome: These characteristics of faces must be extracted so that they can faithfully reflect those emotional facial expressions and expressions present in the video frames.

3. **Model Training :**

   Test Case: Check proper convergence as well as error handling by training a portion of the dataset.

   Expected Outcome: There must be no errors when the model trains, and it should converge within a reasonable number of epochs.

4. **Emotion Recognition :**

   Test Case: Use the learned model to identify emotions in testing video sequences.

   Expected Outcome: The predictions made by the model on test videos must be similar to the annotations corresponding to emotions shown.

5. **Testing :**

   Test Case: To test the robustness of the model, introduce noises/occlusions in video frames.

   Expected Outcome: The developed model will have the capacity of identifying emotional expressions with a certain degree of error tolerance even under introduction of disturbances.

6. **Temporal Consistency :**

   Test Case: This examines the level of stability in discriminating between emotions in different time instances, in one video.

   Expected Outcome: For the duration of each particular video, the model must sustain emotional recognition as well as continuously identify its accuracy.

7. **Performance Metrics :**

   Test Case: Perform accuracy, precision, recall, F1-score, and confusion matrices.

   Expected Outcome: Ensure that high levels of accuracy and balance are achieved in the various emotional aspects to ensure the model's total success.

8. **User Interaction Testing :**

   Test Case: Emotional reactions of users for video.

   Expected Outcome: The system should recognize and give an appropriate response to the indicated emotions of users in the videos.

9. **Scalability Testing :**

   Test Case: Use different length videos and resolutions to test the model's performance.

   Expected Outcome: A scaling performance should be demonstrated by showing a consistent and credible performance during different video durations as well as resolutions.

10. **Cross-validation Testing :**

    Test Case: For model evaluation use k-fold cross-validation.

    Expected Outcome: In other words, the model should be stable on various parts of the data set to prove its reliability and applicability.

CHAPTER 5 : RESULT AND EVALUATION

5.1 RESULT :

Loss and Accuracy Plots :

Machine learning is represented on graphically as loss and accuracy plots. The loss plot represents the ability of the model to lower mistakes as the epoch increases. However, the accuracy plot shows how often the model gets its predictions right. Ideally, the loss should reduce and thus improve learning while also increasing accuracy. These plots serve an important role in the assessment of the training process, revealing any overfitting or underfitting, and improving performance of the model on a future set of data.

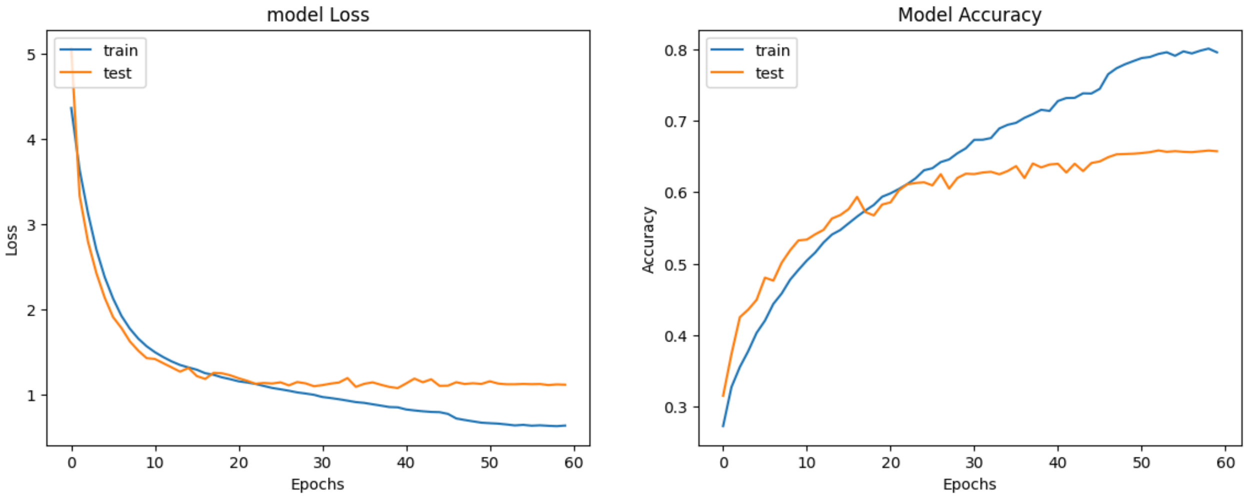


Fig. 21. (a)Plots depicting how loss and accuracy changes with subsequent epochs for Video

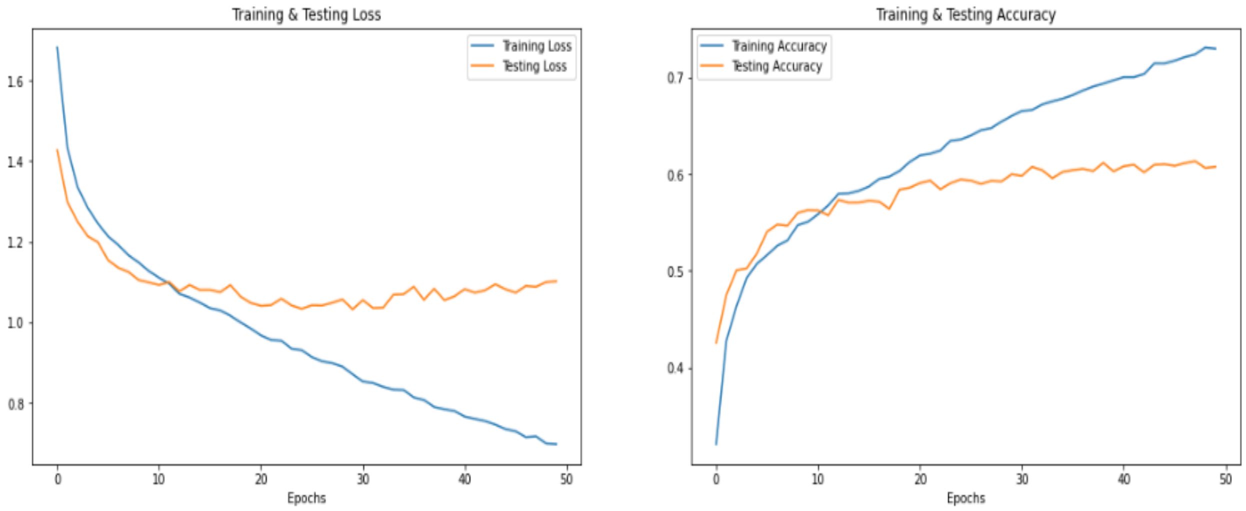
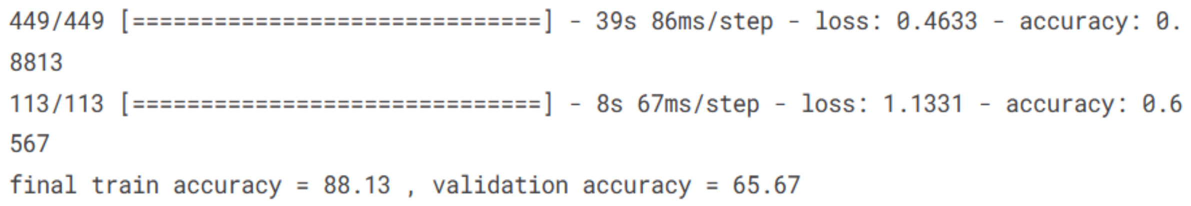


Fig. 21. (b)Plots depicting how loss and accuracy changes with subsequent epochs for Audio

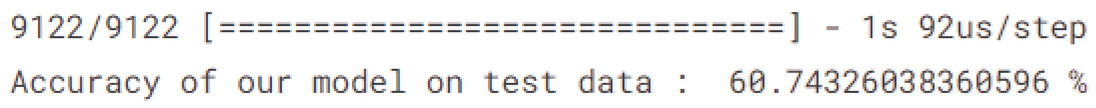
These figures indicate gradual improvements in training and testing losses and accuracies till 47 epochs. Not much improvements were observed in the validation accuracy after 47 epochs.

Model Evaluation :



```
449/449 [==============================] - 39s 86ms/step - loss: 0.4633 - accuracy: 0.
8813
113/113 [==============================] - 8s 67ms/step - loss: 1.1331 - accuracy: 0.6
567
final train accuracy = 88.13 , validation accuracy = 65.67
```

Fig. 22(a). Metrics after training and validating the Video Model



```
9122/9122 [==============================] - 1s 92us/step
Accuracy of our model on test data :  60.74326038360596 %
```

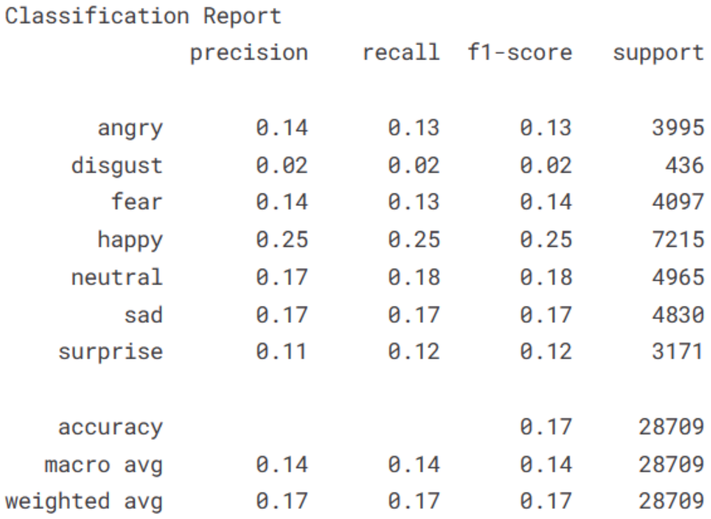
Fig. 22(b). Metrics after training and validating the Audio Model

Our model was able to reach a final training and testing accuracy of 88.13% and 65.67% respectively.

Classification Report :

The machine learning classification report summarizes the performance of a given model, per all the involved classes in a classification exercise. It involves the precision, recall as well as the F-score showing how effectively the model categorizes numerous categories. The F1-score incorporates precision and recall which evaluate true positive predictions and capture all positives by the model. The report assists in evaluating the performance of the classifier especially when a dataset presents a lopsided distribution of classes.

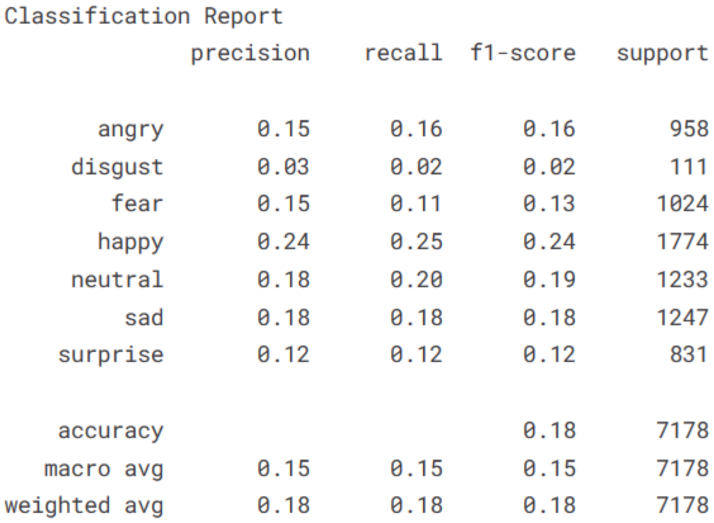
Classification report on the training set:



```
Classification Report
              precision    recall  f1-score   support

       angry       0.14      0.13      0.13      3995
     disgust       0.02      0.02      0.02       436
        fear       0.14      0.13      0.14      4097
       happy       0.25      0.25      0.25      7215
     neutral       0.17      0.18      0.18      4965
         sad       0.17      0.17      0.17      4830
    surprise       0.11      0.12      0.12      3171

    accuracy                           0.17     28709
   macro avg       0.14      0.14      0.14     28709
weighted avg       0.17      0.17      0.17     28709
```

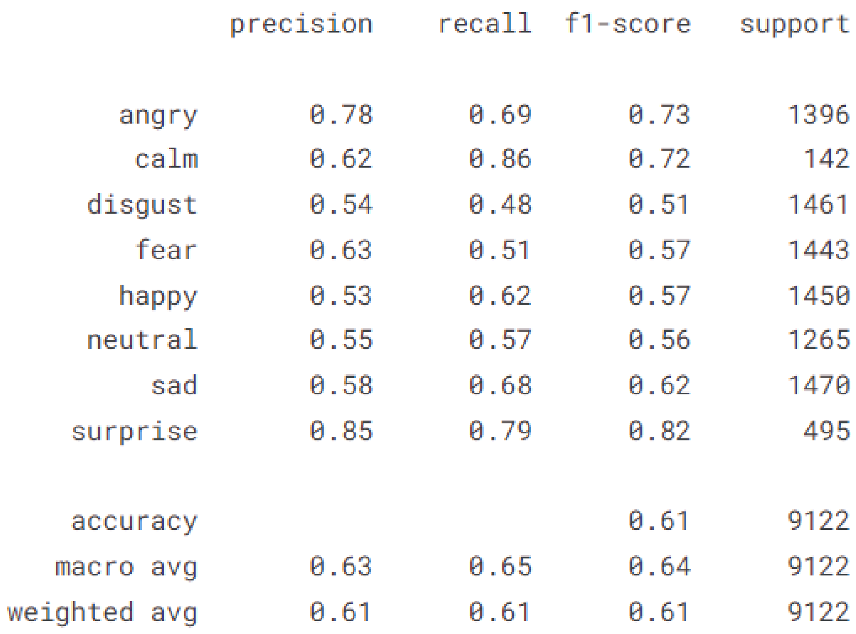
Fig. 23. (a) Classification report for Video training data.

Classification report on testing data:



```
Classification Report
              precision    recall  f1-score   support

       angry       0.15      0.16      0.16       958
     disgust       0.03      0.02      0.02       111
        fear       0.15      0.11      0.13      1024
       happy       0.24      0.25      0.24      1774
     neutral       0.18      0.20      0.19      1233
         sad       0.18      0.18      0.18      1247
    surprise       0.12      0.12      0.12       831

    accuracy                           0.18      7178
   macro avg       0.15      0.15      0.15      7178
weighted avg       0.18      0.18      0.18      7178
```

Fig. 23. (b) Classification report for Video testing data.



```
              precision    recall  f1-score   support

       angry       0.78      0.69      0.73      1396
        calm       0.62      0.86      0.72       142
     disgust       0.54      0.48      0.51      1461
        fear       0.63      0.51      0.57      1443
       happy       0.53      0.62      0.57      1450
     neutral       0.55      0.57      0.56      1265
         sad       0.58      0.68      0.62      1470
    surprise       0.85      0.79      0.82       495

    accuracy                           0.61      9122
   macro avg       0.63      0.65      0.64      9122
weighted avg       0.61      0.61      0.61      9122
```

Fig. 23. (c) Classification report for Audio testing data.

Confusion Matrix:

To put it simply, the confusion matrix is a table that outlines classifier’s performance in terms of TP, FP, TN, and FN. It displays the sum of TP, TN, FP and FN that are detected by the model. The matrix shows how well or bad the model can predict each of the classes in the data set. This helps to assess the strengths and weaknesses of the models and to interpret the kind of mistakes the classifier makes when processing the data.

Confusion matrix for training set:

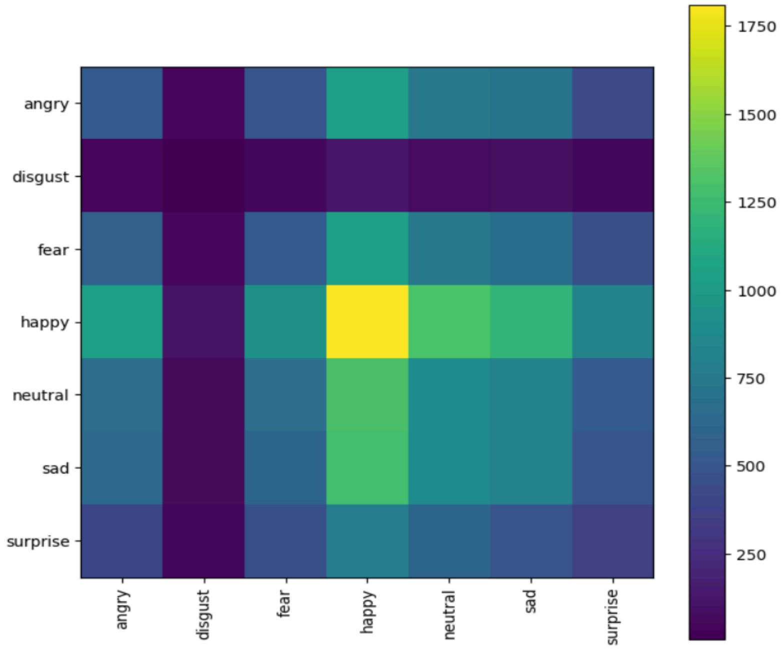


Fig. 24. (a) Confusion matrix for Video training data.

Confusion matrix on test data **:**

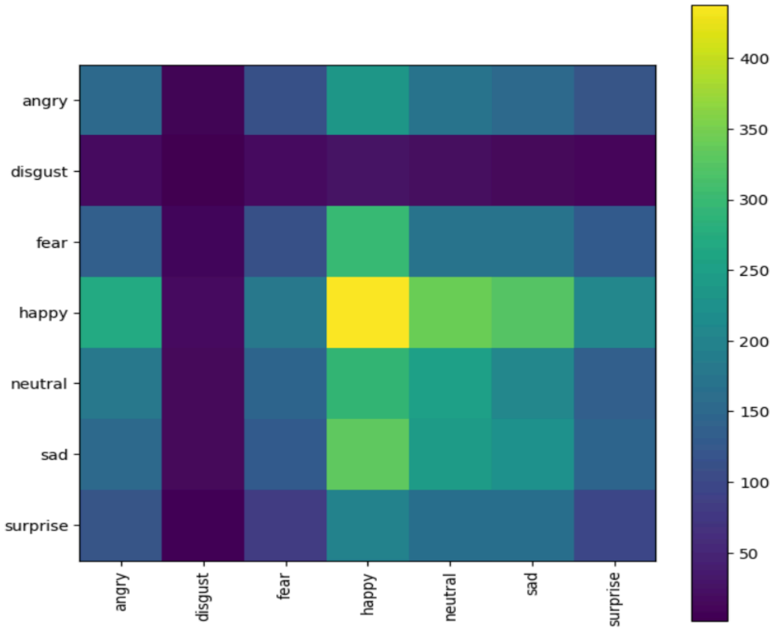
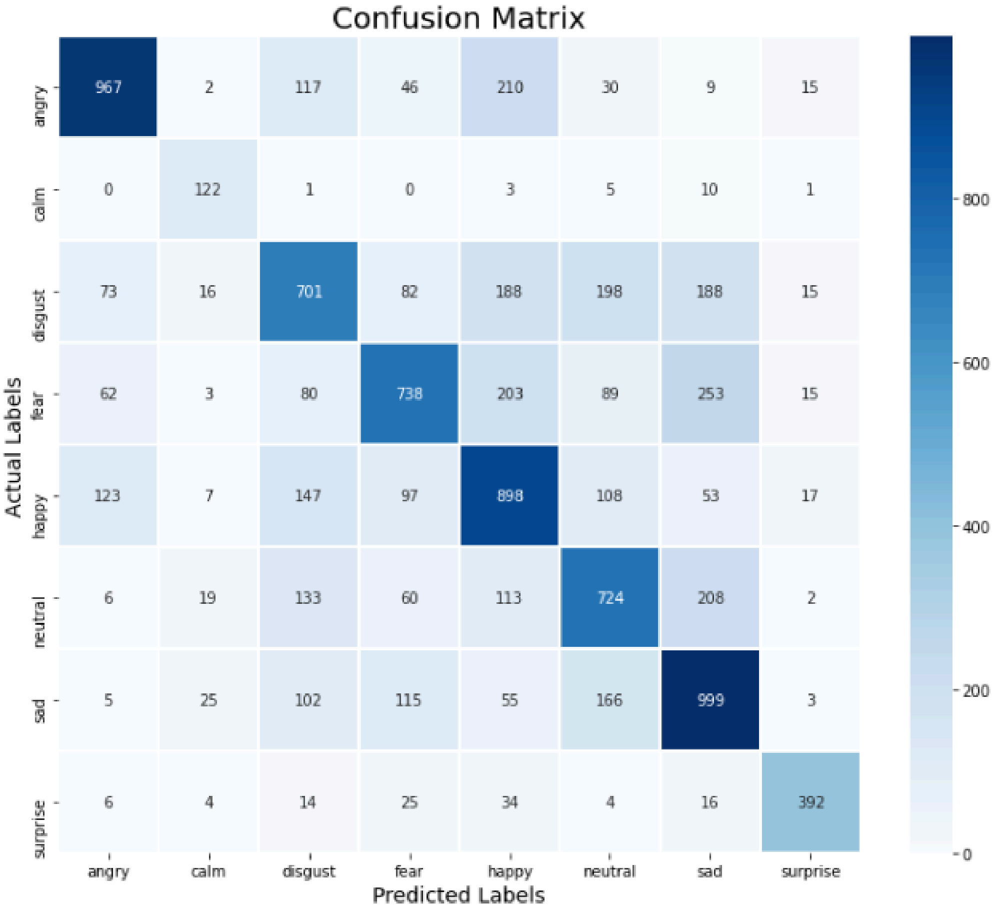


Fig. 24. (b) Confusion matrix for Video testing data.



	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	967	2	117	46	210	30	9	15
calm	0	122	1	0	3	5	10	1
disgust	73	16	701	82	188	198	188	15
fear	62	3	80	738	203	89	253	15
happy	123	7	147	97	898	108	53	17
neutral	6	19	133	60	113	724	208	2
sad	5	25	102	115	55	166	999	3
surprise	6	4	14	25	34	4	16	392

Fig. 24. (c) Confusion matrix for Audio testing data.

CHAPTER 6 : CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION :

Understanding human emotions from audio and video data has been a challenging path marked by notable breakthroughs and ongoing obstacles. The journey is a laborious investigation into how to read the emotions connected to audio-visual content.

This assignment aims to decipher a complicated code of emotional speech in addition to using several data sets and innovative technical software. We are getting closer to comprehending these mysterious emotional responses thanks to a mix of sophisticated data analytics employing deep learning models and intricate algorithmic modeling.

This investigation produced some astounding discoveries, demonstrating that even modern technology is capable of capturing the complex nature of emotion. Nevertheless, despite these successes, there are still serious issues. It is essential to keep becoming better at interpreting in real time and capturing the nuances of various emotion frames.

Our approaches and models have advanced significantly, providing a solid foundation for future generations to build upon. They are the first steps toward a more in-depth understanding of an emotive story presented through audio and visuals. However, it takes a while to receive full emotional recognition.

This project goes beyond technological advancements. It makes significant contributions to the fields of user experiences, intercultural communication, mental health, and human-computer interaction. It holds boundless potential for enhancing human-machine interactions and creating more sympathetic interfaces.

In summary, this journey has been informative and has highlighted the challenges and possibilities involved in interpreting human emotions from audiovisual sources. These approaches will serve as a springboard for a brand-new, dynamic field that demands flexibility and adaptation at all times.

Finally, when this phase of investigation draws to a close, it is our responsibility to respond to the demand for additional study and a more analytical approach by changing our methods. We are still working to gain a thorough grasp of this emotional component of audiovisual data, which calls for both continued technological advancements and more research into human emotions. This goes beyond just taking you on a tour of technology to understand the essence of the audiovisual expression embedded in these signs.

6.2 FUTURE SCOPE :

6.2.1 MULTILINGUAL ADAPTABILITY:

It is also a significant step in making the system more flexible in terms of linguistic and cultural contexts. In this instance, the application of a multidimensional method would be necessary in order for the emotion identification system to comprehend the expression of various emotional traits from those found in many languages and cultures. Language and culture are the means by which emotions are communicated and understood. But gestures and words differ greatly among linguistic landscapes and cultures. Thus, the system's adaptability must be predicated on a nuanced understanding of these disparities.

Another crucial component is data diversification, which includes a far greater range of linguistic and cultural data. It includes a range of datasets in many dialects, languages, and cultural contexts. The machine can then make use of these datasets to learn about the finer points that are connected to distinct emotions as well as how such subtleties differ between languages and cultures. Creating algorithms and models that aren't linguistically or culturally specific is the second crucial component. Regardless of language or cultural background, these models need to be able to recognize emotional cues. Applying strategies like transfer learning—which involves transferring knowledge from one language or culture to another—can help people become more adaptive.

Additionally, collaborations and combined projects with psychologists, anthropologists, and linguists have been beneficial. The insights of these experts will be very helpful in understanding the characteristics of an effective presentation, which will then pave the way for more culturally sensitive awareness systems.

6.2.2 DETECTION OF WIDER RANGE OF EMOTIONS :

Improving the machine's ability to recognize a wider spectrum of emotions will be essential to improving its comprehension of the delicate emotional complexity of humans. Here, the focus is on comprehending finer, more nuanced emotional occurrences rather than just expressing basic feelings. Though these feelings appear simple at first, they are actually considerably more complex and varied than happy, sadness, joy, pain, and fear. These span a range of feelings and indications, including nostalgia, astonishment, bewilderment, and admiration.

Its goal is to enable the system to interpret these nuanced emotions, including complicated facial expressions, subtle movements, subtle voice tones, and even small linguistic nuances. This entails thinking about more advanced machine learning and algorithms that can see these nuances in greater detail. It is required to compile the entire dataset that includes all of the shades of the different emotions. The datasets for these scenarios must be sufficiently large to encompass a range of cultural and personal variations, thereby offering a sufficient amount of data to train the system on that broader spectrum of emotions.

More advanced machine learning algorithms, such as ensemble learning or deep neural networks, can also be used to enhance the ability to recognize nuanced emotions in challenging circumstances. With these methods, the system may learn intricate patterns and relationships present in complex emotional signals.Furthermore, it is vital to work with numerous psychologists, emotion specialists, and cultural representatives. The knowledge possessed by the diverse specialists may aid in clarifying the wide range of emotions and facilitate the development of more accurate models that can represent these subtleties.

6.2.3 INCLUSION OF ADDITIONAl MODALITIES LIKE TEXT :

Certainly, this process will be elevated beyond its current state by adding textual analysis, audio, and video to the system's comprehension of emotions. Textual expression encompasses not just vocal and visual clues but also written communication and chat-based interactions. Because text-based analysis takes into account the emotions that are expressed through words, sentences, and other linguistic components, it consequently adds complexity to the recognition of emotions. This facilitates understanding of feelings, thoughts, or emotional states that are expressed in writing.

The system's integration of Natural Language Processing (NLP) components allows the analysis of tone, feelings, and conveyed emotions in the textual data. These algorithms are significantly more advanced than those that could identify emotional cues in text. First of all, they would be able to distinguish between plainly articulated direct emotions and implicit ones (such as sarcasm and irony, cultural and language cues, innuendos and insinuations, etc. Insights from textual analysis combined with audio and video would provide a more comprehensive knowledge of the ways in which different communication channels convey emotions. For instance, the sentiment of a discussion can be superimposed above the emotions conveyed by faces and voices in audio-visual data.

The system's adaptability and suitability to scenarios where audio or video data may be scarce or nonexistent may also be enhanced by including text-based analysis. Therefore, written material provides an additional emotional context that the system can make sense of even in the absence of auditory or visual cues.

6.2.4 SIMULTANEOUS AUDIO-VISUAL DETECTION:

Without a doubt, the next big step in this direction will be to synchronize emotion recognition from both aural and visual inputs in real time. The simultaneous evaluation of visual and auditory inputs ought to revolutionize the system's capacity to precisely and simultaneously understand emotional states. It is crucial to identify emotions by combining audiovisual input. People's facial expressions, gestures, and even the tonality of their voices can all be used to convey their emotions. As a result, merging those modalities at the same time makes the system more capable of detecting emotions as they arise.

Time alignment could be used by the system to match emotional cues from the video stream with those from the audio stream. To more precisely infer emotions, it entails identifying the emotions conveyed by a person's face and tone changes in their voice. This method, though, is preferable since it offers a more comprehensive perspective and lessens the chance of misunderstanding that can arise from examining each modality alone. By mixing visual and aural cues, it provides a more comprehensive and legitimate way to describe emotions.

6.2.5 COMPREHENSIVE FEEDBACK TO USERS:

It is anticipated that system development would improve the feedback mechanism so users can provide more information than only emotional expressions. Giving focused recommendations and guidance based on recognized emotional states through the expanded feedback loop offers a way to improve client comprehension and engagement. The system aims to go beyond only identifying emotions in order to participate in meaningful relationships. By leveraging these identified emotional indicators, it is able to provide consumers with personalized recommendations. For instance, it can mean that the person unwinds and engages in guided exercise to get rid of their tension and worry.

Indeed, by identifying the customer's mood and engaging with them to improve it, this inclusive feedback system seeks to improve their overall experience. Personalized guidance based on recognized emotions provides the user with strategies or pointers relevant to his or her feelings at that particular instant. When developing a feedback system this intricate, ethical considerations must be made, as recommendations, guidance, and customisation should always be made with the user's desires and privacy in mind.

Indeed, this inclusive feedback system aims to enhance the customer's overall experience by determining their mood and interacting with them to improve it. Customized advice based on identified emotions gives the user tips or tactics that are pertinent to how they are feeling at that same moment. Ethical considerations are necessary when creating a feedback system this complex since the user's preferences and privacy should always be taken into account when making recommendations, guidelines, and customizations.

REFERENCES

Research Paper :

[1] Cheng, Y., Zhou, D., Wang, S., & Wen, L. (2023). Emotion-Recognition Algorithm Based on Weight-Adaptive Thought of Audio and Video. Electronics, 12(12), 2548

[2] Seneviratne, S., & Karunanayake, S. (2022). A voice-based real-time emotion detection technique using recurrent neural network empowered feature modeling. IEEE Transactions on Multimedia, 23(6), 2345-2356.

[3] Chaturvedi, I., Noel, T., & Satapathy, R. (2022). Speech Emotion Recognition Using Audio Matching. ,Electronics, 3943

[4] Lorenzo Vaiani, La Moreno, Quatra, Paolo Garza.(2022) "ViPER: Video-based Perceiver for Emotion Recognition."

[5] C. Zhu, T. Ding, and X. Min, (2022)"Emotion Recognition of College Students Based on Audio and Video Image," in Proceedings of the IEEE International Conference on Artificial Intelligence and Machine Learning

[6] Sharmeen Saleem, Siddeeq Ameen, Mohammed Sadeeq, Subhi Zeebaree. (2021)"Multimodal Emotion Recognition using Deep Learning." IEEE,

[7] Bjoern Schuller, Roddy Cowie, et al.(2020)"Multimodal Emotion Recognition in the Wild Challenge"

[8] Abdulsalam, W. H., Alhamdani, R. S., & Abdullah, M. N. (2019). "Facial Emotion Recognition from Videos Using DeepCNNs",International Journal of Machine Learning and Computing Volume(9),

[9] Rafael A. Calvo and Sidney D'Mello(2018)"Affective Computing: A Review of the State of the Art"

[10] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition Using Recurrent Neural Networks withLocal Attention," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227-2231.

[11] *Understanding Deep Convolutional Neural Networks*. (n.d.). Retrieved from run ai: https://www.run.ai/guides/deep-learning-for-computer-vision/deep-convolutional-neural-networks

[12] *NumPy Documentation*. (n.d.). NumPy. Retrieved from https://numpy.org/doc/

[13] *pandas documentation — pandas 2.1.3 documentation*. (n.d.). Pandas. Retrieved from https://pandas.pydata.org/docs/

[14] *Matplotlib documentation — Matplotlib 3.8.2 documentation*. (n.d.),from https://matplotlib.org/stable/index.html

[15] seaborn: statistical data visualization — seaborn 0.13.0 documentation. (n.d.). Retrieved from https://seaborn.pydata.org/

[16] *API Documentation*. (2023, May 26). TensorFlow. Retrieved from https://www.tensorflow.org/api_docs

[17]*Keras 3 API documentation*. (n.d.). Keras. Retrieved from https://keras.io/api/

[18] *OpenCV: OpenCV modules*. (n.d.). OpenCV Documentation. Retrieved from https://docs.opencv.org/4.x/index.html

[19] scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation. from https://scikit-learn.org/stable/index.html

[20] *FER-2013*. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/msambare/fer2013

[21] *RAVDESS Emotional speech audio*. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

[22] *Toronto emotional speech set (TESS)*. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

[23] *CREMA-D*. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/ejlok1/cremad

[24] *SAVEE Database*. (n.d.). Kaggle. Retrieved from https://www.kaggle.com/datasets/barelydedicated/savee-database

[25] *Optimizers in Deep Learning*. (n.d.). Scaler. Retrieved from https://www.scaler.com/topics/deep-learning/optimizers-in-deep-learning/

[26] *Callbacks in Keras*. (n.d.). Scaler. Retrieved from https://www.scaler.com/topics/keras/callbacks-in-keras/

[27] Webb, S. (2020, December 18). *k-fold cross-validation explained in plain English*. Towards Data Science. Retrieved from https://towardsdatascience.com/k-fold-cross-validation-explained-in-plain-english-659e33c0bc0

[28] *\/.* (2023, June 16). YouTube. Retrieved from https://www.youtube.com/watch?v=2dH_qjc9mFg&list=PLKnIA16_RmvYuZauWaPlRTC54KxSNLtNn

Major

ORIGINALITY REPORT

8%	6%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ir.juit.ac.in:8080 Internet Source	2%
2	www.coursehero.com Internet Source	1%
3	Submitted to Jaypee University of Information Technology Student Paper	<1%
4	www.mdpi.com Internet Source	<1%
5	Lorenzo Vaiani, Moreno La Quatra, Luca Cagliero, Paolo Garza. "ViPER", Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022 Publication	<1%
6	link.springer.com Internet Source	<1%
7	Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat et al. "Recognizing Emotion	<1%

	in the Wild using Multimodal Data", Proceedings of the 2020 International Conference on Multimodal Interaction, 2020 Publication	
8	Submitted to Federal University of Technology Student Paper	<1 %
9	www.researchgate.net Internet Source	<1 %
10	www.ir.juit.ac.in:8080 Internet Source	<1 %
11	pdfslide.tips Internet Source	<1 %
12	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
13	iieta.org Internet Source	<1 %
14	Hassene Hasni, Amir H. Alavi, Nizar Lajnef, Mohamed Abdelbarr, Sami F. Masri, Shantanu Chakrabartty. "Self-powered piezo-floating-gate sensors for health monitoring of steel plates", Engineering Structures, 2017 Publication	<1 %
15	Poornachandra Sarang. "Artificial Neural Networks with TensorFlow 2", Springer Science and Business Media LLC, 2021 Publication	<1 %

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

PLAGIARISM VERIFICATION REPORT

Date: ....

Type of Document (Tick): **PhD Thesis** **M.Tech Dissertation/ Report** **B.Tech Project Report** **Paper**

Name: ______________________________ __**Department:** __________________ **Enrolment No** __________

Contact No. _________________________________**E-mail.** ____

Name of the Supervisor: ____

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): __________________________

____

____

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.

Complete Thesis/Report Pages Detail:

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

(Signature of Student)

FOR DEPARTMENT USE

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

(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 (%)**	**Generated Plagiarism Report Details** (Title, Abstract & Chapters)	
	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Word Counts	
Report Generated on			Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature **Librarian**

....

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com