



**University of
East London**

**AIN SHAMS UNIVERSITY
FACULTY OF ENGINEERING**



**Computer Engineering and Software Systems Program - International
Credit Hours Engineering Programs (I-CHEP)
Multi-Modal Emotion Recognition**

Under the Supervision of:

Dr Mohamed Rehan

Student's names:

Abdel Rahman Mohamed Salah 19P9131

Mahmoud Mohamed Seddik Hassan ElNashar 19P3374

Mohamed Hatem Zakaria Elafifi 19P7582

Omar Ashraf Mabrouk Abdelelwahab 19P8102

Omar Mohamed Ibrahim Alsayed 19P7813

Zakaria Sobhy Abd-Elsalam Soliman Madkour 19P2676

ABSTRACT:

Emotion detection is a critical aspect of human-computer interaction and artificial intelligence applications. This bachelor project explores the combination of preexisting visual and audio models to enhance the accuracy of emotion detection systems. Rather than developing models from scratch, we utilize advanced visual and audio models and investigate their synergies for a more comprehensive understanding of human emotions.

The project examines the interaction between facial expressions, extracted through established computer vision algorithms, and audio features derived from signal processing methods. Notably, this study adopts a strategy of merging a preexisting visual model with an audio model, aiming to capitalize on the strengths of each modality and achieve improved overall accuracy.

Facial landmarks are extracted using cutting-edge computer vision techniques, and audio features such as pitch, intensity, and tempo are analyzed. Machine learning models, including deep neural networks, are then utilized to combine the outputs of the preexisting visual and audio models. The project assesses the impact of this multimodal fusion on the enhancement of emotion detection accuracy.

Results in theory should indicate a significant improvement in accuracy compared to individual modalities, highlighting the efficiency of combining preexisting models. The integrated system demonstrates robustness across diverse emotional expressions and speech patterns. The project underscores the pragmatic approach of combining established visual and audio models to achieve heightened accuracy without reinventing the wheel.

In conclusion, this project advances the field of emotion detection by presenting a pragmatic approach that combines preexisting visual and audio models. The findings emphasize the effectiveness of this integration, offering a pathway for more sophisticated and accurate applications in artificial intelligence and human-computer interaction.

TABLE OF CONTENTS

Abstract:	i
List of Figures:	v
List of Tables	v
List of Abbreviations	v
Chapter One: Introduction	1
Motivation	1
Problem Statement	2
Outline	3
Chapter Two: Literature Review	4
Datasets and Benchmarks.....	4
Popular Datasets for Multimodal Emotion Detection	4
Benchmarks and Challenges.....	5
Impact on Model Development	6
Video Data Availability:	7
Fusion techniques review	9
Wav2vec2.0 (acoustic encoder)	9
RoBERTa _{BASE} (contextual encoder)	10
MMER (Multimodal Multi-task Learning for speech emotion recognition)	11
Efficient Face	12
Chapter Three: NN Architecture Review	13
Visual Model	13
i. FaceNet:	13
ii. EfficientFace:	16
Audio Model	19
i. Wav2Vec.....	20
ii. RoBERTa _{BASE}	20
Chapter Four: Solution Architecture.....	21
Project Timeline	21
Visual Model (FaceNet and EfficientFace Integration)	21
Audio Model (MMER with Wav2Vec2 and RoBERTa _{BASE})	22
Fusion Implementation.....	23

Method 1	24
Method 2.....	25
Method 3.....	26
Method 4.....	27
Method 5.....	28
Chapter Five: Software Requirements Specifications (SRS).....	29
Scope	29
Functional Requirements.....	30
Non-Functional Requirements	31
Performance Requirements	31
Reliability	31
Scalability.....	31
Use Cases	32
Chapter Six: Prototype Results	34
Chapter Seven: Concludin phase 1	35
Chapter Eight: Enhancements to Audio-Visual Emotion Recognition Model	36
Introduction	36
Limitations of Initial Model (Challenges).....	37
Limitations of Initial Model (Issues Encountered)	37
Enhanced Model Architecture with Transformers	37
Key Changes in the Enhanced Model	38
Training Process.....	38
Benefits of the Enhanced Model	39
Chapter Nine: Application Overview and Functionality	40
Application Concept.....	40
Workflow Overview.....	40
Speech-to-Text (STT) Functionality	41
Llama.....	42
Text-to-Speech (TTS) Functionality	43
Emotional Analysis Integration.....	44
Emotional Analysis Impact	44
Benefits of Emotional Analysis.....	44
Effects of Using Emotion in Application	44
Emotion Label Effect on LLM Accuracy	45

Impact of Emotion Labelling.....	45
Clinical Implications.....	45
Future Directions	45
Chapter Ten: Emotion Label effect on the LLM Accuracy	46
Case 7	47
Case 15	48
Case 27	49
Case 56	50
Case 74	51
Conclusion	52
Chapter Eleven: Project Conclusion	53
Phase 2 Summary.....	54
Problem Statement:	54
Project Summary.....	55
Problem Statement:	55
Methodology:	55
References.....	56

LIST OF FIGURES:

Figure 1: RAVDESS dataset distribution	7
Figure 2: FaceNet architecture. You can see the full details of the layer's sizes, The flow of the data and PARM, FLOPS in Table2.....	14
Figure 3: Effiecientface abstract architecture showing its pipeline.....	16
Figure 4: Local-feature extractor used in EffecientFace components.	17
Figure 5: Channel-spatial modulator compnents.....	18
Figure 6: (a) Represents the normal residual block which has a wide -> narrow -> wide structure with the number of channels. (b) Inverted Residual Block follows a narrow -> wide -> narrow approach, hence the inversion.....	18
Figure 7: MMER architecture	19
Figure 8: Cross Modal Encoder	19
Figure 9: Wav2Vec2 ver2.0 pipeline.	20
Figure 10 Fusion 1	24
Figure 11 Fusion 2 (removed last layer of each model)	25
Figure 12 Fusion 3 (Used Average Pool instead of FC in visual model)	26
Figure 13 Fusion 4 (Using ELU and Dropout)	27
Figure 14 Fusion 5 (Using ELU, Dropout, and Max Pooling)	28
Figure 15 Transformers structure.....	37
Figure 16 Training Phases	38
Figure 17 Enhanced Model Arch	39
Figure 18 Applications workflow	41
Figure 19 VOSK STT	41
Figure 20 LLama 3 developed by meta	42
Figure 21 100 Cases in Psychiatry, Second Edition	46

LIST OF TABLES

Table 1: Number of Annotated Images in Each Category.....	8
Table 2: table show the structure of our Zeiler&Fergus based model with 1×1 convolutions. The input and output sizes are described in rows \times cols \times #filters. The kernel is specified as rows \times cols, stride and the maxout pooling $p=2$ [7].	15
Table 4 Results Comparison between using emotions and only text.....	52

LIST OF ABBREVIATIONS

MMER	(Multimodal multi-task learning approach for Speech Emotion Recognition).
FER	Facial emotion recognition
IEMOCAP	Interactive Emotional Dyadic Motion Capture
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
MELD	Multimodal Emotion Lines Dataset
LLM	Large Language Model

CHAPTER ONE: INTRODUCTION

Motivation

The motivation behind this research is deeply embedded in the transformative landscape of artificial intelligence, which permeates diverse facets of human life, reshaping our interactions with technology. As AI systems progressively assimilate into daily routines, the recognition of the vital role that emotions play in human-computer interaction becomes increasingly apparent. Emotion detection stands as a critical linchpin in achieving a harmonious integration of AI into our lives, not only promising enhanced user experiences but also fostering the development of empathetic and contextually aware applications.

While existing emotion detection systems have made notable strides in both visual and audio-based recognition, inherent limitations persist. Unimodal approaches, whether reliant on facial expressions or audio cues, often struggle to encapsulate the intricate tapestry of human emotions, inherently multimodal in nature. It is imperative to note that while there are existing models attempting to combine both modalities, they frequently exhibit suboptimal accuracy, underscoring the need for a more effective and unified strategy.

Furthermore, with the continuous evolution of AI technologies, there is a growing imperative to bridge the gap between human emotions and machine understanding. Applications ranging from virtual assistants to gaming environments and mental health support systems could significantly benefit from emotion detection systems that not only exhibit heightened accuracy but also possess contextual awareness. The motivation driving this research is deeply rooted in the pursuit of a pragmatic solution that acknowledges and addresses the limitations of current models, paving the way for a new era of emotionally intelligent AI that transcends the shortcomings of existing unimodal approaches.

Problem Statement

The predominant challenge in prevailing emotion detection methodologies stems from their reliance on single-modal approaches, exclusively emphasizing either facial expressions or speech patterns. These unimodal systems often prove insufficient in capturing the intricate and multifaceted nature of human emotions, as they disregard the potential synergies between visual and auditory cues. Consequently, the present models may misinterpret or overlook subtle emotional nuances, diminishing accuracy and reliability in real-world applications.

Moreover, unimodal approaches encounter difficulties in scenarios where one modality provides ambiguous or incomplete information. For instance, in noisy environments or situations where facial expressions are obscured, relying solely on visual cues may lead to inaccuracies. Similarly, exclusive dependence on audio signals may face challenges in contexts where speech is absent or lacks discernible emotional cues.

Importantly, it's worth noting that while some models attempt to combine both modalities, their accuracy remains suboptimal. In real-world applications, these integrated models may still struggle to provide reliable emotion detection results. The identified limitations underscore the imperative need for a multimodal approach—one that integrates preexisting visual and audio models to harness the distinct strengths of both modalities. By merging the capabilities of facial expression recognition with the nuances embedded in speech patterns, a more comprehensive and accurate emotion detection system can be realized. This research endeavors to address this critical gap by exploring the integration of preexisting visual and audio models, aiming to contribute to the advancement of emotion detection technologies and their practical applications.

Outline

Literature Review:

- Comprehensive examination of existing emotion detection models, exploring unimodal and multimodal approaches.
- Identification of gaps in current research to establish the need for an integrated visual and audio model.

Neural Network Architecture of utilized Preexisting Models

- In-depth analysis of neural network architectures utilized in established visual and audio emotion detection models.
- Evaluation of the strengths and limitations of these architectures as a foundation for the proposed solution.

Proposed Solution Architecture

- Detailed exposition of the architecture proposed for integrating visual and audio models in emotion detection.
- Discussion of the methodology for combining facial landmarks from computer vision and audio features through signal processing.
- Integration of deep neural networks for multimodal fusion and enhanced accuracy.

System Requirements Specification (SRS)

- Documentation of functional and non-functional requirements essential for system development and deployment.
- Definition of the system's scope, specifying supported emotional expressions, environmental conditions, and user scenarios.

Results

- Evaluation of the integrated system's performance compared to individual visual and audio models.
- Discussion of findings, challenges encountered during implementation, and implications for future research and applications.

Conclusion

- Summation of key findings and contributions to the field.
- Reflection on the proposed solution's effectiveness in advancing emotion detection in AI systems.
- Recommendations for further research and applications, emphasizing the project's role in bridging gaps in current emotion detection methodologies.

CHAPTER TWO: LITERATURE REVIEW

Datasets and Benchmarks

Popular Datasets for Multimodal Emotion Detection

1- IEMOCAP (Interactive Emotional Dyadic Motion Capture):

Modalities: Speech, Facial Expressions.

Description:

- Consists of 151 recorded dialogues with 2 speakers per session, totalling 302 videos.
- Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral).
- Additional annotations include valence, arousal, and dominance.
- Recorded across 5 sessions with 5 pairs of speakers.

Structure: Likely organized into folders representing sessions, each containing subfolders for speaker pairs.

Within each subfolder, video files (dialogues) would be present.

Associated annotation files or metadata providing details on emotions, valence, arousal, and dominance.

2- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):

Modalities: Speech, Facial Expressions.

Description:

- Contains 7356 files with a total size of 24.8 GB.
- Includes 24 professional actors (12 female, 12 male) vocalizing statements in a neutral North American accent.
- Expressions cover emotions such as calm, happy, sad, angry, fearful, surprise, and disgust for speech and song.

Structure: Organized into folders based on actor information and modality (speech or song).

Subfolders for each emotion within the speech and song modalities.

Each subfolder likely contains audio files corresponding to actor and emotion.

3- MELD (Multimodal Emotion Lines Dataset):

Modalities: Speech, Facial Expressions.

Description:

- Multimodal dataset with 13,118 utterances from 1,433 dialogues between 804 speakers.
- Annotated with emotion labels and sentiment scores.

Structure: Folders representing dialogues, each containing subfolders for speakers.

Subfolders may further categorize data based on emotions or sentiment.

Within each subfolder, audio files, transcripts (language), and possibly visual cues.

4- **EmoReact:**

Modalities: Speech, Facial Expressions.

Description:

- Contains 1102 audio-visual clips annotated for 17 different emotional states, including six basic emotions, neutral, valence, and nine complex emotions.

Structure: Likely organized into folders based on emotional categories.

Each folder contains audio-visual clips corresponding to different emotional states.

Metadata or annotation files providing details on the labelled emotions.

5- **CMU-MOSEI (Multimodal Sentiment Intensity):**

Modalities: Speech, Facial Expressions.

Description:

- Dataset with 93 hours of video and audio data, including 23,453 utterances.
- Annotated with sentiment and emotion intensity scores.

Structure: Organized into folders or sessions based on video and audio data.

Likely subfolders for individual utterances or speakers.

Annotation files providing sentiment and emotion intensity scores.

Benchmarks and Challenges

1. **Emotion Recognition Challenges:**

- RAVDESS has been employed in various emotion recognition challenges and competitions, serving as a benchmark for evaluating the performance of models across different emotion categories.
- Challenges often entail tasks such as classifying emotions in both speech and song performances, highlighting the versatility of the dataset in assessing model robustness.

2. **Contributions to Research:**

- RAVDESS has played a pivotal role in advancing emotion detection research, particularly in the realm of audio-visual emotion analysis.
- Researchers have utilized the dataset to explore the effectiveness of models in capturing emotions through both auditory and visual cues, contributing to the development of more sophisticated and accurate emotion recognition systems.

3. **Cross-Modal Studies:**

- The synchronized audio and visual data in RAVDESS facilitate cross-modal studies, encouraging researchers to develop models capable of integrating information from both modalities.
- Cross-modal studies are instrumental in understanding how emotional information is conveyed and perceived through multiple channels, leading to more comprehensive multimodal models.

Impact on Model Development

- **Enhanced Model Robustness:**

The inclusion of both speech and singing performances, along with diverse emotional expressions, enhances the robustness of models trained on RAVDESS.

Models developed on this dataset are better equipped to handle a wide range of emotional scenarios.

- **Comparative Analysis:**

RAVDESS facilitates comparative analyses between unimodal (audio-only or visual-only) and multimodal models, allowing researchers to assess the benefits of combining information from multiple modalities.

- **Real-world Applicability:**

The dataset's focus on naturalistic emotional expressions contributes to the development of models with real-world applicability, particularly in scenarios where emotions are expressed through speech or singing.

- **Innovation in Model Architectures:**

Researchers working with RAVDESS are encouraged to explore innovative model architectures that can effectively fuse information from audio and visual modalities, pushing the boundaries of multimodal emotion detection.

Video Data Availability:

Although all three datasets we came across while working on this project are big in size (in Gigabytes) they still contain few data. For example, IEMOCAP contains only 5 conversations, RAVDESS contains 24 actors only **Figure 1** . The area of emotion detection using video feed faces a huge hurdle due to a scarcity of video-based datasets, making it difficult to train viable algorithms. Despite the growing relevance of understanding and interpreting human emotions, there is a noticeable lack of extensive video datasets designed specifically for emotion identification tasks.

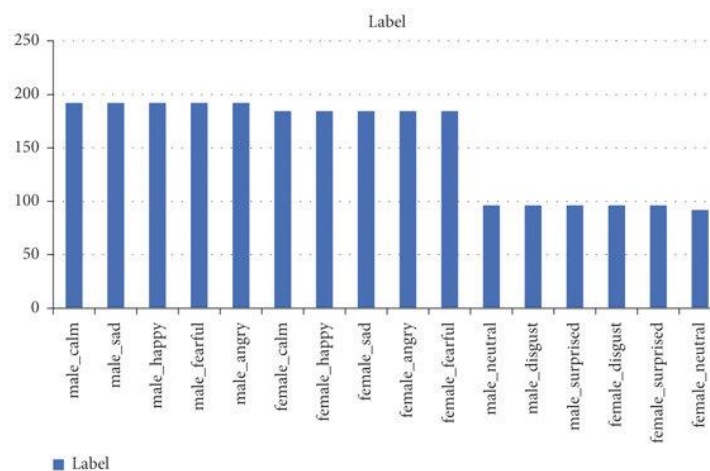


Figure 1: RAVDESS dataset distribution

Existing datasets that address this topic are frequently small, with only a few samples. The lack of varied and comprehensive video datasets impedes the creation and assessment of reliable emotion identification algorithms, stalling progress in this critical field of study. The lack of data not only limits the ability to create advanced algorithms, but it also impedes the investigation of subtle emotional responses across varied demographic groups and cultural situations. Addressing this gap in the availability of high-quality video datasets is critical for improving the capabilities of emotion identification algorithms and promoting their practical application in real-world settings.

To solve the difficulty of data scarcity for video-based emotion recognition (VER), we combined two powerful modalities: visual and auditory with some work arounds which are:

1. Using the capabilities of EfficientFace for visual emotion recognition (ER), an image-based approach, we were able to access substantially bigger picture datasets. EfficientFace was pretrained on AffectNet7. AffectNet contains more than 1M facial images collected from the Internet by querying three major search engines using 1250 emotion related keywords in six different languages [1]. To improve its temporal knowledge, we fine-tuned the model with the RAVDESS dataset, which, being a video database, naturally integrates temporal dynamics. This not only increased the amount of visual data available for training, but also allowed the model to detect subtle emotional expressions over time.

Expression	Number
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
Contempt	5,135
None	35,322
Uncertain	13,163
Non-Face	88,895

Table 1: Number of Annotated Images in Each Category

2. In addition, Connectionist Temporal Classification (CTC) loss function. The use of CTC loss helps to alleviate data scarcity problems by allowing the model to learn from data without needing explicit matching with labelled emotion segments. Our methodology seeks to overcome data restrictions by utilising a multimodal approach and smart modifications [2]. In supervised learning contrastive loss cannot handle many samples belonging to the same class owing to labelling. When the number of positives is increased, additional functionalities become available equations (1),(2).

$$L_{out}^{sup} = \sum_{i \in I} L_{out,i}^{sup} = \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\frac{z_i \cdot z_p}{\tau}}}{\sum_{a \in A(i)} e^{\frac{z_i \cdot z_a}{\tau}}} \quad (1)$$

$$L_{in}^{sup} = \sum_{i \in I} L_{in,i}^{sup} = \sum_{i \in I} -\log \left\{ \sum_{i \in I} -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{\frac{z_i \cdot z_p}{\tau}}}{\sum_{a \in A(i)} e^{\frac{z_i \cdot z_a}{\tau}}} \right\} \quad (2)$$

Fusion techniques review

In their papers [3] and [4], the writers described briefly the different methods of modalities fusion. The problem of multiple modality fusion could be seen as four smaller sub problems as follows: single-modal feature extraction, feature fusion, model classification, and result output. Depending on when fusion takes place between the stated sub stages we can classify fusion into three categories namely early fusion, late fusion, and intermediate fusion. Early fusion takes place when different modalities are fused by a series of dot products or concatenation of features.

The other extreme is to have multiple models one for each modality and fuse between the classification outputs of all the modalities models, which is known as late fusion. Intermediate fusion occurs when each modality runs a feature extraction network then the fusion happens between the extracted features before classification.

It has been widely believed that intermediate fusion is the best approach to address the problem of multiple modalities especially after the effectiveness of deep learning architectures such as CNNs and fusion models. Based on our reading in this topic we decided to apply both late, and intermediate fusion, since it was reported that late fusion in most of the cases yielded good results.

Wav2vec2.0 (acoustic encoder)

Wav2vec2 has gained its unprecedented leadership in the field of processing raw audio data owing to self-supervised learning, and transformer technology. It is designed to address fundamental audio processing challenges such as feature extraction, contextual understanding, and representation learning. Aided with self-supervised learning wav2vec2 was able to be trained on much larger data corpus since it is able to be trained directly on raw audio waveforms without the need of explicit annotations.

1. Feature extraction

Wav2vec2 is feed with raw input audio signal (sequence of amplitudes over time) that passes through multi-layer convolutional neural network that extracts low-level features that capture the primitive acoustic speech properties.

2. Contrastive learning

The goal of contrastive learning isn't classification, but, learning better representations of the acoustic signal in the latent space. Here is how it works: the model creates multiple representations in the latent space per audio segment. These different representations are either similar or dissimilar to the given segment. The goal of contrastive learning is to pull similar latent representations near together, and push

dissimilar pairs away from each other. This way the latent representation is robust meaningful representation of the audio segments.

3. Encoder

This part of the model is composed of the encoder part of a transformer which is feed with the low-level extracted features from the CNN and then uses these features to draw long-range dependencies between different segments within the same audio file.

4. Output layer

Based on the task this output layer will vary accordingly. If the task is automatic speech recognition (ASR) then this layer will be a decoder that outputs the highest probability sequence of textual words representing this acoustic file. In our case we will either use this layer as an emotion classifier where it is trained on one of the famous emotions datasets and then utilize this output emotion in the late fusion layer, or use the latent representation of the transformer encoder for the intermediate fusion layer.

RoBERTa_{BASE} (contextual encoder)

BERT stands for Bidirectional Encoder Representation from Transformers which is a natural language proceeding model. It is used to grasp the meaning of each word in a sentence. BERT utilizes bidirectional context understanding, and parallelized processing of a sequence thanks to the transformer architecture.

BERT analyses a phrase as a whole taking into account preceding and following words, in contrast to other unidirectional sequence models such as RNNs. This gives BERT the power to understand the meaning of the word in its context. Take for example the sentences “Teddy bears are on sale.”, and “Teddy Roosevelt was a great President.” the word Teddy changes its meaning depending on the context of the words following it. The bidirectional nature of BERT’s architecture is capable of capturing such difference in the representation of the word teddy depending on the sequence that follows.

BERT's parallelization is one of its main benefits over RNNs, its predecessor. This implies that although RNNs must process a sentence one word at a time, BERT may process a sentence as a whole at once. This results in great efficiency in its training time. Furthermore, because BERT is parallelized, it can capture long-range relationships between words more readily than RNNs, which are only able to consider a small amount of context at once.

MMER (Multimodal Multi-task Learning for speech emotion recognition)

In their paper [5] Sreyan Ghosh, et al. tried to extract detailed multimodal emotional data from text and audio modalities by utilizing voice and the text transcripts that go along with it. In order to discover the natural harmonic alignment between voice and text as well as the syntactic and semantic information concealed in the text, it first resolves the problem of automatic speech recognition ASR by minimizing the CTC loss.

They state that most of the fusion model between the speech and text modalities were late fusion and they yielded good results, yet this late fusion technique failed to capture cross-modal interactions. They also mention that although early fusion solves this problem that was posed by the late fusion method, they experience suppression of modality-specific interactions. They suggest that a special type of intermediate fusion namely cross-modal attention mechanism where features from one modality could attend to features from the other modality.

The suggested approach to handle this task was as follows:

1. Features encoding: they used wav2vec2 trained on 960 hours of Librispeech as an acoustic encoder, and Roberta_{Base} as the text encoder.

2. Multimodal interaction module:

This module contains three cross-modal encoders (CME) to capture the interactions between text, and speech representations. The **speech-aware word representation** CME takes wav2vec2 embeddings as queries and Roberta as keys and values. Similarly the **word-aware speech representation** CME takes Roberta embeddings as queries, and wav2vec2 embeddings as keys and values. **Acoustic gate** CME dynamically control the contribution of each speech frame embedding to get rid of the effect of random noise encoding.

Finally, after concatenation, sigmoid gate, and element-wise multiplication of the outputs of these three CME, this embedding is used to draw an emotion for the given text, and audio recording.

Results show that this approach is the state-of-the-art for performing emotion analysis on the IEMOCAP dataset with accuracy of 81.2%

Efficient Face

EfficientFace is a facial expression recognition (FER) model that is meant to be parameter-efficient while maintaining accuracy and resilience in recognising facial emotions under a variety of situations, such as those found in real-world settings. To improve the resilience of the lightweight network, the authors include a local-feature extractor and a channel-spatial modulator. The inclusion of depth wise convolution in these components allows the network to record both local and global salient face information, increasing sensitivity to nuanced facial emotions.

In addition to architectural improvements, the study proposes a new training technique termed the Label Distribution Learning (LDL) method. This approach recognises the complexity of emotions, noting that they frequently present as combinations, mixes, or compounds of fundamental emotions. The LDL technique allows the model to learn not only individual emotion classes, but also their connections and combinations. This allows for a more nuanced and thorough grasp of the complicated nature of complex emotional responses. The proposed EfficientFace is rigorously tested on realistic occlusion and posture variation datasets to demonstrate its durability under demanding settings. The experimental findings reveal state-of-the-art performance on benchmark datasets such as RAF-DB, CAER-S, and AffectNet-7, with excellent accuracies of 88.36%, 85.87%, and 63.70%. Furthermore, the model gets a competitive performance on the AffectNet-8 dataset, with an accuracy of 59.89%.

To reduce computational overheads, the authors use the cutting-edge lightweight network ShuffleNet-V2 as a backbone network, which includes Conv1, Stage 2, Stage 3, Stage 4, and Conv5. The model incorporates a local-feature extractor and a channel-spatial modulator to address real-world difficulties like occlusion and posture variation.

The **local-feature extractor** is intended to effectively learn local face characteristics without relying on facial landmarks, hence solving the inefficiency of current approaches. Following Conv1, the global feature maps are divided into four patches, with each patch undergoing depthwise convolution to produce local face features. These local characteristics are subsequently concatenated, increasing the network's sensitivity to facial expressions associated with action units.

To reduce duplicate information in global face characteristics, a **channel-spatial modulator** is added after Stage 2. This modulator, modelled after the BAM architecture, emphasises critical global aspects. Channel and spatial heatmaps are generated, normalised, and merged to modify global characteristics. The final global-local features are created by mixing extracted local features with modulated global features, allowing the network to learn both global-salient and local face characteristics.

CHAPTER THREE: NN ARCHITECTURE REVIEW

Visual Model

Before we start talking about the architecture of the visual model let's first break it down into three main components: Facenet which is used for face detection, Efficientface which is the base of the visual emotion recognition model using images, and the final model which adds the capability to use videos or sequences of images.

i. FaceNet:

Although we use it only for face detection FaceNet is a face recognition system that allows for fast face verification, identification, and grouping at scale (clustering). The model learns to map face pictures into a compact Euclidean space, where distances directly indicate the similarity of faces. It does this by training to optimise a 128-dimensional embedding using a new triplet-based loss [6] function. The network is trained using triplets of generally aligned matching and non-matching face patches obtained by online triplet mining. FaceNet's main benefit is its representational efficiency, which allows it to achieve cutting-edge face recognition performance while consuming just 128 bytes per face. Unlike prior efforts, FaceNet simply trains its output to be a compact embedding, eliminating the need for an intermediate bottleneck layer and resulting in a far more efficient representation per face.

Triplet-based loss allows the network to focus on learning a meaningful and discriminative feature space by explicitly taking into account the links between triplets of examples. This is especially beneficial in applications that need fine-grained similarities or dissimilarities, such as facial recognition or picture retrieval. It consists of the Anchor A which is the reference instance for which we want to learn the representation. Positive P which is a similar instance to the anchor (same class or category). Negative N which is a dissimilar instance to the anchor (different class or category). Triplet-based loss can be calculated using the following formula in:

$$L(A, P, N) = \max(0, d(A, P) - d(A, N) + \alpha) \quad (3)$$

- $d(A, P)$ is the distance between the anchor and positive examples.
- $d(A, N)$ is the distance between the anchor and negative examples.
- α is a margin or threshold that ensures the positive examples are sufficiently closer to the anchor than the negative examples.

FaceNet uses a deep convolutional neural network to generate a compact Euclidean space for face photos, allowing for fast face verification, identification, and grouping at scale. The model directly optimises the embedding with a triplet-based loss function, which consists of an anchor face, a matching positive face, and a non-matching negative face. The training procedure includes online triplet mining, which selects hard triplets to ensure efficient convergence. FaceNet uses harmonic embeddings to ensure compatibility across different network outputs and applies a harmonic triplet loss. The models are trained using Stochastic Gradient Descent and AdaGrad to investigate architectures with different parameters and FLOPS. The approach achieves cutting-edge performance on benchmark datasets, displaying resilience to posture and lighting fluctuations [7].

The FaceNet model uses a deep convolutional neural network (CNN) with two fundamental architectures: the Zeiler&Fergus model and the Inception model. The Zeiler&Fergus-inspired network has 22 layers and uses 1x1 convolutions between typical convolutional layers, yielding 140 million parameters and needing 1.6 billion floating-point operations per picture [8]. The Inception-based models, NNS1 and NNS2, have much fewer parameters (6.6M-7.5M) and fewer floating-point operations (220M-1.6B) such as in **Figure 2**. The network is trained with Stochastic Gradient Descent (SGD) and AdaGrad, and the triplet loss function is used to optimise a 128-dimensional embedding. The live triplet selection technique provides efficient triplet creation by focusing on difficult situations, which contributes to speedy model convergence.



Figure 2: FaceNet architecture. You can see the full details of the layer's sizes, The flow of the data and PARM, FLOPS in Table2

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

Table 2: table show the structure of our Zeiler&Fergus based model with 1×1 convolutions. The input and output sizes are described in rows × cols × #filters. The kernel is specified as rows × cols, stride and the maxout pooling p=2 [7].

ii. EfficientFace:

For emotion recognition we used EfficientFace, a lightweight and resilient facial expression recognition network designed for practical FER in the wild [9]. In the context of feature extraction, a local-feature extractor and a channel-spatial modulator are presented. The learnt face characteristics remain stable under occlusion and posture change.

To address the noise problem on facial expression datasets and the fact that real-world facial expressions are a distribution rather than a single emotion, a new LDL approach is presented based on LDG design. To reduce computational overheads in the FER model, we use the lightweight network ShuffleNet-V2, which includes Conv1, Stage2, Stage3, Stage4, and Conv5. To address occlusion and position variation in real-world scenarios, we created a local-feature extractor and channel-spatial modulator.

Furthermore, a new label distribution learning approach is given that is congruent with the psychologist's theory (Plutchik 1980) [10]. The next subsections introduce the proposed local-feature extractor and channel-spatial modulator, followed by information on the label distribution generator and loss.

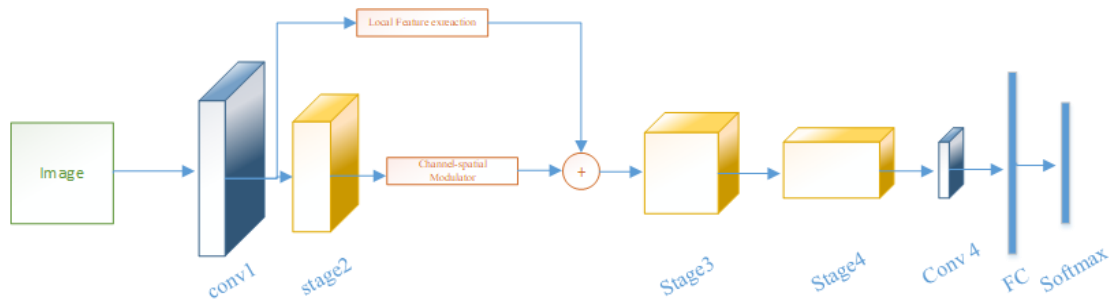


Figure 3: Effiecientface abstract architecture showing its pipeline.

A **local feature extractor** is a component that detects and extracts unique characteristics from various parts of an image. These characteristics are usually tiny, localised patterns or structures that are distinct or easily identifiable, such as corners, edges, or blobs [11]. The purpose is to collect information about significant spots or regions in a picture that may be used to characterise its content in a way that is unaffected by specific transformations such as translation, rotation, and scaling. The architecture of our local feature extractor is in **Figure4**.

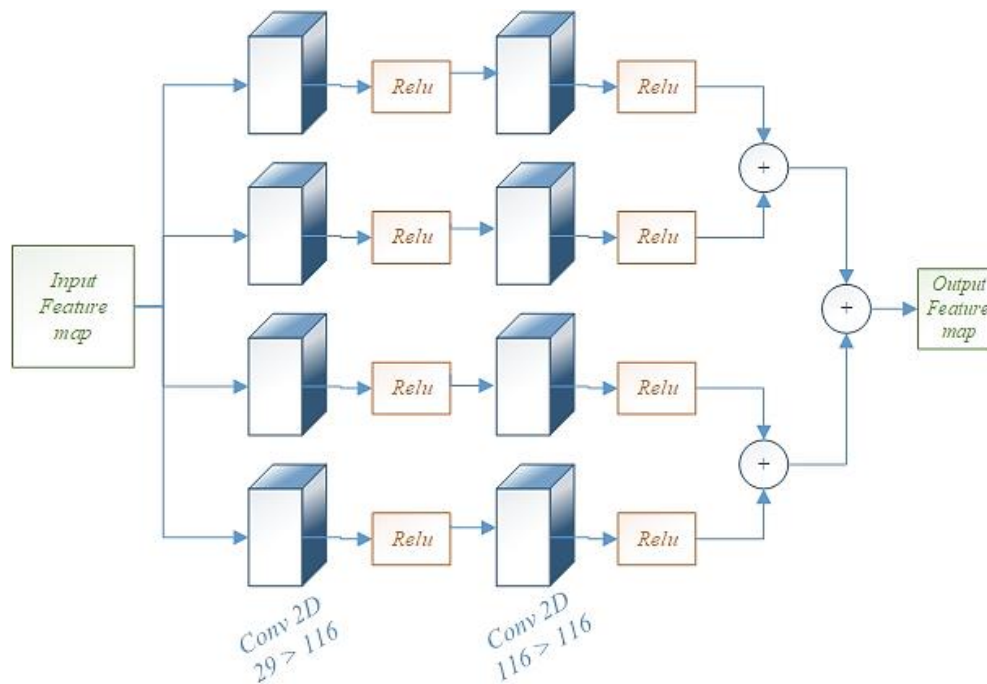


Figure 4: Local-feature extractor used in EffecientFace components.

Previous research suggests that learning local face traits is beneficial for FER in the wild. However, such approaches acquire local characteristics using face landmarks, which is inefficient. The model presents an efficient local-feature extractor that learns local area features at the feature level. These features are then supplemented into global features as residuals. The **channel-spatial modulator** is introduced to highlight the crucial global features after the Stage2. Allow for computational overheads, the channel-spatial modulator is designed based on BAM. The architecture of our local feature extractor is in **Figure5**.

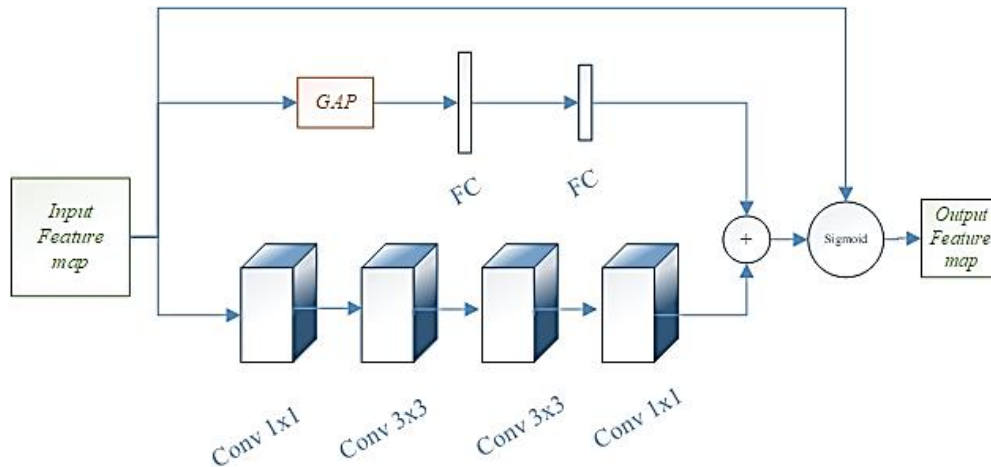


Figure 5: Channel-spatial modulator components

The model consists of several stages, each composed of **inverted residual** blocks. The stages are named 'stage2', 'stage3', and 'stage4', and their architecture is determined by the `stages_repeats` and `stages_out_channels` parameters provided during initialization. The InvertedResidual is a building block commonly used in mobile and efficient neural network architectures, particularly in models like MobileNetV2 and EfficientNet. This block is called "inverted" because it inverts the traditional residual block structure, where the shortcut connection is applied before the non-linear activation [12].

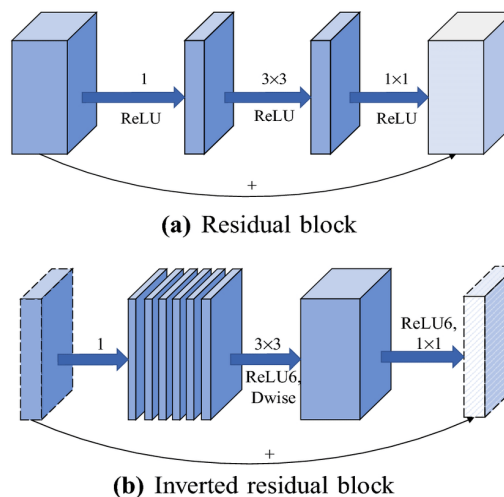


Figure 6: (a) Represents the normal residual block which has a wide \rightarrow narrow \rightarrow wide structure with the number of channels. (b) Inverted Residual Block follows a narrow \rightarrow wide \rightarrow narrow approach, hence the inversion..

Audio Model

For our audio model we used MMER (Multimodal Multi-task learning approach for Speech Emotion Recognition). MMER is based on early fusion and cross-modal self-attention between text and acoustic modalities and solves three auxiliary tasks for learning emotion recognition from spoken utterances. In practice, MMER outperforms all our baselines and achieves state-of-the-art performance on the IEMOCAP benchmark. Additionally, we conduct extensive ablation studies and results analysis to prove the effectiveness of our proposed approach [5].

For the pre-trained acoustic encoding it uses Wav2Vec2 which is a model made by Facebook. For the pre-trained text encoding it uses RoBERTa_{BASE} which uses BERT's model. You can see the full architecture in **Figure7**.

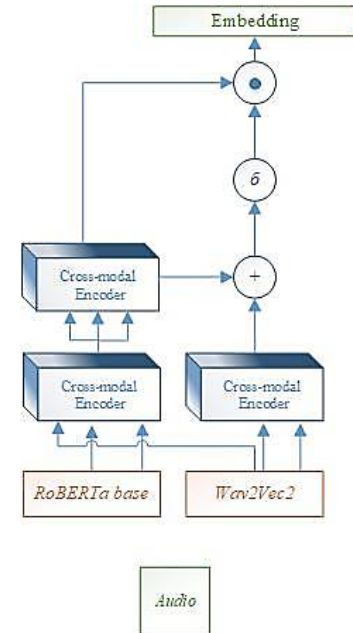


Figure 7: MMER architecture

The cross-modal encoder in Figure7 seeks to create a common representation space in which data from many modalities may be incorporated and compared. This shared representation captures important correlations and similarities between instances of data from many modalities. The fundamental objective is to develop a combined representation that allows for tasks requiring many types of data. You can see the details in **Figure8**.

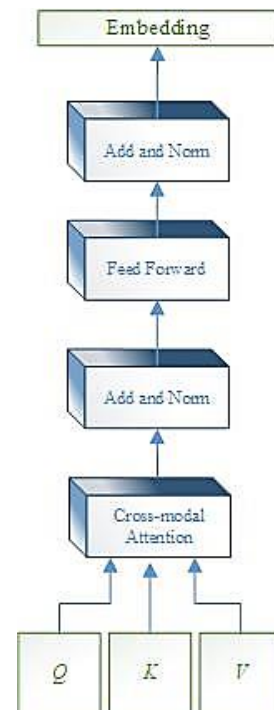


Figure 8: Cross Modal Encoder

i. Wav2Vec

Wav2Vec2 is a speech model that accepts a float array corresponding to the raw waveform of the speech signal. wav2vec 2.0 masks the speech input in the latent space and solves a contrastive task specified by a quantization of the latent representations that are simultaneously trained. Experiments employing all labelled Librispeech data provide 1.8/3.3 WER on clean/other test sets. When the quantity of labelled data is reduced to one hour, wav2vec 2.0 exceeds the prior state of the art on the 100-hour subset while using 100 times less labelled data.

Using only 10 minutes of labelled data and pre-training on 53k hours of unlabelled data yields 4.8/8.2 WER. This illustrates the capability of voice recognition with little labelled data. [13]. The model uses the pre-trained checkpoint released by Facebook, pre-trained on 960 hours of Librispeech, and use the wav2vec-2.0-base architecture for all our experiments [14]. You can see the pipeline of the model in **Figure9**.

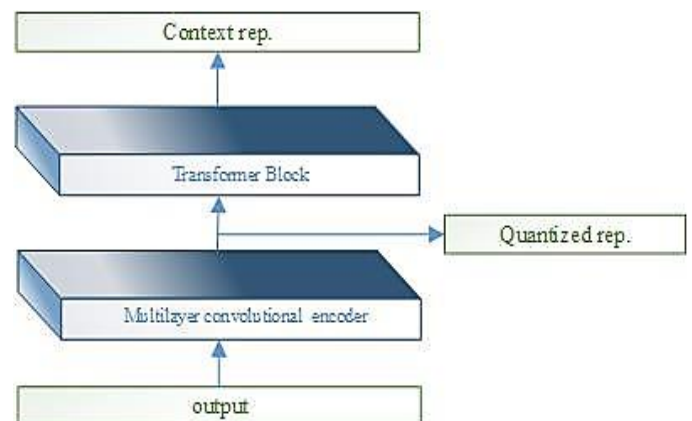


Figure 9: Wav2Vec2 ver2.0 pipeline.

ii. RoBERTa_{BASE}

RoBERTa is a self-supervised transformers model trained on a vast corpus of English data. This implies that it was pretrained solely on raw texts, with no human labelling (which allows it to use a large amount of publicly available data), and that an artificial mechanism generated inputs and labels from those texts. RoBERTa iterates on BERT's pretraining procedure, including training the model longer, with bigger batches over more data; removing the next sentence prediction objective; training on longer sequences; and dynamically changing the masking pattern applied to the training data [15].

CHAPTER FOUR: SOLUTION ARCHITECTURE

The solution architecture for the Emotion Detection Deep Learning Model involves a comprehensive integration of preexisting visual and audio models to enhance accuracy. Fusion technique was used to enhance the performance of the overall model by fusing the two mentioned before models. We used the **Late** fusion as an approach to model fusion, where the predictions of individual models are combined at a later stage, typically after their outputs have been generated.

However, not only we used the outputs of each model directly to fuse the 2 models, but also, we experimented with different ways to apply the late fusion model that are discussed later in this chapter. But first let us talk about our project timeline.

Project Timeline

Visual Model (FaceNet and EfficientFace Integration)

In the early stages of the project, the integration process begins with the exploration of the FaceNet architecture. FaceNet, known for its advancements in face recognition, provides a solid foundation for extracting discriminative facial features from raw video data. The model's pre-trained weights and capabilities in generating face embeddings are crucial components in establishing a comprehensive visual representation for subsequent emotion recognition tasks.

Parallely, the project delves into the integration of EfficientFace, a specialized architecture designed for efficient facial expression recognition directly from raw video frames. The decision to incorporate EfficientFace lies in its capacity to efficiently capture facial expressions, aligning with the project's emphasis on end-to-end learning without the need for separately learned features. This integration involves incorporating the EfficientFace feature extraction component into the visual branch of the overall audiovisual emotion recognition model.

As the integration progresses, the two models are harmoniously fused within the visual branch. The integration point involves strategically combining the extracted facial features from both FaceNet and EfficientFace, creating a rich and comprehensive representation of facial expressions. This fusion is crucial for capturing nuanced facial information that contributes to a more robust understanding of human emotions.

Throughout the integration process, considerations are made to optimize computational efficiency, ensuring that the final visual model performs effectively in real-world scenarios. The collaborative efforts of FaceNet and EfficientFace within the visual branch contribute to the model's ability to handle diverse facial expressions in unconstrained settings, addressing challenges where one modality may be absent or noisy during inference.

Audio Model (MMER with Wav2Vec2 and RoBERTaBASE)

Now the part for the Audio Model within the context of Multi-Modal Emotion Recognition (MMER) involves the integration of advanced models, specifically Wav2Vec2 and RoBERTaBASE, to enhance the system's ability to analyse and recognize emotions from audio data. In this integrated approach, the Wav2Vec2 model is employed for audio feature extraction, capturing intricate patterns in speech signals, and converting them into meaningful representations. This process allows the model to directly learn from raw audio data, providing a foundation for end-to-end training without the need for pre-extracted features.

Complementing the audio feature extraction, the RoBERTaBASE model plays a crucial role in understanding the semantic context of the transcribed audio content. By leveraging RoBERTa's natural language processing capabilities, the model can capture nuanced linguistic cues, aiding in the interpretation of emotional states expressed through speech. The integration of RoBERTaBASE further enriches the overall multi-modal architecture by combining auditory information with contextual understanding derived from textual representations.

Throughout the project timeline, the focus lies on achieving seamless integration between the Wav2Vec2 and RoBERTaBASE models within the MMER framework. This integration not only enables the joint learning of audio and textual features but also ensures that the system remains robust in real-world scenarios where one modality might be incomplete or noisy. The proposed modality dropout mechanism, inspired by the challenges encountered in multi-modal emotion recognition, aims to enhance the model's resilience by simulating scenarios with missing or unreliable data during training. This holistic approach considers the interplay between audio and textual modalities, emphasizing not only performance gains in ideal settings but also the system's adaptability in less-than-ideal conditions.

Fusion Implementation

In the pursuit of refining the Emotion Detection Deep Learning Model, the implementation of fusion strategies emerges as a critical component. Leveraging a late fusion approach, our objective is to seamlessly integrate information from both video and audio models. Late fusion, chosen for its efficacy and reduced training requirements, plays a pivotal role in aggregating diverse outputs into unified emotion predictions. Intermediate fusion was used for to achieve better performance using feature maps for both models.

This section delineates four distinctive fusion strategies each tailored to capture temporal patterns and optimize feature combination for enhanced emotion prediction. From straightforward concatenation to intricate architectures involving average pooling, dropout, and non-linear activations, these fusion methodologies are designed to unravel the complexity of human emotion.

Method 1

This method involves the consolidation of 16 outputs (8 output for each emotion from each model), corresponding to different emotions. These diverse outputs are aggregated to produce 8 final predictions. This straightforward fusion method sets the foundation for the subsequent strategies, providing a baseline for comparison with more intricate architectures.

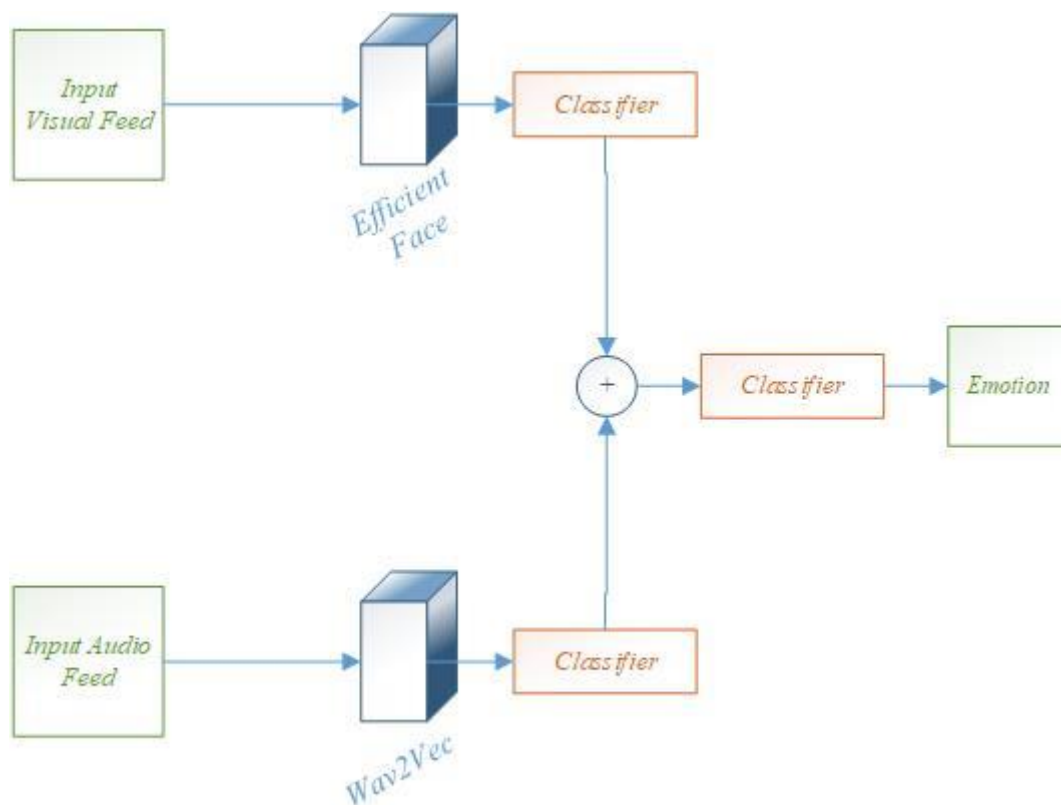


Figure 10 Fusion 1

CHANGES:

The individual models were not changed.

FUSION:

- There are 16 outputs in total, 8 from each model corresponding to different emotions.
- The fusion involves aggregating these 16 outputs into 8 final predictions using 1 additional layer.
- Demonstrates high validation accuracy (95.83%) and solid test accuracy (93.75%).

Method 2

For the second method used, a more complex architecture unfolds. The fusion involves concatenating output representations from both models, also the last layer from the visual model is neglected and the output is received from the previous layer and the same is done for the audio model. Furthermore, a linear layer is employed to reduce the concatenated dimension to 8 outputs for each emotion.

This method allows us to have a broader view of the features capture by each model instead of just using the final output of each model such as the previous model.

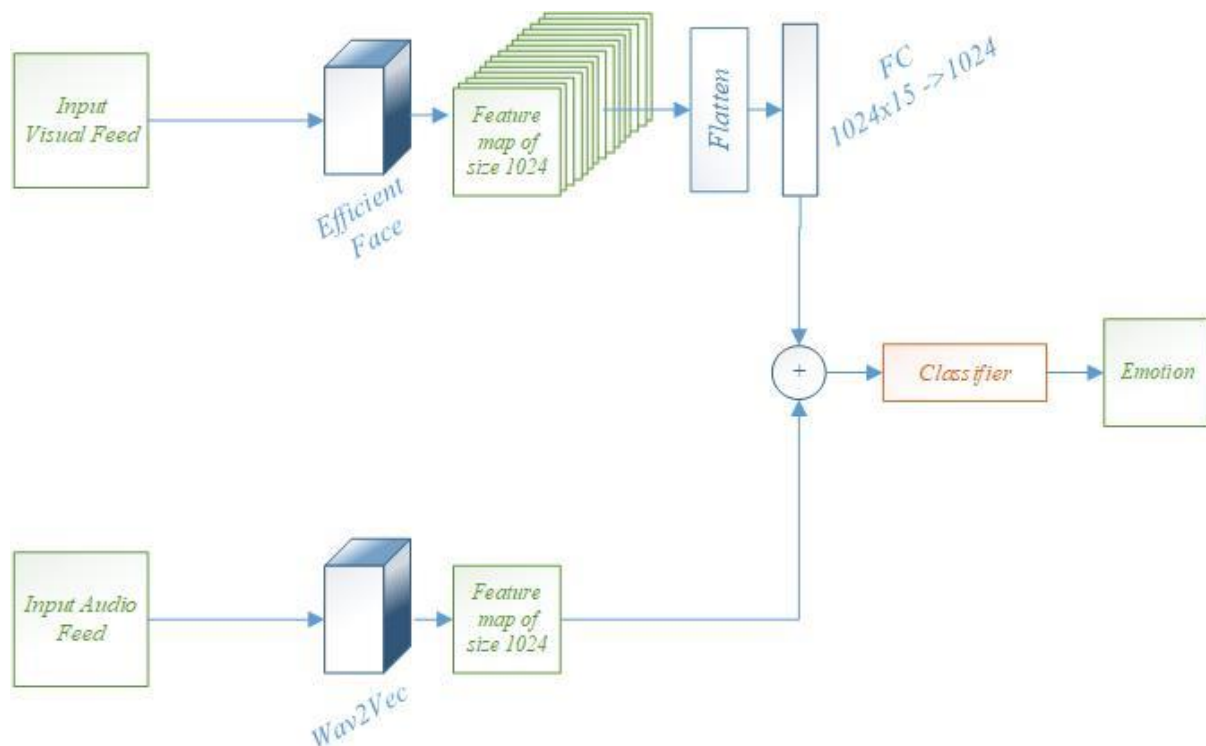


Figure 11 Fusion 2 (removed last layer of each model)

CHANGES:

- Remove last layer from both the audio and video models.
- Added a fully connected layer to the video model to map the (1024×15) to 1024 feature.
- The audio model now outputs 1024 since the last layer is removed

FUSION:

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024×2 to 8.
- Although exhibiting lower validation (81.5%) and test (74.38%) accuracies, lays the groundwork for subsequent, more intricate fusion methods.

Method 3

Here, a similar architecture to the previous method is employed, with the concatenation of output representations from visual and audio models. However, instead of a linear layer, average pooling is applied along the temporal dimension, followed by a subsequent linear layer, capturing temporal patterns in a different manner.

This method helped in reducing the training time, since average pooling helped to reduce our number of captured features and this layer have no learnable parameters so training is not heavily affected by the introduction of this layer.

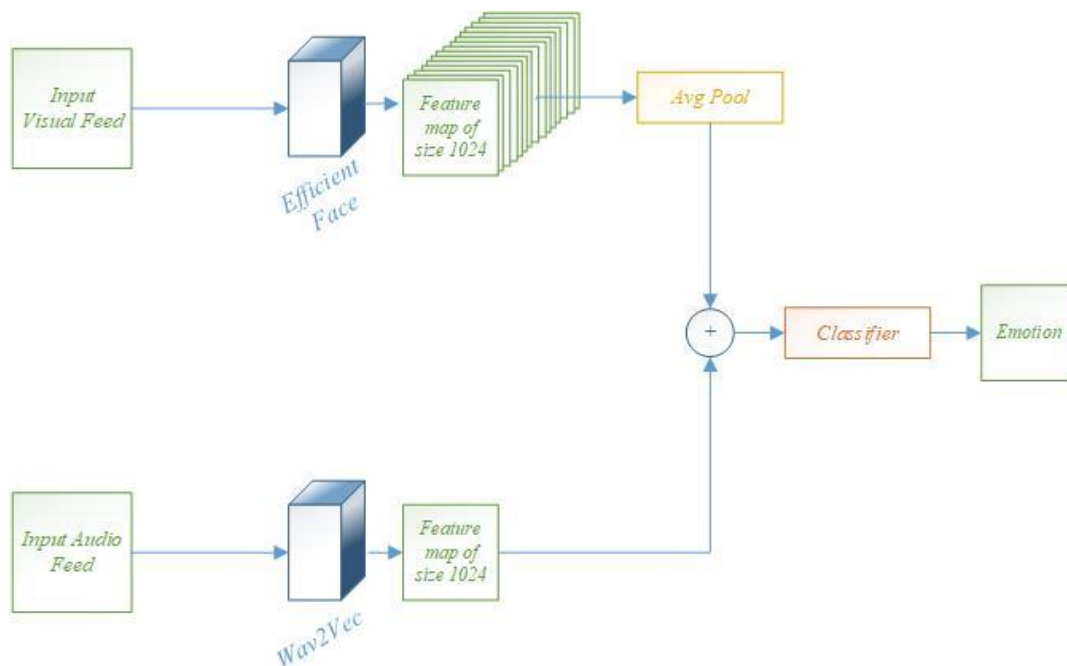


Figure 12 Fusion 3 (Used Average Pool instead of FC in visual model)

CHANGES:

- Remove last layer from both the audio and video models.
- Added an average pooling layer to the video model to map the (1024*15) to 1024 feature.
- The audio model outputs 1024 since the last layer is removed

FUSION:

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024*2 to 8.
- Achieves impressive validation (97.08%) and test (96.04%) accuracies, showcasing the efficacy of average pooling in reducing training time without compromising performance.

Method 4

In this approach, we enhance the fusion strategy by introducing more complexity. We add extra layers, including dropout for regularization and an ELU activation function. These modifications aim to better capture the emotional cues by emphasizing non-linear activations and implementing regularization techniques.

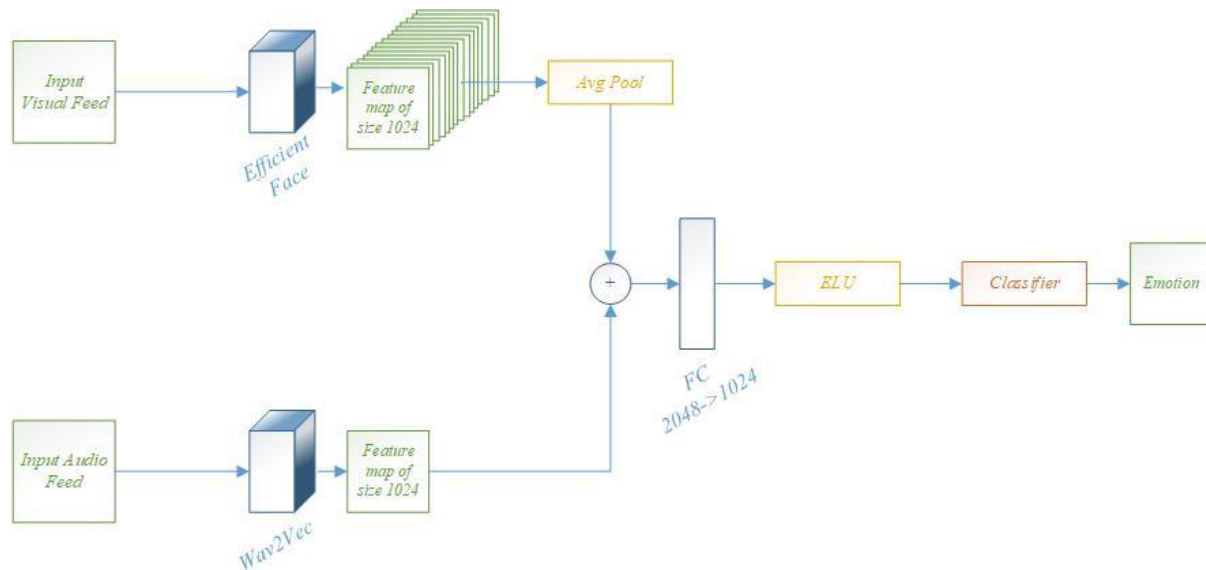


Figure 13 Fusion 4 (Using ELU and Dropout)

CHANGES:

- Remove last layer from both the audio and video models.
- Added an average pooling layer to the video model to map the (1024*15) to 1024 feature.
- The audio model outputs 1024 since the last layer is removed

FUSION:

- The fusion involves aggregating these 2048 (1024 output from each model) outputs
- A linear layer is used to reduce the concatenated dimension from 1024*2 to 1024.
- A dropout layer is introduced for regularization.
- An ELU activation function is applied.
- A linear layer reduces the dimension from 1024 to 8 for the final predictions.
- Introduces additional complexities, leading to excellent validation (97.58%) and test (97.08%) accuracies, emphasizing the significance of non-linear activations and regularization techniques.

Method 5

Similar to the previous method, adds complexity by incorporating additional layers, including dropout for regularization and an ELU activation function. The key distinction is the use of max pooling instead of average pooling, offering an alternative approach to capturing temporal patterns in the fused feature space.

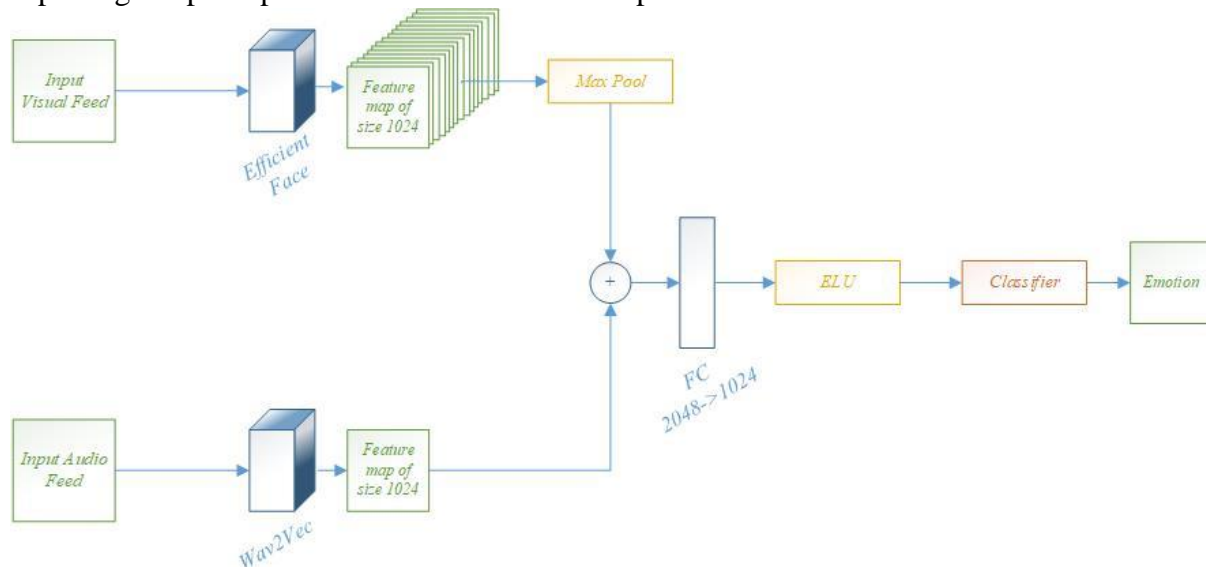


Figure 14 Fusion 5 (Using ELU, Dropout, and Max Pooling)

CHANGES:

- Remove the last layer from both the audio and video models.
- Add an average pooling layer to the video model to map the (1024*15) to 1024 features.
- The audio model outputs 1024 since the last layer is removed.

FUSION:

- Aggregate the 2048 outputs (1024 from each model).
- Apply a linear layer to reduce the concatenated dimension from 2048 to 1024.
- Introduce a dropout layer for regularization.
- Apply an ELU activation function.
- Utilize max pooling instead of average pooling to capture diverse temporal patterns.
- Conclude with a linear layer to reduce the dimension from 1024 to 8 for the final predictions.
- With max pooling, maintains high validation (97.92%) and test (97.08%) accuracies, illustrating the effectiveness of incorporating max pooling for capturing diverse temporal patterns.

CHAPTER FIVE: SOFTWARE REQUIREMENTS SPECIFICATIONS (SRS)

Scope

The proposed project attempts to create an Emotion Recognition Deep Learning Model that uses visual and auditory inputs to increase accuracy. The system consists of two separate models: one for processing facial image sequences using an efficient face model, and another for analysing audio data using the Wave2Vec2 model. These distinct models will be trained using the RAVDESS dataset (Ryerson Audio-Visual Database of Emotional Speech and Song).

The system will be designed to operate in an environment that supports the necessary deep learning frameworks, libraries, and dependencies for both image and audio processing.

It is assumed that the RAVDESS dataset adequately represents a diverse range of emotional expressions and can be used effectively for training both the visual and auditory models.

Dependencies include the availability of libraries and frameworks compatible with the chosen models and the accessibility of required computational resources.

Resource Constraints: The project will be constrained by available computational resources, including GPU capabilities, and will need to optimize model architecture accordingly.

Time Constraints: The project is bound by a timeline corresponding to the academic schedule for the graduation project.

The main aspects of the Emotion Recognition Deep Learning Model are:

- **Efficient Face Model:** An efficient face model is used to analyse facial image sequences in order to extract emotional information.
- **Wave2Vec2 Audio Model:** Using the Wave2Vec2 model to process audio data and extract emotional clues from speech and music.
- **Training on RAVDESS:** The efficient face model and the audio model will be trained using the RAVDESS dataset, which contains a wide range of emotional expressions.
- **Fusion of models:** The combination of visual and aural models to generate a final Emotion Detection Model, which improves accuracy by incorporating input from both modalities.

Functional Requirements

Data Input

- Image Input: The system should accept sequences of facial images as input for emotion prediction.
- Audio Input: The system should accept audio data as input for emotion prediction, where the audio data corresponds to speech or song.

Model Training

- Efficient Face Model: The system should train the efficient face model on the RAVDESS dataset, ensuring the extraction of facial features relevant to emotion recognition.
- Wave2Vec2 Audio Model: The system should train the Wave2Vec2 model on the RAVDESS dataset, capturing emotional cues from the audio data, including speech and song.

Fusion of Models

- Integration: The system should integrate the outputs of the trained efficient face model and the Wave2Vec2 audio model to create a fused Emotion Detection Model.
- Fusion Technique: The system should employ a suitable fusion technique, such as concatenation or ensemble methods, to combine the predictions from the visual and auditory models.

Prediction

- Emotion Prediction: The system should provide the capability to predict emotions based on new input data, using the final Emotion Detection Model.
- Output Format: The predictions should be presented in a structured format, indicating the likelihood or probability of each emotion category.

User Interfaces

- Training Configuration Interface: The system should provide a user interface for configuring training parameters, including options for adjusting hyperparameters, selecting training epochs, and specifying batch sizes.
- Monitoring Interface: The system should offer a monitoring interface for users to track the progress of the training process, visualize training metrics, and identify potential issues.
- Interaction Interface: The system should include an interface for users to interact with the final Emotion Detection Model, providing input for prediction and receiving the model's output.

Hardware Interfaces

- **Compatibility:** The system should be compatible with standard hardware configurations, and the efficient utilization of GPU resources should be considered for model training.

Software Interfaces

- **Deep Learning Frameworks:** The system should interface with deep learning frameworks, including but not limited to TensorFlow or PyTorch, for model development, training, and deployment.
- **Data Storage:** The system should interact with data storage systems for managing and accessing the RAVDESS dataset during training.

Non-Functional Requirements

Performance Requirements

- **Training Time:** The training process for both the efficient face model and the Wave2Vec2 audio model should be completed within a reasonable time frame, considering computational resources.
- **Inference Time:** The Emotion Detection Model should provide predictions for new input data in real-time or within acceptable response time limits.

Reliability

- **Robustness:** The Emotion Detection Model should handle variations in input data gracefully, providing consistent and reliable predictions across different scenarios.

Scalability

- **Model Enhancement:** The system should be designed to accommodate future enhancements, including the integration of additional modalities or improvements to existing models.
- **Dataset Variations:** The Emotion Detection Model should demonstrate scalability in handling variations in input data, such as diverse facial expressions and emotional nuances in speech and song.

Use Cases

1. Train Efficient Face Model

Primary Actor: Machine Learning Engineer

Description: The machine learning engineer starts the training procedure for the efficient face model. The use case consists of loading the RAVDESS dataset, establishing training parameters (such as batch size and learning rate), and running the training script. The system should provide feedback on training progress, show important metrics, and save the learned model for future use.

2. Train Wave2Vec2 Audio Model

Primary Actor: Machine Learning Engineer

Description: The machine learning engineer begins training the Wave2Vec2 audio model. This includes loading the RAVDESS audio dataset, configuring the training settings, and running the training script. The system should give real-time feedback on training progress, show relevant metrics, and preserve the trained audio model.

3. Fuse Models for Emotion Detection

Primary Actor: System

Description: To generate the final Emotion Detection Model, the system combines the trained efficient face model with the Wave2Vec2 audio model. This entails using a fusion technique (such as concatenation) to integrate the visual and audio model outputs. The fused model is then retained for use in future emotion prediction challenges.

4. Predict Emotions from Images and Audio

Primary Actor: End User

Description: The end user interacts with the Emotion Detection Model, which predicts emotions based on input data. The user inputs a sequence of facial photos or audio data, which the system processes using trained models. The result is a structured list of projected emotions, along with related likelihood or probability scores.

5. Configure Training Parameters

Primary Actor: Machine Learning Engineer

Description: The machine learning engineer uses the training configuration interface to specify settings for training the efficient face model and the Wave2Vec2 audio model. This includes modifying hyperparameters, choosing the number of training epochs, and determining batch sizes. The system checks the input parameters for accuracy and consistency.

6. Monitor Training Process

Primary Actor: Machine Learning Engineer

Description: The machine learning engineer utilizes the monitoring interface to track the training progress of both the efficient face model and the Wave2Vec2 audio model. The interface shows real-time indicators like loss and accuracy, allowing the user to recognize possible problems, make informed decisions, and modify the training process as needed.

7. Interact with Emotion Detection Model

Primary Actor: End User

Description: The end user communicates with the Emotion Detection Model via the interaction interface. This comprises supplying input data for emotion prediction, obtaining predictions, and analyzing the model's results. The interface should be user-friendly and allow for easy contact with the model.

8. Handle Unexpected Input Gracefully

Primary Actor: System

Description: The system ensures robustness by gently processing unexpected input data during the prediction process. If the input data deviates from the intended format or contains anomalies, the system should handle the errors and offer relevant feedback to the user.

CHAPTER SIX: PROTOTYPE RESULTS

Our model has yielded promising results across various fusion strategies. When examining individual models, Efficient Face demonstrated a testing accuracy of 68.56. The combination of both modalities in Fusion 1 showcased a substantial improvement, achieving a high testing accuracy, suggesting effective fusion of facial and audio features.

Among the explored fusion strategies, Fusion 3 emerged as particularly successful, employing a combination of 1024-dimensional features from both modalities using average pooling. This fusion approach demonstrated a good accuracy in both training and testing datasets. Moreover, deeper fusion strategies, as seen in Fusion 4 and Fusion 5, involving additional pooling layers, exhibited even higher accuracy in validation (71.5%). In contrast, Fusion 2, which utilized fully connected layers for fusion, showed comparatively lower accuracy, highlighting the importance of choosing an effective fusion strategy for multimodal models.

In summary, the multimodal emotion recognition model's success lies in the strategic fusion of facial and audio features. Fusion 3, Fusion 4, and Fusion 5, characterized by pooling operations, showcased superior accuracy in recognizing emotions, indicating the potential for more sophisticated fusion strategies in multimodal models. These findings contribute valuable insights for optimizing multimodal architectures for emotion recognition tasks.

CHAPTER SEVEN: CONCLUDIN PHASE 1

In this project, we extensively investigated and integrated sophisticated neural network architectures for multimodal emotion identification. The visual model used FaceNet to recognise faces and EfficientFace to recognise facial expressions, whilst the audio model used Wav2Vec2 to recognise voice emotions. The fusion of these modalities was accomplished by late fusion, which combined the strengths of visual and aural information to improve emotion identification.

The visual model proved the usefulness of FaceNet in face identification by using a triplet-based loss function for optimum embedding in a small Euclidean space. EfficientFace, a lightweight and durable facial expression recognition network, tackled real-world problems by integrating local-feature extraction and channel-spatial modulation. Wav2Vec2 is an audio model used for acoustic feature extraction.

The fusion of visual and audio modalities was explored through various strategies. Fusion 3, involving average pooling of 1024-dimensional features, emerged as particularly successful with outstanding accuracy in both validation and testing datasets. Deeper fusion strategies in Fusion 4 and Fusion 5, incorporating additional pooling layers, exhibited even higher accuracy. These results emphasize the importance of strategic fusion methods in achieving superior multimodal emotion recognition.

The project's timeline showcased the integration process of FaceNet and EfficientFace within the visual model, highlighting the harmonious fusion of facial features. Similarly, the integration of Wav2Vec2 and RoBERTaBASE within the MMER framework enriched the audio model's ability to understand emotion expressed through speech. The fusion implementation involved late fusion techniques, and the comparison of different fusion methods demonstrated the superiority of strategies emphasizing pooling operations.

To summarise, this effort proposes a strong multimodal emotion identification model that efficiently combines visual and aural data. The great accuracy in emotion detection achieved using multiple fusion techniques demonstrates the promise for real-world applications in understanding and interpreting human emotions. The findings provide useful insights for optimising multimodal systems and pave the path for breakthroughs in emotion identification technologies.

CHAPTER EIGHT: ENHANCEMENTS TO AUDIO-VISUAL EMOTION RECOGNITION MODEL

Introduction

In the realm of emotion recognition, integrating audio and visual data can significantly improve accuracy. This project began with an initial model architecture that used pre-trained audio and visual models, combined via simple fusion methods. Recognizing limitations in this approach, we enhanced the architecture by integrating transformers and additional training for the visual model. This report details the changes made to the model architecture and their impact.

Initial Model Architecture: Audio and Visual Fusion

The initial model leveraged two state-of-the-art pre-trained models:

- Wav2Vec2 for audio processing, which excels in extracting robust audio features.
- EfficientFace for visual processing, known for its effectiveness in facial expression recognition.

These models **were kept frozen during training**, meaning their weights were not updated. Instead, we only trained the fusion layers added to combine outputs from these models. The fusion methods used included:

1. Aggregating outputs from both modalities.
2. Concatenating output representations and employing a linear layer for dimensionality reduction.
3. Average pooling along the temporal dimension.
4. Adding layers with dropout and ELU activation functions.
5. Max pooling along the temporal dimension followed by a linear layer.

Limitations of Initial Model (Challenges)

The initial approach, while robust in its feature extraction, faced significant challenges:

- **Inefficiency in Capturing Temporal Dependencies:** Simple fusion methods struggled to effectively capture the temporal dependencies between audio and visual data, which are crucial for understanding emotions over time.
- **Limited Complex Interaction Modelling:** The model lacked the ability to model complex interactions over time between the modalities, leading to potential inaccuracies in emotion recognition.

Limitations of Initial Model (Issues Encountered)

Several issues became apparent during testing:

- **Handling Temporal Relationships:** The model was inefficient in handling temporal relationships, resulting in a potential loss of important contextual information during the fusion process.
- **Performance on Real-Life Videos:** While the model performed well on the RAVDESS dataset, it did not perform as expected on real-life videos, indicating overfitting and limited generalization capability.

Enhanced Model Architecture with Transformers

To address these limitations, we introduced significant enhancements to the model architecture:

- **Transformer Integration:** Transformers were introduced to capture temporal dependencies and complex interactions more effectively **Figure 15**. This allowed for a more nuanced understanding of the temporal dynamics in audio-visual data.
- **Training EfficientFace:** Unlike the initial model, EfficientFace was now trained alongside the new architecture. Initially pre-trained on images, EfficientFace was fine-tuned for video data to improve its adaptability and performance.
- **Wav2Vec2:** The Wav2Vec2 model was fine tuned to output the same shape of the vector from the visual model **to avoid bias** and to leverage its pre-trained capabilities for robust audio feature extraction, but its output was input directly to the transformers before the fusion process.

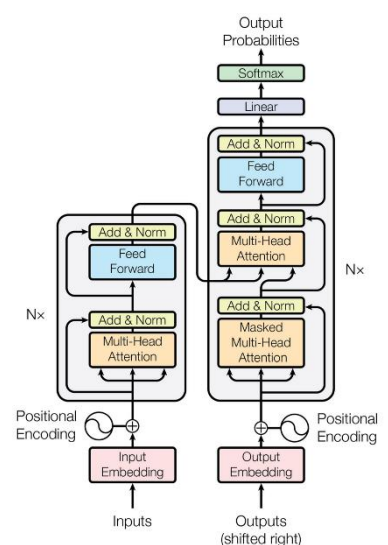


Figure 15 Transformers structure

Key Changes in the Enhanced Model

- Audio Processing: Wav2Vec2 was fine-tuned, focusing on utilizing its pre-trained capabilities for robust audio feature extraction.
- Visual Processing: EfficientFace was trained alongside the new architecture. Initially pre-trained on images, it was now fine-tuned for video data with added **conv1D blocks** to handle temporal data.
- Transformer Integration: **Transformers were used to enhance the fusion process** by capturing temporal dependencies and complex interactions between the audio and visual data.

Training Process

In the enhanced architecture, we implemented a two-phase training process:

1. Training Phase 1: We trained the entire model, including the EfficientFace model and its Conv1D layers. During this phase, the Wav2Vec2 model remained frozen, ensuring that its pre-trained features were effectively utilized without modification.
2. Training Phase 2: The final step involved training everything except the Wav2Vec2 before the projector and classifier. We specifically modified Wav2Vec2's classifier to output a 128-dimensional vector instead of the 8-dimensional vector, matching the output dimension of EfficientFace's Conv1D layers. This alignment facilitated more effective fusion of the two modalities.
3. Training Phase 3: Involved fine tuning the final model with IEMOCAP dataset which introduced data that include noise to be more robust in real life implementations.

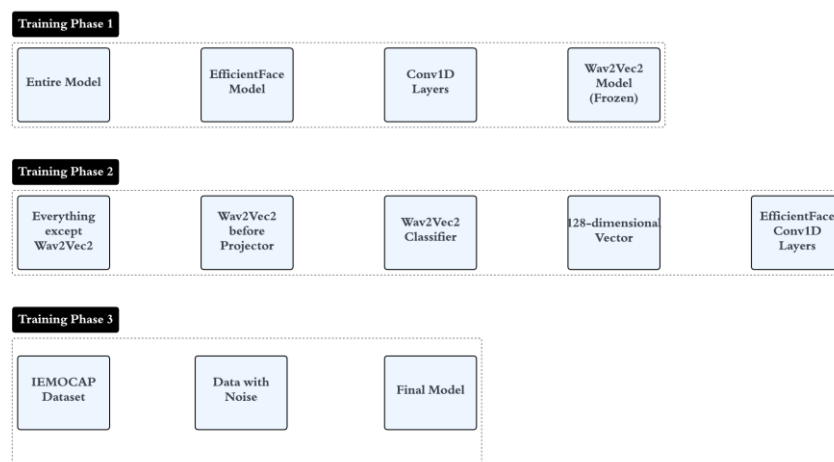


Figure 16 Training Phases

Benefits of the Enhanced Model

- **Improved Temporal Understanding:** The integration of transformers allowed for effective capturing of temporal relationships, leading to better temporal understanding.
- **Enhanced Contextual Modelling:** The model now has a better capability to understand complex interactions between audio and visual data, resulting in more accurate emotion recognition (75% validation accuracy).
- **Robust Emotion Recognition:** These enhancements led to more accurate emotion detection and classification, particularly in real-life video scenarios.

The integration of transformer-based fusion methods significantly improved the temporal and contextual understanding of emotional data. Training the EfficientFace model alongside Wav2Vec2, and adjusting their outputs for better fusion, enhanced the accuracy and reliability of emotion recognition.

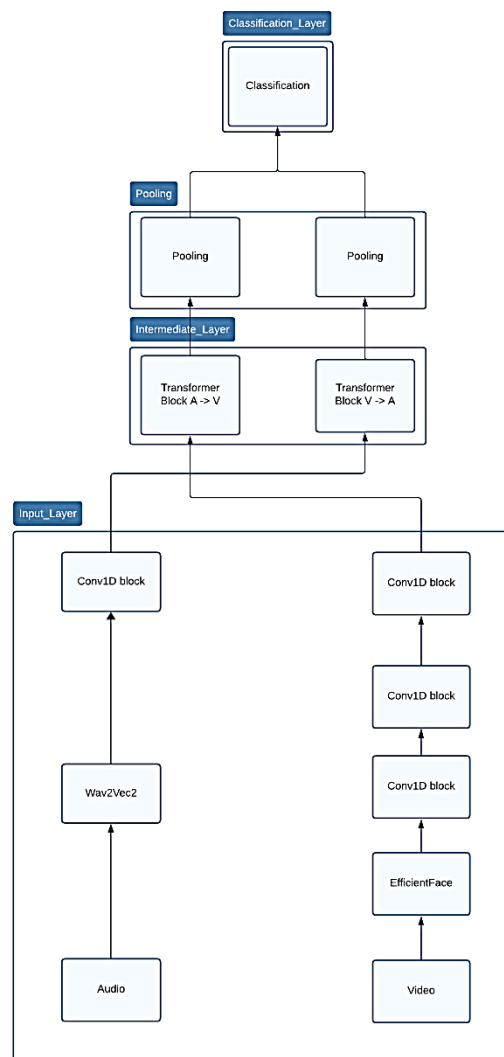


Figure 17 Enhanced Model Arch

CHAPTER NINE: APPLICATION OVERVIEW AND FUNCTIONALITY

Application Concept

The primary goal of this part is to develop an advanced application that aids in diagnosing psychological disorders by LLMs. The application achieves this by analysing user responses to a series of disease-specific questions. It records and analyses both the verbal responses and the emotional states of users, thereby providing comprehensive reports to mental health professionals. This approach not only enhances the diagnostic process but also offers deeper insights into the emotional well-being of individuals.

Workflow Overview

The application's workflow is designed to be intuitive and user-friendly, ensuring a seamless experience for both users and mental health professionals:

1. **User Inputs a Psychological Disease Name:** The process begins with the user entering the name of a psychological disorder they wish to explore. This input sets the stage for the subsequent steps.
2. **Llama AI Generates Questions:** Leveraging Facebook's latest Llama AI model, the application generates a series of tailored questions relevant to the specified disorder. These questions are designed to elicit responses that can provide diagnostic insights.
3. **Questions Presented via TTS and Text:** The generated questions are presented to the user both visually (as text) and audibly (via Text-to-Speech, or TTS). This dual presentation ensures that users receive the questions clearly and can respond in their preferred mode.
4. **User Responds Using Video and Voice:** Users respond to the questions using video, which captures their facial expressions and visual cues, and voice, which captures their spoken responses. This multimodal input is crucial for comprehensive analysis.
5. **STT Captures Responses:** The Speech-to-Text (STT) functionality converts the user's spoken responses into text. This conversion is essential for analysing the content of the responses and ensuring that no detail is missed.
6. **Emotional Analysis of Recorded Responses:** The application conducts an emotional analysis of the recorded responses, examining both the verbal content and the visual emotional cues. This analysis provides deeper insights into the user's emotional state.
7. **AI-Generated Report:** Finally, an AI-generated report is created, synthesizing the user's responses and associated emotions. This report offers a detailed diagnostic overview, aiding mental health professionals in their evaluations.

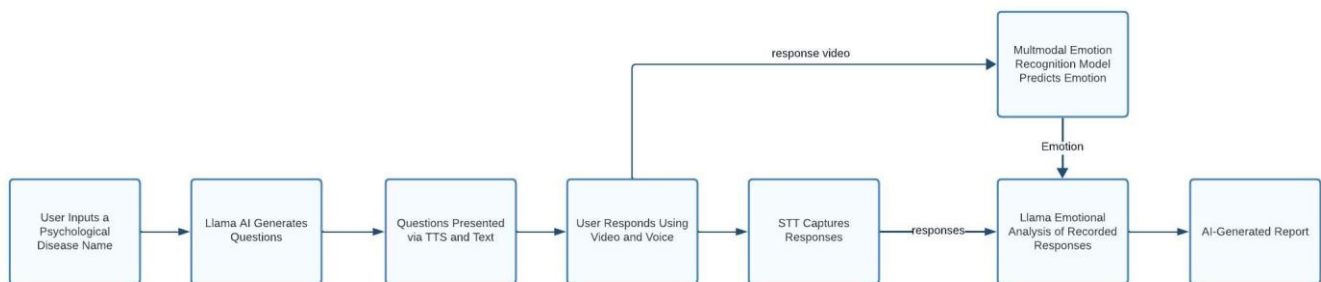


Figure 18 Applications workflow

Speech-to-Text (STT) Functionality

The STT functionality is a critical component of the application, ensuring that user responses are accurately captured and converted into text. This conversion facilitates non-intrusive interaction, where the use of threading allows the STT process to run smoothly in the background. As a result, the user interface remains responsive and fluid, enhancing the overall user experience. High-quality transcription of spoken responses is essential for precise analysis and diagnosis, making the STT process not only a technical necessity but also a key factor in the application's effectiveness.



Figure 19 VOSK STT

In this implementation, the Vosk library is employed to handle the STT tasks. Vosk is an open-source speech recognition toolkit that provides high accuracy and supports multiple languages. The application leverages Vosk's capabilities by loading an appropriate model based on the language requirements, either English or Arabic. The recorded audio is processed through this model to produce text output. By using Vosk, the application benefits from its robust performance and ease of integration, ensuring that the transcription process is both reliable and efficient. Additionally, the conversion of audio files to a standard format (mono PCM) before transcription ensures compatibility and consistency, further contributing to the accuracy of the STT functionality.

The STT functionality is a critical component of the application, ensuring that user responses are accurately captured and converted into text. This conversion facilitates:

- **Non-Intrusive Interaction:** By utilizing threading and multiprocessing, the STT process runs smoothly in the background, ensuring that the user interface remains responsive and fluid.
- **Accurate Text Transcription:** High-quality transcription of spoken responses is essential for precise analysis and diagnosis.

Llama

The Llama model is an integral part of the application's functionality, leveraging advanced language processing capabilities to generate psychological reports and validate inputs. The Llama model, known for its extensive natural language understanding, with **its 70B parameters** it ensures that the questions generated, and the responses analysed are accurate and contextually relevant. This precision is crucial for diagnosing psychological disorders and generating comprehensive reports, enhancing the reliability of the application's assessments.

In practice, the Llama model processes user inputs to validate whether they indicate a psychological disorder, using a pre-trained model to provide accurate Boolean outputs. This validation step is essential for determining the appropriate follow-up questions or actions. The model is also employed to generate a series of questions tailored to different psychological tests. By providing a structured set of questions in JSON format, the Llama model ensures that the user receives a consistent and standardized set of inquiries, facilitating a smooth and efficient assessment process.

Additionally, the Llama model plays a pivotal role in analysing responses and generating detailed psychological reports. The model is used to generate questions and subsequently receive user answers, which are tagged with the corresponding emotions. This feature allows the application to capture nuanced emotional responses, which are crucial for a comprehensive psychological assessment. By interpreting the user's answers and labelling them with corresponding emotions, the model aids in constructing a thorough analysis of the user's psychological state. This report is formatted in markdown, making it easily readable and interpretable by professionals. This comprehensive approach makes the Llama model a cornerstone of the application's functionality, ensuring accurate and reliable psychological assessments.

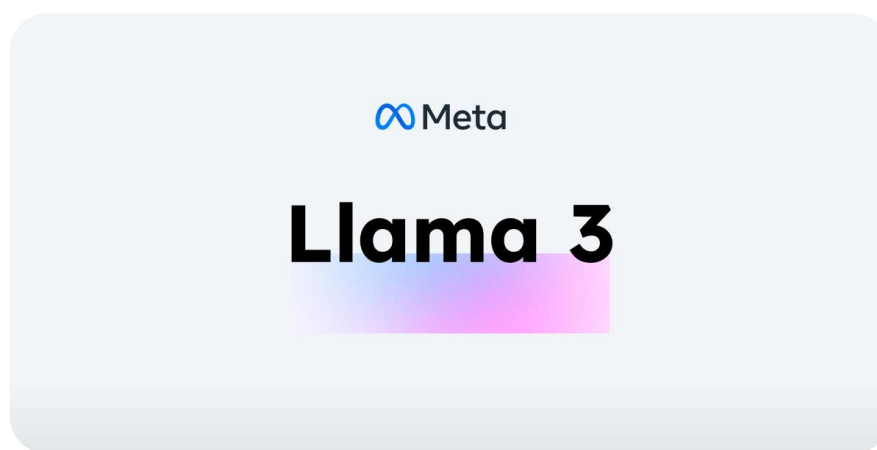


Figure 20 LLama 3 developed by meta

Text-to-Speech (TTS) Functionality

The TTS functionality enhances the user experience by providing a seamless conversational interface. Key aspects include voicing questions, where the TTS system voices the questions generated by the Llama AI, allowing users to hear and understand the questions clearly. This creates a more natural and engaging interaction, encouraging users to respond verbally and more authentically. The TTS system ensures that the communication feels more intuitive and less intrusive, making the overall interaction smoother and more effective.

In our implementation, the Google Text-to-Speech (gTTS) library is used to convert text into spoken words. The system saves the audio files corresponding to the questions in a designated folder and maps phrases to their respective audio files for quick retrieval. The application utilizes the Pygame library to play these audio files, ensuring reliable audio playback. By pre-generating audio files for frequently asked questions, the system can quickly respond to user inputs without delay. Additionally, the ability to generate and play audio dynamically for new phrases ensures that the TTS system is flexible and adaptable to various conversational needs.

The TTS functionality enhances the user experience by providing a seamless conversational interface. Key aspects include:

- **Voicing Questions:** The TTS system voices the questions generated by the Llama AI, allowing users to hear and understand the questions clearly.
- **Conversational Interface:** This creates a more natural and engaging interaction, encouraging users to respond verbally and more authentically.

Comprehensive Diagnostic Insights

The application's combination of STT and TTS functionalities, along with emotional analysis, provides a rich, multifaceted approach to diagnosing psychological disorders. By capturing both the verbal and emotional aspects of user responses, the application offers:

- **Enhanced Diagnostic Accuracy:** The integration of emotional cues with verbal responses leads to more accurate and insightful diagnoses.
- **In-Depth Emotional Analysis:** The detailed emotional analysis helps mental health professionals understand the underlying emotional states of their patients, beyond just their spoken words.
- **User-Friendly Experience:** The application's design ensures that users can interact with it easily and comfortably, making the diagnostic process less intimidating and more accessible.

Emotional Analysis Integration

Emotional Analysis Impact

Integrating emotional analysis into the diagnostic process significantly enhances the application's ability to provide comprehensive insights into a user's mental state. By analyzing the emotional context of user responses in real-time, the application links these emotional cues with each answer, offering deeper diagnostic insights.

This integration demonstrates a marked improvement in diagnostic accuracy compared to assessments that do not consider emotional context. By capturing both the verbal and emotional aspects of user responses, the application provides a richer, more nuanced understanding of the user's mental health, ultimately leading to more accurate and effective diagnoses.

Benefits of Emotional Analysis

The inclusion of emotional analysis in the diagnostic process offers several key benefits. First, it enhances the understanding of a user's mental state during the assessment, providing mental health professionals with a clearer picture of the patient's emotional well-being. Second, it enables the creation of personalized treatment plans based on the emotional nuances observed during the assessment.

This personalization ensures that treatment plans are tailored to the specific emotional needs of the patient, potentially leading to better therapeutic outcomes. Lastly, the integration of emotional analysis highlights the advanced diagnostic capabilities of the application, setting it apart from traditional diagnostic tools that do not consider emotional context.

Effects of Using Emotion in Application

Incorporating emotional analysis into the application has a profound effect on its diagnostic accuracy. A study evaluating the impact of emotion labels on the accuracy of the Llama3 language model (LLM) utilized real cases from "**100 Cases in Psychiatry, Second Edition.**" Emotion labels were derived from the Mental State Examination (MSE) for each case, providing a structured approach to capturing emotional context. This study highlighted the significant benefits of including emotional labels in the diagnostic process, demonstrating that emotion-aware AI can provide more accurate and insightful assessments compared to emotion-neutral approaches.

Emotion Label Effect on LLM Accuracy

The study assessed the impact of emotion labels on the accuracy of the LLM by using historical data from real cases to generate questions and answers. Each set of questions and answers was sent to the LLM twice: once with emotion labels and once without. The true diagnoses were then compared with the LLM's results to measure accuracy.

The findings showed that the accuracy of the LLM increased from **76.66%** without emotion labels to **86.66%** with emotion labels. This significant improvement underscores the value of integrating emotional context into AI-driven diagnostic tools, highlighting how emotion labelling can enhance the accuracy of psychiatric diagnoses.

Impact of Emotion Labelling

The inclusion of emotion labels in the diagnostic process leads to a notable increase in accuracy across selected cases. Specifically, the accuracy of diagnoses made by the LLM increased from 76.66% without emotion labels to 86.66% with emotion labels. This improvement demonstrates that emotion labelling provides nuanced insights into patients' mental states, which are crucial for accurate diagnoses. By capturing the emotional context of user responses, the application can offer a more comprehensive and accurate diagnostic assessment, aiding mental health professionals in making well-informed decisions.

Clinical Implications

The clinical implications of integrating emotion labels into the diagnostic process are substantial. Emotion labels provide nuanced insights into patients' mental states, helping psychiatrists form more comprehensive diagnostic assessments. These insights can also aid in tailoring personalized treatment plans that address the specific emotional needs of patients, potentially leading to better therapeutic outcomes.

By incorporating emotional analysis, the application enhances the ability of mental health professionals to understand and respond to the complex emotional landscapes of their patients, ultimately improving the quality of care.

Future Directions

Looking ahead, there are several promising avenues for further research and development in the integration of emotion labelling in LLM technology. Future studies could explore broader applications of emotion labelling across different models and clinical settings.

Additionally, investigating the impact of emotion-aware AI on therapeutic outcomes could provide valuable insights into the long-term benefits of this technology. Developing guidelines for integrating emotion labels into psychiatric practice will be essential to ensure that these advancements are effectively and ethically applied in clinical settings. By continuing to refine and expand the use of emotion labelling, the field can enhance the precision and effectiveness of the LLM diagnostics, ultimately improving mental health care outcomes.

CHAPTER TEN: EMOTION LABEL EFFECT ON THE LLM ACCURACY

This study investigates the impact of emotion labelling on the performance of the LLM (Llama3) in analyzing real psychiatric cases from the book "100 Cases in Psychiatry, Second Edition" which include expert diagnoses [16]. Each case's History section generates questions and answers, while the Mental State Examination section provides the basis for inferring emotion labels. The study compares the LLM's responses when presented with questions and answers twice: once without emotion labelling and once with emotion labelling. Accuracy is assessed by comparing the LLM's outputs against the true diagnoses.

The utilization of real cases and expert diagnoses ensures the study's robustness in evaluating how emotion labelling influences the LLM's analytical capabilities. By examining both labelled and unlabelled scenarios, the study aims to quantify any improvements or biases introduced by including emotion information. This research not only contributes to understanding the role of emotion in natural language processing tasks but also offers insights into enhancing diagnostic accuracy and interpretability in psychiatric evaluations using advanced language models like Llama3.

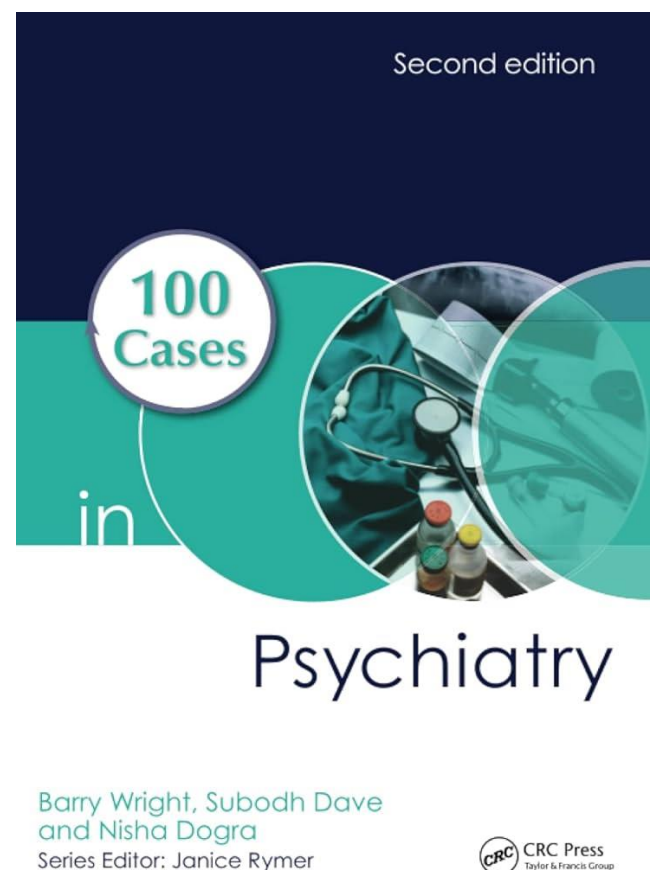


Figure 21 100 Cases in Psychiatry, Second Edition

Case 7

True Diagnoses

This woman is presenting with a mixture of anxiety and depressive symptoms occurring in the context of her son's illness. She is feeling very stressed and has coped well with her husband's support. Detailed history and mental state examination will be needed to establish the diagnosis.

No Emotion

Based on the respondent's answers, it appears that they are experiencing symptoms of anxiety and depression, likely related to the stress of their son's asthma. They are struggling with sleep disturbances, fatigue, demotivation, and changes in mood. However, they have a supportive partner and a healthy coping mechanism (running), which may help them cope with their stress. A diagnosis of Adjustment Disorder with Anxiety and Depressive Symptoms is possible. It is recommended that the respondent seek professional help to manage their symptoms and develop additional coping strategies to address their current situation.

With Emotion

Based on the respondent's answers, it appears that they are experiencing anxiety, worry, and fatigue due to their son's asthma. They are also struggling with motivation, concentration, and mood regulation. While they have a supportive partner and a healthy coping mechanism, they are currently overwhelmed and may benefit from professional help to address their anxiety and potential depression. A thorough diagnostic assessment and treatment plan should be conducted to address these concerns. Additionally, it is essential to monitor their physical health, given the reported fatigue and exhaustion.

Case 15

True Diagnoses

The differential diagnoses are schizophrenia, schizoaffective disorder, drug-related psychosis and organic medical conditions.

No Emotion

Based on the respondent's answers, it appears that they are experiencing a psychotic episode characterized by auditory hallucinations, delusional thinking, and paranoia. The respondent's symptoms are affecting their daily life, including their sleep, behaviour, and decision-making.

A diagnosis of Paranoid Schizophrenia or a related psychotic disorder is likely. Further evaluation and treatment are necessary to address the respondent's symptoms and improve their mental health and well-being.

With Emotion

Based on the responses, it appears that the respondent is experiencing symptoms of a psychotic disorder, likely schizophrenia. The presence of auditory hallucinations, delusions, and paranoia are all indicative of this diagnosis. Additionally, the respondent's sleep disturbances, exhaustion, and feelings of helplessness suggest a significant impact on daily functioning. It is essential to provide the respondent with a comprehensive treatment plan, including medication and therapy, to address these symptoms and improve their overall wellbeing.

Case 27

True Diagnoses

This woman is presenting with posttraumatic stress disorder (PTSD). There is an intense and prolonged reaction to severe trauma.

No Emotion

Based on the respondent's answers, it is clear that they are struggling with PTSD, depression, and suicidal ideation following a traumatic car accident. Their guilt, anxiety, and feelings of worthlessness are contributing to their emotional pain. The self-harming behaviour and substance abuse are coping mechanisms, but they exacerbate their mental health issues. It is essential to provide immediate support and treatment, including therapy and counselling, to help the respondent work through their trauma and develop healthier coping mechanisms.

With Emotion

Based on the respondent's answers, I diagnose post-traumatic stress disorder (PTSD), depression, and anxiety disorder. The self-harming behaviour and suicidal ideation are significant concerns that require immediate attention and treatment. The respondent's history of self-harm and suicide attempts suggests a pattern of emotional dysregulation. Referral to a mental health treatment program, therapy, and support group are highly recommended.

Case 56

True Diagnoses

This stalking behaviour may be due to erotomania (also known as de Clerambault syndrome) in which a person falsely believes, with delusional intensity, that another person is in love with him or her.

No Emotion

Based on the respondent's answers, it appears that they are experiencing symptoms of stalking behaviour, including obsessive attention, harassment, and violent thoughts towards the female student. The respondent's distorted perception of the relationship, social isolation, and lack of friends may be contributing to their behaviour. A comprehensive diagnostic evaluation, including a thorough psychological assessment, is necessary to confirm the diagnosis and develop a treatment plan to address the respondent's behaviour.

With Emotion

Based on the respondent's answers, it appears that they are experiencing chronic fatigue syndrome, which is significantly impacting their daily life and causing emotional distress. The respondent is frustrated, disappointed, and saddened by their situation, and may be experiencing symptoms of depression. The uncertainty and irrational beliefs (e.g., poltergeists) may be a coping mechanism or a manifestation of their emotional state. Further evaluation and therapy are recommended to address the respondent's mental health concerns and provide them with coping strategies to manage their condition.

Case 74

True Diagnoses

This is a description of Capgras syndrome – a rare condition in which the patient believes that an impostor who has taken on a similar appearance has replaced a family member or close acquaintance. Capgras syndrome is more properly referred to as Capgras delusion as it is a particular type of delusional disorder (delusional misrepresentation of personal identity). It is most commonly seen in schizophrenia but a third of cases are thought to be due to organic pathology such as brain injury or dementia.

No Emotion

Based on the respondent's answers, it appears that he may be experiencing a delusional disorder, possibly Capgras syndrome, which involves a false belief that a person, often a family member or spouse, has been replaced by an identical impostor. The respondent's narrative is characterized by a sense of urgency, distress, and fixation on the perceived threat, which may indicate a thought disorder. The respondent's emotional response suggests a significant psychological distress, which should be addressed in conjunction with exploring the underlying causes of his delusional beliefs. A comprehensive diagnostic evaluation and therapeutic intervention are recommended to address the respondent's psychological health and well-being.

With Emotion

Based on the respondent's answers and emotions, it is likely that they are experiencing a delusional disorder, such as Capgras syndrome, which involves the delusional belief that a family member or close friend has been replaced by an identical impostor. This condition is often accompanied by feelings of anxiety, fear, and distress. The respondent's frustration, helplessness, and desperation suggest that they are struggling to cope with the situation and require immediate support and intervention. A thorough psychological evaluation and treatment plan are recommended to address the underlying psychological issues and provide the respondent with the necessary support and guidance.

Conclusion

Case Number	Without Emotion	With Emotion
2	YES	YES
3	YES	YES
4	YES	YES
5	YES	YES
7	YES	YES
12	YES	YES
13	NO	NO
15	YES	YES
16	NO	YES
18	YES	YES
27	YES	YES
37	Partial	YES
45	YES	YES
56	NO	NO
74	YES	YES

Table 3 Results Comparison between using emotions and only text

Without Emotion Accuracy: 76.66%

With Emotion Accuracy: 86.66%

These results show that the LLM accuracy increases when given the emotion.

It was also observed in the case studies that the mood and the emotional state of the patient is vital information for the psychiatrist's diagnoses therefore as an assistive tool the use of the emotion's label is helpful to the LLM and the psychiatrist.

CHAPTER ELEVEN: PROJECT CONCLUSION

Multimodality in machine learning involves the integration and analysis of data from several sources or modalities. This method uses several sources of data, such as text, audio, photos, and video, to build more robust and accurate models. Multimodal machine learning models can collect more information by mixing several modalities, resulting in improved performance and more complete insights. The advantages of multimodality are clear in a variety of applications, including emotion identification and therapeutic treatments.

In our research, we implemented a multimodal emotion recognition model that simultaneously analysed audio and visual data. The audio component analysed features such as pitch, tone, and speech patterns, while the visual component examined facial expressions and movements. By fusing these data streams, the model achieved a higher accuracy in detecting emotions compared to unimodal models. This enhanced capability is particularly useful in applications such as human-computer interaction, where understanding the user's emotional state can lead to more responsive and adaptive systems.

Also in our research, we developed a system that used data from a fused model, integrating audio, visual, and textual information, alongside answers to specific questions. This data was processed using the Llama3 Large Language Model (LLM), which enhanced the system's ability to interpret and respond to the patient's needs. By utilizing multimodal data, the system could consider not only the content of the patient's responses but also their emotional tone and facial expressions.

The results demonstrated that this multimodal approach significantly outperformed traditional methods that relied solely on verbal answers. The integrated system provided therapists with richer insights, enabling more personalized and effective interventions. Patients responded positively, reporting that the system's nuanced understanding of their emotions and behaviours led to more meaningful therapeutic interactions.

To summarise, multimodality in machine learning provides considerable benefits by combining many data sources to produce more accurate, complete, and effective models. The examples of multimodal emotion identification and better therapeutic treatments demonstrate this approach's transformational potential, paving the door for more sophisticated and human-centric artificial intelligence systems.

PHASE 2 SUMMARY

Problem Statement:

This phase focused on enhancing our audio-visual emotion recognition model and developing an advanced application for diagnosing psychological disorders.

1. **Enhanced Model Architecture:** The initial model fused audio and visual data using simple methods but faced challenges in capturing temporal dependencies and complex interactions. The enhanced model integrated transformers, trained EfficientFace for video data, and fine-tuned Wav2Vec2 for robust audio feature extraction. This led to improved temporal understanding, enhanced contextual modelling, and robust emotion recognition, particularly in real-life video scenarios.
2. **Application Implementation:** The application leverages AI and machine learning to analyse user responses to disease-specific questions, recording both verbal and emotional states. It generates tailored questions, presents them via Text-to-Speech (TTS), captures responses using Speech-to-Text (STT), conducts emotional analysis, and generates AI-driven reports for mental health professionals.
 - **STT Functionality:** Critical for accurate transcription of user responses.
 - **Llama AI:** Generates tailored questions and validates inputs for psychological assessments.
 - **TTS Functionality:** Enhances user experience by voicing questions and creating a conversational interface.
 - **Emotional Analysis:** Integrates emotional context with verbal responses for more accurate diagnoses.
3. **Analysis of Impact:** The inclusion of emotional analysis significantly enhances diagnostic accuracy by providing nuanced insights into users' mental states. Studies showed a notable increase in accuracy (from 76.66% to 86.66%) when emotion labels were considered in the diagnostic process, highlighting the value of emotion-aware AI in psychiatric evaluations.

PROJECT SUMMARY

Problem Statement:

Multimodality in machine learning refers to the integration and analysis of data from several sources or modalities. This strategy builds more robust and accurate models by combining data from several sources, including text, audio, images, and video. Multimodal machine learning models may capture more information by combining several modalities, resulting in better performance and more comprehensive insights. The benefits of multimodality are obvious in a range of applications, including emotion recognition and therapeutic treatments.

Methodology:

The methodology employed in our project centres around the implementation of fusion strategies, with a focus on late fusion. Late fusion in method 1 was chosen for its efficacy and reduced training requirements, plays a pivotal role in aggregating diverse outputs into unified emotion predictions. Intermediate fusion was used in methods 2-5 for to achieve better performance using feature maps for both models. Four distinctive methods were developed, each tailored to capture temporal patterns and optimize feature combination for enhanced emotion prediction. Method 1 involves consolidating 16 outputs (8 from each model) to produce 8 final predictions, serving as a baseline. Method 2 employs a more complex architecture by concatenating output representations, omitting the last layer of each model and utilizing a linear layer for dimension reduction. Method 3, similar to Method 2, introduces average pooling along the temporal dimension. Method 4 enhances fusion by introducing more complexity, incorporating dropout for regularization and an ELU activation function. Method 5, similar to Method 4, adds complexity with max pooling instead of average pooling. These fusion strategies aim to address the limitations of unimodal approaches and contribute to the advancement of emotion detection technologies. Accuracies will be discussed in the results section. The early model merged auditory and visual input using basic approaches but struggled to capture temporal interdependence and complicated interconnections. The improved model included transformers, trained EfficientFace for video data, and fine-tuned Wav2Vec2 for reliable audio feature extraction. This resulted in increased temporal comprehension, contextual modelling, and strong emotion identification, particularly in real-world video contexts.

REFERENCES

- [1] B. H. M. H. M. Ali Mollahosseini, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild, arxiv.org, 2017.
- [2] P. T. C. W. A. S. Y. T. P. I. A. M. C. L. D. K. Prannay Khosla, Supervised Contrastive Learning, arxiv.org, 2021.
- [3] H. Wang, X. Li, Z. Ren, M. Wang and C. Ma, Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion., 2023.
- [4] S. & A. A. & M. M. & D. S. Boulahia, Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition, 2021.
- [5] U. T. S. R. H. S. D. M. Sreyan Ghosh, MMER: Multimodal Multi-task Learning for Speech Emotion Recognition, arxiv.org, 2022.
- [6] J. B. L. S. Kilian Q. Weinberger, Distance metric learning for large margin nearest neighbor classification, MIT Press, 2006.
- [7] D. K. J. P. Florian Schroff, FaceNet: A Unified Embedding for Face Recognition and Clustering, arxiv.org, 2015.
- [8] R. F. Matthew D Zeiler, Visualizing and understanding Convolutional Networks, arxiv.org, 2013.
- [9] H. Kim, J.-H. Lee and B. C. Ko, Facial Expression Recognition in the Wild Using Face Graph and Attention, ieeexplore.ieee.org, 2023.
- [10] Q. L. F. Z. Zengqun Zhao, Robust Lightweight Facial Expression Recognition Network with Label Distribution Training, aaai.org, 2021.
- [11] П. C. Anatolij Sergiyenko, Local Feature extraction in images, <http://itvisnyk.kpi.ua/>, 2021.
- [12] A. H. M. Z. A. Z. L.-C. C. Mark Sandler, MobileNetV2: Inverted Residuals and Linear Bottleneck, arxiv.org, 2018.
- [13] H. Z. A. M. M. A. Alexei Baevski, wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, arxiv.org, 2020.
- [14] U. T. S. R. H. S. D. M. Sreyan Ghosh, MMER: Multimodal Multi-task Learning for Speech Emotion Recognition, 2023: arxiv.org.
- [15] M. N. J. M. D. O. M. L. V. YinhanLiu, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arxiv.org, 2019.
- [16] S. D. a. N. D. Barry Wright, 100 Cases in Psychiatry, Second Edition, CRC press, 2017.

EMOTION RECOGNITION

Multimodal emotion recognition and Real-application Implementation

RECAP

Introduction to Emotion Recognition

Enhancing AI with Emotional Intelligence:

- Emotion detection is pivotal for AI integration in daily life.
- Enhances user experience and fosters empathetic, context-aware applications.
- Critical for customer service, mental health monitoring, and other applications.

Challenges of Unimodal Emotion Recognition

Limitations of Single-Modality Approaches:

- Unimodal approaches rely on either facial expressions or audio cues.
- Single modality fails to capture the complex nature of human emotions.
- Emotions involve a combination of visual, auditory, and contextual information.

Importance of Multimodal Approaches

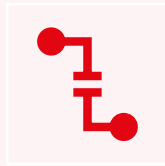
Integrating Visual and Audio Models :

- Addresses limitations of unimodal systems.
- Integrating visual and audio improves accuracy and robustness.
- Enhances reliability for practical applications.

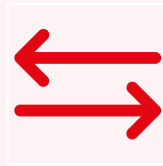
Fusion Techniques



Early Fusion:
Combines modalities
at the feature
extraction stage.



Late Fusion: Combines
outputs of individual
models at a later
stage.



Intermediate Fusion:
Cross-modal attention
mechanisms for better
interaction between
modalities.

DATASETS UTILIZED

- IEMOCAP: Speech and facial expressions, 151 dialogues with nine distinct emotions.
- RAVDESS: 7,356 files of emotional speech and song by 24 actors.
- MELD: 13,118 utterances from 1,433 dialogues, annotated with emotion labels and sentiment scores.

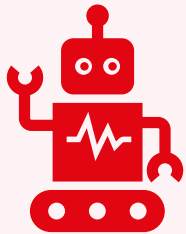
IEMOCAP



RAVDESS



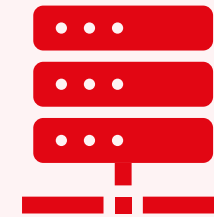
EfficientFace for Visual Emotion Recognition



Lightweight facial
expression recognition
network.



Parameter-efficient with
high accuracy.



Uses local-feature
extractor and channel-
spatial modulator.

Wav2Vec2.0 for Acoustic Emotion Recognition

- Acoustic encoder leveraging self-supervised learning.
- Processes raw audio data to extract features and understand context.
- Effective for emotion recognition from speech.

ROBERTA FOR CONTEXTUAL UNDERSTANDING

Enhancing Emotional Nuances:

- Robust variant of BERT for natural language processing.
- Comprehends contextual meaning of words in sentences.
- Enhances model's ability to capture emotional nuances in speech.

Solution Architecture

Integration of Visual and Audio Models:

Integrates preexisting visual and audio models using late, intermediate fusion techniques.

Combines strengths of both modalities.

Enhances overall model performance.

Fusion Methodologies

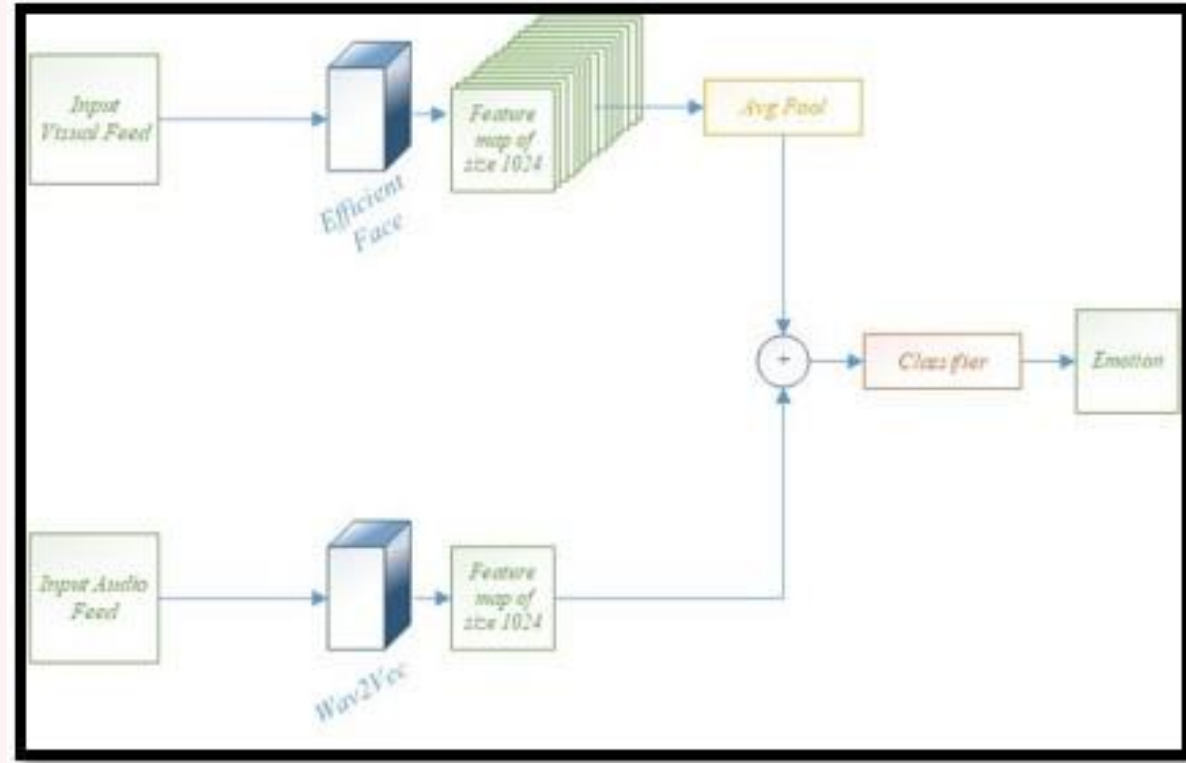
Strategies for Improved Accuracy :

1. Aggregates outputs from both modalities.
2. Concatenates output representations, employs linear layer for dimensionality reduction.
3. Average pooling along the temporal dimension.
4. Additional layers, dropout, ELU activation functions.
5. Max pooling along the temporal dimension followed by a linear layer.

Experiments and Analysis

Architecture:

- Additional pooling layers and regularization techniques further improved accuracy.
- Highlights importance of strategic fusion methods.

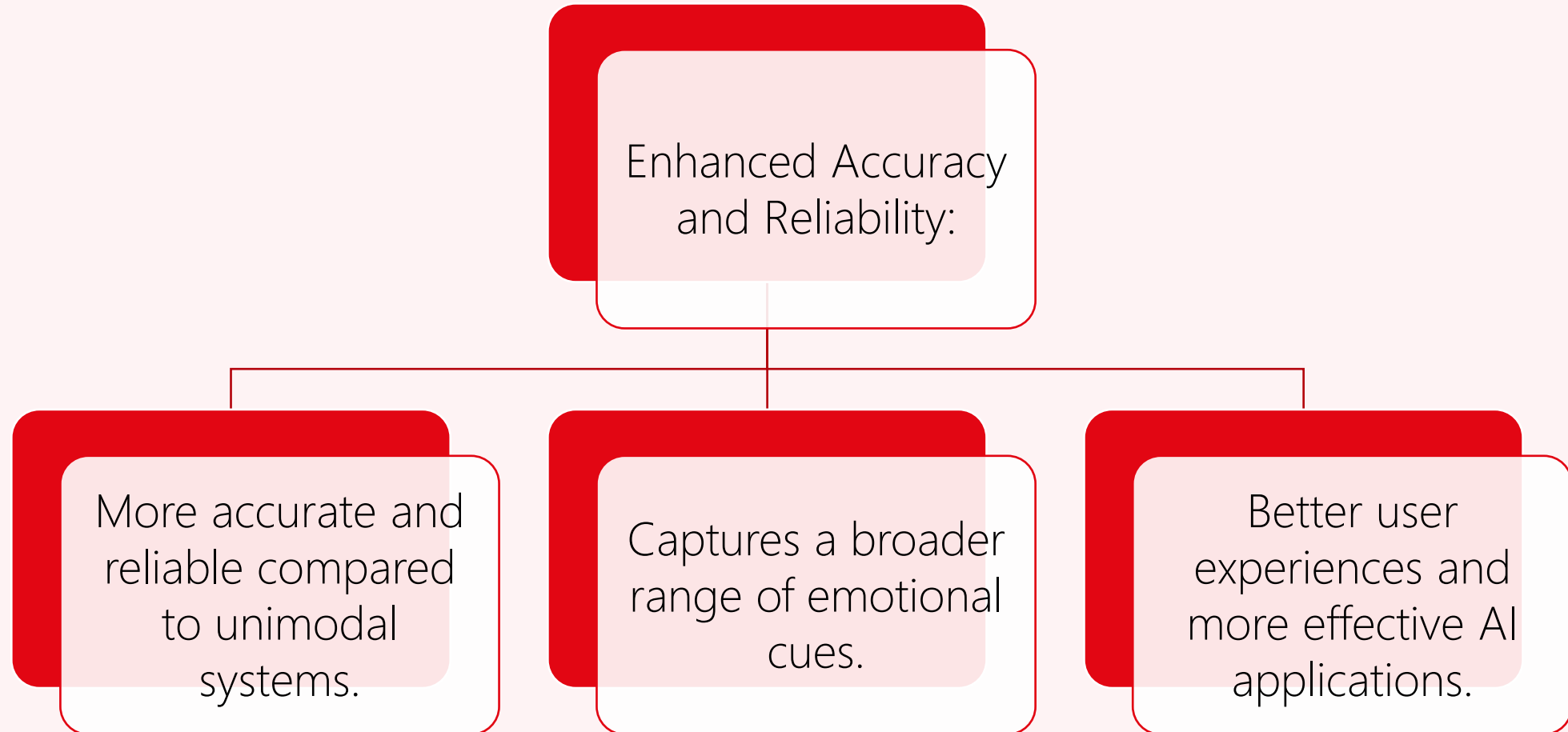


Efficiency and Robustness of the Model

Effective Under Various Conditions :

- EfficientFace and Wav2Vec2.0 ensure model efficiency and robustness.
- Recognizes emotions under various conditions.
- RoBERTa adds contextual understanding for enhanced performance.

Advantages of Multimodal Emotion Recognition



Applications and Practical Implications

Broad Range of Uses :

- Customer service, mental health monitoring, human-computer interaction.
- Provides empathetic and contextually aware responses.
- Improves overall user satisfaction.

Future Directions in Multimodal Emotion Recognition

Expanding Datasets and Fusion Techniques :

- Focus on expanding datasets and improving fusion techniques.
- Capture more subtle and complex emotional cues.
- Explore real-time applications and scalability

Model architecture

Initial Model Architecture: Audio and Visual Fusion

- The initial model architecture utilized two pre-trained models:
 - Wav2Vec2 for audio processing.
 - EfficientFace for visual (facial expression) processing.
- Both models were frozen during training (weights were not updated).
- Only trained the added fusion layers.

Limitations of Initial Model (Challenges)

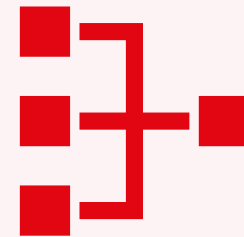
- Simple fusion methods may not effectively capture temporal dependencies between audio and visual data.
- Limited ability to model complex interactions over time between the modalities.
- Overfitting risks on RAVDESS dataset.

Limitations of Initial Model (Issues Encountered)

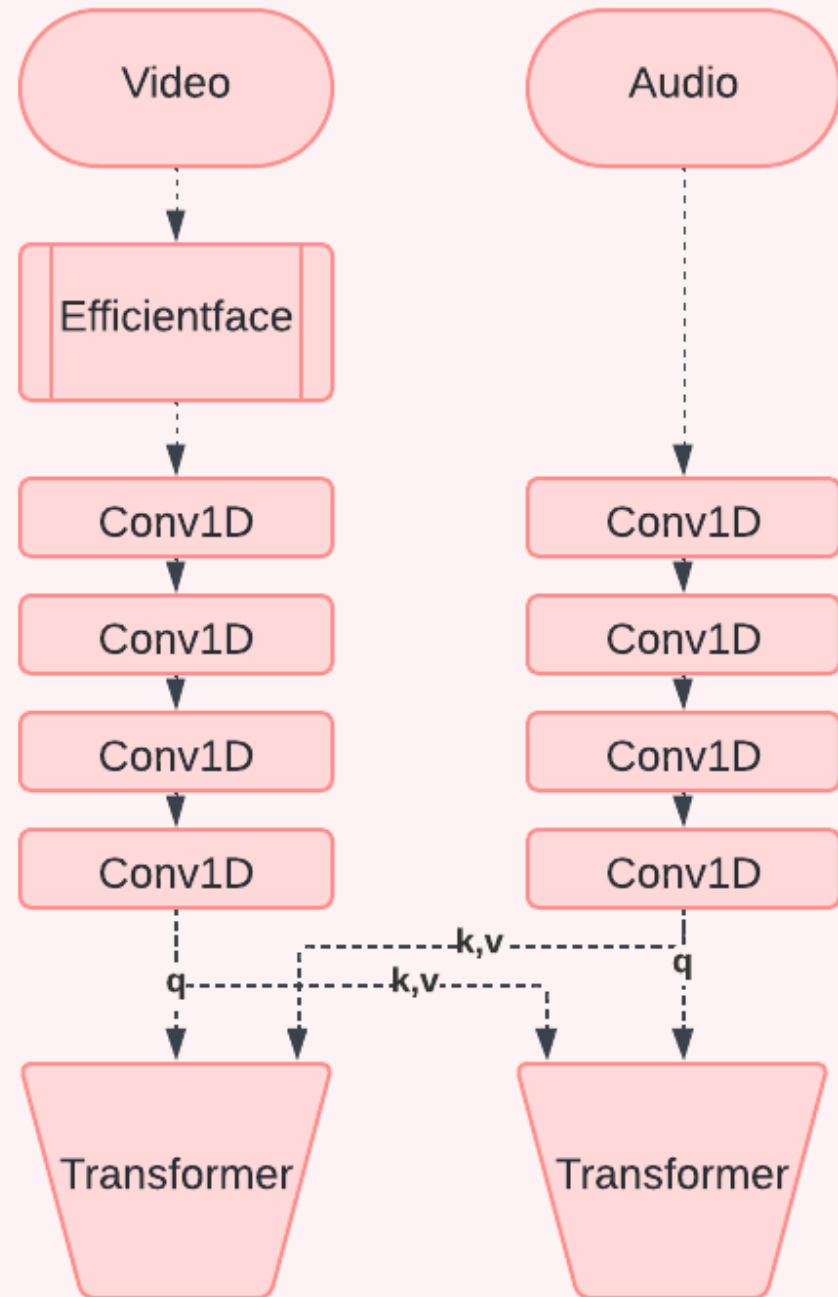
- Inefficiency in handling temporal relationships.
- Potential loss of important contextual information during fusion.
- Did not perform as expected on real life videos

Enhanced Model Architecture with Transformers

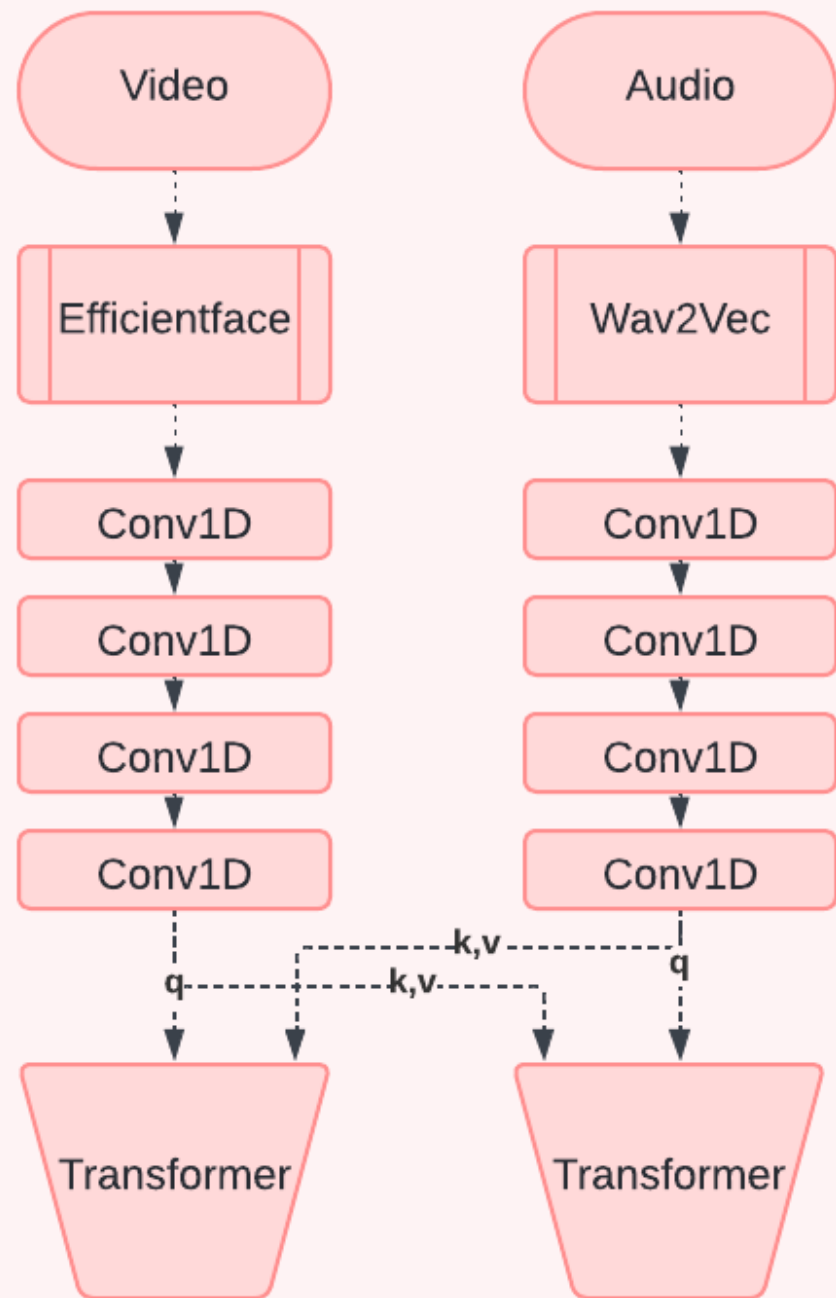
- Introduced transformers to better capture temporal dependencies and complex interactions.
- Enhanced fusion of audio and visual data by training the EfficientFace model along with additional layers, while keeping Wav2Vec2 frozen.



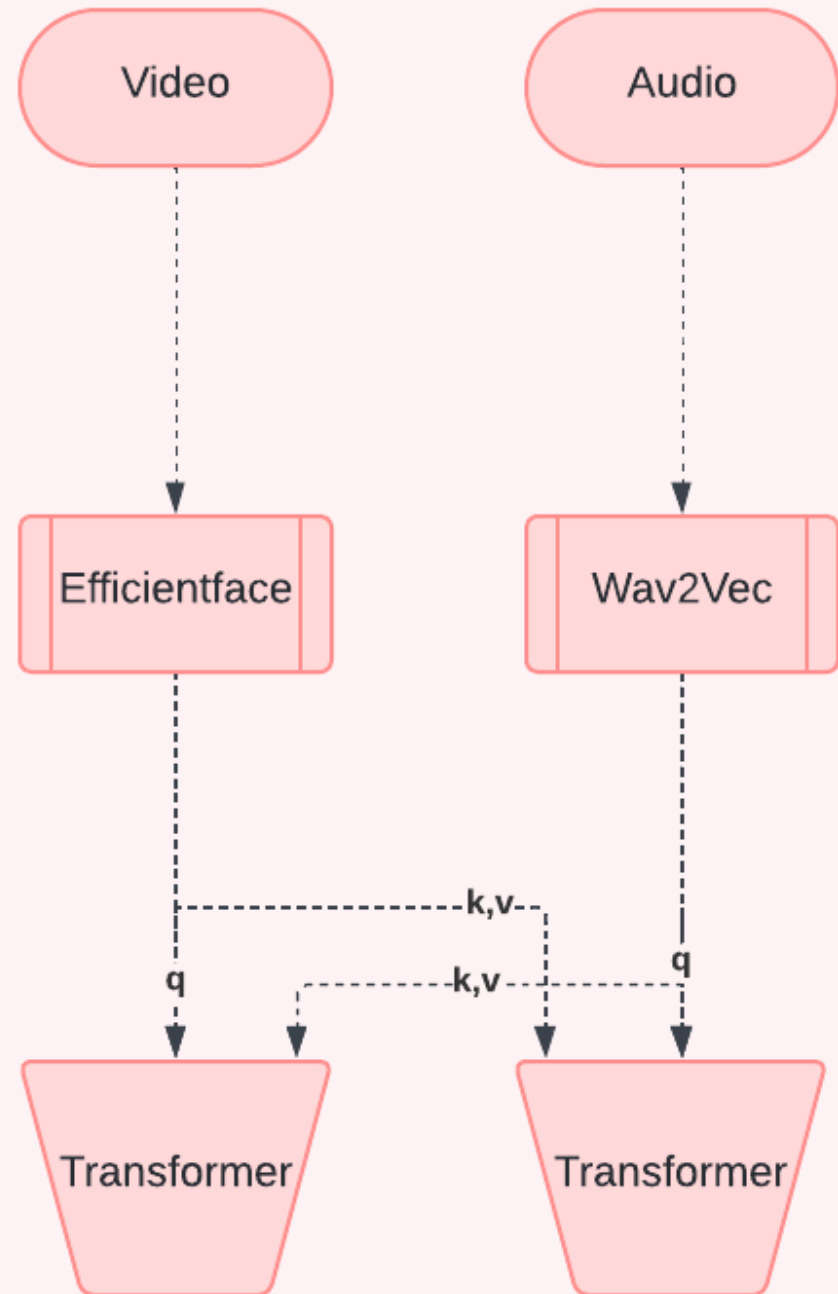
Enhanced Model Architecture with Transformers



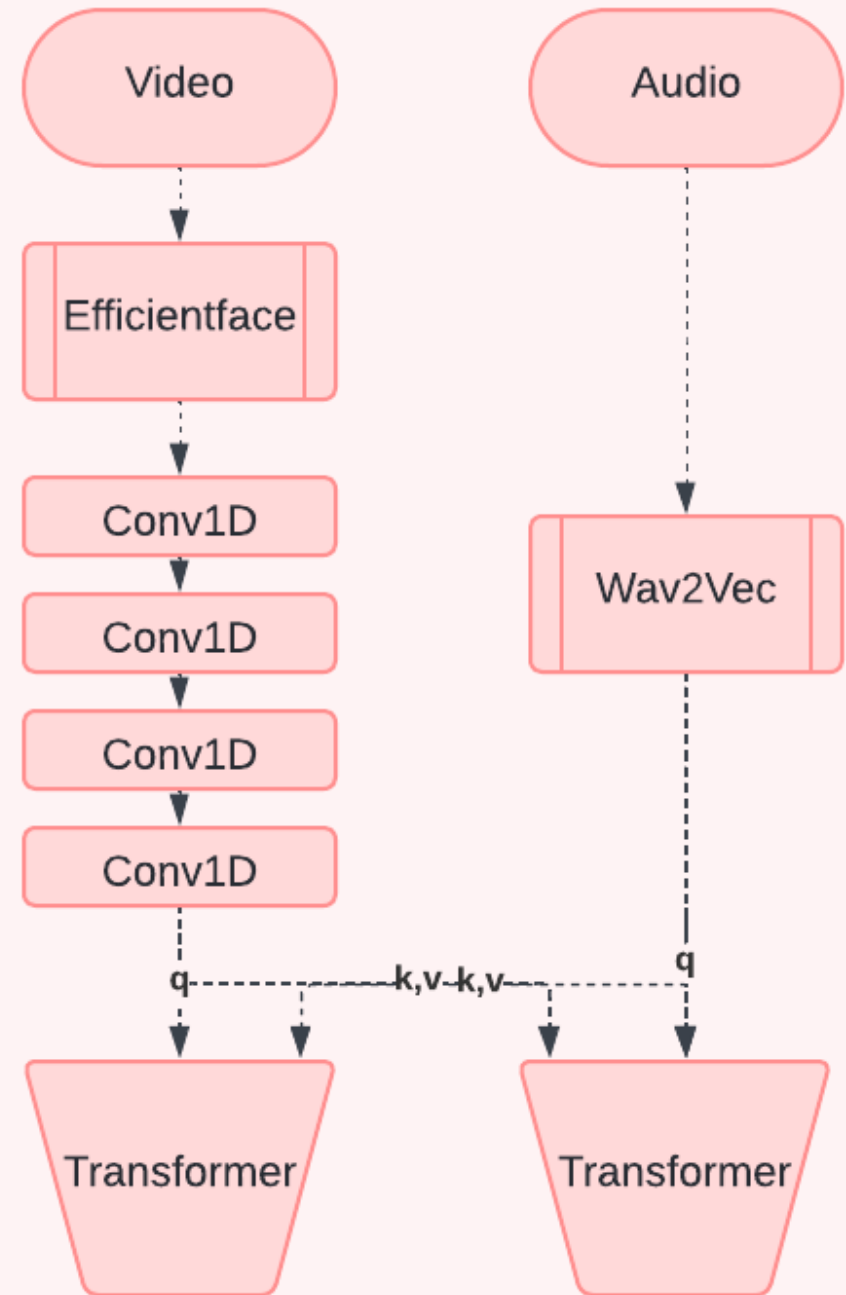
Enhanced Model Architecture with Transformers



Enhanced Model Architecture with Transformers



Enhanced Model Architecture with Transformers



Key Changes



Audio Processing: Wav2Vec2 model remains frozen while focusing on leveraging pre-trained capabilities for robust audio feature extraction.



Visual Processing: EfficientFace model is now trained alongside the new architecture. Initially pre-trained on images, now fine-tuned for video data.



Transformer Integration: Transformers are used to capture temporal dependencies and enhance fusion.

Key Changes



Fine Tuning: Wav2Vec2 model was fine tuned by pruning the last layer to match the 128-dimension visual model.



Visual Model: Added multiple stages of conv1D at the end of the model.

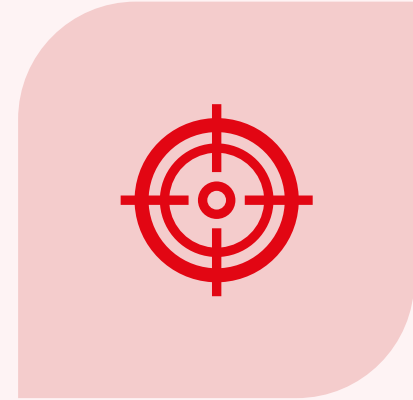
Benefits



IMPROVED TEMPORAL UNDERSTANDING:
CAPTURES TEMPORAL RELATIONSHIPS
EFFECTIVELY.



ENHANCED CONTEXTUAL MODELING: BETTER
UNDERSTANDING OF COMPLEX INTERACTIONS
BETWEEN AUDIO AND VISUAL DATA.



ROBUST EMOTION RECOGNITION: MORE
ACCURATE EMOTION DETECTION AND
CLASSIFICATION.

The background features a light gray circle with a thin white border, centered on a white background. The word "APPLICATION" is written in bold red capital letters across the center of the circle.

APPLICATION

Application Concept



This project aims to develop an application that assists in diagnosing psychological disorders.



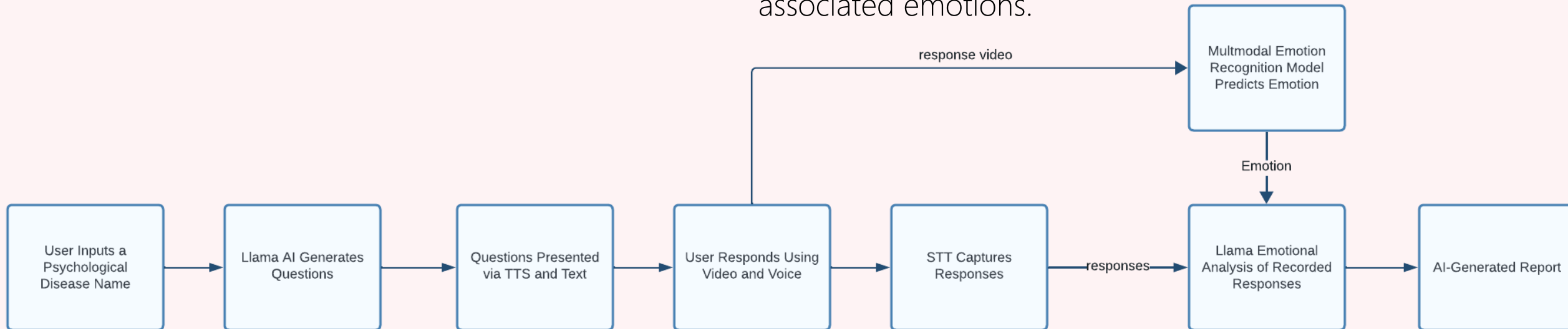
Achieved through analyzing user responses to disease specific questions.



The application will record and analyze both the verbal responses and emotional states of the users, providing comprehensive reports to aid mental health professionals.

Workflow Overview

1. User inputs a psychological disease name.
2. Llama AI generates questions.
3. Questions are presented via TTS and text shown to user.
4. User responds using video (for visual cues) and voice.
5. STT captures responses.
6. Emotional analysis obtained from recorded responses enhances diagnostic insights.
7. AI generated report using responses and associated emotions.



Speech-to-Text (STT) Functionality

- Crucial for the application to capture user responses.
- Converts user's spoken responses into text.
- Facilitates non-intrusive interaction (threaded/multiprocessing) to ensure smooth user experience without UI halts.



Text-to-Speech (TTS) Functionality

- Voices questions generated by Llama AI to the user.
- Provides a seamless conversational interface.
- Essential for users responding to questions verbally.

Large Language Model LLM. (Llama)

- The Llama model is an integral part of the application's functionality, leveraging advanced language processing capabilities to generate psychological reports and validate inputs. The Llama model, known for its extensive natural language understanding, with its 70B parameters it ensures that the questions generated, and the responses analysed are accurate and contextually relevant.
- This precision is crucial for diagnosing psychological disorders and generating comprehensive reports, enhancing the reliability of the application's assessments.



Emotional Analysis Integration

- Emotional analysis impact:
 - Emotional context of user responses analyzed in real-time.
 - Linked with each answer to provide deeper diagnostic insights.
 - Demonstrates improved diagnostic accuracy compared to emotion-neutral assessments.

Benefits of Emotional Analysis



ENHANCES UNDERSTANDING OF USER'S
MENTAL STATE DURING ASSESSMENT.



ENABLES PERSONALIZED TREATMENT
PLANS BASED ON EMOTIONAL
NUANCES.



DEMONSTRATES THE APPLICATION'S
ADVANCED DIAGNOSTIC CAPABILITIES.

Prompt

```
"role": "system",  
  "content": f""Imagine you are a psychiatrist me with a set of {number_of_questions}  
questions as a {questions_type} test to be given they are open questions where the respondent will  
answer by talking  
Provide the response in JSON format only and nothing outside the JSON format.  
Use this as an example:  
{{  
  "Question1": {{  
    "Question": "Your generated Question 1"  
  }},  
  "Question2": {{  
    "Question": "Your generated Question 2"  
  }}  
}}""
```


Prompt

```
"role": "system",
  "content": ""You will be provided with a series of questions and answers with the
respondent's emotion as a label.
Your job is to generate a report on the respondent's psychological health.
The report needs to be structured in markdown.
Make the report structured like this for each question and end it with a conclusion.
# Psychological Report
## Question1
- Question:provided question
- Answer:provided answer
- Emotion:provided emotion
- Analysis: Your analysis

## Conclusion: your diagnoses and remarks

"""
```

Effects of using emotion

Emotion Label effect on the LLM Accuracy



Study evaluating the impact of emotion labels on LLM (Llama3) accuracy.



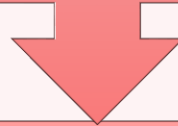
Utilized real cases from "100 Cases in Psychiatry, Second Edition."



Emotion labels derived from Mental State Examination (MSE) for each case.

Methodology Overview

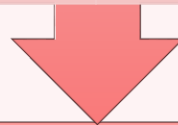
Used history from real cases to generate questions and answers.



Questions/answers sent to LLM twice.

once with emotion labels

once without emotion labels



True diagnoses compared with LLM results to measure accuracy.

Impact of Emotion Labeling

- Accuracy comparison across selected cases.
 - Without Emotion Accuracy: 76.66%
 - With Emotion Accuracy: 86.66%
- Concludes that emotion labeling enhances LLM accuracy in psychiatric diagnoses.

Example case (case 15)

- True Diagnoses:** The differential diagnoses include schizophrenia, schizoaffective disorder, drug-induced psychosis, and organic medical conditions.

- No Emotion Label:** Based on responses, the respondent shows symptoms of a psychotic episode with auditory hallucinations, delusional thinking, and paranoia. These symptoms significantly impact daily life, including sleep, behavior, and decision-making, suggesting a likely diagnosis of Paranoid Schizophrenia or a related disorder.

- With Emotion Label:** Symptoms include auditory hallucinations, delusions, and paranoia, indicative of schizophrenia. Additional issues such as sleep disturbances, exhaustion, and feelings of helplessness further underscore the severity of their condition, emphasizing the need for a comprehensive treatment plan involving medication and therapy to improve overall well-being.

Diagnostic Accuracy comparison

Case Number	Without Emotion	With Emotion
2	YES	YES
7	YES	YES
16	NO	YES
45	YES	YES
56	NO	NO
74	YES	YES

Clinical Implications and Future Directions

Clinical Implications

- Emotion labels provide nuanced insights into patients' mental states.
- Assist psychiatrists in forming comprehensive diagnostic assessments.
- Aid in tailoring personalized treatment plans based on emotional nuances.

Future Directions

- Explore broader applications of emotion labeling in LLM technology.
- Investigate the impact of emotion-aware AI on therapeutic outcomes.
- Develop guidelines for integrating emotion labels into psychiatric practice.

Further Improvements



Exploring Dynamic Fusion techniques



Introduce new modalities to the system



Explore more applications that can benefit from multimodality

Thank you !