

A project report on

EMOTIONAL DISPARITY

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning

by

ROANEK JENA (21BAI1125)

YEDHU KRISHNAN (21BAI1806)



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

11 November, 2024

EMOTIONAL DISPARITY

Submitted in partial fulfillment for the award of the degree of

Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning

by

ROANEK JENA (21BAI1125)

YEDHU KRISHNAN (21BAI1806)



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

11 November, 2024



VIT[®]

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

DECLARATION

I hereby declare that the thesis entitled "EMOTIONAL DISPARITY" submitted by **YEDHU KRISHNAN (21BAI1806)**, for the award of the degree of Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of **Dr. MANJU G (53009)**.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date: 20/11/2024

Signature of the Candidate



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

DECLARATION

I hereby declare that the thesis entitled “**EMOTIONAL DISPARITY**” submitted by **YEDHU KRISHNAN (21BAI1806)**, for the award of the degree of Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning, Vellore Institute of Technology, Chennai is a record of bonafide work carried out by me under the supervision of **Dr. MANJU G (53009)**.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Chennai

Date:

Signature of the Candidate



VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled "EMOTIONAL DISPARITY" is prepared and submitted by **YEDHU KRISHNAN (21BA11806)** to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning** is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr. Manju G.

Date: 18.11.24

Signature of the Examiner

Name: D. Rekha

Date: 20.11.24

Signature of the Examiner

Name: R. DHANALAKSHMI

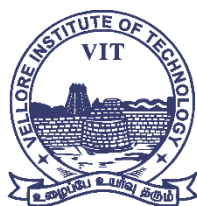
Date: 20.11.24

Approved by the Head of Department,
B.Tech. CSE with Specialization in
Artificial Intelligence and Machine Learning

Name: Dr. Sweetlin Hemalatha C

Date: 20.11.24





VIT[®]
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

School of Computer Science and Engineering

CERTIFICATE

This is to certify that the report entitled “**EMOTIONAL DISPARITY**” is prepared and submitted by **YEDHU KRISHNAN (21BAI1806)** to Vellore Institute of Technology, Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology in Computer Science and Engineering with Specialization in Artificial Intelligence and Machine Learning** is a bonafide record carried out under my guidance. The project fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Signature of the Guide:

Name: Dr. Manju G.

Date:

Signature of the Examiner

Name:

Date:

Signature of the Examiner

Name:

Date:

Approved by the Head of Department,
**B.Tech. CSE with Specialization in
Artificial Intelligence and Machine Learning**

Name: **Dr. Sweetlin Hemalatha C**

Date:

ABSTRACT

The COVID-19 epidemic has increased the use of online treatment, but it has also brought attention to its shortcomings in comparison to in-person therapy sessions. Online therapy is a practical and easily available alternative, but it does not have the same depth of nonverbal clues and subtle emotional expressions as in-person sessions. This study attempts to close this gap by examining the emotional differences that occur between a user's speech and facial expressions during therapy sessions. It is based in the disciplines of Speech Processing, Natural Language Processing (NLP), and Psychology.

The system will independently analyze the other's covert nature-their speech pieces and emotional expressions-teeth to teeth, using a specifically built dataset constructed by machine learning models. These analyses shall then be compared to reveal and characterize the deviations between them, which could signal underlying emotional conflicts or mismatches. Thus, these findings will be summarized further by an assessment model, allowing easier idioms for the therapist reflecting more on clients' emotional states. The project seeks to provide better emotional awareness for online therapy sessions, thus making virtual counseling an even more appropriate option as an alternative to face-to-face therapy.

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. Manju G., Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part to work with an intellectual and expert in the field of Natural Language Processing.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor, Dr. T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

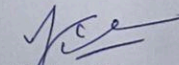
Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. Sweetlin Hemalatha C, Head of the Department, B.Tech. CSE with Specialization in Artificial Intelligence and Machine Learning and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staff at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date: 20/11/2024


Yedhu Krishnan

ACKNOWLEDGEMENT

It is my pleasure to express with deep sense of gratitude to Dr. Manju G., Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, for her constant guidance, continual encouragement, understanding; more than all, she taught me patience in my endeavor. My association with her is not confined to academics only, but it is a great opportunity on my part to work with an intellectual and expert in the field of Natural Language Processing.

It is with gratitude that I would like to extend my thanks to the visionary leader Dr. G. Viswanathan our Honorable Chancellor, Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Dr. G V Selvam Vice Presidents, Dr. Sandhya Pentareddy, Executive Director, Ms. Kadhambari S. Viswanathan, Assistant Vice-President, Dr. V. S. Kanchana Bhaaskaran Vice-Chancellor, Dr. T. Thyagarajan Pro-Vice Chancellor, VIT Chennai and Dr. P. K. Manoharan, Additional Registrar for providing an exceptional working environment and inspiring all of us during the tenure of the course.

Special mention to Dr. Ganesan R, Dean, Dr. Parvathi R, Associate Dean Academics, Dr. Geetha S, Associate Dean Research, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for spending their valuable time and efforts in sharing their knowledge and for helping us in every aspect.

In jubilant state, I express ingeniously my whole-hearted thanks to Dr. Sweetlin Hemalatha C, Head of the Department, B.Tech. CSE with Specialization in Artificial Intelligence and Machine Learning and the Project Coordinators for their valuable support and encouragement to take up and complete the thesis.

My sincere thanks to all the faculties and staff at Vellore Institute of Technology, Chennai who helped me acquire the requisite knowledge. I would like to thank my parents for their support. It is indeed a pleasure to thank my friends who encouraged me to take up and complete this task.

Place: Chennai

Date:

Yedhu Krishnan

CONTENTS

DECLARATION	1
CERTIFICATE	3
ABSTRACT	4
ACKNOWLEDGEMENT	6
CONTENTS	7
LIST OF TABLES	9
LIST OF FIGURES	10
LIST OF ACRONYMS	11
CHAPTER 1	
INTRODUCTION	13
1.1 THE FIELD OF EMOTION ANALYSIS	13
1.2 PSYCHIATRIC CARE IN VIRTUAL MODALITY	13
1.3 EMOTIONAL DISPARITY	14
1.4 SCOPE	14
CHAPTER 2	
RELATED WORK	15
2.1 PSYCHIATRIC SURVEYS	15
2.2 EMOTION ANALYSIS DOMAIN	16
CHAPTER 3	
APPROACH	22
3.1 OUTLINE	22
3.2 INPUT PROCESSING	23
3.3 FACIAL COMPONENT	23
3.4 SPEECH COMPONENT	25
3.5 CONTEXTUAL AND SENTIMENTAL COMPONENT	26
3.6 INFERENCE GENERATION	28
3.7 THE OUTPUT	29
CHAPTER 4	
RESULTS	31
4.1 FACIAL COMPONENT	31
4.2 SPEECH COMPONENT	38
4.3 CONTEXTUAL AND SENTIMENTAL COMPONENTS	42

4.4 INFERENCE GENERATION	42
4.5 THE OUTPUT	43
CHAPTER 5	
CONCLUSION AND FUTURE WORK	50
REFERENCES	51
APPENDIX 1: USING LLAMA LOCALLY	54
APPENDIX 2: LLAMA PROMPTS	56
APPENDIX 3: GUIDE TO CITATIONS	58
APPENDIX 4: RETRAINING THE WAV2VEC2 MODEL	59
APPENDIX 5: RETRAINING YOLO MODELS	64
APPENDIX 6: USING OPENAI WHISPER AND INTEGRATING ROBERTA	66

LIST OF TABLES

Table 4.1: Training Metrics from Wav2Vec2 Model	37
---	----

LIST OF FIGURES

Fig. 3.1: Application Overview	20
Fig 3.2: YOLOV11 Architecture	21
Fig. 3.3: Wav2Vec2 Architecture	23
Fig. 3.4: Speech to text transcription model architecture(Open AI Whisper)	24
Fig. 3.5: Text sentiment analysis model architecture (RoBERTa)	24
Fig. 3.6: Llama 3.1 Architecture Overview	26
Fig 3.7: Llama 3.1 Metrics Comparison	27
Fig. 4.1: Output for Video Facial Emotion recognition using Yolov11	29
Fig. 4.2: Test data (taking 14 sec video from the youtube podcast “anthony padilla”)	30
Fig 4.3: YOLO11 Architecture	31
Fig. 4.4: CNN architecture	32
Fig. 4.5: Train and validation accuracy and loss(CNN)	33
Fig. 4.6: Train and validation accuracy and loss(YOLOv11)	33
Fig. 4.7: Confusion Matrix(CNN)	34
Fig. 4.8: Confusion Matrix(YOLOv11)	34
Fig. 4.9: Confusion Matrix for Wav2Vec2 Model	36
Fig. 4.10: Loss Graph for Speech CNN	37
Fig. 4.11: Accuracy Graph for Speech CNN	38
Fig. 4.12: Speech CNN Architecture	39
Fig 4.13: Output for Speech to text Transcription	40
Fig 4.14: Output for Text sentiment analysis	40
Fig 4.15: Inference Output 1	41
Fig. 4.16: Inference Output 2	41
Fig 4.17: Patient answering question 1	42
Fig 4.18: Output of Facial Emotional Analysis of the patient’s video	42
Fig 4.19: Output of the Speech to text Transcription	43
Fig 4.20: Output of the Text sentiment analysis	43
Fig. 4.21: Output of SER	43
Fig 4.22: Patient answering question 2	45
Fig 4.23: Output of Facial Emotional Analysis of the patient’s video	45
Fig 4.24: Output of the Speech to text Transcription	46
Fig 4.25: Output of the Text sentiment analysis	46
Fig. 4.26: Output of SER	46

LIST OF ACRONYMS

FER	Facial Emotion Recognition
SER	Speech Emotion Recognition
PC	Psychiatric Care
HCI	Human-Computer Interaction
ER	Emotion Recognition
CBT	Cognitive Behavioral Therapy
GNH	Global Mental Health
IEMOCAP	Interactive Emotional Dyadic Motion Capture
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
3DCNN	3D Convolutional Neural Network
FBP	Factorized Bilinear Pooling
SVM	Support Vector Machine
EmotiW	Emotion Recognition in the Wild
EEG	Electroencephalography
CSP1_X	A module in the YOLOv5 model
RAF-DB	Real-World Affective Face Database
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
HMM	Hidden Markov Model
MFCC	Mel-Frequency Cepstral Coefficients
Log-Mel	Logarithmic Mel Spectrogram
SER	Speech Emotion Recognition
EMO-SUPERB	Emotion Speech Understanding and Processing Benchmark

SSLM	Self-Supervised Learning Model
Word2Vec	A model for word embedding (semantic features)
Speech2Vec	A model for speech embedding (semantic features)
SEWA	Socio-emotional Web Affect (dataset)
UA	Unweighted Accuracy
WA	Weighted Accuracy
HuBERT	Hidden-Unit BERT
SOTA	State of the Art
CCC	Concordance Correlation Coefficient
MMD	Multimodal Deep learning
JAFPE	Japanese Female Facial Expression (a dataset)
Berlin Emo	Berlin Emotion (a dataset)
DNN	Deep Neural Network
LDA	Linear Discriminant Analysis
KNN	K-Nearest Neighbors
Bi-RNN	Bi-directional Recurrent Neural Network
DCNN	Deep Convolutional Neural Network
TESS	Toronto Emotional Speech Set (a dataset)
LLM	Large Language Model

Chapter 1

Introduction

1.1 THE FIELD OF EMOTION ANALYSIS

Emotion Analysis is a widely researched and studied field with endless papers, models, etc. Corporate companies use emotion and sentiment analysis as a market research tool to gauge customers' opinions of their company and/or products more metrically for feedback to project teams. This field has endless datasets, models, competitions, and problem statements

There are various subdomains in emotion analysis. One of the most well-known ones is image/video-based emotion detection and analysis, specifically for emotions portrayed by faces. The relationship between facial expression and its change is studied concerning the emotions expressed by an individual. There are many applications of FER which include but are not limited to driver safety in vehicles, healthcare for patients with communication disability, etc.

Text-based sentiment analysis is one of the oldest and most practically applied research domains under the umbrella of emotion analysis. Social media companies use it to gather user sentiment under certain topics of discussion that are trending. Product-based companies process reviews of their products left by customers with the help of text sentiment analysis for market research purposes. Sentimental analysis is generally done to draw out, as the name suggests, the sentiment of a piece of text with respect to the words used in it.

Speech emotion recognition is not as fully-fledged as the earlier two subdomains we discussed due to the many factors plaguing it. Generally, there have been many approaches to study the emotions portrayed in speech. The main approaches seen in papers are either feature-based, content-based, or a combination of both. Feature-based SER refers to using audio features such as spectrograms, envelopes, etc to identify emotions while content-based SER is usually conducted by converting the speech to text followed by sentiment analysis.

1.2 PSYCHIATRIC CARE IN VIRTUAL MODALITY

Psychiatric care refers to the professional evaluation and treatment of mental health conditions. Many people struggling emotionally or having undiagnosed mental disorders view getting mental help negatively from a social standpoint, in (2.1) we look at studies that extensively surveyed such perceived barriers that young people face when deciding whether or not to obtain PC.

With the advent of the Internet an attempt to make PC accessible and more available was and is still being made, our project aims to aid this industry to further expand its effectiveness and alleviate some of the challenges that come with therapy in

virtual modes. As already mentioned we will look into surveys and studies on online therapy later in (2.1). COVID-19 had led to a boom of online therapy as people who attended physical sessions had to move over to attending them online, many challenges were faced in this including a loss of connection between the therapist and the patient. Our project also aims to bring back this connection by giving information to the therapist as they would get in a physical session.

1.3 EMOTIONAL DISPARITY

We define emotional disparity as the difference in emotion portrayed through a person's speech, facial expressions, and other attributes from which we can derive one's emotional state of mind. In a vis-a-vis session, a psychiatrist would be able to discern these nuances and be able to formulate appropriate responses to their patient but in an online session these attributes are lost and not entirely dependable due to many factors such as internet connectivity, audio clarity through microphone and speakers in a call and limited visibility of the patient.

1.4 SCOPE

While this project aims to bridge the gap between therapists and their patients, it can fundamentally be used in any psychiatric application including, but not limited to psychoanalysis of interrogation targets, applicant screening for visa and passport purposes, therapy and market research.

Psychoanalysis of interrogation targets is an application which is very close to our main therapeutic application as it is also mainly about analyzing the nuances of a person and the emotions they portray during their exchanges with another. Applicant screening currently is done through a self declared form in most cases but with our application an applicant can be screened with confirmation and confidence. As far as market research is concerned our application can be used to analyze different emotions a certain piece of media product invokes on a person such as films and games.

Chapter 2

Related Work

Emotion recognition as well as emotion disparity have emerged as emerging research topics especially with the availability of pressing advances in machine learning and multimodal analysis for mutually beneficial HCI as well as mental health aid. The identification of how emotions are conveyed and interpreted through different channels, including vocal and facial structures, is essential when developing ER models, as well as in any situations where emotional expression may vary significantly from person to person or when nonverbal cues are masked. This paper analyzes multiple papers that have relevant information about emotion recognition systems as for speech and faces, concerns, techniques, and the incorporation of novelties. The review also briefly discusses such new learning areas as psychotherapy and, specifically, psychotherapy after COVID-19 when emotions are no longer controlled but constructed through virtual media and other practices.

2.1 PSYCHIATRIC SURVEYS

The pandemic of COVID-19 has influenced the field of psychotherapy by increasing the use of teletherapy, as well as experiencing certain benefits and risks. Writing about patients' views on online therapy, Giordano et al. [16] use qualitative data obtained from 51 patients who have invested in both individual and group therapy. They identify four key domains impacting patient experience: in an online environment, therapy content, outcomes of therapy sessions, therapist-patient interactions, and interactions between the individuals in group therapy. The study has significant results indicating the enhanced relevance of personal space visibility for a group therapy proposed in virtual environments, which implies other relations in new media. Furthermore, they identify important limitations like technical difficulties and loss of paralinguistic cues, and also stress that more investigation on the topic of group online therapy is needed – an area still rather investigated.

In a similar vein, Békés et al. [17] examine the difficulties that therapists experience when confronted with the necessity to perform teletherapy during the pandemic. Based on the survey of 1,257 therapists across the world, they raise issues of intimacy, patients' privacy, and boundaries of interactions. This suggests that while issues of emotional connection are key drivers of fear about internet delivered therapy, therapists who are fearful about losing connection with clients think less favorably about internet delivered therapy. In particular, some barriers, including privacy concerns, seem to decrease over time, while distractions-related problems remain persistent with regards to adaptation concerns. Using the study of Békés et al., it is possible to develop new training and supervision strategies for mental health professionals for the purpose of continuing online therapeutic activities.

Wang et al. [18] use a systematic review and metaanalysis of 25 randomized controlled trials of Internet Based self help interventions for young people, involving 4,480 participants. Their analysis shows that their adolescents and college students who participated with these interventions, have significant reductions in anxiety, depression, and stress symptoms; particularly so for longer interventions, which include more than eight weeks. Wang et al. also highlight the importance of guidance in interventions along these lines, as well as the particular efficacy of third wave CBT interventions. Results from the study suggest that college students derive more benefit from these programs than adolescents, suggesting that demographic factors are an influential determinant of outcomes. However, the study's finding corroborates the possibility of digital self-help programs, but Wang et al. proposed further research to alleviate dissonances in the literature as well as improving the intervention efficiency.

A broad perspective on adolescents and young adult mental health help-seeking behaviors is provided by Gulliver et al. [19] via a systematic review on barriers and facilitators. They analyze data from 22 studies and identify stigma, embarrassment, and low mental health literacy as the most important barriers that prevent help seeking, yet positive past experiences and social support as facilitators. Gulliver et al.'s suggested changes are interventions that increase mental health literacy, reduce stigma, and support young people to do what they want to do. It is important to emphasize that understanding of young people's attitudes towards mental health and increasing help seeking behavior among this demographic is relevant to streamline mental health.

Doan et al. [20] return to this question, exploring the reasons why healthy Canadian secondary school students will not seek the benefits of mental health support in the COMPASS study, where over 47,000 students are surveyed. GNH asks students whether they are afraid to approach adults with mental health questions, and if so, why, and finds that 58 percent of them are reluctant to seek help from adults for mental health problems, wherein these hesitations are stronger among those with lower self rated mental health, less family and peer support as well as weaker school connectedness. In addition, the study shows variations in how likely students are to seek help based on school location and socioeconomic context: students in lower density communities and middle income schools are less averse. Importantly, these insights emphasize the efficacy of creating supportive school environments wherein help seeking behaviors are promoted in order to improve mental health outcomes for youth.

2.2 EMOTION ANALYSIS DOMAIN

Other recent studies have endeavored to propose various approaches and strategies for integrating audio and visual signals, as well as acknowledging their synergism. One of the most significant works in this field is provided by M. Singh and Y. Fang in [1], wherein a framework using a deep neural network for the emotion recognition based on audio and video inputs is presented and described. Their work using the IEMOCAP dataset tested different modes including CNN, CNN+LSTM, and CNN+RNN for audio-based emotion recognition. Of these, CNN+RNN model was determined to be the most accurate in identifying happy, anger, sad and neutral emotions

with an average accuracy of 54 percent. They also expanded their model by including video frames into multiple streams as CNN+RNN+3DCNN and described that the performance improved dramatically to 71.75% of accuracy for three emotions. These studies with the multimodal approach proved the importance of combining audio features with visual features for recognizing emotions, while the authors mentioned the problems of data imbalance especially for the sector of happy emotions and the problems connected with the occlusion of face in the video clips. Furthermore, they explained that noise removal was not beneficial for enhancing the accuracy, and background noise could actually help out in the training process, promising directions for further research about noise control and model scaling in the multimodal learning framework.

Picking up the pace in this direction, H. Zhou et al. in the paper[2] proposed the multilayer structure that involved advanced feature extraction and fusion for audio-video emotion recognition. According to their method, they applied CNN to extract features from the spectrogram of the given audio and also to obtain face features from the frames of the video. To improve the performance of the system, they formulated inter- and intra-modal fusion techniques with familiar techniques like Self Attention, Relation Attention and Transformer Attention concentrating on important emotional aspects. For cross-modal fusion, they presented the factorized bilinear pooling (FBP) as a practical solution for the fusion of audio and visual streams. In their experiments, they were able to achieve 65.5% of validation accuracy when using transformer-attention for intra-modal fusion and FBP for cross-modal fusion would make this approach one of the best performances during the EmotiW 2019 challenge. It focuses on the attention mechanism and new fusion approaches; however, they also highlighted its drawbacks, including the fluctuations in the model results because of the variation in the datasets and ambiguity in determining which model is best suited for a particular task.

Like Y. Fan et al. (2016) [3], we explored the use of hybrid networks consisting of CNNs and RNNs for video-based emotion recognition in a similar way. For example, as part of EmotiW 2016 Challenge, their hybrid approach which consists of CNNs for spatial feature extraction and RNNs (LSTM in particular) for temporal sequence processing made significant improvements in the representation of the coarse recognition. Additionally, 3D Convolutional Networks (C3D) were integrated which encode both spatial and temporal features from video segments to represent the dynamics of facial expressions in time. This is a second place for the hybrid model to achieve a notable accuracy of 59.02%, which beats previous methods. Still extending the recognition framework, this study also introduces audio features processed through SVMs, giving a small accuracy boost. Even though their method worked, they acknowledged the difficulty of these models being on the high computational side, and the limited fusion capacity of the trained models, pointing to future work to make models more computationally efficient, and fusion capabilities.

Works such as that of M. Soleymani et al. (2011) [4] investigate emotion recognition models, looking at the continuous evolution of these models as they explore physiological signals, music video analysis, etc. The work that they did was all around real time emotion detection to the music video, using such signals as EEG, skin response,

heart rate, breathing. Physiological data was recorded from 32 subjects as they watched 40 music video clips, each drawing from a broad spectrum of emotions. They used linear ridge regression models with leave-one-out cross-validation and significantly outperformed random guesses in predicting emotional states like arousal, valence, or dominance. We demonstrated this work by showing the potential for utilizing physiological signals in personalized emotion detection, particularly in multimedia recommendation systems. Despite these challenges (i.e., noise in the physiological signals which may degrade prediction accuracy), the authors suggest further work to improve the accuracy of such systems for more reliable real-time emotion detection.

H. Zhong et al. (2023) [5] also make an important contribution, by proposing a real-time facial expression recognition (FER) system for enhancing teaching quality assessments. Yolo5, a state-of-the-art object detection model, optimized along with attention mechanisms was used in their approach to improve the detection of teachers' facial expressions during lectures. In the CSP1_X module of YOLOv5, they introduced coordinate attention (CA), which improves the accuracy and efficiency of the entire system while also delivering a detection accuracy of 77.1 % with just 25 ms calculation time. Using this approach, the speed and accuracy were proven to be better than Faster-RCNN and Swim-Transformer. In order to train from balanced expression representation and high quality images, they used a subset of the RAF-DB dataset. The system was very efficient for real time facial expression detection but the authors clarified that since the model did not consider temporal information or any additional sources of data, such as audio or text, their model was limiting. In particular, integration of these modalities for improved FER accuracy in educational settings where nuanced emotional detection could aid teaching effectiveness was suggested by them.

It also follows that emotion recognition from speech has achieved large progress as K. Venkataramanan and H. R. Rajamohan's (2019) [6] Emotion recognition based on speech was developed to improve human computer interactions. They used the RAVDESS dataset to investigate various audio features, including Log-Mel Spectrograms, MFCCs, pitch, and so on, and compared several models such as CNNs, LSTMs, and Hidden Markov Models (HMMs). In the 14 class emotion recognition task, their study found that a four layer 2D CNN model based on Log-Mel Spectrograms achieved a top accuracy of 68%. They also pointed to feature selection as key to improving model outcomes, with simpler models with carefully selected features able to outperform more complex models. In addition, their findings indicated that including gender based classification, could further improve the recognition accuracy. The promising results, however, highlighted that the RAVDESS dataset, which is primarily composed of North American English speakers and actors, also exhibits bias that impedes generalizability of the model to other domains with respect to language and culture.

In recent advancements in speech emotion recognition (SER), Wu et al. [7] introduce "EMO-SUPERB: "The An In-depth Look at Speech Emotion Recognition," that will address reproducibility and data standardization issues that a SER faces, putting down an open source benchmark. To prevent data leakage, this benchmark defines standardized data partitioning and processing procedures for consistent dataset splits that

are common causes of data leakage. The authors identified three main obstacles in SER: as it lacked data partitioning consistency, lacked reproducibility (due to closed source implementations), and of course, does not handle nuanced natural language emotion annotations. The authors then show that they can improve the performance of these 16 self-supervised learning models (SSLMs) by relabelling complex emotion annotations with ChatGPT, increasing performance on average by 3.08%. Furthermore, EMO-SUPERB includes a publicly available codebase compatible with 15 SSLMs integrated as well as a leaderboard for comparative model analysis. Despite these limitations, authors admit to using datasets in English and Chinese, and only relying on high resource models, excluding demographic diversity and inclusivity concerning engagements in SER systems.

In their paper entitled "Speech Emotion Recognition Using Semantic Information", Tzirakis et al. [8] explore a novel approach for further improvement of SER in combining semantic and paralinguistic features from the speech. Aligned Word2Vec and Speech2Vec embeddings allow them to achieve this by capturing high level, text-like information from speech while still preserving paralinguistic cues. We developed a new attention based fusion strategy, including a disentangled attention mechanism that unifies these features into a unified representation. The time patterns in such data is then captured by an LSTM network, which is then in turn used to represent the data. On the SEWA dataset this model is evaluated and found to offer state-of-the-art valence and likability but show increased stability and generalizability when compared to previous methods. The authors recognize that despite these advancements, their model requires high computational demands to train necessitating real time applications. Future research could optimize semantic and paralinguistic feature extraction more efficiently, and could instead focus on the development of a unified model.

In 'Speech and Text Based Emotion Recognizer', Sharma [9] tackles the challenges posed by small and imbalanced datasets in SER, by creating a balanced corpus of five available datasets and employing different techniques to augment. We investigate several deep learning architectures and end up with a more complex multi-modal model that outperforms the baseline by a large margin: 157.57 vs. 119.66 in Unweighted Accuracy (UA) and Weighted Accuracy (WA). Our key findings demonstrate that aggregating audio with text embeddings produces significant improvements in emotion classification accuracy. Sharma contributed a standardized, larger dataset and an exploration of transfer learning techniques with models such as HuBERT. But dataset biases could affect generalizability to other emotional contexts and the model's performance may be difficult to replicate for non-attended emotional contexts in real world usage.

As proposed, the paper is a multimodal, modular emotion recognition framework utilizing speech, text and motion capture data. The researchers applied particular deep learning architectures developed for each modality, concatenated them in the last classification layer, and obtained robust results without retraining the model if one modality is missing. Despite training individual modality models before their fusion, their model achieved an accuracy of 71.04%, which is close to SOTA performance in

correlating features and even underlines the usefulness of training the individual modality models in the first place. This framework supports an interesting combination of modalities in any combination they may desire, while their open source code base encourages further research. New model limitations include dataset biases that affect the model's ability to generalize, and computational expense of tuning hyperparameters makes the model not well suited for extremely large datasets with diverse emotional context.

In their paper 'Multimodal Emotion Recognition Using Visual, Audio and Textual Features,' Deng et al.[11] aim to advance long term emotion recognition with the use of visual, audio, and text features. They apply Long Short-Term Memory (LSTM) networks to their visual features and analyze utterance-level audio and text features based on video clips. First, the proposed model outperforms unimodal baselines according to concordance correlation coefficients (CCC) of 0.400 for arousal and 0.353 for valence, implying the advantage of multimodal fusion for emotion recognition. By leveraging this early fusion approach, we effectively merge modalities for enhanced accuracy, and set a foundation for future affective computing research. As a result, the authors also recognize the existence of possible biases in the dataset and significant computational challenge of training a MMD multimodal architecture in particular due to the high memory and processing power demand.

Chennoor et al. [12] propose a model for emotion detection from audio and video signals for positive impact in human-computer interaction (HCI), and focusing on a host of communication disorders including Autism. The approach of this project is based on facial expression recognition and speech emotion analysis where they get 84 % accuracy in facial emotion classification using the JAFFE database and 78.94 % accuracy in speech emotion recognition using the Berlin Emo database. Their framework takes audio and visual analysis together to form a holistic understanding of emotions by looking at frame counts from a video clip and audio signals. The model is robust, but its ability to accurately classify overlapping emotions, and sensitivity to quality of input indicate room for improvement by improving the quality of input.

In [13], Priyasad et al. present an innovative model for speech emotion recognition using acoustic and textual data using a deep learning approach along with SincNet layer for better acoustic feature extraction. In their paper Attention Driven Fusion for Multi-Modal Emotion Recognition, the authors utilize a Deep Convolutional Neural Network (DCNN) for acoustic classification and a Bi-directional Recurrent Neural Network (Bi-RNN) and DCNN with cross attention mechanisms for their textual features. When used on the IEMOCAP dataset, this attention based fusion method leads to a 3.5% weighted accuracy improvement over existing methods. We reaffirm the importance of attention based fusion and SincNet for SER by highlighting their contributions. While they can overcome these problems, challenges remain in recognizing overlapping emotional expressions, and in quality issues with the input signals.

Sharanyaa et al. [14] in “Emotion Recognition using speech processing” use speech processing to train a deep neural network (DNN) based on audio input data with high accuracy (96%) using MFCC, Chromogram, and Spectral contrast. The results show how the DNN outperforms traditional classification algorithms such as LDA and KNN when it comes to automatic emotion detection. This model has applications in robotics, as well as for its potential in human-computer interaction. Limitations regarding the cultural influence on emotion recognition and dataset diversity are recognized by the authors, who suggest that future research will be able to improve the generalizability of the model.

Recent advancements in cross-lingual speech conversion have led to innovative frameworks that enhance communication across language barriers. A notable contribution in this domain is presented by A. Tathe, A. Kamble, and S. Yadav in their paper titled "End to End Hindi to English Speech Conversion Using Bark, mBART, and a Finetuned XLSR Wav2Vec2" [1]. Their research introduces a comprehensive approach that integrates three cutting-edge technologies: XLSR Wav2Vec2 for Automatic Speech Recognition (ASR), mBART for Neural Machine Translation (NMT), and Bark for Text-to-Speech (TTS) synthesis. The authors detail how the XLSR Wav2Vec2 model, trained extensively on multilingual datasets, effectively recognizes spoken Hindi, while mBART translates the recognized text into coherent English. The final component, Bark, synthesizes the translated text into natural-sounding English audio. The framework's end-to-end nature not only streamlines the conversion process but also significantly enhances the accuracy of speech translation compared to traditional methods that rely on separate ASR and NMT systems. The results showcase practical applications in various fields, including media production and real-time communication tools for travelers. However, the authors acknowledge challenges related to computational demands and model efficiency, suggesting avenues for future research that could focus on optimizing these models for broader accessibility and deployment in portable devices. This work exemplifies the potential of integrating advanced machine learning techniques to facilitate seamless multilingual communication in an increasingly interconnected world.

Overall, it is shown that these papers cover many different emotion recognition methodologies, from multimodal approaches merging audio, visual, and textive data for greater classification accuracy and more general applicability in different domains. The studies contribute unique insights and models to coping with dataset limitations, computational demands, and the difficulty of emotion representation. Improving the robustness and accessibility of emotion recognition technologies is dependent upon additional progress in data standardization and multimodal fusion, as well as advancements in feature extraction.

Chapter 3

Approach

3.1 OUTLINE

As highlighted in (1.3) and (1.4), the main objective of this project is to investigate how two aspects of an individual, namely, facial expressions and speech emotions differ. In order to achieve this, a multimodal approach has been chosen. This approach can be viewed in the following general overview described in Fig 3.1 below; further details will be given in the next sections.

The first of such steps involves pre-processing to deduce a successive video stream of a patient's interview before separating its audio segment. Every item is then passed to the respective processing models to produce ER for segment length that is predetermined in advance. These outputs are then aggregated and channeled to an inference generator. The generator merges the results and presents a summary where any differences between the identified emotions from the faces and the one expressed in the speech at the same particular time is accentuated. It helps facilitate an evaluation of affective fluctuations across two formats of emotional communication.

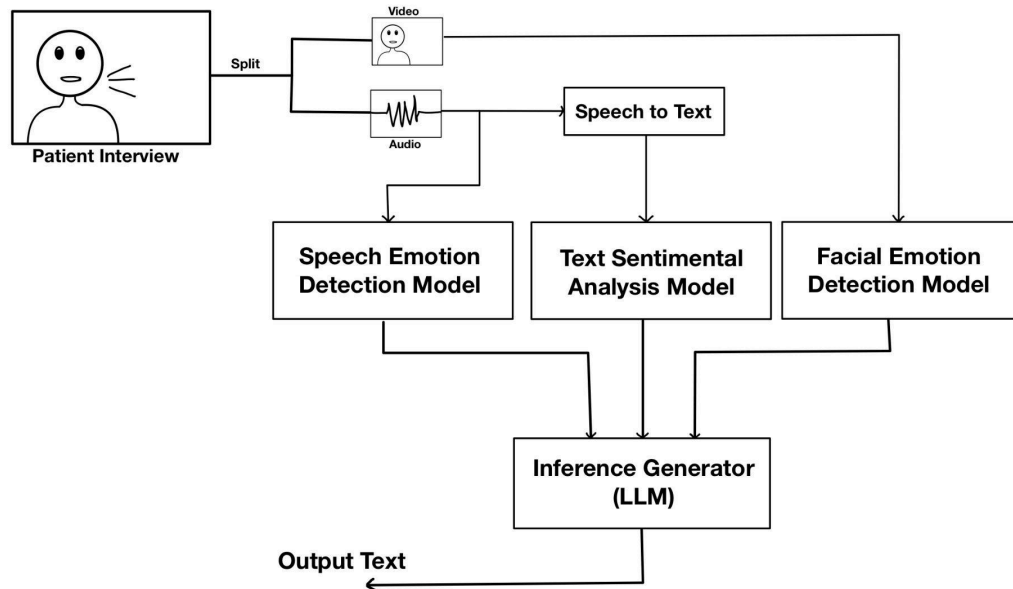


Fig. 3.1: Application Overview

3.2 INPUT PROCESSING

The expected input for this process should be a video of the interview with a patient; although it would be preferable for the video to be limited to the face and only the audio of the patient. This is necessary because of the constraints of the model training process to which reference will be made in section (4). Also, the input should also comprise time divisions which are mentioned above as particular time intervals. If no specific time divisions are assigned then the default interims will be two secs.

Complete time units are complete time divisions within the interview that are used by the model to analyze interview segments. Although it is possible to analyze each fraction of a second of the video in terms of facial expression recognition, this approach is impossible for speech analysis. That is why the number of complete spoken phrases is limited – they contain enough context to identify the emotions conveyed. This way the model assures adequate capturing of verbal and non-verbal information within the divided time segments as to produce accurate emotion outputs for the corresponding segments.

3.3 FACIAL COMPONENT

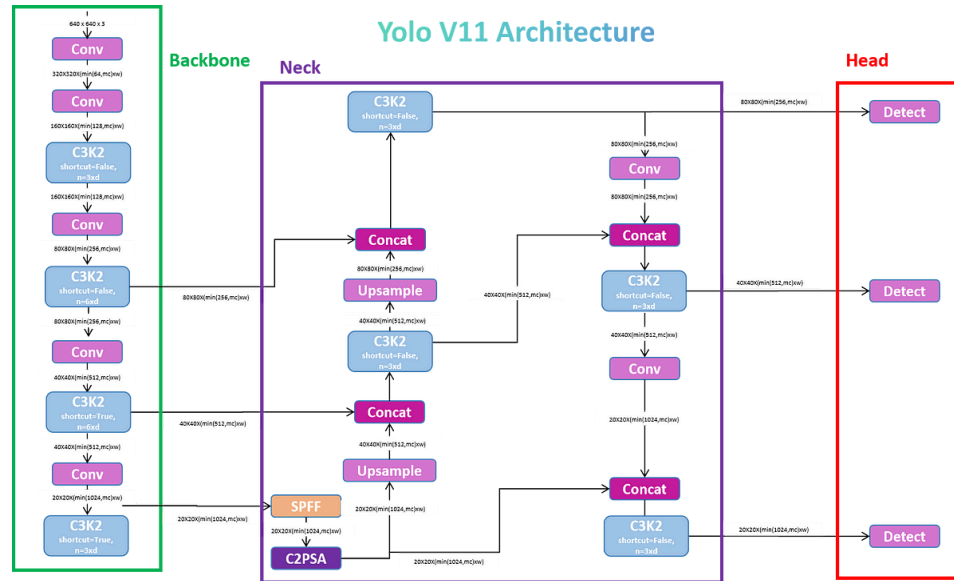


Fig 3.2: YOLOV11 Architecture

As for the architecture of the YoloV11 it relies on previous YOLO (You Only Look Once) models and does really great in the areas of both speed and accuracy of the model. One of the things YOLOv11 exploits is a modified backbone network to produce feature maps from input images (usually a modification of CSPDarknet53 or some other network optimized for performance). Following this, a set of convolutional layers and attention mechanisms are used to improve feature extraction and reduce computational cost relative to the backbone. Moreover this model operates on a multi-scale feature

fusion, which makes the model good at detecting objects at different sizes. The newly introduced detection head in YOLOv11 is more efficient, requiring bounding box coordinates, objectness scores, and class probabilities to be output. It also utilizes techniques such as using PANet (Path Aggregation Network) to improve feature propagation and spatial resolution. In addition, the performance of YOLOv11 is further improved with improvements in non-maximum suppression (NMS) and anchor box optimization, reflecting its powerful attributes for real time object detection.

For the process of facial emotion recognition in videos using the YOLOv11, image classification method begins with comprehensive dataset preparation and segment analysis. The CK+ (Cohn-Kanade) dataset was selected for training the model as it has extensive annotations of facial expressions which makes it ideal for emotion classification tasks and obtain the best performance. Each of the images in the dataset represents 7 sets of emotions including sadness, anger, fear, contempt, surprise, disgust, and happiness. All images were preprocessed by resizing them to a standard input dimension compatible with YOLOv11(416 x416) such that the input size of all images will be the same and will be taken into account during training as well as to ensure best possible model performance.

Specifically, YOLOv11 was trained in the image classification mode, which showed strong performance with high resolution images, which are especially suitable for performing detailed facial emotion classification. This model was trained on the CK+ dataset to identify facial features that correspond to each annotation of emotion. On the training-side we have adjusted the hyperparameters like learning rate, batch size and epoch count for maximizing accuracy. This training was aimed at developing a video based emotion detector model which will generalize effectively to new data, which is crucial for a reliable video based emotion detection system.

As for the inference and emotion profiling, a sliding window technique was applied for video inference, processing video in overlapping 2.5 second segments to improve prediction accuracy and maintain continuous emotion profiling. Frames were resized for each segment to 416x416 and sent through the YOLOv11 model within which probability scores were generated for each emotion in each segment. To produce an emotion profile of the segment, these scores were then averaged. Then scores were normalized, using the Softmax function, to determine the main emotion for each piece of the video. To gain insight over the emotion shifts in the time, a code was written to use robust logging and error handling method and it was done so by implementing it with the “VideoEmotionDetector” class which gives the framework structured and adaptable way of usage in other video analysis applications.

3.4 SPEECH COMPONENT

To compare the performance of methods used in this paper, several models were tested to measure their ability to make emotional inference from speech. After careful consideration, the model that best fit out requirements was found to be the HuggingFace Wav2Vec2.0 model. This model was retrained using the TESS dataset where emotions were categorized into 7 classes of emotions. Details of how well the new retrained model will perform, is outlined in (4.2).

This retraining process was undertaken at a sampling rate of 16,000 Hertz with the maximum input length being equal to 44,100, which in fact is equal to approximately 27 seconds of audio. This duration was chosen because of the versatility that is offered despite a wide input range allowable patrol durations. Given that the TESS dataset is composed of slightly more than 13 thousand audio files, each approximately 2 seconds in length on average, the chosen configuration is more than adequate. Also, the limitation of the input length eliminates situations when a long audio input might hinder the recognition of the overall emotional framework and its components.

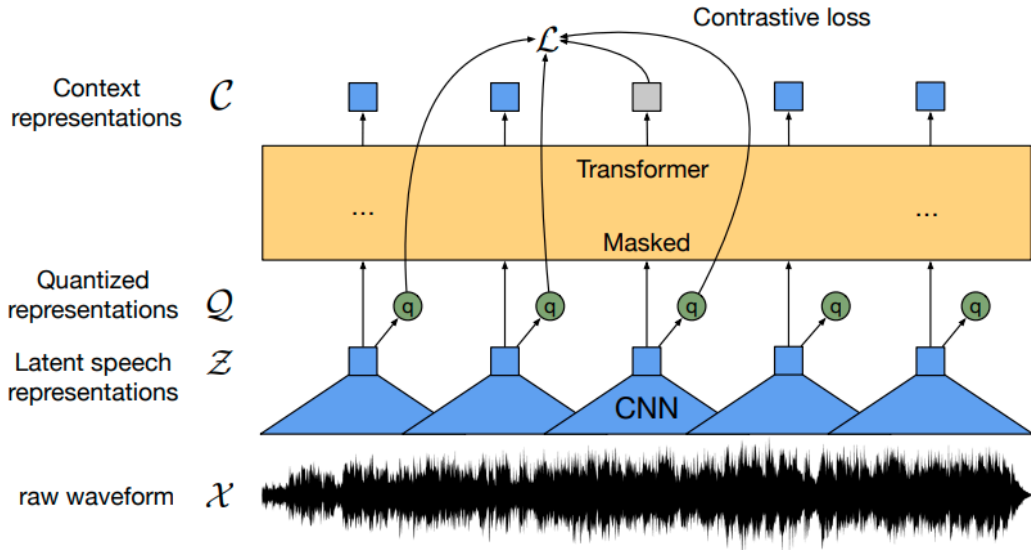


Fig. 3.3: Wav2Vec2 Architecture

3.5 CONTEXTUAL AND SENTIMENTAL COMPONENT

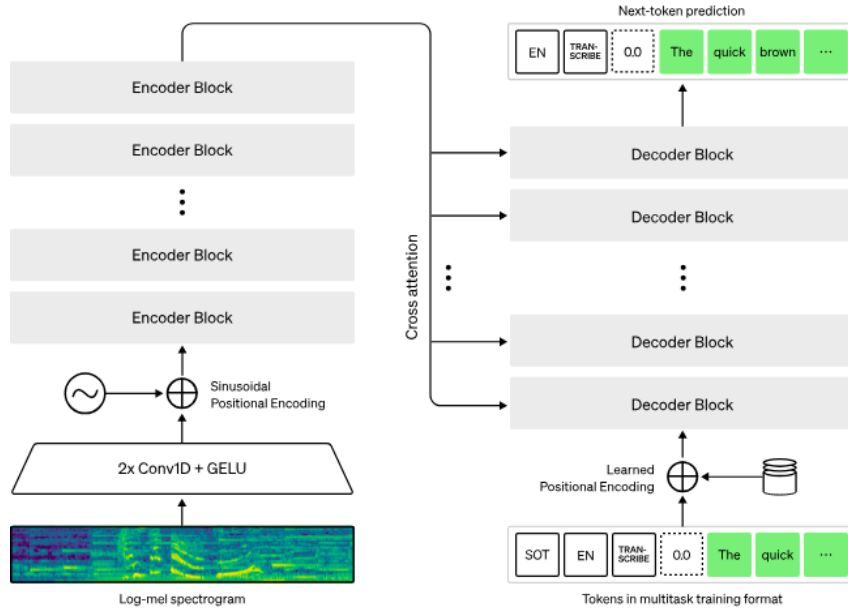


Fig. 3.4: Speech to text transcription model architecture(Open AI Whisper)

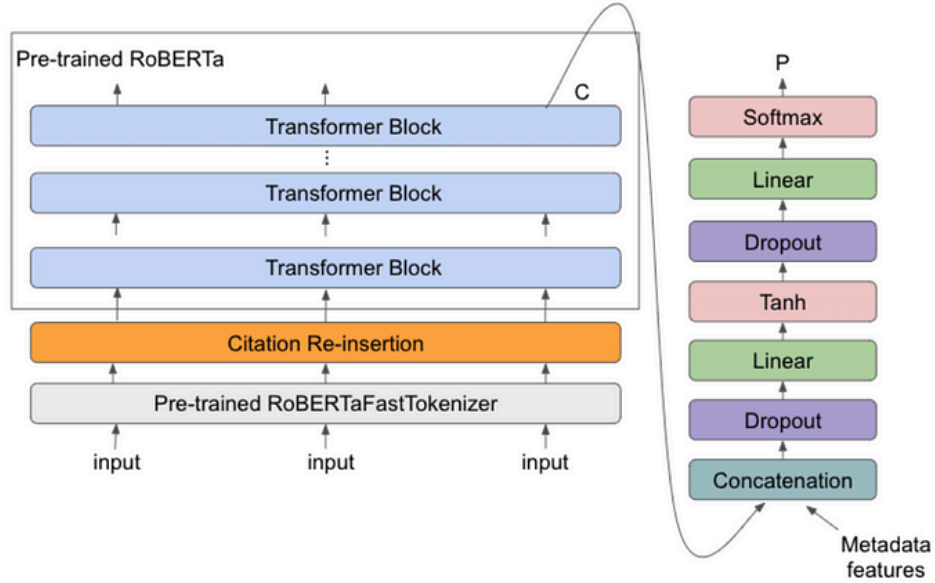


Fig. 3.5: Text sentiment analysis model architecture (RoBERTa)

As for the architecture of OpenAI's Whisper, it is a fast and efficient model used for automatic speech recognition (ASR). Whisper, is based on a transformer architecture that is particularly adept at processing sequential data such as speech, but must capture long range dependencies in the input audio. Following a sequence to sequence

framework, the model consists of an encoder processing the input audio features, and a decoder producing the transcription. A feature extraction network is employed by the encoder to convert raw audio to a set of representations, whose representations are passed through a series of transformer layers to model the underlying structure of speech. Attention mechanisms are used to align input features to output tokens in a stepwise generation of the transcription. Though smaller than other models, Whisper-small is trained on a rich multilingual data set resulting in the model's effectiveness with multiple languages and accents. We designed it both as lightweight and fast for real time transcription, or even in near real time, and while maintaining high accuracy across the board of varying audio inputs.

As for our custom roberta model, it is initially based on the RoBERTa model, an effective variant of the BERT architecture, fine tuned for emotion classification tasks (SamLowe/roberta-base-go_emotions). Its architecture is a transformer and it uses self attention mechanisms to extract contextual relationships between words in the sentence. It pretrains the model on tons of text data and great tunes it knowing the emotions given in the GoEmotions dataset, that includes various emotional tags. This model consists of multiple transformer layers which learn to represent increasingly abstract aspects of the input text. The model's weights are fine-tuned so it learns emotion specific patterns and can tell us happiness, sadness, anger etc. In samlowe/roberta-base-go_emotions classification head on top of transformer layers, output logits predict emotion labels. As this architecture supports the more efficient classification of text into one of several emotional categories based on what the model learns about language and sentiment.

Now for our contextual speech-to-text sentiment analysis, we begin with taking the audio from the recorded video clips that are inputted and transcribed as text for their sentiment evaluation, thus starting the contextual speech-to text sentiment analysis. To generate this transcription process, we're using an automatic speech recognition (ASR) model OpenAI's Whisper('openai/whisper-small') to produce the accurate text representation of the spoken phrases. This captures the speaker's precise words and structure, so cues to emotions, in verbal form, are maintained. This step ensures that phrases are taken into account rather than risk missing conversations due to partial speech segments.

After transcribing, we analyze the text against a sentiment classification model 'SamLowe/roberta-base-go_emotions'. This model applies a detailed classification process, in which sorting of each sentence is done to determine those which might be the expression of probable emotions, through an analysis of the linguistic content of that sentence alone. Sentiment analysis provides a list of emotions with respective confidence scores that indicate the likelihood of expressing that emotion in the text. Using these scores, the system creates an emotional profile which is independent from facial expression, based on the speaker's language.

This emotional data is based on text, and is used as a basic input to a large language model (LLM) in the latter stages of analysis. By integrating text based emotions using facial analysis for obtaining visual emotional cues it allows the LLM to compare and interpret the verbal as well as non verbal signals. It supports contextual integration of the input to discover consistent or conflicting emotional cues between speech and expressions, which help provide precise and complete understanding of a speaker's, rather than prompted, emotional state.

3.6 INFERENCE GENERATION

For inference generation we required an LLM with specific requirements due to reasons like being subject to possible patient privacy laws:

- It should run offline with no communication outside the network of our application.
- Large enough context length to support the vast outputs that are generated by the other connecting model.
- Run practically quick enough for an interactive experience with the user, i.e., the psychiatrist in this case.

An LLM that satisfied all these requirements for us is Meta's Llama 3.1 8B model. This model presents us with a context length of 128,000 tokens which will allow us to input vast outputs from our ER models.

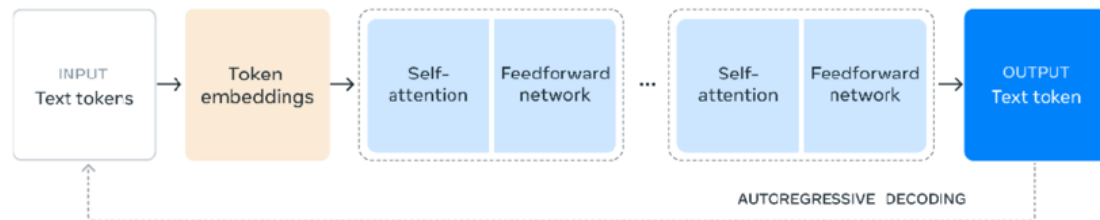


Fig. 3.6: Llama 3.1 Architecture Overview

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mistral 8x22B Instruct	GPT 3.5 Turbo
General						
MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code						
HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math						
GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4

Fig 3.7: Llama 3.1 Metrics Comparison

3.7 THE OUTPUT

The final output is a text which puts together emotional insight from both speech and facial expression analysis to form a holistic understanding of what the actual emotional state of an individual is. An automatic speech recognition (ASR) model is used to initially type out speech (transcribed), which is then split between classification of the emotions expressed during speech (sentiment analysis). In parallel, a separate emotion recognition model analyzes facial expressions in order to determine the emotions expressed through the person's given face.

The LLM then builds with both of the sets of outputs, namely emotion labels generated from the speech transcription, and emotion labels generated from the facial expression analysis. It ingests these multi modal emotional signals, and integrates both verbal and non-verbal signals to strive to solve any discrepancies between the two. For example, if the speech says one thing and the face another, if that's a mismatch like where the speech expresses one emotion, but the face another, the LLM will take that into account and resolve the conflict by figuring out these contextual clues and patterns in the data. Consequently, a detailed text representation of the individual's true emotion in a synthesized form out of numerous sources of information, a form that provides a more detailed understanding than speech or facial expression alone.

With this integrated approach we can have a better, more holistic understanding of that emotional state in increasingly complex situations where emotion cues may not line up perfectly, such as when someone hides their emotion by their words and facial expressions are subtle or ambiguous.

Chapter 4

Results

4.1 FACIAL COMPONENT

```
Final Results Summary:  
Segment 1/28: anger (0.208)  
Segment 2/28: anger (0.178)  
Segment 3/28: anger (0.166)  
Segment 4/28: anger (0.179)  
Segment 5/28: anger (0.175)  
Segment 6/28: happy (0.187)  
Segment 7/28: happy (0.208)  
Segment 8/28: happy (0.211)  
Segment 9/28: happy (0.215)  
Segment 10/28: happy (0.214)  
Segment 11/28: happy (0.198)  
Segment 12/28: fear (0.196)  
Segment 13/28: fear (0.213)  
Segment 14/28: fear (0.187)  
Segment 15/28: sadness (0.169)  
Segment 16/28: sadness (0.179)  
Segment 17/28: sadness (0.174)  
Segment 18/28: happy (0.173)  
Segment 19/28: anger (0.165)  
Segment 20/28: anger (0.184)  
Segment 21/28: anger (0.185)  
Segment 22/28: fear (0.197)  
Segment 23/28: fear (0.242)  
Segment 24/28: fear (0.274)  
Segment 25/28: fear (0.279)  
Segment 26/28: fear (0.265)  
Segment 27/28: fear (0.252)  
Segment 28/28: fear (0.226)
```

Fig. 4.1: Output for Video Facial Emotion recognition using Yolov11

In the facial emotion detection component of our "Emotion Disparity" analysis, we have used the YOLOv11 model to detect and classify different emotions based upon facial expressions in segmented (video) frames. To predict a primary emotional state, each segment was analyzed to classify from one of the 7 emotions like anger, happiness, sadness, and fear. For each emotion, the model further encompassed it with confidence scores indicating the likelihood to be there.

For example, from our dataset, anger and fear were dominant emotions across several segments while happiness and sadness occasionally surfaced indicating changing emotion states.



Fig. 4.2: Test data (taking 14 sec video from the youtube podcast “anthony padilla”)

In this section, we will also do a comparative analysis of two models implemented on the task of facial emotion recognition with the CK+ dataset: The designed MGI class imbalance loss for neural networks, combined with an advanced YOLOv11 based model, and a conventional CNN model. The YOLOv11 model was trained for over 100 epochs, and is our main model in consideration as the model has advanced architecture designed with more targeted object detection problems and truthfully, it is possible that the model is especially designed to be best suited for facial emotion recognition.

4.1.1 COMPARATIVE ANALYSIS

The development stage of the much popular YOLO architecture is YOLOv11. The object detection algorithm it uses is a single-stage paradigm that speeds up localization and classification. With real time location and more emphasis on location for facial emotion recognition, this model is very handy. It is composed of feature extraction architecture with a head that can predict objects given the features extracted from the image. YOLOv11 includes the state-of-the-art components of CSPNet, path aggregation, and the attention mechanism and is expected to promote the capability of facial feature extraction and robustness to changes in facial features.

The enhanced backbone and feature pyramid network (FPN) assist YOLOv11 in the critical factor of perceiving subtle expressions by increasing spatial awareness and fine feature learning. These architectural improvements enable YOLOv11 to surpass regular CNNs in terms of both accuracy and speed and are particularly advantageous in tasks requiring a high level of image analysis.

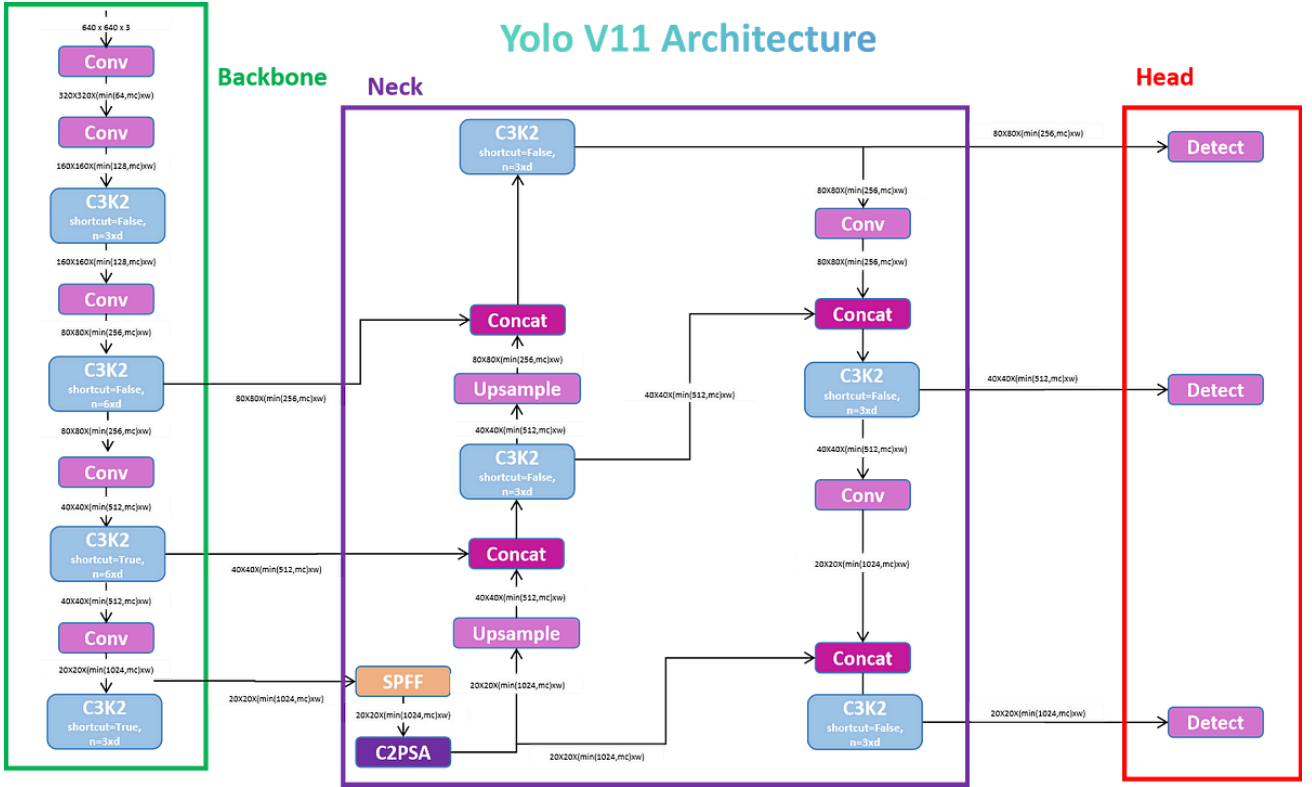


Fig 4.3: YOLO11 Architecture

The architecture shown in Fig. 4.4 is for a custom designed (CNN) model used for classification. There are a sequence of convolutional layers, then a batch normalization and dropout layers to help to stabilize learning and avoid overfitting. Input images are passed through the model using 'Conv2D' layers that increase in filter size (32, 64, 128, and 256) to extract spatial features.

'MaxPooling2D' layers are used after each convolutional layer to reduce spatial dimensions and to provide us with many learnt hierarchical features that make the network much more efficient. A Flatten layer unrolls the 2D feature maps on the 2nd dimension, then the output is fed into a dense layer with 256 units, then we read another output layer with 7 units (for 7 labels of emotion). After a dense linear layer, dropout layers are used to prevent overfitting, making this architecture ideal for complex image classification problems.

Model: "sequential_4"

Layer (type)	Output Shape	Param #
conv2d_16 (Conv2D)	(None, 98, 98, 32)	896
batch_normalization_16 (BatchNormalization)	(None, 98, 98, 32)	128
max_pooling2d_16 (MaxPooling2D)	(None, 49, 49, 32)	0
dropout_20 (Dropout)	(None, 49, 49, 32)	0
conv2d_17 (Conv2D)	(None, 47, 47, 64)	18,496
batch_normalization_17 (BatchNormalization)	(None, 47, 47, 64)	256
max_pooling2d_17 (MaxPooling2D)	(None, 23, 23, 64)	0
dropout_21 (Dropout)	(None, 23, 23, 64)	0
conv2d_18 (Conv2D)	(None, 21, 21, 128)	73,856
batch_normalization_18 (BatchNormalization)	(None, 21, 21, 128)	512
max_pooling2d_18 (MaxPooling2D)	(None, 10, 10, 128)	0
dropout_22 (Dropout)	(None, 10, 10, 128)	0
conv2d_19 (Conv2D)	(None, 8, 8, 256)	295,168
batch_normalization_19 (BatchNormalization)	(None, 8, 8, 256)	1,024
max_pooling2d_19 (MaxPooling2D)	(None, 4, 4, 256)	0
dropout_23 (Dropout)	(None, 4, 4, 256)	0
flatten_4 (Flatten)	(None, 4096)	0
dense_4 (Dense)	(None, 256)	1,048,832
dropout_24 (Dropout)	(None, 256)	0
dense_5 (Dense)	(None, 7)	1,799

Fig. 4.4: CNN architecture

The two models were benchmarked against CK+, an established gold standard benchmark in facial emotion recognition with a large variety of facial expressions in an enormous range of emotional states, making it an

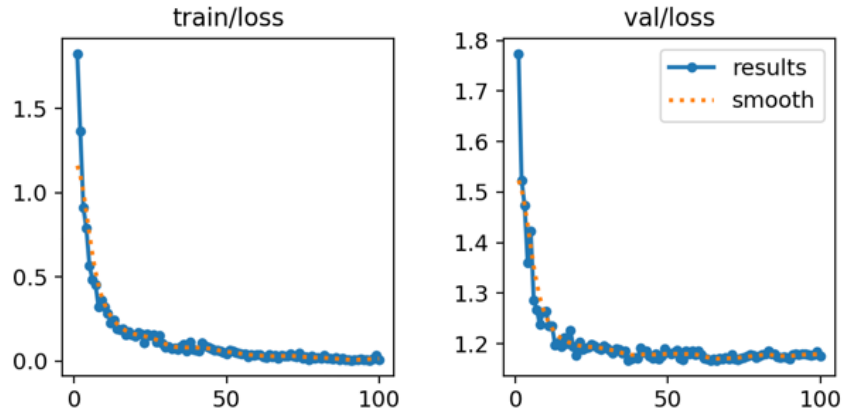


Fig. 4.5: Train and validation accuracy and loss(CNN)

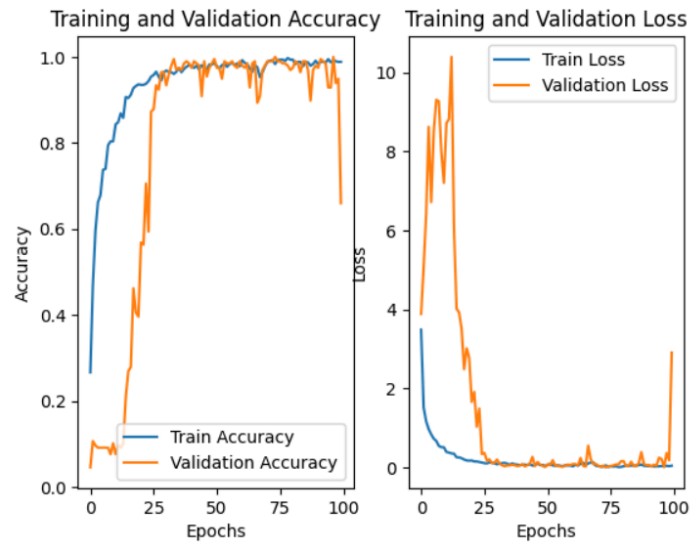


Fig. 4.6: Train and validation accuracy and loss(YOLOv11)

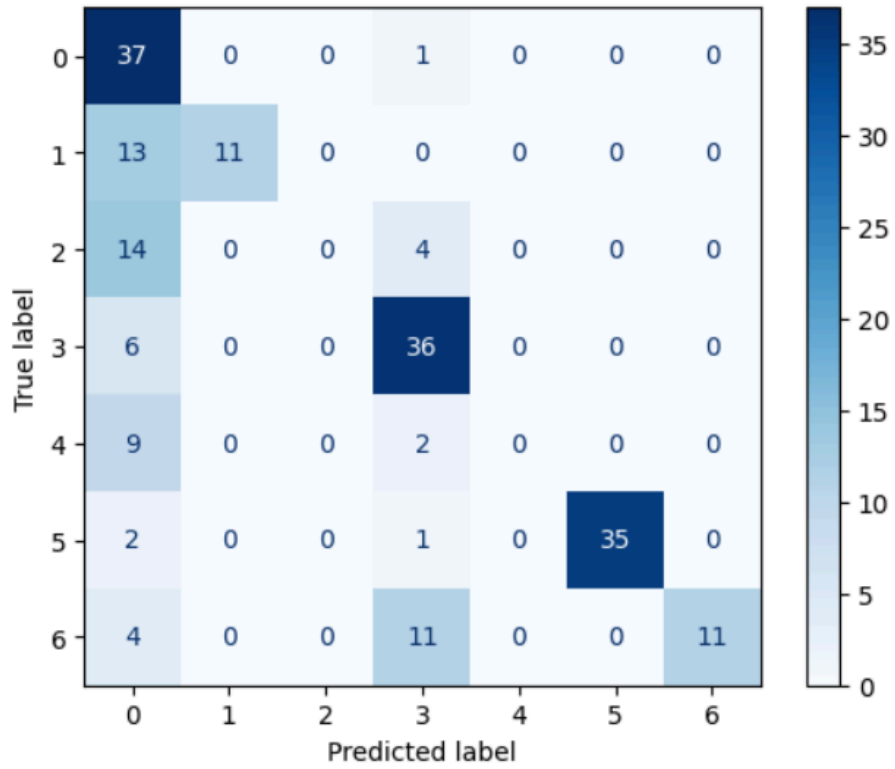


Fig. 4.7: Confusion Matrix(CNN)

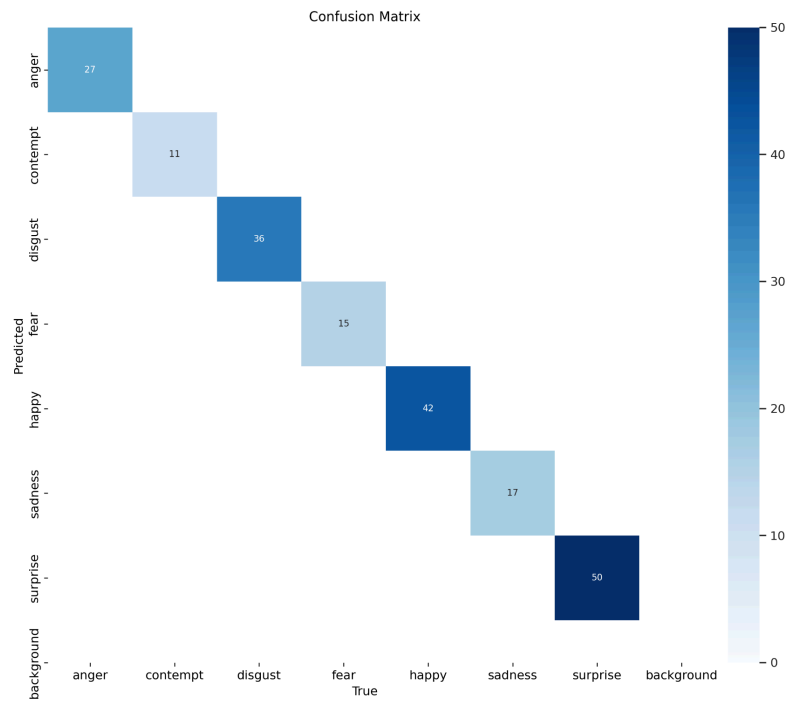


Fig. 4.8: Confusion Matrix(YOLOv11)

- Accuracy: Each of the confusion matrices indicated that the YOLOv11 gave a more accurate result than the CNN model. The confusion matrix CNN Figure 1 shows that many of the emotions such as contempt-label 1, disgust-label 2, fear-label 3, etc. got wrongly classified for some emotion sometime. Compared to this, Figure 2 shows the confusion matrix of YOLOv11, with much accuracy and very few misclassifications and it can distinguish even small differences in facial expression.
- Error Analysis: It's also unable to differentiate close facial expressions that might develop different strengths of emotions. The words 'contempt', 'disgust' and 'fear' can be sighted as errors. It does not have sufficient capabilities to aggregate the high level features and multi scale processing to discriminate the fine differences, and most probably it is due to the two major deficiencies. As almost all categories in the YOLOv11 diagonal confusion matrix have high values almost all along its diagonal, YOLOv11 is far more robust.
- Speed and Real-time Application: In a real time application of emotion recognition, a single stage detector, the proposed YOLOv11, is able to give superior speed in a real time application. Utilizing a sequential structure, the CNN model will need to spend more computation time than optimal for such applications that require quick response as the CNN model does not provide features aggregation simultaneously.
- Feature Sensitivity: The attention mechanisms allowed YOLOv11 to focus more on emotion relevant facial parts such as mouth, eyes and eyebrows and who play a major role in determining the different emotions. It's a brilliant model for the CNN, but it's also sensitive to small variation differences or the small mix-ups which occur in the expressions.

4.2 SPEECH COMPONENT

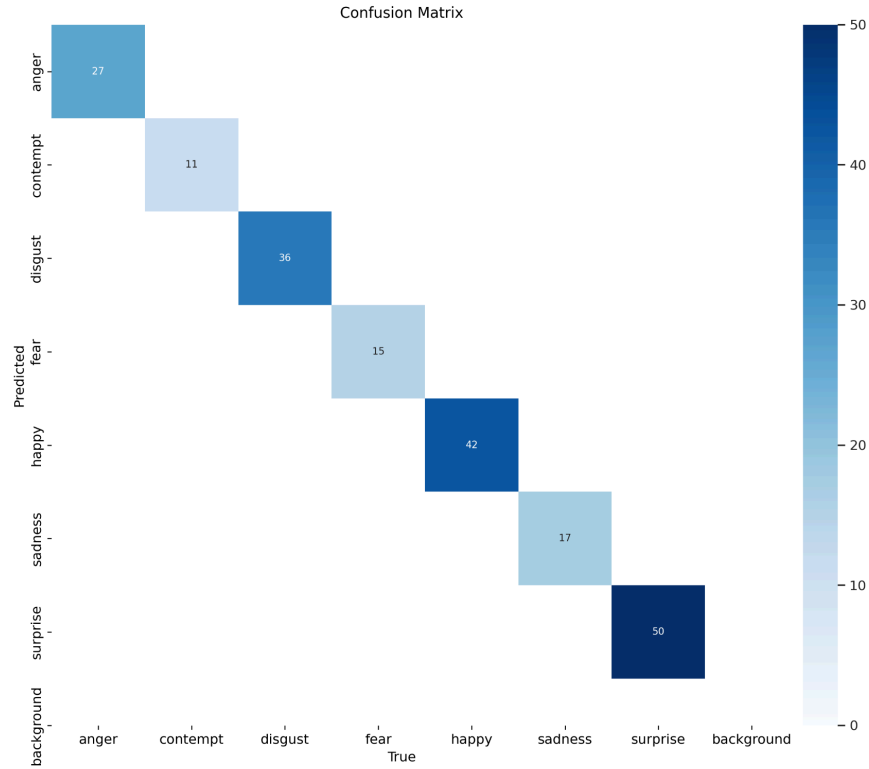


Fig. 4.9: Confusion Matrix for Wav2Vec2 Model

Similar to the video model a CNN model was designed in the beginning with the architecture in Fig. 4.10. It gave an accuracy of 60% over test data which when compared to our Wav2Vec2 model, retrained for 10 epochs gave the following final metrics on the validation data on the last epoch:

- Accuracy: 99.5%
- Precision: 0.99569
- Recall: 0.995238
- F1: 0.995238
- Validation Loss: 0.808182

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	1.888824	0.422619	0.702325	0.422619	0.349180
2	No log	1.757516	0.559524	0.753224	0.559524	0.517443
3	No log	1.578214	0.783333	0.869718	0.783333	0.750756
4	No log	1.353363	0.853571	0.893563	0.853571	0.847287
5	1.739600	1.146761	0.944048	0.954515	0.944048	0.942213
6	1.739600	1.007324	0.975000	0.976415	0.975000	0.974787
7	1.739600	0.909754	0.990476	0.990735	0.990476	0.990455
8	1.739600	0.851228	0.992857	0.993039	0.992857	0.992859
9	1.078500	0.817563	0.996429	0.996479	0.996429	0.996429
10	1.078500	0.808182	0.995238	0.995269	0.995238	0.995238

Table 4.1: Training Metrics for Wav2Vec2 Model

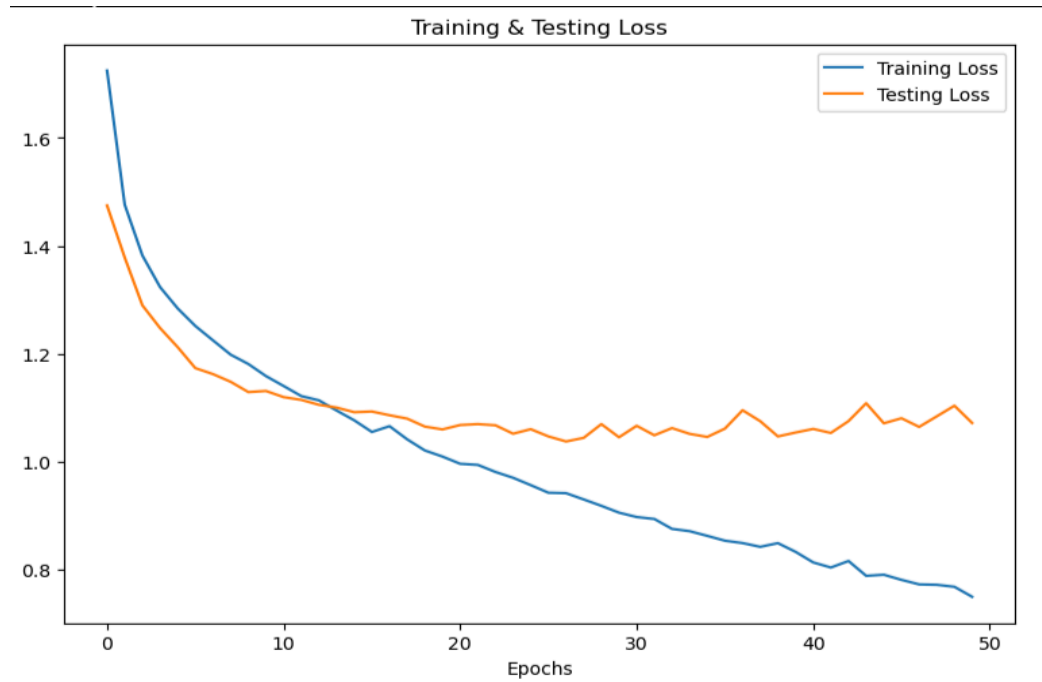


Fig. 4.10: Loss Graph for Speech CNN

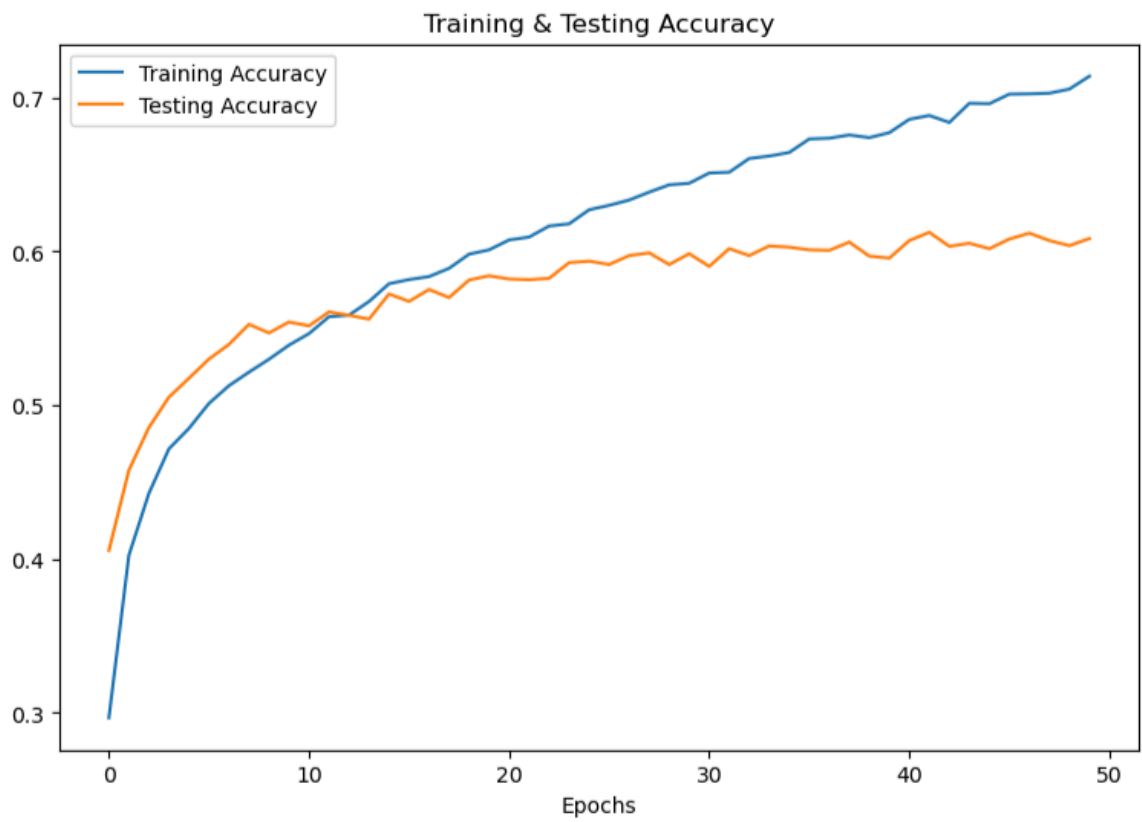
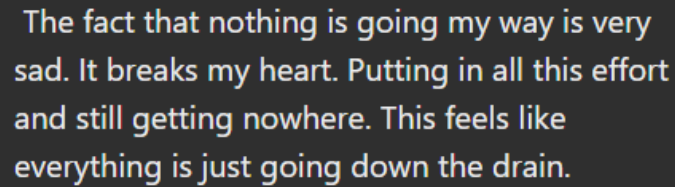


Fig. 4.11: Accuracy Graph for Speech CNN

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 256)	1536
max_pooling1d (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 256)	327936
max_pooling1d_1 (MaxPooling1D)	(None, 41, 256)	0
conv1d_2 (Conv1D)	(None, 41, 128)	163968
max_pooling1d_2 (MaxPooling1D)	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41024
max_pooling1d_3 (MaxPooling1D)	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 8)	264
=====		
Total params: 557,288		
Trainable params: 557,288		
Non-trainable params: 0		

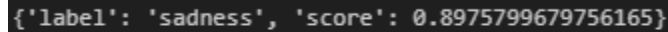
Fig. 4.12: Speech CNN Architecture

4.3 CONTEXTUAL AND SENTIMENTAL COMPONENTS



The fact that nothing is going my way is very sad. It breaks my heart. Putting in all this effort and still getting nowhere. This feels like everything is just going down the drain.

Fig 4.13: Output for Speech to text Transcription



```
{ 'label': 'sadness', 'score': 0.8975799679756165 }
```

Fig 4.14: Output for Text sentiment analysis

On the audio data we will apply a contextual sentiment analysis pipeline of speech to text to analyze the emotional content of transcribed speech. Then, using OpenAI's Whisper model for ASR, the pipeline ran the audio inputs from the mic through and transcribed the text. After this, the transcribed text was then fed through a sentiment classification pipeline (aka "go_emotions", an RoBERTa based model), to classify the emotional tone of the speech. Labels which were accompanied by scores of emotional states like sadness, joy, frustration, etc. were identified as the model modeled the emotion states.

Surprisingly, this was an important feature in understanding the speaker's emotional speaking state, for being combined with facial expression data and taken into an LLM to predict the speaker's true emotional state. As a result, the speech-to-text sentiment analysis was a robust, context aware feature which augmented the overall emotion recognition system.

4.4 INFERENCE GENERATION

Initially when the application was still being developed we tested out Llama 3.1 with the prompt given in (A2.1) and produced the following outputs with the test sample inputs given in (A2.2):

The analysis for the patient interview:

In this particular video segment, there appears to be a discrepancy between the emotions detected on the patient's face and the emotions expressed through their speech. The facial emotion detection algorithm identified 'happy', 'sad', and 'disgust' in 2.5-second segments of the video, respectively.

However, when examining the patient's speech, it reveals a more consistent tone, with both instances showing 'sad'. This suggests that the patient may be experiencing some underlying emotional distress or struggle, which is not necessarily reflected on their facial expression. The facial emotion detection might have picked up subtle moments of discomfort or unease, such as 'disgust', whereas their verbal response conveys a sense of sadness.

The only instance where speech and face match is when the patient expresses happiness, both in terms of facial emotion and speech. This could indicate that during this brief moment, the patient feels a surge of positive emotions, which are reflected both verbally and non-verbally.

Disclaimer: It's essential to note that these findings s complexities of human emotions and interactions.erapeutic purposes without further verification from qualified professionals. The model predictions provided here are based on limited data points and might not accurately capture the

Fig 4.15: Inference Output 1

The analysis for the patient interview:

Upon comparing and contrasting the detected emotions, it is observed that there are discrepancies between the facial expressions and spoken emotions. The facial expression 'happy' at 2.5 seconds is inconsistent with the corresponding speech emotion of 'sad', suggesting a potential incongruity between the patient's outward emotional display and their actual emotional state. Conversely, the 'disgust' facial emotion corresponds with the 'sad' speech emotion, possibly indicating a deeper emotional complexity or underlying distress that the patient may be experiencing.

The contrasting emotions are also reflective of the nuances in human emotional expression, where individuals may suppress or contradict their true feelings through verbal cues. This disparity highlights the complexities of human emotions and the limitations of relying solely on facial expressions or spoken words to accurately infer an individual's emotional state.

Disclaimer: The predictions made by this model should not be relied upon for diagnostic purposes or to inform clinical decisions. Emotions can vary greatly between individuals, contexts, and cultures, and this analysis is intended only as a general insight into the comparison of facial expressions and speech emotions in a single patient interview scenario.

Fig. 4.16: Inference Output 2

4.5 THE OUTPUT

We took a real world example of our friend answering 2 questions:

1. *How are you?*
2. *How is your relationship going?*

4.5.1 QUESTION 1

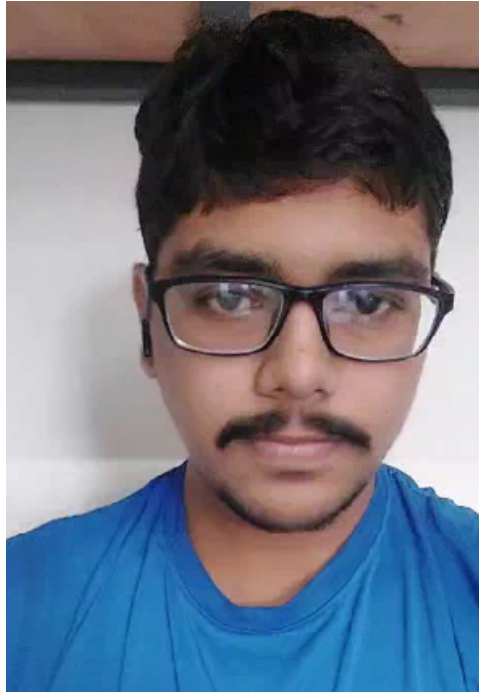


Fig 4.17: Patient answering question 1

Final Results Summary:
Segment 1/11: happy (0.267)
Segment 2/11: happy (0.270)
Segment 3/11: happy (0.244)
Segment 4/11: happy (0.227)
Segment 5/11: happy (0.226)
Segment 6/11: happy (0.259)
Segment 7/11: happy (0.265)
Segment 8/11: happy (0.269)
Segment 9/11: happy (0.268)
Segment 10/11: happy (0.270)
Segment 11/11: happy (0.254)

Fig 4.18: Output of Facial Emotional Analysis of the patient's video


```
Hello, I have been triggered. However, there have been some things which have happened in the past week which have made me very stressed and anxious especially about my work. So hence why I am here.
```

Fig 4.19: Output of the Speech to text Transcription

```
{'label': 'nervousness', 'score': 0.6214368343353271}
```

Fig 4.20: Output of the Text sentiment analysis

```
Final Results Summary:  
["neutral","neutral","happy","neutral","neutral","happy","neutral","neutral","happy","neutral","happy"]
```

Fig. 4.21: Output of SER

The final output:

“

The analysis for the patient interview:

Upon comparing the facial emotion and speech emotion, it is interesting to note that there are some discrepancies. The person's face consistently shows a happy expression throughout the video, which suggests that they may be hiding their true emotions or trying to appear more positive than they actually feel. In contrast, their spoken words reveal a more nuanced emotional state. They start by expressing being triggered and stressed about work, indicating a negative sentiment. However, when they mention past events that have caused them anxiety, the speech emotion shifts to happy, which seems inconsistent with the context. When analyzing the sentiment of their speech, it is detected as nervous, which aligns more closely with the initial statement about being triggered and stressed. This suggests that the person's facial expression may be masking their true emotional state, and they are attempting to present a more optimistic demeanor despite feeling anxious.

Disclaimer: It is essential to note that this analysis is based on machine learning model predictions and should not be solely relied upon for making conclusions or decisions about human emotions and well-being. Human emotions can be complex and multifaceted, and may not always align with the results of facial emotion detection or sentiment analysis models.

“

4.5.2 QUESTION 2



Fig 4.22: Patient answering question 2

```
Final Results Summary:  
Segment 1/17: happy (0.236)  
Segment 2/17: happy (0.233)  
Segment 3/17: happy (0.220)  
Segment 4/17: happy (0.241)  
Segment 5/17: happy (0.261)  
Segment 6/17: happy (0.259)  
Segment 7/17: happy (0.261)  
Segment 8/17: happy (0.242)  
Segment 9/17: happy (0.226)  
Segment 10/17: happy (0.244)  
Segment 11/17: happy (0.267)  
Segment 12/17: happy (0.280)  
Segment 13/17: happy (0.257)  
Segment 14/17: happy (0.227)  
Segment 15/17: happy (0.207)  
Segment 16/17: happy (0.229)  
Segment 17/17: happy (0.257)
```

Fig 4.23: Output of Facial Emotional Analysis of the patient's video

So as far as my relationship status is concerned, I have just recently got a girlfriend and I'm actually very happy about her because she's like one of the prettiest girls I know and she's also really really cute and smart and she supports me in every way that she can.

Fig 4.24: Output of the Speech to text Transcription

```
{'label': 'admiration', 'score': 0.7535213828086853}
```

Fig 4.25: Output of the Text sentiment analysis

```
Final Results Summary:  
["neutral", "neutral", "happy", "neutral", "neutral", "happy", "neutral", "neutral", "happy", "neutral", "happy"]
```

Fig. 4.26: Output of SER

The final output:

“

The analysis for the patient interview:

Upon comparing the facial emotion detected on the person's face with their speech, it appears that there is a notable discrepancy. The facial emotion consistently registers as 'happy', indicating a genuine and prolonged state of happiness in the individual. However, the speech emotion detected varies between 'neutral' and 'happy', but not consistently so. This suggests that while the person may be experiencing happiness internally, their verbal communication does not always accurately reflect this emotional state.

The context provided further supports this observation, as the person is indeed expressing admiration for their new girlfriend in a manner that is both effusive and genuine. The sentiment detected on the speech ('admiration') also aligns with the facial emotion of 'happy', suggesting that the individual's happiness is likely rooted in their positive feelings towards this relationship.

However, it is also possible that the person is exhibiting social desirability bias, where they feel compelled to present themselves as more emotionally expressive than they actually are. In any case, a closer examination of the context and speech content reveals a strong correlation between the facial emotion and sentiment, while the speech emotion detected appears to be more variable.

Disclaimer: It is essential to note that this analysis should not be taken as an absolute truth or used for making definitive conclusions about human emotions. The predictions generated by models like these are inherently probabilistic and may not always accurately capture the complexities of human emotional experiences. As such, they should be treated with caution and not relied upon as the sole basis for decision-making or judgments about individuals.

”

Chapter 5

Conclusion and Future Work

The results demonstrate that our application is capable of producing valuable inferences that can significantly aid in assessing a patient's emotional state. These inferences provide meaningful insights, enhancing the understanding of emotional cues derived from both facial expressions and speech.

Looking ahead, there are several key areas that require further development. First, it is essential to establish a comprehensive evaluation metric that can be used to verify and quantify the accuracy and quality of the inferences generated by the application. Such a metric would provide a standardized way to measure performance and identify areas for improvement.

Additionally, it is necessary to validate our findings through rigorous cross-verification to ensure consistency and reliability. This step would involve testing the application's effectiveness in practical, real-world scenarios, ideally in collaboration with industry experts. Partnering with professionals in clinical or psychological fields would provide valuable feedback and help refine the model to better meet the needs of practitioners and patients. Through these enhancements, we aim to solidify the application's utility and ensure its readiness for real-world implementation.

REFERENCES

- [1] Singh, M., & Fang, Y. (2020). Emotion recognition in audio and video using deep neural networks. arXiv preprint arXiv:2006.08129.
- [2] Zhou, H., Wang, Y., Liu, Y., & Zhang, J. (2019). Exploring emotion features and fusion strategies for audio-video emotion recognition. In Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI '19) (pp. 1-8).
- [3] Fan, Y., Zhao, Y., & Zhang, C. (2016). Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 1-8).
- [4] Soleymani, M., Zafar, A., & Pantic, M. (2011). Continuous emotion detection in response to music videos. In 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG) (pp. 1-6).
- [5] Zhong, H., Liu, X., & Chen, Y. (2023). Research on real-time teachers' facial expression recognition based on YOLOv5 and attention mechanisms. EURASIP Journal on Advances in Signal Processing, 2023(1), 1-15.
- [6] Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. arXiv preprint arXiv:1912.10458.
- [7] Wu, H., Zhang, Y., & Wang, S. (2024). EMO-SUPERB: An in-depth look at speech emotion recognition. arXiv preprint arXiv:2402.13018.
- [8] Tzirakis, P., Schuller, B., & Krajewski, J. (2021). Speech emotion recognition using semantic information. arXiv preprint arXiv:2103.02993.
- [9] Sharma, V. (2023). Speech and text-based emotion recognizer. arXiv preprint arXiv:2312.11503.
- [10] Tripathi, S., Gupta, A., & Kumar, A. (2019). Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv preprint arXiv:1804.05788.
- [11] Deng, D., Zhang, Z., & Wang, X. (2018). Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625.
- [12] Chennoor, S. N., Kumaravelan, R., & Karthikeyan, K. (2020). Human emotion detection from audio and video signals. arXiv preprint arXiv:2006.11871.
- [13] Priyasad, D., Jayasuriya, N., & Fernando, T. (2020). Attention driven fusion for multi-modal emotion recognition. arXiv preprint arXiv:2009.10991.

- [14] Sharanyaa, S., Kumaravelan, R., & Mohanraj, R. (2023). Emotion recognition using speech processing. In 2023 3rd International Conference on Intelligent Technologies (CONIT).
- [15] Bharathi, B., Manikandan, S., & Rajeshkumar, R. (2022). Findings of the shared task on speech recognition for vulnerable individuals in Tamil. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion.
- [16] Giordano, D., de Lima, M.H.M., & de Araújo, J.P.C.(2021). Patients' perspectives on online psychotherapy during the COVID-19 pandemic: A qualitative study.Frontiers in Psychology, 12(730345), 1-12.
- [17] Békés, V., Aafjes-van Doorn, K., Luo, X., Prout, T.A., & Hoffman L.(2021). Psychotherapists' challenges with online therapy during COVID-19: Concerns about connectedness predict therapists' negative view of online therapy and its perceived efficacy over time.Frontiers in Psychology, 12(705699), 1-12.
- [18] Wang Q., Zhang W., & An S.(2023). A systematic review and meta-analysis of internet-based self-help interventions for mental health among adolescents and college students.Internet Interventions, 34(100690), 1-10.
- [19] Gulliver,A., Griffiths,K.M.& Christensen,H.(2010). Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review.BMC Psychiatry, 10(113), 1-11.
- [20] Doan,N.K.A.Patte,K.A.Ferro,M.A.& Leatherdale,S.T.(2020). Reluctancy towards help-seeking for mental health concerns at secondary school among students in the COMPASS study.International Journal of Environmental Research and Public Health, 17(7128), 1-13.
- [21] Tathe, A. Kamble, A .& Yadav, S. (2023). End to end Hindi to English speech conversion using Bark,mBART and a finetuned XLSR Wav2Vec2.Proceedings of the International Conference on Machine Learning and Data Science.
- [22] Abdel-Hamid, O.Mohamed, A.Jiang, H.Deng, L.Penn, G. & Yu ,D.(2014). Convolutional neural networks for speech recognition.IEEE/ACM Transactions on Audio Speech and Language Processing, 22(1533-1545).
- [23] Ahmedani, B.K. (2011). Mental health stigma: Society individuals and the profession.Journal of Social Work Values and Ethics, 8(41–416).
- [24] Zweifel,P.(2021). Mental health: The burden of social stigma.International Journal of Health Planning and Management, 36(2),309-318.

[25] Misra, S.Jackson, V.W.Chong, J.Choe, K.Tay, C.Wong,J. & Yang,L.H. (2021). Systematic review of cultural aspects of stigma and mental illness among racial and ethnic minority groups in the United States: Implications for interventions.American Journal of Community Psychology,68(3-4),486-512.

Appendix 1: Using Llama Locally

Llama can be run locally in many ways, for full fledged access with fine tuning capabilities you will have to request access on Meta's website after which you will be granted access only if you are invited. A more commercial way to use Llama models with prompt engineering would be to use Ollama which is an open source project that allows developers and users to access LLMs such as Llama.

Instructions on installing ollama for your corresponding operating system can be found on their website ollama.com. Once installed it is recommended to run ollama through your terminal and install your desired model, this is recommended as we don't want to install the model we intend to use during a test run of our program.

- Command to start ollama service

```
>ollama serve
```

- Command to install a model, you can find the name of the model you want used by ollama on their site:

```
>ollama run <model-name>
```

This will install the model first into the ollama directory.

To use sophisticated prompts langchain can be used within your python scripts. Here is a sample script which we used to design the prompt we used in this project:

```
from langchain_ollama import OllamaLLM
from langchain_core.prompts import ChatPromptTemplate
```

```
template="""
```

```
Given is the emotion detected on a person's face and corresponding speech separately for
a video of a single person, compare and contrast the respective
emotions. The video that the emotions are extracted from is an interview conducted by a
psychologist on the patient and the emotion listed is for 2.5
second segments of the video. Only generate an explanation text, no need for lists or
tables. After the full inference state a disclaimer to not depend
upon the model predictions.
```

```
Facial Emotion: {face_emotions}
Speech Emotion: {speech_emotion}
```

Response should be in the following format:

'The analysis for the patient interview:'

Insert analysis here

Insert Disclaimer here

Response:

“””

```
model = OllamaLLM(model="llama3.1")
prompt=ChatPromptTemplate.from_template(template)
chain = prompt | model
face=["happy", "sad", "disgust"]
speech=["sad", "sad", "happy"]

for result in chain.stream({"face_emotions": face, "speech_emotion":speech, "context":c,
"sentiment":s}):
    print(result, end="", flush=True)

print()
```

This script will output the text generated by the Llama 3.1 model in a stream format, i.e, it won't wait for the final word to be generated to print, this is helpful to make the use of these models feel more interactive.

Appendix 2: Llama Prompts

Here we list the prompts we used to generate inferences from our emotion model outputs.

A2.1, Prompt 1:

“
Given is the emotion detected on a person's face and corresponding speech separately for a video of a single person, compare and contrast the respective emotions. The video that the emotions are extracted from is an interview conducted by a psychologist on the patient and the emotion listed is for 2.5 second segments of the video. Only generate an explanation text, no need for lists or tables. After the full inference state a disclaimer to not depend upon the model predictions.

*Facial Emotion: {face_emotions}
Speech Emotion: {speech_emotion}*

Response should be in the following format:

'The analysis for the patient interview:'

Insert analysis here

Insert Disclaimer here

Response:
“

A2.2, Sample Outputs 1:

*face=["happy","sad","disgust"]
speech=["sad","sad","happy"]*

A2.3 Prompt 2:

“

Given is the emotion detected on a person's face and corresponding speech separately for a video of a single person and also what they spoke and the sentiment detected on it, compare and contrast the respective segments. Only generate an explanation text, no need for lists or tables. After the full inference state a disclaimer to not depend upon the model predictions.

Facial Emotion: {face_emotions}

Speech Emotion: {speech_emotion}

Context: {context}

Sentiment: {sentiment}

Response should be in the following format:

'The analysis for the patient interview:'

Insert analysis here

Insert Disclaimer here

Response:

“

Appendix 3: Guide To Citations

Within this report there are 2 types of citations:

1. Reference Citation: These citations have been made with [] and an Arabic numeral in it which corresponds to the article being referred to. For example [4] refers to the paper 4 in the References section.
2. Section Citation: These citations have been made with () and an Arabic numeral in it which corresponds to the section and chapter number being referred to. For example (4.1) refers to chapter 4 section 1.

Tables and figures in this report are numbered using the section citation style. Hence Fig. 4.5 refers to the 5th figure in chapter 4.

Appendix 4: Retraining the Wav2Vec2 Model

The following is the code outline to retrain the HuggingFace Wav2Vec2 Model for audio based classification problems. Here we will use the TESS dataset to train the model and we will be using kaggle hence changing directory paths according to your target system.

- Import Libraries

```
import pandas as pd
import numpy as np

import os
import sys

# librosa is a Python library for analyzing audio and music.
import librosa
import librosa.display
import seaborn as sns
import matplotlib.pyplot as plt

import torchaudio
import torch
from torch.utils.data import Dataset, DataLoader
from transformers import Wav2Vec2Model, Wav2Vec2Processor, Trainer,
TrainingArguments, Wav2Vec2ForSequenceClassification

import warnings
warnings.filterwarnings('ignore')
```

- Load Dataset and create Dataset Class

```
Tess = "/kaggle/input/toronto-emotional-speech-set-tess/tess toronto emotional speech set
data/TESS Toronto emotional speech set data/"

#To support the ability to merge multiple datasets we will use this list method
file_emotion = []
file_path = []
```

#The respective code to extract data can be found at your dataset's datacard

```
tess_directory_list = os.listdir(Tess)
for dir in tess_directory_list:
    directories = os.listdir(Tess + dir)
    for file in directories:
        part = file.split('.')[0]
        part = part.split('_')[2]
        if part=='ps':
            file_emotion.append('surprise')
        else:
            file_emotion.append(part)
    file_path.append(Tess + dir + '/' + file)
```

```
paths=file_path
labels=file_emotion
```

```
df=pd.DataFrame()
df["Paths"]=paths
df["Labels"]=labels
```

#Plot the data to check for data imbalance

```
import seaborn as sns
sns.countplot(data=df,x='Labels')
```

```
#Label Encoding
label_map={label:idx for idx, label in enumerate(df['Labels'].unique())}
inverse_label_map={idx:label for label, idx in label_map.items()}
df['Labels']=df['Labels'].map(label_map)
df.head()
```

```
#Split Dataset
from sklearn.model_selection import train_test_split as tts
train_df, test_df= tts(df,test_size=0.3,random_state=42,shuffle=True)
```



```

#Dataset Class
class SpeechData(Dataset):
    def __init__(self, df, processor, max_length=44100):
        self.df=df
        self.processor=processor
        self.max_length=max_length

    def __len__(self):
        return len(self.df)
    def __getitem__(self, idx):
        path=self.df.iloc[idx]['Paths']
        label=self.df.iloc[idx]['Labels']

        speech, sr = librosa.load(path,sr=16000)

        if len(speech) > self.max_length:
            speech=speech[:self.max_length]
        else:
            speech=np.pad(speech, (0,self.max_length-len(speech)), 'constant')

        inputs= self.processor(speech,
                               sampling_rate=16000,
                               return_tensors='pt',
                               padding=True,
                               truncate=True,
                               max_length=self.max_length
                               )
        input_values=inputs.input_values.squeeze()
        return {'input_values':input_values, 'labels': torch.tensor(label, dtype=torch.long)}

#Load the Processor
processor=Wav2Vec2Processor.from_pretrained('facebook/wav2vec2-base')

#Create DataLoaders incase you want to use custom train loop
train_dataset=SpeechData(train_df,processor)
test_dataset=SpeechData(test_df,processor)
batch_size=8
train_dl=DataLoader(train_dataset,batch_size=batch_size, shuffle=True)
test_dl=DataLoader(test_dataset,batch_size=batch_size, shuffle=False)

```

- Loading and training the model

```
#Put you number of labels in parameter num_labels
model=Wav2Vec2ForSequenceClassification.from_pretrained('facebook/wav2vec2-base',
num_labels=7)
```

```
training_args= TrainingArguments(output_dir='./results',
                                evaluation_strategy='epoch',
                                save_strategy='epoch',
                                learning_rate=2e-6,
                                per_device_train_batch_size=16,
                                per_device_eval_batch_size=16,
                                num_train_epochs=10,
                                weight_decay=0.01,
                                report_to=[]
                                )
```

```
#Defining compute metrics
from sklearn.metrics import accuracy_score,precision_recall_fscore_support
```

```
def compute_metrics(pred):
    labels=pred.label_ids
    preds=np.argmax(pred.predictions, axis=1)
    accuracy=accuracy_score(labels, preds)
    precision, recall, f1,
    _=precision_recall_fscore_support(labels,preds,average='weighted')
    print( {
        'accuracy': accuracy,
        'precision': precision,
        'recall': recall,
        'f1': f1
    })
    return {
        'accuracy': accuracy,
        'precision': precision,
        'recall': recall,
        'f1': f1
    }
```

```
#Training the Model  
trainer=Trainer(  
    model=model,  
    args=training_args,  
    train_dataset=train_dataset,  
    eval_dataset=test_dataset,  
    compute_metrics=compute_metrics  
)  
trainer.train()
```

- Evaluating the model

```
results=trainer.evaluate()  
print(results)
```

- Saving the model

```
trainer.save_model("./final")
```

Appendix 5: Retraining YOLO models

YOLO has multiple types of models, in this project we have used YOLO11 in classification mode. There is a prerequisite of directory structure in case you don't have access to a *.yaml* file for your dataset. The directory structure is as follows:

- Root (Folder of your notebook/python script)
 - Data
 - train
 - <images>
 - val
 - <images>
 - test
 - <images>

Training image set should go in train folder, test ones in test folder, validation ones in val folder. The YOLO train function will automatically pick the images with this structure.

- Install Ultralytics

!pip install ultralytics

- Import Ultralytics library and initialize YOLO model

```
from ultralytics import YOLO  
model=YOLO("yolo11n-cls.pt")
```

- Training the Model

```
model.train(data='Data', epochs=<number of epochs>)
```

- Evaluate the Model

```
metrics=model.val()  
print(metrics.top1)  
print(metrics.top5)
```

- Predicting custom inputs

```
results=model.predict(<path to image or list of paths>)  
print(results[i]) #print prediction for image i, i=0 for single image input  
result[i].show() #show image i with prediction
```

- Saving the model

```
model.export()
```

Appendix 6: Using OpenAI Whisper and Integrating RoBERTa

We have used OpenAI Whisper (openai/whisper-small) which is a sequence to sequence model, for English speech to text transcription.

Then we use that text as an input for another model which will perform text sentiment analysis on that input and will give us the output in the form of an array containing emotion expressed in that text.

Both the whisper and RoBERTa model are imported from hugging-face. They are very good models for Automatic Speech Recognition and Text Classification Task respectively.

The models are imported and trained in the following manner:

- Import Libraries

```
#transformers library, versatile open-source library designed for natural language processing (NLP) tasks. It provides pre-trained models and tools for working with state-of-the-art transformer architectures like BERT, GPT, T5, and many others.
from transformers import pipeline
from transformers import WhisperForConditionalGeneration, WhisperProcessor
from transformers import RobertaForSequenceClassification, RobertaTokenizer
from google.colab import files
import shutil
```

- Import Models

```
# Load the pipeline with Wav2Vec 2.0 for ASR.
pipe = pipeline("automatic-speech-recognition", model="openai/whisper-small")

#Load pipeline for RoBERTa model.
classifier = pipeline(task="text-classification",
model="SamLowe/roberta-base-go_emotions", top_k=None)
```

- Generating Results

```
#Transcription of speech to text.(openai whisper model)
result = pipe("/content/sad.mp3")
print(result["text"])

#Sentiment Analysis of the text input.
sentences = result["text"] # This is the transcribed text
```

```

# Step 1: Classify the transcribed text
model_outputs = classifier(sentences)

# Step 2: Access the inner list in model_outputs and get the result with the highest score
if isinstance(model_outputs, list) and len(model_outputs) > 0:
    model_outputs = model_outputs[0] # Access the first (and only) inner list

    # Find the label with the highest score
    top_output = max(model_outputs, key=lambda x: x['score'])
    print(top_output)
else:
    print("Unexpected format in model_outputs:", model_outputs)

```

- Downloading models for offline use
 - OpenAI Whisper-small:

```

model_name = "openai/whisper-small"

# Download the model and processor
model = WhisperForConditionalGeneration.from_pretrained(model_name)
processor = WhisperProcessor.from_pretrained(model_name)

# Save the model and processor locally
model.save_pretrained("./whisper-small")
processor.save_pretrained("./whisper-small")

# Zip the model directory
shutil.make_archive("/content/whisper-small", 'zip', "/content", "whisper-small")

# Download the zip file to your local machine
files.download('/content/whisper-small.zip')

```

- SamLowe/roberta-base-go_emotions:

```

# Define the model and tokenizer
model_name = "SamLowe/roberta-base-go_emotions"

# Download the model and tokenizer
model = RobertaForSequenceClassification.from_pretrained(model_name)
tokenizer = RobertaTokenizer.from_pretrained(model_name)

# Save the model and tokenizer locally
model.save_pretrained("./roberta-base-go_emotions")
tokenizer.save_pretrained("./roberta-base-go_emotions")

```

```
# Zip the model directory
shutil.make_archive("/content/roberta-base-go_emotions", 'zip', "/content",
    "roberta-base-go_emotions")
```

```
# Download the zip file to your local machine
files.download('/content/roberta-base-go_emotions.zip')
```

- Run the offline Model

```
classifier = pipeline(
    task="text-classification",
    model="./roberta-base-go_emotions", #path to model safetensor
    tokenizer="./roberta-base-go_emotions", #path to model tokenizer.json
    top_k=None
)
```